Version 14

Fitting Linear Models

"The real voyage of discovery consists not in seeking new landscapes, but in having new eyes."

Marcel Proust

JMP, A Business Unit of SAS
SAS Campus Drive
Cary, NC 27513

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2018. *JMP® 14 Fitting Linear Models*. Cary, NC: SAS Institute Inc.

JMP® 14 Fitting Linear Models

ISBN 978-1-63526-509-5 (Hardcopy)
ISBN 978-1-63526-510-1 (EPUB)
ISBN 978-1-63526-511-8 (MOBI)
ISBN 978-1-63526-512-5 (Web PDF)

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513-2414.

March 2018

Other brand and product names are trademarks of their respective companies.

SAS software may be provided with certain third-party software, including but not limited to open-source software, which is licensed under its applicable third-party software license agreement. For license information about third-party software distributed with SAS software, refer to http://support.sas.com/thirdpartylicenses.

Technology License Notices

- Scintilla - Copyright © 1998-2017 by Neil Hodgson <neilh@scintilla.org>.

  All Rights Reserved.

  Permission to use, copy, modify, and distribute this software and its documentation for any purpose and without fee is hereby granted, provided that the above copyright notice appear in all copies and that both that copyright notice and this permission notice appear in supporting documentation.

  NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.
- Telerik RadControls: Copyright © 2002-2012, Telerik. Usage of the included Telerik RadControls outside of JMP is not permitted.
- ZLIB Compression Library - Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.
- Packages - Copyright © 2009-2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  Neither the name of the WhiteBox nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- iODBC software - Copyright © 1995-2006, OpenLink Software Inc and Ke Jin (www.iodbc.org). All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

  – Neither the name of OpenLink Software Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL OPENLINK OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- bzip2, the associated library "libbzip2", and all documentation, are Copyright © 1996-2010, Julian R Seward. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.

Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

The name of the author may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHOR "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE AUTHOR BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- R software is Copyright © 1999-2012, R Foundation for Statistical Computing.
- MATLAB software is Copyright © 1984-2012, The MathWorks, Inc. Protected by U.S. and international patents. See www.mathworks.com/patents. MATLAB and Simulink are registered trademarks of The MathWorks, Inc. See www.mathworks.com/trademarks for a list of additional trademarks. Other product or brand names may be trademarks or registered trademarks of their respective holders.
- libopc is Copyright © 2011, Florian Reuter. All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  – Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
  – Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.

– Neither the name of Florian Reuter nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

- libxml2 - Except where otherwise noted in the source code (e.g. the files hash.c, list.c and the trio files, which are covered by a similar license but with different Copyright notices) all the files are:

  Copyright © 1998 - 2003 Daniel Veillard. All Rights Reserved.

  Permission is hereby granted, free of charge, to any person obtaining a copy of this software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation the rights to use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so, subject to the following conditions:

  The above copyright notice and this permission notice shall be included in all copies or substantial portions of the Software.

  THE SOFTWARE IS PROVIDED "AS IS", WITHOUT WARRANTY OF ANY KIND, EXPRESS OR IMPLIED, INCLUDING BUT NOT LIMITED TO THE WARRANTIES OF MERCHANTABILITY, FITNESS FOR A PARTICULAR PURPOSE AND NONINFRINGEMENT. IN NO EVENT SHALL DANIEL VEILLARD BE LIABLE FOR ANY CLAIM, DAMAGES OR OTHER LIABILITY, WHETHER IN AN ACTION OF CONTRACT, TORT OR OTHERWISE, ARISING FROM, OUT OF OR IN CONNECTION WITH THE SOFTWARE OR THE USE OR OTHER DEALINGS IN THE SOFTWARE.

  Except as contained in this notice, the name of Daniel Veillard shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization from him.

- Regarding the decompression algorithm used for UNIX files:

  Copyright © 1985, 1986, 1992, 1993

The Regents of the University of California. All rights reserved.

THIS SOFTWARE IS PROVIDED BY THE REGENTS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE REGENTS OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.

3. Neither the name of the University nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

- Snowball - Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton.

  All rights reserved.

  Redistribution and use in source and binary forms, with or without modification, are permitted provided that the following conditions are met:

  1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.

  2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.

  3. Neither the name of the copyright holder nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

  THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS \"AS IS\" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE

Contents

Fitting Linear Models

Chapter 1

Learn about JMP

Documentation and Additional Resources

This chapter includes details about JMP documentation, such as book conventions, descriptions of each JMP book, the Help system, and where to find other support.

Contents

Formatting Conventions

The following conventions help you relate written material to information that you see on your screen:

- Sample data table names, column names, pathnames, filenames, file extensions, and folders appear in Helvetica font.
- Code appears in `Lucida Sans Typewriter` font.
- Code output appears in *`Lucida Sans Typewriter`* italic font and is indented farther than the preceding code.
- **Helvetica bold** formatting indicates items that you select to complete a task:
  - buttons
  - check boxes
  - commands
  - list names that are selectable
  - menus
  - options
  - tab names
  - text boxes
- The following items appear in italics:
  - words or phrases that are important or have definitions specific to JMP
  - book titles
  - variables
- Features that are for JMP Pro only are noted with the JMP Pro icon JMP PRO. For an overview of JMP Pro features, visit https://www.jmp.com/software/pro/.

Note: Special information and limitations appear within a Note.

Tip: Helpful information appears within a Tip.

JMP Documentation

JMP offers documentation in various formats, from print books and Portable Document Format (PDF) to electronic books (e-books).

- Open the PDF versions from the **Help > Books** menu.
- All books are also combined into one PDF file, called *JMP Documentation Library,* for convenient searching. Open the *JMP Documentation Library* PDF file from the **Help > Books** menu.
- You can also purchase printed documentation and e-books on the SAS website:

  https://www.sas.com/store/search.ep?keyWords=JMP

JMP Documentation Library

The following table describes the purpose and content of each book in the JMP library.

Document Title	Document Purpose	Document Content
Discovering JMP	If you are not familiar with JMP, start here.	Introduces you to JMP and gets you started creating and analyzing data. Also learn how to share your results.
Using JMP	Learn about JMP data tables and how to perform basic operations.	Covers general JMP concepts and features that span across all of JMP, including importing data, modifying columns properties, sorting data, and connecting to SAS.

Document Title	Document Purpose	Document Content
Basic Analysis	Perform basic analysis using this document.	Describes these Analyze menu platforms: • Distribution • Fit Y by X • Tabulate • Text Explorer Covers how to perform bivariate, one-way ANOVA, and contingency analyses through Analyze > Fit Y by X. How to approximate sampling distributions using bootstrapping and how to perform parametric resampling with the Simulate platform are also included.
Essential Graphing	Find the ideal graph for your data.	Describes these Graph menu platforms: • Graph Builder • Scatterplot 3D • Contour Plot • Bubble Plot • Parallel Plot • Cell Plot • Scatterplot Matrix • Ternary Plot • Treemap • Chart • Overlay Plot The book also covers how to create background and custom maps.
Profilers	Learn how to use interactive profiling tools, which enable you to view cross-sections of any response surface.	Covers all profilers listed in the Graph menu. Analyzing noise factors is included along with running simulations using random inputs.

Document Title	Document Purpose	Document Content
Design of Experiments Guide	Learn how to design experiments and determine appropriate sample sizes.	Covers all topics in the DOE menu and the Specialized DOE Models menu item in the Analyze > Specialized Modeling menu.
Fitting Linear Models	Learn about Fit Model platform and many of its personalities.	Describes these personalities, all available within the Analyze menu Fit Model platform: • Standard Least Squares • Stepwise • Generalized Regression • Mixed Model • MANOVA • Loglinear Variance • Nominal Logistic • Ordinal Logistic • Generalized Linear Model

Document Title	Document Purpose	Document Content
Predictive and Specialized Modeling	Learn about additional modeling techniques.	Describes these Analyze > Predictive Modeling menu platforms: • Modeling Utilities • Neural • Partition • Bootstrap Forest • Boosted Tree • K Nearest Neighbors • Naive Bayes • Model Comparison • Formula Depot Describes these Analyze > Specialized Modeling menu platforms: • Fit Curve • Nonlinear • Functional Data Explorer • Gaussian Process • Time Series • Matched Pairs Describes these Analyze > Screening menu platforms: • Response Screening • Process Screening • Predictor Screening • Association Analysis • Process History Explorer The platforms in the Analyze > Specialized Modeling > Specialized DOE Models menu are described in *Design of Experiments Guide*.

Document Title	Document Purpose	Document Content
Multivariate Methods	Read about techniques for analyzing several variables simultaneously.	Describes these Analyze > Multivariate Methods menu platforms: • Multivariate • Principal Components • Discriminant • Partial Least Squares • Multiple Correspondence Analysis • Factor Analysis • Multidimensional Scaling • Item Analysis Describes these Analyze > Clustering menu platforms: • Hierarchical Cluster • K Means Cluster • Normal Mixtures • Latent Class Analysis • Cluster Variables
Quality and Process Methods	Read about tools for evaluating and improving processes.	Describes these Analyze > Quality and Process menu platforms: • Control Chart Builder and individual control charts • Measurement Systems Analysis • Variability / Attribute Gauge Charts • Process Capability • Pareto Plot • Diagram • Manage Spec Limits

Document Title	Document Purpose	Document Content
Reliability and Survival Methods	Learn to evaluate and improve reliability in a product or system and analyze survival data for people and products.	Describes these Analyze > Reliability and Survival menu platforms: • Life Distribution • Fit Life by X • Cumulative Damage • Recurrence Analysis • Degradation and Destructive Degradation • Reliability Forecast • Reliability Growth • Reliability Block Diagram • Repairable Systems Simulation • Survival • Fit Parametric Survival • Fit Proportional Hazards
Consumer Research	Learn about methods for studying consumer preferences and using that insight to create better products and services.	Describes these Analyze > Consumer Research menu platforms: • Categorical • Choice • MaxDiff • Uplift • Multiple Factor Analysis
Scripting Guide	Learn about taking advantage of the powerful JMP Scripting Language (JSL).	Covers a variety of topics, such as writing and debugging scripts, manipulating data tables, constructing display boxes, and creating JMP applications.
JSL Syntax Reference	Read about many JSL functions on functions and their arguments, and messages that you send to objects and display boxes.	Includes syntax, examples, and notes for JSL commands.

Note: The **Books** menu also contains two reference cards that can be printed: The *Menu Card* describes JMP menus, and the *Quick Reference* describes JMP keyboard shortcuts.

JMP Help

JMP Help is an abbreviated version of the documentation library that provides targeted information. You can open JMP Help in several ways:

- On Windows, press the F1 key to open the Help system window.
- Get help on a specific part of a data table or report window. Select the Help tool from the **Tools** menu and then click anywhere in a data table or report window to see the Help for that area.
- Within a JMP window, click the **Help** button.
- Search and view JMP Help on Windows using the **Help > Help Contents**, **Search Help**, and **Help Index** options. On Mac, select **Help > JMP Help**.
- Search the Help at https://jmp.com/support/help/ (English only).

Additional Resources for Learning JMP

In addition to JMP documentation and JMP Help, you can also learn about JMP using the following resources:

- "Tutorials"
- "Sample Data Tables"
- "Learn about Statistical and JSL Terms"
- "Learn JMP Tips and Tricks"
- "Tooltips"
- "JMP User Community"
- "JMP Books by Users"
- "The JMP Starter Window"

Tutorials

You can access JMP tutorials by selecting **Help > Tutorials**. The first item on the **Tutorials** menu is **Tutorials Directory**. This opens a new window with all the tutorials grouped by category.

If you are not familiar with JMP, then start with the **Beginners Tutorial**. It steps you through the JMP interface and explains the basics of using JMP.

The rest of the tutorials help you with specific aspects of JMP, such as designing an experiment and comparing a sample mean to a constant.

Sample Data Tables

All of the examples in the JMP documentation suite use sample data. Select **Help > Sample Data Library** to open the sample data directory.

To view an alphabetized list of sample data tables or view sample data within categories, select **Help > Sample Data**.

Sample data tables are installed in the following directory:

On Windows: C:\Program Files\SAS\JMP\14\Samples\Data

On Macintosh: \Library\Application Support\JMP\14\Samples\Data

In JMP Pro, sample data is installed in the JMPPRO (rather than JMP) directory. In JMP Shrinkwrap, sample data is installed in the JMPSW directory.

To view examples using sample data, select **Help > Sample Data** and navigate to the Teaching Resources section. To learn more about the teaching resources, visit http://jmp.com/tools.

Learn about Statistical and JSL Terms

The **Help** menu contains the following indexes:

Statistics Index Provides definitions of statistical terms.

Scripting Index Lets you search for information about JSL functions, objects, and display boxes. You can also edit and run sample scripts from the Scripting Index.

Learn JMP Tips and Tricks

When you first start JMP, you see the Tip of the Day window. This window provides tips for using JMP.

To turn off the Tip of the Day, clear the **Show tips at startup** check box. To view it again, select **Help > Tip of the Day**. Or, you can turn it off using the Preferences window.

Tooltips

JMP provides descriptive tooltips when you place your cursor over items, such as the following:

- Menu or toolbar options

- Labels in graphs
- Text results in the report window (move your cursor in a circle to reveal)
- Files or windows in the Home Window
- Code in the Script Editor

Tip: On Windows, you can hide tooltips in the JMP Preferences. Select **File > Preferences > General** and then deselect **Show menu tips**. This option is not available on Macintosh.

JMP User Community

The JMP User Community provides a range of options to help you learn more about JMP and connect with other JMP users. The learning library of one-page guides, tutorials, and demos is a good place to start. And you can continue your education by registering for a variety of JMP training courses.

Other resources include a discussion forum, sample data and script file exchange, webcasts, and social networking groups.

To access JMP resources on the website, select **Help > JMP User Community** or visit https://community.jmp.com/.

JMP Books by Users

Additional books about using JMP that are written by JMP users are available on the JMP website:

https://www.jmp.com/en_us/software/books.html

The JMP Starter Window

The JMP Starter window is a good place to begin if you are not familiar with JMP or data analysis. Options are categorized and described, and you launch them by clicking a button. The JMP Starter window covers many of the options found in the Analyze, Graph, Tables, and File menus. The window also lists JMP Pro features and platforms.

- To open the JMP Starter window, select **View** (**Window** on the Macintosh) **> JMP Starter**.
- To display the JMP Starter automatically when you open JMP on Windows, select **File > Preferences > General**, and then select **JMP Starter** from the Initial JMP Window list. On Macintosh, select **JMP > Preferences > Initial JMP Starter Window**.

Technical Support

JMP technical support is provided by statisticians and engineers educated in SAS and JMP, many of whom have graduate degrees in statistics or other technical disciplines.

Many technical support options are provided at https://www.jmp.com/support, including the technical support phone number.

Chapter 2

Model Specification

Specify Linear Models

Using the Fit Model platform, you can specify complex models efficiently. Your task is simplified by Macros, Attributes, and transformations. Fit Model is your gateway to fitting a broad variety of models and effect structures.

These include:

- simple and multiple linear regression
- analysis of variance and covariance
- random effect, nested effect, mixed effect, repeated measures, and split plot models
- nominal and ordinal logistic regression
- multivariate analysis of variance (MANOVA)
- canonical correlation and discriminant analysis
- loglinear variance (to model the mean and the variance)
- generalized linear models (GLM)
- parametric survival and proportional hazards
- response screening, for studying a large number of responses

JMP PRO In JMP Pro, you can also fit the following:

- mixed models with a range of covariance structures
- generalized regression models including the elastic net, lasso, and ridge regression
- partial least squares

The Fit Model platform lets you fit a large variety of types of models by selecting the desired personality. This chapter focuses on the elements of the Model Specification window that are common to most personalities.

Contents

Overview of the Fit Model Platform

The Fit Model platform gives you an efficient way to specify models that have complex effect structures. These effect structures are linear in the model parameters. Once you have specified your model, you can select the appropriate fitting technique from a number of fitting *personalities*. Once you choose a personality, the Fit Model window provides choices that are relevant for the chosen personality. This chapter focuses on the elements of the Model Specification window that are common to most personalities. For a description of all personalities, see "Fit Model Launch Window Elements" on page 40.

Fit Model can be used to specify a wide variety of models that can be fit using various methods. Table 2.1 lists some typical models that can be defined using Fit Model. In the table, the effects X and Z represent columns with a continuous modeling type, while A, B, and C represent columns with a nominal or ordinal modeling type.

Refer to the section "Examples of Model Specifications and Their Model Fits" on page 54 to see the clicking sequences that produce these model effects, plots of the model fits, and some examples.

Table 2.1 Standard Model Types

Type of Model	Model Effects
Simple Linear Regression	X
Polynomial in X to Degree k	X, X*X, ..., X^k
Polynomial in X and Z to Degree k	X, X*X, ..., X^k, Z, Z*Z, ..., Z^k
Multiple Linear Regression	X, Z, and other continuous columns
One-Way Analysis of Variance	A
Two-Way Analysis of Variance	A, B
Two-Way Analysis of Variance with Interaction	A, B, A*B
Three-Way Full Factorial	A, B, C, A*B, A*C, B*C, A*B*C
Analysis of Covariance, Equal Slopes	A, X
Analysis of Covariance, Unequal Slopes	A, X, A*X
Two-Factor Nested Random Effects Model	A, B[A]&Random
Three-Factor Fully Nested Random Effects Model	A, B[A]&Random, C[A,B]&Random

Table 2.1 Standard Model Types *(Continued)*

Type of Model	Model Effects
Simple Split Plot or Repeated Measures Model	A, B[A]&Random, C, C*A
Two-Factor Response Surface Model	X&RS, Z&RS, X*X, X*Z, Z*Z

Example of a Regression Analysis Using Fit Model

You have data resulting from an aerobic fitness study, and you want to predict the oxygen uptake from several continuous variables.

1. Select **Help > Sample Data Library** and open Fitness.jmp.
2. Select **Analyze > Fit Model**. Note that the Personality box is empty.
3. Select Oxy and click **Y**.

   When you specify a continuous response, the Personality defaults to Standard Least Squares, but you are free to choose another personality. Also, the Emphasis defaults to Effect Leverage.
4. Press CTRL, and select Sex, Age, Weight, Runtime, RunPulse, RstPulse, and MaxPulse. Click **Add** to add these to the Construct Model Effects list. Note that you can select **Keep dialog open** if you want to have this window available later on. Your Model Specification window should appear as shown in Figure 2.1.
5. Click **Run**. Figure 2.2 gives a partial view of the report.

Figure 2.1 Model Specification Window for Fitness Regression Model

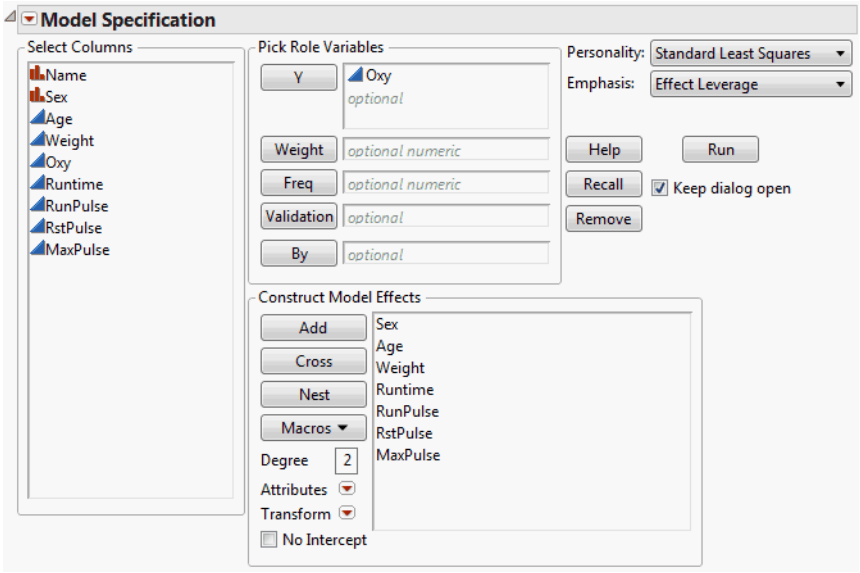
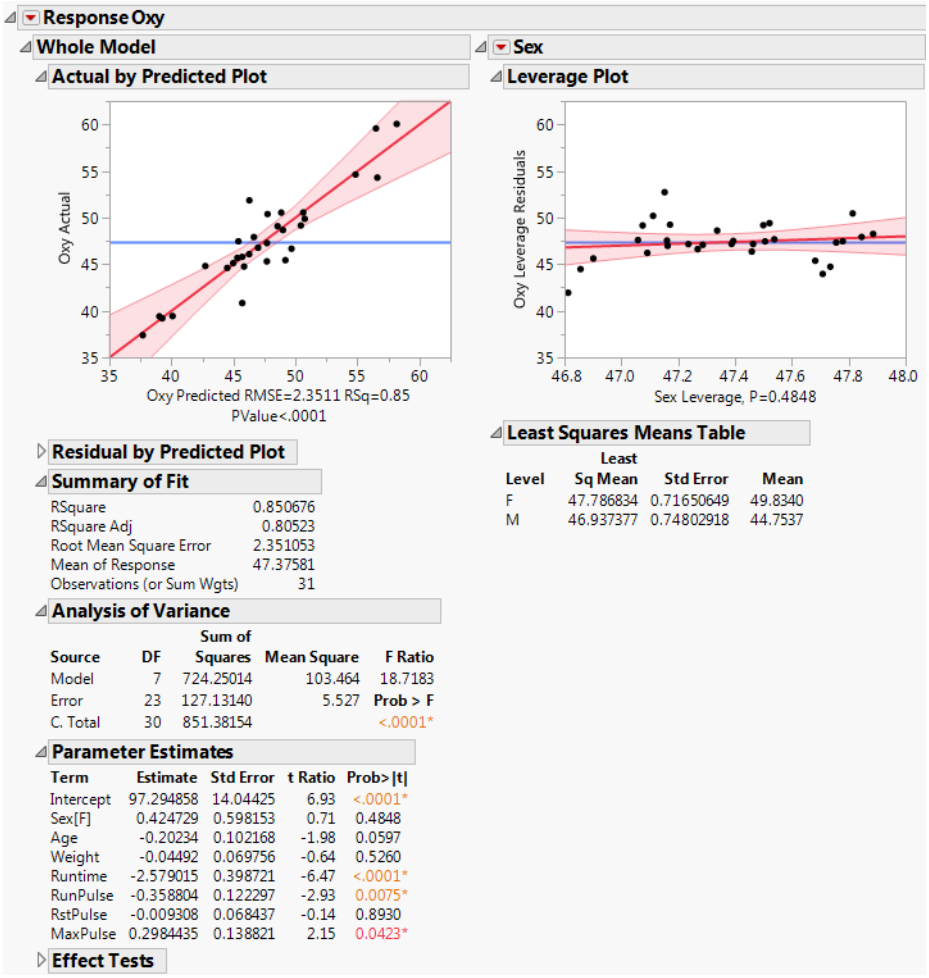


Figure 2.2 Partial View of Standard Least Squares Report for Fitness Data



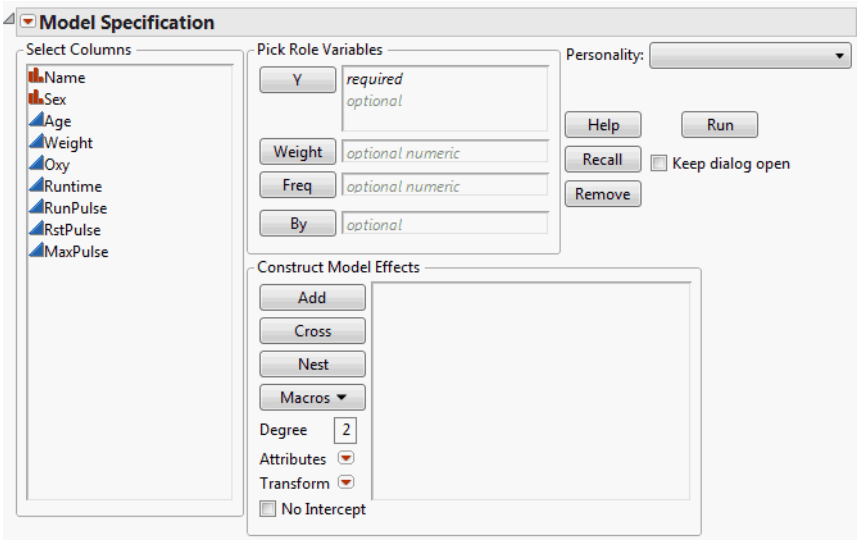
The plot and reports for the whole model appear in the left-most report column. The columns to the right show leverage plots for each of the effects that you specified in the model. Due to space limitations, Figure 2.2 shows only the column for Sex, but the report shows columns for the other six effects as well. The red triangle menus contain additional options that add reports and plots to the report window. For details about the Standard Least Squares report window, see "Fit Least Squares Report" on page 86 in the "Standard Least Squares Report and Options" chapter.

Looking at the p-values in the Parameter Estimates report, you can see that Runtime, RunPulse, and MaxPulse appear to be significant predictors of oxygen uptake. The next step might be to reduce the model by removing insignificant predictors. See the "Stepwise Regression Models" chapter on page 249 for more information.

Launch the Fit Model Platform

You can launch the Fit Model platform by selecting **Analyze > Fit Model**. Figure 2.3 shows an example of the launch window for the Fitness.jmp sample data table.

Figure 2.3 Fit Model Launch Window



Note: When you select **Analyze > Fit Model** in a data table that has a script named *Model* (or *model*), the launch window is automatically filled in based on the script.

The Fit Model launch window contains the following major areas:

- Select Columns is a list of the columns in the current data table. If you have excluded columns, they will not appear in the list.
- Pick Role Variables contains standard JMP launch window buttons. For a description of these buttons, see "Fit Model Launch Window Elements" on page 40.
- Construct Model Effects contains options that you use to enter effects into your model. See "Construct Model Effects" on page 42.
- Personality is a list of the model types you can choose from. Once you have selected a personality, different options appear, depending on the personality that you have selected. See "Fitting Personalities" on page 49.

Fit Model Launch Window Elements

The following Fit Model launch window elements are common to most personalities:

Model Specification The Model Specification menu contains the following options:

Center Polynomials Centers the effects when polynomials are included in the model.

Informative Missing Creates a coding system that accommodates missing values in effects.

Set Alpha Level Sets the alpha level for confidence intervals in the model reports.

Save to Data Table Saves the model specifications to a script in the current data table.

Save to Script Window Saves the model specifications to a script window.

Create SAS Job Saves SAS code for the current model specification to a SAS Program window.

Submit to SAS Submits SAS code for the current model specification to a SAS session.

Convergence Settings Contains options for setting the maximum iterations and convergence limit used in the convergence of models in some personalities.

For more information about any of these options, see "Model Specification Options" on page 51.

Select Columns Lists the unexcluded columns in the current data table.

Y Identifies one or more response variables (the dependent variables) for the model.

Weight Identifies a column whose values assign a weight to each row for the analysis. See "Weight" on page 42.

Freq Identifies a column whose values assign a frequency to each row for the analysis. In general terms, the effect of a frequency column is to expand the data table, so that any row with integer frequency *k* is expanded to *k* rows. You are allowed to specify fractional frequencies. See "Frequency" on page 41.

JMP PRO **Validation** In JMP Pro, for some personalities, you can enter a Validation column. See the appropriate Personality chapter for details. If you click the Validation button with no columns selected in the Select Columns list, you can add a validation column to your data table. For more information about the Make Validation Column utility, see the Modeling Utilities chapter in the *Predictive and Specialized Modeling* book.

By Performs a separate analysis for each level of the variable.

Add Adds effects to the model. See "Add" on page 43.

Cross Creates interaction and polynomial effects by crossing two or more variables. See "Cross" on page 43.

Nest Creates nested effects. See "Nest" on page 43.

Macros Generates effects for commonly used models. See "Macros" on page 44.

Degree Applies the specified degree to models with factorial or polynomial effects generated using Macros. See **Factorial to Degree** and **Polynomial to Degree** in "Macros" on page 44.

Attributes Applies attributes to model effects. These attributes determine how the effects are treated. See "Attributes" on page 45.

Transform Transforms selected continuous effects or *Y* columns. See "Transform" on page 47.

No Intercept Excludes the intercept term from the model.

Personality Specifies the fitting methodology. See "Fit Model Launch Window Elements" on page 40. Different options appear depending on the personality that you select.

Target Level (Available only in certain personalities and when *Y* is binary and has a nominal modeling type.) Specifies the level whose probability you want to model.

Help Takes you to Help topics for the Fit Model launch window.

Recall Populates the launch window with the last model specification that you ran.

Remove Removes the selected variable from the assigned role. Alternatively, double click the effect or select the effect and press the Backspace key.

Run Generates the report window for the specified model and personality.

Keep dialog open Keeps the launch window open after you run the analysis, enabling you to alter and re-run the analysis at any time.

Frequency

Frequency variables, entered in the **Freq** text box, are supported in most Fit Model personalities. In general, a frequency is interpreted as follows. Suppose that a row has a frequency *f*. Then the computed results are identical to those for a data table containing *f* copies of that row, each having a frequency of one.

Rows with zero or missing frequency values are excluded from analyses. Rows with negative frequency values are permitted only for censored observations, otherwise they are excluded

from analyses. When used with censored observations, negative frequency values can be used to fit truncated distributions.

Frequency values need not be integers. The technical details describing how frequency columns, including those with non-integer values, are handled are given in "Frequencies" on page 544.

Weight

Weight variables can be useful in situations where there are observations with different variances. For example, this can happen when one performs regression modeling on data where each row consists of pre-summarized means. Here, rows involving a larger number of observations (smaller variance) should contribute more heavily to the loss function than rows involving a smaller number of observations (larger variance). You can ensure that this occurs by using appropriately defined weights.

Weight variables are supported in many Fit Model personalities. Each personality that supports weight variables uses one of the following methods:

- Variance Scaling
- Frequency Symmetry

Variance Scaling

When estimation is performed using least squares or normal theory maximum likelihood, the weight w for a given row scales that row's contribution to the loss function by $w^{-1/2}$.

Weight variables have an impact on estimates and standard errors. However, unlike frequency variables, they do not affect the degrees of freedom used in hypothesis tests.

Rows with negative or zero values for Weight are excluded from analyses.

Frequency Symmetry

In the Nominal Logistic and Ordinal Logistic personalities, weight variables are handled as if they were frequency variables. If weight and frequency variables are both specified, these personalities handle each observation as if it has a frequency value equal to the product of the weight and frequency values.

Construct Model Effects

This section describes the options that you can use to facilitate entering effects into your model. Examples of how these options can be used to obtain specific types of models are given in "Examples of Model Specifications and Their Model Fits" on page 54.

Add

Adds effects to the model. These effects can either be added directly from the Select Columns list or they can be selected in that list and modified using Macros or Attributes. Effects can also be created and added, or modified, using Cross and Nest. The modeling types of the variables involved in the effect, as well as any Attribute assigned to the effect, determine how that effect is treated in the model.

Note: To remove an effect from the Construct Model Effects list, double-click the effect, or select it and click **Remove** or press the Backspace or Delete key.

Cross

Creates interaction or polynomial effects. Select two or more variables in the Select Columns list and click **Cross**. Or, select one or more variables in the Select Columns list and one or more effects in the Construct Model Effects list and click **Cross**.

See "Statistical Details" on page 521, for a discussion of how crossed effects are parameterized and coded.

Note: You can construct effects that combine up to ten columns as crossed and nested.

Example of Crossed Effects

Suppose that a product coating requires a dye to be applied. Both Dye pH and Dye Concentration are suspected to have an effect on the coating color. To understand their effects, you design an experiment where Dye pH and Dye Concentration are each set at a high and low level. It is possible that the effect of Dye pH on the color is more pronounced at the high level of Dye Concentration than at its low level. This is known as an *interaction*. To model this possible interaction, you include the crossed term, Dye pH * Dye Concentration, in the Construct Model Effects list. This enables JMP to test for an interaction.

Nest

Creates nested effects. If the levels of one effect (B) occur only within a single level of another effect (A), then B is said to be *nested* within A. The notation B[A], which is read as "B nested within A," is typically used. Note that nesting defines a hierarchical relationship. A is called the *outside* effect and B is called the *inside* effect. Nested terms must be categorical.

Note: The nesting terms must be specified in order from outer to inner. For example, if B is nested within A, and C is nested within B, then the model is specified as: A, B[A], C[B,A] (or, equivalently, A, B[A], C[A,B]). You can construct effects that combine up to ten columns as crossed and nested.

Example of Nested Effects

As an illustration of nesting, consider the math teachers in each of two schools. One school has three math teachers; the other school has two math teachers. Each teacher in each school teaches two or three classes consisting of non-overlapping groups of students. In this example, classes (C) are nested within teachers (B), and teachers (B) are nested within schools (A). Enter these effects in the Fit Model launch window as follows:

1. Add both A and B to the Construct Model Effects panel.
2. In the Construct Model Effects panel, click B.
3. In the Select Columns list, click A.
4. Click **Nest**. This converts B to the effect B[A].
5. Add C to the Construct Model Effects panel and click on it.
6. Select A and B in the Select Columns list.
7. Click **Nest**. The converts C to the effect C[A, B].

Macros

In the Macros list, select options to automatically generate the effects for commonly used models and enter them into the Construct Model Effects list:

Full Factorial Creates all main effects and interactions for the columns selected in the Select Columns list. These are entered in an order that is based on the order in which the main effects are listed in the Select Columns list. For an alternate ordering, see **Factorial Sorted**, in this table.

Factorial to Degree Creates all main effects, but only interactions up to a specified degree (order). Specify the degree in the **Degree** box beneath the **Macros** button.

Factorial Sorted Creates the same set of effects as the **Full Factorial** option but lists them in order of degree. All main effects are listed first, followed by all two-way interactions, then all three-way interactions, and so on.

Response Surface Creates main effects, two-way interactions, and quadratic terms. The selected main effects are given the response surface attribute, denoted **RS**. When the RS attribute is applied to main effects and the Standard Least Squares personality is selected, a Response Surface report is provided. This report gives information about critical values and the shape of the response surface.

See also Response Surface Effect in "Attributes" on page 45 and the Response Surface Design chapter in the *Design of Experiments Guide*.

Mixture Response Surface Creates main effects and two-way interactions. Main effects have the response surface (**RS**) and mixture (**Mixture**) attributes. In the Standard Least

Squares personality, the **Mixture** attribute causes a mixture model to be fit. The **RS** attribute creates a Response Surface report that is specific to mixture models.

See also Mixture Effect in "Attributes" on page 45 and the Response Surface Design chapter in the *Design of Experiments Guide*.

Polynomial to Degree Creates main effects and polynomial terms up to a specified degree. Specify the degree in the **Degree** box beneath the **Macros** button.

Scheffé Cubic Creates main effects, interactions, and Scheffé cubic terms, which are useful in specifying response surfaces for mixture experiments. This macro creates a complete cubic model.

When you fit a 3rd degree polynomial model to a mixture, you must not introduce even-powered terms, for example X1*X1*X2, because they are not estimable. However, it turns out that a complete polynomial specification of the surface can include terms of the form X1*X2*(X1 – X2), which are called *Scheffé cubic* terms.

Scheffé cubic terms are also included if you enter a 3 in the **Degree** box and then select the **Mixture Response Surface** macro command.

JMP PRO **Grouped Regressors** Creates a single effect for up to 10 continuous factors that are treated as one effect in various parts of the report. You can add or remove the grouped effect in the Effect Summary. There is one leverage plot and one effect test for each group effect, rather than individual plots and tests for each factor in the group. This can be useful if your data table contains indicator columns that represent levels of a categorical effect.

Note: Non-continuous factors included in a group of regressors are not included in the analysis.

Attributes

In the Attributes list, select attributes that you can assign to an effect selected in the Construct Model Effects list.

Random Effect Assigns the Random attribute to an effect. For details about random effects, see "Specifying Random Effects and Fitting Method" on page 189 in the "Standard Least Squares Report and Options" chapter.

Response Surface Effect Assigns the **RS** attribute to an effect. Note that the relevant model terms must be included in the Construct Model Effects list. The Response Surface option in the Macros list automatically generates these terms and assigns the RS attribute to the main effects. To obtain the Response Surface report, interaction and polynomial terms do not need to have the RS attribute assigned to them. You need only assign this attribute to main effects.

LogVariance Effect Assigns the LogVariance attribute to an effect. This attribute indicates that the effect is to be included in a model of the variance of the response.

To include an effect in models for *both* the mean and variance of the response, you must specify the effect twice. In the tabbed interface, it must appear on both the Mean Effects and Variance Effects tabs. Otherwise, you can enter it twice on the Mean Effects tab, once without the LogVariance Effect attribute and once with the LogVariance Effect attribute.

Mixture Effect Assigns the Mixture attribute to main effects. This is used to specify the main effects involved in the mixture. Note that the Mixture Response Surface option in the Macros list automatically assigns the mixture attribute to selected effects, and provides a Response Surface report when possible.

Excluded Effect Assigns the Excluded attribute to an effect. This excludes the effect from the model fit. However, the effect is used to group observations for lack-of-fit tests. In the Standard Least Squares personality, a table of least square means is provided for this effect.

Knotted Spline Effect Assigns the Knotted attribute to a continuous main effect. This implicitly adds cubic splines for the effect to the model specification. See "Knotted Spline Effect" on page 46.

Knotted Spline Effect

Knotted splines are used to fit a response Y using a flexible function of a predictor. Consider the single predictor X. When the Knotted Spline Effect is assigned to X, and k knots are specified, then k-2 additional effects are implicitly added to the set of predictors. Each of these effects is a piecewise cubic polynomial spline whose segments are defined by the knots. See Stone and Koo (1985).

The number of splines is determined by the number of knots, which you are asked to specify. The coefficients associated with the splines are estimated based on the method used by the personality.

The placement of knots follows guidance given in the literature. In particular, if there are 100 or fewer points, the first and last knots are the fifth smallest and largest points, respectively. Otherwise, the first and last knots are placed at the 0.05 and 0.95 quantiles for 5 or fewer knots, or the 0.025 and 0.975 quantiles for more than 5 knots. The default number of knots is 5 for more than 30 observations, and 3 for fewer than 30 observations.

Note: Knotted splines are implemented only for main-effect continuous terms.

Transform

The Transform options transform selected Y columns or main effects that are selected in the Construct Model Effects text box.

Note: You can also transform a column by right-clicking it in the Select Columns list and selecting Transform. A reference to the transformed column appears in the Select Columns list. You can then use the column in the Fit Model window as you would any data table column. See the Enter and Edit Data chapter in the *Using JMP* book for details.

None Removes any Transform options that have been applied.

Log Applies the natural logarithm transformation to the selected variable.

Sqrt Applies the square root of the values of the selected variable.

Square Applies the square of the values of the selected variable.

Reciprocal Applies the transformation 1/X to the variable X.

Exp Applies the exponential transformation to the selected variable.

Arrhenius Applies the Arrhenius transformation to the variable T (temperature in degrees Centigrade):

$$X = \frac{11605}{T + 273.15}$$

This is the component of the Arrhenius relationship that is multiplied by the activation energy.

ArrheniusInv Applies the inverse of the Arrhenius transformation to the variable X:

$$T = \frac{11605}{X} - 273.15$$

Logit Calculates the inverse of the logistic function for the selected column (where p is in the range of 0 to 1):

$$\text{Logit}(p) = \log\left(\frac{p}{1-p}\right)$$

Logistic Calculates the logistic (also known as Squish and Logist) function for the selected column (where the result is in the range of 0 to 1):

$$\text{Logistic}(x) = \frac{1}{(1+e^{-x})}$$

LogitPct Calculates the logit as a percent for the selected column (where *pct* is a percent in the range of 0 to 100):

$$\text{LogitPct}(pct) = \log\left(\frac{\left(\frac{pct}{100}\right)}{1-\left(\frac{pct}{100}\right)}\right)$$

LogisticPct Calculates the logistic (or logist) as a percent for the selected column (where the result is in the range of 0 to 100):

$$\text{LogisticPct}(x) = \frac{100}{(1+e^{-x})}$$

No Intercept

Select No Intercept if you want to fit a model with no intercept term. Certain modeling structures require no intercept models. For these, the No Intercept box is checked by default.

Construct Model Effects Tabs

For the following personalities, you can enter model effects using a tabbed interface:

Note: If you apply Attributes to effects on the first (main) tab, the attributes determine how the effects are treated in the model. If you run the model and then request Model Dialog from the report's red triangle menu, you find that those effects appear on the appropriate tabs.

Standard Least Squares Enter model effects as follows:

- Fixed Effects tab: Enter effects to be modeled as fixed effects. A fixed effect is one whose specific treatment levels are of interest. You want to compare the mean response across its treatment levels.
- Random Effects tab: Enter effects to be modeled as random effects. A random effect is one whose levels are considered a random sample from a larger population. You want to estimate the variation in the response that is attributable to this effect.

Mixed Model Enter model effects as follows:

- Fixed Effects tab: Enter effects to be modeled as fixed effects. See Standard Least Squares in this table.

- Random Effects tab: Enter effects to be modeled as random effects. Use for variance component models and random coefficients models.
- Repeated Structure tab: Use to select a covariance structure for repeated effects.

LogLinear Variance Enter model effects as follows:

- Mean Effects tab: Enter effects for which you want to model expected values.
- Variance Effects tab: Enter effects for which you want to model variance.

If you want to model both the expected value and variance of an effect, you must enter it on both tabs.

Parametric Survival Enter model effects as follows:

- Location Effects tab: Enter effects that you want to use in modeling the location parameter, mu, or in the case of the Weibull distribution, the shape parameter.
- Scale Effects tab: Enter effects that you want to use in modeling the scale parameter.

Fitting Personalities

In the Fit Model launch window, you select your fitting and analysis method by specifying a Personality. Based on the response (or responses) and the factors that you enter, JMP makes an initial context-based guess at the desired personality, but you can alter this selection in the Personality menu.

The following fitting personalities are available:

Standard Least Squares Fits models where the response is continuous. Techniques include regression, analysis of variance, analysis of covariance, mixed models, and analysis of designed experiments. See the "Standard Least Squares Report and Options" chapter on page 75 and "Emphasis Options for Standard Least Squares" on page 84.

Stepwise Facilitates variable selection for standard least squares and ordinal logistic analyses (or nominal with a binary response). For continuous responses, cross validation, p-value, BIC, and AICc criteria are provided. Also provided are options for fitting all possible models and for model averaging. For logistic fits, p-value, BIC, and AICc criteria are provided. See the "Stepwise Regression Models" chapter on page 249.

JMP PRO **Generalized Regression** Fits generalized linear models using regularized, also known as penalized, regression techniques. The regularization techniques include ridge regression, the lasso, the adaptive lasso, the elastic net, and the adaptive elastic net. The response distributions include the normal, binomial, Poisson, zero-inflated Poisson, negative binomial, zero-inflated negative binomial, and gamma. See the "Generalized Regression Models" chapter on page 285 and "Distribution" on page 292.

JMP PRO **Mixed Model** Fits a wide variety of linear models for continuous-responses with complex covariance structures. The situations addressed include:

- Split plot experiments
- Random coefficients models
- Repeated measures designs
- Spatial data
- Correlated response data

See the "Mixed Models" chapter on page 351.

Manova Fits models that involve multiple continuous Y variables. Techniques include multivariate analysis of variance, repeated measures, discriminant analysis, and canonical correlations. See the "Multivariate Response Models" chapter on page 431.

LogLinear Variance For a continuous Y variable, constructs models for both the mean and the variance. You can specify different sets of effects for the two models. See the "Loglinear Variance Models" chapter on page 457.

Nominal Logistic Fits a logistic regression model to a nominal response. See the "Logistic Regression Models" chapter on page 469.

Ordinal Logistic Fits a logistic regression model to an ordinal response. See the "Logistic Regression Models" chapter on page 469.

Proportional Hazard Fits a semi-parametric regression model (the Cox proportional hazards model) to assess the effect of explanatory variables on survival times, taking censoring into account.

You can also launch this personality by selecting **Analyze > Reliability and Survival > Fit Proportional Hazards**. See the Fit Parametric Survival chapter in the *Reliability and Survival Methods* book.

Parametric Survival Fits a general linear regression model to survival times. Use this option if you have survival times that can be expressed as a function of one or more explanatory variables. Takes into account various survival distributions and censoring.

You can also launch this personality by selecting **Analyze > Reliability and Survival > Fit Parametric Survival**. See the Fit Parametric Survival chapter in the *Reliability and Survival Methods* book.

Generalized Linear Model Fits generalized linear models using various distribution and link functions. Techniques include logistic, Poisson, and exponential regression. See the "Generalized Linear Models" chapter on page 499.

JMP PRO **Partial Least Squares** Fits models to one or more *Y* variables using latent factors. This permits models to be fit when explanatory variables (*X* variables) are highly correlated, or when there are more *X* variables than observations.

You can also launch a partial least squares analysis by selecting **Analyze > Multivariate Methods > Partial Least Squares**. See the Partial Least Squares Models chapter in the *Multivariate Methods* book.

Response Screening Automates the process of conducting tests for linear model effects across a large number of responses. Test results and summary statistics are presented in data tables and plots. A False-Discovery Rate (FDR) approach guards against incorrect declarations of significance. A robust estimation method reduces the sensitivity of tests to outliers.

Note: This personality only allows continuous responses. Response Screening for individual factors is also available by selecting **Analyze > Screening > Response Screening**. This platform supports categorical responses, and also provides equivalence tests and tests of practical significance. See the Response Screening chapter in the *Predictive and Specialized Modeling* book.

Model Specification Options

The red triangle menu next to Model Specification includes the following options:

Center Polynomials Centers by its mean any continuous term that is involved in an effect with a degree greater than one. This option is checked by default, except when a term involved in the effect is assigned the Mixture Effect attribute or has the Mixture column property. Terms with the Coding column property are centered midway between their specified High and Low values.

Centering is useful in making regression coefficients more interpretable and in reducing collinearity between model effects.

Informative Missing Provides a coding system for missing values. This system allows estimation of a predictive model despite the presence of missing values. It is useful in situations where missing data are informative. See "Informative Missing" on page 53 for more details.

This option is available for the following personalities: Standard Least Squares, Stepwise, Generalized Regression, MANOVA, Loglinear Variance, Nominal Logistic, Ordinal Logistic, Proportional Hazard, Parametric Survival, Generalized Linear Model, and Response Screening.

Set Alpha Level Sets the alpha level for confidence intervals in the Fit Model analysis. The default alpha level is 0.05.

Save to Data Table Saves your Fit Model launch window specifications as a script that is attached to the data table. The script is named Model. When a table contains a script called Model, this script automatically populates the launch window when you select **Analyze > Fit Model**. (Simply rename the script if this is not desirable.)

For details about JSL scripting, see the *Scripting Guide*.

Save to Script Window Copies your Fit Model launch window specifications to a script window. You can save the script window and re-create the model at any time by running the script.

Create SAS job Creates a SAS program that can re-create the current analysis and data table in SAS in a script window. Once created, you have several options for submitting the code to SAS.

1. Copy and paste the code into the SAS Program Editor. This method is useful if you are running an older version of SAS (pre-version 8.2).
2. Select **Edit > Submit to SAS**.
3. Save the file and double-click it to open it in a local copy of SAS. This method is useful if you would like to take advantage of SAS ODS options, such as generating HTML or PDF output from the SAS code.

See the Import Your Data chapter in the *Using JMP* book.

Submit to SAS Submits code to SAS and displays the results in JMP. If you are not connected to a SAS server, prompts guide you through the connection process.

See the Import Your Data chapter in the *Using JMP* book.

Convergence Settings The Convergence Settings menu contains the following options:

Maximum Iterations Specifies the maximum number of iterations that are used in the model fitting. By default, the maximum number of iterations is 100. If your model does not readily converge, you might want to increase the number of iterations. If you have a very large data set or a complicated model, you might want to limit the number of iterations.

Convergence Limit Specifies the convergence limit for the model fitting. If you model does not readily converge, you might want to increase the convergence limit. By default, the convergence limit is 0.00000001.

Note: The Convergence Settings menu appears only for certain personalities. In the Standard Least Squares personality, the Convergence Settings menu appears only when there is a random effect and REML is selected as the Method in the launch window.

Informative Missing

The Informative Missing option constructs a coding system that allows estimation of a predictive model despite the presence of missing values. It codes both continuous and categorical model effects.

Continuous Effects

When a continuous main effect has missing values, a new design matrix column is created. This column is an indicator variable, with values of one if the main effect column is missing and zero if it is not missing. In addition, missing values for the continuous main effect are replaced with the mean of the non-missing values for rows included in the analysis. The mean is a neutral value that maintains the interpretability of parameter estimates.

The parameter associated with the indicator variable estimates the difference between the response predicted by the missing value grouping and the predicted response if the covariate is set at its mean.

For a higher-order effect, missing values in the covariates are replaced by the covariate means. This makes the higher-order effect zero for rows with missing values, assuming that Center Polynomials is checked (the default setting). This is because Center Polynomials centers the individual terms involved in a polynomial by their means.

In the Effect Tests report, each continuous main effect with missing values will have Nparm = 2. In the Parameter Estimates report, the parameter for a continuous main effect with missing values is labeled <colname> Or Mean if Missing and the indicator parameter is labeled <colname> Is Missing. Prediction formulas that you save to the data table are given in terms of expressions corresponding to these model parameters.

Categorical Effects

When a nominal or ordinal main effect has missing values, the missing values are coded as a separate level of that effect. As such, in the Effect Tests report, each categorical main effect with missing values will have one additional parameter.

In the Parameter Estimates report, the parameter for a nominal effect is labeled <colname>[ ]. For an ordinal effect, the parameter is labeled <colname>[-x], where x denotes the level with highest value ordering.

As with continuous effects, prediction formulas that you save to the data table are given in terms of expressions corresponding to the model parameters.

Coding Table

When you are using the Standard Least Squares personality, you can view the design matrix columns used in the Informative Missing model by selecting **Save Columns > Save Coding Table.**

Validity Checks

Fit Model checks your model for errors such as duplicate effects or missing effects in a hierarchy. If you receive an alert message, you can either click **Continue** to proceed with fitting, or click **Cancel** to stop the fitting process.

Examples of Model Specifications and Their Model Fits

This section gives templates for entering the effects for various model types that you can specify using the Construct Model Effects panel in the Fit Model platform.

- The model effects X and Z represent continuous columns.
- The model effects A, B, and C represent nominal or ordinal columns.

For most models, visual views of their model fits are also given.

- "Simple Linear Regression"
- "Polynomial in X to Degree k"
- "Polynomial in X and Z to Degree k"
- "Multiple Linear Regression"
- "One-Way Analysis of Variance"
- "Two-Way Analysis of Variance"
- "Two-Way Analysis of Variance with Interaction"
- "Three-Way Full Factorial"
- "Analysis of Covariance, Equal Slopes"
- "Analysis of Covariance, Unequal Slopes"
- "Two-Factor Nested Random Effects Model"

- "Three-Factor Fully Nested Random Effects Model"
- "Simple Split Plot or Repeated Measures Model"
- "Two-Factor Response Surface Model"
- "Knotted Spline Effect"

Simple Linear Regression

Effects to be entered: X

1. In the Select Columns list, select X.
2. Click **Add**.

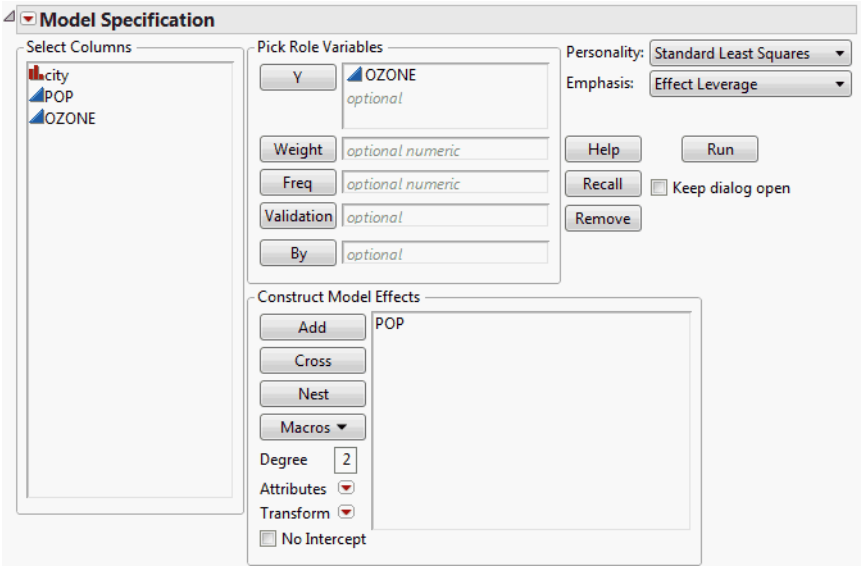
Example of Simple Linear Regression Model

Open Polycity.jmp. You are interested in the relationship between POP, the population in thousands of the given city, and Ozone. Ozone is the response of interest, and POP is the continuous model effect.

1. Select **Analyze > Fit Model**.
2. In the Select Columns list, select Ozone and click Y.
3. In the Select Columns list, select POP.
4. Click **Add**.

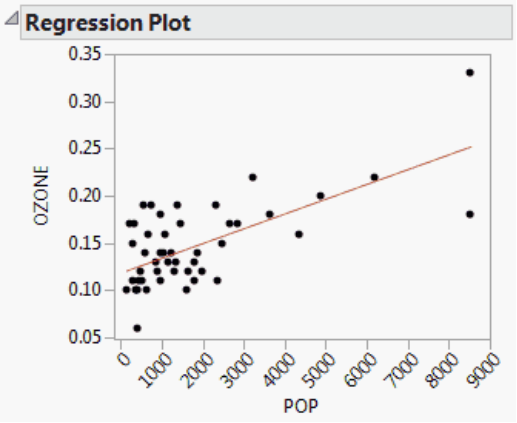
The Fit Model window appears as shown in Figure 2.4

Figure 2.4 Fit Model Window for Simple Linear Regression



When you click **Run**, the Fit Least Squares report appears, showing various results, including a Regression Plot. The Regression Plot shows the data and a simple linear regression model fit to the data (Figure 2.5).

Figure 2.5 Model Fit for Simple Linear Regression



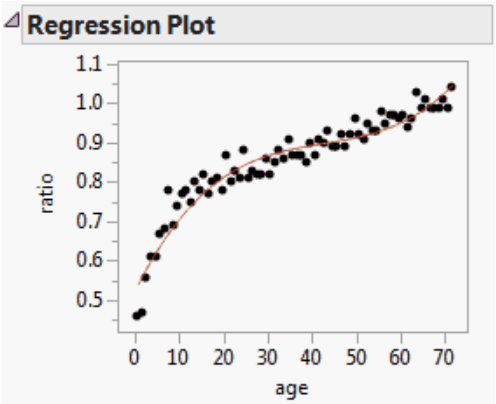
Polynomial in X to Degree k

Effects to be entered: X, X*X, ..., X^k

1. Type *k* into the text box for **Degree**.
2. In the Select Columns list, select X.
3. Select **Macros > Polynomial to Degree**.

Figure 2.6 shows a plot of the data and a cubic polynomial model fit to the data for the Growth.jmp sample data table. This is one of the fits produced when you run the **Bivariate** data table script.

Figure 2.6 Model Fit for a Degree-Three Polynomial in One Variable



Polynomial in X and Z to Degree k

Effects to be entered: X, X*X, ..., X^k, Z, Z*Z, ..., Z^k

1. Type *k* into the text box for **Degree**.
2. In the Select Columns list, select X and Z.
3. Select **Macros > Polynomial to Degree**.

For a plot of a degree-two fit for a polynomial two variables, see Figure 2.21.

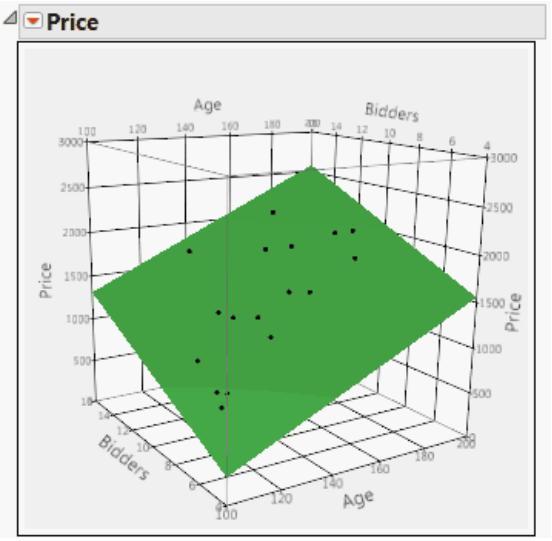
Multiple Linear Regression

Effects to be entered: Selected columns

1. In the Select Columns list, select the continuous effects of interest.
2. Click **Add**.

Figure 2.7 shows a surface profiler plot of the data and of the multiple linear regression fit to the data for the Grandfather Clocks.jmp sample data table. The model effects are Age and Bidders. The response is Price. You can obtain the plot by running the data table script **Fit Model with Surface Profiler Plot**.

Figure 2.7 Model Fit for a Multiple Linear Regression Model with Two Predictors



Example of Multiple Linear Regression Model

See also "Example of a Regression Analysis Using Fit Model" on page 36 for an example with several predictors.

One-Way Analysis of Variance

Effects to be entered: A

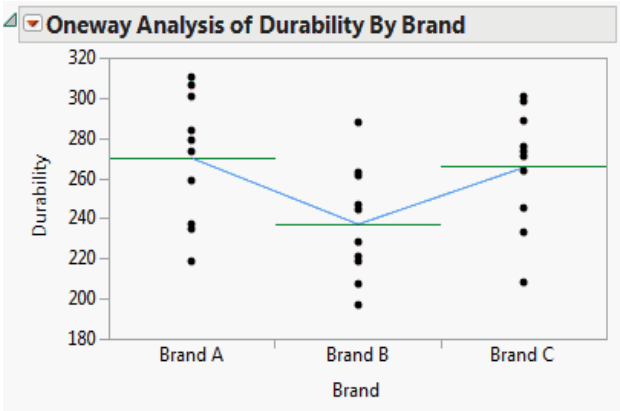
1. In the Select Columns list, select one nominal or ordinal effect, A.
2. Click **Add**.

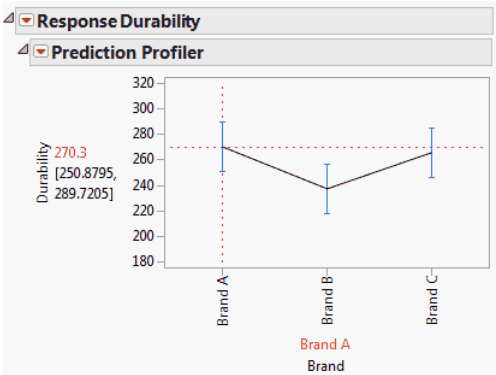
Consider the Golf Balls.jmp sample data table. You are interested in whether Durability varies by Brand. Figure 2.8 shows two plots.

The first is a plot, obtained using Fit Y by X, that shows the data by brand. Horizontal lines are plotted at the mean for each brand and line segments connect the means. To produce this plot, run the script **Oneway: Durability by Brand** in the Golf Balls.jmp sample data table.

The second plot is a profiler plot obtained using Fit Model. This second plot shows the predicted responses for each brand, connected by line segments. To produce this plot, run the script Fit Model: Durability by Brand in the Golf Balls.jmp sample data table. Drag the vertical dashed red line to the brand of interest. The horizontal dashed red line updates to intersect the vertical axis at the predicted response.

Figure 2.8 Model Fit for One-Way Analysis of Variance





Two-Way Analysis of Variance

Effects to be entered: A, B

1. In the Select Columns list, select two nominal or ordinal effects, A and B.
2. Click **Add**.

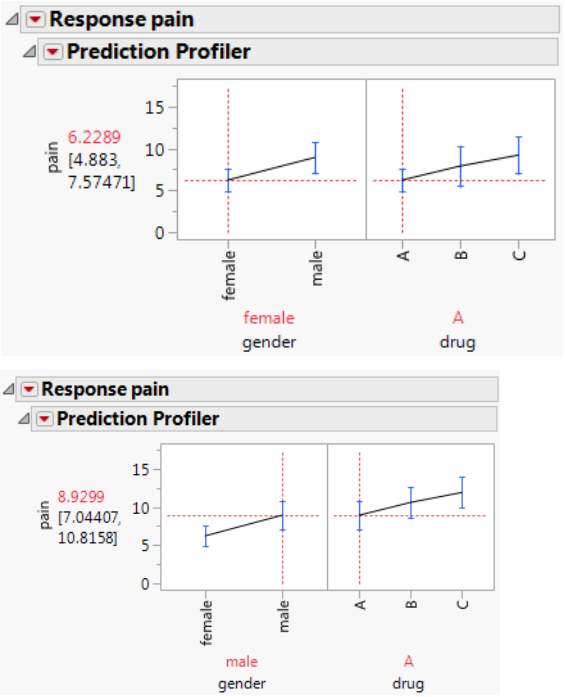
Figure 2.9 shows two profiler plots of the fit to the data for the Analgesics.jmp sample data table. The model effects are gender and drug. The response is pain. To obtain this plot, select **Analyze > Fit Model**, select pain as Y, select gender and drug as model effects, and then click **Run**. From the report's red triangle menu, select **Factor Profiling > Profiler.**

The line segments in each plot connect the predicted values for the settings defined by the vertical dashed red lines. Move these to see predictions at other settings.

The top plot in Figure 2.9 shows predictions for females, while the bottom plot shows predictions for males. Note that the relative effects of the three drugs are consistent across the

levels of gender. This is because there is no interaction term in the model. For an example with interaction, see Figure 2.11.

Figure 2.9 Model Fit for a Two-Way Analysis of Variance with No Interaction



Two-Way Analysis of Variance with Interaction

Effects to be entered: A, B, A*B

1. In the Select Columns list, select two nominal or ordinal effects, A and B.
2. Select **Macros > Full Factorial**.

Or:

1. In the Select Columns list, select two nominal or ordinal effects, A and B.
2. Click **Add**.
3. In the Select Columns list, select A and B again and click **Cross**.

Example of Two-Way Analysis of Variance with Interaction

Open Popcorn.jmp. You are interested in whether popcorn and batch have an effect on yield.

1. Select **Analyze > Fit Model**.

2. In the Select Columns list, select yield and click **Y**.
3. In the Select Columns list, select popcorn and batch.
4. Select **Macros > Full Factorial**.

The Fit Model window appears as shown in Figure 2.10.

Figure 2.10 Fit Model Window for Two-Way Analysis of Variance with Interaction

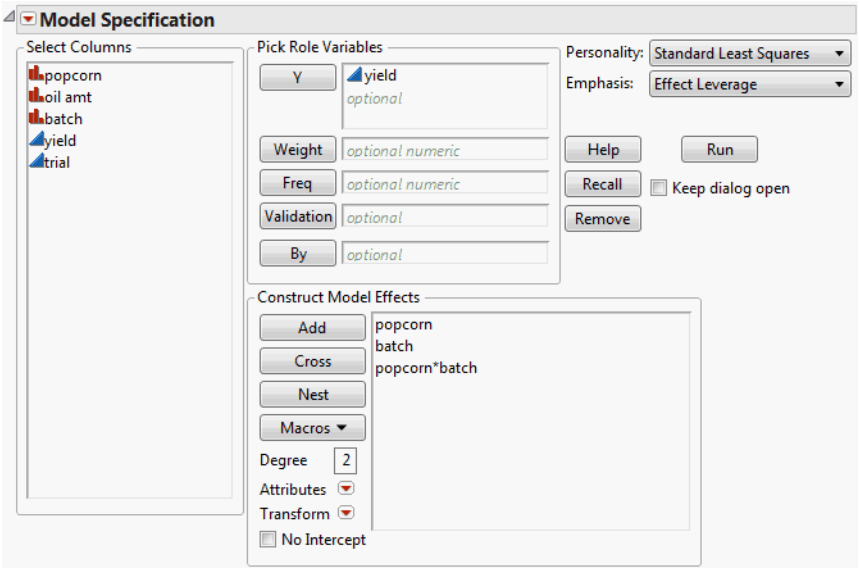
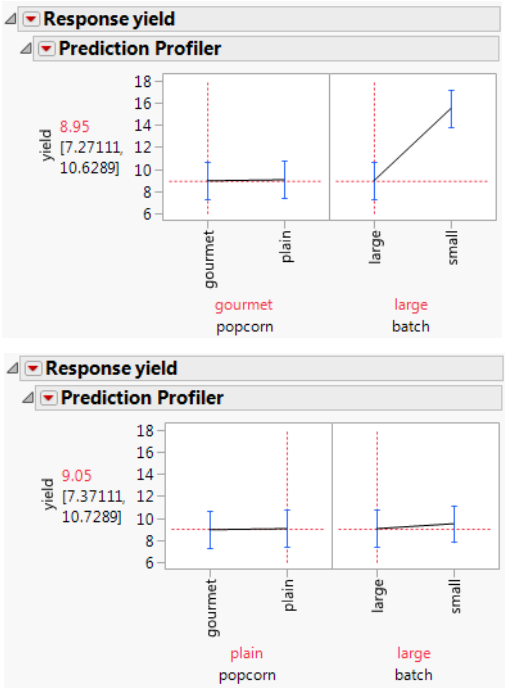


Figure 2.11 shows a profiler plot of the fit for this example. To obtain this plot, click **Run** in the Fit Model window shown in Figure 2.10. Then, from the red triangle menu for the Fit Least Squares report, select **Factor Profiling > Profiler.**

In the top plot, popcorn is set to gourmet, and in the bottom plot, it is set to plain. Note how the predicted values for the settings of batch depend on the type of popcorn. This is a consequence of the interaction between popcorn and batch.

Figure 2.11 Model Fit for a Two-Way Analysis of Variance with Interaction



Three-Way Full Factorial

Effects to be entered: A, B, C, A*B, A*C, B*C, A*B*C

1. In the Select Columns list, select three nominal or ordinal effects, A, B, and C.
2. Select **Macros > Full Factorial**.

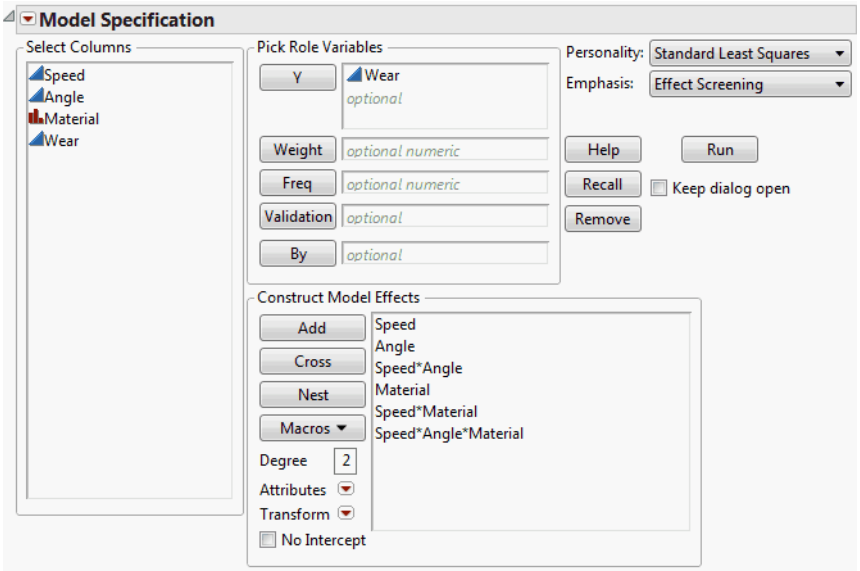
Example of Three-Way Full Factorial

Open Tool Wear.jmp. You are interested in whether Speed, Angle, and Material, or their interactions, have an effect on the Wear of a cutting tool.

1. Select **Analyze > Fit Model**.
2. In the Select Columns list, select Wear and click **Y**.
3. In the Select Columns list, select Speed, Angle, and Material.
4. Select **Macros > Full Factorial**.

The Fit Model window appears as shown in Figure 2.12.

Figure 2.12 Fit Model Window for Three-Way Full Factorial

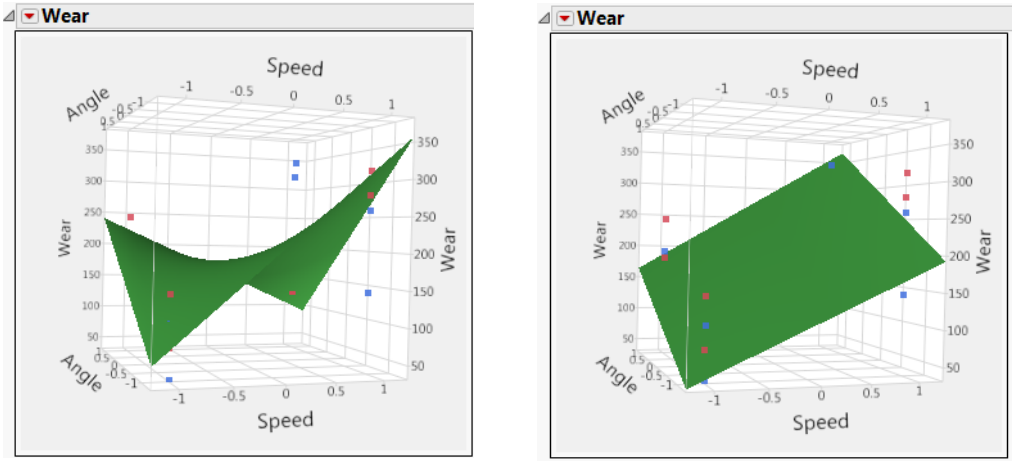


The Surface Profiler plots in Figure 2.13 show the predicted response for Wear in terms of the two continuous effects Speed and Angle. The plot on the left shows the predicted response when Material is A; the plot on the right shows the predicted response when Material is B. The points for which Material is A are colored red, while those for which Material is B are colored blue. The difference in the form of the response surfaces across the levels of Material is a consequence of the three-way interaction.

To obtain Surface Profiler plots, click **Run** in the Fit Model window shown in Figure 2.12. From the red triangle menu for the Fit Least Squares report, select **Factor Profiling > Surface Profiler**. To add points to the plot, open the Appearance panel and click **Actual**. If you want to make the points appear larger, right click in the plot, select **Settings** and adjust the **Marker Size**.

To show plots for both Materials A and B, use the slider marked Material in the Independent Variables panel, setting it at 0 for Material A and 1 for Material B. Note that the table contains two data table scripts that produce Surface Profiler plots: **Prediction and Surface Profilers** and **Surface Profilers for Two Materials**.

Figure 2.13 Model Fit for a Three-Way Full Factorial Design - Material A on Left, Material B on Right



Analysis of Covariance, Equal Slopes

Here you are interested in testing for the effect of A with X as a covariate. Suppose that you have reason to believe that the effect of X on the response does not depend on the level of A.

Effects to be entered: A, X

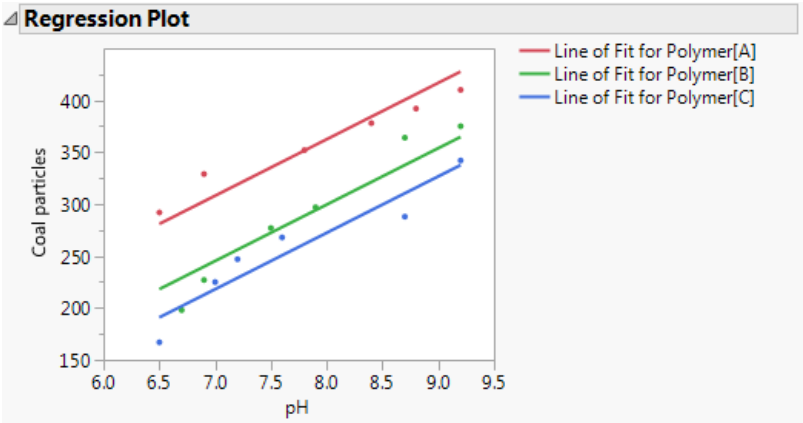
1. In the Select Columns list, select one nominal or ordinal effect, A, and one continuous effect, X.
2. Click **Add**.

Figure 2.14 shows the data from the Cleansing.jmp sample data table. You are interested in which Polymer removes the most Coal particles from a cleansing tank. However, you suspect that the pH of the tank also has an effect on removal. The plot shows an analysis of covariance fit. Here the slopes relating pH and Coal particles are assumed equal across the levels of Polymer.

The plot in Figure 2.14 is obtained as follows. In the Fit Model window, enter Coal particles as **Y** and both pH and Polymer in the **Construct Model Effects** box. Click **Run**. The Regression Plot appears in the Fit Least Squares report. To color the points, select **Rows > Color or Mark by Column**, select Polymer from the **Mark by Column** list, and click **OK**.

However, a more complete analysis indicates that pH and Polymer *do* interact in their effect on Coal particles. The appropriate model fit is shown in "Analysis of Covariance, Unequal Slopes" on page 65.

Figure 2.14 Model Fit for Analysis of Covariance, Equal Slopes



Analysis of Covariance, Unequal Slopes

Here you are again interested in testing for the effect of A with X as a covariate. But you construct your model so as to admit the possibility that the effect of X on the response depends on the level of A.

Effects to be entered: A, X, A*X

1. In the Select Columns list, select one nominal or ordinal effect, A, and one continuous effect, X.
2. Select **Macros > Full Factorial**.

Or:

1. In the Select Columns list, select one nominal or ordinal effect, A, and one continuous effect, X.
2. Click **Add**.
3. In the Select Columns list, select A and X again and click **Cross**.

Example of Analysis of Covariance, Unequal Slopes

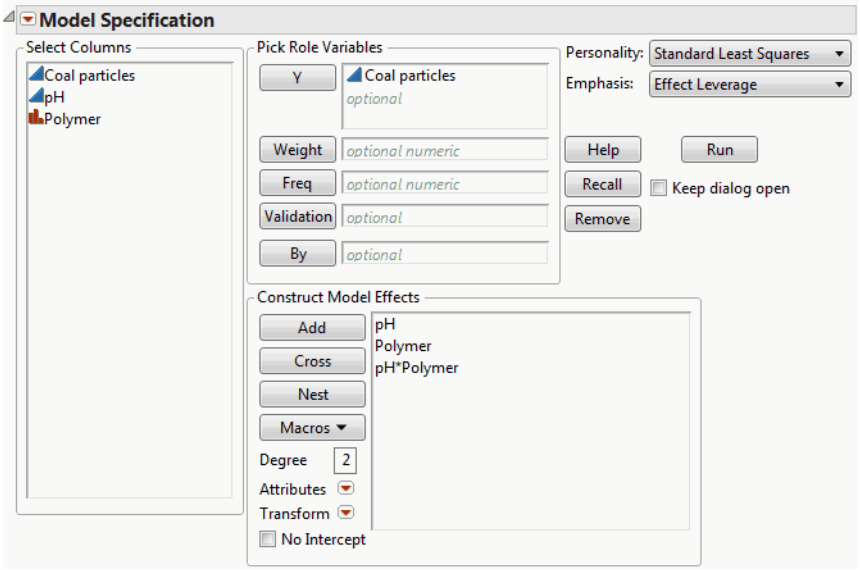
Open Cleansing.jmp. You are interested in whether any of the three polymers (Polymer) has an effect on coal particle removal (Coal particles). The tank pH is included as a covariate as it might affect a polymer's ability to clean the tank. You allow for possibly different slopes when modeling the relationship between pH and Coal particles for the three Polymer types.

1. Select **Analyze > Fit Model**.
2. In the Select Columns list, select Coal particles and click **Y**.
3. In the Select Columns list, select pH and Polymer.

4. Select **Macros > Full Factorial**.

The Fit Model window appears as shown in Figure 2.15.

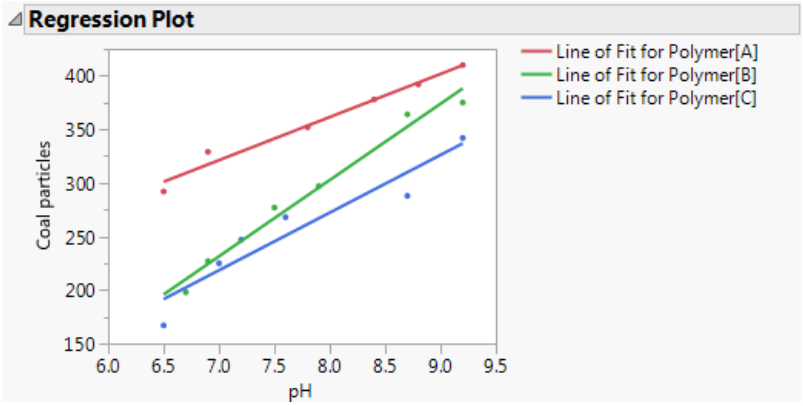
Figure 2.15 Fit Model Window for Analysis of Covariance, Unequal Slopes



When you click **Run**, the Fit Least Squares report appears. The Effect Tests report indicates that the interaction between pH and Polymer is significant and should be included in the model.

The Regression Plot given in the report is shown in Figure 2.16. This plot shows the points and the model fit. The interaction allows the slopes of the lines that relate pH to Coal particles to depend on the Polymer. Note that, despite this interaction, over the range of interest, Polymer A consistently has the highest removal. If you want to color the points as shown in Figure 2.16, select **Rows > Color or Mark by Column**, select Polymer from the **Mark by Column** list, and click **OK**.

Figure 2.16 Model Fit for Analysis of Covariance, Unequal Slopes



Two-Factor Nested Random Effects Model

Consider a model with two factors, A and B, but where B is nested within A. Although there are situations where a nested effect is treated as a fixed effect, in most situations a nested effect is treated as a random effect. For this reason, in the model described below, the nested effect is entered as a random effect.

Effects to be entered: A, B[A]&Random

1. In the Select Columns list, select two nominal or ordinal effects, A and B.
2. Click **Add**.
3. To nest B within A: In the Construct Model Effects list, select B. In the Select Columns list, select A. The two effects should be highlighted.
4. Click **Nest**.
5. With B[A] highlighted in the Construct Model Effects list, select **Attributes > Random Effect**.

Example of Two-Factor Nested Random Effects Model

Open the 2 Factors Nested.jmp sample data table located in the Variability Data subfolder. As part of a measurement systems analysis study, 24 randomly chosen parts are measured. These parts are evenly divided among the six operators who typically measure these parts. Each operator make three independent measurements of each of the four assigned parts.

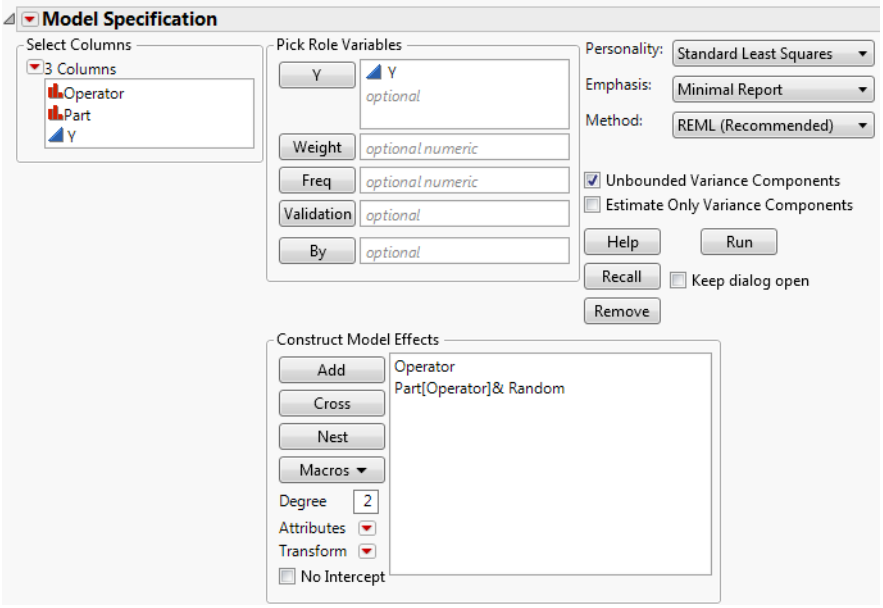
Since the parts measured by one operator are measured only by that specific operator, Part is nested within Operator. Since the parts are a random sample of production, Part is considered a random effect. Since these specific six operators are of interest, Operator is treated as a fixed effect. The appropriate model is specified as follows.

1. Select **Analyze > Fit Model**.

2. In the Select Columns list, select Y and click **Y**.
3. In the Select Columns list, select Operator and Part.
4. Click **Add**.
5. To nest Part within Operator: In the Construct Model Effects list, select Part. In the Select Columns list, select Operator. The two effects should be highlighted.
6. Click **Nest**.
7. With Part[Operator] highlighted in the Construct Model Effects list, select **Attributes > Random Effect**.

   The Fit Model window appears as shown in Figure 2.17.

Figure 2.17 Fit Model Window for Two-Factor Nested Random Effects Model



8. Click **Run** to obtain the Fit Least Squares report.

Figure 2.18 shows two plots. The first is a Variability Chart showing the three measurements by each Operator on each of the four parts. Horizontal line segments show the mean measurement for each Operator.

To construct the Variability Chart in Figure 2.18, in the 2 Factors Nested.jmp sample data table, run the data table script **Variability Chart - Nested**. From the report's red triangle menu, deselect **Show Range Bars** and select **Show Group Means**.

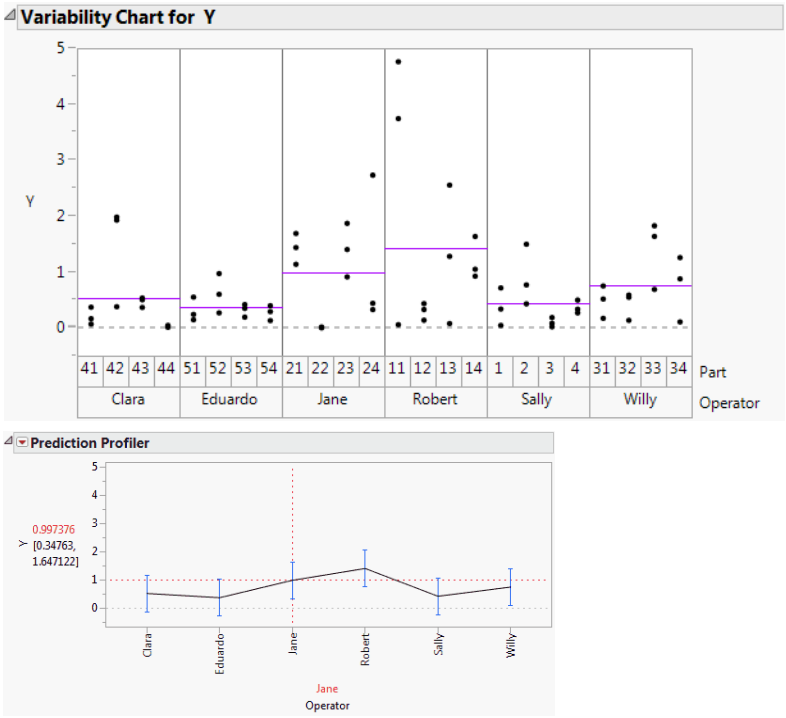
The second plot is the Fit Least Squares report Prediction Profiler plot for Operator. This plot shows the predicted response for each operator. The vertical dashed red line set at Jane indicates that Jane's predicted response is 0.997. You can see the correspondence between the

model predictions given in the Prediction Profiler plot and the raw data in the Variability Chart.

To obtain the Prediction Profiler plot, from the Fit Least Squares report red triangle menu, select **Factor Profiling > Profiler**.

These plots show how the predicted measurements for each Operator are modeled. However, keep in mind that you are not only interested in whether the operators differ in how they measure parts. You are also interested in the variability of the part measurements themselves, which requires estimation of the variance component associated with Part.

Figure 2.18 Model Fit for Two-Factor Nested Random Effects Model



Three-Factor Fully Nested Random Effects Model

Consider a model with three factors, A, B, and C, but where B is nested within A and C is nested within both A and B. Also consider B and C to be random effects.

Effects to be entered: A, B[A]&Random, C[A,B]&Random

1. In the Select Columns list, select three nominal or ordinal effects, A, B, and C.
2. Click **Add**.

3. To nest B within A: In the Construct Model Effects list, select B. In the Select Columns list, select A. The two effects should be highlighted.
4. Click **Nest**.
5. To nest C within A and B: In the Construct Model Effects list, select C. In the Select Columns list, select A and B. The three effects should be highlighted.
6. Click **Nest**.
7. With both B[A] and C[A,B] highlighted in the Construct Model Effects list, select **Attributes > Random Effect**.

Simple Split Plot or Repeated Measures Model

Here A is the whole plot variable, B[A] is the whole plot ID, and C is the split plot, or repeated measures, variable.

Effects to be entered: A, B[A]&Random, C, C*A

1. In the Select Columns list, select two nominal or ordinal effects, A and B.
2. Click **Add**.
3. To nest B within A: In the Construct Model Effects list, select B. In the Select Columns list, select A. The two effects should be highlighted.
4. Click **Nest**.
5. In the Construct Model Effects list, select B[A].
6. Select **Attributes > Random Effect**.
7. In the Select Columns list, select a third nominal or ordinal effect, C.
8. Click **Add**.
9. In the Construct Model Effects list, select C. In the Select Columns list, click A. Both effects should be highlighted.
10. Click **Cross**.

Example of a Simple Repeated Measures Model

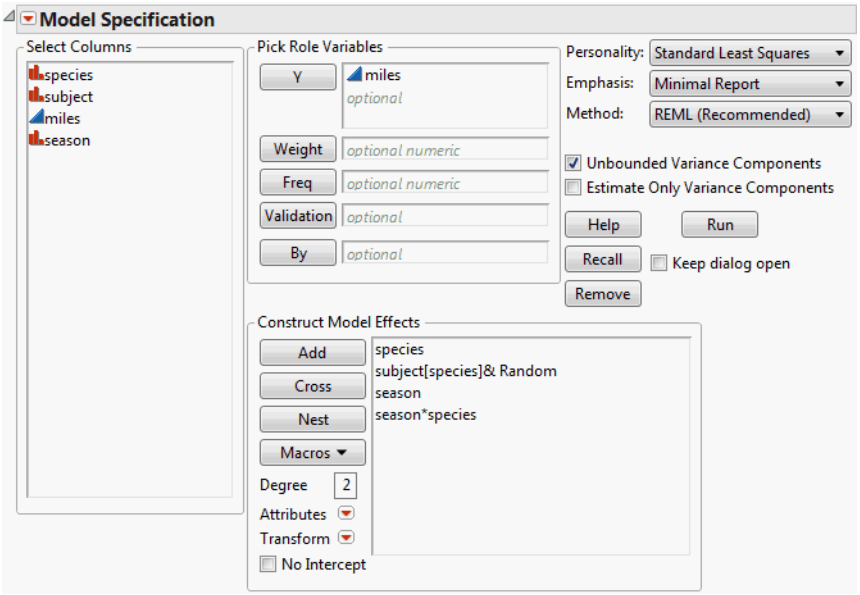
Open the Animals.jmp sample data table. The column Miles gives the distance traveled by each of six animals in each of the four seasons. Note that there are two species and that subject, the animal identifier, is nested within species. Since these six animals are representatives of larger species populations, you decide to treat subject as a random effect. You want to model the response, miles, as a function of species and season, accounting for the fact that there are repeated measures for each animal.

1. Select **Analyze > Fit Model**.
2. In the Select Columns list, select miles and click **Y**.

3. In the Select Columns list, select species and subject.
4. Click **Add**.
5. To nest subject within species: In the Construct Model Effects list, select subject. In the Select Columns list, select species. The two effects should be highlighted.
6. Click **Nest**.
7. In the Construct Model Effects list, select subject[species].
8. Select **Attributes > Random Effect**.
9. In the Select Columns list, select season.
10. Click **Add**.
11. In the Construct Model Effects list, select season. In the Select Columns list, click species. Both effects should be highlighted.
12. Click **Cross**.

The Fit Model window appears as shown in Figure 2.19.

Figure 2.19 Fit Model Window for Simple Repeated Measures Model



Two-Factor Response Surface Model

Effects to be entered: X&RS, Z&RS, X*X, X*Z, Z*Z

1. In the Select Columns list, select two continuous effects, X and Z.
2. Select **Macros > Response Surface**.

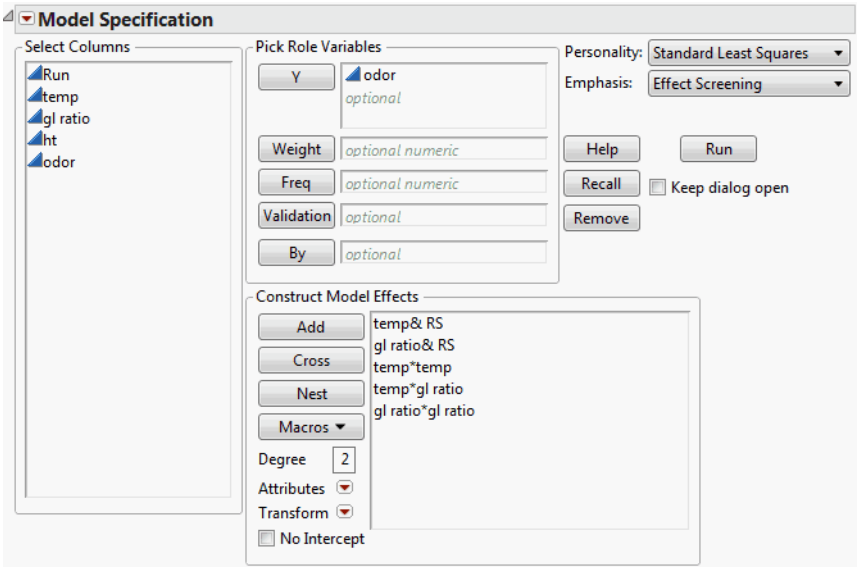
Example of Two-Factor Response Surface Model

Open the Odor Control Original.jmp sample data table. You want to fit a response surface to model the response, odor, as a function of temp and gl ratio. (Although you could include ht, as shown in the data table script Response Surface, for this illustration do not.)

1. Select **Analyze > Fit Model**.
2. In the Select Columns list, select odor and click **Y**.
3. In the Select Columns list, select temp and gl ratio.
4. Select **Macros > Response Surface**.

   The Fit Model window appears as shown in Figure 2.20.

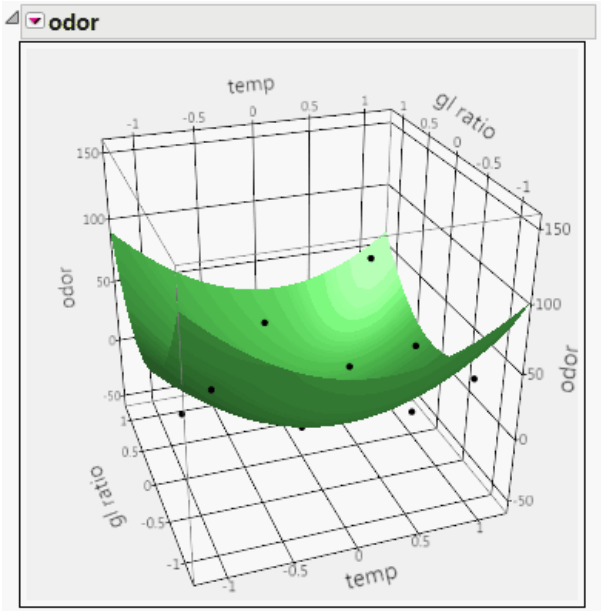
Figure 2.20 Fit Model Window for Two-Factor Response Surface Model



5. Click **Run**.

Figure 2.21 shows a Surface Profiler plot of the data and a quadratic response surface fit to the data for the Odor Control Original.jmp sample data table. To obtain this plot, from the report's red triangle menu, select **Factor Profiling > Surface Profiler**. To show the points, click the disclosure icon to open the **Appearance** panel and click **Actual**. If you want to make the points appear larger, right click in the plot, select **Settings** and adjust the **Marker Size**.

Figure 2.21 Model Fit for a Degree-Two Polynomial in Two Variables



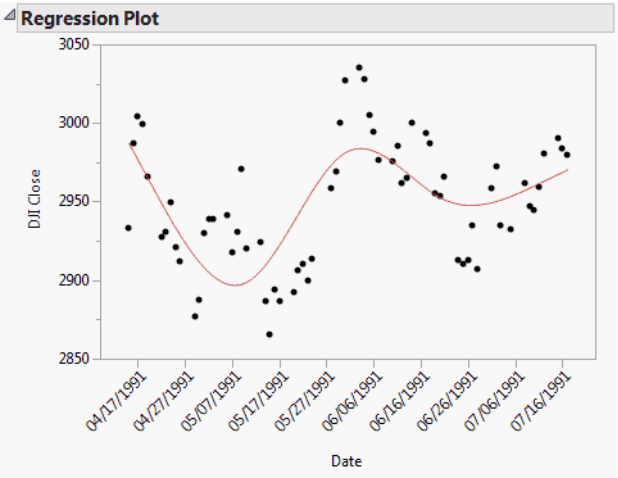
Knotted Spline Effect

Effects to be entered: X&Knotted

1. In the Select Columns list, select a continuous effect, X.
2. Click **Add**.
3. Select X in the Construct Model Effects list.
4. Select **Attributes > Knotted Spline Effect**.
5. In the menu that appears, specify the number of knots or accept the default number.
6. Click **OK**.

Figure 2.22 shows the Regression Plot for a model fit to the data in the XYZ Stock Averages (plots).jmp sample data table. Here Date is assigned the Knotted Spline Effect and five knots are specified. DJI Close is the response.

Figure 2.22 Model Fit for a Knotted Spline with Five Knots



Chapter 3

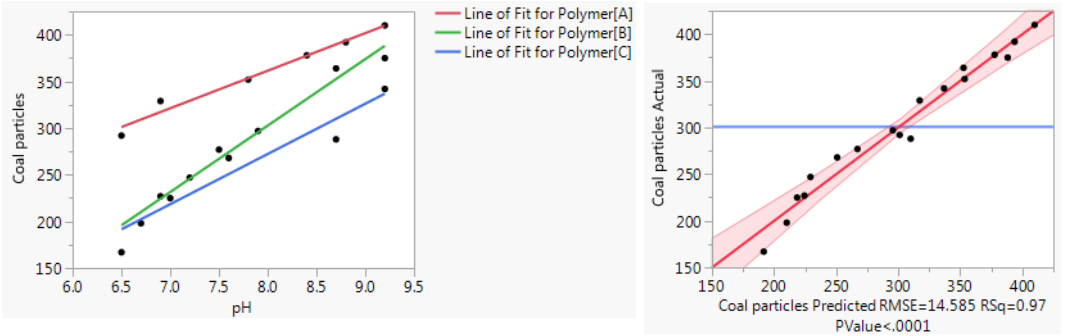
Standard Least Squares Report and Options

Analyze Common Classes of Models

The Standard Least Squares personality within the Fit Model platform fits a wide spectrum of standard models. These models include regression, analysis of variance, analysis of covariance, and mixed models, as well as the models typically used to analyze designed experiments. Use the Standard Least Squares personality to construct linear models for continuous-response data using least squares or, in the case of random effects, restricted maximum likelihood (REML).

Analytic results are supported by compelling dynamic visualization tools such as profilers, contour plots, and surface plots (see the book *Profilers*). These visual displays stimulate, complement, and support your understanding of the model. They enable you to optimize several responses simultaneously and to explore the effect of noise.

Figure 3.1 Examples of Standard Least Squares Plots



Contents

Example Using Standard Least Squares

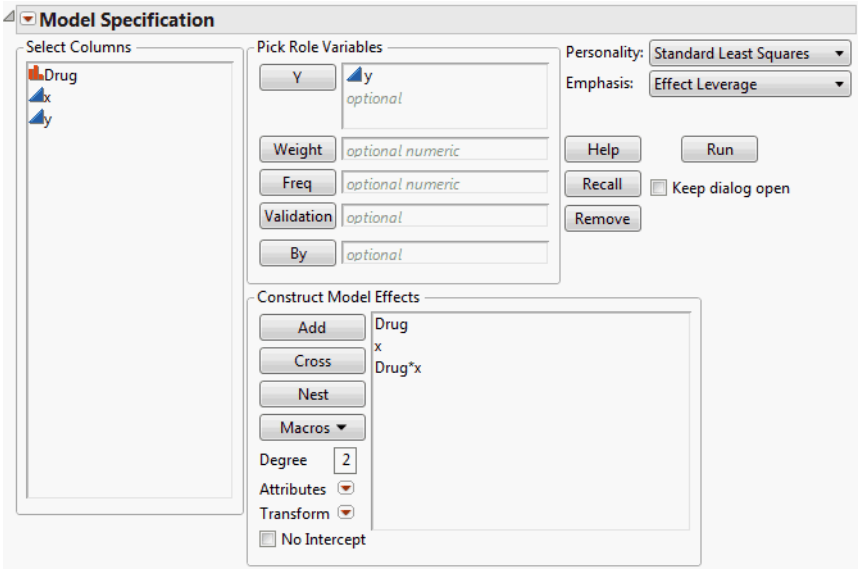
In a study of the effect of drugs in treating a disease, thirty patients are randomly divided into three groups of ten. Two of these groups are administered drugs (Drug a and Drug d), while the third group is administered a placebo (Drug f). A pretreatment measure, x, is taken on each patient, as well as a posttreatment measure, y. The pretreatment score, x, is included as a covariate, to account for differences in the stage of the disease among patients. This example is taken from Snedecor and Cochran (1967, p. 422).

You are interested in determining if there is a difference in the three Drug groups. You construct a model with response y and model effects Drug, x, and the interaction of Drug and x. The interaction might account for a situation where drugs have differential effects, based on the stage of the disease. (For background on the Fit Model window and the various personalities, see the "Model Specification" chapter on page 33.)

1. Select **Help > Sample Data Library** and open Drug.jmp.
2. Select **Analyze > Fit Model**.
3. Select y and click **Y**.

   When you add this column as **Y**, the fitting **Personality** becomes Standard Least Squares. An **Emphasis** option is added with a selection of Effect Leverage, which you can change if desired.
4. Select Drug and x. With these two effects highlighted in the Select Columns list, click **Macros** and select **Full Factorial**. The macro adds the two effects and their two-way interaction to the Construct Model Effects list (Figure 3.2).

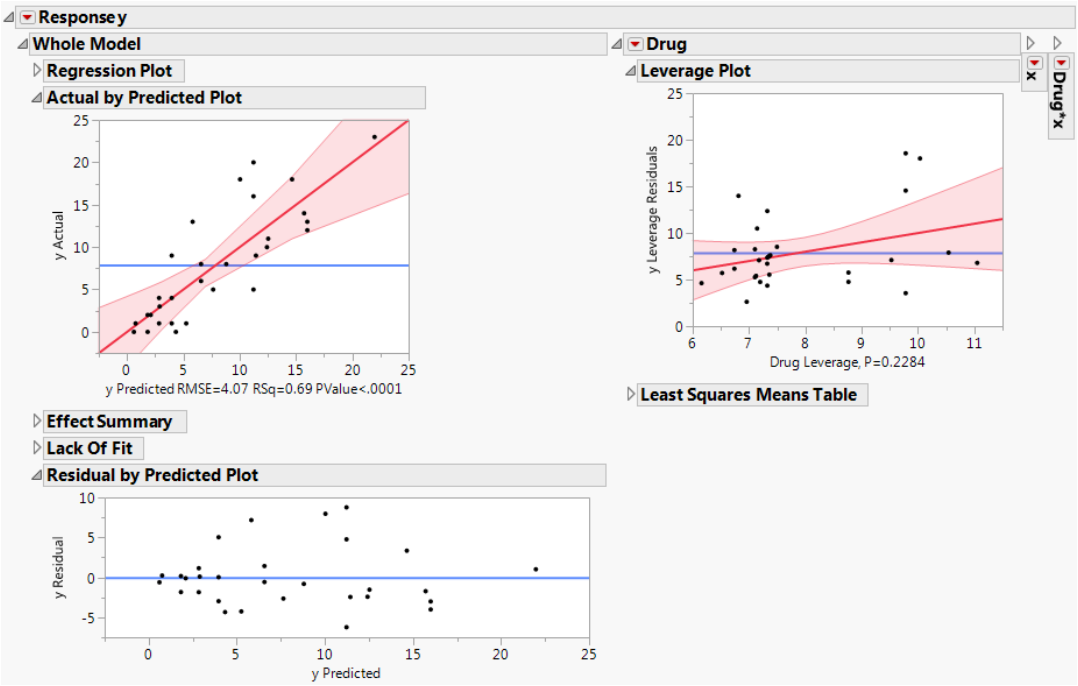
Figure 3.2 Completed Fit Model Launch Window



5. Click **Run**.

The Fit Least Squares report is shown in Figure 3.3. Note that some of the constituent reports are closed because of space considerations. The Actual by Predicted, Residual by Predicted, and Leverage plots show no discrepancies in terms of model fit and underlying assumptions.

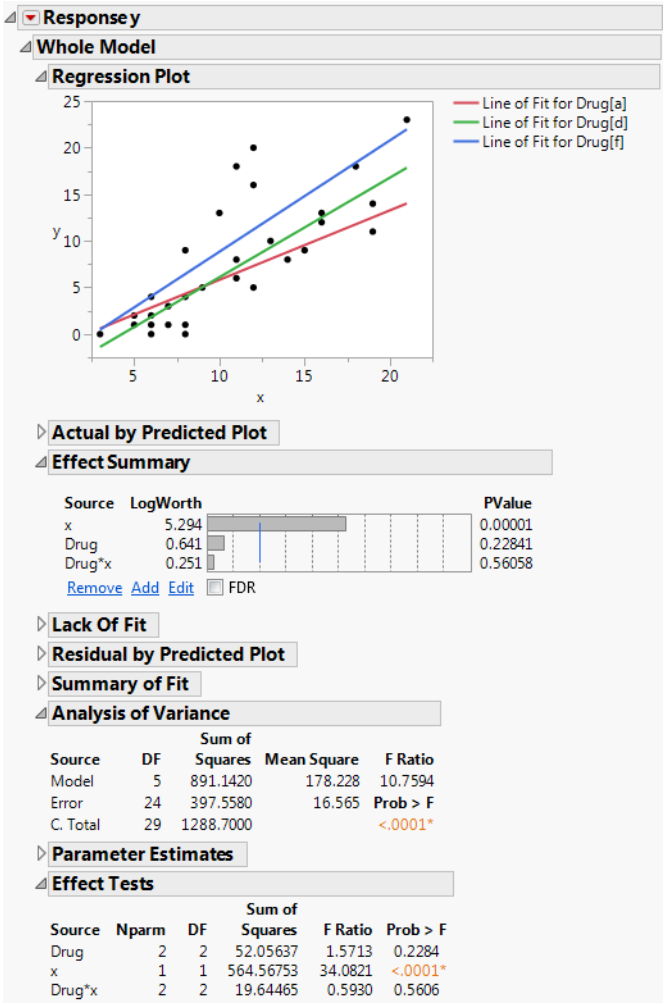
Figure 3.3 Fit Least Squares Report Showing Plots to Assess Model Fit



Since there are no apparent problems with the model fit, you can now interpret the statistical tests. Figure 3.4 shows the relevant reports. The overall model is significant, as shown in the Analysis of Variance report.

Although the Regression Plot suggests that Drug and the pretreatment measure, x, interact, the Prob > F value in the Effect Tests report does not support that conclusion. The Effect Tests report also shows that x is significant in explaining y, but Drug is not significant. The study does not detect a difference among the three groups. However, you cannot conclude that Drug has no effect. The drugs might have different effects, but the study size was not large enough to detect that difference.

Figure 3.4 Fit Least Squares Report Showing Reports to Assess Significance



Launch the Standard Least Squares Personality

Standard least squares is one of several analytic techniques that you can select in the Fit Model launch window.

This section describes how you select standard least squares as your fitting methodology in the Fit Model launch window. Options that are specific to this selection are also covered.

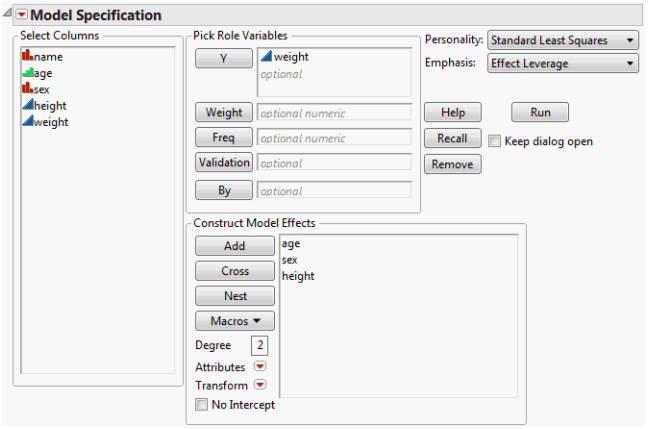
Fit Model Launch Window

You can specify models with both fixed and random effects in the Fit Model launch window. The options differ based on the nature of the model that you specify.

Fixed Effects Only

To fit models using the standard least squares personality, select **Analyze > Fit Model** and then select **Standard Least Squares** from the Personality list. When you enter one or more continuous variables in the Y list, the Personality defaults to Standard Least Squares. Note, however, that other selections are available for continuous Y variables. When you specify only fixed effects for a Standard Least Squares fit, the Fit Model launch window appears as shown in Figure 3.5. This example illustrates the launch window using the Big Class.jmp sample data table.

Figure 3.5 Fit Model Launch Window for a Fixed Effects Model



When the Standard Least Squares personality is selected in the Personality list, an Emphasis option also appears. Emphasis options control the reports that are provided in the initial report window. Based on the model effects that are included, JMP infers which reports you are likely to want. However, any report not shown as part of the initial report can be shown by selecting the appropriate option from the default report's red triangle menu.

For details about reports that are available for each Emphasis option, see "Emphasis Options for Standard Least Squares" on page 84.

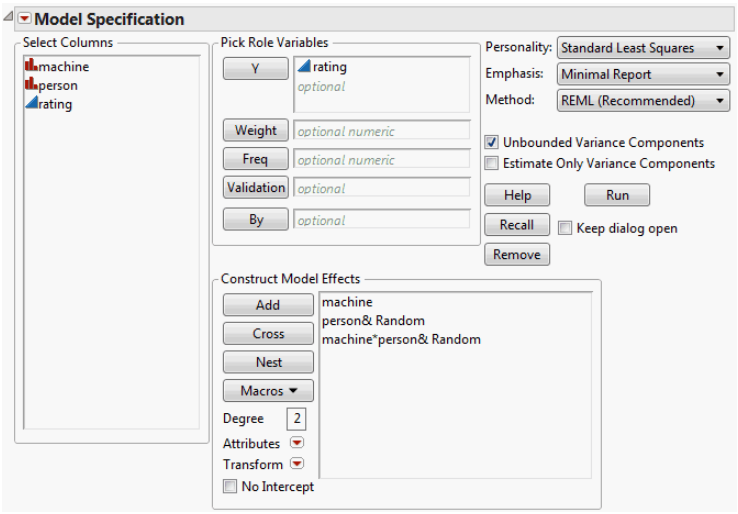
Random Effects

If the specified model contains one or more random effects, then additional options become available in the Fit Model launch window. Consider the Machine.jmp sample data table. Each of six randomly chosen workers performs work at each of three machines and their output is

rated. You are interested in estimating the variation in ratings across the workforce, rather than in determining whether these six specific workers' ratings differ. You need to treat person and machine*person as random effects when you specify the model.

The Fit Model launch window for this model is shown in Figure 3.6. When the Random Effect attribute is applied to person, a Method option and two options relating to variance components appear in the Fit Model Launch window.

Figure 3.6 Fit Model Launch Window for a Model Containing a Random Effect



Standard Least Squares Options in the Fit Model Launch Window

The following Fit Model launch window options are specific to the Standard Least Squares personality.

Emphasis Controls the types of reports and plots that appear in the initial report window. See "Emphasis Options for Standard Least Squares" on page 84.

Method Only appears when random effects are specified. Estimates the model using one of these methods:

- REML. See "REML Variance Component Estimates" on page 193.
- EMS. Expected Mean Squares, also called the Method of Moments. See "EMS (Traditional) Model Fit Reports" on page 196.

Unbounded Variance Components (Appears only when REML is selected as the Method.) Selecting Unbounded Variance Components allows variance component estimates to be negative. This option is selected by default. This option should remain selected if you are

interested in fixed effects, since bounding the variance estimates at zero leads to bias in the tests for fixed effects. See "Negative Variances" on page 190.

Estimate Only Variance Components (Appears only when REML is selected as the Method.) Provides a report that shows only variance component estimates. See "Estimate Only Variance Components" on page 191.

Fit Separately (Appears only for models with multiple Y variables and no random effects.) Fits a separate model for each Y variable using all rows that are nonmissing. If this option is not selected and there are no missing values, the Fit Least Squares report contains individual reports for each of the Y variables. However, some parts of the report are combined for all Y variables. See "Missing Values" on page 85.

Emphasis Options for Standard Least Squares

The three options in the Emphasis list control the types of plots and reports that you see as part of the initial report for the Standard Least Squares personality. See the descriptions below. JMP chooses a default emphasis based on the number of rows in the data table, the number of effects entered in the Construct Model Effects list, and the attributes applied to effects. You can change this choice of emphasis based on your needs. For details about how JMP chooses the emphasis, see "Emphasis Rules" on page 204.

After the initial report opens, you can add other reports and plots from the red triangle menu in the platform report window.

The following emphasis options are available:

Effect Leverage Shows leverage and residual plots, as well as reports with details about the model fit. This option is useful when your main focus is model fitting.

Effect Leverage is the most comprehensive option. This emphasis divides reports into those that relate to the Whole Model and those that relate to individual model effects. The Whole Model reports are in the left corner of the report window under the Whole Model title, with effect reports to the right.

Effect Screening Displays a sorted or scaled parameter estimates report along with a graph (when appropriate), the Prediction Profiler, and reports with details about the model fit. This option is useful when you have many effects and your initial focus is to discover which effects are active, as in screening designs.

Minimal Report Displays only the regression plot and reports with details about the model fit. This Emphasis is the default when the Random Effect attribute is applied to any model effect.

This option is the least detailed and most concise. You can request reports of specific interest to you from the red triangle menus.

To change which reports or plots appear for all of the Emphasis options, use platform preferences. Go to **File > Preferences > Platforms > Fit Least Squares**, and use the **Set** checkboxes as follows:

- To prevent an option from appearing in the report, next to an option, select **Set** but do not select the option.
- To ensure an option appears in the report, select **Set** and select the option.

JMP PRO Validation

If you enter a Validation column, a Crossvalidation report is provided. See "Crossvalidation Report" on page 89.

Missing Values

By default, rows that have missing values for Y or any model effects are excluded from the analysis.

Note: JMP provides an Informative Missing option in the Fit Model window in the Model Specification red triangle menu. Informative Missing enables you to fit models using rows where model effects are missing. See "Informative Missing" on page 53 in the "Model Specification" chapter for details.

When your model contains a random effect, Y values are fit separately by default. The individual reports appear in the Fit Group report.

Suppose that your model contains only fixed effects, and the following statements are true:

- You specified more than one Y response.
- Some of these Y responses have missing values.
- You did not select the Fit Separately option in the Fit Model launch window.

Then, JMP prompts you to select one of the following options:

- Fit Separately fits each Y using all rows that are nonmissing for that particular Y.
- Fit Together fits each Y uses only those rows that are nonmissing for all of the Y variables.

When you select Fit Separately, a Fit Group report contains the individual reports for the Y variables. You can select profilers from the Fit Group red triangle menu to view all the Y variables in the same profiler. Alternatively, you can select a profiler from an individual Y variable report to view only that variable in the profiler.

When you select Fit Together, a Fit Least Squares report contains individual reports for each of the Y variables. However, some parts of the report are combined for all Y variables: the Effect Summary and the Profilers.

Fit Least Squares Report

When you fit a model using the Standard Least Squares personality, you obtain a Fit Least Squares report. The content of the report is driven by the nature of the data and your selections in the Fit Model launch window.

Tip: To always see reports that do not appear by default, select them using **File > Preferences > Platforms > Fit Least Squares**.

Single versus Multiple Responses

When you fit a single response variable Y, the Fit Least Squares window organizes detailed reports in a report entitled “Response Y”. When you fit several responses, reports for individual responses are usually organized in a report entitled “Least Squares Fit”. However, if there is missing response data, and you select the option to Fit Separately, reports for individual responses are organized in a report titled “Fit Group”.

Report Structure Related to Emphasis

When you select the Effect Leverage Emphasis in the Fit Model launch window, the report for a given response is arranged in columns. The left column consists of the Whole Model report, which contains additional reports that pertain to the model. Reports for each effect in the model are shown in the columns to the right of the Whole Model report.

When you select either the Effect Screening or Minimal Report Emphasis in the Fit Model launch window, all reports for each response are arranged in the left column.

Special Reports

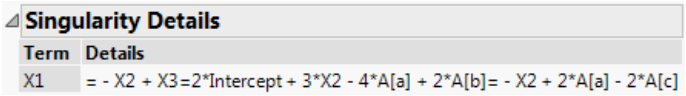
This section describes the reports that are available based on the data structure or choices that you made regarding effect attributes.

Singularity Details

When there are linear dependencies among model effects, the Singularity Details report appears as the first report under the Response report title. It contains a table of the linear

functions that the model terms satisfy. These functions define the aliasing relationships among model terms. Figure 3.7 shows an example for the Singularity.jmp sample data table.

Figure 3.7 Singularity Details Report



Singularity Details

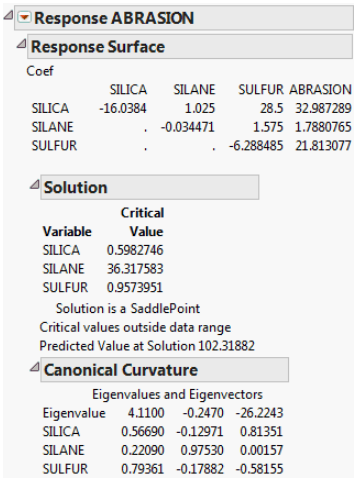
Term	Details
X1	= - X2 + X3=2*Intercept + 3*X2 - 4*A[a] + 2*A[b]= - X2 + 2*A[a] - 2*A[c]

When there are linear dependencies among effects, estimates of some model terms are not unique. See "Models with Linear Dependencies among Model Terms" on page 199.

Response Surface Report

When an effect in a model has the response surface (&RS) or mixture response surface (&RS&Mixture) attribute, a Response Surface report is provided. See Figure 3.8 for an example of a Response Surface report for the Tiretread.jmp sample data table.

Figure 3.8 Response Surface Report



Response ABRASION

Response Surface

Coef

	SILICA	SILANE	SULFUR	ABRASION
SILICA	-16.0384	1.025	28.5	32.987289
SILANE	.	-0.034471	1.575	1.7880765
SULFUR	.	.	-6.288485	21.813077

Solution

Variable	Critical Value
SILICA	0.5982746
SILANE	36.317583
SULFUR	0.9573951

Solution is a SaddlePoint
Critical values outside data range
Predicted Value at Solution 102.31882

Canonical Curvature

Eigenvalues and Eigenvectors

Eigenvalue	4.1100	-0.2470	-26.2243
SILICA	0.56690	-0.12971	0.81351
SILANE	0.22090	0.97530	0.00157
SULFUR	0.79361	-0.17882	-0.58155

Coef Table

The Coef table shown as the first part of the Response Surface report gives a concise summary of the estimated model parameters. The first columns give the coefficients of the second-order terms. The last column gives the coefficients of the linear terms. To see the prediction expression in its entirety, select **Estimates** > **Show Prediction Expression** from the report's red triangle menu.

Solution Report

The Solution report gives a critical value (maximum, minimum, or saddle point), if one exists, along with the predicted value at that point. It also alerts you if the solution falls outside the range of the data.

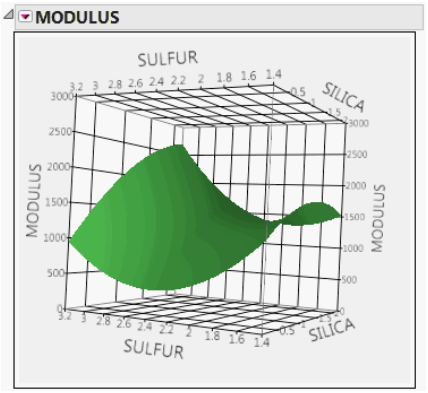
Canonical Curvature Report

The eigenvalues and eigenvectors of the matrix of second-order parameter estimates determine the type of curvature. The eigenvectors show the principal directions of the surface, including the directions of greatest and smallest curvature.

The eigenvalues are provided in the first row of the Canonical Curvature table.

- If the eigenvalues are negative, the response surface curves downward from a maximum.
- If the eigenvalues are positive, the surface shape curves upward from a minimum.
- If there are both positive and negative eigenvalues, the surface is saddle shaped, curving up in one direction and down in another direction. See Figure 3.9 for an example using the Tiretread.jmp sample data table.

Figure 3.9 Surface Profiler Plot with Saddle-Shaped Surface



The eigenvectors listed below the eigenvalues show the orientation of the principal axes. The larger the absolute value of an eigenvalue, the greater the curvature of the response surface in its associated direction. Sometimes a zero eigenvalue occurs. This eigenvalue means that, along the direction described by the corresponding eigenvector, the fitted surface is flat.

Note: The response surface report is not shown for response surface models consisting of more than 20 factors. No error message or alert is given. For more information about response surface designs, see the Response Surface Designs chapter in the *Design of Experiments Guide* book.

Mixed and Random Effect Model Reports

When you specify a random effect in the Fit Model launch window, the Method list appears. This list provides two fitting methods: REML (Recommended) and EMS (Traditional). Additional reports as well as Save Columns and Profiler options are shown, based on the model and the method that you select.

For details about the REML method reports, see "Restricted Maximum Likelihood (REML) Method" on page 191. For details about the EMS method reports, see "EMS (Traditional) Model Fit Reports" on page 196.

JMP PRO Crossvalidation Report

When you enter a Validation column in the Fit Model launch window, a Crossvalidation report is provided. The report gives the following for each of the sets used in validation:

Source Identifies the set as the Training, Validation, or Test set.

RSquare The RSquare value calculated for observations in the given set relative to the model derived using the Training Set. For the Training Set, this is the usual RSquare value.

For each of the Training, Validation, and Test sets, the RSquare value is computed as follows:

- For each observation in the given set, compute the prediction error. This is the difference between the actual response and the response predicted by the Training set model.
- Square and sum the prediction errors to obtain SSE_{Source}, where the subscript *Source* denotes any of the Training, Validation, or Test sets.
- Square and sum the differences between the actual responses for observations in the *Source* set and their mean. Denote this value by SST_{Source}.
- RSquare for the *Source* set is:

$$\text{RSquare}_{Source} = 1 - \frac{SSE_{Source}}{SST_{Source}}$$

Note: It is possible for RSquare values for the Validation and Test sets to be negative.

RASE The square root of the mean squared prediction error. For each of the Training, Validation, and Test sets, RASE is computed as follows:

- For each observation in the given set, calculate the prediction error. This is the difference between the actual response and the response predicted by the Training set model.

- Square and sum the prediction errors to obtain SSE_{Source}, where the subscript *Source* denotes any of the Training, Validation, or Test sets.
- Denote the number of observations by n.
- RASE is:

$$\text{RASE}_{Source} = \sqrt{\frac{SSE_{Source}}{n}}$$

Note: In the Stepwise Fit report, the RASE values for the Validation and Test sets are called RMSE Validation and RMSE Test. See "RMSE Validation" on page 281 in the "Stepwise Regression Models" chapter.

Freq The number of observations in the Source set.

Least Squares Fit Options

You might have more than one Y and no missing response values, or more than one Y with missing values. When you select Fit Together, the responses are grouped in a report called Least Squares Fit. The red triangle menu includes the following options:

Profilers Shows all responses in a single profiler. You can view the effects of model terms on all responses simultaneously and perform multiple optimization. See "Factor Profiling" on page 164.

Model Dialog Shows the completed launch window for the current analysis.

Effect Summary Shows the interactive Effect Summary report that enables you to add or remove effects from the model. See "Effect Summary Report" on page 182.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Fit Group Options

When you have more than one Y and you select Fit Separately, the responses are grouped in a report called Fit Group. The red triangle menu includes the following options:

Profiler Shows all responses in a single profiler. You can view the effects of model terms on all responses simultaneously and perform multiple optimization. See "Profiler" on page 165.

Contour Profiler Shows all responses in a single contour profiler. You can explore the effects of model terms on all responses simultaneously.

Surface Profiler Shows separate surface profiler reports for each response.

Arrange in Rows Rearranges the reports for the platform analyses in a specified number of rows. This would be used mostly to arrange reports so that more reports fit in a window or on the page of output.

Order by Goodness of Fit Sorts the reports by significance of fit (RSquare). This option is available only for platforms that surface the RSquare to the platform level. For example, if you have hundreds of Oneway analyses generated from one launch window, they will appear in a `FitGroup` and you can sort them so that the strongest relationships appear first.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Response Options

Red triangle menu options for the response give you the ability to customize reports according to your needs.

Regression Reports Provides basic reports and report options. See "Regression Reports" on page 93.

Estimates Provides options for further analyses relating to parameter estimates. See "Estimates" on page 115.

Effect Screening Provides reports and plots for identifying significant effects. See "Effect Screening" on page 153.

Factor Profiling Provides profilers, interaction, and cube plots to examine how the response is related to the model terms. Also provides a plot and report for fitting a Box-Cox transformation. See "Factor Profiling" on page 164.

Row Diagnostics Provides plots and reports for examining residuals. Also reports the PRESS statistic and provides a Durbin-Watson test. See "Row Diagnostics" on page 173.

Save Columns Saves model results as columns in the data table, except for Save Coding Table, which saves results in a separate data table. See "Save Columns" on page 179.

Model Dialog Shows the completed Fit Model launch window for the current analysis.

Effect Summary Shows the interactive Effect Summary report that enables you to add or remove effects from the model. See "Effect Summary Report" on page 182.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Regression Reports

The Regression Reports menu provides summary information about model fit, effect significance, and model parameters.

Summary of Fit Shows or hides a summary of model fit. See "Summary of Fit" on page 93.

Analysis of Variance Shows or hides calculations for comparing the fitted model to a simple mean model. See "Analysis of Variance" on page 95.

Parameter Estimates Shows or hides a report containing the parameter estimates and *t* tests for the hypothesis that each parameter is zero. See "Parameter Estimates" on page 96.

Effect Tests Shows or hides tests for the fixed effects in the model. See "Effect Tests" on page 97.

Effect Details Shows or hides a report containing details, plots, and tests for individual effects. See "Effect Details" on page 98.

When the Effect Leverage Emphasis option is selected, each effect has its own report at the top of the Fit Least Squares report window. This report includes effect details options as well as a leverage plot. See "Leverage Plots" on page 175.

Lack of Fit Shows or hides a test assessing if the model has the appropriate effects, when that test can be conducted. See "Lack of Fit" on page 113.

Show All Confidence Intervals Shows or hides confidence intervals for the following statistics:

- Parameter estimates in the Parameter Estimates report
- Least squares means in the Least Squares Means Table

AICc Shows or hides AICc and BIC values in the Summary of Fit report. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

Summary of Fit

The Summary of Fit report provides details such as RSquare calculations and the AICc and BIC values.

RSquare Estimates the proportion of variation in the response that can be attributed to the model rather than to random error. Using quantities from the corresponding Analysis of Variance table, RSquare (also called the *coefficient of multiple determination)* is calculated as:

$$\frac{\text{Sum of Squares(Model)}}{\text{Sum of Squares(C. Total)}}$$

An RSquare closer to 1 indicates a better fit to the data than does an RSquare closer to 0. An RSquare near 0 indicates that the model is not a much better predictor of the response than is the response mean.

Note: A low RSquare value suggests that there may be variables not in the model that account for the unexplained variation. However, if your data are subject to a large amount of inherent variation, even a useful regression model may have a low RSquare value. Read the literature in your research area to learn about typical RSquare values.

Rsquare Adj Adjusts the RSquare statistic for the number of parameters in the model. Rsquare Adj facilitates comparisons among models with different numbers of parameters. The computation uses the degrees of freedom. Using quantities from the corresponding Analysis of Variance table, RSquare Adj is calculated as:

$$1 - \frac{\text{Mean Square(Error)}}{\text{Sum of Squares (C. Total)}/\text{DF(C. Total)}}$$

Root Mean Square Error Estimates the standard deviation of the random error. This quantity is the square root of the Mean Square for Error in the Analysis of Variance report.

Note: Root Mean Square Error is commonly known as *RMSE*.

Mean of Response Shows the overall mean of the response values.

Observations (or Sum Wgts) Shows the number of observations used in the model.

- This value is the same as the number of rows in the data table under the following conditions: there are no missing values, no excluded rows, and no column assigned to the role of Weight or Freq.
- This value is the sum of the positive values in the Weight column if there is a column assigned to the role of Weight.
- This value is the sum of the positive values in the Freq column if there is a column assigned to the role of Freq.

AICc Shows the corrected Akaike Information Criterion value (AICc) and the Bayesian Information Criterion value (BIC). For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

Note: AICc and BIC appear only if you have selected the AICc option from the Regression Reports menu or if you have set AICc as a Fit Least Squares preference.

Analysis of Variance

The Analysis of Variance report provides the calculations for comparing the fitted model to a model where all predicted values equal the response mean.

Note: If either a Frequency or a Weight variable is entered in the Fit Model launch window, the entries in the Analysis of Variance report are adjusted in keeping with the descriptions in "Frequency" on page 41 and "Weight" on page 42.

The Analysis of Variance report contains the following columns:

Source The three sources of variation: Model, Error, and C. Total (Corrected Total).

DF The associated *degrees of freedom* (DF) for each source of variation. The C. Total DF is always one less than the number of observations, and it is partitioned into degrees of freedom for the Model and Error as follows:

- The Model DF is the number of parameters (other than the intercept) used to fit the model.
- The Error DF is the difference between the C. Total DF and the Model DF.

Sum of Squares The associated Sum of Squares (SS) for each source of variation.

- The total (C. Total) SS is the sum of the squared differences between the response values and the sample mean. It represents the total variation in the response values.
- The Error SS is the sum of the squared differences between the fitted values and the actual values. It represents the variability that remains unexplained by the fitted model.
- The Model SS is the difference between C. Total SS and Error SS. It represents the variability explained by the model.

Mean Square The mean square statistics for the Model and Error sources of variation. Each Mean Square is the sum of squares divided by its corresponding DF.

Note: The square root of the Mean Square for Error is the same as RMSE in the Summary of Fit report.

F Ratio The model mean square divided by the error mean square. The F Ratio is the test statistic for a test of whether the model differs significantly from a model where all predicted values are the response mean.

Prob > F The *p*-value for the test. The Prob > F value measures the probability of obtaining an F Ratio as large as what is observed, given that all parameters except the intercept are zero. Small values of Prob > F indicate that the observed *F* Ratio is unlikely. Such values are considered evidence that there is at least one significant effect in the model.

Parameter Estimates

The Parameter Estimates report shows the estimates of the model parameters and, for each parameter, gives a *t* test for the hypothesis that it equals zero.

Note: Estimates are obtained and tested, if possible, even when there are linear dependencies among the model terms. Such estimates are labeled Biased or Zeroed. For details, see "Models with Linear Dependencies among Model Terms" on page 199.

Term The model term corresponding to the estimated parameter. The first term is always the intercept, unless the No Intercept option was checked in the Fit Model launch window. Continuous effects appear with the name of the data table column. Note that continuous columns that are part of higher order terms might be centered. Nominal or ordinal effects appear with values of levels in brackets. See "Coding for Nominal Effects" on page 152 and the "The Factor Models" on page 526 in the "Statistical Details" appendix for information about the coding of nominal and ordinal terms.

Estimate The parameter estimates for each term. These are the estimates of the model coefficients. When there are linear dependencies among model terms, these might be labeled as Biased or Zeroed. See "Models with Linear Dependencies among Model Terms" on page 199.

Std Error The estimates of the standard errors for each of the estimated parameters.

t **Ratio** The tests of whether the true value of each parameter is zero. The *t* Ratio is the ratio of the estimate to its standard error. Given the usual assumptions about the model, the *t* Ratio has a Student's *t* distribution under the null hypothesis.

Prob>|t| The *p*-value for the test that the true parameter value is zero, against the two-sided alternative that it is not.

Lower 95% The lower 95% confidence limit for the parameter estimate. This column appears only if you have the Regression Reports > Show All Confidence Intervals option selected or if you right-click in the report and select Columns > Lower 95%.

Upper 95% The upper 95% confidence limit for the parameter estimate. This column appears only if you have the Regression Reports > Show All Confidence Intervals option selected or if you right-click in the report and select Columns > Upper 95%.

Std Beta The parameter estimates for a regression model where all of the terms have been standardized to a mean of 0 and a variance of 1. This column appears only if you right-click in the report and select Columns > Std Beta.

VIF The variance inflation factor for each term in the model. High VIFs indicate a collinearity issue among the terms in the model.

The VIF for the i^{th} term, x_i, is defined as follows:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the RSquare, or *coefficient of multiple determination*, for the regression of x_i as a function of the other explanatory variables. This column appears only if you right-click in the report and select Columns > VIF.

Design Std Error The square roots of the *relative variances* of the parameter estimates (Goos and Jones 2011, p. 25):

$$\sqrt{diag(X'X)^{-1}}$$

These are the standard errors divided by RMSE. This column appears only if you right-click in the report and select Columns > Design Std Error.

Effect Tests

The Effect Tests report only appears when there are fixed effects in the model. The effect test for a given effect tests the null hypothesis that all parameters associated with that effect are zero. An effect might have only one parameter as for a single continuous explanatory variable. In this case, the test is equivalent to the *t* test for that term in the Parameter Estimates report. A nominal or ordinal effect can have several associated parameters, based on its number of levels. The effect test for such an effect tests whether all of the associated parameters are zero.

Note the following:

- Effect tests are conducted, when possible, for effects whose terms are involved in linear dependencies. For details, see "Models with Linear Dependencies among Model Terms" on page 199.
- Parameterization and handling of singularities differ from the SAS GLM procedure. For details about parameterization and handling of singularities, see the "The Factor Models" on page 526 in the "Statistical Details" appendix.

The Effects Test report contains the following columns:

Source The effects in the model.

Nparm The number of parameters associated with the effect. A continuous effect has one parameter. The number of parameters for a nominal or ordinal effect is one less than its number of levels. The number of parameters for a crossed effect is the product of the number of parameters for each individual effect.

DF The degrees of freedom for the effect test. Ordinarily, Nparm and DF are the same. They can differ if there are linear dependencies among the predictors. In such cases, DF might be less than Nparm, indicating that at least one parameter associated with the effect is not testable. Whenever DF is less than Nparm, the note LostDFs appears to the right of the line in the report. If there are degrees of freedom for error, the test is conducted. For details, see "Effect Tests Report" on page 201.

Sum of Squares The sum of squares for the hypothesis that the effect is zero.

F Ratio The F statistic for testing that the effect is zero. The F Ratio is the ratio of the mean square for the effect divided by the mean square for error. The mean square for the effect is the sum of squares for the effect divided by its degrees of freedom.

Prob > F The *p*-value for the effect test.

Mean Square The mean square for the effect, which is the sum of squares for the effect divided by its DF.

Note: Appears only if you right-click in the report and select **Columns** > **Mean Square**.

Effect Details

The Effect Details report provides details, plots, and tests for individual effects. It consists of separate reports based on the emphasis that you select in the Fit Model launch window.

- Effect Leverage emphasis: Each effect has its own report at the top of the Fit Least Squares report window to the right of the Whole Model report. In this case, the report includes a Leverage Plot for the effect.
- Effect Screening or Minimal Report emphases: The Effect Details report is provided but is initially closed. Click the disclosure icon to show the report.

The initial content of the report is the Table of Least Squares Means. Depending on the nature of the effect, this table might not be appropriate, and the default report might initially show no content. However, certain red triangle options are available.

Table of Effect Options

The red triangle menu next to an effect name provides the following options. For certain modeling types, some of these options might not be appropriate and are therefore not available.

LSMeans Table Shows the statistics that are compared when effects are tested. See "LSMeans Table" on page 99.

This option is not enabled for continuous effects.

LSMeans Plot Shows plots of least squares means for nominal and ordinal effects. If the effect is an interaction, this option displays the Least Squares Means Plot Options window. See "LSMeans Plot" on page 101.

LSMeans Contrast Shows the Contrast Specification window, which enables you to specify and test contrasts to compare levels for nominal and ordinal effects and their interactions. See "LSMeans Contrast" on page 104.

LSMeans Student's t Shows tests and confidence intervals for pairwise comparisons of least squares means using Student's *t* tests. See "LSMeans Student's t and LSMeans Tukey HSD" on page 107.

Note: The significance level applies to individual comparisons and *not* to all comparisons collectively. The error rate for the collection of comparisons is greater than the error rate for individual tests.

LSMeans Tukey HSD Shows tests and confidence intervals for pairwise comparisons of least squares means using the *Tukey-Kramer HSD* (Honestly Significant Difference) test (Tukey 1953; Kramer 1956). See "LSMeans Student's t and LSMeans Tukey HSD" on page 107.

Note: The significance level applies to the collection of pairwise comparisons. The significance level is exact if the sample sizes are equal and conservative if the sample sizes differ (Hayter 1984).

LSMeans Dunnett Shows tests and confidence intervals for pairwise comparisons against a control level that you specify. Also provides a plot of test results. See "LSMeans Dunnett" on page 110.

Test Slices For each level of each column in the interaction, jointly tests pairwise comparisons among all the levels of the other classification columns in the interaction. See "Test Slices" on page 111.

Note: Available only for interactions involving nominal and ordinal effects.

Power Analysis Shows the Power Details report, which enables you to analyze the power for the effect test. For details, see "Power Analysis" on page 112.

LSMeans Table

Least squares means are values predicted by the model for the levels of a categorical effect where the other model factors are set to *neutral* values. The neutral value for a continuous effect is defined to be its sample mean. The neutral value for a nominal effect that is not involved in the effect of interest is the average of the coefficients for that effect. The neutral

value for an uninvolved ordinal effect is defined to be the first level of the effect in the value ordering.

Least squares means are also called *adjusted means* or *population marginal means*. Least squares means can differ from simple means when there are other effects in the model. In fact, it is common for the least squares means to be closer together than the sample means. This situation occurs because of the nature of the neutral values where these predictions are made.

Because least squares means are predictions at specific values of the other model factors, you can compare them. When effects are tested, comparisons are made using the least squares means. For further details about least squares means, see "Least Squares Means across Nominal Factors" on page 531 in the "Statistical Details" appendix and "Ordinal Least Squares Means" on page 541.

For main effects, the Least Squares Means Table also includes the sample mean (Figure 3.10).

Example of a Least Squares Means Table

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Model**.
3. Select weight and click **Y**.
4. Select age, sex, and height and click **Add**.
5. From the **Emphasis** list, select **Effect Screening**.
6. Click **Run**.
7. The Effect Details report appears near the bottom of the Fit Least Squares report and is initially closed. Click the disclosure icon next to the Effect Details report title to show the report.

The Effect Details report, shown in Figure 3.10, shows reports for each of the three effects. Least Squares Means tables are given for age and sex, but not for the continuous effect height. Notice how the least squares means differ from the sample means.

Figure 3.10 Least Squares Mean Table

Response weight

Effect Details

age

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
12	119.38671	5.5302002	99.000
13	105.70112	5.2695143	94.714
14	93.53945	3.9461891	100.833
15	99.61686	5.1729356	108.286
16	109.22769	7.7788438	118.333
17	121.90696	8.0768318	140.667

sex

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
F	121.43592	6.1844278	100.944
M	117.33750	5.7895918	108.318

height

The Least Squares Means report contains the following columns:

Level The categorical levels or combination of levels.

Least Sq Mean An estimate of the least squares mean for each level.

Estimability (Appears only when a least squares mean is not estimable.) A warning if a least squares mean is not estimable.

Std Error Shows the standard error of the least squares mean for each level.

Lower 95% Shows the lower 95% confidence limit for the least squares mean. This column appears only if you have the Regression Reports > Show All Confidence Intervals option selected or if you right-click in the report and select Columns > Lower 95%.

Upper 95% Shows the upper 95% confidence limit for the least squares mean. This column appears only if you have the Regression Reports > Show All Confidence Intervals option selected or if you right-click in the report and select Columns > Upper 95%.

Mean Shows the response sample mean for the given level. This mean differs from the least squares mean if the values for other effects in the model do not balance out across this effect.

LSMeans Plot

The LSMeans Plot option produces a Least Squares Means Plot for nominal and ordinal main effects and their interactions. If the effect is an interaction, this option displays the Least Squares Means Plot Options window. See "Least Squares Means Plot Options" on page 102.

The Least Squares Means Plot red triangle menu contains the following options:

Show Confidence Limits Shows or hides confidence limits for each estimate in the plot.

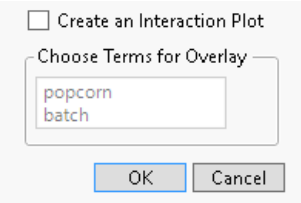
Show Connected Points Shows or hides one or more lines that connect the least squares means for each level in the plot.

Remove Removes the Least Squares Means Plot report for the specified effect.

Least Squares Means Plot Options

When you select the LSMeans Plot option from the red triangle menu of an interaction effect, the Least Squares Means Plot Options window appears.

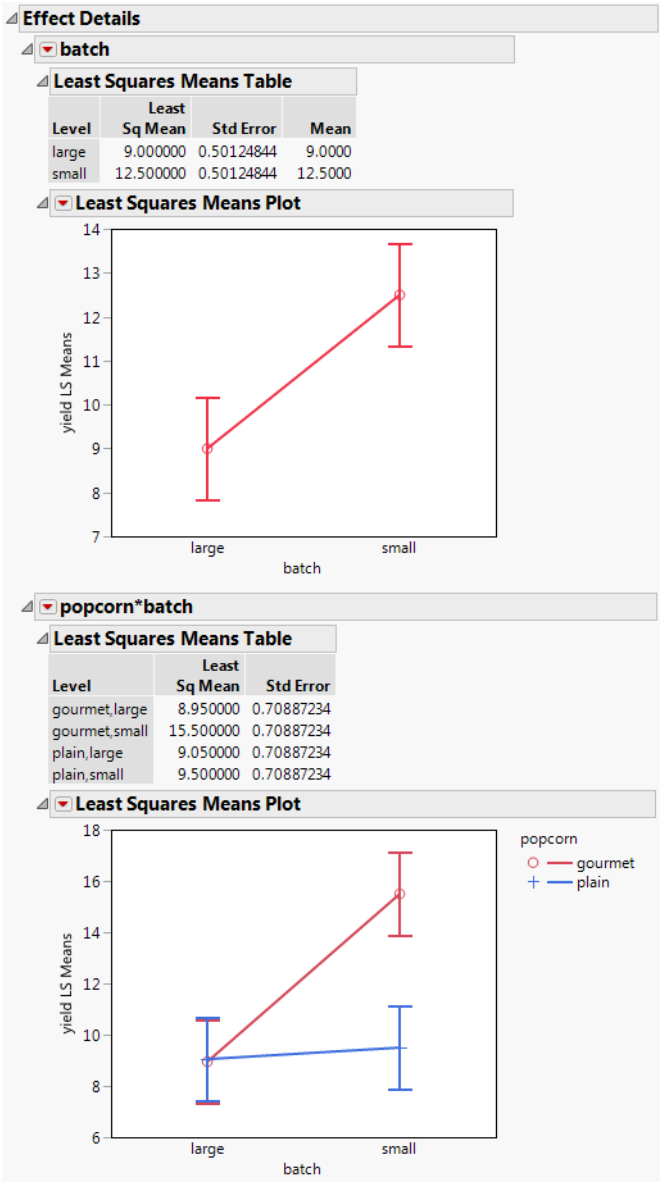
Figure 3.11 Least Squares Means Plot Options Window



If you click OK without selecting anything in the window, one LSMeans Plot appears. The horizontal axis of the plot consists of the levels of the factors nested to obtain a separate effect for each combination. To create an interaction plot, select the box next to Create an Interaction Plot. The Choose Terms for Overlay option enables you to select which effect is displayed as the overlay variable in the interaction plot.

For a three-way interaction term, a second panel of options appears after you choose an overlay variable for the interaction plot. If you click OK without selecting anything in the second panel, one interaction plot appears. Alternatively, use the second panel of options to create separate interaction plots for each level of the effect that you select under Choose Terms for Separate Plots.

Figure 3.12 Least Squares Means Tables and Plots for Two Effects



Example of an LS Means Plot

To create the report in Figure 3.12, follow these steps:

1. Select **Help > Sample Data Library** and open Popcorn.jmp.
2. Select **Analyze > Fit Model**.
3. Select yield and click **Y**.

4. Select popcorn, oil amt, and batch and click **Macros > Full Factorial**. Note that the Emphasis changes to Effect Screening.
5. Click **Run**.
6. Click the Effect Details disclosure icon to show the details for the seven model effects.
7. Select **LSMeans Plot** from the batch red triangle menu.
8. Select **LSMeans Plot** from the popcorn*batch red triangle menu. The Least Squares Means Plot Options window appears.
9. In the Least Squares Means Plot Options window, click the box next to Create an Interaction Plot.
10. Under Choose Terms for Overlay, select popcorn.
11. Click **OK**.
12. To transpose the factors in the plot for popcorn*batch, repeat step 8 and step 9.
13. Under the Choose Terms for Overlay, select batch and click **OK**.

Figure 3.13 LSMeans Plot for Interaction with Factors Transposed

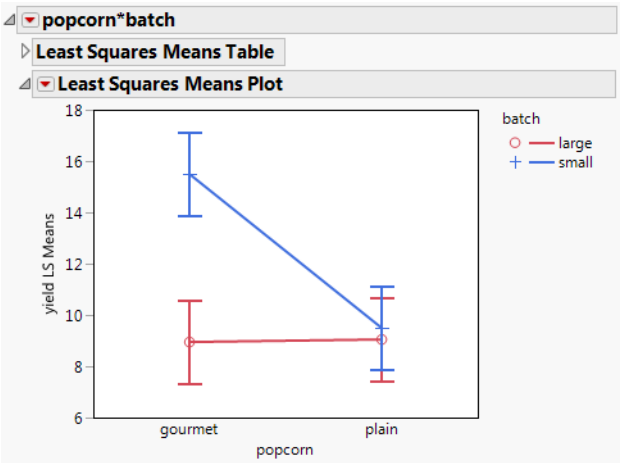


Figure 3.13 shows the popcorn*batch interaction plot with the factors transposed. Compare it with the plot in Figure 3.12. These plots depict the same information but, depending on your interest, one might be more intuitive than the other.

LSMeans Contrast

A *contrast* is a linear combination of parameter values. In the Contrast Specification window, you can specify multiple contrasts and jointly test whether they are zero (Figure 3.14).

JMP builds contrasts in terms of the least squares means of the effect. Each column of the contrast is normalized to have sum zero and so that the sum of the absolute values equals two.

If a contrast involves a covariate, you can specify the value of the covariate at which to test the contrast.

The Contrast Specification box shows the name of the effect and the names of the levels in the effect. The contrast values, which are initially set to zero, appear next to cells containing + and - signs. Click these buttons to compare levels.

Each time you click the + or - button, the contrast coefficients are normalized to make their sum zero and their absolute sum equal to two, if possible. To compare additional levels, click the **New Column** button. A new column appears in which you define a new contrast. After you are finished, click **Done**. The Contrast report appears (Figure 3.15). The overall test is a joint *F* test for all contrasts.

Note: If you attempt to specify more than the maximum number of contrasts possible, the test automatically evaluates.

The Contrast report provides the following details about the joint *F* test:

SS sum of squares for the joint test

NumDF numerator degrees of freedom

DenDF denominator degrees of freedom

F Ratio ratio of SS divided by NumDF divided by the mean square error

Prob > F *p*-value for the significance test

Test Detail Report

The Test Detail report (Figure 3.15) shows a column for each contrast that you tested. For each contrast, the report gives its estimated value, its standard error, a *t* ratio for a test of that single contrast, the corresponding *p*-value, and its sum of squares.

Parameter Function Report

The Parameter Function report (Figure 3.15) shows the contrasts that you specified expressed as linear combinations of the terms of the model.

Example of LSMeans Contrast

To illustrate the LSMeans Contrast option, form a contrast that compares the first two age levels with the next two levels.

Follow these steps to create the report shown in Figure 3.14.

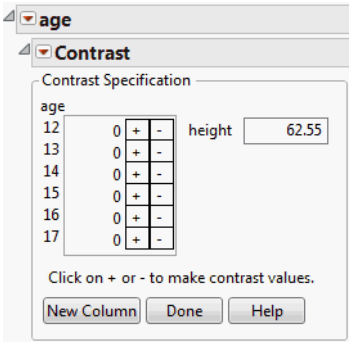
1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Model**.

3. Select weight and click **Y**.
4. Select age, sex, and height, and click **Add**.
5. Select age in the Select Columns list, select height in the Construct Model Effects list, and click **Cross**.
6. Click **Run**.

   The Fit Least Squares report appears.
7. From the red triangle menu next to age, select **LSMeans Contrast**.

   The Contrast Specification report shown in Figure 3.14 appears.

Figure 3.14 LSMeans Contrast Specification for age

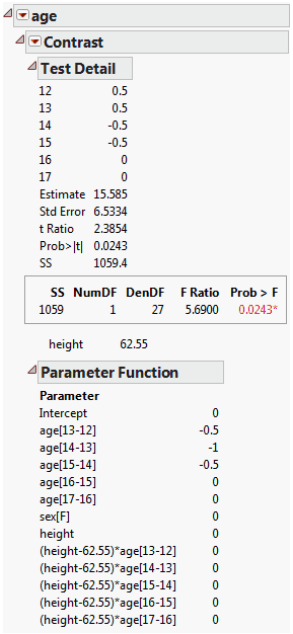


8. Click "+" for the ages 12 and 13.
9. Click "-" for ages 14 and 15.

   This contrast tests whether the mean weights differ for the two age groups, based on predicted values at a height of 62.55.
10. Note that there is a text box next to the continuous effect height. The default value is the mean of the continuous effect.
11. Click **Done**.
12. Open the Test Detail and Parameter Function reports.

The Contrast report is shown in Figure 3.15. The test for the contrast is significant at the 0.05 level. You conclude that the predicted weight for age 12 and 13 children differs statistically from the predicted weight for age 14 and 15 children at the mean height of 62.55.

Figure 3.15 LSMeans Contrast Report



LSMeans Student's t and LSMeans Tukey HSD

The LSMeans Student's t and LSMeans Tukey HSD (*honestly significant difference*) options test pairwise comparisons of model effects.

- The LSMeans Student's t option is based on the usual independent samples, equal variance *t* test. Each comparison is based on the specified significance level. The overall error rate resulting from conducting multiple comparisons exceeds that specified significance level.
- The LSMeans Tukey HSD option conducts Tukey HSD tests. For these comparisons, the significance level applies to the entire collection of pairwise comparisons. For this reason, confidence intervals for LS Means Tukey HSD are wider than those for LSMeans Student's t. The significance level is exact if the sample sizes are equal and conservative if the sample sizes differ (Hayter 1984).

Figure 3.16 shows the LSMeans Tukey report for the effect age in the Big Class.jmp sample data table. (You can obtain this report by running the **Fit Model** data table script and selecting **LS Means Tukey HSD** from the red triangle menu for age.) By default, the report shows the Crosstab Report and the Connecting Letters Report.

Figure 3.16 LSMeans Tukey HSD Report

LSMeans Differences Tukey HSD

α= 0.050 Q= 3.02917

LSMean[i]	Mean[i]-Mean[j] Std Err Dif Lower CL Dif Upper CL Dif	LSMean[j] 12	13	14	15	16	17
12		0	13.6856	25.8473	19.7698	10.159	-2.5203
		0	6.96659	7.24776	8.04937	9.96733	10.5955
		0	-7.4174	3.89261	-4.613	-20.034	-34.616
		0	34.7886	47.8019	44.1527	40.3517	29.5754
13		-13.686	0	12.1617	6.08426	-3.5266	-16.206
		6.96659	0	6.82786	7.59798	9.69566	10.0859
		-34.789	0	-8.521	-16.931	-32.896	-46.758
		7.41737	0	32.8444	29.0998	25.8432	14.346
14		-25.847	-12.162	0	-6.0774	-15.688	-28.368
		7.24776	6.82786	0	6.28014	8.57237	8.65355
		-47.802	-32.844	0	-25.101	-41.655	-54.581
		-3.8926	8.52105	0	12.9462	10.2789	-2.1545
15		-19.77	-6.0843	6.07741	0	-9.6108	-22.29
		8.04937	7.59798	6.28014	0	9.23282	9.21068
		-44.153	-29.1	-12.946	0	-37.579	-50.191
		4.61302	16.9313	25.101	0	18.3569	5.61058
16		-10.159	3.52657	15.6882	9.61083	0	-12.679
		9.96733	9.69566	8.57237	9.23282	0	10.8916
		-40.352	-25.843	-10.279	-18.357	0	-45.672
		20.0337	32.8963	41.6554	37.5786	0	20.3131
17		2.52026	16.2058	28.3675	22.2901	12.6793	0
		10.5955	10.0859	8.65355	9.21068	10.8916	0
		-29.575	-14.346	2.15448	-5.6106	-20.313	0
		34.6159	46.7577	54.5806	50.1908	45.6716	0

Level		Least Sq Mean
17	A	121.90696
12	A	119.38671
16	A B	109.22769
13	A B	105.70112
15	A B	99.61686
14	B	93.53945

The Crosstab Report

Both options display a matrix, called the Crosstab Report, where each cell contains the difference in means, the standard error of the difference, and lower and upper confidence limits. The significance level and corresponding critical value are given above the matrix. The default significance level is 0.05, but you can specify a different significance level in the Fit Model launch window. Cells that correspond to pairs of means that differ statistically are shown in red.

The Connecting Letters Report

A Connecting Letters Report appears by default beneath the Crosstab matrix. Levels that share, or are connected by, the same letter do not differ statistically. Levels that are not connected by a common letter do differ statistically.

In Figure 3.16, levels 17, 12, 16, 13, and 15 are connected by the letter A. The connection indicates that these levels do not differ at the 0.05 significance level. Also, levels 16, 13, 15, and 14 are connected by the letter B, indicating that they do not differ statistically. However, ages

17 and 14, and ages 12 and 14, are not connected by a common letter, indicating that these two pairs of levels are statistically different.

Tip: Right-click in the connecting letters report and select Columns to add columns containing connecting letters (Letters), standard errors (Std Error), and confidence interval limits (Lower X% and Upper X%). In the Letters column, the connecting letters are concatenated into a single column. The significance and confidence levels are determined by the significance level you specify in the Fit Model launch window using the Set Alpha Option.

LSMeans Student's t and LSMeans Tukey HSD Options

The red triangle options that appear in each report window show or hide optional reports. All of the options below are available for LSMeans Student's t. The first four options are available for LSMeans Tukey HSD. For both LSMeans Student's t and LSMeans Tukey HSD, the Crosstab Report and the Connecting Letters Report are shown by default.

Crosstab Report Shows a two-way table that provides, for each pair of levels, the difference in means, the standard error of the difference, and confidence limits for the difference. The contents of cells containing significant differences are highlighted in red.

Connecting Letters Report Illustrates significant and non-significant comparisons with connecting letters. Levels not connected by the same letter are significantly different. Levels connected by the same letter are not significantly different.

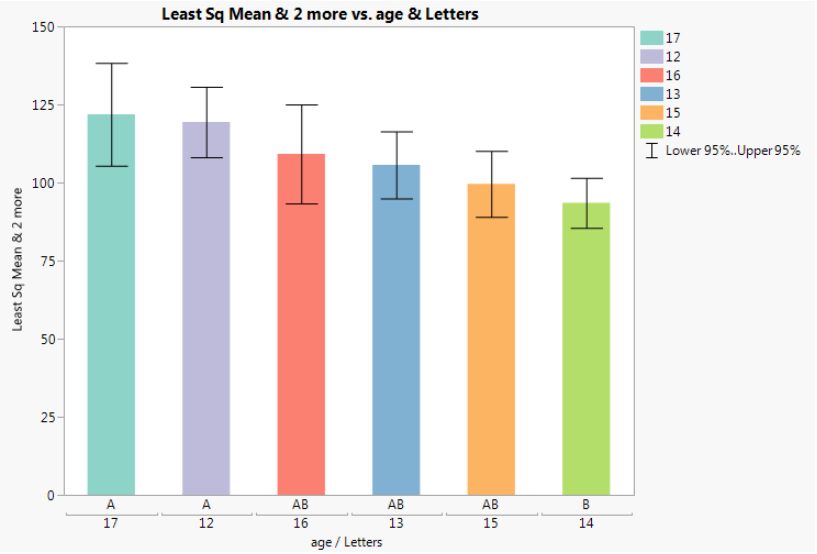
Save Connecting Letters Table Creates a data table whose columns give the levels of the effect, the connecting letters, the least squares means, their standard errors, and confidence intervals. The table contains a script called Bar Chart that produces a colored bar graph of the least squares means with their confidence intervals superimposed. The levels are arranged in decreasing order of least squares means.

Figure 3.17 shows the bar chart for an example based on Big Class.jmp. Run the **Fit Model** data table script, select **LSMeans Tukey HSD** from the red triangle menu for age. Select **Save Connecting Letters Table** from the LSMeans Differences Tukey HSD report. Run the **Bar Chart** script in the data table that appears.

Ordered Differences Report Ranks the differences from largest to smallest, giving standard errors, confidence limits, and *p*-values. Also plots the differences on a bar chart with overlaid confidence intervals.

Detailed Comparisons Gives individual detailed reports for each comparison. For a given comparison, the report shows the estimated difference, standard error, confidence interval, *t* ratio, degrees of freedom, and *p*-values for one- and two-sided tests. Also shown is a plot of the *t* distribution, which illustrates the significance test for the comparison. The area of the shaded portion is the *p*-value for a two-sided test.

Figure 3.17 Bar Chart from LSMeans Differences HSD Connecting Letters Table



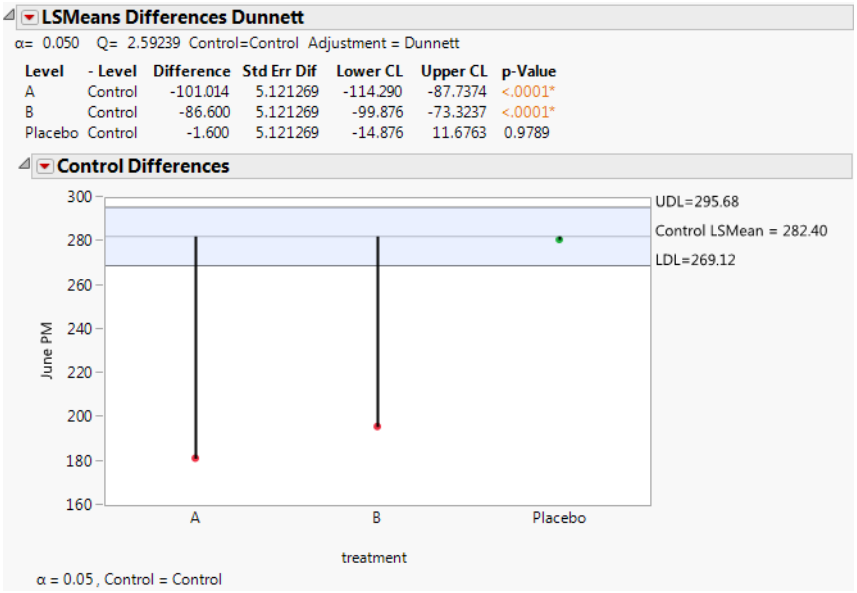
LSMeans Dunnett

Dunnett's test (Dunnett 1955) compares a set of means against the mean of a control group. The error rate applies to the collection of pairwise comparisons. The LSMeans Dunnett option conducts Dunnett's test for the levels of the given effect. Hsu's factor analytical approximation is used for the calculation of p-values and confidence intervals (Hsu 1992).

When you select LSMeans Dunnett, you are prompted to enter a control level for the effect. The LS Means Differences Hsu-Dunnett report shows the significance level, the value of the test statistic (Q), and the control level.

A report for the LSMeans Dunnett option for effect treatment in the Cholesterol.jmp sample data table is shown in Figure 3.18. Here, the response is June PM and the level of treatment called Control is specified as the control level.

Figure 3.18 LSMeans Dunnett Report



The report has two options:

Control Differences Report The Control Differences report is shown by default. For each level of the effect, a table shows the following information: the level being compared to the control level, the estimated difference, the standard error of the difference, a confidence interval, and the *p*-value for the comparison.

Control Differences Chart For each level other than the control, a point shows the difference between the LS Mean for that level and the LS Mean for the control level. Upper and lower decision limits (UDL, LDL) are plotted. The report has a Show Summary Report option and Display options. The Show Summary Report option gives the plot detail. The Display options enable you to modify the plot appearance.

Test Slices

The Test Slices option is enabled for interaction effects composed of nominal or ordinal columns. For each level of each nominal or ordinal column in the interaction, this option produces a report that jointly tests all pairwise comparisons of settings involving that level. The test is effectively a test of differences within the specified “slice” of the interaction.

Suppose that you are interested in an A*B*C interaction, where one of the levels of A is “Small”. The Test Slice report for the slice A = Small jointly tests all pairwise comparisons of the B*C levels when A = Small. It allows you to detect differences in levels within an interaction.

The Test Slice reports follow the same format as do the LSMeans Contrast reports. See "LSMeans Contrast" on page 104.

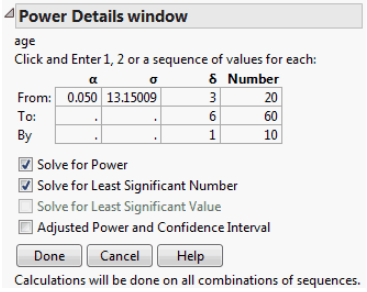
Power Analysis

Opens the Power Details window, where you can enter information to obtain retrospective or prospective details for the *F* test of a specific effect.

Note: To ensure that your study includes sufficiently many observations to detect the required differences, use information about power when you *design* your experiment. Such an analysis is called a *prospective* power analysis. Consider using the DOE platform to design your study. Both DOE > Sample Size and Power and DOE > Evaluate Design are useful for prospective power analysis. For an example of a prospective power analysis using standard least squares, see "Prospective Power Analysis" on page 215.

Figure 3.19 shows an example of the Power Details window for the Big Class.jmp sample data table. Using the Power Details window, you can explore power for values of alpha (α), sigma (σ), delta (δ), and Number (study size). Enter a single value (From only), two values (From and To), or the start (From), stop (To), and increment (By) for a sequence of values. Power calculations are reported for all possible combinations of the values that you specify.

Figure 3.19 Power Details Window



For further details, see "Power Analysis" on page 209.

The Power Details window report contains the following columns and options:

Alpha (α) The significance level of the test. This value is between 0 and 1, and is often 0.05, 0.01, or 0.10. The initial value for Alpha, shown in the first row, is 0.05, unless you have selected Set Alpha Level and set a different value in the Fit Model launch window.

Sigma (σ) An estimate of the residual error in the model. The initial value shown in the first row, provided for guidance, is the RMSE (the square root of the mean square error).

Delta (δ) The effect size of interest. See "Effect Size" on page 210 for details. The initial value, shown in the first row, is the square root of the sum of squares for the hypothesis divided by the square root of the number of observations in the study; that is, $\delta = \sqrt{SS/n}$.

Number (*n*) The sample size. The initial value, shown in the first row, is the number of observations in the current study.

Solve for Power Solves for the power as a function of α, σ, δ, and *n*. The power is the probability of detecting a difference of size δ by seeing a test result that is significant at level α, for the specified σ and *n*. For more details, see "Computations for the Power" on page 553 in the "Statistical Details" appendix.

Solve for Least Significant Number Solves for the smallest number of observations required to obtain a test result that is significant at level α, for the specified δ and σ. For more details, see "Computations for the LSN" on page 552 in the "Statistical Details" appendix.

Solve for Least Significant Value Solves for the smallest positive value of a parameter or linear function of the parameters that produces a *p*-value of α. The least significant value is a function of α, σ, and *n*. This option is available only for one-degree-of-freedom tests. For more details, see "Computations for the LSV" on page 552 in the "Statistical Details" appendix.

Adjusted Power and Confidence Interval Retrospective power calculations use estimates of the standard error and the test parameters in estimating the *F* distribution's noncentrality parameter. *Adjusted power* is retrospective power calculation based on an estimate of the noncentrality parameter from which positive bias has been removed (Wright and O'Brien 1988).

The confidence interval for the adjusted power is based on the confidence interval for the noncentrality estimate.

The adjusted power deals with a sample estimate, so it and its confidence limits are computed only for the δ estimated in the current study. For more details, see "Computations for the Adjusted Power" on page 554 in the "Statistical Details" appendix.

Lack of Fit

The Lack of Fit report gives details for a test that assesses whether the model fits the data well. The Lack of Fit report only appears when it is possible to conduct this test. The test relies on the ability to estimate the variance of the response using an estimate that is independent of the model. Constructing this estimate requires that response values are available at replicated values of the model effects. The test involves computing an estimate of *pure error*, based on a sum of squares, using these replicated observations.

In the following situations, the Lack of Fit report does not appear because the test statistic cannot be computed:

- There are no replicated points with respect to the *X* variables, so it is impossible to calculate a pure error sum of squares.
- The model is *saturated*, meaning that there are as many estimated parameters as there are observations. Such a model fits perfectly, so it is impossible to assess lack of fit.

The difference between the error sum of squares from the model and the pure error sum of squares is called the *lack of fit* sum of squares. The lack of fit variation can be significantly greater than pure error variation if the model is not adequate. For example, you might have the wrong functional form for a predictor, or you might not have enough, or the correct, interaction effects in your model.

The Lack of Fit report contains the following columns:

Source The three sources of variation: Lack of Fit, Pure Error, and Total Error.

DF The degrees of freedom (DF) for each source of error:

- The DF for Total Error is the same as the DF value found on the Error line of the Analysis of Variance table. Based on the sum of squares decomposition, the Total Error DF is partitioned into degrees of freedom for Lack of Fit and for Pure Error.
- The Pure Error DF is pooled from each replicated group of observations. In general, if there are g groups, each with identical settings for each effect, the pure error DF, denoted $\mathrm{DF}_{PE,}$ is as follows:

$$\mathrm{DF}_{PE} = \sum_{i=1}^{g} (n_i - 1)$$

  where n_i is the number of replicates in the i^{th} group.
- The Lack of Fit DF is the difference between the Total Error and Pure Error DFs.

Sum of Squares The associated sum of squares (SS) for each source of error:

- The Total Error SS is the sum of squares found on the Error line of the corresponding Analysis of Variance table.
- The Pure Error SS is the total of the sum of squares values for each replicated group of observations. The Pure Error SS divided by its DF estimates the variance of the response at a given predictor setting. This estimate is unaffected by the model. In general, if there are g groups, each with identical settings for each effect, the Pure Error SS, denoted $\mathrm{SS}_{\mathrm{PE},}$ is as follows:

$$\mathrm{SS}_{PE} = \sum_{i=1}^{g} \mathrm{SS}_i$$

where SS_i is the sum of the squared differences between each observed response and the mean response for the i^{th} group.

– The Lack of Fit SS is the difference between the Total Error and Pure Error sum of squares.

Mean Square The mean square for the Source, which is the Sum of Squares divided by the DF. A Lack of Fit mean square that is large compared to the Pure Error mean square suggests that the model is not fitting well. The *F* ratio provides a formal test.

F Ratio The ratio of the Mean Square for Lack of Fit to the Mean Square for Pure Error. The *F* Ratio tests the hypothesis that the variances estimated by the Lack of Fit and Pure Error mean squares are equal, which is interpreted as representing "no lack of fit".

Prob > F The *p*-value for the Lack of Fit test. A small *p*-value indicates a significant lack of fit.

Max RSq The maximum RSquare that can be achieved by a model based only on these effects. The Pure Error Sum of Squares is invariant to the form of the model. So the largest amount of variation that a model with these replicated effects can explain equals:

$$\frac{SS(\text{C. Total}) - SS(\text{Pure Error})}{SS(\text{C. Total})} = 1 - \frac{SS(\text{Pure Error})}{SS(\text{C. Total})}$$

This formula defines the Max RSq.

Estimates

The Estimates menu provides additional detail about model parameters. To better understand estimates, you might want to review JMP's approach to coding nominal and ordinal effects. See "Details of Custom Test Example" on page 204, "Nominal Factors" on page 527 in the "Statistical Details" appendix, and "Ordinal Factors" on page 538 in the "Statistical Details" appendix).

If your model contains random effects, then only the options below that are appropriate are available from the Estimates menu.

The Estimates menu provides the following options:

Show Prediction Expression Shows or hides the Prediction Expression report, which contains the equation for the estimated model. See "Show Prediction Expression" on page 117 for an example.

Sorted Estimates Shows or hides the Sorted Parameter Estimates report, which can be useful in screening situations. If the design is not saturated, this report is the Parameter Estimates report with the terms, other than the Intercept, sorted in decreasing order of

significance. If the design is saturated, then Pseudo *t* tests are provided. See "Sorted Estimates" on page 117.

Expanded Estimates Shows or hides the Expanded Estimates report, which expands the Parameter Estimates report by giving parameter estimates for all levels of nominal effects. See "Expanded Estimates" on page 121.

Indicator Parameterization Estimates (Available only when there are nominal columns among the model effects.) Shows or hides the Indicator Function Parameterization report, which contains parameter estimates with the nominal effects in the model parametrized using the classical indicator functions. See "Indicator Parameterization Estimates" on page 123.

Sequential Tests Shows or hides the Sequential (Type 1) Tests report that contains the sums of squares as effects are added to the model sequentially. Conducts *F* tests based on the sequential sums of squares. See "Sequential Tests" on page 124.

Custom Test Enables you to test a custom hypothesis. See "Custom Test" on page 125.

Multiple Comparisons Enables you to specify comparisons among effect levels. These comparisons can involve a single effect or you can define flexible custom comparisons. You can compare to the overall mean, to a control mean, or you can obtain all pairwise comparisons using Tukey HSD or Student's *t*. When you specify the Student's *t* method, you can also perform equivalence tests to identify pairwise differences that are of practical importance. See "Multiple Comparisons" on page 128.

Compare Slopes (Available only when there is one nominal term, one continuous term, and their interaction effect for the fixed effects.) Produces a report that enables you to compare the slopes of each level of the interaction effect in an analysis of covariance (ANCOVA) model. See "Compare Slopes" on page 142.

Joint Factor Tests (Available only when the model contains interactions.) For each main effect in the model, shows or hides a joint test on all of the parameters involving that main effect. See "Joint Factor Tests" on page 142.

Inverse Prediction Enables you to predict values of explanatory variables for one or more values of the response. See "Inverse Prediction" on page 143.

Cox Mixtures (Available only when the model contains mixture effects.) Produces parameter estimates for the Cox mixture model. Using these to derive factor effects and estimate the response surface shape relative to a reference point in the design space. See "Cox Mixtures" on page 147.

Parameter Power Adds columns to the Parameter Estimates report that give power and other details relating to the corresponding hypothesis tests. See "Parameter Power" on page 149.

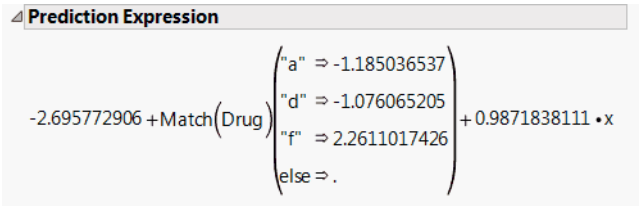
Correlation of Estimates Shows or hides a correlation matrix for all parameter estimates in the model. See "Correlation of Estimates" on page 151.

Show Prediction Expression

The Show Prediction Expression option shows the equation used to predict the response. Figure 3.20 shows an example for the Drug.jmp sample data table. This expression is given as a typical JMP formula. For example, to predict the response for someone on Drug a with x = 10, you would calculate, with some rounding: –2.696 - 1.185 + 0.987(10) = 5.99.

Tip: To specify the number of digits in the prediction formula, go to **File** > **Preferences** > **Tables** and change the **Default Field Width** value.

Figure 3.20 Prediction Expression



To obtain the preceding report, follow these steps:

Example of a Prediction Expression

1. Select **Help > Sample Data Library** and open Drug.jmp.
2. Select **Analyze > Fit Model**.
3. Select y and click **Y**.
4. Select Drug and x, and then click **Add**.
5. Click **Run**.
6. From the red triangle menu next to Response y, select **Estimates > Show Prediction Expression**. The report in Figure 3.20 appears.

Sorted Estimates

The Sorted Estimates option produces a version of the Parameter Estimates report that is useful in screening situations. If the design is not saturated, the Sorted Estimates report gives the information found in the Parameter Estimates report, but with the terms, other than the Intercept, sorted in decreasing order of significance (second report in Figure 3.21). If the

design is saturated, then Pseudo *t* tests are provided. These are based on Lenth's *pseudo standard error* (Lenth 1989). See "Lenth's PSE" on page 119.

Example of a Sorted Estimates Report

1. Select **Help > Sample Data Library** and open Reactor.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Make sure that 2 appears in the **Degree** box near the bottom of the window.
5. Select F, Ct, A, T, and Cn and click **Macros > Factorial to Degree**.
6. Click **Run**.
7. Select **Estimates > Sorted Estimates** from the red triangle menu.

Figure 3.21 Sorted Parameter Estimates

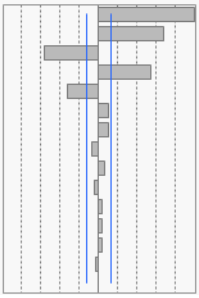
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	65.5	0.565962	115.73	<.0001*
F	-0.6875	0.565962	-1.21	0.2421
Ct	9.75	0.565962	17.23	<.0001*
A	-0.3125	0.565962	-0.55	0.5885
T	5.375	0.565962	9.50	<.0001*
Cn	-3.125	0.565962	-5.52	<.0001*
F*Ct	0.6875	0.565962	1.21	0.2421
F*A	0.375	0.565962	0.66	0.5170
F*T	-0.4375	0.565962	-0.77	0.4508
F*Cn	0.0625	0.565962	0.11	0.9134
Ct*A	0.4375	0.565962	0.77	0.4508
Ct*T	6.625	0.565962	11.71	<.0001*
Ct*Cn	1	0.565962	1.77	0.0963
A*T	1.0625	0.565962	1.88	0.0788
A*Cn	0.4375	0.565962	0.77	0.4508
T*Cn	-5.5	0.565962	-9.72	<.0001*

Effect Tests

Sorted Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Ct	9.75	0.565962	17.23	<.0001*
Ct*T	6.625	0.565962	11.71	<.0001*
T*Cn	-5.5	0.565962	-9.72	<.0001*
T	5.375	0.565962	9.50	<.0001*
Cn	-3.125	0.565962	-5.52	<.0001*
A*T	1.0625	0.565962	1.88	0.0788
Ct*Cn	1	0.565962	1.77	0.0963
F	-0.6875	0.565962	-1.21	0.2421
F*Ct	0.6875	0.565962	1.21	0.2421
F*T	-0.4375	0.565962	-0.77	0.4508
Ct*A	0.4375	0.565962	0.77	0.4508
A*Cn	0.4375	0.565962	0.77	0.4508
F*A	0.375	0.565962	0.66	0.5170
A	-0.3125	0.565962	-0.55	0.5885
F*Cn	0.0625	0.565962	0.11	0.9134



The Sorted Parameter Estimates report also appears automatically if the Emphasis is set to Effect Screening and all of the effects have only one parameter.

Note the following differences between the Parameter Estimates report and the Sorted Parameter Estimates report (both shown in Figure 3.21):

- The Sorted Parameter Estimates report does not show the intercept.
- The effects are sorted by the absolute value of the *t* ratio, showing the most significant effects at the top.
- A bar chart shows the *t* ratio with vertical lines showing critical values for the 0.05 significance level.

Sorted Estimates Report for Saturated Models

Screening experiments often involve fully saturated models, where there are not enough degrees of freedom to estimate error. In these cases, the Sorted Estimates report (Figure 3.21) gives relative standard errors and constructs *t* ratios and *p*-values using Lenth's pseudo standard error (PSE). These quantities are labeled with *Pseudo* in their names. See "Lenth's PSE" on page 119 and "Pseudo t-Ratios" on page 120. A note explains the change and shows the PSE.

The report contains the following columns:

Term The model term whose coefficient is of interest.

Estimate The parameter estimates are presented in sorted order, with smallest *p*-values listed first.

Relative Standard Error If there are no degrees of freedom for residual error, the report gives relative standard errors. The relative standard error is computed by setting the root mean square error equal to 1.

Pseudo t-Ratio A *t* ratio for the estimate, computed using pseudo standard error. The value of Lenth PSE is shown in a note at the bottom of the report.

Pseudo p-Value A *p*-value computed using an error degrees of freedom value (DFE) of $m/3$, where m is the number of parameters other than the intercept. The value of DFE is shown in a note at the bottom of the report.

Lenth's PSE

Lenth's *pseudo standard error* (PSE) is an estimate of residual error due to Lenth (1989). It is based on the principle of effect sparsity: in a screening experiment, relatively few effects are active. The inactive effects represent random noise and form the basis for Lenth's estimate.

The value is computed as follows:

1. Consider the absolute values of all non-intercept parameters.
2. Remove all parameter estimates whose absolute values exceed 3.75 times the median absolute estimate.
3. Multiply the median of the remaining absolute values of parameter estimates by 1.5.

Pseudo t-Ratios

When relative standard errors are equal, Lenth's PSE is shown in a note at the bottom of the report. The Pseudo *t*-Ratio is calculated as follows:

$$\text{Pseudo } t\text{-Ratio} = \frac{\text{Estimate}}{\text{PSE}}$$

When relative standard errors are not equal, the TScale Lenth PSE is computed. This value is the PSE of the estimates divided by their relative standard errors. The Pseudo *t*-Ratio is calculated as follows:

$$\text{Pseudo } t\text{-Ratio} = \frac{\text{Estimate}}{\text{TScale Lenth PSE} \times \text{Relative Std Error}}$$

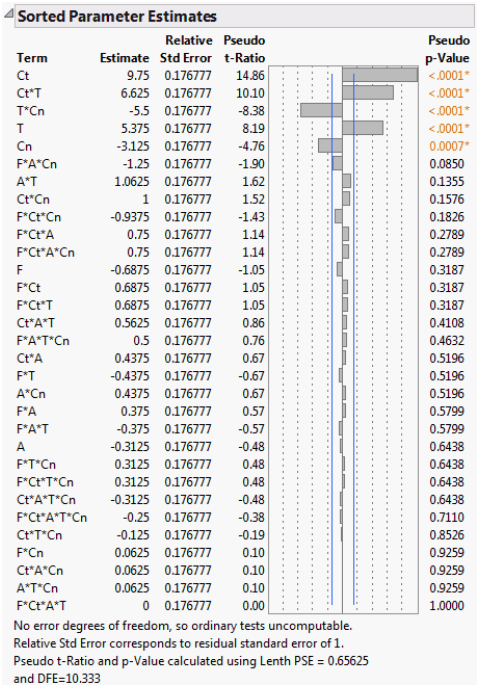
Note that, to estimate the standard error for a given estimate, TScale Lenth PSE is adjusted by multiplying it by the estimate's relative standard error.

Example of a Saturated Model

1. Select **Help > Sample Data Library** and open Reactor.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Select the following five columns: F, Ct, A, T, and Cn.
5. Click the **Macros** button and select **Full Factorial**.
6. Click **Run**.
7. From the red triangle menu next to Response Y, select **Estimates > Sorted Estimates**.

   Note that Lenth's PSE and the degrees of freedom used are given at the bottom of the report. The report indicates that, based on their Pseudo p-Values, the effects Ct, Ct*T, T*Cn, T, and Cn are highly significant.

Figure 3.22 Sorted Parameter Estimates Report for Saturated Model



Sorted Parameter Estimates

Term	Estimate	Relative Std Error	Pseudo t-Ratio	Pseudo p-Value
Ct	9.75	0.176777	14.86	<.0001*
Ct*T	6.625	0.176777	10.10	<.0001*
T*Cn	-5.5	0.176777	-8.38	<.0001*
T	5.375	0.176777	8.19	<.0001*
Cn	-3.125	0.176777	-4.76	0.0007*
F*A*Cn	-1.25	0.176777	-1.90	0.0850
A*T	1.0625	0.176777	1.62	0.1355
Ct*Cn	1	0.176777	1.52	0.1576
F*Ct*Cn	-0.9375	0.176777	-1.43	0.1826
F*Ct*A	0.75	0.176777	1.14	0.2789
F*Ct*A*Cn	0.75	0.176777	1.14	0.2789
F	-0.6875	0.176777	-1.05	0.3187
F*Ct	0.6875	0.176777	1.05	0.3187
F*Ct*T	0.6875	0.176777	1.05	0.3187
Ct*A*T	0.5625	0.176777	0.86	0.4108
F*A*T*Cn	0.5	0.176777	0.76	0.4632
Ct*A	0.4375	0.176777	0.67	0.5196
F*T	-0.4375	0.176777	-0.67	0.5196
A*Cn	0.4375	0.176777	0.67	0.5196
F*A	0.375	0.176777	0.57	0.5799
F*A*T	-0.375	0.176777	-0.57	0.5799
A	-0.3125	0.176777	-0.48	0.6438
F*T*Cn	0.3125	0.176777	0.48	0.6438
F*Ct*T*Cn	0.3125	0.176777	0.48	0.6438
Ct*A*T*Cn	-0.3125	0.176777	-0.48	0.6438
F*Ct*A*T*Cn	-0.25	0.176777	-0.38	0.7110
Ct*T*Cn	-0.125	0.176777	-0.19	0.8526
F*Cn	0.0625	0.176777	0.10	0.9259
Ct*A*Cn	0.0625	0.176777	0.10	0.9259
A*T*Cn	0.0625	0.176777	0.10	0.9259
F*Ct*A*T	0	0.176777	0.00	1.0000

No error degrees of freedom, so ordinary tests uncomputable.
Relative Std Error corresponds to residual standard error of 1.
Pseudo t-Ratio and p-Value calculated using Lenth PSE = 0.65625 and DFE=10.333

Expanded Estimates

In dealing with parameter estimates, you must understand how JMP codes nominal and ordinal columns. For details about how nominal columns are coded, see "Details of Custom Test Example" on page 204. For complete details about how ordinal columns are coded and modeled, see "Nominal Factors" on page 527 in the "Statistical Details" appendix and "Ordinal Factors" on page 538 in the "Statistical Details" appendix.

Use the Expanded Estimates option when there are nominal terms in the model and you want to see details for the full set of estimates. The Expanded Estimates option provides the estimates, their standard errors, *t* ratios, and *p*-values.

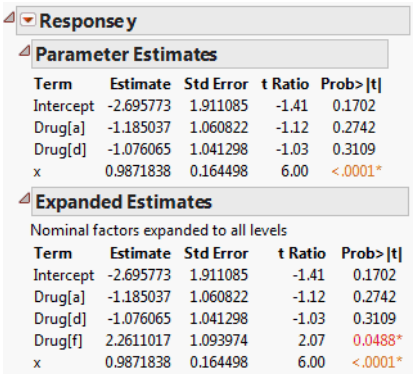
Example of an Expanded Estimates Report

1. Select **Help > Sample Data Library** and open Drug.jmp.
2. Select **Analyze > Fit Model**.
3. Select y and click **Y**.
4. Select Drug and x, and then click **Add**.
5. Click **Run**.

6. From the red triangle menu next to Response y, select **Estimates > Expanded Estimates**.

The Expanded Estimates report, along with the Parameter Estimates report, is shown in Figure 3.23. Note that an estimate for the term Drug[f] appears in the Expanded Estimates report. The null hypothesis for the test is that the mean for the Drug f group does not differ from the overall mean. The test for Drug[f] is significant at the 0.05 level, suggesting that the mean response for the Drug f group differs from the overall response. See "Interpretation of Tests for Expanded Estimates" on page 122 for more details.

Figure 3.23 Comparison of Parameter Estimates and Expanded Estimates



Responsey

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.695773	1.911085	-1.41	0.1702
Drug[a]	-1.185037	1.060822	-1.12	0.2742
Drug[d]	-1.076065	1.041298	-1.03	0.3109
x	0.9871838	0.164498	6.00	<.0001*

Expanded Estimates

Nominal factors expanded to all levels

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	-2.695773	1.911085	-1.41	0.1702
Drug[a]	-1.185037	1.060822	-1.12	0.2742
Drug[d]	-1.076065	1.041298	-1.03	0.3109
Drug[f]	2.2611017	1.093974	2.07	0.0488*
x	0.9871838	0.164498	6.00	<.0001*

Interpretation of Tests for Expanded Estimates

Suppose that your model consists of a single nominal factor that has *n* levels. That factor is represented by *n-1* indicator variables, one for each of *n-1* levels. The parameter estimate corresponding to any one of these *n-1* indicator variables is the difference between the mean response for that level and the average response across all levels. This representation is due to the way that JMP codes nominal variables (see "Details of Custom Test Example" on page 204). The parameter estimate is often interpreted as the *effect* of that level.

For example, in the Cholesterol.jmp sample data table, consider the single factor treatment and the response June PM. The parameter estimate associated with the term, or indicator variable, treatment[A] is the difference between the mean of June PM for treatment A and the overall mean of June PM.

The effects across *all* levels of a nominal variable are constrained to sum to zero. Consider the effect of the last level in the level ordering, namely, the level that is coded with –1s. The effect of this level is the negative of the sum of the effects across the other *n-1* levels. It follows that the effect of the last level is the negative of the sum of the parameter estimates across the other *n-1* levels.

The Expanded Estimates option in the Estimates menu calculates missing estimates, tests for all effects that involve nominal columns, and shows them in a text report. You can verify that the mean (or sum) of the estimates across the levels of any such effect is zero. In particular, this

relationship indicates that these estimates, and their associated tests, are not independent of each other.

In the Drug.jmp report shown in Figure 3.23, the estimates for the terms associated with Drug are based on a model that includes the covariate x.

Notes:

- The estimate for Drug[a] is the difference between the least squares mean for Drug a and the overall mean of y.
- The estimate for Drug[f], given in the Expanded Estimates report, is the negative of the sum of the estimates for Drug[a] and Drug[d].
- The *t* test for Drug [f] presented in the Expanded Estimates report tests whether the response for the Drug f group differs from the overall mean response.
- If nominal factors are involved in high-degree interactions, the Expanded Estimates report can be lengthy. For example, a five-way interaction of two-level nominal factors produces only one parameter estimate but has $2^5 = 32$ expanded effects, which are all identical up to sign changes.

Indicator Parameterization Estimates

This option displays the Indicator Function Parameterization report, which gives parameter estimates for the model where nominal columns are coded using indicator (SAS GLM) parameterization and are treated as continuous. Ordinal columns remain coded using the usual JMP coding scheme. The SAS GLM and JMP coding schemes are described in "The Factor Models" on page 526 in the "Statistical Details" appendix.

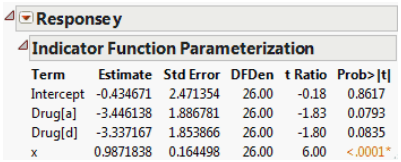
In the JMP coding scheme, the estimate that corresponds to the indicator for a level of a nominal variable is an estimate of the difference between the mean response at that level and the mean response over all the levels. To see the JMP coding, select **Save Columns** > **Save Coding Table** from the Standard Least Squares report's red triangle menu.

In the indicator coding scheme, the estimate that corresponds to the indicator for a level of a nominal variable is an estimate of the difference between the mean response at that level and the mean response at the last level. The last level is the level with the highest value order coding; it is the level whose indicator function is not included in the model.

Caution: Standard errors and t-ratios given in the Indicator Function Parameterization report will differ from those in the Parameter Estimates report. This is because the estimates are estimating different parameters.

To create the report in Figure 3.24, follow the steps in "Example of an Expanded Estimates Report" on page 121. But instead of selecting Expanded Estimates, select **Indicator Parameterization Estimates**.

Figure 3.24 Indicator Parameterization Estimates



Responsey

Indicator Function Parameterization

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	-0.434671	2.471354	26.00	-0.18	0.8617
Drug[a]	-3.446138	1.886781	26.00	-1.83	0.0793
Drug[d]	-3.337167	1.853866	26.00	-1.80	0.0835
x	0.9871838	0.164498	26.00	6.00	<.0001*

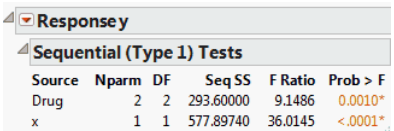
The JMP coding scheme is used for nominal variables throughout JMP with the exception of the Generalized Regression personality of Fit Model. For more information about the coding scheme for nominal variables in the Generalized Regression personality, see "Launch the Generalized Regression Personality" on page 291 in the "Generalized Regression Models" chapter.

Note that there might be differences in models derived using the JMP versus the SAS GLM parameterization. Some models are equivalent. Other models (such as no-intercept models, models with missing cells, models with nominal or ordinal effects, and mixture models) might show differences.

Sequential Tests

The Sequential Tests report shows sums of squares and tests as effects are added to the model sequentially. The order of entry is defined by the order of effects as they appear in the Fit Model launch window's Construct Model Effects list. The report in Figure 3.25 is for the Drug.jmp sample data table.

Figure 3.25 Sequential Tests Report



Responsey

Sequential (Type 1) Tests

Source	Nparm	DF	Seq SS	F Ratio	Prob > F
Drug	2	2	293.60000	9.1486	0.0010*
x	1	1	577.89740	36.0145	<.0001*

The sums of squares that form the basis for sequential tests are also called *Type I Sums of Squares*. They are computed by fitting models in steps following the specified entry order of effects. Consider a specific effect. Compute the model sum of squares for a model containing all effects entered *prior* to that effect. Then compute the model sum of squares for a model containing those effects *and* the specified effect. The sequential sum of squares for the specified effect is the increase in the model sum of squares.

Refer to Figure 3.25, showing sequential sums of squares for the Drug.jmp sample data table. In the Fit Model launch window, Drug was entered first, followed by x. A model consisting only of Drug has model sum of squares equal to 293.6. When x is added to the model, the model sum of squares becomes 871.4974. The increase of 577.8974 is the sequential sum of squares for x.

The tests shown in the Sequential (Type 1) Tests report are *F* tests based on sequential sums of squares, also called *Type I Tests*. The *F* Ratio tests the specified effect, where the model contains only that effect and the effects listed above it in the Source column.

The sequential sums of squares sum to the model sum of squares. Another nice feature is that, under the usual model assumptions, the values are statistically independent of each other. However, they do depend on the order of terms in the model and, as such, are not appropriate in many situations.

Sequential tests are considered appropriate in the following situations:

- balanced analysis of variance models specified in proper sequence (that is, two-way interactions follow main effects in the effects list, and so on)
- purely nested models specified in the proper sequence
- polynomial regression models specified in the proper sequence.

The tests given in the Parameter Estimates and Effect Tests reports are based on *Type III Sums of Squares*. Here the sum of squares for an effect is the extra sum of squares explained by the effect after all other effects have been entered in the model.

Custom Test

To test one or more custom hypotheses involving any model parameters, select **Custom Test** from the Estimates menu. In this window, you can specify one or more linear functions, or *contrasts*, of the model parameters.

The results include individual tests for each contrast and a joint test for all contrasts. See Figure 3.26. The report for the individual contrasts gives the estimated value of the specified linear function of the parameters and its standard error. A *t* ratio, its *p*-value, and the associated sum of squares are also provided. Below the individual contrast results, the joint test for all contrasts gives the sum of squares, the numerator degrees of freedom, the *F* ratio, and its *p*-value.

Caution: These tests are conducted using residual error. If you have random effects in your model and if you use EMS instead of REML, then these tests might not be appropriate.

Note: If you are testing for effects that are involved in higher-order effects, consider using a test for least squares means, rather than a custom test. Least squares means are adjusted for other model effects. You can test least squares means contrasts under Effect Details.

Custom Test Report Components

The Custom Test specification window has the following components:

Editable text box The space beneath the Custom Test title bar is an editable area for entering a test name.

Parameter Lists the model terms. To the right of the list of terms are columns of zeros corresponding to the corresponding parameters. Enter values in these cells to specify the linear functions for your tests.

The "=" sign The last line in the Parameter list is labeled =. Enter a constant into this cell to complete the specification for each contrast.

Add Column Adds columns of zeros so that you can jointly test several linear functions of the parameters.

Done Click the Done button to perform the tests. The report changes to show the test statistic value, the standard error, and other statistics for each test column. The joint *F* test for all columns is given in a box at the bottom of the report.

Custom Test Report Options

The red triangle menu for the Custom Test report has two options:

Power Analysis Provides a power analysis for the joint test. This option is available only after the test has been conducted. For details, see "Parameter Power" on page 149.

Remove Removes the Custom Test report.

Note: Select **Estimates > Custom Test** repeatedly to conduct several joint custom tests.

Figure 3.26 shows an example of the specification window with three contrasts, using the Cholesterol.jmp sample data table. Note that the constant is set to zero for all three tests. The report for these tests is shown in Figure 3.27.

Example of a Custom Test

The Cholesterol.jmp sample data table gives repeated measures on 20 patients at six time periods. Four treatment groups are studied. Typically, this data should be properly analyzed

using all repeated measures as responses. This example considers only the response for June PM.

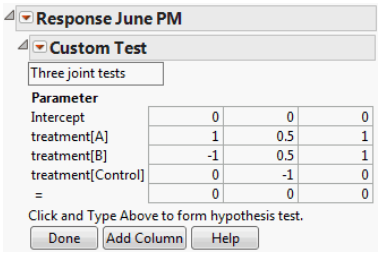
Suppose that you want to test three contrasts. You want to compare the mean responses for: treatment A to treatment B, treatments A and B to the control group, and treatments A and B to the control and placebo groups.

To test these contrasts using Custom Test:

1. Select **Help > Sample Data Library** and open Cholesterol.jmp.
2. Select **Analyze > Fit Model**.
3. Select June PM and click **Y**.
4. Select treatment and click **Add**.
5. Click **Run**.
6. From the red triangle next to Response June PM, select **Estimates > Custom Test**.
7. In the Custom Test specification window, click **Add Column** twice to create three columns.
8. Fill in the editable area with a test name and enter values in the three columns as shown in Figure 3.26.

   To see how to obtain these values, particularly those in the third column, see "Interpretation of Parameters" on page 528 in the "Statistical Details" appendix.

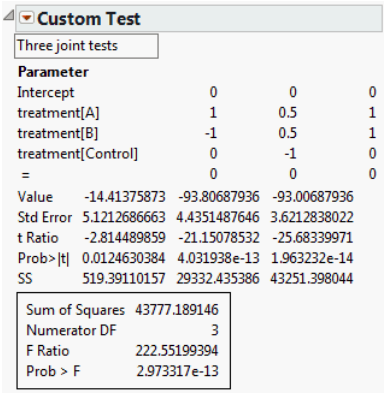
Figure 3.26 Custom Test Specification Window for Three Contrasts



9. Click **Done**.

The results shown in Figure 3.27 indicate that all three hypotheses are individually, as well as jointly, significant.

Figure 3.27 Custom Test Report Showing Tests for Three Contrasts



Custom Test

Three joint tests

Parameter			
Intercept	0	0	0
treatment[A]	1	0.5	1
treatment[B]	-1	0.5	1
treatment[Control]	0	-1	0
=	0	0	0
Value	-14.41375873	-93.80687936	-93.00687936
Std Error	5.1212686663	4.4351487646	3.6212838022
t Ratio	-2.814489859	-21.15078532	-25.68339971
Prob>|t|	0.0124630384	4.031938e-13	1.963232e-14
SS	519.39110157	29332.435386	43251.398044

Sum of Squares	43777.189146
Numerator DF	3
F Ratio	222.55199394
Prob > F	2.973317e-13

Multiple Comparisons

Use this option to obtain tests and confidence levels that compare means defined by levels of your model effects. The goal of multiple comparisons methods is to determine whether group means differ, while controlling the probability of reaching an incorrect conclusion. The Multiple Comparisons option lets you compare group means with the overall average (Analysis of Means) and with a control group mean. You can also conduct pairwise comparisons using either Tukey HSD or Student's *t*. When you specify the Student's *t* method, you can also perform equivalence tests to identify pairwise differences that are of practical importance.

The Student's *t* method controls only the error rate for an individual comparison. As such, it is not a true multiple comparison procedure. All other methods provided control the overall error rate for all comparisons of interest. Each of these methods uses a multiple comparison *adjustment* in calculating p-values and confidence limits.

If your model contains nominal and ordinal effects, you can conduct comparisons using Least Squares Means estimates, or you can define specific comparisons using User-Defined Estimates. If your model contains only continuous effects, you can compare means using User-Defined Estimates.

Tip: Suppose that a continuous effect consists of relatively few levels. If you are interested in comparisons using Least Squares Means Estimates, consider assigning that effect an ordinal (or nominal) modeling type.

Launch the Option

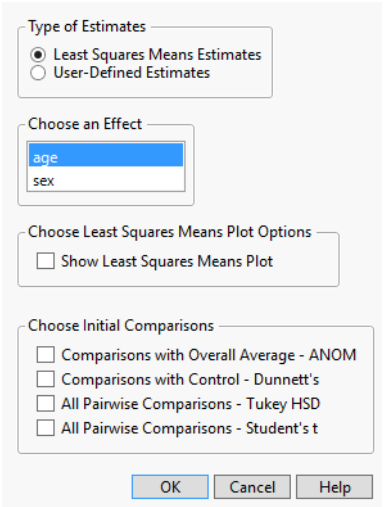
An example of the control window for the Multiple Comparisons option is shown in Figure 3.28. This example is based on the Big Class.jmp data table, with weight as Y and age,

sex, and height as model effects. Two classes of estimates are available for comparisons: Least Squares Means Estimates and User-Defined Estimates.

Least Squares Means Estimates

This option compares least squares means and is available only if there are nominal or ordinal effects in the model. Recall that least squares means are means computed at some neutral value of the other effects in the model. (For a definition of least squares means, see "LSMeans Table" on page 99.) You must select the effect of interest. In Figure 3.28, Least Squares Means Estimates for age are specified. There is an option to show the least squares means plot. See "Least Squares Means Plot Options" on page 102.

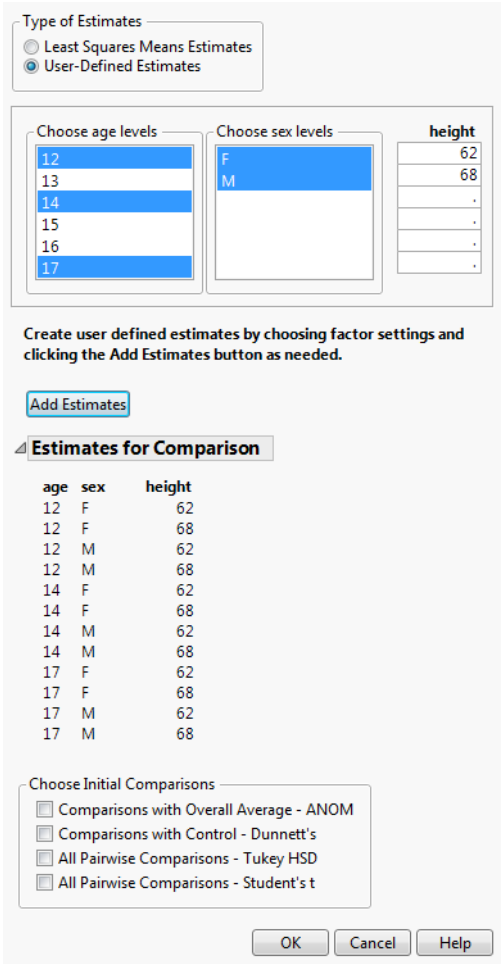
Figure 3.28 Launch Window for Least Squares Means Estimates



User-Defined Estimates

The specification of User-Defined Estimates is illustrated in Figure 3.29. Three levels of age and both levels of sex have been selected. Also, two values of height have been manually entered. The Add Estimates button has been clicked, resulting in the listing of all possible combinations of the specified levels. At this point, you can specify more estimates and click the Estimates button again to add them to the list of Estimates for Comparison.

Figure 3.29 Launch Window for User-Defined Estimates



When you use User-Defined Estimates, effects with no specified levels are set as follows:

- Continuous effects are set to the mean of the effect.
- Nominal and ordinal effects are set to the first level in the value ordering.

Note: In this section, we use the term *mean* to refer to either estimates of least squares means or user-defined estimates.

Choose Least Squares Means Plot Options

Select **Show Least Squares Means Plot** to obtain a least square means plot. If your effect is an interaction term, then you have the option to create an interaction plot. You select the term for

the overlay. If you do not select the interaction plot, then the least squares plot will nest the effect terms. See "Least Squares Means Plot Options" on page 102.

Choose Initial Comparisons

Once you have specified estimates, you can choose the types of comparisons that you would like to see in your initial report by making selections under Choose Initial Comparisons. Or click OK without making any selections.

Comparisons with Overall Average - ANOM Compares each effect least squares mean with the overall least squares mean. (Analysis of Means).

Comparisons with Control - Dunnett's Compares each effect least squares mean with the least squares mean of a control level.

All Pairwise Comparisons - Tukey HSD Tests all pairwise comparisons of the effect least squares means using the Tukey HSD adjustment for multiplicity.

All Pairwise Comparisons - Student's t Tests all pairwise comparisons of the effect least squares means with no multiplicity adjustment.

Each of these selections opens a report with an area at the top that shows details specific to the report. This information includes the quantile, or critical value. For the true multiple comparisons procedures, the method used for the multiple comparison adjustment is shown. If you have specified User-Defined Estimates, the report displays a list of effects that do not vary relative to the specified estimates and the levels at which these effects are set. Unless you have specified otherwise, any continuous effect is set to its mean. Any nominal or ordinal effect is set to the first level in its value ordering.

If you click OK without selecting from the Choose Initial Comparisons list, the Multiple Comparisons report opens, showing the Least Squares Means Estimates table or the User-Defined Estimates table. From the Multiple Comparison's red triangle menu, all of the options listed above are available. The available reports and options are described below.

Least Squares Means or User-Defined Estimates Report

By default, the Multiple Comparisons option displays a Least Squares Means Estimates report or a User-Defined Estimates report, depending on the type of estimates that you selected in the launch window. For each combination of levels of interest, this table gives an estimate of the mean, as well as a test and confidence interval. Specifically, this table contains the following:

Levels of the Categorical Effects The first columns in the report identify the effect or effects of interest. The values in the columns specify the groups being analyzed.

Estimate An estimate of the mean for each group.

Std Error The standard error of the mean for each group.

DF The degrees of freedom for a test of whether the mean is 0.

Lower 95% The lower confidence limit for the mean. You can change the confidence level by selecting Set Alpha Level in the Fit Model window.

Upper 95% The upper confidence limit for the mean.

t Ratio The *t* ratio for the significance test. This column appears only if you right-click in the report and select Columns > t Ratio.

Prob>|t| The *p*-value for the significance test. This column appears only if you right-click in the report and select Columns > Prob>|t|.

Arithmetic Mean Estimate (Appears only in the Least Squares Means Estimates report.) An estimate of the arithmetic mean for each group.

Note: You can obtain *t* ratios and *p*-values by right-clicking in the table and selecting Columns.

Comparisons with Overall Average

This option compares the means for the specified levels specified to the overall mean for these levels. It displays a table showing confidence intervals for differences from the overall mean and a chart showing decision limits. The method used to make the comparisons is called analysis of means (ANOM) (Nelson et al. 2005). ANOM is a multiple comparison procedure that controls the joint error rate for all pairwise comparisons to the overall mean. See Figure 3.30 for a report based on the Lipid Data.jmp sample data table.

ANOM might appear similar to analysis of variance. However, it is fundamentally different in that it identifies levels with means that differ from the overall mean for all levels. In contrast, analysis of variance tests for differences in the means themselves.

At the top of the Comparisons with Overall Average report, you find:

Quantile The value of Nelson's *h* statistic used in constructing the decision limits.

Adjusted DF The degrees of freedom used in constructing the decision limits.

Avg The average mean. For least squares estimates, the average mean is a weighted average of the group least squares means. This weighted average represents the overall mean at the neutral settings where the group least squares means are calculated.

Specifically, the average least squares mean is a weighted average with weights inversely proportional to the diagonal entries of the matrix $L(X'X)^{-1}L'$. Here L is the matrix of coefficients used to compute the group least squares means. For a technical definition of

least squares means and the average least squares mean, see the GLM Procedure chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. 2017).

For user-defined estimates, the average mean is defined similarly. However, in this case, L is the matrix of coefficients used to define the estimates.

Adjustment Describes the method used to obtain the critical value:

- Nelson: Provides exact critical values and p-values. Used whenever possible, in particular, when the estimates are uncorrelated.
- Nelson-Hsu: Provides approximate critical values and p-values based on Hsu's factor analytical approximation is used (Hsu 1992). Used when exact values cannot be obtained.
- Sidak: Used when both Nelson and Nelson-Hsu fail.

For technical details, see the GLM Procedure chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. 2017).

Three options are available from the Comparisons with Overall Average report menu:

Differences from Overall Average

For each comparison of a group's mean to the overall mean, this report provides the following details:

- The levels being compared
- Difference - the estimated difference
- Std Error - the standard error of the difference
- Lower and Upper limits for the confidence interval
- t Ratio - the ratio of the Difference and Std Error columns

Comparisons with Overall Average Decision Chart

This decision chart plots a point at the mean for each group. A horizontal line is plotted at the average mean. Upper and lower decision limits are plotted. Suppose that a point corresponding to a group mean falls outside these limits. This occurrence indicates that the group mean differs from the overall mean, based on the analysis of means test at the specified significance level. The significance level is shown below the chart.

The Comparisons with Overall Average Decision Chart report menu has these options:

Show Summary Report Produces a table showing the estimate, decision limits, and the limit exceeded for each group

Display Options Gives several options for controlling the display of the chart.

Calculate Adjusted P-Values

Adds a column that contains *p*-values (Prob>|t|) to the Comparisons with Overall Average report. Note that computing exact critical values and *p*-values for unbalanced designs requires complex integration and can be computationally challenging. When calculations for such a quantile fail, the Sidak quantile is computed but *p*-values are not available.

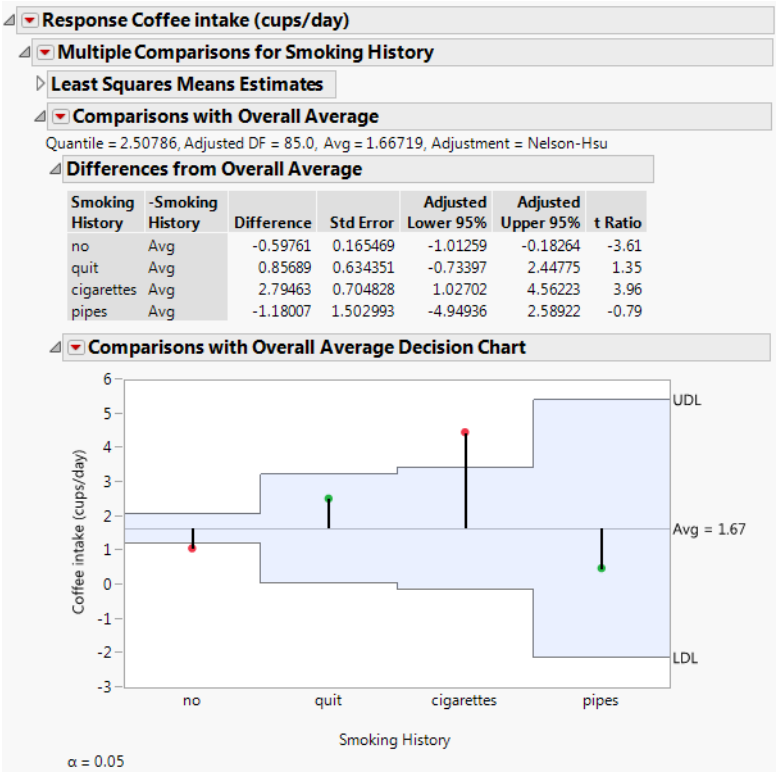
Example of Comparisons with Overall Average

Consider the Lipid Data.jmp sample data table. You are interested in whether any of the four Smoking History categories are unusual in that their mean Coffee intake (cups/day) differ from the overall average coffee intake while controlling for alcohol use and heart history. You specify a model with Coffee intake (cups/day) as the response and Smoking History, Alcohol Use, and Heart History as model effects.

1. Select **Help > Sample Data Library** and open Lipid Data.jmp.
2. Select **Analyze > Fit Model**.
3. Select Coffee intake (cups/day) and click **Y**.
4. Select Smoking History, Alcohol Use, and Heart History, and click **Add**.
5. Click **Run**.
6. From the red triangle next to Response Coffee intake (cups/day), select **Estimates > Multiple Comparisons**.
7. From the Choose an Effect list, select Smoking History.
8. In the Choose Initial Comparisons list, select **Comparisons with Overall Average - ANOM**.
9. Click **OK**.

The results shown in Figure 3.30 indicate that the least squares means for non-smokers and cigarette smokers differ significantly from the overall average in terms of coffee intake.

Figure 3.30 Comparisons with Overall Average for Ratings



Comparisons with Control

If you select Comparisons with Control - Dunnett's, a window opens, asking you to specify a control group. If you selected Least Squares Means Estimates, the list consists of all levels of the effect you that you selected. If you selected User-Defined Estimates, the list consists of the combinations of effect levels that you specified.

After you choose a control group and click OK, the Comparisons with Control report appears in your Fit Least Squares report. This option compares the means for the specified settings to the control group mean. It displays a table showing confidence intervals for differences from the control group and a chart showing decision limits. Dunnett's method is used to make the comparisons. Dunnett's method is a multiple comparison procedure that controls the error rate over all comparisons (Hsu 1996; Westfall et al. 2011).

When exact calculation of p-values and confidence intervals is not possible, Hsu's factor analytical approximation is used (Hsu 1992). Note that computing exact critical values and p-values for unbalanced designs requires complex integration and can be computationally intensive. When calculations for such a quantile fail, the Sidak quantile is computed.

In addition to the list of effects that do not vary for the specified estimates, at the top of the Comparisons with Control report you also find:

Quantile The critical value for Dunnett's test.

Adjusted DF The degrees of freedom used in constructing the confidence intervals.

Control The setting that defines the control group. This is a single level if you have selected a single effect; it is a combination of levels if you specified a user-defined combination of more than one effect.

Adjustment The method used to obtain the critical value:

- Dunnett: Provides exact critical values and *p*-values. Used whenever possible, in particular, when the estimates are uncorrelated.
- Dunnett-Hsu: Provides approximate critical values and *p*-values based on Hsu's factor analytical approximation (Hsu 1992). Used when exact values cannot be obtained.
- Sidak: Used when both Dunnett and Dunnett-Hsu fail.

For technical details, see the GLM Procedure chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. 2017).

Three options are available from the Comparisons with Control report menu:

Differences from Control

For each comparison of a group mean to the control mean, this report provides the following details:

- The levels being compared
- Difference - the estimated difference
- Std Error - the standard error of the difference
- Lower and Upper limits for the confidence interval
- t Ratio - the ratio of the Difference and Std Error columns

Comparisons with Control Decision Chart

This decision chart plots a point at the mean for each group being compared to the control group. A horizontal line shows the mean for the control group. Upper and lower decision limits are plotted. When a point falls outside these limits, it corresponds to a group whose mean differs from the control group mean based on Dunnett's test at the specified significance level. That level is shown beneath the chart.

The Comparisons with Control Decision Chart report menu has these options:

Show Summary Report Produces a table showing the estimate, decision limits, and the limit exceeded for each group

Display Options Gives several options for controlling the display of the chart.

Calculate Adjusted P-Values

Adds a column that contains p-values (Prob>|t|) to the Comparisons with Control report. Note that computing exact critical values and p-values for unbalanced designs requires complex integration and can be computationally challenging. When calculations for such a quantile fail, the Sidak quantile is computed but p-values are not available.

All Pairwise Comparisons

The All Pairwise Comparisons option shows either a Tukey HSD All Pairwise Comparisons or Student's t All Pairwise Comparisons report (Hsu 1996; Westfall et al. 2011). Tukey HSD comparisons are constructed so that the significance level applies jointly to all pairwise comparisons. In contrast, for Student's t comparisons, the significance level applies to each individual comparison. When making several pairwise comparisons using Student's t tests, the risk that one of the comparisons incorrectly signals a difference can well exceed the stated significance level.

At the top of the Tukey HSD All Pairwise Comparisons report you find:

Quantile The critical value for the test. Note that, for Tukey HSD, the quantile is $q/(\sqrt{2})$, where q is the appropriate percentage point of the Studentized range statistic.

Adjusted DF The degrees of freedom used in constructing the confidence intervals.

Adjustment Describes the method used to obtain the critical value:

- Tukey: Provides exact critical values and p-values. Used when the means are uncorrelated and have equal variances, or when the design is variance-balanced.
- Tukey-Kramer: Provides approximate critical values and p-values. Used when exact values cannot be obtained.

For technical details, see the GLM Procedure chapter in the *SAS/STAT 14.3 User's Guid*e (SAS Institute Inc. 2017).

At the top of the Student's t All Pairwise Comparisons report you find the Quantile, or critical value, for the t test and DF, the degrees of freedom used for the t test.

All Pairwise Differences Report

Both Tukey HSD and Student's *t* compare all pairs of levels. For each pairwise comparison, the All Pairwise Differences report shows:

- The levels being compared
- Difference - the estimated difference between the means
- Std Error - the standard error of the difference
- t Ratio - the *t* ratio for the test of whether the difference is zero
- Prob > |t| - the *p*-value for the test
- Lower and Upper limits for a confidence interval for the difference in means

All Pairwise Comparisons Scatterplot

This plot, sometimes called a *diffogram* or a *mean-mean scatterplot,* displays the confidence intervals for all means pairwise differences. (See Figure 3.32 for an example.) Colors indicate which differences are significant.

The plot shows a reference line as an upwardly sloping line on the diagonal. This line represents points where the two means are equal. Each line segment corresponds to a confidence interval for a pairwise comparison. The coordinates of the point displayed on the line segment are the means for the corresponding groups. Placing your cursor over one of these points displays a tooltip identifying the groups being compared and showing the estimated difference. If a line segment crosses the line on the diagonal, then the means can be equal and the comparison is not significant.

The Pairwise Comparisons Scatterplot has the following option:

Show Reference Lines Displays reference grid lines for the points on the scatterplot. This is not recommended if there are many points in the scatterplot. If there are many points, it is better to place your cursor over the points to view the tooltip label.

All Pairwise Differences Connecting Letters

Use this option to display a report that illustrates significant and non-significant comparisons with connecting letters. Levels not connected by the same letter are significantly different. Levels connected by the same letter are not significantly different.

Save All Pairwise Differences Connecting Letters Table

This option creates a data table whose columns contain the levels of the effect, the connecting letters, the least squares means, their standard errors, and confidence intervals. The data table contains a script called Bar Chart that produces a colored bar chart of the least squares means with their confidence intervals superimposed. The levels are arranged in decreasing order of least squares means.

Equivalence Tests

Use this option to conduct one or more equivalence tests. Equivalence tests are useful when you want to detect differences that are of *practical* interest. You are asked to specify a threshold difference for group means for which smaller differences are considered practically equivalent. In other words, if two group means differ by this amount or less, you are willing to consider them equivalent.

Once you have specified this value, the Equivalence Tests report appears. The bounds that you have specified are given at the top of the report. The report consists of a table giving the equivalence tests and a scatterplot that displays them. The equivalence tests and confidence intervals are based on Student's *t* critical values.

Note: Equivalence tests are available only for the Student's *t* method.

Equivalence TOST Tests

The Two One-Sided Tests (TOST) method is used to test for a practical difference between the means (Schuirmann 1987). Two one-sided pooled-variance *t* tests are constructed for the null hypotheses that the true difference exceeds the threshold values. If both tests reject, the difference in the means does not statistically exceed either threshold value. Therefore, the groups are considered practically equivalent. If only one or neither test rejects, then the groups might not be practically equivalent.

For each comparison, the Equivalence TOST Tests report gives the following information:

- Difference - the estimated difference in the means
- Lower Bound t Ratio, Upper Bound t Ratio - the lower and upper bound *t* ratios for the two one-sided pooled-variance significance tests
- Lower Bound p-Value, Upper Bound p-value - *p*-values corresponding to the lower and upper bound *t* ratios
- Maximum p-Value - the maximum of the lower and upper bound *p*-values
- Lower and Upper limits for a $1-2\alpha$ confidence interval for the difference in the means.

Note: Equivalence TOST tests are available only for the Student's *t* method.

Equivalence Tests Scatterplot

Using colors, this scatterplot indicates which means are practically equivalent and which are not as determined by the equivalence test.

The plot shows a solid reference line on the diagonal as well as a shaded reference band. The width of the band is twice the practical difference. Each line segment corresponds to a $1-2\alpha$ confidence interval for a pairwise comparison. The coordinates of the point on the line segment are the means for the corresponding groups. Placing your cursor over one of these

points displays a tooltip indicating the groups being compared and the estimated difference. When a line segment is entirely contained within the diagonal band, it follows that the means are practically equivalent.

Note: Equivalence tests scatterplots are available only for the Student's *t* method.

The Equivalence Tests Scatterplot has the following option:

Show Reference Lines Displays reference lines for the points on the scatterplot. This is not recommended if there are many points in the scatterplot. If there are many points, it is better to place your cursor over the points to view the tooltip label.

Remove

This option removes the Equivalence Tests report from the Student's t All Pairwise Comparisons report.

Example of Tukey HSD All Pairwise Comparisons

Consider the Lipid Data.jmp sample data table. You are interested in Cholesterol differences for gender and non-smokers versus former smokers (Smoking History equal to no and quit, respectively) across two ages (25 and 35) and average height.

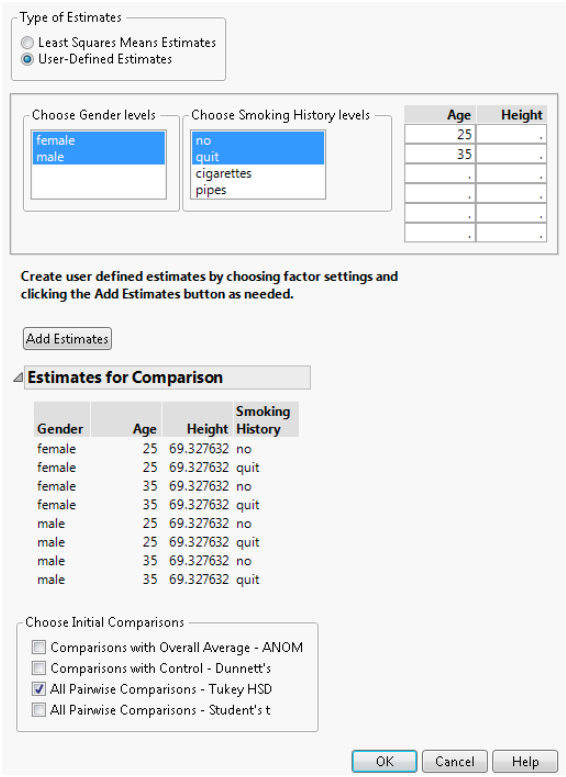
1. Select **Help > Sample Data Library** and open Lipid Data.jmp.
2. Select **Analyze > Fit Model**.
3. Select Cholesterol and click **Y**.
4. Select Gender, Age, Height, and Smoking History, and click **Add.**
5. Click **Run**.
6. From the red triangle next to Response Cholesterol, select **Estimates > Multiple Comparisons**.
7. From the Type of Estimates list, click **User-Defined Estimates**.
8. From the Choose Gender levels list, select female (it should already be selected by default) and male.
9. From the Choose Smoking History levels list, select no and quit.
10. In the Age list, enter the ages 25 and 35 in the first two rows.

    Do not enter any values in the list entitled Height. Because no values for Height are specified, the mean value of the Height column is used in the multiple comparisons report.
11. Click **Add Estimates**.

    Note that all possible combinations of the levels that you specified appear in the Estimates for Comparison report.
12. In the Choose Initial Comparisons list, select **All Pairwise Comparisons - Tukey HSD**.

Check that your window is completed as shown in Figure 3.31.

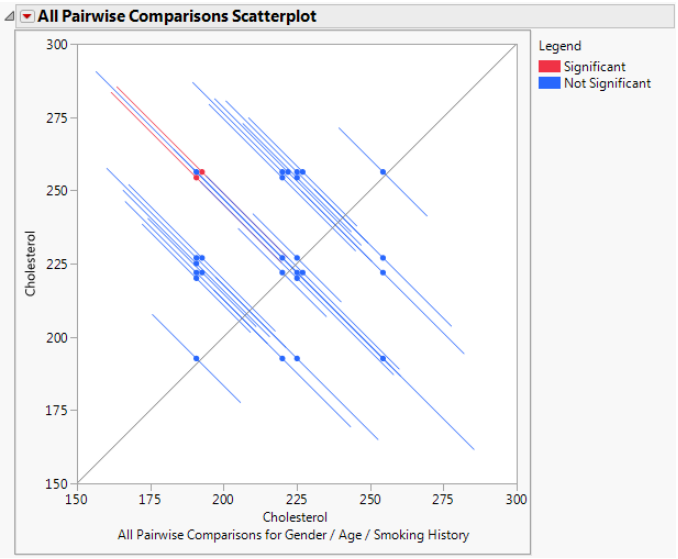
Figure 3.31 Populated Used-Defined Estimates Window



13. Click **OK**.

The All Pairwise Differences report indicates that two of the 28 pairwise comparisons are significant. The All Pairwise Comparisons Scatterplot, shown in Figure 3.32, shows the confidence intervals for these comparisons in red. You can place your pointer over any of the points to determine which pairwise comparison the point represents. The tooltips also contain the difference between the two levels in the comparison. The two red points in Figure 3.32 represent the points comparing 35-year-old former smokers to 25-year-old non-smokers, for both females and males.

Figure 3.32 All Pairwise Comparisons Scatterplot for User-Defined Comparisons



Compare Slopes

The Compare Slopes option appears when there is one nominal term, one continuous term, and their interaction effect for the fixed effects. This option produces a report that enables you to compare the slopes in an analysis of covariance (ANCOVA) model. The report compares the slopes of each level of the interaction effect to the overall slope. The comparison uses analysis of means (ANOM) with the overall average. For more information about the analysis of means (ANOM) report, see "Comparisons with Overall Average" on page 132.

The overall average slope is a weighted average of the slopes, where the weights are inversely proportional to the variances of the slope estimates. These variances are the squared values of the Std Error column in the Differences from Overall Average Slope table.

Joint Factor Tests

The Joint Factor Test option appears when interaction effects are present. For each main effect in the model, JMP produces a joint test of whether all the coefficients for terms involving that main effect are zero. This test is conditional on all other effects being in the model. Specifically, the joint test is a general linear hypothesis test of a restricted model. In that model, all parameters that correspond to the specified effect and the interactions that contain it are set to zero.

Example of a Joint Factor Tests Report

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Model**.
3. Select weight and click **Y**.
4. Verify that 2 appears in the **Degree** box.
5. Select age, sex, and height and click **Macros > Factorial to Degree**.
6. Click **Run**.
7. From the red triangle next to Response weight, select **Estimates > Joint Factor Tests**.

   The report shown in Figure 3.33 appears.

Note that the test for age has 15 degrees of freedom. This test involves the five parameters for age, the five parameters for age*sex, and the five parameters for height*age. The null hypothesis for this test is that all 15 parameters are zero.

Figure 3.33 Joint Factor Tests Report

Joint Factor Tests

Term	DF	Sum of Squares	F Ratio	Prob > F
age	15	6116.2127	2.8488	0.0139*
sex	7	2113.1080	2.1091	0.0879
height	7	9217.8156	9.2002	<.0001*

Inverse Prediction

Inverse prediction occurs when you use a statistical model to infer the value of an explanatory variable, given a value of the response variable. Inverse prediction is sometimes referred to as *calibration*.

By selecting Inverse Prediction on the Estimates menu, you can estimate values of an independent variable, *X*, that correspond to specified values of the response (Figure 3.37). In addition, you can specify values for other explanatory variables in the model (Figure 3.37). The inverse prediction computation provides confidence limits for values of X that correspond to the specified response value. You can specify the response value to be the mean response or simply an individual response. For an example, see "Example of Inverse Prediction" on page 144.

Analyzing Multiple Explanatory Variables

When the model includes multiple explanatory variables, you can predict the value of X for the specified values of the other variables. You might want to predict the amount of running time that results in an oxygen uptake of 50 when one's resting pulse rate is 60. You might want separate inverse predictions for both males and females. Specify these requirements using the inverse prediction option.

The inverse prediction window shows the list of explanatory variables to the left. (See Figure 3.37 for an example.) Each continuous variable is initially set to its mean. Each nominal or ordinal variable is set to its lowest level (in terms of value ordering). You must remove the value for the variable that you want to predict, setting it to missing. Also, you must specify the values of the other variables for which you want your inverse prediction to hold (if these differ from the default settings). In the list to the right in the window, you can supply one or more response values of interest. For an example, see "Example of Predicting a Single X Value with Multiple Model Effects" on page 146.

Note: The confidence limits for inverse prediction can sometimes result in a one-sided or even an infinite interval. For technical details, see "Inverse Prediction with Confidence Limits" on page 555 in the "Statistical Details" appendix.

Example of Inverse Prediction

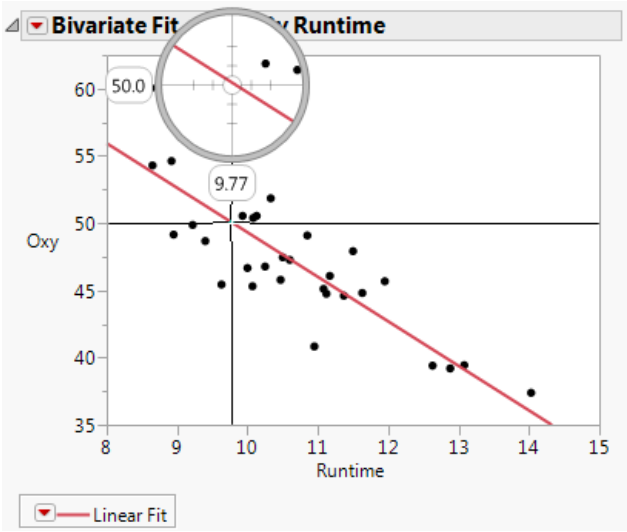
In this example, you fit a regression model that predicts oxygen uptake from Runtime. Then you estimate the Runtime values that result in specified oxygen uptake values. There is only a single *X*, Runtime, so you start by using the Fit Y by X platform to obtain a visual approximation of the inverse prediction values.

1. Select **Help > Sample Data Library** and open Fitness.jmp.
2. Select **Analyze > Fit Y by X**.
3. Select Oxy and click **Y, Response**.
4. Select Runtime and click **X, Factor**.
5. Click **OK**.
6. From the red triangle menu, select **Fit Line**.

   Use the crosshair tool as described below to approximate the Runtime value that results in a mean Oxy value of 50.
7. Select **Tools > Crosshairs**.
8. Click the Oxy axis at about 50 and then drag the cursor so that the crosshairs intersect with the prediction line.

Figure 3.34 shows that a Runtime of about 9.77 gives an inverse prediction of about 50 for Oxy.

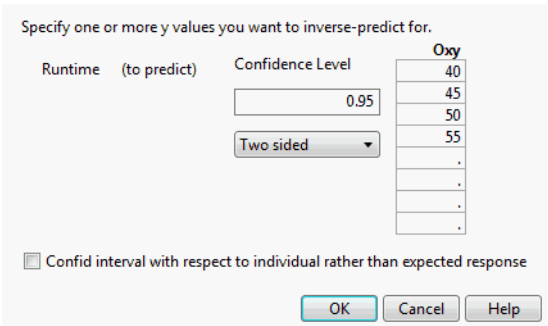
Figure 3.34 Bivariate Fit for Fitness.jmp



To obtain an exact prediction for Runtime, along with a confidence interval, use the Fit Model launch window as follows:

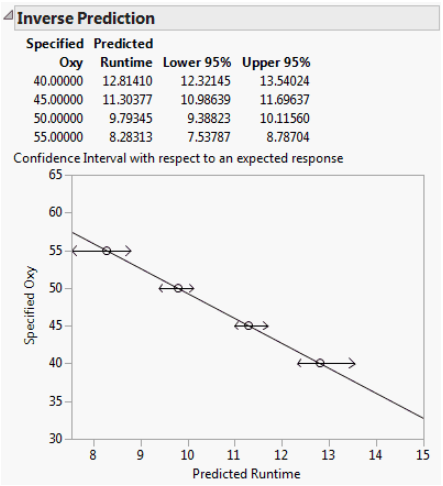
1. From the Fitness.jmp sample data table, select **Analyze > Fit Model**.
2. Select Oxy and click **Y**.
3. Select Runtime and then click **Add**.
4. Click **Run**.
5. From the red triangle menu next to Response Oxy, select **Estimates > Inverse Prediction**.
6. Enter four values for Oxy as shown in Figure 3.35.
7. Click **OK**.

Figure 3.35 Completed Inverse Prediction Specification Window



The Inverse Prediction report (Figure 3.36) gives predicted Runtime values that correspond to each specified Oxy value. The report also shows upper and lower 95% confidence limits for these Runtime values, relative to obtaining the mean response.

Figure 3.36 Inverse Prediction Report



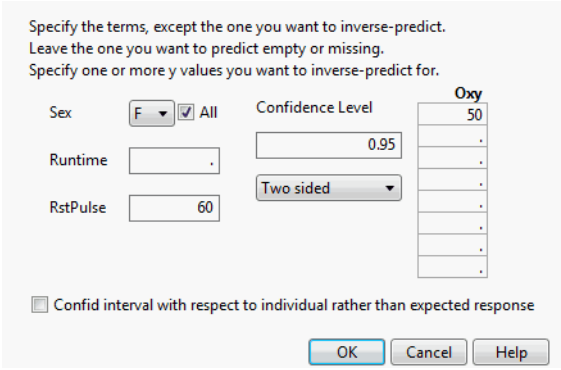
The exact predicted Runtime resulting in an Oxy value of 50 is 9.7935. This value is close to the approximate Runtime value of 9.77 found in the Bivariate Fit report shown in Figure 3.34. The Inverse Prediction report also gives a plot showing the linear relationship between Oxy and Runtime and the confidence intervals.

Example of Predicting a Single X Value with Multiple Model Effects

This example predicts the Runtime that results in oxygen uptake of 50 when RstPulse is 60. The Runtime is predicted for both males and females.

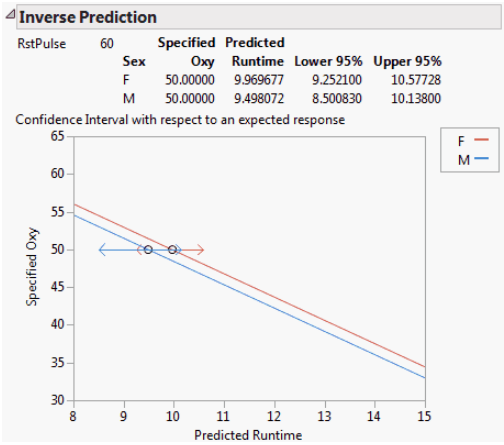
1. From the Fitness.jmp sample data table, select **Analyze > Fit Model**.
2. Select Oxy and click **Y**.
3. Select Sex, Runtime, and RstPulse and then select **Add**.
4. Click **Run**.
5. From the red triangle menu next to Response Oxy, select **Estimates > Inverse Prediction**.
6. Delete the value for Runtime, because you want to predict that value.
7. Select the **All** box next to Sex to estimate Runtime for all levels of Sex.
8. Replace the mean for RstPulse with 60.
9. Enter the value 50 for Oxy as shown in Figure 3.37.
10. Click **OK**.

Figure 3.37 Inverse Prediction Specification for a Multiple Regression Model



The report, shown in Figure 3.38, gives the predicted values of Runtime for both females and males. The report also includes 95% confidence intervals for Runtime values that give a mean response of 50.

Figure 3.38 Inverse Prediction Report for a Multiple Regression Model



The plot shows the linear fits for females and males, given that RstPulse is 60. The two confidence intervals are shown in red and blue, respectively. Note that the intervals overlap, indicating that the true values of Runtime leading to an Oxy value of 50 might be identical for both males and females.

Cox Mixtures

Note: This option is available only for mixture models.

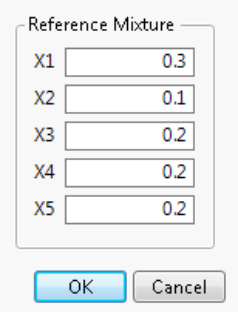
Standard least squares fits mixture models using the parameterization suggested in Scheffé (1958). The parameters for this model cannot easily be used to judge the effects of the mixture components. The Cox Mixture model is a reparameterized and constrained version of the Scheffé model. Using its parameter estimates, you can derive factor effects and the response surface shape relative to a reference point in the design space. See Cornell (1990) for a complete discussion.

The Cox Mixture option opens a window where you enter the reference mixture. If you enter components for the reference mixture that do not sum to one, then the components are proportionately scaled so that they do sum to one. The rescaled mixture is shown in the report as the Reference Mixture. The component effects also appear in the report. A Cox component effect is the difference in the predicted response as the factor goes from its minimum to maximum values along the Cox effect direction.

Example of Cox Mixtures

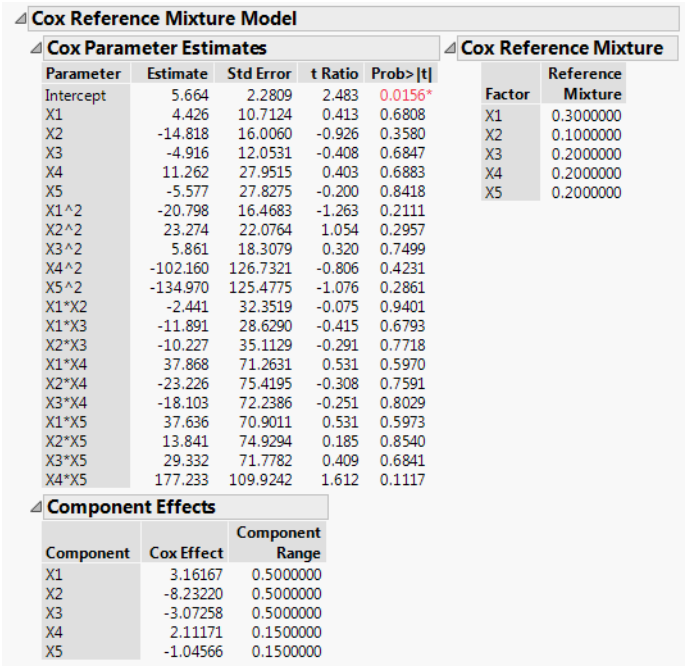
1. Select **Help > Sample Data Library** and open Five Factor Mixture.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y1 and click **Y**.
4. Select X1 through X5.
5. Select **Macros > Mixture Response Surface**.
6. Click **Run**.
7. From the red triangle menu next to Response Y1, select **Estimates > Cox Mixtures**.

Figure 3.39 Cox Reference Mixture Window



8. Type the reference mixture points shown in Figure 3.39.
9. Click **OK**.

Figure 3.40 Cox Mixtures



Cox Reference Mixture Model

Cox Parameter Estimates

Parameter	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.664	2.2809	2.483	0.0156*
X1	4.426	10.7124	0.413	0.6808
X2	-14.818	16.0060	-0.926	0.3580
X3	-4.916	12.0531	-0.408	0.6847
X4	11.262	27.9515	0.403	0.6883
X5	-5.577	27.8275	-0.200	0.8418
X1^2	-20.798	16.4683	-1.263	0.2111
X2^2	23.274	22.0764	1.054	0.2957
X3^2	5.861	18.3079	0.320	0.7499
X4^2	-102.160	126.7321	-0.806	0.4231
X5^2	-134.970	125.4775	-1.076	0.2861
X1*X2	-2.441	32.3519	-0.075	0.9401
X1*X3	-11.891	28.6290	-0.415	0.6793
X2*X3	-10.227	35.1129	-0.291	0.7718
X1*X4	37.868	71.2631	0.531	0.5970
X2*X4	-23.226	75.4195	-0.308	0.7591
X3*X4	-18.103	72.2386	-0.251	0.8029
X1*X5	37.636	70.9011	0.531	0.5973
X2*X5	13.841	74.9294	0.185	0.8540
X3*X5	29.332	71.7782	0.409	0.6841
X4*X5	177.233	109.9242	1.612	0.1117

Cox Reference Mixture

Factor	Reference Mixture
X1	0.3000000
X2	0.1000000
X3	0.2000000
X4	0.2000000
X5	0.2000000

Component Effects

Component	Cox Effect	Component Range
X1	3.16167	0.5000000
X2	-8.23220	0.5000000
X3	-3.07258	0.5000000
X4	2.11171	0.1500000
X5	-1.04566	0.1500000

The report shows the parameter estimates for the Cox mixture model, along with standard errors, and hypothesis tests. The reference mixture appears on the right. The component effects appear below, along with the range.

Parameter Power

The *power* of a statistical test is the probability that the test will be significant, if a difference actually exists. The power of the test indicates how likely your study is to declare a true effect to be significant. The Parameter Power option addresses *retrospective* power analysis.

Note: To ensure that your study includes sufficiently many observations to detect the required differences, use information about power when you *design* your experiment. This type of analysis is called *prospective* power analysis. Consider using the DOE platform to design your study. Both DOE > Sample Size and Power and DOE > Evaluate Design are useful for prospective power analysis. For an example of a prospective power analysis using standard least squares, see "Prospective Power Analysis" on page 215.

The power of a test to detect a difference is affected by the following factors:

- the sample size
- the unknown residual error variance

- the significance level of the test
- the size of the effect to be detected

Suppose that you have already conducted your study, analyzed your data, and found that an effect of interest is not significant. You might be interested in the size of the difference that you might have been likely to detect or the power of the test that you conducted. Or you might want to know the number of observations that you would have needed to detect a difference of a given size with high probability.

The Parameter Power option inserts three columns of values relating to retrospective power analysis in the Parameter Estimates report. The least significant value (LSV0.05), the least significant number (LSN0.05), and a power calculation (AdjPower0.05) are provided.

The Parameter Power calculations apply to a new sample that has the same variability profile as the observed sample.

Caution: The results provided by the LSV0.05, LSN, and AdjPower0.05 should not be used in prospective power analysis. They do not reflect the uncertainty inherent in a future study.

- LSV0.05 is the *least significant value*. This number is the smallest absolute value of the estimate that would make this test significant at significance level 0.05. To be more specific, suppose that the number of observations, the mean square error and that the sum of squares and cross-products matrix for the design remain unchanged. Then, if the absolute value of the estimate had been less than LSV0.05, the Prob>|t| value would have exceeded 0.05. (For more details, see "The Least Significant Value (LSV)" on page 213.)
- LSN is the *least significant number*. This number is the number of observations that would make this test significant at significance level 0.05. Specifically, suppose that the estimate of the parameter, the mean square error, and the sum of squares and cross-products matrix for the design remain unchanged. Then, if the number of observations had been less than the LSN, the Prob>|t| value would have exceeded 0.05. (For more details, see "The Least Significant Number (LSN)" on page 212.)
- AdjPower0.05 is the *adjusted power* value. This number is an estimate of the probability that this test will be significant. Sample values from the current study are substituted for the parameter values typically used in a power calculation. The *adjusted* power calculation adjusts for bias that results from direct substitution of sample estimates into the formula for the non-centrality parameter (Wright and O'Brien 1988). (For more details, see "The Adjusted Power and Confidence Intervals" on page 213.)

The LSV, LSN, and adjusted power are useful in assessing a test's sensitivity. These retrospective calculations also provide an enlightening instructional tool. However, you must be cautious in interpreting these values (Hoenig and Heisey 2001).

For further details about LSV, LSN, and adjusted power, see "Power Analysis" on page 209. For an example of a retrospective analysis, see "Example of Retrospective Power Analysis" on page 214.

Correlation of Estimates

The Correlation of Estimates command on the Estimates menu computes the correlation matrix for the parameter estimates. These correlations indicate whether collinearity is present.

For insight on the construction of this matrix, consider the typical least squares regression formulation. Here, the response (Y) is a linear function of predictors (x's) plus error (ε):

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \varepsilon$$

Each row of the data table contains a response value and values for the p predictors. For each observation, the predictor values are considered fixed. However, the response value is considered to be a realization of a random variable.

Considering the values of the predictors fixed, for any set of Y values, the coefficients, $\beta_0, \beta_1, \ldots, \beta_p$, can be estimated. In general, different sets of Y values lead to different estimates of the coefficients. The Correlation of Estimates option calculates the theoretical correlation of these parameter estimates. (For technical details, see "Details of Custom Test Example" on page 204.)

The correlations of the parameter estimates depend solely on the predictor values and a term representing the intercept. The correlation between two parameter estimates is not affected by the values of the response.

A high positive correlation between two estimates suggests that a collinear relationship might exist between the two corresponding predictors. Note, though, that you need to interpret these correlations with caution (Belsley et al. 1980, p. 185, 92–94). Also, a rescaling of a predictor that shifts its mean changes the correlation of its parameter estimate with the intercept's value.

Example of Correlation of Estimates

1. Select **Help > Sample Data Library** and open Socioeconomic.jmp.
2. Select **Analyze > Fit Model**.
3. Select Median House Value and click **Y**.
4. Select Total Population, Median School Years, Total Employment, and Professional Services and click **Add**.
5. In the Emphasis list, select **Minimal Report**.
6. Click **Run**.
7. From the Response red triangle menu, select **Estimates > Correlation of Estimates**.

Figure 3.41 Correlation of Estimates Report

Correlation of Estimates

Corr

	Intercept	Total Population	Median School Years	Total Employment	Professional Services
Intercept	1.0000	-0.5743	-0.9818	0.4719	0.6903
Total Population	-0.5743	1.0000	0.5871	-0.9746	-0.2398
Median School Years	-0.9818	0.5871	1.0000	-0.5132	-0.7085
Total Employment	0.4719	-0.9746	-0.5132	1.0000	0.1094
Professional Services	0.6903	-0.2398	-0.7085	0.1094	1.0000

The report (Figure 3.41) shows high negative correlations between the parameter estimates for the Intercept and Median School Years (–0.9818). High negative correlations also exist between Total Population and Total Employment (–0.9746).

Coding for Nominal Effects

When you enter a column with a nominal modeling type into your model, JMP represents it internally as a set of continuous indicator variables. Each variable assumes only the values –1, 0, and 1. (Note that this coding is one of many ways to use indicator variables to code nominal variables.) If your nominal column has *n* levels, then *n-1* of these indicator variables are needed to represent it. (The need for *n-1* indicator variables relates directly to the fact that the main effect associated with the nominal column has *n-1* degrees of freedom.) Full details are covered in "Nominal Factors" on page 527 in the "Statistical Details" appendix.

Tip: You can view the coding by selecting Save Columns > Save Coding Table from the red-triangle menu for the main report. See "Save Coding Table" on page 181.

Suppose that you have a nominal column with four levels. Take, as an example, the treatment column in the Cholesterol.jmp sample data table. The treatment column has four levels: A, B, Control, and Placebo. Each of the first three levels is represented by an indicator variable. These indicator variables are named treatment[A], treatment[B], and treatment[Control].

The indicator variable for a given level assigns the values 1 to that level, –1 to the last level, and 0 to the remaining levels. Table 3.1 shows the definitions of the treatment[A], treatment[B], and treatment[Control] indicator variables for this example. For example, consider the indicator variable treatment[A]. As shown in Table 3.1, this variable assigns values as follows:

- The value 1 is assigned to rows that have treatment = A
- The value 0 is assigned to rows that have treatment = B or Control
- The value –1 is assigned to rows that have treatment = Placebo

Table 3.1 Illustration of Indicator Variables for treatment in Cholesterol.jmp

Treatment Assigned to Row	treatment[A]	treatment[B]	treatment[Control]
A	1	0	0
B	0	1	0
Control	0	0	1
Placebo	-1	-1	-1

The order of the levels is determined either by the Value Ordering column property, if you have assigned one, or by the default ordering assigned by JMP. The default ordering is typically the numeric sorting order for numbers and the alphanumeric sorting order for character data. However, certain categorical values, such as the names of months, are sorted appropriately by default. For details about value ordering, see The Column Info Window chapter in the *Using JMP* book.

These variables are used to parametrize the model. They do not typically appear in the data table, but the estimated coefficients for these variables are given in the Parameter Estimates and other reports. Although many other codings are possible, this coding has proven to be practical and interpretable.

For information about the coding of ordinal effects, see "Ordinal Factors" on page 538 in the "Statistical Details" appendix.

Effect Screening

A screening design often provides no degrees of freedom for error. So classical tests for effects are not available. In such cases, Effect Screening is particularly useful.

For these designs, most inferences about effect sizes assume that the estimates for non-intercept parameters are uncorrelated and have equal variances. These assumptions hold for the models associated with many classical experimental designs. However, there are situations where these assumptions do not hold. In both of these situations, the Effect Screening platform guides you in determining which effects are significant.

The Effect Screening platform uses the principle of *effect sparsity* (Box and Meyer 1986). This principle asserts that relatively few of the effects that you study in a screening design are active. Most are inactive, meaning that their true effects are zero and that their estimates are random error.

The following Effect Screening options are available:

Scaled Estimates Gives parameter estimates corresponding to factors that are scaled to have a mean of zero and a range of two. See "Scaled Estimates and the Coding of Continuous Terms" on page 154.

Normal Plot Identifies parameter estimates that deviate from normality, helping you determine which effects are active. See "Normal Plot Report" on page 160.

Bayes Plot Computes posterior probabilities for all model terms using a Bayesian approach. See "Bayes Plot Report" on page 161.

Pareto Plot Plots the absolute values of the orthogonalized and standardized parameter estimates, relating these to the sum of their absolute values. See "Pareto Plot Report" on page 163.

Scaled Estimates and the Coding of Continuous Terms

A parameter estimate is highly dependent on the scale of its corresponding factor. When you convert a factor from grams to kilograms, its parameter estimate changes by a multiple of a thousand. When you apply the same change to a squared (quadratic) term, its parameter estimate changes by a multiple of a million.

To better understand and compare effect sizes, you should examine parameter estimates in a scale-invariant fashion. It makes sense to use a scale that relates the size of a parameter estimate to the size of the effect of its corresponding term on the response. There are many approaches to doing this.

The Effect Screening > Scaled Estimates command on the report's red triangle menu gives coefficients corresponding to scaled factors. The factors are scaled to have a mean of zero and a range of two. Figure 3.42 shows a report for Drug.jmp.

If the sample values for the factor are such that the maximum and minimum values are equidistant from the sample mean, then the scaled factor ranges from –1 to 1. This scaling corresponds to the traditional coding used in the design of experiments. In the case of regression with a single factor, the scaled parameter estimate is half of the predicted response change as the factor travels its whole range.

Scaled estimates are important in assessing the impact of model terms when the data involve uncoded values. For orthogonal designs, the scaled estimates are identical to the estimates for the uncoded data.

Note: The Coding column property scales factor values linearly so that their coded values range from –1 to 1. Parameter estimates are given in terms of these coded values, so that scaled estimates are not required in this situation. (Unlike the transformation used to obtain scaled estimates, the coded values might not have mean zero.)

Example of Scaled Estimates

1. Select **Help > Sample Data Library** and open Drug.jmp.
2. Select **Analyze > Fit Model**.
3. Select y and click **Y**.
4. Select Drug and x and add these to the **Construct Model Effects** list.
5. Change the Emphasis to **Minimal Report**.
6. Click **Run**.
7. From the red triangle menu next to Response y, select **Effect Screening > Scaled Estimates**.

The report (Figure 3.42) indicates that the continuous factor, x, is centered by its mean and scaled by its half-range.

Figure 3.42 Scaled Estimates Report

Scaled Estimates

Nominal factors expanded to all levels
Continuous factors centered by mean, scaled by range/2

Term	Scaled Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.9	0.731352	10.80	<.0001*
Drug[a]	-1.185037	1.060822	-1.12	0.2742
Drug[d]	-1.076065	1.041298	-1.03	0.3109
Drug[f]	2.2611017	1.093974	2.07	0.0488*
x	8.8846543	1.480478	6.00	<.0001*

The model fits three parallel lines, one for each Drug group. The x values range from 3 to 21. The Scaled Estimate for x, 8.8846543, is half the difference between the predicted value for x = 21 and the predicted value for x = 3 for any one of the Drug groups. You can verify this fact by selecting **Save Columns > Prediction Formula** from the report's red triangle menu. Then add rows to obtain predicted values for each of the Drug groups at x = 21 and x = 3.

So, over the range of x values in this particular data set, the impact of x is to vary the response over a range of about 17.8 units. Note that the parameter estimate for x based on the raw data is 0.9871838; it does not permit direct interpretation in terms of the response.

Plot Options

The Normal, Bayes, and Pareto Plot options provide reports that appear as part of the Effect Screening report. These reports can be constructed so that they correct for unequal variances and correlations among the estimates.

Note: The Normal, Bayes, and Pareto Plot options require that your model involves at least four parameters, one of which can be the intercept.

Transformations

When you select any of the plot options, the following information appears directly beneath the Effect Screening report title:

- If the estimates have equal variances and are uncorrelated, these two notes appear:
  - The parameter estimates have equal variances.
  - The parameter estimates are not correlated.
- If the estimates have unequal variances or are correlated, then an option list replaces the relevant note. The list items selected by default show that JMP has transformed the estimates. If you want to undo either or both transformations, select the appropriate list items.

Lenth PSE Values

A Lenth PSE (*pseudo standard error*) table appears directly beneath the notes or option lists. (For a description of the PSE, see "Lenth's PSE" on page 119.) The statistics that appear in the Lenth table depend on the variances and correlation of the parameter estimates.

When the parameter estimates have equal variances and are uncorrelated, the Lenth table provides the following statistic:

Lenth PSE The Lenth pseudo standard error for the estimates.

When the parameter estimates have unequal variances or are correlated or both, the Lenth table provides the following statistics:

t-Test Scale Lenth PSE The Lenth pseudo standard error computed for the transformed parameter estimates divided by their standard errors in the transformed scale.

Coded Scale Lenth PSE The Lenth pseudo standard error for the transformed parameter estimates.

Parameter Estimate Population Report

The Parameter Estimate Population report gives tests for the parameter estimates. The tests are based on transformations as specified in the option lists.

- The option **Using estimates standardized to have equal variances** applies a normalizing transformation to standardize the variances. This option is selected by default when the variances are unequal.

- The option **Using estimates orthogonalized to be uncorrelated** applies a transformation to remove correlation. This option is selected by default when the estimates are correlated. The transformation that is applied is identical to the transformation that is used to calculate sequential sums of squares. The estimates measure the additional contribution of the variable after all previous variables have been entered into the model.
- If the notes indicate that the estimates have equal variances and are not correlated, no transformation is applied.

The columns that appear in the table depend on the selections initially described in the notes or option lists. The report highlights any row corresponding to an estimate with a *p*-value of 0.20 or less. All versions of the report give Term, Estimate, and either *t*-Ratio and Prob>|t| or Pseudo *t*-Ratio and Pseudo *p*-Value.

Term Gives the model term whose parameter estimate is of interest.

Estimate Gives the estimate for the parameter. Estimate sizes can be compared only when the model effects have identical scaling.

t-Ratio Appears if there are degrees of freedom for error. This value is the parameter estimate divided by its standard error.

Prob>|t| Gives the *p*-value for the test. If a transformation is applied, this option gives the *p*-value for a test using the transformed data.

Pseudo t-Ratio Appears when there are no degrees of freedom for error. If the relative standard errors of the parameters are equal, Pseudo t-Ratio is the parameter estimate divided by Lenth's PSE. If the relative standard errors vary, it is calculated as shown in "Pseudo t-Ratios" on page 120.

Pseudo p-Value Appears when there are no degrees of freedom for error. The *p*-value is derived using a *t* distribution. The degrees of freedom are *m*/3, rounded down to the nearest integer, where *m* is the number of parameters.

If **Using estimates standardized to have equal variances** is selected and the note indicating that the parameter estimates are not correlated appears, the report shows a column called Standardized Estimate. This column provides estimates of the parameters resulting from the transformation used to transform the estimates to have equal variances.

If both **Using estimates standardized to have equal variances** and **Using estimates orthogonalized to be uncorrelated** are selected, the report gives a column called Orthog Coded. The following information is provided:

Orthog Coded Gives the estimate of the parameter resulting from the transformation that is used to orthogonalize the estimates.

Orthog t-Ratio Appears if there are degrees of freedom for error. Gives the *t* ratio for the transformed estimates.

Pseudo Orthog t-Ratio Appears if there are no degrees of freedom for error. It is a *t* ratio computed by dividing the orthogonalized estimate, Orthog Coded, by Coded Scale Lenth PSE.

Effect Screening Report

Figure 3.43 shows the Effect Screening report that you create by running the Fit Model script in the Bicycle.jmp sample data table. Note that you would select **Effect Screening > Normal Plot** in order to obtain this form of the report. The notes directly beneath the report title indicate that no transformation is required. Consequently, the Lenth PSE is displayed. Because there are no degrees of freedom for error, no estimate of residual error can be constructed. For this reason, Lenth's PSE is used as an estimate of residual error to obtain pseudo *t* ratios. Pseudo *p*-values are given for these *t* ratios. Rows for non-intercept terms corresponding to the three estimates with *p*-values of 0.20 or less are highlighted.

Figure 3.43 Effect Screening Report for Equal Variance and Uncorrelated Estimates

Effect Screening

The parameter estimates have equal variances.
The parameter estimates are not correlated.

Lenth PSE
0.75

Parameter Estimate Population

Term	Estimate	Pseudo t-Ratio	Pseudo p-Value
Intercept	66.5000	88.6667	<.0001*
HBars[down]	0.5000	0.6667	0.5649
Dynamo[off]	-6.0000	-8.0000	0.0096*
Seat[down]	1.7500	2.3333	0.1266
Tires[hard]	-1.2500	-1.6667	0.2195
Gear[low]	-11.2500	-15.0000	0.0023*
Raincoat[off]	0.2500	0.3333	0.7665
Brkfast[no]	0.5000	0.6667	0.5649

Orthog t Test used Pseudo Standard Error

Normal Plot

Effect Screening Report for Unequal Variances and Correlated Estimates

In the Odor.jmp sample data table, run the Model script and click **Run**. To create the report shown in Figure 3.44, select **Effect Screening > Normal Plot** from the Response Y red triangle menu. You can also create the report by selecting the Bayes Plot or Pareto Plot options in the Response Y red triangle menu.

The notes directly beneath the report title indicate that transformations were required both to standardize and orthogonalize the estimates. The correlation matrix is shown in the Correlation of Estimates report.

The report shows the *t*-Test Scale and Coded Scale Lenth PSEs. But, because there are degrees of freedom for error, the tests in the Parameter Estimate Population report do not use the

Lenth PSEs. Rows for non-intercept terms corresponding to the three estimates with *p*-values of 0.20 or less are highlighted. A note at the bottom of the Parameter Estimate Population report indicates that orthogonalized estimates depend on their order of entry into the model.

Figure 3.44 Effect Screening Report for Unequal Variances and Correlated Estimates

Effect Screening

Using estimates standardized to have equal variances

Using estimates orthogonalized to be uncorrelated

	Lenth PSE
t-Test Scale	1.7815489
Coded Scale	14.253752

Parameter Estimate Population

Term	Estimate	t Ratio	Orthog Coded	Orthog t-Ratio	Prob>|t|
Intercept	-4.3333	-0.2422	15.2000	1.8998	0.1159
temp	-25.8750	-2.3618	-18.8964	-2.3618	0.0646
gl ratio	-23.5000	-2.1450	-17.1620	-2.1450	0.0848
ht	-10.6250	-0.9698	-7.7594	-0.9698	0.3767
temp*gl ratio	-22.0000	-1.4200	-11.3608	-1.4200	0.2149
temp*ht	-5.2500	-0.3389	-2.7111	-0.3389	0.7485
gl ratio*ht	-5.0000	-0.3227	-2.5820	-0.3227	0.7600
temp*temp	25.2917	1.5684	12.2138	1.5266	0.1874
gl ratio*gl ratio	18.5417	1.1498	9.5025	1.1877	0.2883
ht*ht	-7.2083	-0.4470	-3.5763	-0.4470	0.6736

Each Orthog Estimate is conditioned on the effects before it

Correlations of Estimates

Correlation										
Intercept	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.5547	-0.5547	-0.5547
temp	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
gl ratio	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
ht	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
temp*gl ratio	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
temp*ht	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
gl ratio*ht	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000
temp*temp	-0.5547	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	1.0000	0.0769	0.0769
gl ratio*gl ratio	-0.5547	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0769	1.0000	0.0769
ht*ht	-0.5547	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0769	0.0769	1.0000

Transformation to make uncorrelated

Correlations of Estimates Report

The Correlations of Estimates report appears only if the estimates are correlated (Figure 3.44). The report provides the correlation matrix for the parameter estimates. This matrix is similar to the one that you obtain by selecting the Estimates > Correlation of Estimates red triangle option. However, to provide a more compact representation, the report does not show column headings. See "Correlation of Estimates" on page 151 for details.

"Transformation to make uncorrelated" Report

The "Transformation to make uncorrelated" report appears only if the estimates are correlated. The report gives the matrix used to transform the design matrix to produce uncorrelated parameter estimates. The transformed, or *orthogonally coded*, estimates are obtained by pre-multiplying the original estimates with this matrix and dividing the result by 2.

The transformation matrix can be obtained using the Cholesky decomposition. Express X′X as LL′, where L is the lower triangular matrix in the Cholesky decomposition. Then the transformation matrix is given by L′.

This transformation orthonormalizes each regressor with respect to the regressors that precede it in the ordering of model terms. The transformation produces a diagonal covariance matrix with equal diagonal elements. The coded estimates are a result of this iterative process.

Note: The orthogonally coded estimates depend on the order of terms in the model. Each effect's contribution is estimated only after it is made orthogonal to previously entered effects. Consider entering main effects first, followed by two-way interactions, then three-way interactions, and so on.

Normal Plot Report

Below the Normal Plot report title, select either a normal plot or a half-normal plot (Daniel 1959). Both plots are predicated on the principle of effect sparsity, namely, the idea that relatively few effects are active. Those effects that are inactive represent random noise. Their estimates can be assumed to have a normal distribution with mean 0 and variance σ^2, where σ^2 represents the residual error variance. It follows that, on a normal probability plot, estimates representing inactive effects fall close to a line with slope σ.

Normal Plot

If no transformation is required, the vertical coordinate of the normal plot represents the value of the estimate and the horizontal coordinate represents its normal quantile. Points that represent inactive effects should follow a line with slope of σ. Lenth's PSE is used to estimate σ and a blue line with this slope is shown on the plot.

If a transformation to orthogonality has been applied, the vertical axis represents the Normalized Estimates. These are the Orthog *t*-Ratio values found in the Parameter Estimate Population report. (The Orthog *t*-Ratio values are the Orthog Coded estimates divided by the Coded Scale Lenth PSE.)

Because the estimates are normalized by an estimate of σ, the points corresponding to inactive effects should fall along a line of slope 1. A red line with slope 1 is shown on the plot, as well as a blue line with slope equal to the t-Test Scale Lenth PSE.

In all cases, estimates that deviate from normality at the 0.20 level, based on the *p*-values in the Parameter Estimate Population report, are labeled on the plot.

Half-Normal Plot

The half normal plot shows the absolute values of effects. The construction of the axes and the lines displayed mirror those aspects of normal plot.

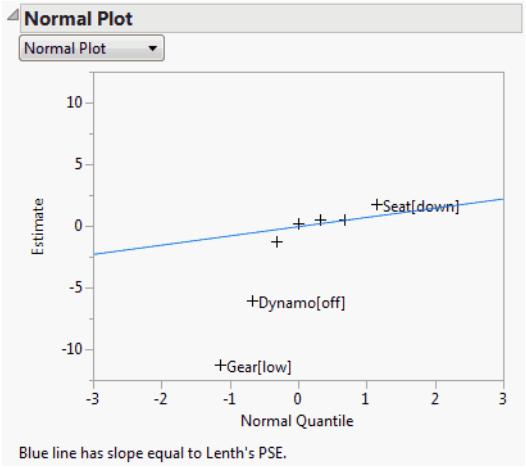
Figure 3.45 shows the Normal Plot report for the Bicycle.jmp sample data table. No transformation is needed for this model, so the plot shows the raw estimates plotted against their normal quantiles. A line with slope equal to Lenth's PSE is shown on the plot. The plot suggests that Gear, Dynamo, and Seat are active factors.

Example of a Normal Plot

1. Select **Help > Sample Data Library** and open Bicycle.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Select HBars through Brkfast and click **Add**.
5. Click **Run**.
6. From the red triangle menu next to Response Y, select **Effect Screening > Normal Plot**.

   The following Normal Plot appears.

Figure 3.45 Normal Plot



Bayes Plot Report

The Bayes Plot report gives another approach to determining which effects are active. This report helps you compute posterior probabilities using a Bayesian approach. This method, due to Box and Meyer (1986), assumes that the estimates are a mixture from two distributions. The majority of the estimates, corresponding to inactive effects, are assumed to be pure random normal noise with variance σ^2. The remaining estimates, the active ones, are assumed to come from a *contaminating* distribution that has a variance K times larger than σ^2.

Term Gives the model term corresponding to the parameter estimate.

Estimate Gives the parameter estimate. The Bayes plot is constructed with respect to estimates that have estimated standard deviation equal to 1. If the estimates are not correlated, the *t*-Ratio is used. If the estimates are correlated, the Orthog *t*-Ratio is used.

Prior Prob Enables you to specify a probability that the estimate is nonzero (equivalently, that the estimate is in the contaminating distribution). Prior probabilities for estimates are usually set to equal values. The commonly recommended value of 0.2 appears initially, though you can change it.

K Contam The value of the contamination coefficient, representing the ratio of the contaminating distribution variance to the error variance. *K* is commonly set to 10, which is the default value.

Std Err Scale If there are degrees of freedom for an estimate of standard error, this value is set to 1. JMP uses this value because the estimates used in the report are transformed and scaled to unit variance. The value is set to 0 for a saturated model with no estimate of error. If you specify a different value, think of it in terms of a scale factor of the RMSE estimate.

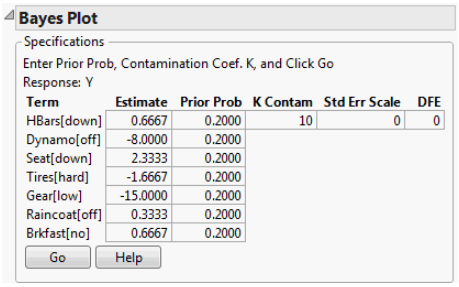
DFE Gives the error degrees of freedom.

The specifications window, showing default settings for a Bayes Plot for the Bicycle.jmp sample data table, is shown in Figure 3.46. Clicking Go in this window updates the report to show Posterior probabilities for each of the terms and a bar chart (Figure 3.47).

Example of a Bayes Plot

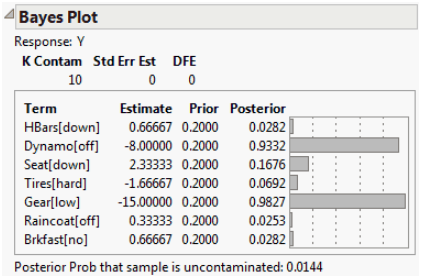
1. Select **Help > Sample Data Library** and open Bicycle.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Select HBars through Brkfast and click **Add**.
5. Click **Run**.
6. From the red triangle menu next to Response Y, select **Effect Screening > Bayes Plot**.

Figure 3.46 Bayes Plot Specifications



7. Click **Go** to calculate the posterior probabilities.

Figure 3.47 Bayes Plot Report



The note beneath the plot in the Bayes Plot report gives the Posterior Prob that the sample is uncontaminated. Posterior Prob is the probability, based on the priors and the data, that there are no active effects whatsoever. The probability is small, 0.0144, indicating that it is likely that there are active effects. The Posterior probability column suggests that at least Dynamo and Gear are active effects.

Pareto Plot Report

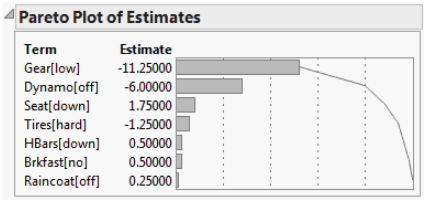
The Pareto Plot report presents a Pareto chart of the absolute values of the estimates. Figure 3.48 shows a Pareto Plot report for the Bicycle.jmp sample data table.

- If the original estimates have equal variances and are not correlated, the original estimates are plotted.
- If the original estimates have unequal variances and are not correlated, the *t* ratios are plotted.
- If the original estimate have unequal variances and are correlated, the Orthog Coded estimates are plotted.

The cumulative sum line in the plot sums the absolute values of the estimates. Keep in mind that the orthogonal estimates depend on the order of entry of terms into the model.

Note: The estimates that appear in the plot are standardized to have equal variances and are transformed to be orthogonal. You have the option of undoing these transformations. See "Transformations" on page 156.

Figure 3.48 Pareto Plot



Factor Profiling

The Factor Profiling menu helps you explore and visualize your estimated model. You can explore the shape of the response surface, find optimum settings, simulate response data based on your specified noise assumptions, and transform the response if needed.

The Profiler, Contour Profiler, Mixture Profiler, and Surface Profiler are extremely versatile tools whose use extends beyond standard least squares models. For details about their interpretation and use, see the *Profilers* book.

If the response column in your model contains an invertible transformation of one variable, the profilers and Interaction Plots show the response on the untransformed scale.

The following Factor Profiling options are available:

Note: If your model contains an expression or vector as an effect, most of these options are not available.

Profiler Shows prediction traces for each X variable. Enables you to find optimum settings for one or more responses and to explore response distributions using simulation. See "Profiler" on page 165 and the Profiler chapter in the *Profilers* book.

Interaction Plots Shows a matrix of interaction plots, when there are interaction effects in the model. See "Interaction Plots" on page 166.

Contour Profiler Provides an interactive contour profiler, which is useful for optimizing response surfaces graphically. See "Contour Profiler" on page 167 and the Contour Profiler chapter in the *Profilers* book.

Mixture Profiler Shows response contours of mixture experiment models on a ternary plot. See "Mixture Profiler" on page 168 and the Mixture Profiler chapter in the *Profilers* book.

Note: This option appears only if you apply the Mixture Effect attribute to three or more effects or the Mixture property to three or more columns.

Cube Plots Shows predicted values for the extremes of the factor ranges laid out on the vertices of cubes. See "Cube Plots" on page 169.

Box Cox Y Transformation Finds a Box-Cox power transformation of the response that is best in terms of satisfying the normality and homogeneity of variance assumptions. See "Box Cox Y Transformation" on page 170.

Surface Profiler Shows a three-dimensional surface plot of the response surface. See "Surface Profiler" on page 172 and the Surface Plot chapter in the *Profilers* book.

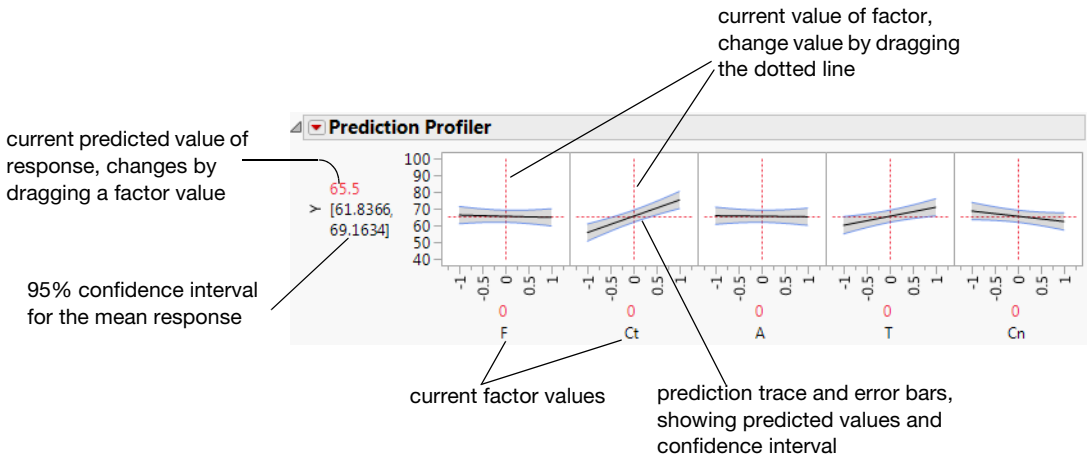
Profiler

Note: For complete details, see the Profiler chapter in the *Profilers* book.

The **Profiler** (or Prediction Profiler) shows prediction traces for each *X* variable.

Figure 3.49 illustrates part of the profiler for the Reactor.jmp sample data table. The vertical dotted line for each *X* variable shows its *current value* or *current setting*. Use the Profiler to change one variable at a time in order to examine the effect on the predicted response.

Figure 3.49 Illustration of Prediction Traces



The factors F and Ct in Figure 3.49 are continuous. If the factor is nominal, the levels are displayed on the horizontal axis.

For each X variable, the value above the factor name is its current value. You change the current value by clicking in the graph or by dragging the dotted line where you want the new current value to be.

- The horizontal dotted line shows the *current predicted value* of each Y variable for the current values of the X variables.
- The black lines within the plots show how the predicted value changes when you change the current value of an individual X variable. The 95% confidence interval for the predicted values is shown by a dotted curve surrounding the prediction trace (for continuous variables) or an error bar (for categorical variables).

Interaction Plots

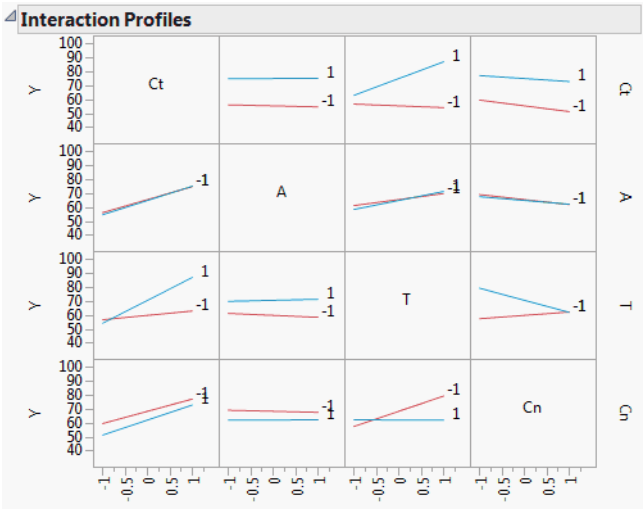
When there are interaction effects in the model, the Interaction Plots option shows a matrix of interaction plots. Each cell of the matrix contains a plot whose horizontal axis is scaled for the effect displayed in the column in which the plot appears. Line segments are plotted for the interaction of that effect with the effect displayed in the corresponding row. So, an interaction plot shows the interaction of the row effect with the column effect.

A line segment is plotted for each level of the row effect. Response values predicted by the model are joined by line segments. Non-parallel line segments give visual evidence of possible interactions. However, the *p*-value for such a suggested interaction should be checked before concluding that it exists. Figure 3.50 gives an interaction plot matrix for the Reactor.jmp sample data table.

Example of Interaction Plots

1. Select **Help > Sample Data Library** and open Reactor.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Make sure that the **Degree** box has a 2 in it.
5. Select Ct, A, T, and Cn and click **Macros > Factorial to Degree**.
6. Click **Run**.
7. From the red triangle menu next to Response Y, select **Factor Profiling > Interaction Plots**.

Figure 3.50 Interaction Plots



The plot corresponding to the T*Cn interaction is the third plot in the bottom row of plots or equivalently, the third plot in the last column of plots. Either plot shows that the effect of Cn on Y is fairly constant at the low level of T, whether Cn is set at a high or low level. However, at the high level of T, the effect of Cn on Y differs based on its level. Cn at –1 leads to a higher predicted Y than Cn at 1. Note that this interaction is significant with a *p*-value < 0.0001.

In certain designs, two-way interactions are aliased with other two-way interactions. When this aliasing occurs, cells in the Interaction Profiles plot corresponding to these interactions are dimmed.

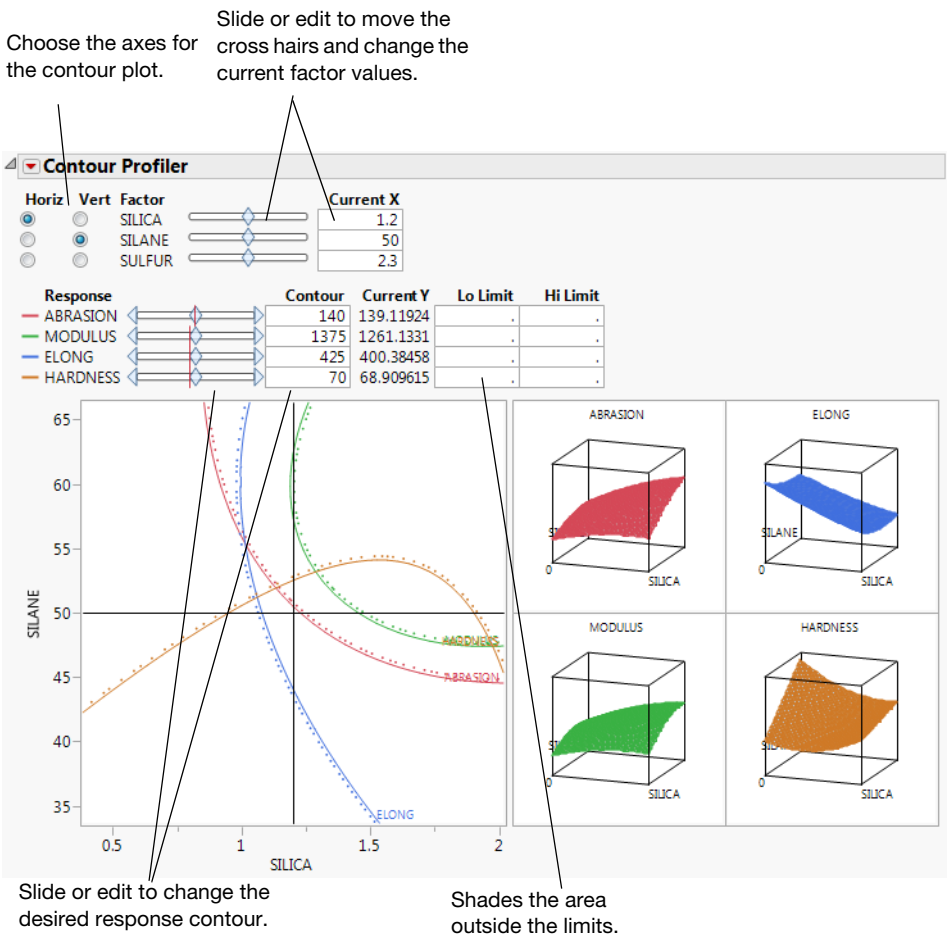
Contour Profiler

Note: For complete details, see the Contour Profiler chapter in the *Profilers* book.

Use the interactive Contour Profiler for optimizing response surfaces graphically. The Contour Profiler shows contours for the fitted model for two factors at a time. The report also includes a surface plot.

Figure 3.51 shows a contour profiler view for the Tiretread.jmp sample data table. Run the data table script **RSM for 4 Responses** and select **Profilers > Contour Profiler** from the Least Squares Fit report menu.

Figure 3.51 Contour Profiler



Mixture Profiler

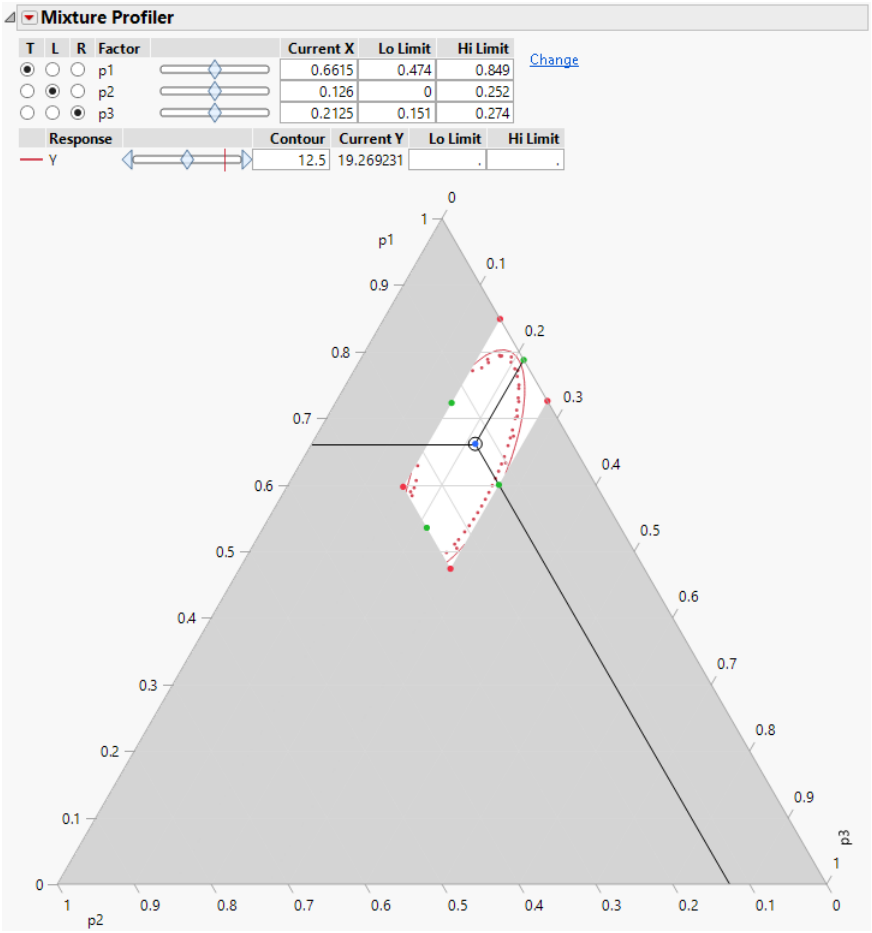
Note: This option appears only if you specify the **Macros > Mixture Response Surface** option for an effect. For complete details, see the Mixture Profiler chapter in the *Profilers* book.

The Mixture Profiler shows response contours of mixture experiment models on a ternary plot. Use the Mixture Profiler when three or more factors in your experiment are components in a mixture. The Mixture Profiler helps you visualize and optimize the response surfaces of your experiment.

Figure 3.52 shows the Mixture Profiler for the model in the Plasticizer.jmp sample data table. Run the **Model** data table script and then select **Factor Profiling > Mixture Profiler** from the

report's red triangle menu. You modify plot axes for the factors by selecting different radio buttons at the top left of the plot. The Lo and Hi Limit columns at the upper right of the plot let you enter constraints for both the factors and the response.

Figure 3.52 Mixture Profiler



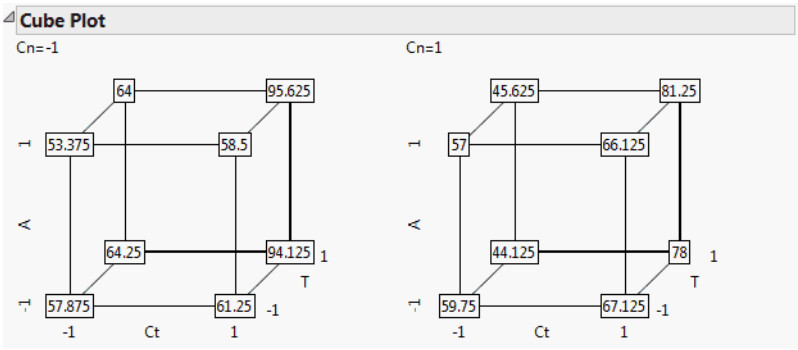
Cube Plots

The Cube Plots option displays predicted values for the extremes of the factor ranges. These values appear on the vertices of cubes (Figure 3.53). The vertices are defined by the smallest and largest observed values of the factor. When you have multiple responses, the multiple responses are shown stacked at each vertex.

Example of Cube Plots

1. Select **Help > Sample Data Library** and open Reactor.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Make sure that the **Degree** box has a 2 in it.
5. Select Ct, A, T, and Cn and click **Macros > Factorial to Degree**.
6. Click **Run**.
7. From the Response Y red triangle menu, select **Factor Profiling > Cube Plots**.

Figure 3.53 Cube Plots



Note that there is one cube for Cn = –1 and one for Cn =1. To change the layout so that the factors are mapped to different cube coordinates, click a factor name in the first cube. Drag it to cover the factor name for the desired axis. For example, in Figure 3.53, if you click T and drag it over Ct, then T and Ct (and their corresponding coordinates) exchange places. To see the levels of Cn in a single cube, exchange it with another factor in the first cube by dragging it over that factor.

Box Cox Y Transformation

A Box-Cox power transformation is used to transform the response so that the usual regression assumptions of normality and homogeneity of variance are more closely satisfied. The transformed response can then be fit using a regression model. However, you can also use the Box-Cox power transformation to transform a variable for other reasons. This transformation is appropriate only when the response, Y, is strictly positive.

A commonly used transformation raises the response to some power. Box and Cox (1964) formalized and described this family of power transformations. The formula for the transformation is constructed to provide a continuous definition in terms of the parameter λ,

and so that the error sums of squares are comparable. Specifically, the following equation provides the family of transformations:

$$Y_\lambda = \begin{cases} \dfrac{y^\lambda - 1}{\lambda \dot{y}^{\lambda-1}} & \text{if } \lambda \neq 0 \\ \dot{y}\ln(y) & \text{if } \lambda = 0 \end{cases}$$

Here, $\dot{y}$ denotes the geometric mean.

The Box Cox Y Transformation option fits transformations from $\lambda = -2$ to 2 in increments of 0.2. To choose a proper value of λ, the likelihood function for each of these transformations is computed. They are computed under the assumption that the errors are independent and normal with mean zero and variance σ^2. The value of λ that maximizes the likelihood is selected. This value also minimizes the SSE over the values of λ. The value of λ that maximizes the likelihood is found using a quadratic interpolation between the two incremental grid points surrounding the grid point with the smallest SSE.

The Box-Cox Transformations report displays a plot showing the sum of squared errors (SSE) values against the values of λ. The horizontal red line on the plot represents a one-sided 95% confidence interval for λ. This confidence interval is based on the confidence region defined in Box and Cox (1964, p. 216). The confidence region is defined by the following inequality:

$$SSE(\lambda) < SSE(\lambda_{best}) * \exp(ChiSquareQuantile(0.95,1) / dfe)$$

where

$SSE(\lambda_{best})$ is the SSE calculated using the reported Best λ

ChiSquareQuantile(0.95,1) is the 0.95^{th} quantile of a χ^2 distribution with 1 degree of freedom

dfe is the error degrees of freedom in the Analysis of Variance table for the regression model

The Box Cox Transformations report provides the following options:

Refit with Transform Enables you to specify a value for lambda to define a transformed Y variable and then provides a least squares fit to the transformed variable.

Replace with Transform Enables you to specify a value for lambda to define a transformed Y variable and then replaces the existing least squares fit with a fit to the transformed variable. If you have multiple responses, Replace with Transform only replaces the report for the response you are transforming.

Save Best Transformation Creates a new column in the data table and saves the formula for the best transformation.

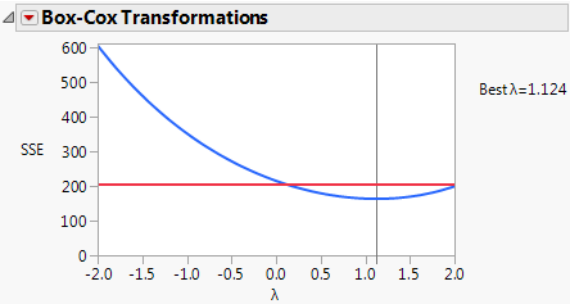
Save Specific Transformation Enables you to specify a value for lambda and creates a column in the data table with the formula for your specified transformation.

Table of Estimates Creates a new data table containing parameter estimates and SSE values for all λ from –2 to 2, in increments of 0.2.

Example of Box-Cox Y Transformation

1. Select **Help > Sample Data Library** and open Reactor.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Make sure that the **Degree** box has a 2 in it.
5. Select F, Ct, A, T, and Cn and click **Macros > Factorial to Degree**.
6. Click **Run**.

Figure 3.54 Box Cox Y Transformation



The plot shows that the best values of λ are between 0.1 and 2.0. The value that JMP selects, using interpolation between the best two values in the 0.2-unit grid of λ values, is 1.124.

7. (Optional) To see the SSE values used to construct the graph, select **Table of Estimates**.

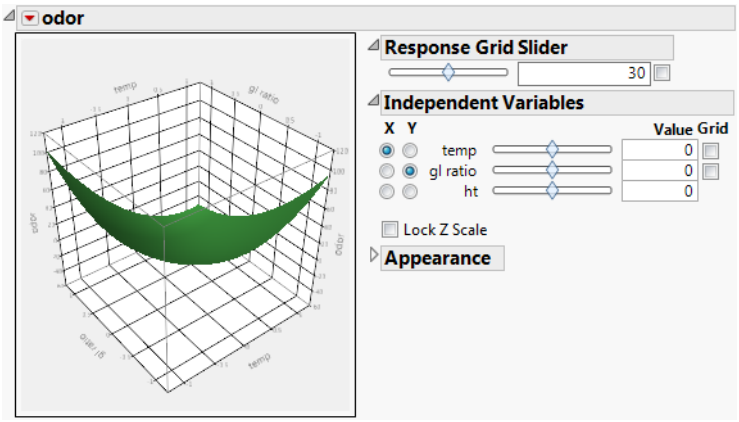
Surface Profiler

Note: For complete details, see the Surface Plot chapter in the *Profilers* book.

The Surface Profiler shows a three-dimensional surface plot of the response surface.

Figure 3.55 shows the Surface Profiler for the model in the Odor.jmp sample data table. Run the **Model** data table script and then select **Factor Profiling > Surface Profiler** from the report's red triangle menu. You can change the variables on the axes using the radio buttons under Independent Variables. Also, you can plot points by clicking Actual under Appearance.

Figure 3.55 Surface Plot



Row Diagnostics

The row diagnostics menu addresses issues specific to rows, or observations.

Plot Regression Shows a Regression Plot report, displaying a scatterplot of the data and regression lines for each level of the categorical effect.

This option only appears if there is exactly one continuous effect and no more than one categorical effect in the model. In that case, the Regression Plot report is provided by default.

Plot Actual by Predicted Shows an Actual by Predicted plot, which plots the observed values of Y against the predicted values of Y. This plot is the leverage plot for the whole model. See "Leverage Plots" on page 175.

Note: The Actual by Predicted Plot is shown by default when Effect Leverage or Effect Screening is selected as the Emphasis in the Fit Model launch window and the RSquare value is less than 0.999.

Plot Effect Leverage Shows a Leverage Plot report for each effect in the model. The plot shows how observations influence the test for that effect and gives insight on multicollinearity. See "Leverage Plots" on page 175.

Note: Effect Leverage Plots are shown by default when Effect Leverage is selected as the Emphasis in the Fit Model launch window and the RSquare value is less than 0.999. They appear to the right of the Whole Model report. When another Emphasis is selected, the Effect Leverage Plots appear in the Effect Details report. In all cases, the option Regression Reports > Effect Details must be selected in order for Effect Leverage plots to display.

Plot Residual by Predicted Shows a Residual by Predicted Plot report. The plot shows the residuals plotted against the predicted values of Y. You typically want to see the residual values scattered randomly about zero.

Note: The Residual by Predicted Plot is shown by default when Effect Leverage or Effect Screening is selected as the Emphasis in the Fit Model launch window and the RSquare value is less than 0.999.

Plot Residual by Row Shows a Residual by Row Plot report. The residual values are plotted against the row numbers. This plot can help you detect patterns that result from the row ordering of the observations.

Plot Studentized Residuals Shows a Studentized Residuals plot. Each point on the plot is computed using an estimate of its standard deviation obtained with the current observation deleted. These residuals are also called *RStudent* or *externally Studentized* residuals.

The limits that appear on the plot are 95% Bonferroni limits, placed at ± tQuantile(0.025/n, n-p-1), where n is the number of observations and p is the number of predictors. The confidence level is not affected by your selection in the Set Alpha Level option in the Model Specification window.

The residuals saved using Save Columns > Studentized Residuals are *not* externally Studentized.

Note: If the model contains random effects and REML is the specified Method in the launch window, the Studentized Residuals plot does not contain limits and the points that are plotted are *not* externally Studentized.

Plot Residual by Normal Quantiles Shows a Residual Normal Quantile Plot. The residual values are plotted against quantiles of the normal distribution. This plot can help you assess the assumption of normality of the residuals.

Press Shows a Press Report giving the Press statistic and its root mean square error (RMSE). The Press statistic is useful when comparing multiple models. Models with lower Press statistics are favored. (For details, see "Press" on page 178.)

Durbin-Watson Test (Not available when you specify a Frequency column.) Shows the Durbin-Watson report, which gives a statistic to test whether the residuals have first-order autocorrelation. The report also displays the autocorrelation of the residuals. This option is appropriate only for time series data and assumes that your observations are in time order.

The Durbin-Watson report contains a menu with the following option:

Significance P Value Computes and displays Prob<DW, the exact probability associated with the statistic. The computation of this exact probability can be memory and time-intensive if there are many observations.

Leverage Plots

An effect leverage plot for X is useful in the following ways:

- You can see which points might be exerting influence on the hypothesis test for X.
- You can spot unusual patterns and violations of the model assumptions.
- You can spot multicollinearity issues.

Construction

A leverage plot for an effect shows the impact of adding this effect to the model, given the other effects already in the model. For illustration, consider the construction of an effect leverage plot for a single continuous effect X. See "Horizontal Axis Scaling" on page 177 for information about the scaling of the horizontal axis in more general situations.

The response Y is regressed on all the predictors except X, and the residuals are obtained. Call these residuals the Y-residuals. Then X is regressed on all the other predictors in the model and the residuals are computed. Call these residuals the X-residuals. The X-residuals might contain information beyond what is present in the Y-residuals, which were obtained without X in the model.

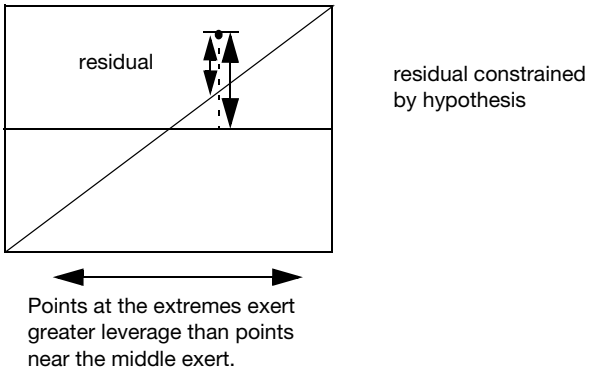
The effect leverage plot for X is essentially a scatterplot of the X-residuals against the Y-residuals (Figure 3.58). To help interpretation and comparison with other plots that you might construct, JMP adds the mean of Y to the Y-residuals and the mean of X to the X-residuals. The translated Y-residuals are called the Y Leverage Residuals and the translated X-residuals are called X Leverage values. The points on the Effect Leverage plots are these X Leverage and Y Leverage Residual pairs.

JMP fits a least squares line to these points as well as confidence bands for the mean; the line of fit is solid red and the confidence bands are shaded red. The slope of the least squares line is precisely the estimate of the coefficient on X in the model where Y is regressed on X and the other predictors. The dashed horizontal blue line is set at the mean of the Y Leverage Residuals. This line describes a situation where the X residuals are not linearly related to the Y

residuals. If the line of fit has nonzero slope, then adding X to the model can be useful in terms of explaining variation.

Figure 3.56 shows how residuals are depicted in the leverage plot. The distance from a point to the line of fit is the residual for a model that includes the effect. The distance from the point to the horizontal line is what the residual error would be without the effect in the model. In other words, the mean line in the leverage plot represents the model where the hypothesized value of the parameter (effect) is constrained to zero.

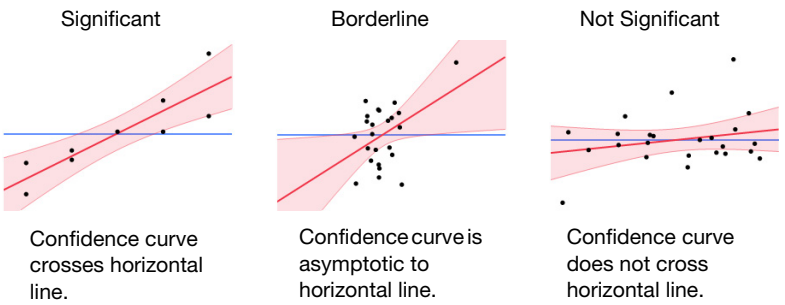
Figure 3.56 Illustration of a Generic Leverage Plot



Confidence Curves

Confidence curves for the line of fit are shown on leverage plots. These curves provide a visual indication of whether the test of interest is significant at the 5% level (or at the Set Alpha Level that you specified in the Fit Model launch window). If the confidence region between the curves contains the horizontal line representing the hypothesis, then the effect is not significant. If the curves cross the line, the effect is significant. See the examples in Figure 3.57.

Figure 3.57 Comparison of Significance Shown in Leverage Plots



Horizontal Axis Scaling

If the modeling type of a predictor X is continuous, then the horizontal axis is scaled in terms of the units of the X. The horizontal axis range mirrors the range of X values. The slope of the line of fit in the leverage plot is the parameter estimate for X. See the left illustration in Figure 3.58.

If the effect is nominal or ordinal, or if the effect is a complex effect such as an interaction, then the horizontal axis cannot represent the values of the effect directly. In this case the horizontal axis is scaled in units of the response, and the line of fit is a diagonal with a slope of 1. The Whole Model leverage plot, where the hypothesis of interest is that all parameter values are zero, uses this scaling. (See "Leverage Plot Details" on page 205.) For this plot, the horizontal axis is scaled in terms of predicted response values for the whole model, as illustrated by the right-hand plot in Figure 3.58.

The leverage plot for the linear effect in a simple regression is the same as the traditional plot of actual response values against the predictor.

Leverage

The term *leverage* is used because these plots help you visualize the influence of points on the test for including the effect in the model. A point that is horizontally distant from the center of the plot exerts more influence on the effect test than does a point that is close to the center. Recall that the test for an effect involves comparing the sum of squared residuals to the sum of squared residuals of the model with that effect removed. At the extremes, the differences of the residuals before and after being constrained by the hypothesis tend to be comparatively larger. Therefore, these residuals tend to have larger contributions to the sums of squares for that effect's hypothesis test.

Multicollinearity

Multicollinearity is a condition where two or more predictors are highly related, or more technically, involved in a nearly linear dependent relationship. When multicollinearity is present, standard errors can be inflated and parameters estimates can be unstable. If an effect is collinear with other predictors, the vertical axis values are very close to the horizontal line at the mean, because the effect brings no new information. Because of the dependency, the horizontal axis values also tend to cluster toward the middle of the plot. This situation indicates that the slope of the line of fit is unstable.

The Whole Model Actual by Predicted Plot

The Plot Effect Leverage option produces a leverage plot for each effect in the model. In addition, the Actual by Predicted plot can be considered to be a leverage plot. This plot lets you visualize the test that all the parameters in the model (except the intercept) are zero. The

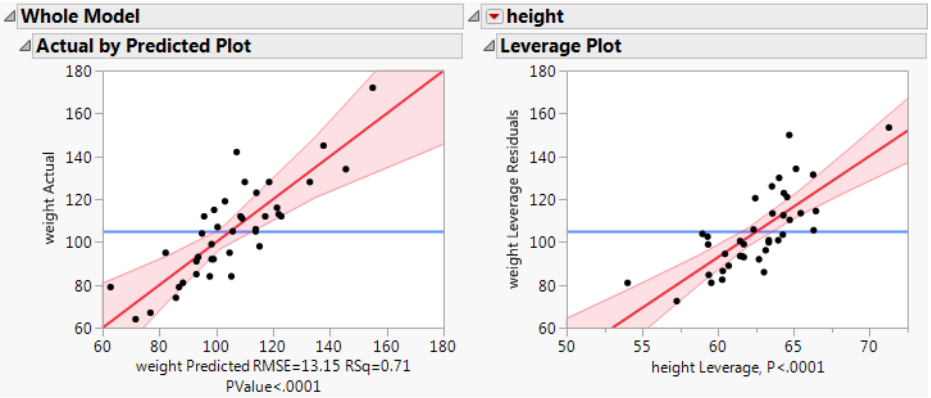
same test is conducted analytically in the Analysis of Variance report. (See "Leverage Plot Details" on page 205 for details about this plot.)

Example of a Leverage Plot for a Linear Effect

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Model**.
3. Select weight and click **Y**.
4. Select height, age, and sex, and click **Add**.
5. Click **Run**.

The Whole Model Actual by Predicted Plot and the effect Leverage Plot for height are shown in Figure 3.58. The Whole Model plot, on the left, tests for all effects. You can infer that the model is significant because the confidence curves cross the horizontal line at the mean of the response, weight. The Leverage Plot for height, on the right, also shows that height is significant, even with age and sex in the model. Neither plot suggests concerns relative to influential points or multicollinearity.

Figure 3.58 Whole Model and Effect Leverage Plots



Press

The *Press*, or *prediction error sum of squares*, statistic is an estimate of prediction error computed using leave-one-out cross validation. In leave-one-out cross validation, each observation, in turn, is removed. Consider a specific observation. The model is fit with that observation withheld and then a predicted value is obtained for that observation. The residual for that observation is computed. This procedure is applied to all observations and the residuals are squared and summed to give the Press value.

Specifically, the Press statistic is given by the following:

$$\text{Press} = \sum_{i=1}^{n} (\hat{y}_{(i)} - y_i)^2$$

where n is the number of observations, y_i is the observed response value for the i^{th} observation, and $\hat{y}_{(i)}$ is the predicted response value for the i^{th} observation. These values are based on a model fit without including that observation.

The Press RMSE is defined as $\sqrt{Press/n}$.

The Press RSquare is defined as 1 - $Press/SS_{\text{Total}}$.

Save Columns

Each Save Columns option adds one or more new columns to the current data table. Additional Save Columns options appear when the fitting method is REML. These are detailed in "REML Save Columns Options" on page 195.

Note the following:

- When formulas are created, they are entered as Formula column properties.
- For many of the new columns, a Notes column property is added describing the column and indicating that Fit Least Squares created it.
- For the Predicted Formula and Predicted Values options, a Predicting column property is added. This property is used internally by JMP in conducting model comparisons (Analyze > Predictive Modeling > Model Comparison). When you fit many models, it is also useful to you because it documents the origin of the column.

The following Save Columns options are available:

Prediction Formula Creates a new column called Pred Formula <colname> that contains both the formula and the predicted values. A Predicting column property is added, noting the source of the prediction.

Note: Pred Formula <colname> inherits certain properties from <colname>. These include Response Limits, Spec Limits, and Control Limits. If you change these properties for <colname> after saving Pred Formula <colname>, they will not update in Pred Formula <colname>.

See "Prediction Formula" on page 182 for more details.

Predicted Values Creates a new column called Predicted <colname> that contains the predicted values computed by the specified model. Both a Notes and a Predicting column property are added, noting the source of the prediction.

Note: Predicted <colname> inherits certain properties from <colname>. These include Response Limits, Spec Limits, and Control Limits. If you change these properties for <colname> after saving Predicted <colname>, they will not update in Predicted <colname>.

Residuals Creates a new column called Residual <colname> that contains the observed response values minus their predicted values.

Mean Confidence Interval Creates two new columns called Lower 95% Mean <colname> and Upper 95% Mean <colname>. These columns contain the lower and upper 95% confidence limits for the mean response.

Note: If you hold the Shift key while selecting the option, you are prompted to enter an α level for the computations.

Indiv Confidence Interval Creates two new columns called Lower 95% Indiv <colname> and Upper 95% Indiv <colname>. These columns contain lower and upper 95% confidence limits for individual response values.

Note: If you hold the Shift key while selecting the option, you are prompted to enter an α level for the computations.

Studentized Residuals Creates a new column called Studentized Resid <colname>, which contains the residuals divided by their standard errors.

Hats Creates a new column called h <colname>. The column values are the diagonal values of the matrix $X(X'X)^{-1}X'$, sometimes called hat values.

Std Error of Predicted Creates a new column called StdErr Pred <colname> that contains the standard errors of the predicted mean response.

Std Error of Residual Creates a new column called StdErr Resid <colname> that contains the standard errors of the residual values.

Std Error of Individual Creates a new column called StdErr Indiv <colname> that contains the standard errors of the individual predicted response values.

Effect Leverage Pairs Creates a set of new columns that contain the X Leverage values and Y Leverage Residuals for each leverage plot. For each effect in the model, two columns are added. If the response column name is R and the effect is X, the new column names are:

- X Leverage of X for R
- Y Leverage of X for R

In the columns panel, these columns are organized in a columns group called Leverage.

Cook's D Influence Creates a new column called Cook's D Influence <colname>, which contains values of the Cook's *D* influence statistic.

StdErr Pred Formula Creates a new column called PredSE <colname> that contains both the formula and the values for the standard error of the predicted values.

Note: The saved formula can be large. If you do not need the formula, use the Std Error of Predicted option.

Mean Confidence Limit Formula Creates two new columns called Lower 95% Mean <colname> and Upper 95% Mean <colname>. These columns contain both the formulas and the values for lower and upper 95% confidence limits for the mean response.

Note: If you hold the Shift key while selecting the option, you are prompted to enter an α level for the computations.

Indiv Confidence Limit Formula Creates two new columns called Lower 95% Indiv <colname> and Upper 95% Indiv <colname>. These columns contain both the formulas and the values for lower and upper 95% confidence limits for individual response values.

Note: If you hold the Shift key while selecting the option, you are prompted to enter an α level for the computations.

Save Coding Table Creates a new data table whose first columns show the JMP coding for all model parameters. The last column gives the values of the response. If you entered more than one response column, all of these columns appear as the last columns in the coding table.

Note: The coding data table contains a table variable called Original Data that gives the name of the data table that was used for the analysis. In the case where a By variable is specified, the Original Data table variable gives the By variable and its level.

JMP PRO **Publish Prediction Formula** Creates a prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. Seethe Formula Depot chapter in the *Predictive and Specialized Modeling* book.

JMP PRO **Publish Standard Error Formula** Creates a standard error formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

JMP PRO **Publish Mean Confid Limit Formula** Creates confidence limit formulas for the mean response and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

JMP PRO **Publish Indiv Confid Formula** Creates confidence formulas for individual response values and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. Seethe Formula Depot chapter in the *Predictive and Specialized Modeling* book.

Prediction Formula

Pred Formula <colname> differs from Predicted <colname> in that it contains the prediction formula. Right-click in the Pred Formula <colname> column heading and select **Formula** to see the prediction formula. The prediction formula can require considerable space if the model is large. If you do not need the formula with the column of predicted values, use the **Save Columns > Predicted Values** option. For information about formulas, see the Formula Editor chapter in the *Using JMP* book.

The Prediction Formula option is useful for predicting values in new rows or for use with the profilers. The profilers are available in the Fit Least Squares report (Factor Profiling). However, when your data table includes formula columns, you can also use the profilers provided in the **Graph** menu. When you are analyzing multiple responses, accessing the profilers from the **Graph** menu can be useful.

Note: If you select **Graph > Profiler** to access the profilers, first save the formula columns to the data table using Prediction Formula and StdErr Pred Formula. Then place both of these formulas into the Y, Prediction Formula role in the Profiler window. After you click **OK**, specify whether you want to use PredSE <colname> to construct confidence intervals for Pred Formula <colname>. Otherwise, JMP creates a separate profiler plot for PredSE <colname>.

Effect Summary Report

The Effect Summary option shows an interactive report. It gives a plot of the LogWorth (or FDR LogWorth) values for the effects in the model. The report also provides controls that enable you to add or remove effects from the model. The model fit report updates automatically based on the changes made in the Effects Summary report.

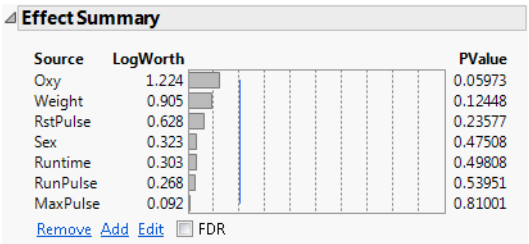
The Effect Summary report is available in the following personalities:

- Standard Least Squares

- Nominal Logistic
- Ordinal Logistic
- Proportional Hazard
- Parametric Survival
- Generalized Linear Model

Figure 3.59 shows the initial view of the Effect Summary report for the Fitness.jmp data table. The check box labeled **FDR** controls the columns that appear in the summary table.

Figure 3.59 Effect Summary Report



Effect Summary Table Columns

The Effect Summary table contains the following columns:

Source Lists the model effects, sorted by ascending *p*-values.

LogWorth Shows the LogWorth for each model effect, defined as -$\log_{10}$(*p*-value). This transformation adjusts *p*-values to provide an appropriate scale for graphing. A value that exceeds 2 is significant at the 0.01 level (because $-\log_{10}(0.01) = 2$).

FDR LogWorth Shows the False Discovery Rate LogWorth for each model effect, defined as -$\log_{10}$(FDR PValue). This is the best statistic for plotting and assessing significance. However, it is highly dependent on the ordering of the significances, is conservative for positively correlated tests, and does not give experiment-wise protection at the alpha level. Select the **FDR** check box to replace the LogWorth column with the **FDR LogWorth** column.

Bar Graph Shows a bar graph of the LogWorth (or FDR LogWorth) values. The graph has dashed vertical lines at integer values and a blue reference line at 2.

PValue Shows the *p*-value for each model effect. This is generally the *p*-value corresponding to the significance test displayed in the Effect Tests table or Effect Likelihood Ratio Tests table of the model report.

FDR PValue Shows the False Discovery Rate *p*-value for each model effect calculated using the Benjamini-Hochberg technique. This technique adjusts the *p*-values to control the false

discovery rate for multiple tests. Select the **FDR** check box to replace the **PValue** column with the **FDR PValue** column.

For more information about the FDR correction, see Benjamini and Hochberg (1995). For more information about the false discovery rate, see the Response Screening chapter in the *Predictive and Specialized Modeling* book or Westfall et al. (2011).

Effect Heredity Column Identifies lower-order effects that are components of more significant higher-order effects. The lower-order effects are identified with a caret. See "Effect Heredity" on page 185.

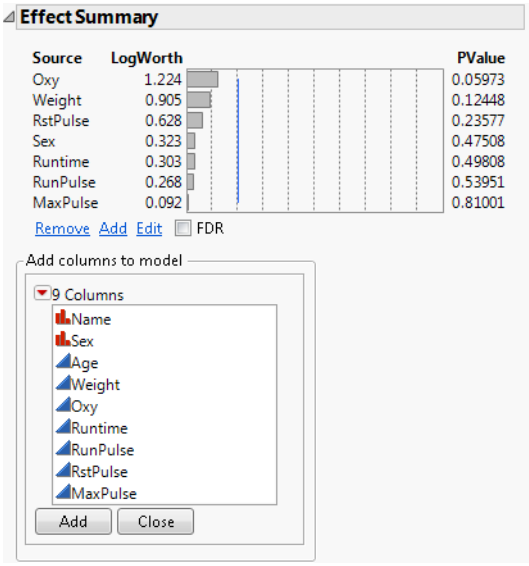
Effect Summary Table Options

The options below the summary table enable you to add and remove effects:

Remove Removes the selected effects from the model. To remove one or more effects, select the rows corresponding to the effects and click the **Remove** button.

Add Opens a panel that contains a list of all columns in the data table. Select columns that you want to add to the model, and then click **Add** below the column selection list to add the columns to the model. Click **Close** to close the panel. Figure 3.60 shows the Add Columns panel.

Figure 3.60 Effect Summary Add Columns Panel



Edit Opens the Edit Model panel, which contains a Select Columns list and an Effects specification panel. The Effects panel resembles the Construct Model Effects panel in the Fit Model launch window. The Edit Model panel enables you to add individual, crossed,

nested, and transformed effects. You can also add multiple effects using the Macros menu. For details on how to construct effects using Add, Cross, Nest, Macros, and Transform, see "Construct Model Effects" on page 42.

The options to the right of the Effects panel do the following:

– **Commit** applies your updates to the model.

– **Close** closes the panel without making changes to the model.

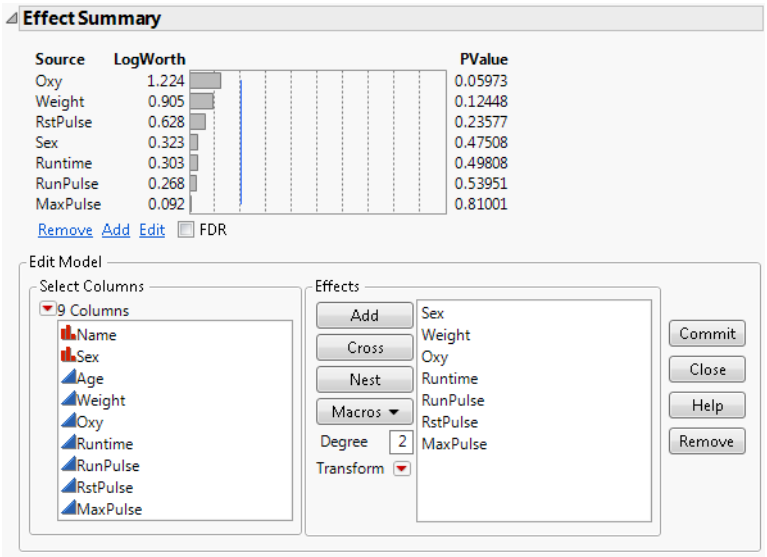
– **Remove** removes one or more selected effects from the Effects list.

Figure 3.61 shows the Edit Model panel.

Tip: The Edit button gives you the greatest degree of control over updates to your model. It includes the functionality of the Remove and Add buttons and allows you to construct effects to add to your model.

Undo Enables you to undo changes to the effects in the model.

Figure 3.61 Effect Summary Edit Model Panel



Effect Heredity

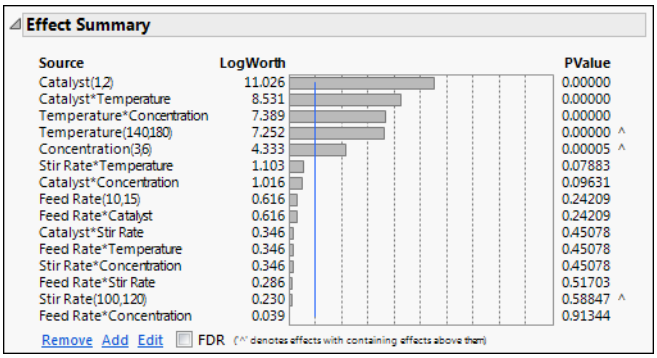
When a model contains significant higher-order effects, you may want to retain some or all of their lower-order components, even though these are not significant. The principle of *strong effect heredity* states that, if a higher-order effect is included in the model, all of its lower-order components should be included as well. The principle of *weak effect heredity* indicates that a chain of components should be included.

When a lower-order component of a higher-order effect appears in the Effect Summary table below the higher-order effect, a caret appears in the right-most column. The caret indicates that the containing higher-order effect is more significant than the lower-order effect. If all higher-order effects that contain a lower-order effect are less significant than the lower-order effect, no caret appears in the row for the lower-order effect.

If you remove an effect marked with a caret, you can choose one of two approaches for removing effects. Choose **Remove all selected effects** to remove all the selected effects, including the ones marked with a caret. Choose **Remove only non-contained effects** to remove only the selected effects that do not have a higher-order effect that still remains in the model.

Figure 3.62 show an example of an Effect Summary table where three lower-order effects appear below higher-order effects that contain the lower-order effects. For example, Stir Rate(100,120) appears below Stir Rate*Temperature.

Figure 3.62 Effect Summary Table with Effect Heredity for Reactor 32 Runs.jmp



Multiple Responses

In the case of multiple responses, each effect appears in each response model, but only one Effect Summary report appears. For each effect, the table shows the minimum *p*-value among the *p*-values for that effect. Adding or removing an effect applies to the models for all of the responses.

Mixed and Random Effect Model Reports and Options

Mixed and random effect models can be specified in the Fit Model launch window. The Standard Least Squares personality fits the variance component covariance structure. Two methods, REML and EMS, are provided for fitting such models.

Note: JMP PRO JMP Pro users are encouraged to use the Mixed Model personality in the Fit Model window. The Mixed Model personality offers a broader set of covariance structures than does Standard Least Squares.

Mixed Models and Random Effect Models

A *random effect model* is a model all of whose factors represent random effects. (See "Random Effects" on page 187.) Such models are also called *variance component models*. Random effect models are often hierarchical models. A model that contains both fixed and random effects is called a *mixed model*. Repeated measures and split-plot models are special cases of mixed models. Often the term *mixed model* is used to subsume random effect models.

To fit a mixed model, you must specify the random effects in the Fit Model launch window. However, if all of your model effects are random, you can also fit your model in the Variability / Attribute Gauge Chart platform. Only certain models can be fit in this platform. Note that the fitting methods used in the Variability / Attribute Gauge Chart platform do not allow variance component estimates to be negative. For details about how the Variability / Attribute Gauge Chart platform fits variance components models, see the Variability Gauge Charts and Attribute Gauge Charts chapter in the *Quality and Process Methods* book.

Random Effects

A random effect is a factor whose levels are considered a random sample from some population. Often, the precise levels of the random effect are not of interest, rather it is the variation reflected by the levels that is of interest (the *variance components*). However, there are also situations where you want to predict the response for a given level of the random effect. Technically, a random effect is considered to have a normal distribution with mean zero and nonzero variance.

Suppose that you are interested in whether two specific ovens differ in their effect on mold shrinkage. An oven can process only one batch of 50 molds at a time. You design a study where three randomly selected batches of 50 molds are consecutively placed in each of the two ovens. Once the batches are processed, shrinkage is measured for five parts randomly selected from each batch.

Note that Batch is a factor with six levels, once for each batch. So, in your model, you include two factors, Oven and Batch. Because you are specifically interested in comparing the effect of each oven on shrinkage, Oven is a fixed effect. But you are not interested in the effect of these

specific six batches on the mean shrinkage. These batches are representative of a whole population of batches that could have been chosen for this experiment and to which the results of the analysis must generalize. Batch is considered a random effect. In this experiment, the Batch factor is of interest in terms of the variation in shrinkage among all possible batches. Your interest is in estimating the amount of variation in shrinkage that it explains. (Note that Batch is also nested within Oven, because only one batch can be processed once in one oven.)

Now suppose that you are interested in the weight of eggs for hens subjected to two feed regimes. Ten hens are randomly assigned to feed regimes: Five are given Feed regime A and five are given Feed regime B. However, these ten hens have some genetic differences that are not accounted for in the design of the study. In this case, you are interested the predicted weight of the eggs from certain specific hens as well as in the variance of the weights of eggs among hens.

The Classical Linear Mixed Model

JMP fits the classical linear mixed effects model:

$$Y = X\beta + Z\gamma + \varepsilon$$
$$\gamma \sim N(0, G)$$
$$\varepsilon \sim N(0, \sigma^2 I_n)$$

Here,

- Y is an n x 1 vector of responses
- **X** is the **n** x **p** design matrix for the fixed effects
- β is a p x 1 vector of unknown fixed effects with design matrix X
- Z is the **n** x **s** design matrix for the random effects
- γ is an s x 1 vector of unknown random effects with design matrix Z
- ε is an n x 1 vector of unknown random errors
- G is an **s** x **s** diagonal matrix with identical entries for each for each level of the random effect
- I_n is an **n** x **n** identity matrix
- γ and ε are independent

The diagonal elements of G, as well as σ^2, are called *variance components*. These variance components, together with the vector of fixed effects β and the vector of random effects γ, are the model parameters that must be estimated.

The covariance structure for this model is sometimes called the *variance component* structure (SAS Institute Inc. 2017, ch. 79). This covariance structure is the only one available in the Standard Least Squares personality.

JMP PRO The Mixed Model personality fits a variety of covariance structures, including Residual, First-order Autoregressive (or *AR(1)*), Unstructured, and Spatial. See "Repeated Structure Tab" on page 362 in the "Mixed Models" chapter for more information.

REML versus EMS for Fitting Models with Random Effects

JMP provides two methods for fitting models with random effects:

- REML, which stands for *restricted maximum likelihood* (always the recommended method)
- EMS, which stands for *expected mean square*s (use only for teaching from old textbooks)

The REML method is now the mainstream fitting methodology, replacing the traditional EMS method. REML is considerably more general in terms of applicability than the EMS method. The REML approach was pioneered by Patterson and Thompson (1974). See also Wolfinger et al. (1994) and Searle et al. (1992).

The EMS method, also called the *method of moments*, was developed before the availability of powerful computers. Researchers restricted themselves to balanced situations and used the EMS methodology, which provided computational shortcuts to compute estimates for random effect and mixed models. Because many textbooks still in use today use the EMS method to introduce models containing random effects, JMP provides an option for EMS. (See, for example, McCulloch et al., 2008; Poduri, 1997; Searle et al., 1992.)

The REML methodology performs maximum likelihood estimation of a restricted likelihood function that does not depend on the fixed-effect parameters. This yields estimates of the variance components that are then used to obtain estimates of the fixed effects. Estimates of precision are based on estimates of the covariance matrix for the parameters. Even when the data are unbalanced, REML provides useful estimates, tests, and confidence intervals.

The EMS methodology solves for estimates of the variance components by equating observed mean squares to expected mean squares. For balanced designs, a complex set of rules specifies how estimates are obtained. There are problems in applying this technique to unbalanced data.

For balanced data, REML estimates are identical to EMS estimates. But, unlike EMS, REML performs well with unbalanced data.

Specifying Random Effects and Fitting Method

Models with random effects are specified in the Fit Model launch window. To specify a random effect, highlight it in the Construct Model Effects list and select **Attributes > Random Effect**. This appends &Random to the effect name in the model effect list. (For a definition of

random effects, see "Random Effects" on page 187.) Random effects can also be specified in a separate effects tab. (See "Construct Model Effects Tabs" on page 48 in the "Model Specification" chapter.)

In the Fit Model launch window, once the &Random attribute has been appended to an effect, you are given a choice of fitting Method: REML (Recommended) or EMS (Traditional).

Caution: You must declare crossed and nested relationships explicitly. For example, a subject ID might also identify the group that contains the subject, as when each subject is in only one group. In such a situation, subject ID must still be declared as nested within group. Take care to be explicit in defining the design structure.

Unrestricted Parameterization for Variance Components

There are two different approaches to parameterizing the variance components: the *unrestricted* and the *restricted* approaches. The issue arises when there are mixed effects in the model, such as the interaction of a fixed effect with a random effect. Such an interaction term is considered to be a random effect.

In the restricted approach, for each level of the random effect, the sum of the interaction effects across the levels of the fixed effect is assumed to be zero. In the unrestricted approach, the mixed terms are simply assumed to be independent random realizations of a normal distribution with mean 0 and common variance. (This assumption is analogous to the assumption typically applied to residual error.)

JMP and SAS use the unrestricted approach. This distinction is important because many statistics textbooks use the restricted approach. Both approaches have been widely taught for 60 years. For a discussion of both approaches, see Cobb (1998, Section 13.3).

Negative Variances

Though variances are always positive, it is possible to have a situation where the unbiased estimate of the variance is negative. Negative estimates can occur in experiments when an effect is very weak or when there are very few levels corresponding to a variance component. By chance, the observed data can result in an estimate that is negative.

Unbounded Variance Components

JMP can produce negative estimates for both REML and EMS. For REML, there are two options in the Fit Model launch window: Unbounded Variance Components and Estimate Only Variance Components. The Unbounded Variance Components option is selected by default. Deselecting this option constrains variance component estimates to be nonnegative.

You should leave the Unbounded Variance Components option selected if you are interested in fixed effects. *Constraining the variance estimates to be nonnegative leads to bias in the tests for the fixed effects.*

Estimate Only Variance Components

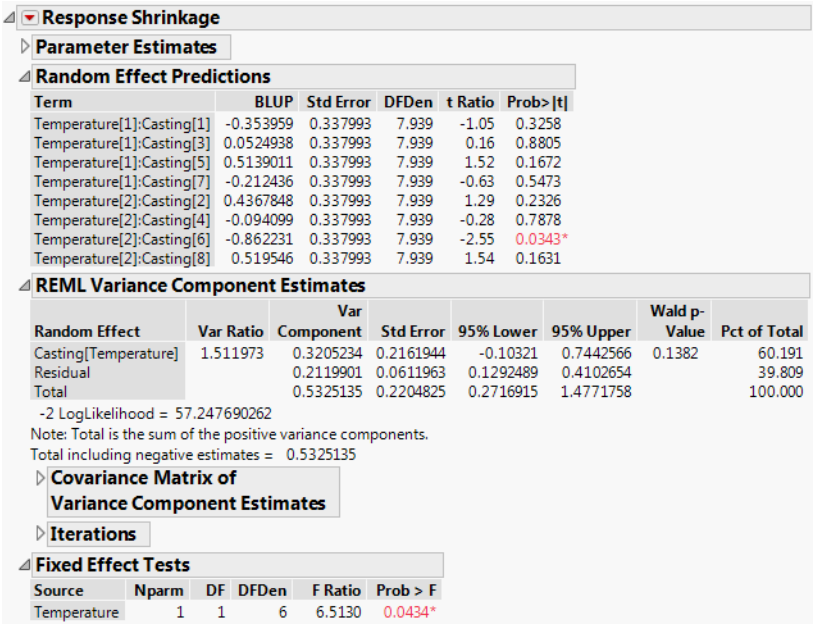
Select this option if you want to see only the REML Variance Component Estimates report. If you are interested only in variance components, you might want to constrain variance components to be nonnegative. Deselecting the Unbounded Variance Components option and selecting the Estimate Only Variance Components option might be appropriate.

Restricted Maximum Likelihood (REML) Method

Based on the fitting method selected, the Fit Least Squares report provides different analysis results and provide additional menu options for Save Columns and Profiler. In particular, the analysis of variance report is not shown because variances and degrees of freedom do not partition in the usual way. You can obtain the residual variance estimate from the REML Variance Component Estimates report. (See "REML Variance Component Estimates" on page 193.) The Effect Tests report is replaced by the Fixed Effect Tests report where fixed effects are tested. Additional reports give predicted values for the random effects and details about the variance components.

Figure 3.63 shows the report obtained for a fit to the Investment Castings.jmp sample data using the REML method. Run the script **Model - REML**, and then fit the model. Note that Casting is a random effect and is nested within Temperature.

Figure 3.63 Fit Least Squares Report for REML Method



Response Shrinkage

Parameter Estimates

Random Effect Predictions

Term	BLUP	Std Error	DFDen	t Ratio	Prob>|t|
Temperature[1]:Casting[1]	-0.353959	0.337993	7.939	-1.05	0.3258
Temperature[1]:Casting[3]	0.0524938	0.337993	7.939	0.16	0.8805
Temperature[1]:Casting[5]	0.5139011	0.337993	7.939	1.52	0.1672
Temperature[1]:Casting[7]	-0.212436	0.337993	7.939	-0.63	0.5473
Temperature[2]:Casting[2]	0.4367848	0.337993	7.939	1.29	0.2326
Temperature[2]:Casting[4]	-0.094099	0.337993	7.939	-0.28	0.7878
Temperature[2]:Casting[6]	-0.862231	0.337993	7.939	-2.55	0.0343*
Temperature[2]:Casting[8]	0.519546	0.337993	7.939	1.54	0.1631

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Wald p-Value	Pct of Total
Casting[Temperature]	1.511973	0.3205234	0.2161944	-0.10321	0.7442566	0.1382	60.191
Residual		0.2119901	0.0611963	0.1292489	0.4102654		39.809
Total		0.5325135	0.2204825	0.2716915	1.4771758		100.000

-2 LogLikelihood = 57.247690262

Note: Total is the sum of the positive variance components.

Total including negative estimates = 0.5325135

Covariance Matrix of Variance Component Estimates

Iterations

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
Temperature	1	1	6	6.5130	0.0434*

Random Effect Predictions

For each term in the model, this report gives an empirical estimate of its *best linear unbiased predictor* (BLUP) and a test for whether the corresponding coefficient is zero.

Note: The Regression Reports > Parameter Estimates option must be selected for the Random Effect Predictions report to appear.

Term The terms in the model that correspond to random effects.

BLUP The empirical estimate of the *best linear unbiased predictor* (BLUP) for each random effect. See "Best Linear Unbiased Predictors" on page 193.

Std Error The standard error of the BLUP.

DFDen The denominator degrees of freedom for a test that the effect is zero. In most cases, the degrees of freedom for the *t* test is fractional.

t Ratio The *t* ratio for testing that the effect is zero. The *t* ratio is obtained by dividing the BLUP by its standard error.

Prob>|t| The *p*-value for the test.

Lower 95% The lower 95% confidence limit for the BLUP. This column appears only if you have the Regression Reports > Show All Confidence Intervals option selected or if you right-click in the report and select Columns > Lower 95%.

Upper 95% The upper 95% confidence limit for the BLUP. This column appears only if you have the Regression Reports > Show All Confidence Intervals option selected or if you right-click in the report and select Columns > Upper 95%.

Best Linear Unbiased Predictors

The term *best linear unbiased predictor* (BLUP) refers to an estimator of a random effect. Specifically, it is an estimator that, among all unbiased estimators, minimizes mean square prediction error. The Random Effect Predictions report gives estimates of the BLUPs, or *empirical* BLUPs. These are empirical because the BLUPs depend on the values of the variance components, which are unknown. The estimated values of the variance components are substituted into the formulas for the BLUPs, resulting in the estimates shown in the report.

REML Variance Component Estimates

When REML is selected as the fitting method in the Fit Model launch window, the REML Variance Component Estimates report is provided. This report contains the following columns:

Random Effect The random effects in the model.

Var Ratio The ratio of the variance component for the effect to the variance component for the residual. It compares the effect's estimated variance to the model's estimated error variance.

Var Component The estimated variance component for the effect. Note that the variance component for the Total is the sum of the positive variance components only. The sum of all variance components is given beneath the table.

Std Error The standard error for the variance component estimate.

95% Lower The lower 95% confidence limit for the variance component. See "Confidence Intervals for Variance Components" on page 194.

95% Upper The upper 95% confidence limit for the variance component. See "Confidence Intervals for Variance Components" on page 194.

Wald p-Value The *p*-value for the test that the covariance parameter is equal to zero. This column appears only when you have selected Unbounded Variance Components in the Fit Model launch window.

Pct of Total The ratio of the variance component for the effect to the variance component for the total as a percentage.

Sqrt Variance Component The square root of the corresponding variance component. It is an estimate of the standard deviation for the effect. This column appears only if you right-click in the report and select Columns > Sqrt Variance Component.

CV The coefficient of variation for the variance component. It is 100 times the square root of the variance component, divided by the mean response. This column appears only if you right-click in the report and select Columns > CV.

Norm KHC The Kackar-Harville correction. See "Kackar-Harville Correction" on page 194. This column appears only if you right-click in the report and select Columns > Norm KHC.

Confidence Intervals for Variance Components

The method used to calculate the confidence limits depends on whether you have selected Unbounded Variance Components in the Fit Model launch window. Note that Unbounded Variance Components is selected by default.

- If Unbounded Variance Components is selected, Wald-based confidence intervals are computed. These are valid asymptotically but note that they can be unreliable with small samples.
- If Unbounded Variance Components is not selected, meaning that parameters have a lower boundary constraint of zero, a Satterthwaite approximation is used (Satterthwaite 1946).

Kackar-Harville Correction

In the REML method, the standard errors of the fixed effects are estimated using estimates of the variance components. However, if variability in these estimates is not taken into account, the standard error is underestimated. To account for the increased variability, the covariance matrix of the fixed effects is adjusted using the Kackar-Harville correction (Kackar and Harville 1984 and Kenward and Roger 1997). All calculations that involve the covariance matrix of the fixed effects use this correction. These include least squares means, fixed effect tests, confidence intervals, and prediction variances. For statistical details, see "The Kackar-Harville Correction" on page 208.

Norm KHC is the Frobenius (matrix) norm of the Kackar-Harville correction. In cases where the design is fairly well balanced, Norm KHC tends to be small.

Covariance Matrix of Variance Components Estimates

This report gives an estimate of the asymptotic covariance matrix for the variance components. It is the inverse of the observed Fisher information matrix.

Iterations

The estimates of σ^2 and the variance components in *G* are obtained by maximizing a residual log-likelihood function that depends on only these parameters. An iterative procedure attempts to maximize the residual log-likelihood function, or equivalently, to minimize twice the negative of the residual log-likelihood (–2LogLike). The Iterations report provides details about this procedure.

Iter Iteration number.

-2LogLike Twice the negative log-likelihood. It is the objective function.

Norm Gradient The norm of the gradient (first derivative) of the objective function

Parameters The column labeled Parameters and the remaining columns each correspond to a random effect. The order of the columns follows the order in which random effects are listed in the REML Variance Component Estimates report. At each iteration, the value in the column is the estimate of the variance component at that point.

The convergence criterion is based on the gradient, with a default tolerance of 10^{-8}. You can change the criterion in the Fit Model launch window by selecting the option Convergence Settings > Convergence Limit and specifying the desired tolerance.

Fixed Effect Tests

When REML is used, the Effect Tests report provides tests for the fixed effects. This report contains the following columns:

Source The fixed effects in the model.

Nparm The number of parameters associated with the effect.

DF The degrees of freedom associated with the effect.

DFDen The denominator degrees of freedom. These are based on an approximation to the distribution of the statistic obtained when the covariance matrix is adjusted using the Kenward-Roger correction. See "Kackar-Harville Correction" on page 194 and "Random Effects" on page 82.

FRatio The computed *F* ratio.

Prob > F The *p*-value for the effect test.

REML Save Columns Options

When you use the REML method, six additional options appear in the Save Columns menu. These option names start with the adjective *Conditional*. This prefix indicates that the

calculations for these columns use the predicted values for the terms associated with the random effects, rather than their expected values of zero.

Conditional Pred Formula Saves the prediction formula to a new column in the data table.

Conditional Pred Values Saves the predicted values to a new column in the data table.

Conditional Residuals Saves the residuals to a new column in the data table.

Conditional Mean CI Saves the confidence interval for the mean.

Conditional Indiv CI Saves the confidence interval for individuals.

JMP PRO **Publish Conditional Formula** Creates a conditional prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

REML Profiler Option

When you use the REML method and select Factor Profiling > Profiler, a new option, Conditional Predictions, appears on the red triangle menu next to Prediction Profiler. Note that the conditional values use the predicted values for the random effects, rather than their zero expected values.

Note: The profiler displays conditional predicted values and conditional mean confidence intervals for all combinations of factors levels, some of which might not be meaningful due to nesting.

EMS (Traditional) Model Fit Reports

Caution: The use of EMS is not recommended. REML is the recommended method.

When EMS is selected as the fitting method, four new reports are displayed. The Effect Tests report is not shown, as tests for both fixed and random effects are conducted in the Tests wrt Random Effects report.

Expected Mean Squares

The expected mean square for a model effect is a linear combination of variance components and fixed effect values, including the residual error variance. This table gives the coefficients that define each model effect's expected mean square. The rows of the matrix correspond to the effects, listed to the left. The columns correspond to the variance components, identified

across the top. Each expected mean square includes the residual variance with a coefficient of one. This information is given beneath the table.

Figure 3.64 shows the Expected Mean Squares report for the Investment Castings.jmp sample data table. Run the Model - EMS script and then run the model.

Figure 3.64 Expected Mean Squares Report

Expected Mean Squares

The Mean Square per row by the Variance Component per column

EMS

	Intercept	Temperature	Casting[Temperature]&Random
Intercept	0	0	0
Temperature	0	16	4
Casting[Temperature]&Random	0	0	4

plus 1.0 times Residual Error Variance

As indicated by the table, the expected mean square for Treatment is

$$16\theta^2_{Temperature} + 4\sigma^2_{Casting[Temperature]} + \sigma^2_{Error}$$

where $\theta^2_{Temperature}$ is the sum of the squares of the effects for Treatment divided by the number of levels of Treatment minus one.

Variance Component Estimates

Estimates of the variance components are obtained by equating the expected mean squares to the corresponding observed mean squares and solving. The Variance Component Estimates report gives the estimated variance components.

Component Lists the random effects.

Var Comp Est Gives the estimate of the variance component.

Percent of Total Gives the ratio of the variance component to the sum of the variance components.

CV Gives the coefficient of variation for the variance component. It is 100 times the square root of the variance component, divided by the mean response.

Note: Appears only if you right-click in the report and select Columns > CV.

Test Denominator Synthesis

For each effect to be tested, an F statistic is constructed. The denominator for this statistic is the mean square whose expectation is that of the numerator mean square under the null hypothesis. This denominator is constructed, or *synthesized*, from variance components and values associated with fixed effects.

Source Shows the effect to be tested.

MS Den Gives the estimated mean square for the denominator of the *F* test.

DF Den Gives the degrees of freedom for the synthesized denominator. These are constructed using Satterthwaite's method (Satterthwaite 1946).

Denom MS Synthesis Gives the variance components used in the denominator synthesis. The residual error variance is always part of this synthesis.

Tests wrt Random Effects

Tests for fixed and random effects are presented in this report.

Source Lists the effects to be tested. These include fixed and random effects.

SS Gives the sum of squares for the effect.

MS Num Gives the numerator mean square.

DF Num Gives the numerator degrees of freedom.

F Ratio Gives the *F* ratio for the test. It is the ratio of the numerator mean square to the denominator mean square. The denominator mean square can be obtained from the Test Denominator Synthesis report.

Prob > F Gives the *p*-value for the effect test.

Caution: Standard errors for least squares means and denominators for contrast *F* tests use the synthesized denominator. In certain situations, such as tests involving crossed effects compared at common levels, these tests might not be appropriate. Custom tests are conducted using residual error, and leverage plots are constructed using the residual error, so these also might not be appropriate.

EMS Profiler

When you use the EMS method and select Factor Profiling > Profiler, the profiler gives predictions and conditional mean confidence intervals based on the fixed-effects model. These values are not based on the predicted values for the random effects.

Models with Linear Dependencies among Model Terms

When there are linear dependencies among the columns of the matrix of predictors, several standard least squares reports are affected.

- "Singularity Details"
- "Parameter Estimates Report"
- "Effect Tests Report"
- "Examples"

Singularity Details

The linear regression model is formulated as $Y = X\beta + \varepsilon$. Here X is a matrix whose first column consists of *1*s, and whose remaining columns are the values of the non-intercept terms in the model. If the model consists of p terms, including the intercept, then X is an n by p matrix, where n is the number of observations. The parameter estimates, denoted by the vector b, are typically given by the formula:

$$b = (X'X)^{-1}X'Y$$

However, this formula presumes that $X'X^{-1}$ exists, in other words, that the p x p matrix $X'X$ is invertible, or equivalently, of full rank. Situations often arise when $X'X$ is not invertible because there are linear dependencies among the columns of X.

In such cases, the matrix $X'X$ is singular, and the Fit Least Squares report contains the Singularity Details report. This report contains a table of expressions that describe the linear dependencies. The terms involved in these linear dependencies are aliased (confounded).

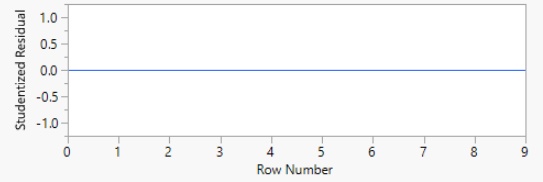
Figure 3.65 shows reports for the Reactor 8 Runs.jmp sample data table. To obtain these reports, fit a model with Percent Reacted as Y. Enter Feed Rate, Catalyst, Stir Rate, Temperature, Concentration, Catalyst*Stir Rate, Catalyst*Concentration, and Feed Rate*Catalyst as model effects.

Figure 3.65 Singularity and Parameter Estimates Report for Model with Linear Dependencies

Singularity Details

Term	Details
Concentration(3,6)	=Catalyst*Stir Rate
Stir Rate(100,120)	=Catalyst*Concentration

Studentized Residuals



Externally Studentized Residuals with 95% Simultaneous Limits (Bonferroni)

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		65.875	6.125	10.76	0.0590
Feed Rate(10,15)		-4.875	6.125	-0.80	0.5720
Catalyst(1,2)		10.125	6.125	1.65	0.3463
Stir Rate(100,120)	Biased	1.875	6.125	0.31	0.8109
Temperature(140,180)		5.875	6.125	0.96	0.5133
Concentration(3,6)	Biased	-3.375	6.125	-0.55	0.6794
Catalyst*Stir Rate	Zeroed	0	0	.	.
Catalyst*Concentration	Zeroed	0	0	.	.
Feed Rate*Catalyst		0.375	6.125	0.06	0.9611

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
Feed Rate(10,15)	1	1	190.12500	0.6335	0.5720	
Catalyst(1,2)	1	1	820.12500	2.7326	0.3463	
Stir Rate(100,120)	1	0	0.00000	.	.	LostDFs
Temperature(140,180)	1	1	276.12500	0.9200	0.5133	
Concentration(3,6)	1	0	0.00000	.	.	LostDFs
Catalyst*Stir Rate	1	0	0.00000	.	.	LostDFs
Catalyst*Concentration	1	0	0.00000	.	.	LostDFs
Feed Rate*Catalyst	1	1	1.12500	0.0037	0.9611	

Parameter Estimates Report

When $X'X$ is singular, a generalized inverse is used to obtain estimates. This approach permits some, but not all, of the parameters involved in a linear dependency to be estimated. Parameters are estimated based on the order of entry of their associated terms into the model, so that the last terms entered are the ones whose parameters are not estimated. Estimates are given in the Parameter Estimates report, and parameters that cannot be estimated are given estimates of 0.

However, estimates of parameters for terms involved in linear dependencies are not unique. Because the associated terms are aliased, there are infinitely many vectors of estimates that satisfy the least squares criterion. In these cases, “Biased” appears to the left of these estimates in the Parameter Estimates report. “Zeroed” appears to the left of the estimates of 0 in the Parameter Estimates report for terms involved in a linear dependency whose parameters cannot be estimated. For an example, see Figure 3.65.

If there are degrees of freedom available for an estimate of error, *t* tests for parameters estimated using biased estimates are conducted. These tests should be interpreted with caution, though, given that the estimates are not unique.

Effect Tests Report

In a standard least squares fit, only as many parameters are estimable as there are model degrees of freedom. In conducting the tests in the Effect Tests report, each effect is considered to be the last effect entered into the model.

- If all the Model degrees of freedom are used by the other effects, an effect shows DF equal to 0. When DF equals 0, no sum of squares can be computed. Therefore, the effect cannot be tested.
- If not all Model degrees of freedom are used by the other effects, then that effect has nonzero DF. However, its DF might be less than its number of parameters (Nparm), indicating that only some of its associated parameters are testable.

An *F* test is conducted if the degrees of freedom for an effect are nonzero, assuming that there are degrees of freedom for error. Whenever DF is less than Nparm, the description LostDFs is displayed to the far right in the row corresponding to the effect (Figure 3.65). These effects have the opportunity to explain only model sums of squares that have not been attributed to the aliased effects that have absorbed their lost degrees of freedom. It follows that the sum of squares given in the Effect Tests report most likely under represents the "true" sum of squares associated with the effect. If the test is significant, its significance is meaningful. But lack of significance should be interpreted with caution.

For more details, see the section "Statistical Background" in the "Introduction to Statistical Modeling with SAS/STAT Software" chapter of *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. 2017).

Examples

Open the Singularity.jmp sample data table. There is a response Y, four predictors X1, X2, X3, and A, and five observations. The predictors are continuous except for A, which is nominal with four levels. Also note that there is a linear dependency among the continuous effects, namely, X3 = X1 + X2.

Non-Uniqueness of Estimates

To see that estimates are not unique when there are linear dependencies:

1. Select **Help > Sample Data Library** and open Singularity.jmp.
2. Run the script **Model 1**. The script opens a Fit Model launch window where the effects are entered in the order X1, X2, X3.
3. Click **Run** and leave the report window open.
4. Run the script **Model 2**. The script opens a Fit Model launch window where the effects are entered in the order X1, X3, X2.

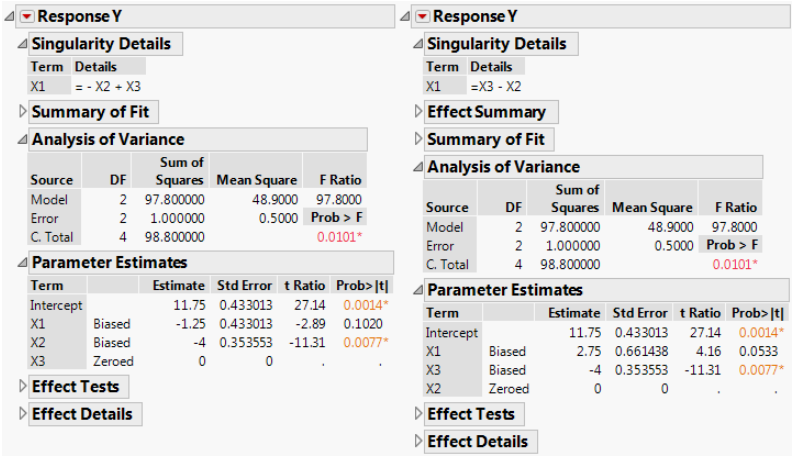
5. Click **Run** and leave the report window open.

Compare the two reports (Figure 3.66). The Singularity Details report at the top of both reports displays the linear dependency, indicating that X1 = X3 - X2.

Now compare the Parameter Estimates reports for both models. Note, for example, that the estimate for X1 for Model 1 is –1.25 while for Model 2 it is 2.75. In both models, only two of the terms associated with effects are estimated, because there are only two model degrees of freedom. See the Analysis of Variance report. The estimates of the two terms that are estimated are labeled Biased while the remaining estimate is set to 0 and labeled Zeroed.

The Effect Tests report shows that no tests are conducted. Each row is labeled LostDFs. The reason this happens is as follows. The effect test for any one of these effects requires it to be entered into the model last. However, the other two effects entirely account for the model sum of squares associated with the two model degrees of freedom. So there are no degrees of freedom or associated sum of squares left for the effect of interest.

Figure 3.66 Fit Least Squares Reports for Model 1 (on left) and Model 2 (on right)



Response Y

Singularity Details

Term	Details
X1	= - X2 + X3

Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	97.800000	48.9000	97.8000
Error	2	1.000000	0.5000	Prob > F
C. Total	4	98.800000		0.0101*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		11.75	0.433013	27.14	0.0014*
X1	Biased	-1.25	0.433013	-2.89	0.1020
X2	Biased	-4	0.353553	-11.31	0.0077*
X3	Zeroed	0	0	.	.

Effect Tests

Effect Details

Response Y

Singularity Details

Term	Details
X1	=X3 - X2

Effect Summary

Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	2	97.800000	48.9000	97.8000
Error	2	1.000000	0.5000	Prob > F
C. Total	4	98.800000		0.0101*

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		11.75	0.433013	27.14	0.0014*
X1	Biased	2.75	0.661438	4.16	0.0533
X3	Biased	-4	0.353553	-11.31	0.0077*
X2	Zeroed	0	0	.	.

Effect Tests

Effect Details

LostDFs

To gain more insight on LostDFs, follow the steps below or run the data table script Fit Model Report:

1. Select **Help > Sample Data Library** and open Singularity.jmp.
2. Click **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Select X1 and A and click **Add**.
5. Set the **Emphasis** to **Minimal Report**.
6. Click **Run**.

Portions of the report are shown in (Figure 3.67). The Singularity Details report shows that there is a linear dependency involving X1 and the three terms associated with the effect A. (For details about how a nominal effect is coded, see "Details of Custom Test Example" on page 204). The Analysis of Variance report shows that there are three model degrees of freedom. The Parameter Estimates report shows Biased estimates for the three terms X1, A[a], and A[b] and a Zeroed estimate for the fourth, A[c].

The Effect Tests report shows that X1 cannot be tested, because A must be entered first and A accounts for the three model degrees of freedom. However, A can be tested, but with only two degrees of freedom. (X1 must be entered first and it accounts for one of the model degrees of freedom.) The test for A is partial, so it must be interpreted with care.

Figure 3.67 Fit Least Squares Report for Model with X1 and A

Response Y

Singularity Details

Term	Details
Intercept	=2*X1 - A[a] - A[b] + 3*A[c]

Regression Plot

Summary of Fit

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	3	98.300000	32.7667	65.5333
Error	1	0.500000	0.5000	Prob > F
C. Total	4	98.800000		0.0905

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept	Biased	14.416667	0.399653	36.07	0.0176*
X1	Biased	-2.583333	0.399653	-6.46	0.0977
A[a]	Biased	-5.333333	0.471405	-11.31	0.0561
A[b]	Biased	3.1666667	0.552771	5.73	0.1100
A[c]	Zeroed	0	0	.	.

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
X1	1	0	0.000000	.	.	LostDFs
A	3	2	64.500000	64.5000	0.0877	LostDFs

Effect Details

Statistical Details for the Standard Least Squares Personality

Emphasis Rules

The default emphasis in the Fit Model launch window is based on the number of rows, n, the number of effects (k) entered in the Construct Model Effects list, and the attributes applied to effects.

- If $n > 1000$, the Emphasis is set to Minimal Report.
- If $n \leq 1000$ and $k \leq 4$, the Emphasis is set at Effect Leverage
- If $n \leq 1000$ and $k \geq 10$, the Emphasis is set at Effect Screening.
- If $n \leq 1000$ and $4 < k < 10$ and $n - k > 20$, the Emphasis is set at Effect Leverage.
- If any effect has a Random Effect attribute, the Emphasis is set to Minimal Report.
- If none of these conditions hold, the Emphasis is set at Effect Screening.

Details of Custom Test Example

In "Example of a Custom Test" on page 126, you are interested in testing three contrasts using the Cholesterol.jmp sample data table. Specifically, you want to compare:

- the mean responses for treatments A and B,
- the mean response for treatments A and B combined to the mean response for the control group,
- the mean response for treatments A and B combined to the mean response for the combined control and placebo groups.

To derive the contrast coefficients that you enter into the Custom Test columns, do the following. Denote the theoretical effects for the four treatment groups as: α_A, α_B, $\alpha_{Control}$, and $\alpha_{Placebo}$. These are the treatment effects, so they are constrained to sum to 0. Because the parameters associated with the indicator variables represent only the first three effects, you need to formulate your contrasts in terms of these first three effects. See "Details of Custom Test Example" on page 204 and "Interpretation of Parameters" on page 528 in the "Statistical Details" appendix for more information.

The hypotheses that you want to test can be written in terms of model effects as follows:

- Compare treatment A to treatment B: $\alpha_A - \alpha_B = 0$
- Compare treatments A and B to the control group: $0.5(\alpha_A + \alpha_B) - \alpha_{Control} = 0$
- Compare treatments A and B to the control and placebo groups:

  $$0.5(\alpha_A + \alpha_B) - 0.5(\alpha_{Control} + \alpha_{Placebo}) = \alpha_A + \alpha_B = 0$$

  To obtain contrast coefficients for this contrast, you need to write the placebo effect in terms of the model effects. Specifically, use the fact that $\alpha_A + \alpha_B + \alpha_{Control} + \alpha_{Placebo} = 0$. Then $\alpha_{Placebo} = -\alpha_A - \alpha_B - \alpha_{Control}$. It follows that:

$$\begin{aligned} 0.5(\alpha_A + \alpha_B) - 0.5(\alpha_{Control} + \alpha_{Placebo}) &= 0.5(\alpha_A + \alpha_B) - 0.5(\alpha_{Control} - \alpha_A - \alpha_B - \alpha_{Control}) \\ &= 0.5(\alpha_A + \alpha_B) - 0.5\alpha_{Control} + 0.5(\alpha_A + \alpha_B + \alpha_{Control}) \\ &= \alpha_A + \alpha_B \\ &= 0 \end{aligned}$$

Correlation of Estimates

Consider a data set with n observations and $p - 1$ predictors. Define the matrix **X** to be the design matrix. That is, **X** is the n by p matrix whose first column consists of *1s* and whose remaining $p - 1$ columns consist of the $p - 1$ predictor values. (Nominal columns are coded in terms of indicator predictors. Each of these is a column in the matrix X.)

The estimate of the vector of regression coefficients is

$$\hat{\beta} = (X'X)^{-1}X'Y$$

where Y represents the vector of response values.

Under the usual regression assumptions, the covariance matrix of $\hat{\beta}$ is

$$Cov(\hat{\beta}) = \sigma^2(X'X)^{-1}$$

where σ^2 represents the variance of the response.

The correlation matrix for the estimates is obtained by dividing each entry in the covariance matrix by the product of the square roots of the diagonal entries. Define V to be the diagonal matrix whose entries are the square roots of the diagonal entries of the covariance matrix:

$$V = Sqrt(Diag(Cov(\hat{\beta})))$$

Then the correlation matrix for the parameter estimates is given by:

$$Corr(\hat{\beta}) = \sigma^2 V^{-1}(X'X)^{-1}V^{-1}$$

Leverage Plot Details

Effect leverage plots are also referred to as *partial-regression residual leverage plots* (Belsley et al. 1980) or *added variable plots* (Cook and Weisberg 1982). Sall (1990) generalized these plots to apply to any linear hypothesis.

JMP provides two types of leverage plots:

- Effect Leverage plots show observations relative to the hypothesis that the effect is not in the model, given that all other effects are in the model.
- The Whole Model leverage plot, given in the Actual by Predicted Plot report, shows the observations relative to the hypothesis of no factor effects.

In the Effect leverage plot, only one effect is hypothesized to be zero. However, in the Whole Model Actual by Predicted plot, all effects are hypothesized to be zero. Sall (1990) generalizes the idea of a leverage plot to arbitrary linear hypotheses, of which the Whole Model leverage plot is an example. The details from that paper, summarized in this section, specialize to the two types of plots found in JMP.

Construction

Suppose that the estimable hypothesis of interest is

$$L\beta = 0$$

The leverage plot characterizes this test by plotting points so that the distance of each point to the sloped regression line displays the unconstrained residual. The distance to the horizontal line at 0 displays the residual when the fit is constrained by the hypothesis. The difference between the sums of squares of these two sets of residuals is the sum of squares due to the hypothesis. This value becomes the main component of the *F* test.

The parameter estimates constrained by the hypothesis can be written

$$b_0 = b - (X'X)^{-1}L'\lambda$$

Here b is the least squares estimate

$$b = (X'X)^{-1}X'y$$

and lambda is the Lagrangian multiplier for the hypothesis constraint, calculated by

$$\lambda = (L(X'X)^{-1}L')^{-1}Lb)$$

The unconstrained and hypothesis-constrained residuals are, respectively,

$$r = y - Xb$$
$$r_0 = r + X(X'X)^{-1}L'\lambda$$

For each observation, consider the point with horizontal axis value v_x and vertical axis value v_y where:

- v_x is the constrained residual minus the unconstrained residual, r_0 - r, reflecting information left over once the constraint is applied

- v_y is the horizontal axis value plus the unconstrained residual

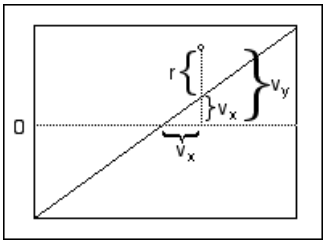
Thus, these points have x and y coordinates

$$v_x = X(X'X)^{-1}\mathbf{L}'\lambda \text{ and } v_y = r + v_x$$

These points form the basis for the leverage plot. This construction is illustrated in Figure 3.68, where the response mean is 0 and slope of the solid line is 1.

Leverage plots in JMP have a dotted horizontal line at the mean of the response, $\bar{y}$. The plotted points are given by $(v_x + \bar{y}, v_y)$.

Figure 3.68 Construction of Leverage Plot



Superimposing a Test on the Leverage Plot

In simple linear regression, you can plot the confidence limits for the expected value of the response as a smooth function of the predictor variable **x**

$$\text{Upper (x)} = \mathbf{xb} + t_{\alpha/2} s \sqrt{x(X'X)^{-1}x'}$$

$$\text{Lower (x)} = \mathbf{xb} - t_{\alpha/2} s \sqrt{x(X'X)^{-1}x'}$$

where **x** = [1 x] is the 2-vector of predictors.

These confidence curves give a visual assessment of the significance of the corresponding hypothesis test, illustrated in Figure 3.57:

- Significant: If the slope parameter is significantly different from zero, the confidence curves cross the horizontal line at the response mean.
- Borderline: If the t test for the slope parameter is sitting right on the margin of significance, the confidence curve is asymptotic to the horizontal line at the response mean.
- Not Significant: If the slope parameter is not significantly different from zero, the confidence curve does not cross the horizontal line at the response mean.

Leverage plots mirror this thinking by displaying confidence curves. These are adjusted so that the plots are suitably centered. Denote a point on the horizontal axis by *z*. Define the functions

$$\text{Upper}(z) = z + \sqrt{s^2 t^2_{\alpha/2} \bar{h} + (F_\alpha / F) z^2}$$

and

$$\text{Lower}(z) = z - \sqrt{s^2 t^2_{\alpha/2} \bar{h} + (F_\alpha / F) z^2}$$

where *F* is the *F* statistic for the hypothesis and F_α is the reference value for significance level α.

And $\bar{h} = \bar{x}(X'X)^{-1}\bar{x}'$, where $\bar{x}$ is a row vector consisting of suitable middle values for the predictors, such as their means.

These functions behave in the same fashion as do the confidence curves for simple linear regression:

- If the *F* statistic is greater than the reference value, the confidence functions cross the horizontal axis.
- If the *F* statistic is equal to the reference value, the confidence functions have the horizontal axis as an asymptote.
- If the *F* statistic is less than the reference value, the confidence functions do not cross.

Also, it is important that Upper(*z*) - Lower(*z*) is a valid confidence interval for the predicted value at z.

The Kackar-Harville Correction

The variance matrix of the fixed effects is always modified to include a Kackar-Harville correction. The variance matrix of the BLUPs, and the covariances between the BLUPs and the fixed effects, are not Kackar-Harville corrected. The rationale for this approach is that corrections for BLUPs can be computationally and memory intensive when the random effects have many levels.

In SAS, the Kackar-Harville correction is done for both fixed effects and BLUPs only when the `DDFM=KENWARDROGER` is set.

- Standard errors for linear combinations involving only fixed effects parameters match `PROC MIXED DDFM=KENWARDROGER`. This case assumes that one has taken care to transform between the different parameterizations used by `PROC MIXED` and JMP.

- Standard errors for linear combinations involving only BLUP parameters match `PROC MIXED DDFM=SATTERTHWAITE`.
- Standard errors for linear combinations involving both fixed effects and BLUPS do not match `PROC MIXED` for any `DDFM` option if the data are unbalanced. However, these standard errors are between what you get with the `DDFM=SATTERTHWAITE` and `DDFM=KENWARDROGER` options. If the data are balanced, JMP matches SAS for balanced data, regardless of the `DDFM` option, because the Kackar-Harville correction is null.

Degrees of Freedom

The degrees of freedom for tests involving only linear combinations of fixed effect parameters are calculated using the Kenward and Roger correction. So JMP's results for these tests match `PROC MIXED` using the `DDFM=KENWARDROGER` option. If there are BLUPs in the linear combination, JMP uses a Satterthwaite approximation to get the degrees of freedom. The results then follow a pattern similar to what is described for standard errors in the preceding paragraph.

For more details about the Kackar-Harville correction and the Kenward-Roger DF approach, see Kenward and Roger (1997). The Satterthwaite method is described in detail in the SAS `PROC MIXED` documentation (SAS Institute Inc. 2017, ch. 79).

Power Analysis

Options relating to power calculations are available only for continuous-response models. These are the contexts in which power and related test details are available:

Parameter Estimate

To obtain retrospective test details for each parameter estimate, select **Estimates > Parameter Power** from the report's red triangle menu. This option displays the least significant value, the least significant number, and the adjusted power for the 0.05 significance level test for each parameter based on current study data.

Effect or Effect Details

To obtain either prospective or retrospective details for the *F* test of a specific effect, select **Power Analysis** from the effect's red triangle menu. Keep in mind that, for the Effect Screening and Minimal Report personalities, the report for each effect is found under Effect Details. For the Effect Leverage personality, the report for an effect is found to the right of the first (Whole Model) column in the report.

LS Means Contrast

To obtain either prospective or retrospective details for a test of one or more contrasts, select **LSMeans Contrast** from the effect's red triangle menu. Define the contrasts of interest and click Done. From the Contrast red triangle menu, select **Power Analysis**.

Custom Test

To obtain either prospective or retrospective details for a custom test, select **Estimates > Custom Test** from the response's red triangle menu. Define the contrasts of interest and click **Done**. From the Custom Test red triangle menu, select **Power Analysis**.

In all cases except the first, selecting Power Analysis opens the Power Details window. You then enter information in the Power Details window to modify the calculations according to your needs.

Effect Size

The effect size, denoted by δ, is a measure of the difference between the null hypothesis and the true values of the parameters involved. The null hypothesis might be formulated in terms of a single linear contrast that is set equal to zero, or of several such contrasts. The value of δ reflects the difference between the true values of the contrasts and their hypothesized values of 0.

In general terms, the effect size is given by:

$$\delta = \sqrt{SS_{Hyp(Pop)}/n}$$

where $SS_{Hyp(Pop)}$ is the sum of squares for the hypothesis being tested given in terms of population parameters and n is the total number of observations.

When observations are available, the estimated effect size is calculated by substituting the calculated sum of squares for the hypothesis into the formula for δ.

Balanced One-Way Layout

For example, in the special case of a balanced one-way layout with k levels where the i^{th} group has mean response α_i,

$$\delta^2 = \frac{\sum(\alpha_i - \bar{\alpha})^2}{k}$$

Recall that JMP codes parameters so that, for i =1, 2, ..., k-1

$$\beta_i = (\alpha_i - \bar{\alpha})$$

and

$$\beta_k = -\sum_{m=1}^{k-1} \alpha_m$$

So, in terms of these parameters, δ for a two-level balanced layout is given by:

$$\delta^2 = \frac{\beta_1^2 + (-\beta_1)^2}{2} = \beta_1^2$$

or $\delta = |\beta_1|$

Unbalanced One-Way Layout

In the case of an unbalanced one-way layout with k levels, and where the i^{th} group has mean response α_i and n_i observations, and where $n = \sum n_i$:

$$\delta^2 = \sum \frac{n_i}{n}(\alpha_i - \bar{\alpha})^2$$

Effect Size and Power

The power is the probability that the F test of a hypothesis is significant at the α significance level, when the true effect size is a specified value. If the true effect size equals δ, then the test statistic has a noncentral F distribution with noncentrality parameter

$$\lambda = (n\delta^2)/\sigma^2$$

When the null hypothesis is true (that is, when the effect size is zero), the noncentrality parameter is zero and the test statistic has a central F distribution.

The power of the test increases with λ. In particular, the power increases with sample size n and effect size δ, and decreases with error variance σ^2.

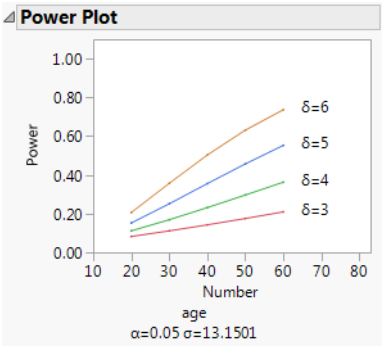
Some books, such as Cohen (1977), use a standardized effect size, $\Delta = \delta/\sigma$, rather than the raw effect size used by JMP. For the standardized effect size, the noncentrality parameter equals $\lambda = n\Delta^2$.

In the Power Details window, δ is initially set to $\sqrt{SS_{Hyp}/n}$. SS_{Hyp} is the sum of squares for the hypothesis, and n is the number of observations in the current study. SS_{Hyp} is an estimate of δ computed from the data, but such estimates are biased (Wright and O'Brien 1988). To calculate power using a sample estimate for δ, you might want to use the Adjusted Power and Confidence Interval calculation rather than the Solve for Power calculation. The adjusted power calculation uses an estimate of δ that is partially corrected for bias. See "Computations for the Adjusted Power" on page 554 in the "Statistical Details" appendix.

Plot of Power by Sample Size

To see a plot of power by sample size, select the Power Plot option from the red triangle menu at the bottom of the Power report. JMP plots the Power and Number columns from the Power table. The plot shown in Figure 3.69 results from plotting the Power table obtained in "Example of Retrospective Power Analysis" on page 214.

Figure 3.69 Plot of Power by Sample Size



The Least Significant Number (LSN)

The *least significant number* (LSN) is the smallest number of observations that leads to a significant test result, given the specified values of delta, sigma, and alpha. Recall that delta, sigma, and alpha represent, respectively, the effect size, the error standard deviation, and the significance level.

Note: LSN is *not* a recommendation of how large a sample to take because it does not take into account the probability of significance. It is computed based on specified values of delta and sigma.

The LSN has these characteristics:

- If the LSN is less than the actual sample size n, then the effect is significant.
- If the LSN is greater than n, the effect is not significant. If you believe that more data will show essentially the same structural results as does the current sample, the LSN suggests how much data you would need to achieve significance.
- If the LSN is equal to n, then the p-value is equal to the significance level alpha. The test is on the border of significance.
- The power of the test for the effect size, calculated when n = LSN, is always greater than or equal to 0.5. Note, however, that the power can be close to 0.5, which is considered low for planning purposes.

The Least Significant Value (LSV)

The LSV, or *least significant value*, is computed for single-degree-of-freedom hypothesis tests. These include tests for the significance of individual model parameters, as well as more general linear contrasts. The LSV is the smallest effect size, in absolute value, that would be significant at level alpha. The LSV gives a measure of the sensitivity of the test on the scale of the parameter, rather than on a probability scale.

The LSV has these characteristics:

- If the absolute value of the parameter estimate or contrast is greater than or equal to the LSV, then the *p*-value of the significance test is less than or equal to alpha.
- The absolute value of the parameter estimate or contrast is equal to the LSV if and only if its significance test has *p*-value equal to alpha.
- The LSV is the radius of a $1-\alpha$ confidence interval for the parameter or linear combination of parameters. The $1-\alpha$ confidence interval is centered at the estimate of the parameter or contrast.

Power

The power of a test is the probability that the test gives a significant result. The power is a function of the effect size δ, the significance level α, the error standard deviation σ, and the sample size n. The power is the probability that you will detect a specified effect size at a given significance level. In general, you would like to design studies that have high power of detecting differences that are of practical or scientific importance.

Power has these characteristics:

- If the true value of the parameter is in fact the hypothesized value, the power equals the significance level of the test. The significance level is usually a small value, such as 0.05. The small value is appropriate, because you want a low probability of seeing a significant result when the postulated hypothesis is true.
- If the true value of the parameter is *not* the hypothesized value, in general, you want the power to be as large as possible.
- Power increases as: sample size increases; error variance decreases; the difference between the true parameter value and the hypothesized value increases.

The Adjusted Power and Confidence Intervals

In retrospective power analysis, you typically substitute sample estimates for the population parameters involved in power calculations. This substitution causes the noncentrality parameter estimate to have a positive bias (Wright and O'Brien 1988). The adjusted power calculation is based on a form of the estimated noncentrality parameter that is partially corrected for this bias.

You can also construct a confidence interval for the adjusted power. Such confidence intervals tend to be wide. See Wright and O'Brien (1988).

Note that the adjusted power and confidence interval calculations are relevant only for the value of δ estimated from the data (the value provided by default). For other values of delta, the adjusted power and confidence interval are not provided.

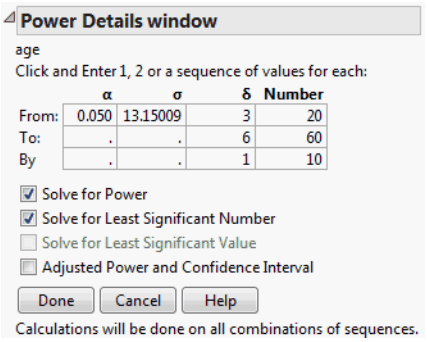
For more details, see "Computations for the Adjusted Power" on page 554 in the "Statistical Details" appendix.

Example of Retrospective Power Analysis

This example illustrates a retrospective power analysis using the Big Class.jmp sample data table. The Power Details window (Figure 3.70) permits exploration of various quantities over ranges of values for α, σ, δ, and Number, or study size. Clicking Done replaces the window with the results of the calculations.

1. Select **Help > Sample Data Library** and open Big Class.jmp.
2. Select **Analyze > Fit Model**.
3. Select weight and click **Y**.
4. Add age, sex, and height as the effects.
5. Click **Run**.
6. From the red triangle next to age, select **Power Analysis**.

Figure 3.70 Power Details Window for Age



7. Replace the δ value in the **From** box with 3, and enter 6 and 1 in the **To** and **By** boxes as shown in Figure 3.70.
8. Replace the **Number** value in the **From** box with 20, and enter 60 and 10 in the **To** and **By** boxes as shown in Figure 3.70.
9. Select **Solve for Power** and **Solve for Least Significant Number**.
10. Click **Done**.

11. The Power Details window is replaced by the Power Details report shown in Figure 3.71

Figure 3.71 Power Details Report for Age

Power Details

Test age

Power

α	σ	δ	Number	Power
0.0500	13.15009	3	20	0.0828
0.0500	13.15009	3	30	0.1117
0.0500	13.15009	3	40	0.1426
0.0500	13.15009	3	50	0.1755
0.0500	13.15009	3	60	0.2099
0.0500	13.15009	4	20	0.1117
0.0500	13.15009	4	30	0.1694
0.0500	13.15009	4	40	0.2317
0.0500	13.15009	4	50	0.2969
0.0500	13.15009	4	60	0.3630
0.0500	13.15009	5	20	0.1524
0.0500	13.15009	5	30	0.2515
0.0500	13.15009	5	40	0.3554
0.0500	13.15009	5	50	0.4575
0.0500	13.15009	5	60	0.5529
0.0500	13.15009	6	20	0.2063
0.0500	13.15009	6	30	0.3572
0.0500	13.15009	6	40	0.5035
0.0500	13.15009	6	50	0.6320
0.0500	13.15009	6	60	0.7368

Least Significant Number

α	σ	δ	Number(LSN)
0.0500	13.15009	3	216.8578
0.0500	13.15009	4	123.8892
0.0500	13.15009	5	80.93391
0.0500	13.15009	6	57.6814

This analysis is a retrospective power analysis because the calculations assume a study with a structure identical to that of the Big Class.jmp sample data table. For example, the calculation of power in this example depends on the effects entered into the model and the number of participants in each age and sex grouping. Also, the value of σ was derived from the current study, though you could have replaced it with a value that would be representative of a future study.

For details about the power results shown in Figure 3.71, see "Power" on page 213. For details about the least significant number (LSN), see "The Least Significant Number (LSN)" on page 212.

Prospective Power Analysis

Prospective analysis helps you answer the question, "If differences of a specified size exist, will I detect them given my proposed sample size, alpha level, and estimate of error variance?" In a prospective power analysis, you must provide estimates of the group means and sample sizes in a data table. You must also provide an estimate of the error standard deviation σ in the Power Details window.

Equal Group Sizes

Consider a situation where you are comparing the means of three independent groups. To obtain sample sizes to achieve a given power, select **DOE > Sample Size and Power** and then select **k Sample Means**. Next to Std Dev, enter your estimate of the error standard deviation. In the Prospective Means list, enter means that reflect the smallest differences that you want to detect. If, for example, you want to detect a difference of 8 units between any two means, enter the extreme values of the means, say, 40, 40, and 48. Because the power is based on deviations from the grand mean, you can enter only values that reflect the desired differences (for example 0, 0, and 8).

If you click **Continue**, you obtain a graph of power versus sample size. If instead you specify either power or sample size in the Sample Size window, the other quantity is computed and displayed in the Sample Size window. In particular, if you specify power, the sample size that is provided is the total required sample size. The k Sample Means calculation assumes equal group sizes. For three groups, you would divide the sample size by 3 to obtain the individual group sizes. For more information about k Sample Means, see the Prospective Sample Size and Power chapter in the *Design of Experiments Guide* book.

Unequal Group Sizes

Suppose that you want to design a study that uses groups of different sizes. You need to plan an experiment to study two treatments that reportedly reduce bacterial counts. You want to compare the effect of these treatments with results from a control group that receives no treatment. You also want to detect a difference of at least 8 units between the means of either treatment group and the control group. But the control group must be twice as large as either treatment group. The two treatment groups also must be equal in size. Previous studies suggest that the error standard deviation is on the order of 5 or 6.

To obtain a prospective power analysis for this situation, create a data table containing some basic information, as shown in the Bacteria.jmp sample data table (Figure 3.72).

Figure 3.72 Bacteria.jmp Data Table

	Group	Group Means	Relative Sizes
1	Control	40	2
2	Treatment A	40	1
3	Treatment B	48	1

- The Group column identifies the groups.
- The Means column reflects the smallest difference among the columns that it is important to detect. Here, it is assumed that the control group has a mean of about 40. You want the test to be significant if either treatment group has a mean that is at least 8 units higher than the mean of the control group. For this reason, you assign a mean of 48 to one of the two treatment groups. Set the mean of the other treatment group equal to that of the control group. (Alternatively, you could assign the control group and one of the treatment groups

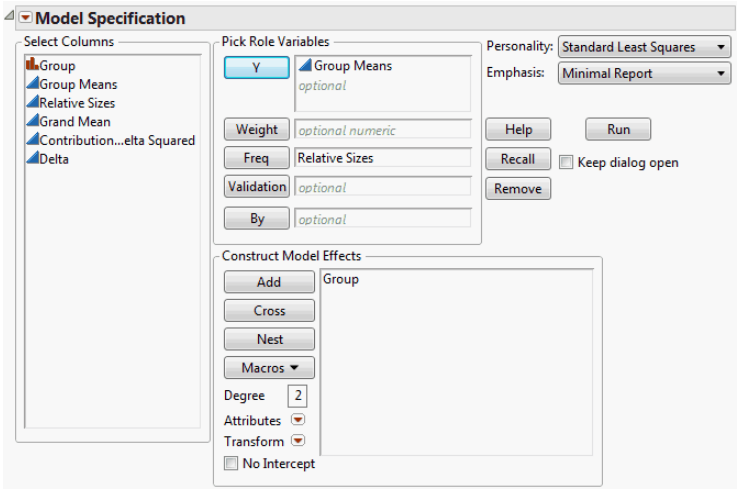
means of 0 and the remaining treatment group a mean of 8.) Note that the differences in the group means are population values.

- The Relative Sizes column shows the desired relative sizes of the treatment groups. This column indicates that the control group needs to be twice as large as each of the treatment groups. (Alternatively, you could start out with an initial guess for the treatment sizes that respects the relative size criterion.)

Note: The Relative Sizes column must be assigned the role of a Freq (frequency). See the symbol to the right of the column name in the Columns panel.

Next, use Fit Model to fit a one-way analysis of variance model (Figure 3.73). Note that Relative Sizes is declared as **Freq** in the launch window. Also, the Minimal Report emphasis option is selected.

Figure 3.73 Fit Model Launch Window for Bacteria Study

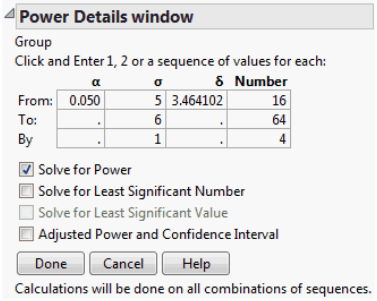


Click **Run** to obtain the Fit Least Squares report. The report shows Root Mean Square Error and Sum of Squares for Error as 0.0, because you specified a data table with no error variation within the groups. You must enter a proposed range of values for the error variation to obtain the power analysis. Specifically, you have information that the error variation will be about 5 but might be as large as 6.

1. Click the disclosure icon next to Effect Details to open this report.
2. From the red triangle menu next to Group, select **Power Analysis**.
3. To explore the range of error variation suspected by the scientist, under σ, enter 5 in the first box and 6 in the second box (Figure 3.74).

4. Note that δ is entered as 3.464102. This is the effect size that corresponds to the specified difference in the group means. The data table contains three hidden columns that illustrate the calculation of the effect size. (See "Unbalanced One-Way Layout" on page 211.)
5. To explore power over a range of study sizes, under **Number**, enter 16 in the first box, 64 in the second box, and an increment of 4 in the third box (Figure 3.74).
6. Select **Solve for Power**.
7. Click **Done**.

Figure 3.74 Power Details Window for Bacteria Study



The Power Details report, shown in Figure 3.75, replaces the Power Details window. This report gives power calculations for $\alpha = 0.05$, for all combinations of $\sigma = 5$ and 6, and sample sizes of 16 to 64 in increments of size 4. When σ is 5, to obtain about 90% power, you need a total sample size of about 32. You need 16 participants in the control group and 8 in each of the treatment groups. On the other hand, if σ is 6, then a total of 44 participants is required.

Figure 3.75 Power Details Report for Bacteria Study

Power Details

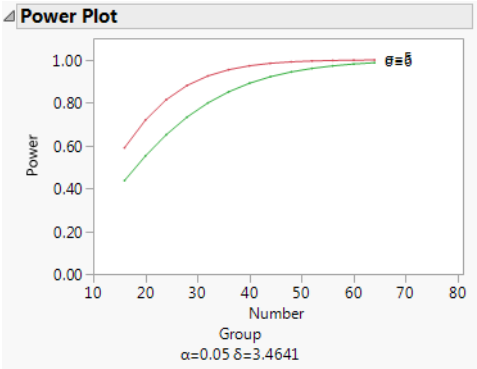
Test Group

Power

α	σ	δ	Number	Power
0.0500	5	3.464102	16	0.5894
0.0500	5	3.464102	20	0.7184
0.0500	5	3.464102	24	0.8134
0.0500	5	3.464102	28	0.8799
0.0500	5	3.464102	32	0.9246
0.0500	5	3.464102	36	0.9537
0.0500	5	3.464102	40	0.9721
0.0500	5	3.464102	44	0.9835
0.0500	5	3.464102	48	0.9903
0.0500	5	3.464102	52	0.9944
0.0500	5	3.464102	56	0.9968
0.0500	5	3.464102	60	0.9982
0.0500	5	3.464102	64	0.9990
0.0500	6	3.464102	16	0.4361
0.0500	6	3.464102	20	0.5510
0.0500	6	3.464102	24	0.6501
0.0500	6	3.464102	28	0.7323
0.0500	6	3.464102	32	0.7986
0.0500	6	3.464102	36	0.8506
0.0500	6	3.464102	40	0.8906
0.0500	6	3.464102	44	0.9208
0.0500	6	3.464102	48	0.9433
0.0500	6	3.464102	52	0.9598
0.0500	6	3.464102	56	0.9717
0.0500	6	3.464102	60	0.9803
0.0500	6	3.464102	64	0.9864

Click the arrow at the bottom of the table in the Power Details report to obtain a plot of power versus sample size for the two values of σ, shown in Figure 3.76. Here, the red markers correspond to σ = 5 and the green correspond to σ = 6.

Figure 3.76 Power Plot for Bacteria Study



Chapter 4

Standard Least Squares Examples

Analyze Common Classes of Models

This chapter provides examples with instructional material for several standard least squares topics. These include analysis of variance, analysis of covariance, a response surface model, a split plot design, estimation of random effect parameters, a knotted spline fit, and the identification of active factors using a script that provides a Bayesian approach.

Contents

One-Way Analysis of Variance Example

In a one-way analysis of variance, a different mean is fit to each of the different groups, as identified by a nominal variable. To specify the model for JMP, select a continuous response column and a nominal effect column. This example uses the data in Drug.jmp.

1. Select **Help > Sample Data Library** and open Drug.jmp.
2. Select **Analyze > Fit Model**.
3. Select y and click **Y**.
4. Select Drug and click **Add**.
5. Click **Run**.

In this example, Drug has three levels, a, d, and f. The standard least squares fitting method translates this specification into a linear model as follows: The nominal variables define a sequence of indicator variables, which only assume the values 1, 0, and –1. The linear model is written as follows:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$$

where:

- y_i is the observed response for the i^{th} observation
- x_{1i} is the value of the first indicator variable for the i^{th} observation
- x_{2i} is the value of the second indicator variable for the i^{th} observation
- β_0, β_1, and β_2 are parameters for the intercept, the first indicator variable, and the second indicator variable, respectively
- ε_i are the independent and normally distributed error terms

The first indicator variable, x_1, is defined as follows. Note that Drug=a contributes a value 1, Drug=d contributes a value 0, and Drug=f contributes a value –1 to the indicator variable:

$$x_{1i} = \begin{cases} 1, & \text{if Drug = a} \\ 0, & \text{if Drug = d} \\ -1, & \text{if Drug = f} \end{cases}$$

The second indicator variable, x_2, is given the following values:

$$x_{2i} = \begin{cases} 0, & \text{if Drug = a} \\ 1 & \text{if Drug = d} \\ -1, & \text{if Drug = f} \end{cases}$$

The estimates of the means for the three levels in terms of this parameterization are as follows:

$$\mu_a = \beta_0 + \beta_1$$

$$\mu_d = \beta_0 + \beta_2$$

$$\mu_f = \beta_0 - \beta_1 - \beta_2$$

Solving for β_i yields the following:

$$\beta_0 = \frac{(\mu_a + \mu_d + \mu_f)}{3} = \mu \text{ (the average over levels)}$$

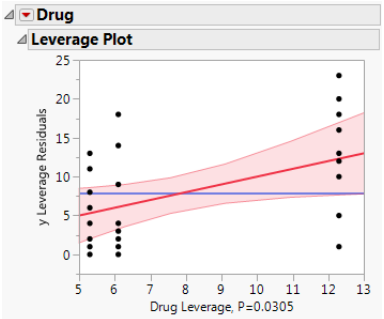
$$\beta_1 = \mu_a - \mu$$

$$\beta_2 = \mu_d - \mu$$

Therefore, if regressor variables are coded as indicators for each level minus the indicator for the last level, then the parameter for a level is interpreted as the difference between that level's response and the average response across all levels. See the appendix "Statistical Details" on page 521 for additional information about the interpretation of the parameters for nominal factors.

Figure 4.1 shows the Leverage Plot and the LS Means Table for the Drug effect. Figure 4.2 shows the Parameter Estimates and the Effect Tests reports for the one-way analysis of the drug data.

Figure 4.1 Leverage Plot and LS Means Table for Drug



Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
a	5.300000	1.9197801	5.3000
d	6.100000	1.9197801	6.1000
f	12.300000	1.9197801	12.3000

Figure 4.2 Parameter Estimates and Effect Tests for Drug.jmp

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	7.9	1.108386	7.13	<.0001*
Drug[a]	-2.6	1.567494	-1.66	0.1088
Drug[d]	-1.8	1.567494	-1.15	0.2609

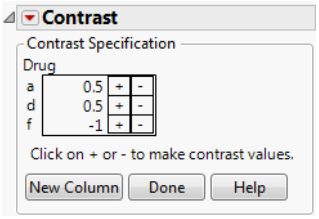
Effect Tests

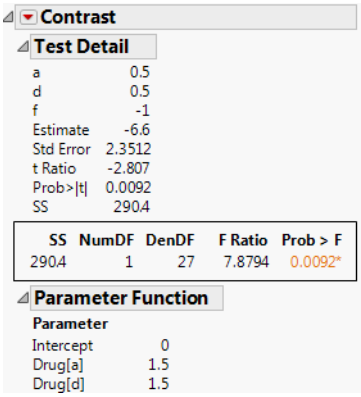
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Drug	2	2	293.60000	3.9831	0.0305*

The Drug effect can be studied in more detail by using a contrast of the least squares means, as follows:

1. From the red triangle menu next to Drug, select **LSMeans Contrast**.
2. Click the + boxes for drugs a and d, and the - box for drug f to define the contrast that compares the average of drugs a and d to f (shown in Figure 4.3).
3. Click **Done**.

Figure 4.3 Contrast Example for the Drug Experiment





The Contrast report shows that the LSMean for drug f is significantly different from the average of the LSMeans of the other two drugs.

Analysis of Covariance with Equal Slopes Example

An analysis of variance model with a continuous regressor term is called an *analysis of covariance*. In the Drug.jmp sample data table, the column x is a covariate.

The covariate adds an additional term, x_{3i}, to the model equation. The model analysis of covariance model is written this way:

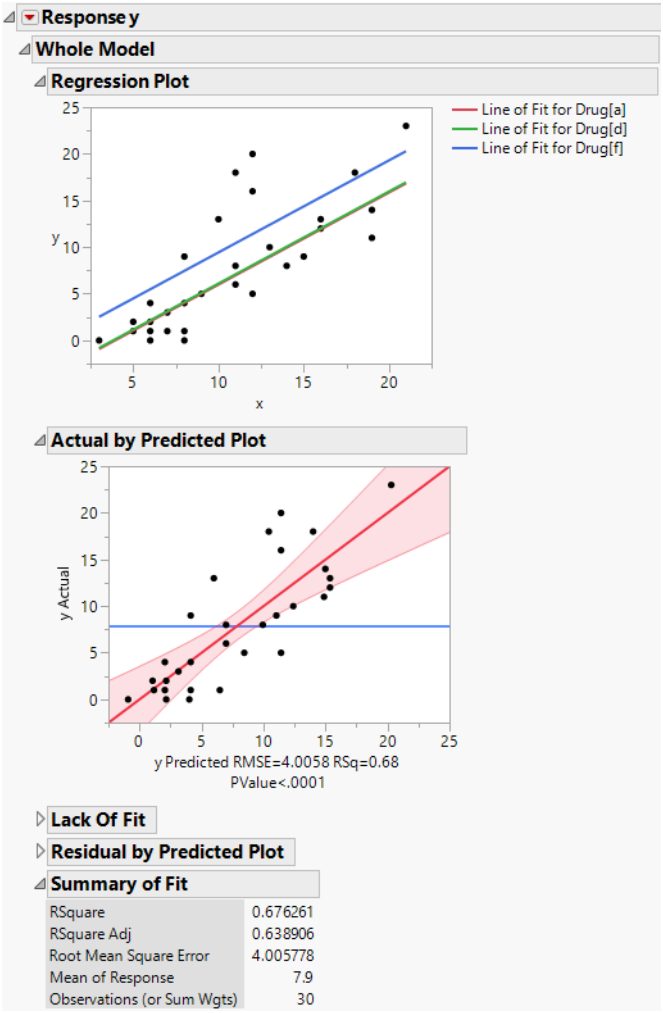
$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \varepsilon_i$$

There are two model effects: one is a nominal main effect involving two parameters, and the other is continuous covariate associated with one parameter.

1. Select **Help > Sample Data Library** and open Drug.jmp.
2. Select **Analyze > Fit Model**.
3. Select y and click **Y**.
4. Select both Drug and x and click **Add**.
5. Click **Run**.

The Regression Plot in the report shows that you have fit a model with equal slopes (Figure 4.4). Compared to the main effects model (Drug effect only), RSquare increases from 22.8% to 67.6%. The Root Mean Square Error decreases from 6.07 to 4.0. As shown in Figure 4.4, the *F*-test significance probability for the whole model decreases from 0.03 to less than 0.0001.

Figure 4.4 Analysis of Covariance for Drug Data



The drug data table contains replicated observations. For example, rows 1 and 11 both have Drug = a and x = 11. In modeling fitting, replicated observations can be used to construct a *pure error* estimate of variation. Another estimate of error can be constructed for unspecified functional forms of covariates, or interactions of nominal effects. These estimates form the basis for a lack of fit test. If the lack of fit error is significant, this indicates that there is some effect in your data not explained by your model.

The Lack of Fit report shows the results of this test for the drug data. The lack of fit error is not significant, as seen by the Prob > F value of 0.7507.

The covariate, x, accounts for much of the variation in the response previously accounted for by the Drug variable. Thus, even though the model is fit with much less error, the Drug effect is

no longer significant. The significance of Drug observed in the main effects model appears to be explained by the covariate.

The least squares means in the covariance model differ from the ordinary means. This is because they are adjusted for the effect of x, the covariate, on the response, y. The least squares means are values predicted for each of the three levels of Drug, when the covariate, x, is held at some neutral value. The neutral value is chosen to be the mean of the covariate, which is 10.7333.

The least squares means are calculated as follows, using the parameter estimates given in the Parameter Estimates report:

Prediction Expression: `-2.696 - 1.185*Drug[a] - 1.0761*Drug[d] + 0.98718*x`

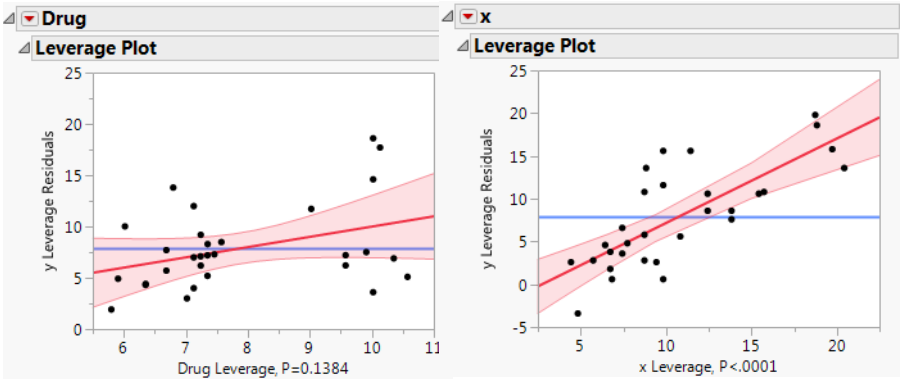
For a: `-2.696 - 1.185*(1) -1.0761*(0) + 0.98718*(10.7333) = 6.71`

For d: `-2.696 - 1.185*(0) -1.0761*(1) + 0.98718*(10.7333) = 6.82`

For f: `-2.696 - 1.185*(-1) -1.0761*(-1) + 0.98718*(10.7333) = 10.16`

Figure 4.5 shows a leverage plot for each effect. Because the covariate is significant, the leverage values for Drug are dispersed somewhat from their least squares means.

Figure 4.5 Comparison of Leverage Plots for Drug Test Data



Analysis of Covariance with Unequal Slopes Example

Continuing with the Drug.jmp sample data table, this example fits a model where the slope for the covariate depends on the level of Drug. After fitting the model, the example compares the least square means for the levels of Drug at a specific value of the covariate x.

1. Select **Help > Sample Data Library** and open Drug.jmp.
2. Select **Analyze > Fit Model**.
3. Select y and click **Y**.
4. Select both Drug and x and click **Macros > Factorial to Degree**.

   This adds terms up to the degree specified in the **Degree** box to the model. The default value for **Degree** is 2. Thus the main effects of Drug and x, and their interaction, Drug*x, are added to the model effects list.
5. Click **Run**.

This specification adds two columns to the linear model (call them x_{4i} and x_{5i}) that allow the slopes for the covariate to differ by Drug level. The new variables are formed by multiplying the indicator variables for Drug by the covariate values, giving the following formula:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + \beta_5 x_{5i} + \varepsilon_i$$

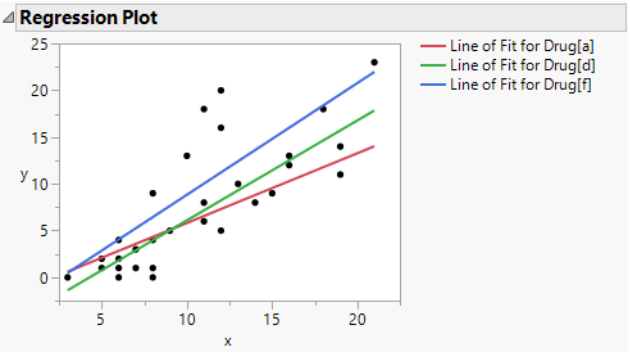
Table 4.1, shows the coding for this model. The mean of X, which is 10.7333, is used in centering continuous terms.

Table 4.1 Coding of Analysis of Covariance with Separate Slopes

Regressor	Effect	Values
X_1	Drug[a]	+1 if a, 0 if d, –1 if f
X_2	Drug[d]	0 if a, +1 if d, –1 if f
X_3	X	the values of X
X_4	Drug[a]*(X - 10.733)	X – 10.7333 if a, 0 if d, –(X – 10.7333) if f
X_5	Drug[d]*(X - 10.733)	0 if a, X – 10.7333 if d, –(X – 10.7333) if f

A portion of the report is shown in Figure 4.6. The Regression Plot shows fitted lines with different slopes. The Effect Tests report gives a p-value for the interaction of 0.56. This is not significant, indicating the model does not need to include different slopes.

Figure 4.6 Plot with Interaction



Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F
Drug	2	2	52.05637	1.5713	0.2284
x	1	1	564.56753	34.0821	<.0001*
Drug*x	2	2	19.64465	0.5930	0.5606

Perform a Spotlight Analysis

You now want to compare the least square means for the levels of Drug at a specific value of the covariate x. This type of comparison in an analysis of covariance model is sometimes referred to as *spotlight analysis*. For more information about spotlight analysis, see Spiller et al. (2013).

1. From the red triangle menu next to Response y, select **Estimates > Multiple Comparisons**.
2. In the Multiple Comparisons window, select **User-Defined Estimates**.
3. Select all three values below **Choose Drug Levels**.
4. In the first box below x, enter 12.5.
5. Click **Add Estimates**.

   This adds comparisons of the three levels of Drug at x = 12.5.
6. Click **OK**.

   The User-Defined Estimates report shows least square means estimates for each level of Drug with the covariate x set to 12.5. The red triangle menu next to Multiple Comparisons for User-Defined Estimates contains options that enable you to test for differences among the estimates.

Figure 4.7 User-Defined Estimates Report

Multiple Comparisons for User-Defined Estimates

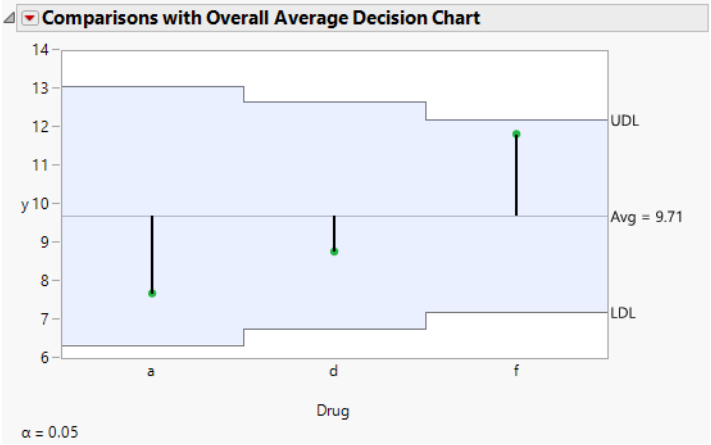
User-Defined Estimates

x = 12.5

Drug	Estimate	Std Error	DF	Lower 95%	Upper 95%
a	7.684713	1.5772057	24	4.4295208	10.939906
d	8.771371	1.4401229	24	5.7991033	11.743639
f	11.822214	1.2943345	24	9.1508392	14.493590

7. From the red triangle menu next to Multiple Comparisons for User-Defined Estimates, select **Comparisons with Overall Average**.

Figure 4.8 Comparisons with Overall Average Decision Chart



The Comparisons with Overall Average option creates an analysis of means (ANOM) chart for differences between the average and the three least squares means. From the ANOM chart, you conclude that there is not a significant effect of Drug on the response at x = 12.5.

Response Surface Model Example

This example fits a response surface model. Your objective is to minimize the response.

- "Fit the Full Response Surface Model"
- "Reduce the Model"
- "Examine the Response Surface Report"
- "Find the Critical Point Using the Prediction Profiler"
- "View the Surface Using the Contour Profiler"

Fit the Full Response Surface Model

1. Select **Help > Sample Data Library** and open Design Experiment/Custom RSM.jmp.
2. Select **Analyze > Fit Model**.

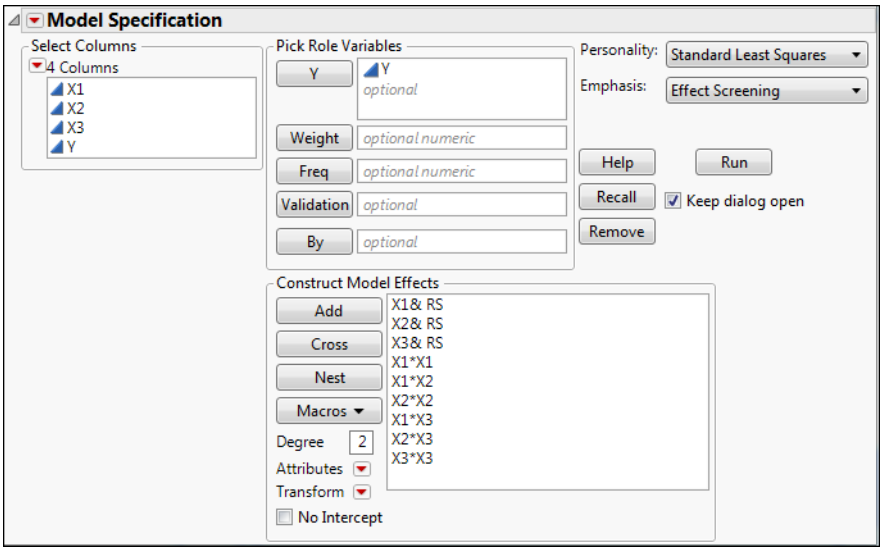
   Because the data table contains a **Model** script, the Model Specification window is filled out as specified in the **Model** script. Note the following:

   – Main effects appear in the **Construct Model Effects** list with a **&RS** suffix, indicating that the Response Surface macro has been applied.

   – The effects are those for a full response surface in the three predictors X1, X2, and X3.

   – Because the model contains terms with the **&RS** suffix, the analysis results will include a Response Surface report.

Figure 4.9 Fit Model Launch Window for the Response Surface Analysis

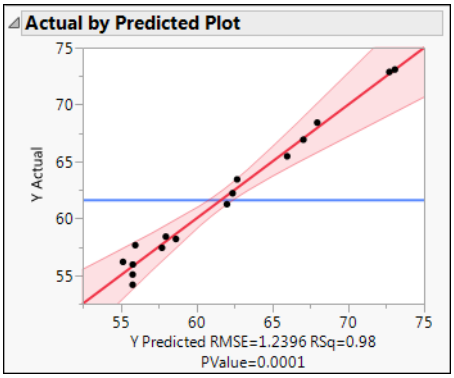


3. Click **Run**.

Reduce the Model

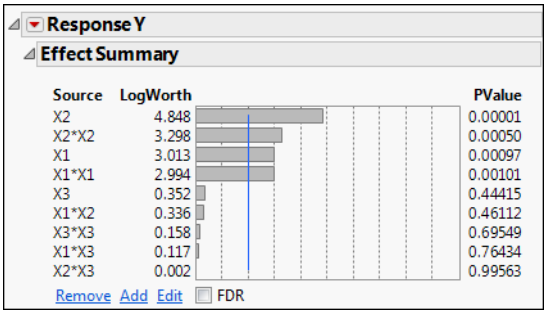
The Actual by Predicted Plot shows that the model is significant. There is no evidence of lack of fit.

Figure 4.10 Actual by Predicted Plot for Full Model



The Effect Summary report suggests that a number of effects are not significant. In particular, X2*X3 is the least significant effect with a PValue of 0.99563. Next, you will systematically reduce the model using the Effect Summary report interactively.

Figure 4.11 Effect Summary Report



1. In the Effect Summary report, click on X2*X3 and click **Remove**.

   The model updates.

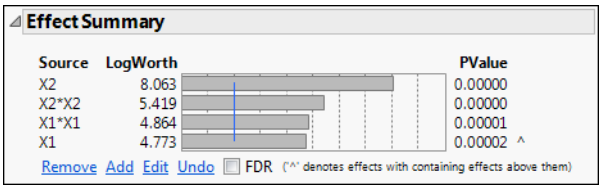
   The PValue column in the Effect Summary report indicates that X1*X3 is not significant.
2. Click on X1*X3 and click **Remove**.

   The PValue for X3*X3 indicates that it is not significant.
3. Click on X3*X3 and click **Remove**.
4. Click on X1*X2 and click **Remove**.

Notice that X3 is not significant. It is not contained in any higher-order effects, so you can remove it without violating the Effect Heredity principle. See "Effect Heredity" on page 185 in the "Standard Least Squares Report and Options" chapter.

5. Click on X3 and click **Remove**.

Figure 4.12 Effect Summary Report after Reducing Model

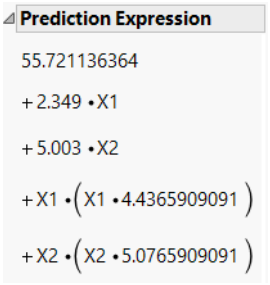


All remaining effects are significant.

Examine the Response Surface Report

1. Click the Response Y red triangle and select **Estimates > Show Prediction Expression**.

Figure 4.13 Prediction Expression

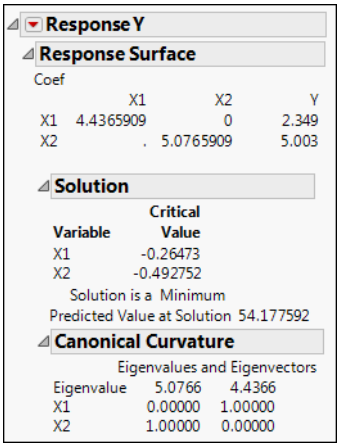


Tip: The Prediction Expression is the prediction formula. You can also obtain this formula by selecting **Save Columns > Prediction Formula**.

In the following steps, you can refer to the prediction expression to see the model coefficients.

2. Open the Response Surface report and then the Canonical Curvature report.

Figure 4.14 Response Surface Report



Response Y

Response Surface

Coef

	X1	X2	Y
X1	4.4365909	0	2.349
X2	.	5.0765909	5.003

Solution

Variable	Critical Value
X1	-0.26473
X2	-0.492752

Solution is a Minimum

Predicted Value at Solution 54.177592

Canonical Curvature

Eigenvalues and Eigenvectors

Eigenvalue	5.0766	4.4366
X1	0.00000	1.00000
X2	1.00000	0.00000

The first table gives the second-order model coefficients in matrix form. The coefficient of X1*X1 is 4.4365909, the coefficient of X2*X2 is 5.0765909, and the coefficient of X1*X2 is 0. The coefficients of the linear effects, 2.349 for X1 and 5.003 for X2, are given in the column labeled Y.

The Solution report shows the critical values. These are the values where a maximum, a minimum, or a saddle point occur. In this example, the Solution report indicates that the response surface achieves a minimum of 54.18 at the critical value, where X1 = -0.265 and X2 = -0.493.

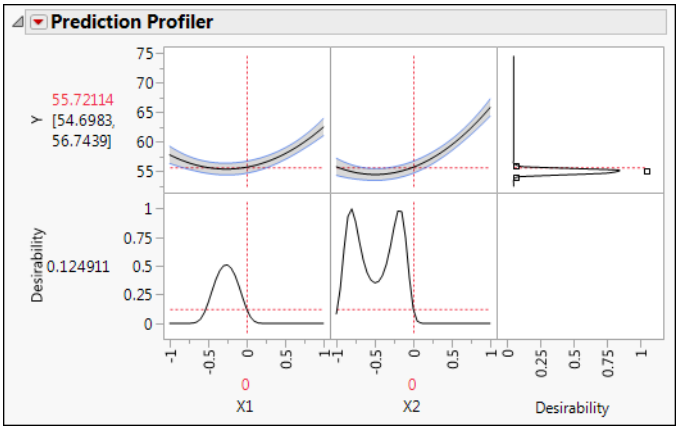
The Canonical Curvature report shows the eigenstructure of the matrix of second-order parameter estimates. The eigenstructure is useful for identifying the shape and orientation of the curvature. See "Canonical Curvature Report" on page 88 in the "Standard Least Squares Report and Options" chapter.

In this example, both eigenvalues are positive, which indicates that the surface achieves a minimum. The direction of greatest curvature corresponds to the largest eigenvalue (5.0766). That direction is defined by the corresponding eigenvector components. For the first direction, X2, with an eigenvector value of 1.00, determines the direction. The second direction is determined by X1, also with an eigenvector value of 1.00.

Find the Critical Point Using the Prediction Profiler

The Prediction Profiler report shows the quadratic behavior of the response surface along traces for X1 and X2. Because the Response Limits column property is set for Y, the profiler also shows desirability functions.

Figure 4.15 Prediction Profiler with Match Target as Goal

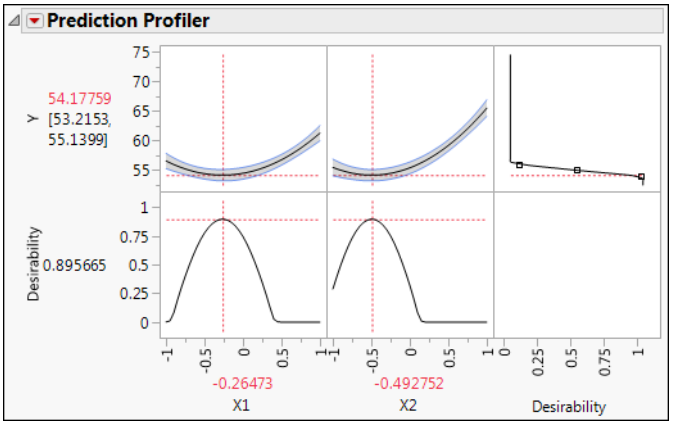


The goal for the Response Limits column property is set to Match Target. But for this example, you are interested in minimizing Y, not matching a target. Change this setting as follows:

1. Press Ctrl and click in the top right cell of the Prediction Profiler.
2. In the Response Goal dialog, select **Minimize** from the list of options.
3. Click **OK**.

   The desirability function now reflects your goal of minimizing Y.
4. Click the Prediction Profiler red triangle and select **Optimization and Desirability > Maximize Desirability**.

Figure 4.16 Prediction Profiler with Minimize as Goal and Desirability Maximized



Settings within the design region that minimize Y appear under the profiler. Note that these are precisely the Critical Values given in the Solution report.

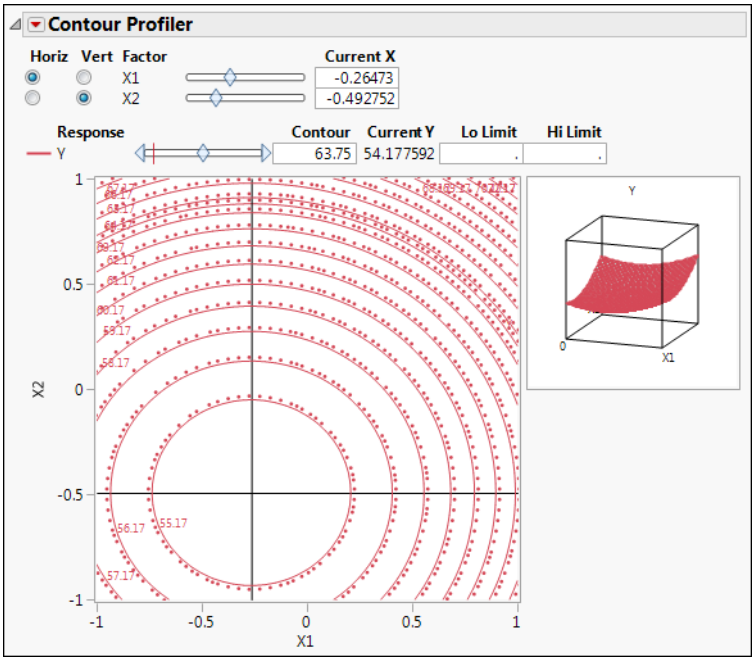
View the Surface Using the Contour Profiler

The Contour Profiler shows contours of the response surface. It gives an alternate profiler visualization of the predicted response in the area of the critical point.

1. Click the Response Y red triangle and select **Factor Profiling > Contour Profiler**.
2. Click the Contour Profiler red triangle menu and select **Contour Grid**.
3. For **Increment**, type 1.
4. Click **OK**.

   The contours are plotted at one unit intervals. See Figure 4.17.
5. Click the Prediction Profiler red triangle menu and select **Factor Settings > Link Profilers**.

Figure 4.17 Contour Profiler with Crosshairs at Critical Point



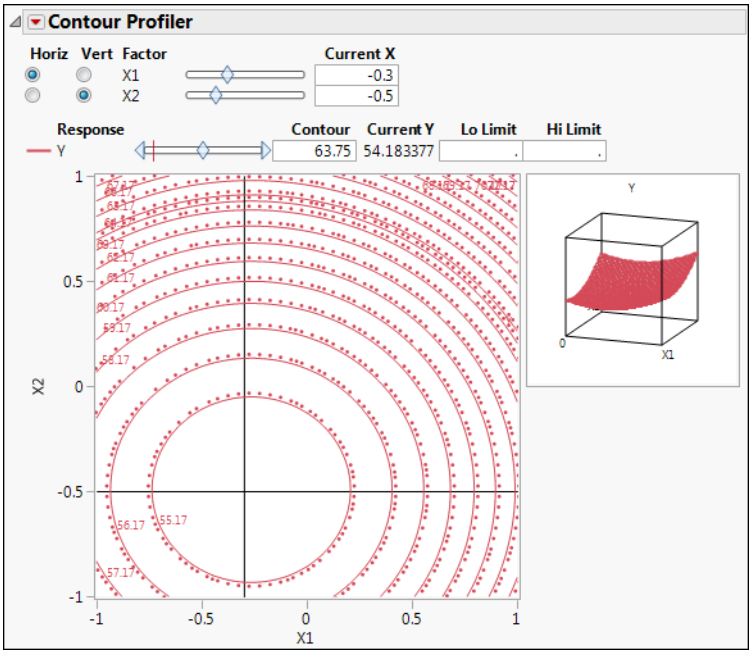
Linking the Contour Profiler to the Prediction Profiler links the **Current X** values in the Contour Profiler to the X values shown in the Prediction Profiler. The X values in the Prediction Profiler give the critical point where Y is minimized. The crosshairs in the Contour Profiler show the critical point. Notice that the **Current Y** value is 54.177592, the predicted minimum value according to the Prediction Profiler.

Often, it is not possible to set your process factors to exactly the values that optimize the response. The Contour Profiler can help you identify alternate settings of the process factors. In the next steps, suppose that you can only set your process to X1 and X2 values

with one decimal place precision, and that your process settings may vary by one decimal place in either direction of those settings.

6. In the Contour Profiler report, under **Current X**, type -0.3 next to X1 and -0.5 next to X2.

Figure 4.18 Contour Profiler Showing X1 = -0.3 and X2 = -0.5



The crosshairs are well within the innermost contour and the Current Y (the predict Y value at the Current X settings) is 54.183377, only slightly different from the predicted minimum of 54.177592.

7. In the Contour Profiler, click and drag the crosshairs to explore Current X values within a 0.1 unit radius of X1 = -0.3 and X2 = -0.5.

   The predicted Y values are all below 54.4. In fact, if the settings wander to any point within the innermost contour, the predicted Y is less than the contour value of 55.17.

Split Plot Design Example

Levels of random effects are randomly selected from a larger population of levels. For the purpose of inference, the distribution of a random effect is assumed to be normal, with mean zero and some variance (called a *variance component*).

In a sense, every model has at least one random effect, which is the effect that makes up the residual error. The individual observations are assumed to be randomly selected from a much larger population, and the error term is assumed to have a mean of zero and variance σ^2.

The most common random effects model is the repeated measures or split plot model. Table 4.2 lists the types of effects in a split plot model. In these models, the experiment has two layers. Some effects are applied on the whole plots or subjects of the experiment. Then these plots are divided or the subjects are measured at different times and other effects are applied within those subunits. The effects describing the whole plots or subjects are whole plot effects, and the subplots or repeated measures are subplot effects. Usually the subunit effect is omitted from the model and absorbed as residual error.

Table 4.2 Types of Effects in a Split Plot Model

Split Plot Model	Type of Effect	Repeated Measures Model
whole plot treatment	fixed effect	across subjects treatment
whole plot ID	random effect	subject ID
subplot treatment	fixed effect	within subject treatment
subplot ID	random effect	repeated measures ID

Each of these cases can be treated as a layered model, and there are several traditional ways to fit them in a fair way. The situation is treated as two different experiments:

1. The whole plot experiment has whole plot or subjects as the experimental unit to form its error term.
2. Subplot treatment has individual measurements for the experimental units to form its error term (left as residual error).

The older, traditional way to test whole plots is to do any one of the following:

- Take means across the measurements and fit these means to the whole plot effects.
- Form an *F*-ratio by dividing the whole plot mean squares by the whole plot ID mean squares.
- Organize the data so that the split or repeated measures form different columns. Fit a MANOVA model, and use the univariate statistics.

These approaches work if the structure is simple and the data are complete and balanced. However, there is a more general model that works for any structure of random effects. This more generalized model is called the *mixed model*, because it has both fixed and random effects.

The most common type of layered design is a balanced split plot, often in the form of repeated measures across time. One experimental unit for some of the effects is subdivided (sometimes by time period) and other effects are applied to these subunits.

Consider the data in the Animals.jmp sample data table (the data are fictional). The study collected information about differences in the seasonal hunting habits of foxes and coyotes. Each season for one year, three foxes and three coyotes were marked and observed periodically. The average number of miles that they wandered from their dens during different seasons of the year was recorded (rounded to the nearest mile). The model is defined by the following aspects:

- The continuous response variable called miles
- The species effect with values fox or coyote
- The season effect with values fall, winter, spring, and summer
- An animal identification code called subject, with nominal values 1, 2, and 3 for both foxes and coyotes

There are two layers to the model:

1. The top layer is the between-subject layer, in which the effect of being a fox or coyote (species effect) is tested with respect to the variation from subject to subject.
2. The bottom layer is the within-subject layer, in which the repeated-measures factor for the four seasons (season effect) is tested with respect to the variation from season to season within a subject. The within-subject variability is reflected in the residual error.

The season effect can use the residual error for the denominator of its *F*-statistics. However, the between-subject variability is not measured by residual error and must be captured with the subject within species (subject[species]) effect in the model. The *F*-statistic for the between-subject effect species uses this nested effect instead of residual error for its *F*-ratio denominator.

Note: JMP Pro users can construct this model using the Mixed Model personality.

To specify the split plot model for this data, follow these steps:

1. Select **Help > Sample Data Library** and open Animals.jmp.
2. Select **Analyze > Fit Model**.
3. Select miles and click **Y**.
4. Select species and subject and click **Add**.
5. In the Select Columns list, select species.
6. In the Construct Model Effects list, select subject.
7. Click **Nest.**

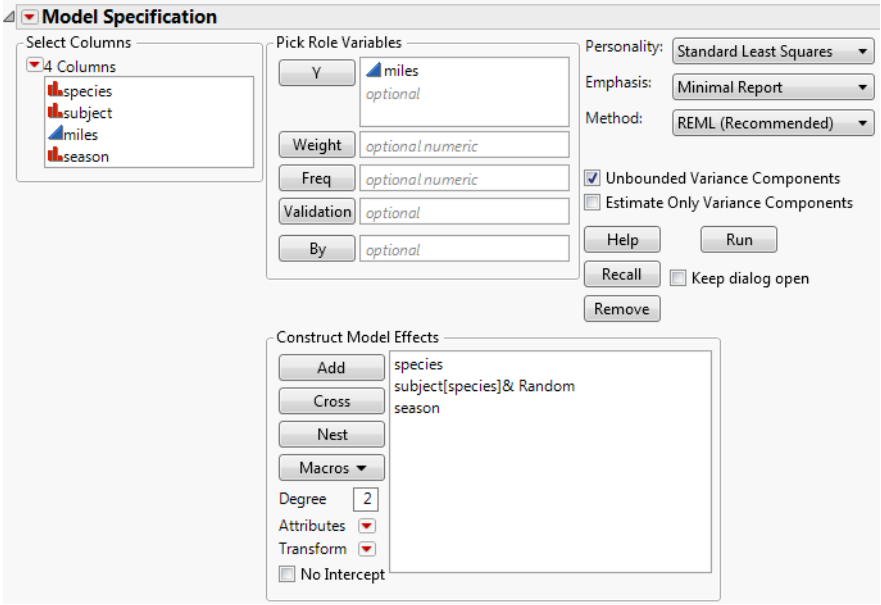
   This adds the subject within species (subject[species]) effect to the model.

8. Select the nested effect subject[species].
9. Select **Attributes > Random Effect**.

   This nested effect is now identified as an error term for the species effect and appears as subject[species]&Random.
10. In the Select Columns list, select season and click **Add**.

    When you define an effect as random using the **Attributes** menu, the Method options (**REML** and **EMS**) appear at the top right of the dialog, with **REML** selected as the default. The completed launch window is shown in Figure 4.19.

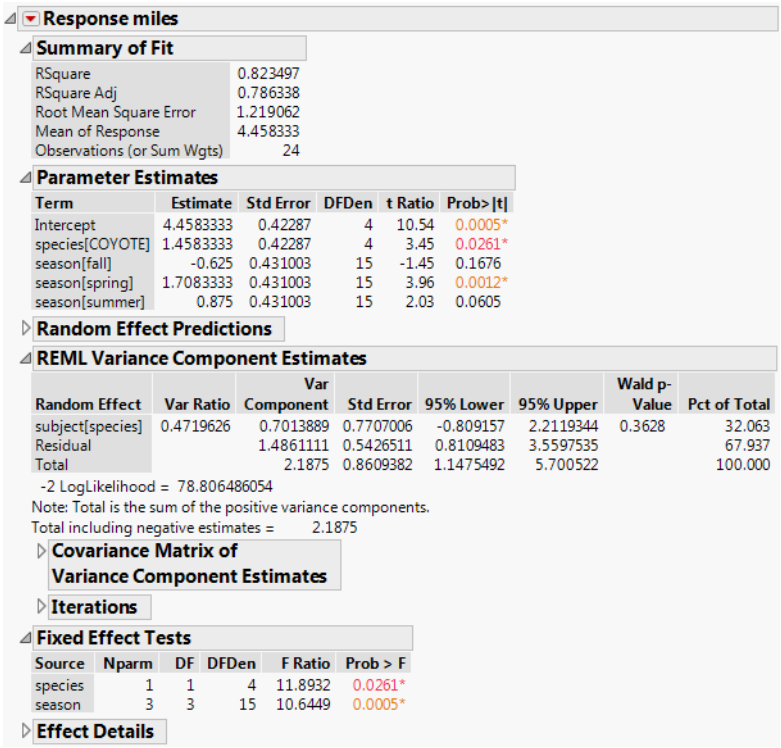
Figure 4.19 Fit Model Dialog



11. Click **Run**.

    The report is shown in Figure 4.20. Both fixed effects, species and season, are significant. The REML Variance Component Estimates report gives estimates of the subject within species and residual variances.

Figure 4.20 Partial Report of REML Analysis



Response miles

Summary of Fit

RSquare	0.823497
RSquare Adj	0.786338
Root Mean Square Error	1.219062
Mean of Response	4.458333
Observations (or Sum Wgts)	24

Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|
Intercept	4.4583333	0.42287	4	10.54	0.0005*
species[COYOTE]	1.4583333	0.42287	4	3.45	0.0261*
season[fall]	-0.625	0.431003	15	-1.45	0.1676
season[spring]	1.7083333	0.431003	15	3.96	0.0012*
season[summer]	0.875	0.431003	15	2.03	0.0605

Random Effect Predictions

REML Variance Component Estimates

Random Effect	Var Ratio	Var Component	Std Error	95% Lower	95% Upper	Wald p-Value	Pct of Total
subject[species]	0.4719626	0.7013889	0.7707006	-0.809157	2.2119344	0.3628	32.063
Residual		1.4861111	0.5426511	0.8109483	3.5597535		67.937
Total		2.1875	0.8609382	1.1475492	5.700522		100.000

-2 LogLikelihood = 78.806486054
Note: Total is the sum of the positive variance components.
Total including negative estimates = 2.1875

Covariance Matrix of Variance Component Estimates

Iterations

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
species	1	1	4	11.8932	0.0261*
season	3	3	15	10.6449	0.0005*

Effect Details

Estimation of Random Effect Parameters Example

Random effects have a dual character. In one characterization, they represent residual error, such as the error associated with a whole-plot experimental unit. In another characterization, they are like fixed effects, associating a parameter to each level of the random effect. As parameters, you have extra information about them—they are derived from a normal distribution with mean zero and the variance estimated by the variance component. The effect of this extra information is that the estimates of the parameters are shrunken toward zero.

The parameter estimates associated with random effects are called *BLUPs* (Best Linear Unbiased Predictors). Some researchers consider these BLUPs as parameters of interest, and others consider them uninteresting by-products of the methodology.

BLUP parameter estimates are used to estimate random-effect least squares means, which are therefore also shrunken toward the grand mean. The degree of shrinkage depends on the variance of the effect and the number of observations per level in the effect. With large variance estimates, there is little shrinkage. If the variance component is small, then more shrinkage takes place. If the variance component is zero, the effect levels are shrunk to exactly

zero. It is even possible to obtain highly negative variance components where the shrinkage is reversed. You can consider fixed effects as a special case of random effects where the variance component is very large.

The REML method balances the information about each individual level with the information about the variances across levels. If the number of observations per level is large, the estimates shrink less. If there are very few observations per level, the estimates shrink more. If there are infinitely many observations, there is no shrinkage and the estimates are identical to fixed effects.

Suppose that you have batting averages for different baseball players. The variance component for the batting performance across players describes how much variation is typical between players in their batting averages. Suppose that the player only plays a few times and that the batting average is unusually small or large. Then you tend not to trust that estimate, because it is based on only a few at-bats. But if you mix that estimate with the grand mean, that is, shrink the estimate toward the grand mean, you would trust the estimate more. For players who have a long batting record, you would shrink much less than those with a short record.

You can explore this behavior yourself.

1. Select **Help > Sample Data Library** and open Baseball.jmp.
2. Select **Analyze > Fit Model**.
3. Select Batting and click **Y, Response**.
4. Select Player and click **Add**.
5. Select Player in the Construct Model Effects box, and select **Random Effect** from the Attributes list.
6. Click **Run**.

Table 4.3 shows the Least Squares Means from the Player report for a REML (Recommended) fit. Also shown are the Method of Moment estimates, obtained using the EMS Method. The Method of Moment estimates are the ordinary Player means. Note that the REML estimate for Suarez, who has only three at-bats, is shrunken more towards the grand mean than estimates for other players with more at-bats.

Table 4.3 Comparison of Estimates between Method of Moments and REML

	Method of Moments	REML	N
Variance Component	0.01765	0.019648	
Anderson	0.29500000	0.29640407	6
Jones	0.20227273	0.20389793	11
Mitchell	0.32333333	0.32426295	6
Rodriguez	0.55000000	0.54713393	6
Smith	0.35681818	0.35702094	11
Suarez	0.55000000	0.54436227	3
Least Squares Means	same as ordinary means	shrunken from means	

Knotted Spline Effect Example

Use the **Knotted Spline Effect** option to have JMP fit a segmentation of smooth polynomials to a specified effect. When you select this attribute, a window appears, enabling you to specify the number of knot points. (Knotted splines are only implemented for main-effect continuous terms.)

JMP follows the advice in the literature in positioning the points. The knotted spline is also referred to as a Stone spline or a Stone-Koo spline. See Stone and Koo (1985). If there are 100 or fewer points, the first and last knots are the fifth point inside the minimum and maximum, respectively. Otherwise, the first and last knots are placed at the 0.05 and 0.95 quantiles if there are 5 or fewer knots, or the 0.025 and 0.975 quantiles for more than 5 knots. The default number of knots is 5 unless there are 30 or fewer points. In that case, the default is 3 knots.

Knotted splines have the following properties in contrast to smoothing splines:

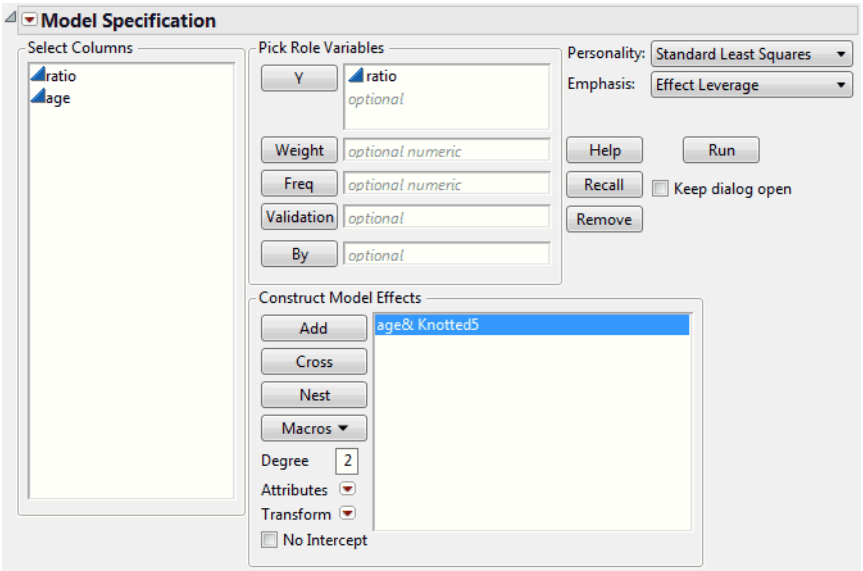
- Knotted splines work inside of general models with many terms, whereas smoothing splines are for bivariate regressions.
- The regression basis is not a function of the response.
- Knotted splines are parsimonious, adding only $k - 2$ terms for curvature for k knot points.
- Knotted splines are conservative compared to pure polynomials in the sense that the extrapolation outside the range of the data is a straight line, rather than a polynomial.
- There is an easy test for curvature.

Example Using the Knotted Spline Effect to Test for Curvature

To test for curvature, follow these steps:

1. Select **Help > Sample Data Library** and open Growth.jmp.
2. Select **Analyze > Fit Model**.
3. Select the ratio column and click **Y**.
4. Select the age column and click **Add**.
5. Select age in the **Effects** pane and select **Attributes > Knotted Spline Effect**.
6. For the number of knots, type 5 and click **OK**.

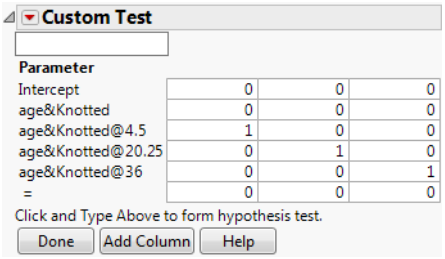
Figure 4.21 Fit Model Launch Window



7. Click **Run**.
8. From the report's red triangle menu, select **Estimates > Custom Test**.

   In the Custom Test report, notice that there is only one column. You want three columns.
9. Click the **Add Column** button twice to produce a total of three columns.
10. In the first column, type 1 for **age&Knotted@4.5**.
11. In the second column, type 1 for **age&Knotted@20.5**.
12. In the third column, type 1 for **age&Knotted@36**.

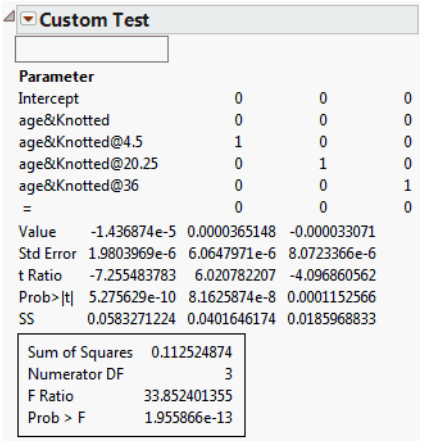
Figure 4.22 Values for the Custom Test for Curvature



13. Click **Done**.

    The report in Figure 4.23 appears. The small Prob > F value indicates that there is curvature.

Figure 4.23 Curvature Report



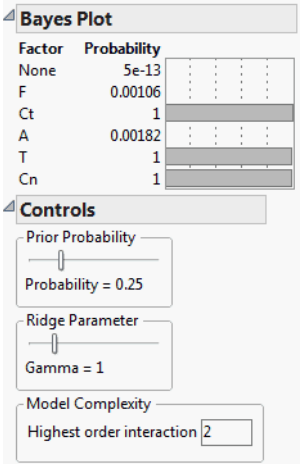
Bayes Plot for Active Factors Example

Suppose that you conduct an experimental design and want to determine which factors are active. You can address this in several ways using JMP. This example illustrates a script that can be used to identify active factors using a Bayesian approach.

1. Select **Help > Sample Data Library** and open Reactor.jmp.
2. In the Samples > Scripts folder, open the BayesPlotforFactors.jsl sample script.
3. Select **Edit > Run Script**.
4. Select Y and click **Y, Response**.
5. Select F, Ct, A, T, and Cn and click **X, Factor**.

6. Click **OK**.

Figure 4.24 Bayes Plot for Factor Activity



The Model Complexity indicates that the highest order interaction to consider is two. Therefore, all possible models that include up to second-order interactions are constructed. Based on the value assigned to Prior Probability, a posterior probability is computed for each of the possible models. The probability for a factor is the sum of the probabilities for each of the models where it was involved.

This approach identifies Ct, T, and Cn as active factors, and A and F as inactive.

If the ridge parameter were zero (not allowed), all the models would be fit by least squares. As the ridge parameter increases, the parameter estimates for any model shrink toward zero. Details on the ridge parameter, and why it cannot be zero, are given in Box and Meyer (1993).

Chapter 5

Stepwise Regression Models

Find a Model Using Variable Selection

Stepwise regression is an approach to selecting a subset of effects for a regression model. It can be useful in the following situations:

- There is little theory to guide the selection of terms for a model.
- You want to interactively explore which predictors seem to provide a good fit.
- You want to improve a model's prediction performance by reducing the variance caused by estimating unnecessary terms.

For categorical predictors, you can do the following:

- Choose from among various rules to determine how associated terms enter the model.
- Enforce effect heredity.

The Stepwise platform also enables you to explore all possible models and to conduct model averaging.

Contents

Overview of Stepwise Regression

In JMP, stepwise regression is a personality of the Fit Model platform. The **Stepwise** feature computes estimates that are the same as those of other least squares platforms, but it facilitates searching and selecting among many models.

The approach has side effects of which you need to be aware. The significance levels on the statistics for selected models violate the standard statistical assumptions because the model has been selected rather than tested within a fixed model. On the positive side, the approach has been helpful for 30 years in reducing the number of terms. The book *Subset Selection in Regression* (Miller 1990) brings statistical sense to model selection statistics.

This chapter uses the term *significance probability* in a mechanical way to represent that the calculation would be valid in a fixed model, recognizing that the true significance probability could be nowhere near the reported one.

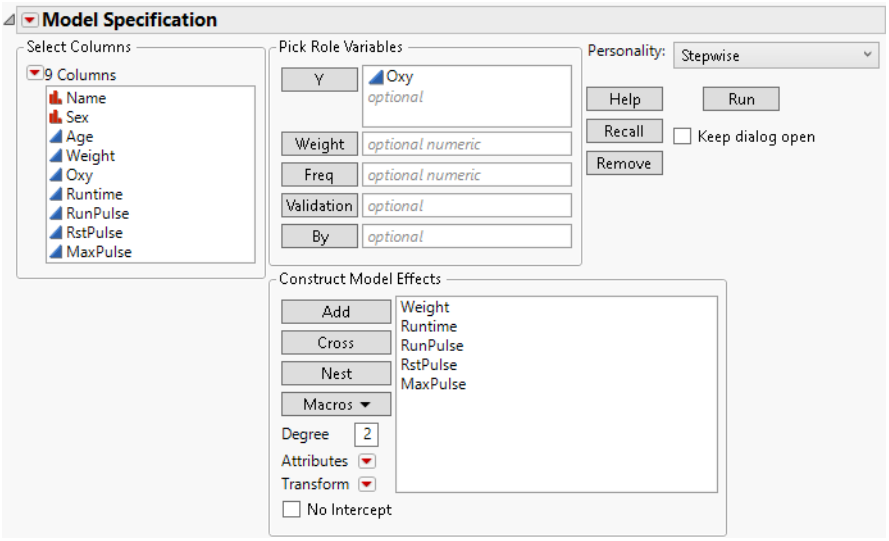
Example Using Stepwise Regression

The Fitness.jmp data table contains the results of an aerobic fitness study. Aerobic fitness can be evaluated using a special test that measures the oxygen uptake of a person running on a treadmill for a prescribed distance. However, it would be more economical to find a formula that uses simpler measurements that evaluate fitness and predict oxygen uptake. To identify such an equation, measurements of age, weight, run time, and pulse were taken for 31 participants who ran 1.5 miles.

Note: For purposes of illustration, certain values of MaxPulse and RunPulse have been changed from data reported by Rawlings (1988, p.105).

1. Select **Help > Sample Data Library** and open Fitness.jmp.
2. Select **Analyze > Fit Model**.
3. Select Oxy and click **Y**.
4. Select Weight, Runtime, RunPulse, RstPulse, MaxPulse, and click **Add**.
5. For **Personality**, select **Stepwise**.

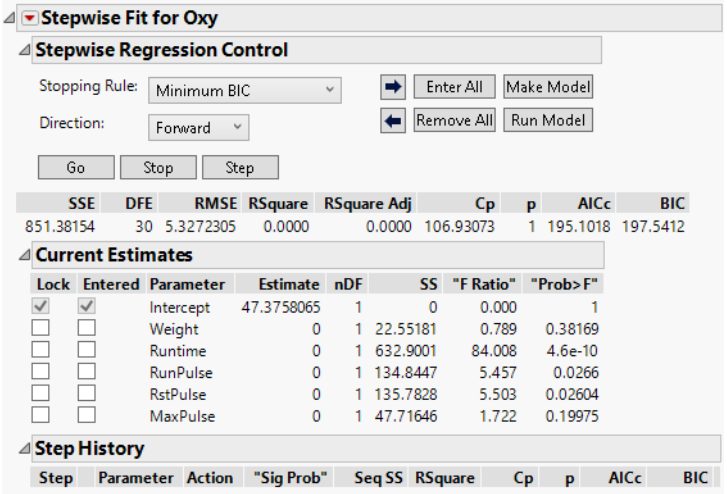
Figure 5.1 Completed Fit Model Launch Window



JMP PRO **Validation** is available only in JMP Pro.

6. Click **Run**.

Figure 5.2 Stepwise Report Window



To find a good oxygen uptake prediction equation, you need to compare many different regression models. Use the options in the Stepwise report window to search through models with combinations of effects and choose the model that you want.

The Stepwise Report

The Stepwise report window contains the following elements:

Platform options The red triangle menu next to Stepwise Fit contains options that affect all of the variables. See "Stepwise Platform Options" on page 253.

Stepwise Regression Control Controls that limit regressor effect probabilities, determine the method of selecting effects, start or stop the selection process, and create a model. See "Stepwise Regression Control Panel" on page 254.

Current Estimates A report that enables you to add, remove, and lock in model effects. See "Current Estimates Report" on page 261.

Step History A report that records the effect of adding a term to the model. See "Step History Report" on page 262.

Stepwise Platform Options

The red triangle menu next to Stepwise Fit contains the following platform options:

K-Fold Crossvalidation Performs K-Fold cross validation in the selection process. When selected, this option enables the Max K-Fold RSquare stopping rule ("Stepwise Regression Control Panel" on page 254).

Available only for continuous responses. For more information about validation, see "Using Validation" on page 279.

All Possible Models Fits all possible models up to specified limits and shows the best models for each number of terms. Enter values for the maximum number of terms to fit in any one model. Also enter values for the maximum number of best model results to show for each number of terms in the model. Categorical variables are represented using indicator variables. See "Models with Nominal and Ordinal Effects" on page 265. You can restrict the models that appear to those that satisfy strong effect heredity. See "The All Possible Models Option" on page 276.

This option is available only for continuous responses.

Model Averaging (Available only for continuous responses.) Enables you to average the fits for a number of models, instead of picking a single best model. See "The Model Averaging Option" on page 278.

Plot Criterion History Creates a plot of AICc and BIC versus the number of parameters. The Criterion History plot contains two shaded zones. Define the minimum AICc value as V^{best}. The green zone is defined by the range [V^{best}, V^{best}+4]. The yellow zone is defined by the range (V^{best}+4, V^{best}+10].

Plot RSquare History (Available only for continuous responses.) Creates a plot of training and validation R-square versus the number of parameters.

Clear History Clears and resets the step history.

Export Model with Validation (Available only when you have entered a Validation column in the Stepwise launch window.) Adds the Validation column to the Model Specification window when you select Make Model. Runs the model with the Validation column when you select Run Model.

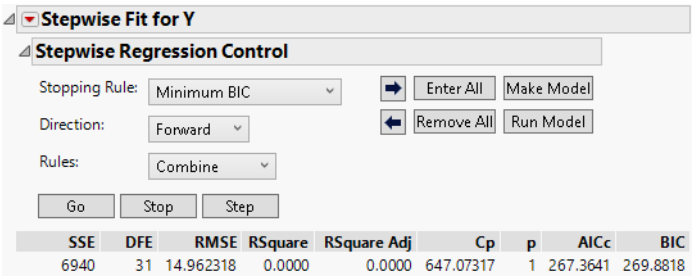
This option is selected by default.

Model Dialog Shows the completed Fit Model launch window for the current model.

Stepwise Regression Control Panel

Use the Stepwise Regression Control panel to limit regressor effect probabilities, determine the method of selecting effects, begin or stop the selection process, and run a model. A note appears beneath the Go button to indicate if you have excluded or missing rows.

Figure 5.3 Stepwise Regression Control Panel



Stopping Rule

The Stopping Rule determines which model is selected. For all stopping rules other than P-value Threshold, only the Forward and Backward directions are allowed. The only stopping rules that use validation are Max Validation RSquare and Max K-Fold RSquare. See "Using Validation" on page 279.

P-value Threshold Uses p-values (significance levels) to enter and remove effects from the model. Two other options appear when you choose P-value Threshold:

- **Prob to Enter** is the maximum p-value that an effect must have to be entered into the model during a forward step.
- **Prob to Leave** is the minimum p-value that an effect must have to be removed from the model during a backward step.

Minimum AICc Uses the minimum corrected Akaike Information Criterion to choose the best model. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

Minimum BIC Uses the minimum Bayesian Information Criterion to choose the best model. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

JMP PRO **Max Validation RSquare** Uses the maximum R-square from the validation set to choose the best model. This is available only when you use a validation column with two or three distinct values. For more information about validation, see "Validation Set with Two or Three Values" on page 279.

Max K-Fold RSquare Uses the maximum RSquare from K-fold cross validation to choose the best model. You can access the Max K-Fold RSquare stopping rule by selecting this option from the Stepwise red triangle menu. JMP Pro users can access the option by using a validation set with four or more values. When you select this option, you are asked to specify the number of folds. For more information about validation, see "K-Fold Cross Validation" on page 283.

Direction

The Direction you choose controls how effects enter and leave the model. Select one of the following options:

Forward Enters the term with the smallest *p*-value. If the P-value Threshold stopping rule is selected, that term must be significant at the level specified by **Prob to Enter**. See "Forward Selection Example" on page 259.

Backward Removes the term with the largest *p*-value. If the P-value Threshold stopping rule is selected, that term must not be significant at the level specified in **Prob to Leave**. See "Backward Selection Example" on page 260.

Note: When Backward is selected as the Direction, you must click Enter All before clicking Go or Step.

Mixed Available only when the P-value Stopping Rule is selected. It alternates the forward and backward steps. It includes the most significant term that satisfies **Prob to Enter** and

removes the least significant term satisfying **Prob to Leave**. It continues removing terms until the remaining terms are significant and then it changes to the forward direction.

Go, Stop, Step Buttons

The Go, Stop, and Step buttons enable you to control how terms are entered or removed from the model.

Note: All Stopping Rules only consider models defined by *p*-value entry (Forward direction) or removal (Backward direction). Stopping rules do not consider all possible models.

Go Automates the process of entering (Forward direction) or removing (Backward direction) terms. Among the fitted models, the model that is considered best based on the selected Stopping Rule is listed last. Except for the P-value Threshold stopping rule, the model selected as Best is one that overlooks local dips in the behavior of the stopping rule statistic. The button to the right the Best model selects it for the Make Model and Run Model options, but you are free to change this selection.

- For P-value Threshold, the best model is based on the Prob to Enter and Prob to Leave criteria. See "P-value Threshold" on page 254.
- For Min AICc and Min BIC, the automatic fits continue until a Best model is found. The Best model is one with a minimum AICc or BIC that can be followed by as many as ten models with larger values of AICc or BIC, respectively. This model is designated by the terms Best in the Parameter column and Specific in the Action column.
- For Max Validation RSquare (JMP Pro only) and Max K-Fold RSquare, the automatic fits continue until a Best model is found. The Best model is one with an RSquare Validation or RSquare K-Fold value that can be followed by as many as ten models with smaller values of RSquare Validation or RSquare K-Fold, respectively. This model is designated by the terms Best in the Parameter column and Specific in the Action column.

Stop Stops the automatic selection process started by the Go button.

Step Enters terms one-by-one in the Forward direction or removes them one-by one in the Backward direction. At any point, you can select a model by clicking its button on the right in the Step History report. The selection of model terms is updated in the Current Estimates report. This is the model that is used once you click Make Model or Run Model.

Rules

Note: Appears only if your model contains related terms. When you have a nominal or ordinal variable, related terms are constructed and appear in the Current Estimates table.

Use Rules to change the rules that are applied when there is a hierarchy of terms in the model. A hierarchy can occur in the following ways:

- A hierarchy results when a variable is a component of another variable. For example, if your model contains variables A, B, and A*B, then A and B are *precedent* terms to A*B in the hierarchy.
- A hierarchy also results when you include nominal or ordinal variables. A term that is above another term in the tree structure is a *precedent* term. See "Construction of Hierarchical Terms" on page 266.

Select one of the following options:

Combine Calculates *p*-values for two separate tests when considering entry for a term that has precedents. The first *p*-value, p_1, is calculated by grouping the term with its precedent terms and testing the group's significance probability for entry as a joint *F* test. The second *p*-value, p_2, is the result of testing the term's significance probability for entry after the precedent terms have already entered into the model. The final significance probability for entry for the term that has precedents is $\max(p_1, p_2)$.

Tip: The Combine rule avoids including non-significant interaction terms, whose precedent terms may have particularly strong effects. In this scenario, the strong main effects may make the group's significance probability for entry, p_1, very small. However, the second test finds that the interaction by itself is not significant. As a result, p_2 is large and is used as the final significance probability for entry.

Caution: The degrees of freedom value for a term that has precedents depends on which of the two significance probabilities for entry is larger. The test used for the final significance probability for entry determines the degrees of freedom, nDF, in the Current Estimates table. Therefore, if p_1 is used, nDF will be the number of terms in the group for the joint test, and if p_2 is used, nDF will be equal to 1.

The Combine option is the default rule. See "Models with Crossed, Interaction, or Polynomial Terms" on page 263.

Restrict Restricts the terms that have precedents so that they cannot be entered until their precedents are entered. See "Models with Nominal and Ordinal Effects" on page 265 and "Example of the Restrict Rule for Hierarchical Terms" on page 271.

No Rules Gives the selection routine complete freedom to choose terms, regardless of whether the routine breaks a hierarchy or not.

Whole Effects Enters only whole effects, when terms involving that effect are significant. This rule applies only when categorical variables with more than two levels are entered as possible model effects. See "Rules" on page 271.

Buttons

The Stepwise Control Panel contains the following buttons:

Go Automates the selection process to completion.

Stop Stops the selection process.

Step Increments the selection process one step at a time.

Arrow buttons Step forward and backward one step in the selection process.

Enter All Enters all unlocked terms into the model.

Remove All Removes all unlocked terms from the model.

Make Model Creates a model for the Fit Model window from the model currently showing in the Current Estimates table. In cases where there are nominal or ordinal terms, **Make Model** creates temporary transform columns that contain terms that are needed for the model.

Run Model Runs the model currently showing in the Current Estimates table. In cases where there are nominal or ordinal terms, **Run Model** creates temporary transform columns that contain terms that are needed for the model.

Statistics

The following statistics appear below the Stepwise Regression Control panel.

SSE Sum of squared errors for the current model.

DFE Error degrees of freedom for the current model.

RMSE Root mean square error (residual) for the current model.

RSquare Proportion of the variation in the response that can be attributed to terms in the model rather than to random error.

RSquare Adj Adjusts R^2 to make it more comparable over models with different numbers of parameters by using the degrees of freedom in its computation. The adjusted R^2 is useful in stepwise procedure because you are looking at many different models and want to adjust for the number of terms in the model.

Cp Mallow's C_p criterion for selecting a model. It is an alternative measure of total squared error and can be defined as follows:

$$C_p = \left(\frac{\text{SSE}_p}{s^2}\right) - (N - 2p)$$

where s^2 is the MSE for the full model and SSE_p is the sum-of-squares error for a model with p variables, including the intercept. Note that p is the number of x-variables+1. If C_p is graphed with p, Mallows (1973) recommends choosing the model where C_p first approaches p.

p Number of parameters in the model, including the intercept.

AICc Corrected Akaike's Information Criterion. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

BIC Bayesian Information Criterion. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

Forward Selection Example

In forward selection, terms are entered into the model and most significant terms are added until all of the terms are significant.

1. Complete the steps in "Example Using Stepwise Regression" on page 251.

   Notice that the default selection for **Direction** is Forward.

2. Click **Step**.

   In Figure 5.4, you can see that after one step, the most significant term, Runtime, is entered into the model.

3. Click **Go**.

   In Figure 5.5 you can see that all of the terms have been added, except RstPulse and Weight.

Figure 5.4 Current Estimates Table for Forward Selection After One Step

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	82.4217727	1	0	0.000	1
☐	☐	Weight	0	1	1.323628	0.171	0.68267
☐	☑	Runtime	-3.3105554	1	632.9001	84.008	4.6e-10
☐	☐	RunPulse	0	1	15.36208	2.118	0.15673
☐	☐	RstPulse	0	1	0.130138	0.017	0.89814
☐	☐	MaxPulse	0	1	1.567361	0.202	0.65632

Figure 5.5 Current Estimates Table for Forward Selection After Three Steps

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	80.9007896	1	0	0.000	1
☐	☐	Weight	0	1	4.989591	0.827	0.37137
☐	☑	Runtime	-2.9701867	1	443.2028	73.971	3.25e-9
☐	☑	RunPulse	-0.3751142	1	55.14175	9.203	0.00529
☐	☐	RstPulse	0	1	0.350744	0.056	0.81399
☐	☑	MaxPulse	0.35421891	1	41.34703	6.901	0.01403

Backward Selection Example

In backward selection, all terms are entered into the model and then the least significant terms are removed until all of the remaining terms are significant.

1. Complete the steps in "Example Using Stepwise Regression" on page 251.
2. Click **Enter All**.

Figure 5.6 All Effects Entered Into the Model

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	82.3936054	1	0	0.000	1
☐	☑	Weight	-0.0509071	1	4.83788	0.772	0.38784
☐	☑	Runtime	-2.9518165	1	366.3375	58.489	5.29e-8
☐	☑	RunPulse	-0.3970425	1	59.51519	9.502	0.00495
☐	☑	RstPulse	0.01239004	1	0.199033	0.032	0.85995
☐	☑	MaxPulse	0.38479281	1	45.83023	7.317	0.01212

3. For **Direction**, select Backward.
4. Click **Step** two times.

   The first backward step removes RstPulse and the second backward step removes Weight.

Figure 5.7 Current Estimates with Terms Removed and Step History Table

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	80.9007896	1	0	0.000	1
☐	☐	Weight	0	1	4.989591	0.827	0.37137
☐	☑	Runtime	-2.9701867	1	443.2028	73.971	3.25e-9
☐	☑	RunPulse	-0.3751142	1	55.14175	9.203	0.00529
☐	☐	RstPulse	0	1	0.350744	0.056	0.81399
☐	☑	MaxPulse	0.35421891	1	41.34703	6.901	0.01403

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	All	Entered	.	.	0.8161	6	6	157.051	162.22	○
2	RstPulse	Removed	0.8600	0.199033	0.8158	4.0318	5	153.721	158.825	○
3	Weight	Removed	0.3714	4.989591	0.8100	2.8284	4	151.592	156.362	◉

The Current Estimates and Step History tables shown in Figure 5.7 summarize the backward stepwise selection process. Note the BIC value of 156.362 for the third step in the Step History table. If you click Step again to remove another parameter from the model, the BIC value

increases to 159.984. For this reason, you choose the step 3 model. This is also the model that the Go button produces.

Current Estimates Report

Use the Current Estimates report to enter, remove, and lock in model effects. (The intercept is permanently locked into the model.) This report contains the following columns:

Figure 5.8 Current Estimates Table

SSE	DFE	RMSE	RSquare	RSquare Adj	Cp	p	AICc	BIC
851.38154	30	5.3272305	0.0000	0.0000	106.93073	1	195.1018	197.5412

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	47.3758065	1	0	0.000	1
☐	☐	Weight	0	1	22.55181	0.789	0.38169
☐	☐	Runtime	0	1	632.9001	84.008	4.6e-10
☐	☐	RunPulse	0	1	134.8447	5.457	0.0266
☐	☐	RstPulse	0	1	135.7828	5.503	0.02604
☐	☐	MaxPulse	0	1	47.71646	1.722	0.19975

Lock Locks a term in or out of the model. A checked term cannot be entered or removed from the model.

Entered Indicates whether a term is currently in the model. You can click a term's check box to manually bring an effect in or out of the model.

Parameter The effect names.

Estimate The current parameter estimate, which is zero if the effect is not currently in the model.

nDF The number of degrees of freedom for a term. A term has more than one degree of freedom if its entry into a model also forces other terms into the model.

Wald/Score ChiSq (Shown only when response is categorical.) The test statistic based on a score test of including the term in the model.

"Sig Prob" (Shown only when response is categorical.) The significance level associated with the Wald/Score ChiSq test statistic based on nDF degrees of freedom. The "Sig Prob" is used to determine the next term to be included in the model.

SS (Shown only when response is continuous.) The reduction in the error (residual) sum of squares (SS) if the term is entered into the model or the increase in the error SS if the term is removed from the model. If a term is restricted in some fashion, it could have a reported SS of zero.

"F Ratio" (Shown only when response is continuous.) The traditional test statistic to test that the term effect is zero. It is the square of a *t*-ratio. It is in quotation marks because it does not have an *F*-distribution for testing the term because the model was selected as it was fit.

"Prob>F" (Shown only when response is continuous.) The significance level associated with the *F* statistic. Like the "F Ratio," it is in quotation marks because it is not to be trusted as a real significance probability.

R The multiple correlation with the other effects in the model. This column appears only if you right-click in the report and select Columns > R.

Step History Report

As each step is taken, the Step History report records the effect of adding a term to the model. For example, the Step History report for the Fitness.jmp example shows the order in which the terms entered the model and shows the statistics for each model. See "Example Using Stepwise Regression" on page 251.

Use the radio buttons on the right to choose a model.

Figure 5.9 Step History Report

Step History

Step	Parameter	Action	"Sig Prob"	Seq SS	RSquare	Cp	p	AICc	BIC	
1	Runtime	Entered	0.0000	632.9001	0.7434	7.8825	2	155.397	158.81	○
2	RunPulse	Entered	0.1567	15.36208	0.7614	7.4298	3	155.787	159.984	○
3	MaxPulse	Entered	0.0140	41.34703	0.8100	2.8284	4	151.592	156.362	◉

Step History for Categorical Responses

The Step History report for models with a categorical response contains four additional columns: L-R ChiSquare, "Sig Prob", Entry ChiSquare, and Entry "Sig Prob".

The L-R ChiSquare and "Sig Prob" columns contain the full-versus-reduced likelihood ratio test statistic and *p*-value. Here, the full model is the one that contains the specified term and the reduced model does not contain the specified term.

The Entry ChiSquare and Entry "Sig Prob" columns contain the Wald/Score ChiSquare and "Sig Prob" values that were used to choose the most recent term to include in the model.

Figure 5.10 Step History for a Categorical Response Before First Step

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	Wald/Score ChiSq	"Sig Prob"
☑	☑	Intercept[F]	0.06453852	1	0	1
☐	☐	Oxy	0	1	7.275453	0.00699
☐	☐	Weight	0	1	10.13425	0.00146
☐	☐	Runtime	0	1	5.764797	0.01635
☐	☐	RunPulse	0	1	1.865024	0.17205
☐	☐	RstPulse	0	1	2.693096	0.10078
☐	☐	MaxPulse	0	1	1.467541	0.22573

Step History

Step	Parameter	Action	L-R ChiSquare	"Sig Prob"	Entry ChiSquare	Entry "Sig Prob"	RSquare	p	AICc	BIC

Figure 5.11 Step History for a Categorical Response After First Step

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	Wald/Score ChiSq	"Sig Prob"
☑	☑	Intercept[F]	16.228294	1	0	1
☐	☐	Oxy	0	1	8.750599	0.0031
☐	☑	Weight	-0.2079227	1	6.994297	0.00818
☐	☐	Runtime	0	1	6.371395	0.0116
☐	☐	RunPulse	0	1	1.428737	0.23197
☐	☐	RstPulse	0	1	3.510878	0.06097
☐	☐	MaxPulse	0	1	0.365726	0.54534

Step History

Step	Parameter	Action	L-R ChiSquare	"Sig Prob"	Entry ChiSquare	Entry "Sig Prob"	RSquare	p	AICc	BIC
1	Weight	Entered	12.16669	0.0005	10.1342	0.00146	0.2833	2	35.2047	37.6441

Note that the Entry ChiSquare and Entry "Sig Prob" values for Weight in the Step History table in Figure 5.11 match the Wald/Score ChiSq and "Sig Prob" values for Weight in Figure 5.10.

Models with Crossed, Interaction, or Polynomial Terms

Some models, especially those associated with experimental designs, involve interaction terms. For continuous factors, these are products of the columns representing the effects. For nominal and ordinal factors, interactions are defined by model terms that involve products of terms representing the categorical levels.

When there are interaction terms, you often want to impose a restriction on the model selection process so that lower-order components of higher-order effects are included in the model. This is suggested by the principle of Effect Heredity. See the Starting Out chapter in the *Design of Experiments Guide*. For example, if a two-way interaction is included in a model, its component main effects (*precedents*) should be included as well.

Example of the Combine Rule

1. Select **Help > Sample Data Library** and open Reactor.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. In the Degree box, type 2.
5. Select F, Ct, A, T, and Cn and click **Macros > Factorial to Degree**.
6. For **Personality**, select **Stepwise**.
7. Click **Run**.

Figure 5.12 Initial Current Estimates Report Using Combine Rule

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	65.5	1	0	0.000	1
☐	☐	F	0	1	15.125	0.066	0.79972
☐	☐	Ct	0	1	3042	23.412	3.67e-5
☐	☐	A	0	1	3.125	0.014	0.90823
☐	☐	T	0	1	924.5	4.611	0.03998
☐	☐	Cn	0	1	312.5	1.415	0.24363
☐	☐	F*Ct	0	1	15.125	0.066	0.79972
☐	☐	F*A	0	3	22.75	0.031	0.9926
☐	☐	F*T	0	1	6.125	0.027	0.87178
☐	☐	F*Cn	0	1	0.125	0.001	0.98161
☐	☐	Ct*A	0	1	6.125	0.027	0.87178
☐	☐	Ct*T	0	1	1404.5	7.612	0.00979
☐	☐	Ct*Cn	0	1	32	0.139	0.71193
☐	☐	A*T	0	1	36.125	0.157	0.69476
☐	☐	A*Cn	0	1	6.125	0.027	0.87178
☐	☐	T*Cn	0	1	968	4.863	0.03525

The model in Figure 5.12 contains all terms for up to two-factor interactions for the five continuous factors. The Combine, Restrict, and Whole Effects rules described in "Rules" on page 256 enable you to control entry of interaction terms.

The Combine rule determines the entry of interaction terms based on two tests. See "Combine" on page 257. You can determine which of the two tests was used for the *p*-value based on the degrees of freedom, nDF. For example, the interaction term F*A has an nDF value of 3. This means that F*A is grouped with its precedent terms F and A, and is considered for entry based on the 3 degree of freedom joint *F* test. In contrast, the interaction term Ct*T has an nDF value of 1. This means that Ct*T is considered for entry based on the 1 degree of freedom *F* test that tests the significance of Ct*T after its precedent terms, Ct and T, are already included in the model.Click **Step** once to see that Ct is entered by itself.

8. Click **Step** again to see that Ct*T is entered, along with T (Ct is already in the model).

Figure 5.13 Current Estimates Report Using Combine Rule, One Step

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	65.5	1	0	0.000	1
☐	☐	F	0	1	15.125	0.263	0.61236
☐	☑	Ct	9.75	2	4446.5	39.676	6.74e-9
☐	☐	A	0	1	3.125	0.054	0.81819
☐	☑	T	5.375	2	2329	20.781	2.93e-6
☐	☐	Cn	0	1	312.5	6.715	0.01524
☐	☐	F*Ct	0	2	30.25	0.256	0.7764
☐	☐	F*A	0	3	22.75	0.123	0.9459
☐	☐	F*T	0	2	21.25	0.178	0.83755
☐	☐	F*Cn	0	1	0.125	0.002	0.96335
☐	☐	Ct*A	0	2	9.25	0.077	0.92601
☐	☑	Ct*T	6.625	1	1404.5	25.064	2.72e-5
☐	☐	Ct*Cn	0	1	32	0.562	0.45989
☐	☐	A*T	0	2	39.25	0.334	0.7194
☐	☐	A*Cn	0	1	6.125	0.106	0.74747
☐	☐	T*Cn	0	1	968	43.488	4.49e-7

When there are significant interaction terms, several terms can enter at the same step. If the **Step** button is clicked twice, Ct*T is entered along with its two contained effects Ct and T. However, a step back is not symmetric because a crossed term can be removed without removing its two component terms. Notice that Ct and T now each have 2 degrees of freedom. This is because if Stepwise removes Ct or T, it must also remove Ct*T. If you change the Direction to **Backward** and click **Step**, Ct*T is removed and the degrees of freedom for Ct and T change to 1.

Models with Nominal and Ordinal Effects

Traditionally, stepwise regression has not addressed the situation where there are categorical effects in the model. Note the following:

- When a regression model contains nominal or ordinal effects, those effects are represented by sets of indicator columns.
- When a categorical effect has only two levels, that effect is represented by a single column.
- When a categorical effect has k levels, where $k > 2$, then it must be represented by k-1 columns.

The convention in JMP for standard platforms is to represent nominal variables by terms whose parameter estimates average to zero across all the levels.

In the Stepwise platform, categorical variables (nominal and ordinal) are coded in a *hierarchical* fashion. This differs from coding in other least squares fitting platforms. In hierarchical coding, the levels of the categorical variable are successively split into groups of levels that most separate the means of the response. The splitting process achieves the goal of representing a k-level categorical variable by k - 1 terms.

Note: In hierarchical coding, the initial terms that are constructed represent the groups responsible for the greatest separation. The advantage of this coding scheme is that these informative terms have the potential to enter the model early.

Construction of Hierarchical Terms

Hierarchical terms are constructed using a tree structure that is analogous to a Partition analysis. However, the criterion that is maximized is the sum of squares between groups (SSB).

For a nominal variable with *k* levels, the *k* levels are split into two groups of levels that have maximum SSB. Call these two groups of levels A1 and A2, where A1 has the smaller mean and A2 has the larger mean. The two groups of levels in A1 and A2 are used to define an indicator variable with values of 1 for the levels in A1 and -1 for the levels in A2. This variable is the first hierarchical term for the nominal variable.

For the levels within each of the initial two groups A1 and A2, the split into two groups of levels with the maximum SSB is identified. Suppose that the groups of levels with maximum SSB are among the levels in A1. Call the two groups B1 and B2, where A1 has the smaller mean and A2 has the larger mean. The two groups of levels in B1 and B2 are used to define a hierarchical variable with values of 1 for the levels in B1, -1 for the levels in B2, and 0 for the levels in A2. To construct the next variable, splits of the levels in B1, B2, and A2 are considered. The split that maximizes SSB defines the next hierarchical variable. The process continues until *k-1* hierarchical terms are constructed.

For an ordinal variable, the groups of levels considered in splitting contain only levels that are contiguous in the ordering. This ensures that the constructed terms respect the level ordering.

Rules and Hierarchical Terms

When you use the **Combine** rule or the **Restrict** rule, a term cannot enter the model unless all the terms above it in the hierarchy have been entered. When you use the **Whole Effects** rule and enter a term for a categorical variable, all of its associated terms are entered. For an example, see "Construction of Hierarchical Terms in Example" on page 269.

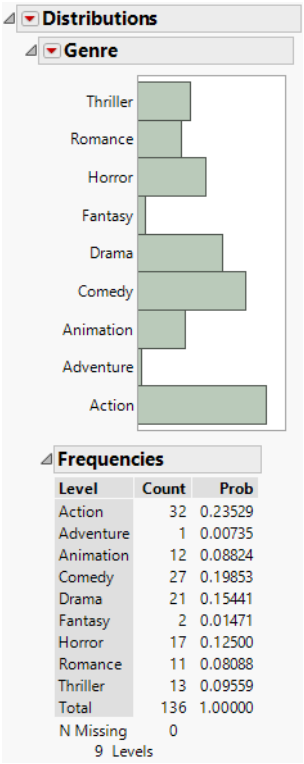
Example of a Model with a Nominal Term

This example uses data on movies that were released in 2011. You are particularly interested in the World Gross values, which represent the gross receipts. Your potential predictors are Rotten Tomatoes Score, Audience Score, and Genre. The two score variables are continuous, but Genre is nominal. Before you attempt to reduce your model using Stepwise, you want to explore the variables of interest.

1. Select **Help > Sample Data Library** and open Hollywood Movies.jmp.

2. Select **Analyze > Distribution**.
3. Select Genre and click **Y, Columns**.
4. Click **OK**.

Figure 5.14 Distribution of Genre



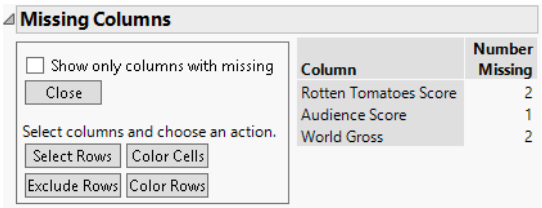
Level	Count	Prob
Action	32	0.23529
Adventure	1	0.00735
Animation	12	0.08824
Comedy	27	0.19853
Drama	21	0.15441
Fantasy	2	0.01471
Horror	17	0.12500
Romance	11	0.08088
Thriller	13	0.09559
Total	136	1.00000

N Missing 0
9 Levels

Note that Genre has nine levels, and so would be represented by eight model terms. Further data exploration will reveal that, because of missing data, only eight levels are considered by Stepwise.

5. In the data table's Columns panel, select the columns of interest: Rotten Tomatoes Score, Audience Score, and World Gross.
6. Select **Analyze > Screening > Explore Missing Values**.
7. Click **Y, Columns** and click **OK**.

Figure 5.15 Missing Columns Report



Note that Rotten Tomatoes Score is missing in 2 rows, Audience Score is missing in 1 row, and World Gross is missing in 2 rows.

8. In the Missing Columns report, select the three columns listed under **Column**.
9. Click **Select Rows**.

   In the data table's Rows panel, you can see that three rows are selected. Because these three rows contain missing data on the predictors or response, they will be automatically excluded from the Stepwise analysis. Note that row 134 is the only entry in the Adventure category, which means that category will be entirely removed from the analysis. For the purposes of the Stepwise analysis, it follows that Genre has only eight categories. Now that you have seen the effect of the missing data, you will conduct the Stepwise analysis.
10. Select **Analyze > Fit Model**.
11. Select Rotten Tomatoes Score, Audience Score, and Genre and click **Add**.

    If you fit a standard least squares model to World Gross using Rotten Tomatoes Score, Audience Score, and Genre as predictors, the residuals are highly heteroskedastic. (This is typical of financial data.) Use a log transformation to better satisfy the regression assumption of equal variance.
12. Right-click on World Gross in the Select Columns list and select **Transform > Log**.

    The transformed variable *Log[World Gross]* appears at the bottom of the Select Columns list.
13. Select *Log[World Gross]* and click **Y**.
14. Select **Stepwise** from the Personality list.
15. Click **Run**.

Figure 5.16 Current Estimates Table Showing List of Model Terms

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	4.05798507	1	0	0.000	1
☐	☐	Rotten Tomatoes Score	0	1	8.732189	2.709	0.10219
☐	☐	Audience Score	0	1	7.509642	2.323	0.12989
☐	☐	Genre{Drama&Thriller&Horror&Fantasy&Romance&Comedy-Action&Animation}	0	1	53.40235	18.526	3.26e-5
☐	☐	Genre{Drama-Thriller&Horror&Fantasy&Romance&Comedy}	0	1	11.02293	3.438	0.06595
☐	☐	Genre{Thriller&Horror&Fantasy&Romance-Comedy}	0	1	4.435756	1.362	0.24528
☐	☐	Genre{Thriller&Horror-Fantasy&Romance}	0	1	0.318501	0.097	0.75611
☐	☐	Genre{Thriller-Horror}	0	1	0.000104	0.000	0.99553
☐	☐	Genre{Fantasy-Romance}	0	1	0.000376	0.000	0.99149
☐	☐	Genre{Action-Animation}	0	1	1.505589	0.459	0.49919

In the Current Estimates table, note that Genre is represented by 7 terms. You will construct a model using two of these to see how these terms are defined.

16. Check the boxes under **Entered** next to the first two terms for Genre:
    - Genre{Drama&Thriller&Horror&Fantasy&Romance&Comedy-Action&Animation}
    - Genre{Drama-Thriller&Horror&Fantasy&Romance&Comedy}
17. Click **Make Model**.

    Notice that the two terms are added as temporary transform columns to the Model Effects list in the Model Specification window. These columns are discussed in the next section.

Construction of Hierarchical Terms in Example

Recall that because of missing values, Genre is a nominal variable with eight levels. In the Current Estimates table, Genre is represented by seven terms. This is appropriate, because Genre has eight levels. The first two terms that represent Genre are described below. Subsequent terms are defined in a similar fashion.

First Term

The first term that appears is Genre{Drama&Thriller&Horror&Fantasy&Romance&Comedy-Action&Animation}. This variable has the form Genre{A1 - A2}, where A1 and A2 are separated by a minus sign. The notation indicates that the maximum separation in terms of sum of squares between groups occurs between the following two sets of levels:

- Drama, Thriller, Horror, Fantasy, Romance, and Comedy (represented by A1)
- Action and Animation (represented by A2)

If you include the term Genre{Drama&Thriller&Horror&Fantasy&Romance&Comedy-Action&Animation} in a model, a temporary transform column representing that term is used in the model. The column contains the following values:

- 1 for Drama, Thriller, Horror, Fantasy, Romance, and Comedy

- -1 for Action and Animation

Second Term

The second term that appears is Genre{Drama-Thriller&Horror&Fantasy&Romance&Comedy}. This set of levels is entirely contained in the first split for the first term (A1). The notation contrasts the levels:

- Drama
- Thriller, Horror, Fantasy, Romance, and Comedy

Among all the splits of the levels of Drama, Thriller, Horror, Fantasy, Romance, and Comedy (A1) and of the levels of Action and Animation (A2), the algorithm determines that this split has the largest sum of squares between groups.

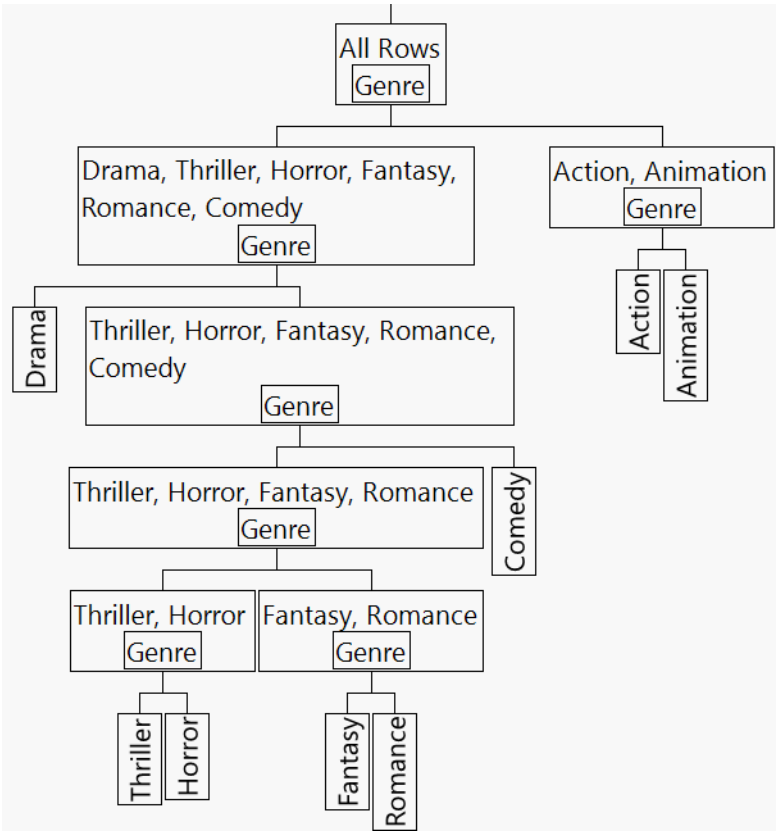
If you include this term in a model, a temporary transform column representing that term is used in the model. The column contains the following values:

- 1 for Drama
- -1 for Thriller, Horror, Fantasy, Romance, and Comedy
- 0 for Action and Animation

Hierarchy of Terms

The splitting of terms continues, based on the sum of squares between groups criterion. The hierarchy that leads to the definition of the terms is illustrated in Figure 5.17.

Figure 5.17 Tree Showing Splits Used in Hierarchical Coding



Rules

When you use the **Combine** rule or the **Restrict** rule, a term cannot enter the model unless all the terms above it in the hierarchy have been entered. For example, if you enter Genre{Action-Animation}, then JMP will enter Genre{Drama&Thriller&Horror&Fantasy&Romance&Comedy-Action&Animation} as well.

When you use the **Whole Effects** rule and enter any one of the Genre terms, all of the Genre terms are entered.

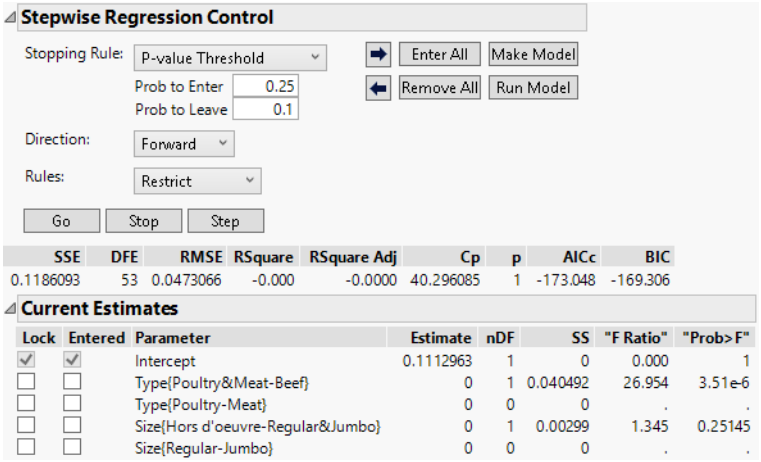
Example of the Restrict Rule for Hierarchical Terms

If you have a model with nominal or ordinal terms, when you make or run the model, temporary transform columns containing the hierarchical terms involved in the model are used in the model fit. The model itself appears in a new Fit Model window. This example further illustrates how **Stepwise** constructs a model with hierarchical effects.

A simple model examines at the cost per ounce ($/oz) of hot dogs as a function of the Type of hot dog (Meat, Beef, Poultry) and the Size of the hot dog (Jumbo, Regular, Hors d'oeuvre).

1. Select **Help > Sample Data Library** and open Hot Dogs2.jmp.
2. Select **Analyze > Fit Model**.
3. Select $/oz and click **Y**.
4. Select Type and Size and click **Add**.
5. For **Personality**, select **Stepwise**.
6. Click **Run**.
7. For **Stopping Rule**, select **P-value Threshold**.
8. For **Rules**, select **Restrict**.

Figure 5.18 Stepwise Control Panel with P-value Threshold and Restrict Rule



Notice that when you change from the default Rule of Combine to Restrict, the F Ratio and Prob > F values for two terms are shown as missing. These are the terms Type{Poultry-Meat} and Size{Regular-Jumbo}. This is because these two terms cannot enter the model until their precedent terms enter.

9. Click **Step**.

   The term Type{Poultry&Meat-Beef} enters the model. This term has the smallest Prob>F value, and that value falls below the Prob to Enter threshold of 0.25.

Figure 5.19 Stepwise Control Panel with One Term Entered

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	0.11864706	1	0	0.000	1
☐	☑	Type{Poultry&Meat-Beef}	-0.0283529	1	0.040492	26.954	3.51e-6
☐	☐	Type{Poultry-Meat}	0	1	0.012426	9.648	0.00309
☐	☐	Size{Hors d'oeuvre-Regular&Jumbo}	0	1	0.001038	0.687	0.4112
☐	☐	Size{Regular-Jumbo}	0	0	0	.	.

The the F Ratio and Prob > F values for the term Type{Poultry-Meat} appear. Since its precedent term has entered the model, Type{Poultry-Meat} is now allowed to enter.

10. Click **Step**.

    Since Type{Poultry-Meat} has the smallest Prob>F value among the remaining terms, and that value is below the Prob to Enter threshold, it is the next term to enter the model.

11. Click **Step**.

    The term Size{Hors d'oeuvre-Regular&Jumbo} enters the model, since its Prob>F value is 0.1577. Because its precedent term is now in the model, the term Size{Regular-Jumbo} is allowed to enter the model and its Prob>F value appears.

    However, the Prob>F value for the term Size{Regular-Jumbo} is 0.7566, which exceeds the Prob to Enter value of 0.25. For this reason, if you click Step again, it is not entered into the model.

Figure 5.20 Current Estimates Report for the Final Model

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	SS	"F Ratio"	"Prob>F"
☑	☑	Intercept	0.10849381	1	0	0.000	1
☐	☑	Type{Poultry&Meat-Beef}	-0.0275278	0	0	.	.
☐	☑	Type{Poultry-Meat}	-0.0205527	1	0.013985	11.083	0.00164
☐	☑	Size{Hors d'oeuvre-Regular&Jumbo}	-0.0121982	1	0.002596	2.057	0.1577
☐	☐	Size{Regular-Jumbo}	0	1	0.000125	0.097	0.7566

Tip: Use the Go button to run the entire stepwise process automatically. To see this in action, click **Remove All**. Then click **Go**.

12. Click **Make Model**.

    After you click Make Model, a Fit Model launch window appears, containing only the three model effects that were selected in the stepwise process. In the launch window, temporary transform columns that define the three hierarchical effects entered into the model are included in the model.

Performing Binary and Ordinal Logistic Stepwise Regression

The Stepwise personality of Fit Model performs ordinal logistic stepwise regression when the response is ordinal or nominal. Nominal responses are treated as ordinal responses in the logistic stepwise regression fitting procedure. When a response has only two levels, ordinal logistic regression models are equivalent to nominal logistic regression models. To run a logistic stepwise regression, specify an ordinal or nominal response, add terms to the model as usual, and choose Stepwise from the Personality menu.

The Stepwise reports for a logistic model are similar to those provided when the response is continuous. The following elements are specific to logistic regression results:

- When the response is categorical, the overall fit of the model is given by its negative log-likelihood (-LogLikelihood). This value is calculated based on the full iterative maximum likelihood fit.
- The Current Estimates report shows chi-square statistics (Wald/Score ChiSq) and their *p*-values (Sig Prob). The test statistic column shows score statistics for parameters that are not in the current model and shows Wald statistics for parameters that are in the current model. The regression estimates (Estimate) are based on the full iterative maximum likelihood fit.
- The Step History report shows the L-R ChiSquare. This value is the test statistic for the likelihood ratio test of the hypothesis that the corresponding regression parameter is zero, given the other terms in the model. The Sig Prob is the *p*-value for this test.

Note: If the response is nominal, you can fit the current model using the Nominal Logistic personality of Fit Model by clicking the Make Model button. In the Fit Model launch window that appears, click Run.

Example Using Logistic Stepwise Regression

1. Select **Help > Sample Data Library** and open Fitness.jmp.
2. Select **Analyze > Fit Model**.
3. Select Sex and click **Y**.
4. Select Weight, Runtime, RunPulse, RstPulse, and MaxPulse and click **Add**.
5. For **Personality**, select **Stepwise**.
6. Click **Run**.
7. Click **Go**.

Figure 5.21 Logistic Stepwise Report

Current Estimates

Lock	Entered	Parameter	Estimate	nDF	Wald/Score ChiSq	"Sig Prob"
☑	☑	Intercept[F]	35.8078459	1	0	1
☐	☑	Weight	-0.2822965	1	5.471429	0.01933
☐	☑	Runtime	-1.3268037	1	2.545732	0.11059
☐	☐	RunPulse	0	1	0.470408	0.4928
☐	☐	RstPulse	0	1	0.007216	0.9323
☐	☐	MaxPulse	0	1	0.076497	0.7821

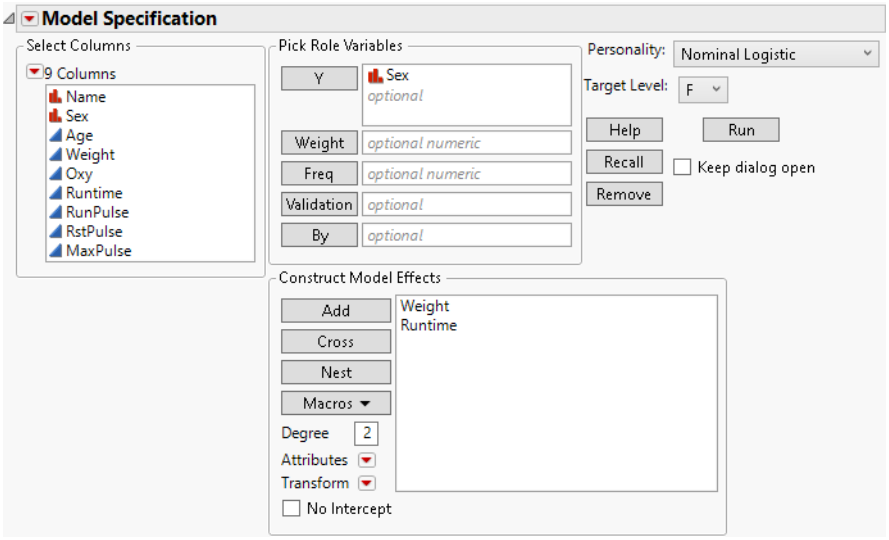
Step History

Step	Parameter	Action	L-R ChiSquare	"Sig Prob"	Entry ChiSquare	Entry "Sig Prob"	RSquare	p	AICc	BIC
1	Weight	Entered	12.16669	0.0005	10.1342	0.00146	0.2833	2	35.2047	37.6441
2	Runtime	Entered	8.017826	0.0046	6.37139	0.0116	0.4700	3	29.6472	33.0603
3	RunPulse	Entered	1.440088	0.2301	1.44684	0.22903	0.5036	4	30.8567	35.0542
4	MaxPulse	Entered	0.071809	0.7887	0.06994	0.79142	0.5052	5	33.6464	38.4164
5	RstPulse	Entered	0.007156	0.9326	0.00722	0.93228	0.5054	6	36.7393	41.8432
6	Best	Specific	.	.	0.00722	0.93228	0.4700	3	29.6472	33.0603

The two variables Weight and Runtime are entered into the model based on the Stopping Rule.

8. Click **Make Model**.

Figure 5.22 Model Specification Window for Reduced Model



A model specification window opens containing the two variables as model effects. Note that the Personality is Nominal Logistic. If the response had been ordinal, the Personality would be Ordinal Logistic.

The All Possible Models Option

Use the All Possible Models option to investigate all models that can be constructed using your predictors. This option is accessed from the red triangle menu next to Stepwise.

Note the following:

- This option is not practical for large problems, when the number of models is greater than 5 million.
- Categorical predictors are represented by indicator variables. See "Models with Nominal and Ordinal Effects" on page 265.

The following options restrict the number of models that appear:

Maximum number of terms in a model Enter a value for the maximum number of terms in a model.

Number of best models to see Enter the maximum number of models of each size to display. The best models according to RSquare value appear.

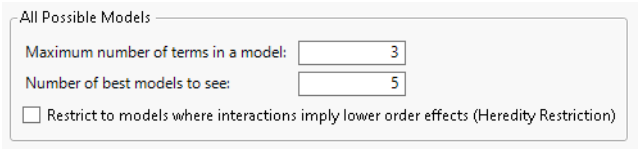
Restrict to models where interactions imply lower order effects (Heredity Restriction)

Shows only models that contain all lower-order effects when a higher-order effect is included. These models satisfy strong effect heredity. This option is useful when your predictors include interaction or polynomial terms.

Example Using the All Possible Models Option

1. Select **Help > Sample Data Library** and open Fitness.jmp.
2. Select **Analyze > Fit Model**.
3. Select Oxy and click **Y**.
4. Select Runtime, RunPulse, RstPulse, and MaxPulse and click **Add**.
5. For **Personality**, select **Stepwise**.
6. Click **Run**.
7. From the red triangle menu next to Stepwise, select **All Possible Models**.
8. Enter 3 for the maximum number of terms, and enter 5 for the number of best models.

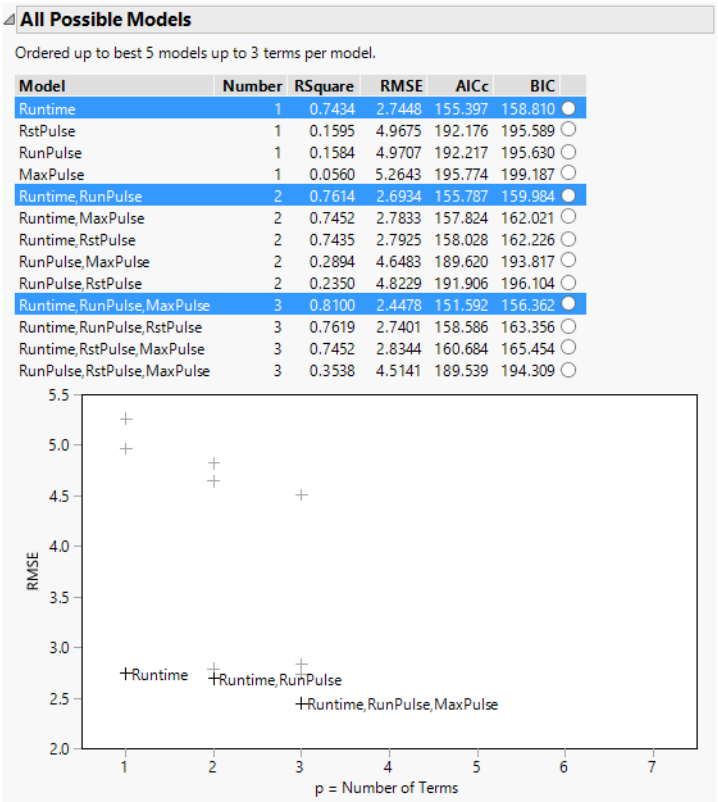
Figure 5.23 All Possible Models Popup Dialog



9. Click **OK**.

All possible models (up to three terms in a model) are fit.

Figure 5.24 All Possible Models Report



The models are listed in increasing order of the number of parameters that they contain. The model with the highest R^2 for each number of parameters is highlighted. The radio button column at the right of the table enables you to select one model at a time and check the results.

Note: The recommended criterion for selecting a model is to choose the one corresponding to the smallest BIC or AICc value. Some analysts also want to see the C_p statistic. Mallow's C_p statistic is computed, but initially hidden in the table. To make it visible, right-click in the table and select **Columns > Cp**.

The Model Averaging Option

The model averaging technique enables you to average the fits for a number of models, instead of picking a single best model. The result is a model with excellent prediction capability. This feature is particularly useful for new and unfamiliar models that you do not want to overfit. When many terms are selected into a model, the fit tends to inflate the estimates. Model averaging tends to shrink the estimates on the weaker terms, yielding better predictions. The models are averaged with respect to the AICc weight, calculated as follows:

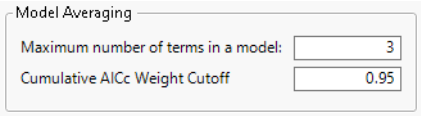
AICcWeight = exp[-0.5(AICc - AICcBest)]

AICcBest is the smallest AICc value among the fitted models. The AICc Weights are then sorted in decreasing order. The AICc weights cumulating to less than one minus the cutoff of the total AICc weight are set to zero, allowing the very weak terms to have true zero coefficients instead of extremely small coefficient estimates.

Example Using the Model Averaging Option

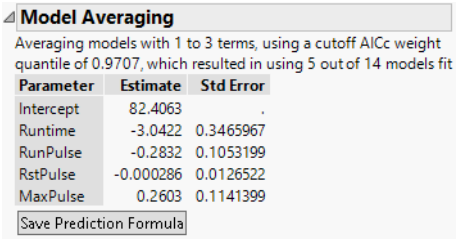
1. Select **Help > Sample Data Library** and open Fitness.jmp.
2. Select **Analyze > Fit Model**.
3. Select Oxy and click **Y**.
4. Select Runtime, RunPulse, RstPulse, and MaxPulse and click **Add**.
5. For **Personality**, select **Stepwise**.
6. Click **Run**.
7. From the red triangle menu next to Stepwise, select **Model Averaging**.
8. Enter 3 for the maximum number of terms, and keep 0.95 for the weight cutoff.

Figure 5.25 Model Averaging Window



9. Click **OK**.

Figure 5.26 Model Averaging Report



Model Averaging

Averaging models with 1 to 3 terms, using a cutoff AICc weight quantile of 0.9707, which resulted in using 5 out of 14 models fit

Parameter	Estimate	Std Error
Intercept	82.4063	.
Runtime	-3.0422	0.3465967
RunPulse	-0.2832	0.1053199
RstPulse	-0.000286	0.0126522
MaxPulse	0.2603	0.1141399

Save Prediction Formula

In the Model Averaging report, average estimates and standard errors appear for each parameter. The standard errors shown reflect the bias of the estimates toward zero.

10. Click **Save Prediction Formula** to save the prediction formula in the original data table.

Using Validation

In JMP, you can perform cross-validation by selecting the **K-Fold Crossvalidation** option from the Stepwise Fit red triangle menu. See "K-Fold Cross Validation" on page 283.

JMP PRO In JMP Pro, you can specify a Validation column in the Fit Model window. A validation column must have a numeric data type and should contain at least two distinct values.

- If the column contains two values, the smaller value defines the training set and the larger value defines the validation set.
- If the column contains three values, the values define the training, validation, and test sets in order of increasing size.
- If the column contains four or more distinct values and the response is continuous, these values define folds for k-fold validation.

For more information about using a Validation column, see "Validation Set with Two or Three Values" on page 279.

Validation Set with Two or Three Values

JMP PRO If you specify a Validation column with two or three values, Stepwise fits models based on the training set. Model fit statistics are reported for the validation and test sets. See "Validation and Test Set Statistic Definitions" on page 281 for details on how these statistics are defined.

If the response is continuous, the following statistics appear in the Stepwise Regression Control panel:

- RSquare Validation (also shown in the Step History report)
- RMSE Validation
- RSquare Test (if there is a test set)
- RMSE Test (if there is a test set)

If the response is binary nominal or ordinal, the following statistics appear in the Stepwise Regression Control panel:

- RSquare Validation (also shown in the Step History report)
- Avg Log Error Validation
- RSquare Test (if there is a test set)
- Avg Log Error Test (if there is a test set)

Max Validation RSquare

If you specify a validation column with two or three values in the Fit Model window, the Stopping Rule defaults to Max Validation RSquare. This rule attempts to find a model that maximizes the RSquare statistic for the validation set. The rule can be applied with the Direction set to Forward or Backward.

Note: Max Validation RSquare considers only the models defined by *p*-value entry (Forward direction) or removal (Backward direction). It does not consider all possible models.

You can use the Step button to enter terms one-by-one in the Forward direction or to remove them one-by one in the Backward direction. At any point, you can select a model by clicking the button to the right of RSquare Validation in the Step History report. The selection of model terms is updated in the Current Estimates report. This is the model that is used once you click Make Model or Run Model.

Forward Direction

In the Forward direction, Stepwise constructs successive models by adding terms based on the next smallest p-value.

If you click Go rather than Step, the process of entering terms proceeds automatically. Among the fitted models, the model that is considered best is listed last. This model is obtained by overlooking local dips in RSquare Validation. Specifically, it is the model with the largest RSquare Validation that can be followed by as many as ten models with lower RSquare Validation values. This model is designated by the terms Best in the Parameter column and Specific in the Action column. The button to the right of RSquare Validation selects this Best model, though you are free to change this selection.

Backward Direction

In the Backward direction, Stepwise constructs successive models by removing terms based on the next largest p-value.

To use the Backward direction, you must first click Enter All to enter all of the terms into the model. The Backward direction behaves in a similar fashion to the Forward direction. If you click Go rather than Step, the process of removing terms proceeds automatically. The model designated as Best is the one with the largest RSquare Validation that can be followed by as many as ten models with lower RSquare Validation values.

Validation and Test Set Statistic Definitions

RSquare Validation and RMSE Validation are defined in this section. RSquare Test and RMSE Test are computed for the test set in a completely analogous fashion.

Continuous Response

RSquare Validation An RSquare measure for the validation set computed as follows:

- For each observation in the validation set, compute the prediction error. This is the difference between the actual response and the response predicted by the training set model.
- Square and sum the prediction errors to obtain $SSE_{Validation}$.
- Square and sum the differences between the actual responses in the validation set and their mean. This is the $SST_{Validation}$.
- RSquare Validation is:

$$\text{RSquare Validation} = 1 - \frac{SSE_{Validation}}{SST_{Validation}}$$

Note: It is possible for RSquare Validation to be negative.

RMSE Validation The square root of the mean squared prediction error for the validation set. This is computed as follows:

- For each observation in the validation set, compute the prediction error. This is the difference between the actual response and the response predicted by the training set model.
- Square and sum the prediction errors to obtain the $SSE_{Validation}$.
- Denote the number of observations in the validation set by $n_{Validation}$.
- RMSE Validation is:

$$\text{RMSE Validation} = \sqrt{\frac{SSE_{Validation}}{n_{Validation}}}$$

Note: In the Fit Least Squares Crossvalidation report, the entries in the RASE (Root Average Squared Error) column for the Validation Set and Test Set are the RMSE Validation and RMSE Test values computed in the Stepwise report. See "RASE" on page 89.

Binary Nominal or Ordinal Response

RSquare Validation An Entropy RSquare measure (also known as McFadden's R^2) for the validation set computed as follows:

- A model is fit using the training set.
- Predicted probabilities are obtained for all observations.
- Using the predicted probabilities based on the training set model, the likelihood for the model is computed for observations in the validation set. Call this quantity $Likelihood\_Full_{Validation}$.
- Using the data in the validation set, the likelihood of the reduced model (no predictors) is computed. Call this quantity $Likelihood\_Reduced_{Validation}$.
- RSquare Validation is:

$$\text{RSquare Validation} = 1 - \frac{\log(Likelihood\_Full_{Validation})}{\log(Likelihood\_Reduced_{Validation})}$$

Note: It is possible for RSquare Validation to be negative.

Avg Log Error Validation The average log error for the validation set is computed as follows:

- For each observation in the validation set, compute the log of its predicted probability as determined by the model based on the training set.
- Sum these logs, divide by the number of observations in the validation set, and take the negative of the resulting value.

Tip: Smaller values of Avg Log Error Validation are desirable.

K-Fold Cross Validation

K-fold cross validation randomly divides the data into *k* subsets. In turn, each of the *k* sets is used as a validation set while the remaining data are used as a training set to fit the model. In total, *k* models are fit and *k* validation statistics are obtained. The model giving the best validation statistic is chosen as the final model. This method is useful for small data sets, because it makes efficient use of limited amounts of data.

Note: K-fold cross validation is only available for continuous responses.

In JMP, select **K-Fold Crossvalidation** from the red triangle options for Stepwise Fit.

In JMP Pro, you can access *k*-fold cross validation in two ways:

- From the red triangle options for Stepwise Fit, select **K-Fold Crossvalidation**.
- Specify a validation column with four or more distinct values.

RSquare K-Fold Statistic

If you conduct k-fold cross validation, the RSquare K-Fold statistic appears to the right of the other statistics in the Stepwise Regression Control panel. RSquare K-Fold is the average of the RSquare Validation values for the *k* folds.

Max K-Fold RSquare

When you use *k*-fold cross validation, the Stopping Rule defaults to Max K-Fold RSquare. This rule attempts to maximize the RSquare K-Fold statistic.

Note: Max K-Fold RSquare considers only the models defined by *p*-value entry (Forward direction) or removal (Backward direction). It does not consider all possible models.

The Max K-Fold RSquare stopping rule behaves in a fashion similar to the Max Validation RSquare stopping rule. See "Max Validation RSquare" on page 280. Replace references to RSquare Validation with RSquare K-Fold.

Chapter 6

JMP PRO Generalized Regression Models

Build Models Using Variable Selection Techniques

The Generalized Regression personality of the Fit Model platform is available only in JMP Pro.

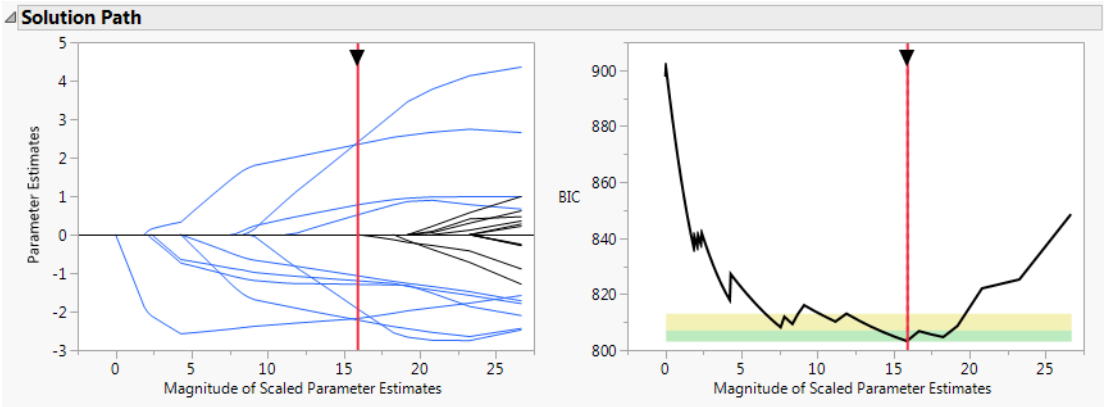
In JMP Pro, the Fit Model platform's Generalized Regression personality provides variable selection techniques, including shrinkage techniques, that specifically address modeling correlated and high-dimensional data. Two of these techniques, the Lasso and the Elastic Net, perform variable selection as part of the modeling procedure.

Large data sets that contain many variables typically exhibit multicollinearity issues. Modern data sets can include more variables than observations, requiring variable selection if traditional modeling techniques are to be used. The presence of multicollinearity and a profusion of predictors exposes the shortcomings of classical techniques.

Even for small data sets with little or no correlation, including designed experiments, the Lasso and Elastic Net are useful. They can be used to build predictive models or to select variables for model reduction or for future study.

The Generalized Regression personality is useful for many modeling situations. This personality enables you to specify a variety of distributions for your response variable. Use it when your response is continuous, binomial, a count, or zero-inflated. Use it when you are interested in variable selection or when you suspect collinearity in your predictors. More generally, use it to fit models that you compare to models obtained using other techniques.

Figure 6.1 The Solution Path for an Elastic Net Fit



Contents

JMP PRO Overview of the Generalized Regression Personality

The Generalized Regression personality features regularized, or penalized, regression techniques. Such techniques attempt to fit better models by shrinking the model coefficients toward zero. The resulting estimates are biased. This increase in bias can result in decreased prediction variance, thus lowering overall prediction error compared to non-penalized models. Two of these techniques, the Elastic Net and the Lasso, include variable selection as part of the modeling procedure.

Modeling techniques such as the Elastic Net and the Lasso are particularly useful for large data sets, where collinearity is typically a problem. In addition, modern data sets often include more variables than observations. This situation is sometimes referred to as the $p > n$ problem, where n is the number of observations and p is the number of predictors. Such data sets require variable selection if traditional modeling techniques are to be used.

The Elastic Net and Lasso can also be used for small data sets with little correlation, including designed experiments. They can be used to build predictive models or to select variables for model reduction or for future study.

The personality provides the following classes of modeling techniques:

- Maximum Likelihood
- Step-Based Estimation
- Penalized Regression

The Elastic Net and Lasso are relatively recent techniques (Tibshirani 1996; Zou and Hastie 2005). Both techniques penalize the size of the model coefficients, resulting in a continuous shrinkage. The amount of shrinkage is determined by a *tuning parameter*. An optimal level of shrinkage is determined by one of several validation methods. Both techniques have the ability to shrink coefficients to zero. In this way, variable selection is built into the modeling procedure. The Elastic Net model subsumes both the Lasso and ridge regression as special cases. For details, see "Statistical Details for Estimation Methods" on page 332.

Details about Generalized Regression Modeling Techniques

- The Maximum Likelihood method is a classical approach. It provides a baseline to which you can compare the other techniques.
- Forward Selection is a method of stepwise regression. In forward selection, terms are entered into the model. The most significant terms are added until all of the terms are in the model or there are no degrees of freedom left.

- The Lasso has two shortcomings. When several variables are highly correlated, it tends to select only one variable from that group. When the number of variables, p, exceeds the number of observations, n, the Lasso selects at most n predictors.
- The Elastic Net, on the other hand, tends to select all variables from a correlated group, fitting appropriate coefficients. It can also select more than n predictors when $p > n$.
- Ridge regression was among the first of the penalized regression methods proposed (Hoerl 1962; Hoerl and Kennard 1970). Ridge regression does not shrink coefficients to zero, so it does not perform variable selection.
- The Double Lasso attempts to separate the selection and shrinkage steps by performing variable selection with an initial Lasso model. The variables selected in the initial model are then used as the input variables for a second Lasso model.
- Two-Stage Forward Selection performs two stages of forward stepwise regression. It performs variable selection on the main effects in the first stage. Then, higher-order effects are allowed to enter the model in the second stage.

The Generalized Regression personality also fits an *adaptive* version of the Lasso and the Elastic Net. These adaptive versions attempt to penalize variables in the true active set less than variables not contained in the true active set. The *true active set* refers to the set of terms in a model that have an actual effect on the response. The adaptive versions of the Lasso and Elastic Net were developed to ensure that the oracle property holds. The oracle property guarantees the following: Asymptotically, your estimates are what they would have been had you fit the model to the true active set of predictors. More specifically, your model correctly identifies the predictors that should have zero coefficients. Your estimates converge to those that would have been obtained had you started with only the true active set. See "Adaptive Methods" on page 334.

The Generalized Regression personality enables you to specify a variety of distributions for your response variable. The distributions fit include normal, Cauchy, exponential, gamma, Weibull, lognormal, beta, binomial, beta binomial, Poisson, negative binomial, zero-inflated binomial, zero-inflated beta binomial, zero-inflated Poisson, zero-inflated negative binomial, and zero-inflated gamma. This flexibility enables you to fit categorical and count responses, as well as continuous responses, and specifically, right-skewed continuous responses. You can also fit quantile regression and Cox proportional hazards models. For some of the distributions, you can fit models to censored data. The personality provides a variety of validation criteria for model selection and supports training, validation, and test columns. See "Distribution" on page 292.

JMP PRO Example of Generalized Regression

The data in the Diabetes.jmp sample data table consist of measurements on 442 diabetics. The response of interest is Y, disease progression measured one year after a baseline measure was

taken. Ten variables thought to be related to disease progression are also measured at baseline. This example shows how to develop a predictive model using generalized regression techniques.

1. Select **Help > Sample Data Library** and open Diabetes.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y from the Select Columns list and click **Y**.
4. Select Age through Glucose and click **Macros > Factorial to Degree**.

   This adds all terms up to degree 2 (the default in the **Degree** box) to the model.
5. Select Validation from the Select Columns list and click **Validation**.
6. From the Personality list, select **Generalized Regression**.
7. Click **Run**.

   The Generalized Regression report that appears contains a Model Launch control panel and a Standard Least Squares with Validation Column report.

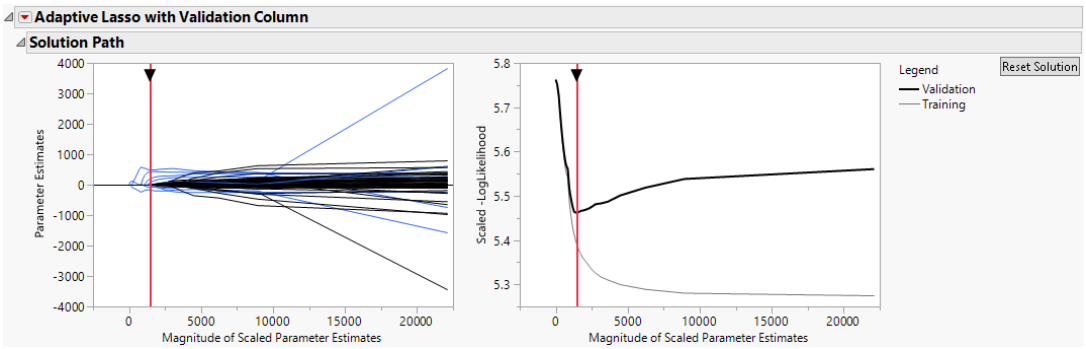
   In the Model Launch control panel, note the following:

   – The default estimation method is the Adaptive Lasso.

   – The Validation Method is set to Validation Column because you specified a validation column in the Fit Model window.
8. Click **Go**.

   An Adaptive Lasso with Validation Column report appears. The Solution Path report (Figure 6.2) shows plots of the parameter estimates and scaled negative log-likelihood. The shrinkage increases as the Magnitude of Scaled Parameter Estimates decreases. The estimates at the far right of the plot are the maximum likelihood estimates. A vertical red line indicates those parameter values selected by the validation criterion, in this case, the holdback sample defined by the column Validation.

Figure 6.2 Solution Path Plot



9. Select the option **Select Nonzero Terms** from the Adaptive Lasso with Validation Column report's red triangle menu.

   This option highlights the nonzero terms in the Parameter Estimates for Original Predictors report (Figure 6.3) and their paths in the Solution Path Plot. The corresponding columns in the data table are also selected. Note that only 7 of the 55 parameter estimates are nonzero. Also note that the scale parameter for the normal distribution (sigma) is estimated and shown in a separate table at the bottom of the Parameter Estimates for Original Data report.

Figure 6.3 Portion of Parameter Estimates for Original Predictors Report

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-247.27	41.051839	36.280859	<.0001*	-327.7301	-166.8099
Age	0	0	0	1.0000	0	0
Gender[1-2]	8.3631279	7.0586509	1.4037639	0.2361	-5.471574	22.197829
BMI	5.4813511	0.9003203	37.066466	<.0001*	3.7167558	7.2459464
BP	0.7654792	0.2570233	8.869965	0.0029*	0.2617229	1.2692356
Total Cholesterol	-0.137311	0.1162886	1.3942409	0.2377	-0.365233	0.0906102
LDL	0	0	0	1.0000	0	0
HDL	-0.810991	0.2860433	8.0383783	0.0046*	-1.371626	-0.250356
TCH	0	0	0	1.0000	0	0
LTG	53.103251	8.5783104	38.321165	<.0001*	36.290071	69.91643
Glucose	0	0	0	1.0000	0	0
(Age-48.5181)*Gender[1-2]	-0.008683	0.3632509	0.0005713	0.9809	-0.720641	0.703276
(Age-48.5181)*(BMI-26.3758)	0	0	0	1.0000	0	0
(Age-48.5181)*(BP-94.647)	0	0	0	1.0000	0	0
(Age-48.5181)*(Total Cholesterol-189.14)	0	0	0	1.0000	0	0
(Age-48.5181)*(LDL-115.439)	0	0	0	1.0000	0	0
(Age-48.5181)*(HDL-49.7885)	0	0	0	1.0000	0	0
(Age-48.5181)*(TCH-4.07025)	0	0	0	1.0000	0	0
(Age-48.5181)*(LTG-4.64141)	0	0	0	1.0000	0	0
(Age-48.5181)*(Glucose-91.2602)	0	0	0	1.0000	0	0
Gender[1-2]*(BMI-26.3758)	0	0	0	1.0000	0	0
Gender[1-2]*(BP-94.647)	0	0	0	1.0000	0	0
Gender[1-2]*(Total Cholesterol-189.14)	0	0	0	1.0000	0	0
Gender[1-2]*(LDL-115.439)	0	0	0	1.0000	0	0
Gender[1-2]*(HDL-49.7885)	0	0	0	1.0000	0	0
Gender[1-2]*(TCH-4.07025)	0	0	0	1.0000	0	0
Gender[1-2]*(LTG-4.64141)	0	0	0	1.0000	0	0
Gender[1-2]*(Glucose-91.2602)	0	0	0	1.0000	0	0
(BMI-26.3758)*(BP-94.647)	0	0	0	1.0000	0	0
(BMI-26.3758)*(Total Cholesterol-189.14)	0	0	0	1.0000	0	0
(BMI-26.3758)*(LDL-115.439)	0	0	0	1.0000	0	0

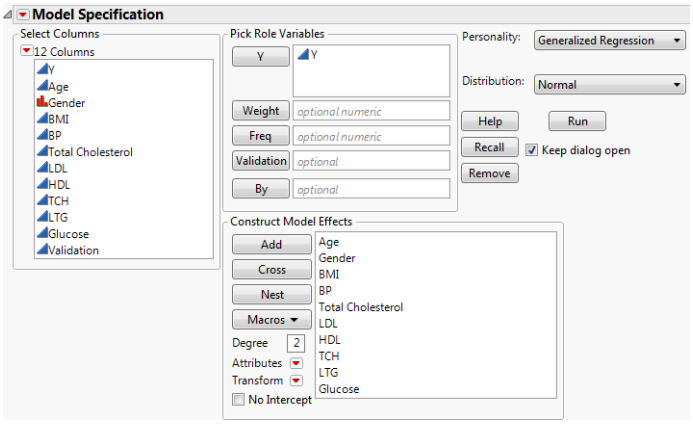
Normal Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Scale	53.159423	2.2864654	540.54426	<.0001*	48.678033	57.640813

To save the prediction formula, select **Save Columns > Save Prediction Formula** from the red triangle menu for the Adaptive Lasso with Validation Column report.

JMP PRO Launch the Generalized Regression Personality

Launch the Generalized Regression personality by selecting **Analyze > Fit Model**, entering one or more columns for **Y**, and selecting **Generalized Regression** from the **Personality** menu (Figure 6.4).

Figure 6.4 Fit Model Launch Window with Generalized Regression Selected



For more information about aspects of the Fit Model window that are common to all personalities, see the "Model Specification" chapter on page 33. Information specific to the Generalized Regression personality are presented here.

The parameterization of nominal variables used in the Generalized Regression personality differs from their parameterization using other Fit Model personalities. The Generalized Regression personality uses indicator function parameterization. In this parameterization, the estimate that corresponds to the indicator for a level of a nominal variable is an estimate of the difference between the mean response at that level and the mean response at the last level. The last level is the level with the highest value order coding; it is the level whose indicator function is not included in the model.

If your model effects have missing values, you can treat these missing values as informative categories. Select the Informative Missing option from the Model Specification red triangle menu.

To specify a model without an intercept term, select the No Intercept option in the Construct Model Effects panel of the Fit Model window. If you select this option, note the following:

- The predictors are not centered and scaled.
- Odds ratios, hazard ratios, and incidence rate ratios are not available in the report window.

Caution: Using the No Intercept option with the Lasso or Elastic Net is not recommended because the results are sensitive to the scale of the model effects. The adaptive versions of these estimation methods are recommended instead.

Censoring

You can specify censoring for your response variable in one of the following ways:

- For right-censored responses, specify a column that contains indicators for right-censored observations as a Censor column in the launch window. Select the value in that column that designates right-censored observations from the Censor Code list.
- For interval-censored and left-censored responses, specify two columns that define the censoring interval in the Y column role:
  - For interval-censored responses, the first Y variable gives the lower limit.
  - For left-censored responses, the first Y variable contains a missing value.
  - For both interval and left censoring, the second Y variable gives the upper limit for each response.

If you specify two columns for Y and a Distribution that supports censoring, an Alert appears that asks whether the columns represent censoring. If you choose No, the columns are treated as separate responses.

Note: You can specify the default behavior for two responses using the Treatment of Two Response Columns preference in Generalized Regression platform preferences.

Censoring is available when the specified Distribution is Normal, Exponential, Gamma, Weibull, Lognormal, or Cox Proportional Hazards.

JMP PRO Distribution

When you select **Generalized Regression** from the **Personality** menu, the **Distribution** option appears. Here you can specify a distribution for Y. The abbreviation *ZI* means *zero-inflated*. The distributions are separated into three categories based on their response: continuous, discrete, and zero-inflated. The options are described below.

Note: If you specify multiple Y variables in the Model Specification window, the same response distribution must be used for all of the specified Y variables. If you want to fit separate distributions to different response variables in the same Generalized Regression report, you must use a script.

JMP PRO Continuous

Normal Y has a normal distribution with mean μ and standard deviation σ. The normal distribution is symmetric and with a large enough sample size, can approximate a large variety of other distributions using the Central Limit Theorem. The link function for μ is the identity. That is, the mean of Y is expressed as a linear model.

Note: When the specified Distribution is Normal, Standard Least Squares replaces the Maximum Likelihood Estimation method.

The scale parameter for the normal distribution is σ. When there is no penalty in the estimation method, the estimate of the scale parameter σ is the root mean square error (RMSE). The RMSE is the square root of the usual unbiased estimator of σ^2. The results shown are equivalent to a standard least squares fit unless censored observations are involved.

Note: The parameterization of nominal variables used in the Generalized Regression personality differs from their parameterization using the Standard Least Squares personality. Because of this difference, parameter estimates differ for models containing nominal or ordinal effects.

See "Statistical Details for Distributions" on page 335.

Cauchy Y has a Cauchy distribution with location parameter μ and scale parameter σ. The Cauchy distribution has an undefined mean and standard deviation. The median and mode are both μ. Most data do not inherently follow a Cauchy distribution. However, it is useful for conducting a robust regression on data that contain a large proportion of outliers (up to 50%). The link function for μ is the identity. See "Statistical Details for Distributions" on page 335.

Exponential Y has an exponential distribution with mean parameter μ. The exponential distribution is right-skewed and is often used to model lifetimes or the time between successive events. The link function for μ is the logarithm. See "Statistical Details for Distributions" on page 335.

Gamma Y has a gamma distribution with mean parameter μ and dispersion parameter σ. The gamma is a flexible distribution and contains a family of other widely used distributions. For example, the exponential distribution is a special case of the gamma distribution where σ = μ. The chi squared distribution can also be derived from the gamma distribution. The link function for μ is the logarithm. See "Statistical Details for Distributions" on page 335.

Weibull Y has a Weibull distribution with mean parameter μ and scale parameter σ. The Weibull distribution is flexible and is often used to model lifetimes or the time until an

event. The link function for μ is the identity. See "Statistical Details for Distributions" on page 335.

LogNormal Y has a Lognormal distribution with mean parameter μ and scale parameter σ. The Lognormal distribution is right-skewed and is often used to model lifetimes or the time until an event. The link function for μ is the identity. See "Statistical Details for Distributions" on page 335.

Beta Y has a beta distribution with mean parameter μ and dispersion parameter σ. The response for the beta is between 0 and 1 (not inclusive) and is often used to model proportions or rates. The link function for μ is the logit. See "Statistical Details for Distributions" on page 335.

Quantile Regression Quantile regression models a specified conditional quantile of the response. No assumption is made about the form of the underlying distribution. When you select Quantile Regression, a Quantile box appears beneath the Distribution menu. Specify the desired quantile.

If you specify 0.5 (the default) for the Quantile on the Model Dialog window, quantile regression models the conditional median of the response. Quantile regression is particularly useful when the rate of change in the conditional quantile, expressed by the regression coefficients, depends on the quantile. An advantage of quantile regression over least squares regression is its flexibility for modeling data with heterogeneous conditional distributions.

Quantile Regression is fit by minimizing an objective function using an iterative approach. For more information about quantile regression, see Koenker and Hallock (2001) and Portnoy and Koenker (1997).

When you choose Quantile Regression, Maximum Likelihood is the only available Estimation Method, and None is the only available Validation Method.

Note: If a quantile regression fit is time intensive, a progress bar appears. The progress bar shows the relative change in the objective function. When you click Accept Current Estimates, the calculation stops and the reported parameter estimates correspond to the best model fit at that point.

Cox Proportional Hazards The Cox proportional hazards model is a regression model for time-to-event data with predictors. It is based on a multiplicative relationship between the predictors and the hazard function. It can be used to examine the effect of predictors on survival times. The model involves an arbitrary baseline hazard function that is scaled by the predictors to give a general hazard function. The proportional hazards model produces parameter estimates and standard errors for each predictor. The Cox

proportional hazards model was first proposed by D. R. Cox (1972). For more information about proportional hazards models, see Kalbfleisch and Prentice (2002).

When you choose Cox Proportional Hazards, the only available Validation Methods are BIC and AICc. Also, the Ridge Estimation Method is not available.

Note: When there are ties in the response, the Efron likelihood is used. See Efron (1977). This is a different method for handling ties than is used in the Proportional Hazard personality of the Fit Model platform or in the Fit Proportional Hazards platform.

JMP PRO Discrete

Binomial Y has a binomial distribution with parameters p and n. The response, Y, indicates the total number of successes in n independent trials with a fixed probability, p, for all trials. This distribution allows for the use of a sample size column. If no column is listed, it is assumed that the sample size is one. The link function for p is the logit. When you select a binary response variable that has a Nominal modeling type, Binomial is the only available response distribution. See "Statistical Details for Distributions" on page 335.

When you select Binomial as the Distribution, the response variable must be specified in one of the following ways.

- Unsummarized: If your data are not summarized as frequencies of events, specify a single binary column as the response. If this column has a modeling type of Nominal, you can designate one of the levels to be the Target Level.
- Summarized with Freq column: If your data are summarized as frequencies of successes and failures, specify a single binary column as the response. If this column has a modeling type of Nominal, you can designate one of the levels to be the Target Level. Assign the frequency column to the **Freq** role.
- Summarized with sample size column entered as second Y: If your data are summarized as frequencies of events (successes) and trials, specify two continuous columns as Y in this order: the count of the number of successes, and the count of the number of trials.

Note: When the specified Distribution is Binomial, Logistic Regression replaces the Maximum Likelihood Estimation method.

Beta Binomial Y has a beta binomial distribution with the probability of success, p, the number of trails, n, and overdispersion parameter, δ. This distribution is an overdispersed version of the binomial distribution.

Run demoBetaBinomial.jsl in the JMP Samples/Scripts folder to compare a beta binomial distribution with dispersion parameter δ to a binomial distribution with parameters p and $n = 20$.

The beta binomial distribution requires a sample size greater than one for each observation. Thus, the user must specify a sample size column. To insert a sample size column, specify two continuous columns as Y in this order: the count of the number of successes, and the count of the number of trials. The link function for p is the logit. See "Statistical Details for Distributions" on page 335.

Multinomial Y has a multinomial distribution with three or more discrete levels. The response variable must have a nominal or ordinal modeling type. The model fits separate intercepts and effects parameters for each level of the response variable. If the response variable has k levels, the model contains k - 1 intercepts and effects parameters. The link function for the Multinomial distribution is the multinomial logit. See "Nominal Responses" on page 524 in the "Statistical Details" chapter.

Ordinal Logistic Y has a multinomial distribution with ordinal levels. The response variable must have an ordinal modeling type. The model fits an intercept for each level of the response variable. The effects parameters are common across all levels of the response variable. The link function for the Ordinal Logistic distribution is the ordered logit. See "Ordinal Responses" on page 525 in the "Statistical Details" chapter.

Note: The intercept parameterization for Ordinal Logistic in Generalized Regression differs from that in the Ordinal Logistic personality of Fit Model. The first Intercept term in Generalized Regression corresponds to the first Intercept term in the Ordinal Logistic personality. The subsequent Intercept terms in Generalized Regression are the successive differences between the intercept terms for the ordered levels of the response variable.

Poisson Y has a Poisson distribution with mean λ. The Poisson distribution typically models the number of events in a given interval and is often expressed as count data. The link function for λ is the logarithm. Poisson regression is permitted even if Y assumes non-integer values. See "Statistical Details for Distributions" on page 335.

Negative Binomial Y has a negative binomial distribution with mean μ and dispersion parameter σ. The negative binomial distribution typically models the number of successes before a specified number of failures. The negative binomial distribution is also equivalent to the Gamma Poisson distribution under certain conditions. For more details about the

connection between negative binomial and Gamma Poisson, see the Distributions chapter in the *Basic Analysis* book.

Run demoGammaPoisson.jsl in the JMP Samples/Scripts folder to compare a Gamma Poisson distribution with mean λ and dispersion parameter σ to a Poisson distribution with mean λ.

The link function for μ is the logarithm. Negative binomial regression is permitted even if Y assumes non-integer values. See "Statistical Details for Distributions" on page 335.

JMP PRO Zero-Inflated

ZI Binomial Y has a zero-inflated binomial distribution with parameters *p*, *n*, and zero-inflation parameter π. The response, Y, indicates the total number of successes in *n* independent trials with a fixed probability, *p*, for all trials. This distribution allows for the use of a sample size column. If no column is listed, it is assumed that the sample size is one. The link function for *p* is the logit. See "Statistical Details for Distributions" on page 335.

ZI Beta Binomial Y has a beta binomial distribution with the probability of success, *p*, the number of trails, *n*, overdispersion parameter, δ, and zero-inflation parameter π. This distribution is an overdispersed version of the ZI binomial distribution. The ZI beta binomial distribution requires a sample size greater than one for each observation. Thus, the user must specify a sample size column. To insert a sample size column, specify two continuous columns as Y in this order: the count of the number of successes, and the count of the number of trials. The link function for *p* is the logit. See "Statistical Details for Distributions" on page 335.

ZI Poisson Y has a zero-inflated Poisson distribution with mean parameter λ and zero-inflation parameter π. The parameter λ is the conditional mean based on the observations coming from the Poisson distribution and not the inflating zeros. The link function for λ is the logarithm. ZI Poisson regression is permitted even if Y assumes no observed zeros or non-integer values. See "Statistical Details for Distributions" on page 335.

ZI Negative Binomial Y has a zero-inflated negative binomial with location parameter μ, dispersion parameter σ, and zero-inflation parameter π. The parameter μ is the conditional mean based on the observations coming from the negative binomial distribution and not the inflating zeros. The link function for μ is the logarithm. ZI negative binomial regression is permitted even if Y assumes no observed zeros or non-integer values. See "Statistical Details for Distributions" on page 335.

ZI Gamma Y has a zero-inflated gamma distribution with mean parameter μ and zero-inflation parameter π. Many times, we might believe that our nonzero responses are gamma distributed. This is true for insurance claims: claim values are approximately

gamma distributed but there are also zeros in the data for policies that do not have any claims. The zero-inflated gamma could handle such data directly without having to split the data into zero and nonzero responses. The parameter μ is the conditional mean based on observations coming from the gamma distribution and not the inflating zeros. The link function for μ is the logarithm. See "Statistical Details for Distributions" on page 335.

Table 6.1 gives the Data Types, Modeling Types, and other requirements for Y variables assigned the various distributions.

Table 6.1 Requirements for Y for Distributions

Distribution	Data Type	Modeling Type	Other
Normal	Numeric	Continuous	
Cauchy	Numeric	Continuous	
Exponential	Numeric	Continuous	Positive
Gamma	Numeric	Continuous	Positive
Weibull	Numeric	Continuous	Positive
LogNormal	Numeric	Continuous	Positive
Beta	Numeric	Continuous	Between 0 and 1
Quantile Regression	Numeric	Continuous	
Cox Proportional Hazards	Numeric	Continuous	Nonnegative
Binomial, unsummarized	Any	Any	Binary
Binomial, summarized with Freq column	Any	Any	Binary
Binomial, summarized with count column entered as second Y	Numeric	Continuous	Nonnegative
Beta Binomial	Numeric	Continuous	Nonnegative
Multinomial	Any	Ordinal or Nominal	
Ordinal Logistic	Any	Ordinal	
Poisson	Numeric	Any	Nonnegative
Negative Binomial	Numeric	Any	Nonnegative

Table 6.1 Requirements for Y for Distributions *(Continued)*

Distribution	Data Type	Modeling Type	Other
Zero-Inflated Binomial	Numeric	Any	Nonnegative
Zero-Inflated Beta Binomial	Numeric	Any	Nonnegative
Zero-Inflated Poisson	Numeric	Any	Nonnegative
Zero-Inflated Negative Binomial	Numeric	Any	Nonnegative
Zero-Inflated Gamma	Numeric	Continuous	Nonnegative

For more information on how these distributions are parameterized, see "Statistical Details for Distributions" on page 335. Table 6.2 summarizes the details.

Table 6.2 Distributions, Parameters, and Link Functions

Distribution	Parameters	Mean Model Link Function
Normal	μ, σ	$Identity(\mu)$
Cauchy	μ, σ	$Identity(\mu)$
Exponential	μ	$Log(\mu)$
Gamma	μ, σ	$Log(\mu)$
Weibull	μ, σ	$Identity(\mu)$
LogNormal	μ, σ	$Identity(\mu)$
Beta	μ	$Logit(\mu)$
Binomial	n, p	$Logit(p)$
Beta Binomial	n, p, δ	$Logit(p)$
Poisson	λ	$Log(\lambda)$
Negative Binomial	μ, σ	$Log(\mu)$

Table 6.2 Distributions, Parameters, and Link Functions *(Continued)*

Distribution	Parameters	Mean Model Link Function
Zero-Inflated Binomial	n, p, π (zero-inflation)	$Logit(p)$
Zero-Inflated Beta Binomial	n, p, δ, π (zero-inflation)	$Logit(p)$
Zero-Inflated Poisson	λ, π (zero-inflation)	$Log(\lambda)$
Zero-Inflated Negative Binomial	μ, σ, π (zero-inflation)	$Log(\mu)$
Zero-Inflated Gamma	μ, σ, π (zero-inflation)	$Log(\mu)$

After selecting an appropriate Distribution, click **Run**. The Generalized Regression report window appears.

JMP PRO Generalized Regression Report Window

When you click **Run** in the Fit Model launch window, the Generalized Regression report window that appears contains the following items:

- A Model Launch control panel for fitting models. See "Model Launch Control Panel" on page 301. As you fit models, outlines are added with titles that describe the types of models that you have fit. See "Model Fit Reports" on page 312 and "Model Fit Options" on page 323.
  - If there are linear dependencies among the model terms, the Model Launch control panel contains a Singularity Details report that shows the linear functions that the model terms satisfy. See "Models with Linear Dependencies among Model Terms" on page 199 in the "Standard Least Squares Report and Options" chapter.
- (Shown only if there are no linear dependencies among the model terms and there are not more predictors than observations.) A Maximum Likelihood report that shows the results of a model that was fit using maximum likelihood estimation. The Maximum Likelihood method is labeled Standard Least Squares when the specified Distribution is Normal. The Maximum Likelihood method is labeled Logistic Regression when the specified Distribution is Binomial. See "Maximum Likelihood" on page 302.

JMP PRO Generalized Regression Report Options

Model Dialog Shows the completed Fit Model launch window for the current analysis.

Set Random Seed Sets the seed for the randomization process used for KFold and Holdback validation. This is useful if you want to reproduce an analysis. Set the seed to a positive value, save the script, and the seed is automatically saved in the script. Running the script always produces the same cross validation analysis.

Save Coding Table Creates a new data table whose first columns show the JMP coding for all model parameters. The last column shows the values of the response variable. If you specified a sample size column for a binomial response, both the response and sample size columns are shown in the table. If you used two response columns to specify censoring, both response columns are shown in the table.

Note: The coding data table contains a table variable called Original Data that gives the name of the data table that was used for the analysis. In the case where a By variable is specified, the Original Data table variable also gives the By variable and its level.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

JMP PRO Model Launch Control Panel

The Model Launch control panel provides options for the following:

JMP PRO Estimation Method Options

The available estimation methods can be grouped into techniques with no selection and no penalty, step-based model selection techniques, and penalized regression techniques.

The Maximum Likelihood, Standard Least Squares, and Logistic Regression methods fit the entire model that is specified in the Fit Model launch window. No variable selection is performed. These models can serve as baselines for comparison to other methods.

Note: Only one of Maximum Likelihood, Standard Least Squares, and Logistic Regression is available for a given report. The name of this estimation method depends on the Distribution specified in the Fit Model launch window.

The Backward Elimination, Forward Selection, Pruned Forward Selection, Best Subset, and Two Stage Forward Selection methods are based on variables entering or leaving the model at each step. However, they do not impose a penalty on the regression coefficients.

The Dantzig Selector, Lasso, Elastic Net, Ridge, and Double Lasso methods are penalized regression techniques. They shrink the size of regression coefficients and reduce the variance of the estimates, in order to improve predictive ability of the model. By default, Generalized Regression fits adaptive versions of the Lasso and Elastic Net.

Note: When your data are highly collinear, the adaptive versions of Lasso and Elastic Net might not provide good solutions. This is because the adaptive versions presume that the MLE provides a good estimate. Uncheck the Adaptive option in such cases.

Two types of penalties are used in these techniques:

- the l_1 penalty, which penalizes the sum of the *absolute values* of the regression coefficients
- the l_2 penalty, which penalizes the sum of the *squares* of the regression coefficients

The default Estimation Method for observational data is the Lasso. If the data table contains a DOE script and no singularities, the default Estimation Method is Forward Selection with the Effect Heredity option enabled. If the data table contains a DOE script and a singularity in the design matrix, the default Estimation Method is Two-Stage Forward Selection with the Effect Heredity option enabled.

The following methods are available for model fitting:

Estimation Methods with No Selection and No Penalty

Maximum Likelihood (Not available when the specified Distribution is Normal or Binomial.) Computes maximum likelihood estimates (MLEs) for model parameters. No penalty is imposed. Maximum Likelihood is the only estimation method available for Quantile Regression. A maximum likelihood model report appears by default, unless there are linear dependencies among the predictors or there are more predictors than

observations. If you specified a Validation column in the Fit Model launch window, the maximum likelihood model is fit to the Training set.

The Maximum Likelihood option gives you a way to construct classical models for the response distributions supported by the Generalized Regression personality. In addition, a model based on maximum likelihood can serve as a baseline for model comparison.

Standard Least Squares (Available only when the specified Distribution is Normal.) When the Normal distribution is specified, the Maximum Likelihood estimation method is replaced with the Standard Least Squares estimation method. The default report is a Standard Least Squares report that gives the usual standard least squares results.

Logistic Regression (Available only when the specified Distribution is Binomial.) When the Binomial distribution is specified, the Maximum Likelihood estimation method is replaced with the Logistic Regression estimation method. The default report is a Logistic Regression report. The logistic results are identical to maximum likelihood results.

Step-Based Estimation Methods

Backward Elimination Computes parameter estimates using backward elimination regression. The model chosen provides the best solution relative to the selected Validation Method. Backward elimination starts by including all parameters in the model and removing one effect at each step until reaching the intercept-only model. At each step, the Wald tests for each parameter is used to determine which parameter is removed.

Caution: The horizontal axis of the Solution Path for Backward Elimination is the reverse of the same axis in other estimation methods. Therefore, as you move left to right in the Solution Path for the Backward Elimination estimation method, terms are being removed from the model, rather than added.

Forward Selection Computes parameter estimates using forward stepwise regression. At each step, the effect with the most significant score test is added to the model. The model chosen is the one that provides the best solution relative to the selected Validation Method.

When there are interactions and the Effect Heredity option is enabled, compound effects are handled in the following manner. If the effect with the most significant score test at a given step is one that would violate effect heredity, then a compound effect is created. The compound effect contains the effect with the most significant score test as well as any other inactive effects that are needed to satisfy effect heredity. If the compound effect has the most significant score test, then all of the effects in the compound effect are added to the model.

Pruned Forward Selection Computes parameter estimates using a mixture of forward and backward steps. The algorithm starts with an intercept-only model. At the first step, the

effect with the most significant score test is added to the model. After the first step, the algorithm considers the following three possibilities at each step:

1. From the effects not in the model, add the effect that has the most significant score test.
2. From the effects in the model, remove the effect that has the least significant Wald test.
3. Do both of the above actions in a single step.

To choose the action taken at each step, the algorithm uses the specified Validation Method. For example, if the Validation Method is BIC, the algorithm chooses the action that results in the smallest BIC value. When there are interactions and the Effect Heredity option is enabled, compound effects are considered for adding effects, but they are not considered for removing effects.

When the model becomes saturated, the algorithm attempts a backward step to check if that improves the model. The maximum number of steps in the algorithm is 5 times the number of parameters. The model chosen is the one that provides the best solution relative to the selected Validation Method.

Pruned Forward Selection is an alternative to the Mixed Step option in the Stepwise Regression personality. However, it does not use the *p*-value to determine which variables enter or leave the model.

Tip: The Early Stopping option is not recommended for the Pruned Forward Selection Estimation Method.

Best Subset Computes parameter estimates by increasing the number of active effects in the model at each step. In each step, the model is chosen among all possible models with a number of effects given by the step number. The values on the horizontal axes of the Solution Path plots represent the number of active effects in the model. Step 0 corresponds to the intercept-only model. Step 1 corresponds to the best model of the ones that contain only one active effect. The steps continue up to the value of Max Number of Effects specified in the Advanced Controls in the Model Launch report. See "Advanced Controls" on page 307.

Tip: The Best Subset Estimation Method is computationally intensive. It is not recommended for large problems.

Two Stage Forward Selection (Available only when there are second- or higher-order effects in the model.) Computes parameter estimates in two stages. In the first stage, a forward stepwise regression model is run on the main effects to determine which to retain in the model. In the second stage, a forward stepwise regression model is run on all of the

higher-order effects that are composed entirely of the main effects chosen in the first stage. This method assumes strong effect heredity.

Terms that are not retained from the first stage still appear in the Parameter Estimates reports as zeroed terms; however, they are ignored in the fitting of the second stage model. Terms that are selected in the first stage are not forced into the second stage; they are available for selection in the second stage.

Penalized Estimation Methods

Dantzig Selector (Available only when the specified Distribution is Normal and the No Intercept option is not selected.) Computes parameter estimates by applying an l_1 penalty using a linear programming approach. See Candes and Tao (2007). The Dantzig Selector is useful for analyzing the results of designed experiments. For orthogonal problems, the Dantzig Selector and Lasso give identical results. For more information, see "Dantzig Selector" on page 333.

Lasso Computes parameter estimates by applying an l_1 penalty. Due to the l_1 penalty, some coefficients can be estimated as zero. Thus, variable selection is performed as part of the fitting procedure. In the ordinary Lasso, all coefficients are equally penalized.

Adaptive Lasso Computes parameter estimates by penalizing a weighted sum of the absolute values of the regression coefficients. The weights in the l_1 penalty are determined by the data in such as way as to guarantee the oracle property (Zou 2006). This option uses the MLEs to weight the l_1 penalty. MLEs cannot be computed when the number of predictors exceeds the number of observations or when there are strict linear dependencies among the predictors. If MLEs for the regression parameters cannot be computed, a generalized inverse solution or a ridge solution is used for the l_1 penalty weights. See "Adaptive Methods" on page 334.

The Lasso and the adaptive Lasso options generally choose parsimonious models when predictors are highly correlated. These techniques tend to select only one of a group of correlated predictors. High-dimensional data tend to have highly correlated predictors. For this type of data, the Elastic Net might be a better choice than the Lasso. For more information, see "Lasso Regression" on page 333.

Elastic Net Computes parameter estimates by applying both an l_1 penalty and an l_2 penalty. The l_1 penalty ensures that variable selection is performed. The l_2 penalty improves predictive ability by shrinking the coefficients as ridge does.

Adaptive Elastic Net Computes parameter estimates using an adaptive l_1 penalty as well as an l_2 penalty. This option uses the MLEs to weight the l_1 penalty. MLEs cannot be computed when the number of predictors exceeds the number of observations or when there are strict linear dependencies among the predictors. If MLEs for the

regression parameters cannot be computed, a generalized inverse solution or a ridge solution is used for the l_1 penalty weights. You can set a value for the Elastic Net Alpha in the Advanced Controls panel. See "Adaptive Methods" on page 334.

The Elastic Net tends to provide better prediction accuracy than the Lasso when predictors are highly correlated. (In fact, both Ridge and the Lasso are special cases of the Elastic Net.) In terms of predictive ability, the adaptive Elastic Net often outperforms both the Elastic Net and the adaptive Lasso. The Elastic Net has the ability to select groups of correlated predictors and to assign appropriate parameter estimates to the predictors involved. For more information, see "Elastic Net" on page 333.

Note: If you select an Elastic Net fit and set the Elastic Net Alpha to missing, the algorithm computes the Lasso, Elastic Net, and Ridge fits, in that order. If a fit is time intensive, a progress bar appears. When you click Accept Current Estimates, the calculation stops and the reported parameter estimates correspond to the best model fit at that point. The progress bar indicates when the algorithm is fitting Lasso, Elastic Net, and Ridge. You can use this information to decide when to click Accept Current Estimates.

Ridge Computes parameter estimates using ridge regression. Ridge regression is a biased regression technique that applies an l_2 penalty and does not result in zero parameter estimates. It is useful when you want to retain all predictors in your model. For more details, see "Ridge Regression" on page 332.

Double Lasso Computes parameter estimates in two stages. In the first stage, a Lasso model is fit to determine the terms to be used in the second stage. In the second stage, a Lasso model is fit using the terms from the first stage. The Solution Path results and the parameter estimate reports that appear are for the second-stage fit. If none of the variables enters the model in the first stage, there is no second stage, and the results of the first stage appear in the report.

The Double Lasso is especially useful when the number of observations is less than the number of predictors. By breaking the variable selection and shrinkage operations into two stages, the Lasso in the second stage is less likely to overly penalize the terms that should be included in the model. The double lasso is similar to the relaxed lasso, which is described in Hastie et al. (2009, p. 91).

Adaptive Double Lasso Computes parameter estimates in two stages. In the first stage, an adaptive Lasso model is fit to determine the terms to be used in the second stage. In the second stage, an adaptive Lasso model is fit using the terms from the first stage. The second stage considers only the terms that are included in the first stage model and uses weights based on the parameter estimates in the first stage. You can choose the method of calculating the weights using the Adaptive Penalty Weights option in the Advanced Controls. See "Advanced Control Options" on page 308. The results that are

shown are for the second-stage fit. If none of the variables enters the model in the first stage, there is no second stage, and the results of the first stage appear in the report. See "Adaptive Methods" on page 334.

JMP PRO Advanced Controls

Use the Advanced Controls options to adjust various aspects of the model fitting process. A number of controls relate to the grid for the tuning parameter.

Tuning Parameter

The solution paths for the Lasso and Ridge Estimation Methods depend on a single tuning parameter. The solution path for the Elastic Net depends on a tuning parameter for the penalty on the likelihood as well as the Elastic Net Alpha. The penalty on the likelihood for the Elastic Net is a weighted sum of the penalties associated with the Lasso and Ridge Estimation Methods. The Elastic Net Alpha determines the weights of these two penalties. See "Statistical Details for Estimation Methods" on page 332 and "Statistical Details for Advanced Controls" on page 334.

When the tuning parameter is zero, the solution is unpenalized and maximum likelihood estimates are obtained. As the tuning parameter increases, the penalty increases.

The solution is the set of parameter estimates that minimizes the penalized negative log-likelihood function relative to the selected validation method. The current solution is designated by the solid red vertical line in the Solution Path Plots.

Note: The value of the tuning parameter increases as the Magnitude of Scaled Parameter Estimates in the Solution Path Plot decreases. Estimates close to the MLE are associated with large magnitudes and estimates that are heavily penalized are associated with small magnitudes.

It is important to be mindful of the following:

- When the tuning parameter is too small, the data are typically overfit and result in models with high variance.
- When the tuning parameter is too large, the data are typically underfit.

The Tuning Parameter Grid

To obtain a solution, the tuning parameter is increased over a fine grid.

- For the Lasso, Elastic Net with Elastic Net Alpha specified, and Ridge, the value of the tuning parameter that gives the solution is the one that provides the best fit over the grid of tuning parameters.

Note: Elastic Net Alpha is set to 0.99 by default.

- If you do not set a value for the Elastic Net Alpha, the value of alpha is also increased over a fine grid. For a fixed value of the tuning parameter, alpha is varied until ten consecutive values of alpha fail to improve upon the best fit as determined by the validation method. This process is repeated for the entire grid of tuning parameter values. The final values of the tuning parameter and alpha are the values that provide the best fit over the grid of tuning parameters.

The grid of tuning parameter values ranges from zero, in most cases, to the smallest value for which all of the non-intercept terms are zero. Define the smallest value of the tuning parameter for which all non-intercept terms are zero to be its *upper bound*. The *lower bound* for the tuning parameter is zero except in the following two cases where it is set to 0.0001:

- If the design matrix is singular, the maximum likelihood estimates cannot be computed. The lower bound of 0.0001 allows estimates close to the MLEs to be computed.
- If the selected distribution is binomial or multinomial, the lower bound of 0.0001 helps prevent separation.

Advanced Control Options

Enforce effect heredity Requires lower-order effects to enter the model before their related higher order effects. In most cases, this means that X^2 is not in the model unless X is in the model. For estimation methods other than Forward Selection, however, it is possible for X^2 to enter the model and X to leave the model in the same step. If the data table contains a DOE script, this option is enabled, but it is off by default.

Elastic Net Alpha Sets the α parameter for the Elastic Net. This α parameter determines the mix of the l_1 and l_2 penalty tuning parameters in estimating the Elastic Net coefficients. The default value is α = 0.99, which sets the coefficient on the l_1 penalty to 0.99 and the coefficient on the l_2 penalty to 0.01. This option is available only when Elastic Net is selected as the Estimation Method. See "Statistical Details for Estimation Methods" on page 332.

Number of Grid Points Specifies the number of grid points between the lower and upper bounds for the tuning parameter. At each grid point value, parameter estimates for that value of the tuning parameter are obtained. The default value is 150 grid points.

Minimum Penalty Fraction Indicates the minimum value for the ratio of the lower bound of the tuning parameter to its upper bound. When the lower bound for the tuning parameter is 0, the solution provides the MLE. In cases where you do not want to include the MLE or solutions very close to it, you can set the Minimum Penalty Fraction to a nonzero value. For the Double Lasso estimation method, the specified value of this option

is used only in the first stage of the fit. When there is a singularity in the design matrix, the default value is 0.0001. Otherwise, the default value is 0.

Grid Scale Provides options for choosing the distribution of the grid scale. You can choose between a linear, square root, or log scale. Grid points equal in number to the specified Number of Grid Points are distributed according to the selected scale between the lower and upper bounds of the tuning parameter. The default grid scale is square root. See "Statistical Details for Advanced Controls" on page 334.

First Stage Solution Provides options for choosing the solution in the first stage of the Double Lasso and Two Stage Forward Selection. By default, the solution that is the best fit according to the specified Validation Method is selected and is the solution initially shown (Best Fit). You can choose to initially display models with larger or smaller l_1 norm values that lie in the green or yellow zones. For example, if you choose Smallest in Yellow Zone, the initially displayed solution is the model in the yellow zone that has the smallest l_1 norm. See "Comparable Model Zones" on page 318.

Max Number of Effects Specifies the maximum number of effects to consider in models for the Best Subset estimation method. You can use this to limit the number of computations needed to fit the model. The default value is 10.

Initial Displayed Solution Provides options for choosing the solution that is initially displayed as the current model in the Solution Path report. The current model is identified by a solid vertical line. See "Current Model Indicator" on page 316. The best fit solution is identified by a dotted vertical line. By default, the displayed solution is the one that is considered the best fit according to the specified Validation Method.

You can choose to initially display models with larger or smaller l_1 norm values that still lie in the green or yellow zones. For example, if you choose Smallest in Yellow Zone, the initially displayed solution is the model in the yellow zone that has the smallest l_1 norm. See "Comparable Model Zones" on page 318.

Adaptive Penalty Weights Provides options for the calculation of the penalty weights that are used in the second stage of the Adaptive Double Lasso. By default, the Inverse Solution option is selected. This option calculates the penalty weights using the parameter estimates from the first stage fit.

The Inverse Model Average option calculates the penalty weights using the parameter estimates from a solution that is the weighted average of the AICc or BIC models. The AICc models are used if the AICc Validation Method is selected. Otherwise, the BIC models are used. If you use the Inverse Model Average option, the maximum likelihood solution, if it exists, appears as the right-most point in the Solution Path for the Adaptive Double Lasso model.

Force Terms Enables you to select which terms, if any, you want to force into the model. The terms that are forced into the model are not included in the penalty.

JMP PRO Validation Method Options

The following methods are available for validation of the model fit.

Note: The only Validation Method allowed for Quantile Regression is None. The only Validation Methods allowed for the Maximum Likelihood Estimation Method are None and Validation Column. The only Validation Methods allowed for Cox Proportional Hazards are BIC, AICc, and None. The only Validation Methods allowed for the Dantzig Selector Estimation Method are BIC and AICc.

KFold For each value of the tuning parameter, the following steps are conducted:

- The observations are partitioned into k subsets, or *folds*.
- In turn, each fold is used as a validation set. A model is fit to the observations *not* in the fold. The log-likelihood based on that model is calculated for the observations *in* the fold, providing a *validation* log-likelihood.
- The mean of the validation log-likelihoods for the k folds is calculated. This value serves as a validation log-likelihood for the value of the tuning parameter.

The value of the tuning parameter that has the maximum validation log-likelihood is used to construct the final solution. To obtain the final model, all k models derived for the optimal value of the tuning parameter are fit to the entire data set. Of these, the model that has the highest validation log-likelihood is selected as the final model. The training set used for that final model is designated as the Training set and the holdout fold for that model is the Validation set. These are the Training and Validation sets used in plots and in the reported results for the final solution.

Holdback Randomly selects the specified proportion of the data for a validation set, and uses the other portion of the data to fit the model. The final solution is the one that minimizes the negative log-likelihood for the validation set. This method is useful for large data sets.

Leave-One-Out Performs leave-one-out cross validation. This is equivalent to KFold, with the number of folds equal to the number of rows. This option should not be used on moderate or large data sets. It can require long processing time for even a moderate number of observations. The Training and (one-row) Validation sets used in plots and in the reported results for the final solution are determined as is done for KFold validation.

BIC Minimizes the Bayesian Information Criterion (BIC) over the solution path. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

AICc Minimizes the corrected Akaike Information Criterion (AICc) over the solution path. AICc is the default setting for Validation Method. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

Note: The AICc is not defined when the number of parameters approaches or exceeds the sample size.

ERIC Minimizes the Extended Regularized Information Criterion (ERIC) over the solution path. See "Model Fit Detail" on page 313. Available only for exponential family distributions and for the Lasso and adaptive Lasso estimation methods.

None Does not use validation. Available only for the Maximum Likelihood Estimation Method and Quantile Regression.

Validation Column Uses the column specified in the Fit Model window as having the Validation role. The final solution is the one that minimizes the negative log-likelihood for the validation set. This option is not available when the specified Estimation Method is Dantzig Selector or when the specified Distribution is Quantile Regression or Cox Proportional Hazards.

JMP PRO Early Stopping

Early Stopping adds an early stopping rule:

- For Forward Selection, the algorithm terminates when 10 consecutive steps of adding variables to the model fail to improve upon the validation measure. The solution is the model at the step that precedes the 10 consecutive steps.
- For Lasso, Elastic Net, and Ridge, the algorithm terminates when 10 consecutive values of the tuning parameter fail to improve upon the best fit as determined by the validation method. The solution is the estimate corresponding to the tuning parameter value that precedes the 10 consecutive values.

Note: For the AICc and BIC validation methods, early stopping does not occur until at least four predictors have entered the model.

JMP PRO Go

When you click **Go**, a report opens. The title of the report specifies the fitting and validation methods that you selected. You can return to the Model Launch control panel to perform additional analyses and choose other estimation and validation methods.

JMP PRO Model Fit Reports

For each Estimation Method and Validation Method that you specify in the Model Launch panel, a report is produced. The report specifies your selected Estimation and Validation methods in its title.

The following reports are presented by default:

- Regression Plot
- Model Summary
- Estimation Details (shown only for Lasso, Elastic Net, and Ridge)
- Solution Path (shown for all but the Maximum Likelihood Estimation Method and Quantile Regression)
- Parameter Estimates for Centered and Scaled Predictors
- Parameter Estimates for Original Predictors
- Effect Tests

JMP PRO Regression Plot

Note: The Regression Plot report appears only when there is one continuous predictor and no more than one categorical predictor. It is not available if the Distribution option is set to Multinomial, Ordinal Logistic, or Cox Proportional Hazards. The response must be continuous.

The Regression Plot report shows a plot of the response values on the vertical axis and the continuous predictor on the horizontal axis. A regression line is shown over the points. If there is a categorical predictor in the model, each level of the categorical predictor has a separate regression line and a legend appears next to the plot. If the response is specified as the counts of the number of successes and the number of trials, the number of successes divided by the number of trials is plotted on the vertical axis.

JMP PRO Model Summary

The Model Summary report describes the model that you have fit and provides summary information about the fit itself.

JMP PRO Model Description Detail

The first part of the Model Summary report gives information that describes the model that you have fit.

Response The column assigned to the Y role in the Fit Model window. When two columns are used to specify interval censoring, both column names are listed.

Distribution The Distribution selected in the Fit Model window. For Quantile Regression, the value of the specified quantile for the response is also displayed.

Estimation Method The Estimation Method selected in the Model Launch panel.

Validation Method The Validation Method selected in the Model Launch panel.

Mean Model Link The link function for the model for the mean, based on the Distribution selected in the Fit Model window.

Location Model Link The link function for the model for the location parameter, shown when Cauchy is selected as the Distribution in the Fit Model window.

Scale Model Link The link function for the model for the scale parameter, based on the Distribution selected in the Fit Model window.

Probability Model Link The link function for the model for the probability, based on the Distribution selected in the Fit Model window.

Dispersion Model Link The link function for the model for the dispersion parameter, based on the Distribution selected in the Fit Model window.

Zero Inflation Model Link The link function for the model for the zero inflation parameter, based on the Distribution selected in the Fit Model window.

JMP PRO Model Fit Detail

The second part of the Model Summary report gives statistics related to the model fit. If either Holdback or Validation Column is selected as the Validation Method, these statistics are computed separately for the training and validation sets. This part of the Model Summary report is not available if either KFold or Leave-One-Out is selected as the Validation Method.

Number of rows The number of rows.

Sum of Frequencies The sum of the values of a column assigned to the Freq or Weight role in the Fit Model window.

Note: For -LogLikelihood, BIC, AICc, and ERIC, smaller is better. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

-LogLikelihood The negative of the natural logarithm of the likelihood function for the current model.

Note: -LogLikelihood is not available for Quantile Regression.

Objective Function (Available only for Quantile Regression.) The value of the function that is minimized to fit the specified quantile regression model. The function that is minimized is the check-loss function.

Number of Parameters The number of nonzero parameters in the current model.

BIC The Bayesian Information Criterion: $\text{BIC} = -2LogLikelihood + k\ln(n)$.

AICc The corrected Akaike Information Criterion:
$\text{AICc} = -2LogLikelihood + 2k + 2k(k+1)/(n-k-1)$.

ERIC (Available only for exponential family distributions and when the Lasso or adaptive Lasso estimation method is specified.) The Extended Regularization Information Criterion: $\text{ERIC} = -2LogLikelihood + (k-2)\ln(n\phi/\lambda)$ where λ is the value of the tuning parameter and ϕ is the nuisance parameter. See Hui et al. (2015).

Generalized RSquare (Not available for Quantile Regression.) An extension of the RSquare measure that can be applied to general regression models. Generalized RSquare compares the likelihood of the fitted model (L_M) to the likelihood of the intercept-only (constant) model (L_0). It is scaled to have a maximum of 1. For distributions other than Binomial, the Generalized RSquare is defined as follows:

$$\text{Generalized RSquare} = 1 - (L_0/L_M)^{2/n}$$

When Binomial is the specified distribution, the Generalized RSquare is defined as follows:

$$\text{Generalized RSquare} = \frac{1 - (L_0/L_M)^{2/n}}{1 - L_0^{2/n}}$$

A Generalized RSquare value of 1 indicates a perfect model; a value of 0 indicates a model that is no better than a constant model. The Generalized RSquare measure simplifies to the

traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the Nagelkerke or Craig and Uhler R^2, which is a normalized version of Cox and Snell's pseudo R^2. See Nagelkerke (1991).

Note: Generalized RSquare is replaced by RSquare when the Normal distribution is specified.

Caution: You should not compare Generalized RSquare values for models that use different response distributions. The comparison being made is to the intercept-only model with a given response distribution.

RSquare (Available only when the Normal distribution is specified.) Estimates the proportion of variation in the response that can be attributed to the model rather than to random error. An RSquare value of 1 indicates a perfect model; a value of 0 indicates a model that is no better than a constant model. The RSquare value is calculated as follows:

$$1 - \frac{\sum (Y - \hat{Y})^2}{\sum (Y - \bar{Y})^2}$$

RSquare Adj (Available only when the Normal distribution is specified and the estimation method does not involve a penalty.) Adjusts the RSquare statistic for the number of parameters in the model. Rsquare Adj facilitates comparisons among models with different numbers of parameters. The computation uses the degrees of freedom. The RSquare Adj value is calculated as follows:

$$1 - \frac{\frac{1}{N-(p-1)}\sum (Y - \hat{Y})^2}{\frac{1}{N-1}\sum (Y - \bar{Y})^2}$$

where N is the number of observations and p is the number of parameters.

Note: When there is a Validation set, the adjusted RSquare statistic is reported only for the Training set.

RMSE (Available only when the Normal distribution is specified.) The Root Mean Square Error (RMSE) estimates the standard deviation of the random error. This quantity is the square root of the mean of the sum of squared errors in the current model.

Lambda Penalty (Available only for the Dantzig Selector, Lasso, Elastic Net, Ridge, and Double Lasso estimation methods.) The value of the tuning parameter λ for the current model. See "Statistical Details for Estimation Methods" on page 332.

JMP PRO Estimation Details

The Estimation Details report shows the settings of the Advanced Controls for the Dantzig Selector, Lasso, Elastic Net, Ridge, and Double Lasso estimation methods. For more information about these controls, see "Advanced Controls" on page 307.

JMP PRO Solution Path

Note: The Solution Path report appears for all Estimation Methods except Maximum Likelihood and all Distributions except Quantile Regression.

The Solution Path report shows two plots:

- The Solution Path Plot displays values of the estimated parameters.
- The Validation Plot displays values of the validation statistic corresponding to the selected validation method.

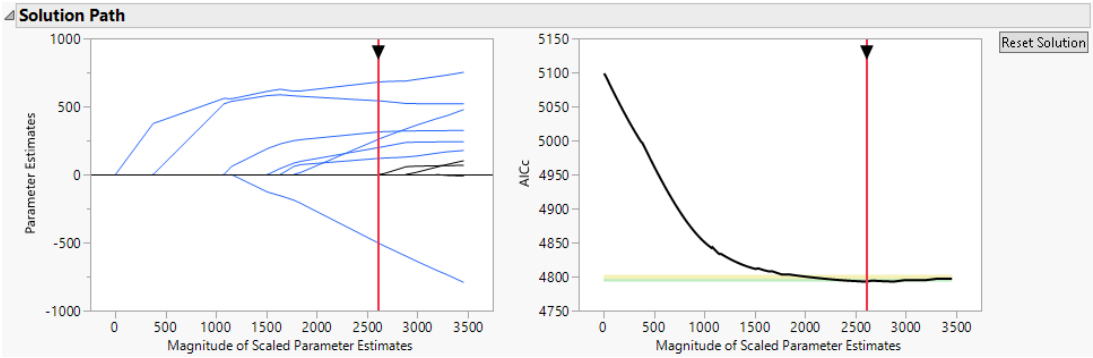
The horizontal scaling for both plots is given in terms of the Magnitude of Scaled Parameter Estimates. This is the l_1 norm, defined as the sum of the absolute values of the scaled parameter estimates for the model for the mean. (Estimates corresponding to the intercept, dispersion parameters, and zero-inflation parameters are excluded from the calculation of the l_1 norm.) Note the following:

- Estimates with large values of the l_1 norm are close to the MLE.
- Estimates with small values of the l_1 norm are heavily penalized.
- The value of the tuning parameter increases as the l_1 norm decreases.

JMP PRO Current Model Indicator

A solid vertical red line is placed in both plots at the value of the l_1 norm for the solution displayed in the Parameter Estimates for Original Predictors report. You can drag the arrow at the top of the vertical red line in either plot to change the magnitude of the penalty, indicating a new current model. In the Validation Plot, you can also click anywhere in the plot to change the model. As you drag the vertical red line to indicate a new model, the results in the report update to reflect the currently selected model. A dashed vertical line remains at the best fit model. You can click the **Reset Solution** button next to the Validation Plot to return the vertical red line and corresponding results to the initial solution. For some validation methods, the Validation Plot provides zones that identify comparable models. See "Comparable Model Zones" on page 318.

Figure 6.5 Solution Path Report for Diabetes.jmp, Lasso with AICc Validation



For more information about the Solution Path Plot, see "Solution Path Plot" on page 317. For more information about the Validation Plot, see "Validation Plot" on page 318.

JMP PRO Solution Path Plot

You can select paths in the Solution Path Plot to highlight the corresponding terms in the Parameter Estimates reports. This action also selects the corresponding columns in the data table. Selecting rows in either of the reports highlights the corresponding rows in the other report and the corresponding paths in the Solution Path Plot. Press Shift and click to select multiple paths or rows.

The Parameter Estimates are plotted using the vertical axis of the Solution Path Plot. These are the *scaled* parameter estimates. They are derived for a model expressed in terms of centered and scaled predictors (see "Parameter Estimates for Centered and Scaled Predictors" on page 320).

When the number of predictors is less than the number of observations, the Solution Path Plot usually shows the entire range of estimates from zero to the unpenalized fit given by the MLE. Otherwise, the plot extends to a magnitude that is close to the unpenalized solution. This occurs when the jump from the next-to-last grid point to the MLE solution is so large that the detail for solutions up to the next-to-last grid point is obscured. When this happens, as long as the MLE is not the final solution, the Solution Path Plot is rescaled so that the axis extends only to the next-to-last grid point.

JMP PRO The Solution ID

Internally, each solution in the Solution Path is assigned a Solution ID. When you adjust the tuning parameter to select a solution other than the one initially presented, the corresponding Solution ID appears in scripts created by the Save Script options. The Solution ID is the value N in the `Set Solution ID( N )` command. Saving the Solution ID ensures that you can re-create your selected solution when you run the script.

JMP PRO Validation Plot

The Validation Plot shows plots of statistics that describe how well models fit across the values of the tuning parameter, or equivalently, across the values of the Magnitude of the Scaled Parameter Estimates. The statistics plotted depend on the selected Validation Method. For each Validation Method, Table 6.3 lists the statistic that is plotted. For all validation methods, smaller values are better. For the KFold and Leave-One-Out validation methods, and for a Validation Column with more than three values, the statistic that is plotted is the mean of the scaled negative log-likelihood values across the folds.

The *Scaled -LogLikelihood* in Table 6.3 is the negative log-likelihood divided by the number of observations in the set for which the negative log-likelihood is computed.

Table 6.3 Validation Methods and Corresponding Validation Statistics and Zones

Validation Method	Validation Statistic	Tuning Parameter Regions
KFold	Mean of the Scaled -LogLikelihood values across the K folds	Two
Holdback	Scaled -LogLikelihood	None
Leave-One-Out	Mean of the Scaled -LogLikelihood values across all folds	Two
BIC	BIC for training data	Two
AICc	AICc for training data	Two
ERIC	ERIC for training data	Two
Validation Column with two or three values	Scaled -LogLikelihood	None
Validation Column with $K > 3$ values	Mean of the Scaled -LogLikelihood values across the K folds	Two

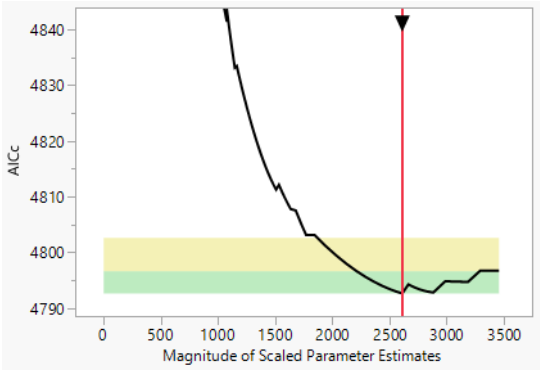
JMP PRO Comparable Model Zones

Although a model is estimated to be the best model, there can be uncertainty relative to this selection. Competing models might fit nearly as well and can contain useful information. For the AICc, BIC, KFold, and Leave-One-Out validation methods, and for a Validation Column with more than three values, the Validation Plot provides zones that identify competing models that might deserve consideration. Models that fall outside the zones are not recommended. See Burnham and Anderson (2004) and Burnham et al. (2011).

A zone is an interval of values of the validation statistics. The zones are plotted as green or yellow rectangles that span the entire horizontal axis. A model falls in a zone if the value of its validation statistic falls in the zone. You can drag the solid vertical red line to explore solutions within the zones. See "Current Model Indicator" on page 316.

Figure 6.6 shows a Validation Plot for Diabetes.jmp with the vertical axis expanded to show the two zones.

Figure 6.6 Validation Plot for Diabetes.jmp, Lasso with AICc Validation



Zones for BIC, AICc, and ERIC Validation

For these validation methods, two regions are shown in the plot. Denote the validation BIC, AICc, and ERIC values for the best solutions by V^{best}.

- The green zone identifies models for which there is strong evidence that a model is comparable to the best model. The green zone is the interval $[V^{best}, V^{best}+4]$.
- The yellow zone identifies models for which there is weak evidence that a model is comparable to the best model. The yellow zone is the interval $(V^{best}+4, V^{best}+10]$.

Zones for KFold Validation, Leave-One-Out Validation, and Validation Column with More Than Three Values

For these validation methods, two regions are shown in the plot. At the solution for the best model, the scaled negative log-likelihood functions are evaluated for each validation set. Denote the standard error of these values as L^{SE}. Denote the scaled negative log-likelihood for the best solution by L^{best}.

- The green zone identifies models for which there is strong evidence that a model is comparable to the best model. The green zone is the interval $[L^{best}, L^{best}+L^{SE}]$.
- The yellow zone identifies models for which there is weak evidence that a model is comparable to the best model. The yellow zone is the interval $(L^{best}+L^{SE}, L^{best}+2.5*L^{SE}]$.

JMP PRO Parameter Estimates for Centered and Scaled Predictors

The Parameter Estimates for Centered and Scaled Predictors report gives estimates and other results for all parameters in the model. The initial table includes the coefficients for the predictors in the model. An additional table includes other model parameters such as scale, dispersion, or zero inflation parameters. See "Distribution" on page 292. Both tables include the same columns of results.

Tip: You can click on terms in the Parameter Estimates for Centered and Scaled Predictors report to highlight the corresponding paths in the Solution Path Plot. The corresponding columns in the data table are also selected. This is useful in terms of running further analyses. Press Shift and click the terms to select multiple rows.

For all fits in the Generalized Regression personality, every predictor is centered to have mean zero and scaled to have sum of squares equal to one:

- The mean is subtracted from each observation.
- Each difference is then divided by the square root of the sum of the squared differences from the mean.

This puts all predictors on an equal footing relative to the penalties applied.

Note: When the No Intercept option is selected in the launch window, the predictors are not centered and scaled.

The Parameter Estimates for Centered and Scaled Predictors report gives parameter estimates for the model expressed in terms of the centered and scaled predictors. The estimates are determined by the Validation Method that you specified. The estimates are depicted in the Solution Path Plots by a vertical red line.

The report provides the following information:

Term A list of the model terms. "Forced in" appears next to any terms that were forced into the model using the Advanced Controls option.

Estimate The parameter estimate corresponding to the centered and scaled model term.

Std Error The standard error of the estimate. This is obtained using M-estimation and a sandwich formula (Zou 2006; Huber and Ronchetti 2009).

Wald ChiSquare The ChiSquare value for a Wald test of whether the parameter is zero.

Prob > ChiSquare The *p*-value for the Wald test.

Lower 95% The lower bound for a 95% confidence interval for the parameter. You can change the α level in the Fit Model window by selecting Set Alpha Level from the red triangle menu.

Upper 95% The upper bound for a 95% confidence interval for the parameter. You can change the α level in the Fit Model window by selecting Set Alpha Level from the red triangle menu.

Singularity Details (Available only if there are linear dependencies among the model terms.) The linear function that the model term satisfies.

JMP PRO Parameter Estimates for Original Predictors

The Parameter Estimates for Original Predictors report gives estimates and other results for all parameters in the model. The initial table includes the coefficients for the predictors in the model. An additional table includes other model parameters such as scale, dispersion, or zero inflation parameters. See "Distribution" on page 292. Both tables include the same columns of results.

Tip: You can select terms in the Parameter Estimates for Original Predictors report to highlight the corresponding paths in the Solution Path Plot. The corresponding columns in the data table are also selected. This is useful when running further analyses. Press Shift and click the terms to select multiple rows.

The Parameter Estimates for Original Predictors report gives parameter estimates for the model expressed in terms of the original (uncentered and unscaled) predictors.

The report provides the following information:

Term A list of the model terms. "Forced in" appears next to any terms that were forced into the model using the Advanced Controls option.

Estimate The parameter estimate corresponding to the model term given in terms of the original measurements.

Std Error The standard error of the estimate. This is obtained using M-estimation and a sandwich formula (Zou 2006 and Huber and Ronchetti 2009).

Wald ChiSquare The ChiSquare value for a Wald test of whether the parameter is zero.

Prob > ChiSquare The *p*-value for the Wald test.

Lower 95% The lower bound for a 95% confidence interval for the parameter. You can change the α level in the Fit Model window by selecting Set Alpha Level from the red triangle menu.

Upper 95% The upper bound for a 95% confidence interval for the parameter. You can change the α level in the Fit Model window by selecting Set Alpha Level from the red triangle menu.

Singularity Details (Available only if there are linear dependencies among the model terms.) The linear function that the model term satisfies.

VIF (Available only when the distribution is Normal. Appears only if you right-click in the parameter estimates table and select Columns > VIF.) The variance inflation factor (VIF) for each term in the model. High VIF values indicate a collinearity issue among the terms in the model.

The VIF for the i^{th} term, x_i, is calculated as follows:

$$VIF_i = \frac{1}{1 - R_i^2}$$

where R_i^2 is the RSquare for the regression of x_i as a function of the other explanatory variables.

JMP PRO Active Parameter Estimates

The Active Parameter Estimates report gives a subset of the Parameter Estimates for Original Predictors report. The Active Parameter Estimates report shows only the nonzero parameters.

JMP PRO Effect Tests

The Effect Tests report gives the following information:

Source A list of the effects in the model.

Nparm The number of parameters associated with the effect.

DF The degrees of freedom for the Wald ChiSquare test. This is the number of nonzero parameter estimates associated with the effect in the model.

Wald ChiSquare The ChiSquare value for a Wald test of whether all parameters associated with the effect are zero.

Prob > ChiSquare The *p*-value for the Wald ChiSquare test.

The following columns appear in the report in place of Wald ChiSquare and Prob > ChiSquare when the Distribution is Normal and there is no penalty in the estimation method:

Sum of Squares The sum of squares for the hypothesis that the effect is zero.

F Ratio The F statistic for testing that the effect is zero. The F Ratio is the ratio of the mean square for the effect divided by the mean square for error. The mean square for the effect is the sum of squares for the effect divided by its degrees of freedom.

Prob > F The *p*-value for the effect test.

If the coefficient for an effect has been estimated as zero, then:

- If the effect has one degree of freedom, the word "Removed" appears at the far right in the row for that effect.
- If the effect has multiple degrees of freedom, the phrase "Levels removed" appears, followed by the number of levels that correspond to terms with parameter estimates of zero.

JMP PRO Model Fit Options

Each model fit report has a red triangle menu with these options:

Caution: Many options in the platform are not available if you specify a column that has the Expression data type or Vector modeling type in the launch window.

Regression Reports Enables you to customize the reports that are shown for the specified model fit. All of the following reports are shown by default except for the Parameter Estimates for Centered and Scaled Parameter Estimates report and the Active Parameter Estimates report.

Model Summary Shows or hides the Model Summary report that includes information about the specification and goodness of fit statistics for the model. This option also displays the Estimation Details report for applicable models. See "Model Summary" on page 313 and "Estimation Details" on page 316.

Solution Path (Not available for Maximum Likelihood models.) Shows or hides the Solution Path and Validation Path plots. See "Solution Path" on page 316.

Parameter Estimates for Centered and Scaled Predictors Shows or hides a table of centered and scaled parameter estimates. See "Parameter Estimates for Centered and Scaled Predictors" on page 320.

Parameter Estimates for Original Predictors Shows or hides a table of parameter estimates in the original scale of the data. See "Parameter Estimates for Original Predictors" on page 321.

Active Parameter Estimates (Not available for Maximum Likelihood or Ridge Regression models.) Shows or hides a table of active, or nonzero, parameter estimates for the currently selected model.

Effect Tests Shows or hides tests for each effect. The effect test for a given effect tests the null hypothesis that all parameters associated with that effect are zero. A nominal or ordinal effect can have several associated parameters, based on its number of levels. The effect test for such an effect tests whether all of the associated parameters are zero. When the Distribution is Multinomial, the effects are combined over the levels of the response. See "Effect Tests" on page 322.

Show Prediction Expression Displays the Prediction Expression report that contains the equation for the estimated model. See "Show Prediction Expression" on page 117 in the "Standard Least Squares Report and Options" chapter for an example.

Select Nonzero Terms Highlights terms with nonzero coefficients in the report. Also selects all associated columns in the data table. This option is not available when Ridge Regression is selected as the Estimation Method.

Select Zeroed Terms Highlights terms with zero coefficients in the report. Also selects all associated columns in the data table. This option is not available when Ridge Regression is selected as the Estimation Method.

Relaunch with Active Effects (Not available for responses that have the Vector modeling type.) Opens a Fit Model launch window where the Construct Model Effects list contains only the terms that have nonzero parameter estimates (*active* effects). All other specifications are those used in the original analysis.

Note: If you select Relaunch with Active Effects in a report that contains a By variable, the By variable is not added to the Fit Model launch window.

Hide Inactive Paths Adjusts the transparency of the inactive paths in the Solution Path Parameter Estimates plot so that the paths that are not currently active appear faded.

Odds Ratios (Available only when the specified Distribution is Binomial and the model contains an intercept. Not available for responses that have the Vector modeling type.) Displays a report that contains odds ratios for categorical predictors, and unit odds ratios and range odds ratios for continuous predictors. An *odds ratio* is the ratio of the odds for two events. The *odds* of an event is the probability that the event of interest occurs versus the probability that it does not occur. The event of interest is defined by the Target Level in the Fit Model launch window.

For each categorical predictor, an Odds Ratios report appears. Odds ratios are shown for all combinations of levels of a categorical model term.

If there are continuous predictors, two additional reports appear:

– Unit Odds Ratios Report. The *unit odds ratio* is calculated over a one-unit change in a continuous model term.

– Range Odds Ratios Report. The *range odds ratio* is calculated over the entire range of a continuous model term.

The confidence intervals in the Odds Ratios report are Wald-based intervals. Note that the odds ratio for a model term is meaningful only if the model term is not involved in any higher-order effects.

Note: If there are interactions in the model, you can use the Multiple Comparisons option to obtain odds ratios. See "Multiple Comparisons" on page 326.

Incidence Rate Ratios (Available only when the specified Distribution is Poisson or Negative Binomial and the model contains an intercept.) Displays a report that contains incidence rate ratios for categorical predictors, and unit incidence rate ratios and range incidence rate ratios for continuous predictors. An *incidence rate ratio* is the ratio of the incidence rate for two events. The *incidence rate* for a model term is the number of new events that occur over a given time period.

For each categorical predictor, an Incidence Rate Ratios report appears. Incidence rate ratios are shown for all combinations of levels of a categorical model term.

If there are continuous predictors, two additional reports appear:

– Unit Incidence Rate Ratios Report. The *unit incidence rate ratio* is calculated over a one-unit change in a continuous model term.

– Range Incidence Rate Ratios Report. The *range incidence rate ratio* is calculated over the entire range of a continuous model term.

The confidence intervals in the Incidence Rate Ratios report are Wald-based intervals. Note that the incidence rate ratio for a model term is meaningful only if the model term is not involved in any higher-order effects.

Hazard Ratios (Available only when the specified Distribution is Cox Proportional Hazards and the model contains an intercept.) Displays a report that contains hazard ratios for categorical predictors, and unit hazard ratios and range hazard ratios for continuous predictors. A *hazard ratio* is the ratio of the hazard rate for two events. The *hazard rate* at time t for an event is the conditional probability that the event will not survive an additional amount of time, given that it has survived to time t.

For each categorical predictor, a Hazard Ratios report appears. Hazard ratios are shown for all combinations of levels of a categorical model term.

If there are continuous predictors, two additional reports appear:

– Unit Hazard Ratios Report. The *unit hazard ratio* is calculated over a one-unit change in a continuous model term.

– Range Hazard Ratios Report. The *range hazard ratio* is calculated over the entire range of a continuous model term.

The confidence intervals in the Hazard Ratios report are Wald-based intervals. Note that the hazard ratio for a model term is meaningful only if the model term is not involved in any higher-order effects.

Covariance of Estimates Displays a matrix showing the covariances of the parameter estimates. These are calculated using M-estimation and a sandwich formula (Zou 2006 and Huber and Ronchetti 2009).

Correlation of Estimates Displays a matrix showing the correlations of the parameter estimates. These are calculated using M-estimation and a sandwich formula (Zou 2006 and Huber and Ronchetti 2009).

Inverse Prediction (Not available for responses that have the Vector modeling type.) Predicts an X value, given specific values for Y and the other X variables. This can be used to predict continuous variables only. For more information about Inverse Prediction, see "Inverse Prediction" on page 143 in the "Standard Least Squares Report and Options" chapter.

Multiple Comparisons (Not available for responses that have the Vector modeling type or for models that do not contain any categorical predictors.) Displays the Multiple Comparisons launch window. For more information about the Multiple Comparisons launch window and report, see "Multiple Comparisons" on page 128 in the "Standard Least Squares Report and Options" chapter. Note that the multiple comparisons are performed on the linear predictor scale. When the specified Distribution is Binomial, the multiple comparisons are performed on the odds ratios. When the specified Distribution is Poisson, the multiple comparisons are performed on the incidence rate ratios. When the specified Distribution is Cox Proportional Hazards, the multiple comparisons are performed on the hazard ratios.

Confusion Matrix (Available only when the specified Distribution is Binomial, Multinomial, or Ordinal Logistic.) Displays a matrix that tabulates the actual response levels and the predicted response levels. For a good model, predicted response levels should be the same as the actual response levels. The confusion matrix enables you to assess how the predicted responses align with the actual responses. If you used validation, a confusion matrix is shown for each of the Training, Validation, and Test sets.

Set Probability Threshold (Available only when the specified Distribution is Binomial.) Specify a cutoff probability for classifying the response. By default, an observation is classified into the Target Level when its predicted probability exceeds 0.5. Change the

threshold to specify a value other than 0.5 as the cutoff for classification into the Target Level. The Predicted Rate in the confusion matrix is updated to reflect classification according to the specified threshold.

If the response has a Profit Matrix column property, the initial value for the probability threshold is determined by the profit matrix.

Profilers (Not available for responses that have the Vector modeling type.) Provides various profilers that enable you to explore the fitted model.

Note: When the number of rows is less than or equal to 500 and the number of predictors is less than or equal to 30, the Profiler plots update continuously as you drag the current model indicator in either Solution Path plot. Otherwise, they update when you release the mouse button.

Profiler Displays the Prediction Profiler. Predictors that have parameter estimates of zero and that are not involved in any interaction terms with nonzero coefficients do not appear in the profiler. For details about the prediction profiler, see the Profiler chapter in the *Profilers* book.

Distribution Profiler (Not available when the specified distribution is Binomial or Quantile Regression.) Displays a profiler of the cumulative distribution function of the predictors and the response. The response is shown in the right-most cell.

Quantile Profiler (Not available when the specified distribution is Binomial or Quantile Regression.) Displays a profiler that shows the predicted response as a function of the predictors and the quantile of the cumulative distribution function. The quantile is called Probability and is shown in the right-most cell.

Survival Profiler (Available only when the specified Distribution is Normal, Exponential, Weibull, Lognormal, or Cox Proportional Hazards.) Displays a profiler that shows the survival function as a function of the predictors and the response. The response is shown in the right-most cell.

Hazard Profiler (Available only when the specified Distribution is Normal, Exponential, Weibull, Lognormal, or Cox Proportional Hazards.) Displays a profiler that shows the hazard rate as a function of the predictors and the response. The response is shown in the right-most cell.

Custom Test Displays a Custom Test report that enables you to test a custom hypothesis. If the model has a Solution Path, the custom test results update as you update the solution. For more information about custom tests, see "Custom Test" on page 125 in the "Standard Least Squares Report and Options" chapter. The Custom Test red triangle menu contains an option to remove the Custom Test report.

Diagnostic Plots Provides various plots to help assess how well the current model fits. If a Validation column is specified or if KFold, Holdback, or Leave-One-Out is selected as the Validation Method, the options below enable you to view the training, validation, and, if applicable, test sets, or they construct separate plots for these sets. If KFold or Leave-One-Out is selected, then the plots correspond to the validation set that optimizes prediction error, and its corresponding training set. See "KFold" on page 310.

Note: All Diagnostic plots update continuously as you drag the current model indicator in either Solution Path plot.

Diagnostic Bundle (Not available when the specified Distribution is Binomial, Multinomial, Ordinal Logistic, or Cox Proportional Hazards.) Displays a set of four graphs including a plot of residuals by predicted values, residuals by row number, a histogram of the residuals, and a histogram of the probability of observing a response larger than the observed response.

The graphs are constructed using all observations. If you used a Validation Column or if you selected KFold, Holdback, or Leave-One-Out as the Validation Method, check boxes enable you to select the Training, Validation, and, if applicable, Test sets. Rows corresponding to these sets are selected in the data table and the corresponding points and areas are highlighted in the graphs. Use this option to determine whether the model fit is similar across the sets.

The Fitted Probability of Observing a Larger Response histogram helps you assess goodness of fit of the model. Different criteria apply based on the distribution:

- For distributions other than zero-inflated distributions and quantile regression, the "correct" model should display an approximately uniform distribution of values.
- For a zero-inflated distribution, the histogram should display a point mass at zero and an approximately uniform distribution elsewhere.
- For quantile regression, the histogram should display an approximately uniform distribution of values to the left of the specified quantile and an approximately uniform distribution of slightly higher values to the right of the specified quantile.

Plot Baseline Survival and Hazard (Available only when the specified distribution is Cox Proportional Hazards.) Displays the Baseline Survival and Hazard plots, which plot the survival and hazard functions for the baseline proportional hazards function versus the response variable. Below the plots, there is a table that contains the plotted values.

Note: If the specified Distribution is Cox Proportional Hazards, the Plot Baseline Survival and Hazard option is the only available Diagnostic Plot.

ROC Curve (Available only when the specified Distribution is Binomial, Multinomial, or Ordinal Logistic.) Displays the Receiver Operating Characteristic (ROC) curve. If you used validation, an ROC curve is shown for each of the Training, Validation, and Test sets.

The ROC curve measures the ability of the fitted probabilities to classify response levels correctly. The further the curve from the diagonal, the better the fit. An introduction to ROC curves is found in the Logistic Analysis chapter in the *Basic Analysis* book.

If the response has more than two levels, the ROC Curve plot displays an ROC curve for each response level. For a given response level, this curve is the ROC curve for correct classification into that level. See the Partition chapter in the *Predictive and Specialized Modeling* book for more information about ROC curves.

Lift Curve (Available only when the specified Distribution is Binomial, Multinomial, or Ordinal Logistic.) Displays the lift curve for the model. If you used validation, an ROC curve is shown for each of the Training, Validation, and Test sets.

A lift curve shows how effectively response levels are classified as their fitted probabilities decrease. The fitted probabilities are plotted along the horizontal axis in descending order. The vertical coordinate for a fitted probability is the proportion of correct classifications for that probability or higher, divided by the overall correct classification rate. Use the lift curve to see whether you can correctly classify a large proportion of observations if you select only those with a fitted probability that exceeds a threshold value.

If the response has more than two levels, the Lift Curve plot displays a lift curve for each response level. For a given response level, this curve is the lift curve for correct classification into that level. See the Partition Models chapter in the *Predictive and Specialized Modeling* book for more information about lift curves.

Plot Actual by Predicted (Not available when the specified Distribution is Binomial, Multinomial, Ordinal Logistic, or Cox Proportional Hazards.) Plots actual Y values on the vertical axis and predicted Y values on the horizontal axis. If you used validation, a plot is shown for each of the Training, Validation, and Test sets.

Plot Residual by Predicted (Not available when the specified Distribution is Binomial, Multinomial, Ordinal Logistic, or Cox Proportional Hazards.) Plots the residuals on the vertical axis and the predicted Y values on the horizontal axis. If you used validation, a plot is shown for each of the Training, Validation, and Test sets.

Plot Residual by Predictor (Not available when the specified Distribution is Binomial, Multinomial, Ordinal Logistic, or Cox Proportional Hazards. Not available for responses that have the Vector modeling type.) For each predictor in the model, plots

the residuals on the vertical axis and the predictor values on the horizontal axis. There is a plot for each of the predictors in the model. If you used validation, a set of plots is shown for each of the Training, Validation, and Test sets.

Normal Quantile Plot (Available only when the specified Distribution is Normal and there is no censoring.) Plots normal quantiles on the vertical axis and standardized residuals on the horizontal axis. If you used validation, a plot is shown for each of the Training, Validation, and Test sets.

Save Columns Enables you to save columns based on the fitted model to the data table. See "Save Columns Options for Cox Proportional Hazards Models" on page 331 for the options that are available if Cox Proportional Hazards is selected as the Distribution. For all other Distributions, the following columns can be saved to the data table:

Save Prediction Formula Saves a column to the data table that contains the prediction formula, given in terms of the observed (unstandardized) data values. The prediction formula does not contain zeroed terms. See "Statistical Details for Distributions" on page 335 for mean formulas.

When the response column is categorical, this option creates a probability column for each response level as well as a column that contains the most likely response. The Most Likely response column contains the level with the highest probability based on the model. If the Probability Threshold is a value other than 0.5, this option creates an additional column that contains the most likely response based on the probability threshold value.

Mean Confidence Interval Saves two columns to the data table that contain the lower and upper 95% confidence limits for the mean response.

Note: You can change the α level for the confidence interval in the Fit Model window by selecting Set Alpha Level from the red triangle menu.

Std Error of Predicted Saves a column to the data table that contains the standard errors of the predicted mean response.

Save Residual Formula Saves a column to the data table that contains a formula for the residuals, given in the form Y minus the prediction formula. The residual formula does not contain zeroed terms. Not available if Binomial is selected as the Distribution.

Save Variance Formula Saves a column to the data table that contains a formula for the variance of the prediction. The variance of the prediction is calculated using the formula for the variance of the selected Distribution. The value of the parameter involved in the link function is estimated by applying the inverse of the link function to

the estimated linear component. Other parameters are replaced by their estimates. See "Statistical Details for Distributions" on page 335 for variance formulas. Not available if Binomial is selected as the Distribution.

Save Linear Predictor Saves a column to the data table that contains a formula for the product of the design matrix and the vector of parameter estimates. This is commonly referred to as $\mathbf{X}\beta$. The formula does not contain zeroed terms.

Save Validation Column (Available only if the specified Validation Method is KFold, Holdback, or Leave-One-Out.) Saves a column that describes the assignment of rows to folds. For KFold, the column lists the fold to which the row was assigned. For Holdback, each row is identified as belonging to the Training or Validation set. For Leave-One-Out, the row's value indicates its order in being left out.

Note: If you selected a Validation column in the launch window, the Save Validation Column option does not appear.

Save Simulation Formula Saves a column to the data table that contains a formula that generates simulated values using the estimated parameters for the model that you fit. This column can be used in the Simulate utility as a Column to Switch In. See the Simulate chapter in *Basic Analysis*.

Cook's D Influence (Available only if the specified Distribution is Normal and the specified Estimation Method is Standard Least Squares.) Saves a column to the data table that contains the values for Cook's *D* Influence statistic.

Hats (Available only if the specified Distribution is Normal and the specified Estimation Method is Standard Least Squares.) Saves a column to the data table that contains the diagonal elements of $\mathbf{X(X'X)^{-1}X'}$. These values are sometimes called *hat values*.

Publish Prediction Formula Creates prediction formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

Save Columns Options for Cox Proportional Hazards Models

Save Survival Formula Saves a column to the data table that contains a formula for the probability of survival at the observed time.

Save Cox-Snell Residual Formula Saves a column to the data table that contains a formula for the Cox-Snell residuals. The Cox-Snell residuals are strictly positive. See Meeker and Escobar (1998, sec. 17.6.1) for a discussion of Cox-Snell residuals.

Save Martingale Residual Formula Saves a column to the data table that contains a formula for the martingale residuals. The martingale residual is defined as the difference between the observed number of events for an individual and a conditionally expected number of events. The martingale residuals have a mean of zero and range between negative infinity and 1. See Fleming and Harrington (1991).

Save Linear Predictor Saves a column to the data table that contains a formula for the product of the design matrix and the vector of parameter estimates. This is commonly referred to as $\mathbf{X}\beta$. The formula does not contain zeroed terms.

Remove Fit Removes the report for the fit.

JMP PRO Statistical Details for the Generalized Regression Personality

JMP PRO Statistical Details for Estimation Methods

Penalized regression methods introduce bias to the regression coefficients by penalizing them.

JMP PRO Ridge Regression

An l_2 penalty is applied to the regression coefficients during ridge regression. Ridge regression coefficient estimates are given by the following:

$$\hat{\beta}^{ridge} = \text{argmin}_{\beta}\left\{ \sum_{i=1}^{N} -LogLikelihood(\beta;y_i) + \frac{\lambda}{2}\sum_{j=1}^{p}\beta_j^2 \right\},$$

where

$\sum_{j=1}^{p}\beta_j^2$ is the l_2 penalty, λ is the tuning parameter, N is the number of rows, and p is the number of variables.

JMP PRO Dantzig Selector

An l_∞ penalty is applied to the regression coefficients during Dantzig Selector. Coefficient estimates for the Dantzig Selector satisfy the following criterion:

$$\min_\beta \left\| X^T(y - X\beta) \right\|_\infty \quad \text{subject to } \|\beta\|_1 \le t$$

where

$\|v\|_\infty$ denotes the l_∞ norm, which is the maximum absolute value of the components of the vector v.

JMP PRO Lasso Regression

An l_1 penalty is applied to the regression coefficients during Lasso. Coefficient estimates for the Lasso are given by the following:

$$\hat{\beta}^{lasso} = \text{argmin}_\beta \left\{ \frac{1}{2} \sum_{i=1}^{N} -LogLikelihood(\beta; y_i) + \lambda \sum_{j=1}^{p} |\beta_j| \right\},$$

where

$\sum_{j=1}^{p} |\beta_j|$ is the l_1 penalty, λ is the tuning parameter, N is the number of rows, and p is the number of variables

JMP PRO Elastic Net

The Elastic Net combines both l_1 and l_2 penalties. Coefficient estimates for the Elastic Net are given by the following:

$$\hat{\beta}^{enet} = \text{argmin}_\beta \left\{ \sum_{i=1}^{N} -LogLikelihood(\beta; y_i) + \lambda \sum_{j=1}^{p} \left(\alpha |\beta_j| + \frac{(1-\alpha)}{2} \beta_j^2 \right) \right\},$$

The notation used in this equation is as follows:

- $\sum_{j=1}^{p} |\beta_j|$ is the l_1 penalty
- $\sum_{j=1}^{p} \beta_j^2$ is the l_2 penalty
- λ is the tuning parameter
- α is a parameter that determines the mix of the l_1 and l_2 penalties
- N is the number of rows
- p is the number of variables

Tip: There are two sample scripts that illustrate the shrinkage effect of varying α and λ in the Elastic Net for a single predictor. Select **Help > Sample Data**, click **Open the Sample Scripts Directory**, and select demoElasticNetAlphaLambda.jsl or demoElasticNetAlphaLambda2.jsl. Each script contains a description of how to use it and what it illustrates.

JMP PRO Adaptive Methods

The adaptive Lasso method uses weighted penalties to provide consistent estimates of coefficients. The weighted form of the l_1 penalty is

$$\sum_{j=1}^{p} \frac{|\beta_j|}{|\tilde{\beta}_j|},$$

where

$\tilde{\beta}_j$ is the MLE when the MLE exists. If the MLE does not exist and the response distribution is normal, estimation is done using least squares and $\tilde{\beta}_j$ is the solution obtained using a generalized inverse. If the response distribution is not normal, $\tilde{\beta}_j$ is the ridge solution.

For the adaptive Lasso, this weighted form of the l_1 penalty is used in determining the $\hat{\beta}^{lasso}$ coefficients.

The adaptive Elastic Net uses this weighted form of the l_1 penalty and also imposes a weighted form of the l_2 penalty. The weighted form of the l_2 penalty for the adaptive Elastic Net is

$$\sum_{j=1}^{p} \left(\frac{\beta_j}{\tilde{\beta}_j}\right)^2,$$

where

$\tilde{\beta}_j$ is the MLE when the MLE exists. If the MLE does not exist and the response distribution is normal, estimation is done using least squares and $\tilde{\beta}_j$ is the solution obtained using a generalized inverse. If the response distribution is not normal, $\tilde{\beta}_j$ is the ridge solution.

JMP PRO Statistical Details for Advanced Controls

JMP PRO Grid

The tuning parameters for ridge regression and the Lasso that best minimize the penalized likelihood are found by searching a grid of tuning parameter values. This grid of values lies

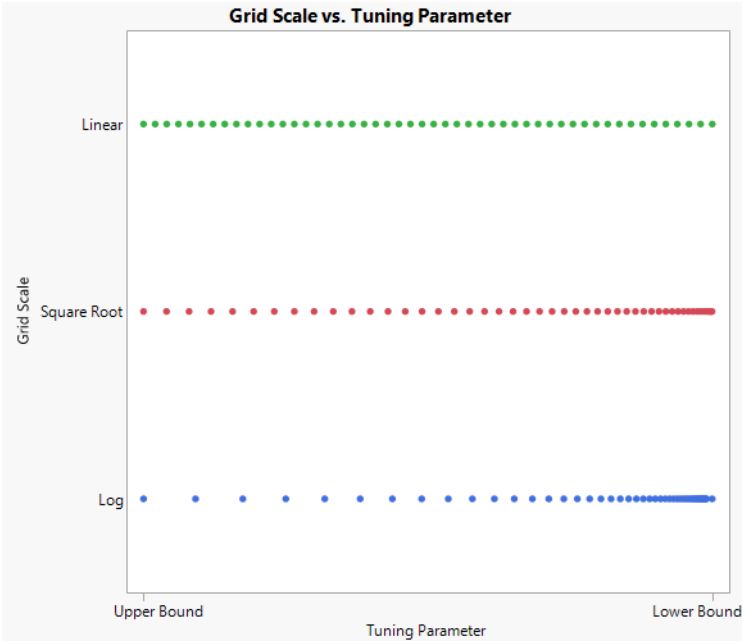
between a lower and an upper bound for the tuning parameter. You can specify the number of grid points under Advanced Controls.

The lower bound is zero except in special cases where it is set to 0.0001. See "Tuning Parameter" on page 307. When the lower bound for the tuning parameter is zero, the solution is unpenalized and the coefficients are the MLEs. The upper bound is the smallest value for which all of the non-intercept terms are zero.

The grid of values between the lower and upper bounds is iteratively searched to determine the best value of the tuning parameter. The grid of possible tuning parameters can be set up in three different scales: linear, log, and square root. The default grid scale is square root.

In some cases, there is a large gap between the unpenalized estimates and the previous step. This large gap can distort the solution path. The log scale focuses its search on small tuning parameter values with few large values, whereas the linear scale evenly disperses the search from the minimum to the maximum value. The square root scale is a compromise between the other two scales. Figure 6.7 shows the different grid scales.

Figure 6.7 Options for Tuning Parameter Grid Scale



JMP PRO Statistical Details for Distributions

The distributions fit by the Generalized Regression personality are given below in terms of the parameters used in model fitting. Although it is not specifically stated as part of their

descriptions, the Generalized Regression personality enables you to specify non-integer values for the discrete distributions.

JMP PRO Continuous Distributions

Normal Distribution

$$f(y|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma^2}(y-\mu)^2\right], -\infty < y < \infty$$

$$E(Y) = \mu$$

$$Var(Y) = \sigma^2$$

Cauchy Distribution

$$f(y|\mu,\sigma) = \left\{\pi\sigma\left[1+\left(\frac{y-\mu}{\sigma}\right)^2\right]\right\}^{-1}, -\infty < y < \infty$$

$$E(Y) = \text{undefined}$$

$$Var(Y) = \text{undefined}$$

Exponential Distribution

$$f(y|\theta) = \frac{1}{\mu}\exp\left[-\frac{y}{\mu}\right], y > 0$$

$$E(Y) = \mu$$

$$Var(Y) = \mu^2$$

Gamma Distribution

$$f(y|\mu,\sigma) = \frac{y^{(\mu/\sigma)-1}\exp[-y/\sigma]}{\Gamma[\mu/\sigma]\sigma^{\mu/\sigma}}, y > 0$$

$$E(Y) = \mu$$

$$Var(Y) = \mu\sigma$$

Weibull Distribution

$$f(y|\mu,\sigma) = \frac{1}{y\sigma}\exp\left[\frac{\log(y)-\mu}{\sigma}\right]\exp\left\{-\exp\left[\frac{\log(y)-\mu}{\sigma}\right]\right\},\ y>0$$

$$E(Y) = \exp(\mu)\Gamma[1+\sigma]$$

$$Var(Y) = \exp(2\mu)\{\Gamma[1+2\sigma]-(\Gamma[1+\sigma])^2\}$$

LogNormal Distribution

$$f(y|\mu,\sigma) = \frac{1}{y\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2\sigma^2}[\log(y)-\mu]^2\right\},\ y>0$$

$$E(Y) = \exp\left(\mu+\frac{\sigma^2}{2}\right)$$

$$Var(Y) = [\exp(\sigma^2)-1]\exp(2\mu+\sigma^2)$$

Beta Distribution

$$f(y|\mu,\sigma) = \frac{\Gamma[1/\sigma]}{\Gamma[\mu/\sigma]\Gamma[(1-\mu)/\sigma]}y^{\mu/\sigma-1}(1-y)^{\frac{(1-\mu)}{\sigma}-1},\ y\in(0,1)$$

$$E(Y) = \mu$$

$$Var(Y) = \mu(1-\mu)\frac{\sigma}{\sigma+1}$$

JMP PRO Discrete Distributions

Binomial Distribution

$$f(y|n,p) = \binom{n}{y}p^y(1-p)^{n-y},\ y = 0, 1, 2, \ldots, n$$

$$E(Y) = np$$

$$Var(Y) = np(1-p)$$

Beta Binomial Distribution

$$f(y|n,p,\delta) = \binom{n}{y}\frac{\Gamma\left[\frac{1}{\delta}-1\right]\Gamma\left[y+p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[n-y+(1-p)\left(\frac{1}{\delta}-1\right)\right]}{\Gamma\left[p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[(1-p)\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[n-1+\frac{1}{\delta}\right]}, \; y = 0, 1, 2, \ldots, n$$

$$E(Y) = np$$

$$Var(Y) = np(1-p)[1+(n-1)\delta]$$

Poisson Distribution

$$f(y|\lambda) = \frac{\lambda^y}{y!}\exp(-\lambda), \; y = 0, 1, 2, \ldots$$

$$E(Y) = \lambda$$

$$Var(Y) = \lambda$$

Negative Binomial Distribution

$$f(y|\mu,\sigma) = \frac{\Gamma[y+(1/\sigma)]}{\Gamma[y+1]\Gamma[1/\sigma]}\left[\frac{(\mu\sigma)^y}{(1+\mu\sigma)^{y+(1/\sigma)}}\right], \; y = 0, 1, 2, \ldots$$

$$E(Y) = \mu$$

$$Var(Y) = \mu + \sigma\mu^2$$

JMP PRO Zero-Inflated Distributions

Zero-Inflated Binomial Distribution

$$f(y|n,p,\pi) = \begin{cases} \pi + (1-\pi)(1-p)^n, \text{ for } y = 0 \\ (1-\pi)\binom{n}{y}p^y(1-p)^{n-y}, \text{ for } y = 1, 2, \ldots, n \end{cases}$$

$$E(Y) = np(1-\pi)$$

$$Var(Y) = (1-\pi)np(1-p) + n^2p^2(1-\pi)^2$$

Zero-Inflated Beta Binomial Distribution

$$f(y|n,p,\delta,\pi) = \begin{cases} \pi + (1-\pi)\dfrac{\Gamma\left[\frac{1}{\delta}-1\right]\Gamma\left[p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[n+(1-p)\left(\frac{1}{\delta}-1\right)\right]}{\Gamma\left[p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[(1-p)\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[n-1+\frac{1}{\delta}\right]}, \text{ for } y=0 \\ (1-\pi)\binom{n}{y}\dfrac{\Gamma\left[\frac{1}{\delta}-1\right]\Gamma\left[y+p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[n-y+(1-p)\left(\frac{1}{\delta}-1\right)\right]}{\Gamma\left[p\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[(1-p)\left(\frac{1}{\delta}-1\right)\right]\Gamma\left[n-1+\frac{1}{\delta}\right]}, y=1,2,\ldots,n \end{cases}$$

$$E(Y) = np(1-\pi)$$

$$Var(Y) = (1-\pi)np\{1+(1-p)[1+(n-1)\delta]\} - [np(1-\pi)]^2$$

Zero-Inflated Poisson Distribution

$$f(y|\lambda,\pi) = \begin{cases} \pi + (1-\pi)\exp[-\lambda], \text{ for } y=0 \\ (1-\pi)\dfrac{\lambda^y}{y!}\exp[-\lambda], \text{ for } y=1,2,\ldots \end{cases}$$

$$E(Y) = (1-\pi)\lambda$$

$$Var(Y) = \lambda(1-\pi)(1+\lambda\pi)$$

Zero-Inflated Negative Binomial Distribution

$$f(y|\mu,\sigma,\pi) = \begin{cases} \pi + (1-\pi)(1+\mu\sigma)^{-(1/\sigma)}, \text{ for } y=0 \\ (1-\pi)\dfrac{\Gamma[y+(1/\sigma)]}{\Gamma[y+1]\Gamma[1/\sigma]}\left[\dfrac{(\mu\sigma)^y}{(1+\mu\sigma)^{y+(1/\sigma)}}\right], \text{ for } y=1,2,\ldots \end{cases}$$

$$E(Y) = (1-\pi)\mu$$

$$Var(Y) = \mu(1-\pi)[1+\mu(\sigma+\pi)]$$

Zero-Inflated Gamma Distribution

$$f(y|\mu, \sigma, \pi) = \begin{cases} \pi \text{ , for } y = 0 \\ (1-\pi)\dfrac{\exp(-y/\sigma)}{\Gamma[\mu/\sigma]\sigma^{\mu/\sigma}y^{1-\mu/\sigma}} \text{ , for } y > 0 \end{cases}$$

$$E(Y) = \mu(1-\mu)$$

$$Var(Y) = \mu(1-\pi)(\sigma+\mu) - (1-\pi)^2\mu^2$$

Chapter 7

Generalized Regression Examples

Build Models Using Regularization Techniques

This chapter provides examples with instructional material for several models fit using the Generalized Regression platform.

Contents

Poisson Generalized Regression Example

The Liver Cancer.jmp sample data table contains liver cancer Node Count values for 136 patients. It also includes measurements on six potentially related variables: BMI, Age, Time, Markers, Hepatitis, and Jaundice. These columns are described in Column Notes in the data table.

This example develops a prediction model for Node Count using the six predictors. Node Count is modeled using a Poisson distribution.

1. Select **Help > Sample Data Library** and open Liver Cancer.jmp.
2. Select **Analyze > Fit Model**.
3. Select Node Count from the Select Columns list and click **Y**.
4. Select BMI through Jaundice and click **Macros > Factorial to Degree**.

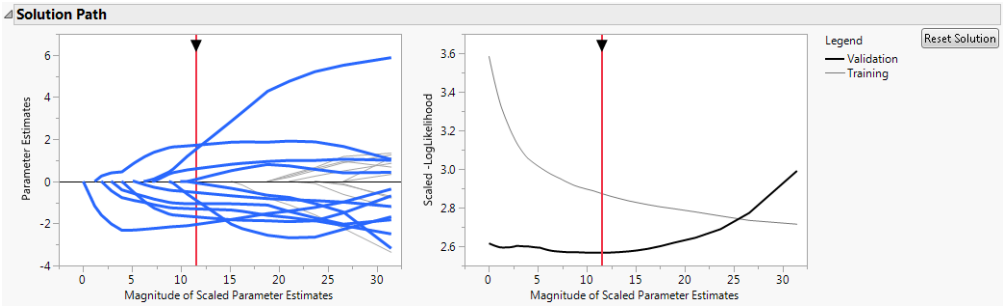
   This adds all terms up to degree 2 (the default in the **Degree** box) to the model.
5. Select Validation from the Select Columns list and click **Validation**.
6. From the Personality list, select **Generalized Regression**.
7. From the Distribution list, select **Poisson**.
8. Click **Run**.

   The Generalized Regression report that appears contains a Model Launch control panel and a Maximum Likelihood with Validation Column report. Note that the default estimation method is the Adaptive Lasso.
9. Click **Go**.
10. Click the red triangle next to Adaptive Lasso with Validation Column and select **Select Nonzero Terms**.

    The Solution Path is shown in Figure 7.1. The paths for terms that have nonzero coefficients are highlighted. Think of the solution paths as moving from right to left across the plot, as the solutions shrink farther from the MLE. A number of terms have paths that shrink them to zero fairly early.

    The vertical axis in the Solution Path Plot represents the values of the parameter estimates for the standardized predictors. The vertical red line indicates their values at the optimal shrinkage, as determined by cross validation. At this point, 11 terms have nonzero coefficients. Notice that the vertical red line indicates the minimum Scaled –LogLikelihood value.

Figure 7.1 Solution Path for Lasso Fit with Nonzero Terms Highlighted



The Parameter Estimates for Original Predictors report (Figure 7.2) shows the parameter estimates for the uncentered and unscaled data. The 11 terms with nonzero parameter estimates are highlighted. These include interaction effects. In the data table, all six predictor columns are selected because every predictor column appears in a term that has a nonzero coefficient.

In the Effect Tests report, the 10 effects with zero coefficient estimates are designated as Removed. The Effect Tests report indicates that only one effect is significant at the 0.05 level: the Age*Markers interaction.

11. Click on the row for (Age - 56.3994)*Markers[0-1] in the Parameter Estimates for Original Predictors report.

    This action highlights that effect's path in the Solution Path Plot and selects the columns Age and Markers in the data table.

Figure 7.2 Parameter Estimates Report with Nonzero Terms Highlighted

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	1.9658064	1.5268652	1.6576014	0.1979	-1.026794	4.9584071
BMI	-0.014557	0.0655906	0.0492589	0.8244	-0.143113	0.1139977
Age	-0.000344	0.0052439	0.0043119	0.9476	-0.010622	0.0099335
Time	0	0	0	1.0000	0	0
Markers[0-1]	0.3853241	0.3347727	1.3248061	0.2497	-0.270818	1.0414665
Hepatitis[0-1]	0	0	0	1.0000	0	0
Jaundice[0-1]	-0.220764	0.1645347	1.8002793	0.1797	-0.543246	0.1017185
(BMI-23.1737)*(Age-56.3994)	0.0007336	0.0014188	0.267356	0.6051	-0.002047	0.0035144
(BMI-23.1737)*(Time-9.00945)	0	0	0	1.0000	0	0
(BMI-23.1737)*Markers[0-1]	-0.066574	0.0511526	1.6938495	0.1931	-0.166831	0.0336832
(BMI-23.1737)*Hepatitis[0-1]	0.0269878	0.0827496	0.1063662	0.7443	-0.135198	0.1891741
(BMI-23.1737)*Jaundice[0-1]	0	0	0	1.0000	0	0
(Age-56.3994)*(Time-9.00945)	0	0	0	1.0000	0	0
(Age-56.3994)*Markers[0-1]	-0.023498	0.0083023	8.0106599	0.0047*	-0.03977	-0.007226
(Age-56.3994)*Hepatitis[0-1]	0	0	0	1.0000	0	0
(Age-56.3994)*Jaundice[0-1]	0	0	0	1.0000	0	0
(Time-9.00945)*Markers[0-1]	-0.007529	0.0104633	0.5178189	0.4718	-0.028037	0.0129784
(Time-9.00945)*Hepatitis[0-1]	0	0	0	1.0000	0	0
(Time-9.00945)*Jaundice[0-1]	0.0009619	0.0093476	0.0105895	0.9180	-0.017359	0.0192828
Markers[0-1]*Hepatitis[0-1]	-0.422619	0.397399	1.1309549	0.2876	-1.201507	0.3562684
Markers[0-1]*Jaundice[0-1]	0	0	0	1.0000	0	0
Hepatitis[0-1]*Jaundice[0-1]	0	0	0	1.0000	0	0

12. Click the red triangle next to Adaptive Lasso with Validation Column and select **Save Columns > Save Prediction Formula** and **Save Columns > Save Variance Formula**.

    Two columns are added to the data table: Node Count Prediction Formula and Node Count Variance.

13. Right-click either column heading and select **Formula** to view the formula. Alternatively, click on the plus sign to the right of the column name in the Columns panel.

    The prediction formula in the **Save Prediction Formula** column applies the exponential function to the estimated linear part of the model. The prediction variance formula in Node Count Variance is given by the identical formula, because the variance of a Poisson distribution equals its mean.

Binomial Generalized Regression Example

This example shows how to develop a prediction model for the binomial response, Severity, in the Liver Cancer.jmp sample data table.

1. Select **Help > Sample Data Library** and open Liver Cancer.jmp.
2. Select **Analyze > Fit Model**.
3. Select Severity from the Select Columns list and click **Y**.
4. Select BMI through Jaundice and click **Macros > Factorial to Degree**.

   All terms up to degree 2 (the default in the **Degree** box) are added to the model.

5. From the Personality list, select **Generalized Regression**.

   The Distribution list automatically shows the Binomial distribution. This is the only distribution available when Y is binary and has a Nominal modeling type.

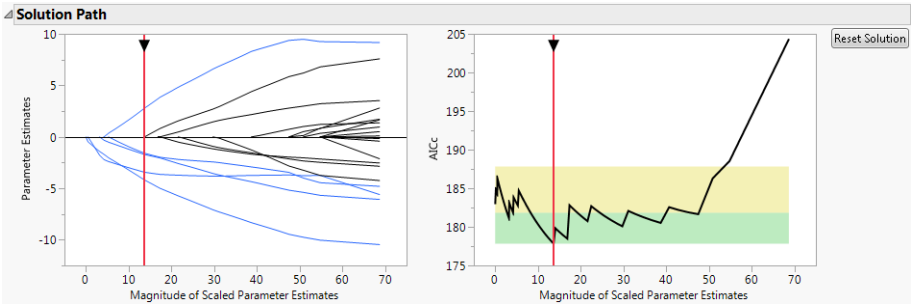
6. Click **Run**.

   The Generalized Regression report that appears contains a Model Launch control panel and a Logistic Regression report. Note that the default estimation method is the adaptive Lasso.

7. Select **Elastic Net** as the Estimation Method.
8. Click **Go**.

   An Adaptive Elastic Net with AICc Validation report appears. The Solution Path is shown in Figure 7.3.

Figure 7.3 Solution Path Plot



The paths for terms that have nonzero coefficients are shown in blue. The optimal parameter values are substantially shrunken away from the MLE. The Validation Plot to the right indicates that several models can be considered as good as the best model. To view those models, slide the vertical red bar around in the region the black line is in the green area.

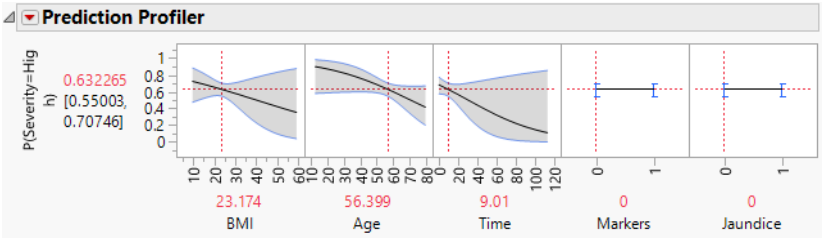
9. Click the red triangle next to Adaptive Elastic Net with AICc Validation and select the **Select Zeroed Terms** option.

   The 16 terms that have coefficient estimates of zero are highlighted in the Parameter Estimates for Original Predictors report. The Effect Tests report designates these terms as Removed.

   The Effect Tests report also shows that there are no significant terms at the 0.05 level. However, the Time*Markers interaction has a small p-value of 0.0626 and the Time effect has a small p-value of 0.1459.

10. Click the red triangle next to Adaptive Elastic Net with AICc Validation and select **Profilers > Profiler**.

Figure 7.4 Profiler for Probability That Severity = High, Time Low

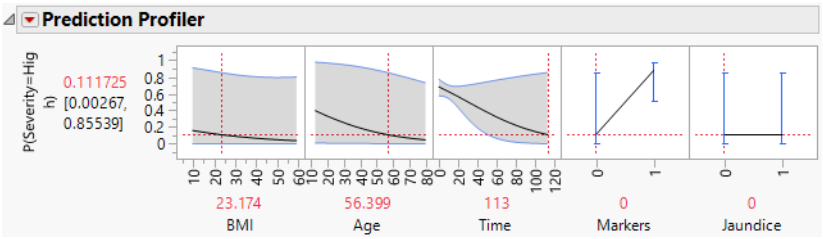


Examine the Prediction Profiler to see how Time and the Time*Markers interaction affect Severity.

Note: The predictor Hepatitis is not shown in the profiler because it does not appear in any active (nonzero) terms. Because Markers and Jaundice appear in active interaction terms, they appear in the profiler even though, as main effects, they are not active.

11. Move the red dashed line for Time from left to right to see its interaction with Markers (Figure 7.4 and Figure 7.5). For patients who enter the study with small values of Time since diagnosis, Markers have little impact on Severity. But for patients who enter the study having been diagnosed for a longer time, Markers are important. For those patients, normal markers suggest a lower probability of high Severity.

Figure 7.5 Profiler for Probability That Severity = High, Time High



Zero-Inflated Poisson Regression Example

The Fishing.jmp sample data table contains fictional data for a study of various factors that affect the number of fish caught by groups visiting a park. The data table contains 250 responses from families or groups of traveling companions. This example models the number of Fish Caught as a function of Live Bait, Fishing Poles, Camper, People, and Children. These columns are described in Column Notes in the data table.

The data table contains a hidden column called Fished. During data collection, it was never determined whether anyone in the group had actually fished. However, the Fished column is included in the table to emphasize the point that catching zero fish can happen in one of two ways: Either no one in the group fished, or everyone who fished in the group was unlucky.

Therefore, zero responses can come from two sources. To address this issue, you can fit a zero-inflated distribution. Because a Poisson distribution is appropriate for the count data resulting from people who fished, you fit a zero-inflated Poisson distribution.

1. Select **Help > Sample Data Library** and open Fishing.jmp.
2. Select **Analyze > Fit Model**.
3. Select Fish Caught from the Select Columns list and click **Y**.
4. Select Live Bait through Children and click **Macros > Factorial to Degree**.

   Terms up to degree 2 (the default in the **Degree** box) are added to the model.

5. Select Validation from the Select Columns list and click **Validation**.
6. From the Personality list, select **Generalized Regression**.
7. From the Distribution list, select **ZI Poisson**.
8. Click **Run**.

   The Generalized Regression report that appears contains a Model Launch control panel and a Maximum Likelihood with Validation Column report. Note that the default estimation method is the Adaptive Lasso.
9. From the Estimation Method List, select **Elastic Net**.
10. Click **Go**.

    An Adaptive Elastic Net with Validation Column report appears. The Solution Path, the Parameter Estimates for Original Predictors report, and the Effect Tests report indicate that a number of terms are zeroed. The Zero Inflation parameter, whose estimate is shown on the last line of both Parameter Estimates reports, is highly significant. This indicates that some of the variation in the response, Fish Caught, might be due to the fact that some groups did not fish.

Figure 7.6 Parameter Estimates for Original Predictors Report

Parameter Estimates for Original Predictors

Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	1.5677918	0.1194977	172.1304	<.0001*	1.3335805	1.802003
Live Bait[0-1]	-0.392752	0.1781431	4.8606954	0.0275*	-0.741906	-0.043598
Fishing Poles	0.2946742	0.0468731	39.521875	<.0001*	0.2028047	0.3865437
Camper[0-1]	-0.438133	0.2007789	4.7618458	0.0291*	-0.831652	-0.044613
People	0	0	0	1.0000	0	0
Children	-0.188872	0.1062779	3.1582561	0.0755	-0.397172	0.0194293
Live Bait[0-1]*(Fishing Poles-0.856)	-0.024329	0.0657001	0.1371288	0.7112	-0.153099	0.1044406
Live Bait[0-1]*Camper[0-1]	-0.258491	0.4663182	0.307275	0.5794	-1.172458	0.6554755
Live Bait[0-1]*(People-2.516)	-0.090833	0.1748961	0.2697262	0.6035	-0.433623	0.2519575
Live Bait[0-1]*(Children-0.936)	0.106422	0.2349365	0.2051927	0.6506	-0.354045	0.5668891
(Fishing Poles-0.856)*Camper[0-1]	-0.425751	0.1100387	14.96992	0.0001*	-0.641423	-0.210079
(Fishing Poles-0.856)*(People-2.516)	0.007093	0.0907279	0.0061119	0.9377	-0.170731	0.1849164
(Fishing Poles-0.856)*(Children-0.936)	-0.176494	0.0814661	4.6935903	0.0303*	-0.336165	-0.016823
Camper[0-1]*(People-2.516)	-0.224584	0.3075565	0.533224	0.4653	-0.827384	0.3782151
Camper[0-1]*(Children-0.936)	0.2714388	0.3407212	0.6346662	0.4256	-0.396363	0.9392401
(People-2.516)*(Children-0.936)	0.0216352	0.0420678	0.2644969	0.6070	-0.060816	0.1040867

ZI Poisson Distribution Parameters	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Zero Inflation	0.7819268	0.0352589	491.80711	<.0001*	0.7128207	0.851033

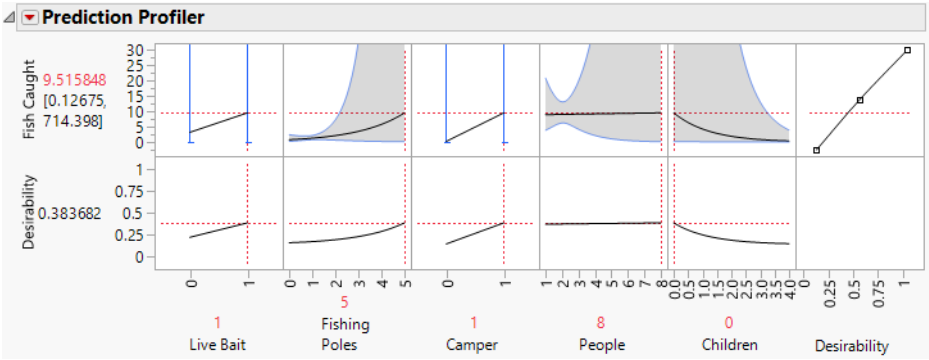
The Effect Tests report indicates that five terms are significant at the 0.05 level: Live Bait, Fishing Poles, Camper, Fishing Poles*Camper, and Fishing Poles*Children.

11. Click the red triangle next to Adaptive Elastic Net with Validation Column and select **Profilers > Profiler**.
12. Click the Prediction Profiler red triangle menu and select **Optimization and Desirability > Desirability Functions**.

A function is imposed on the response, which indicates that maximizing the number of Fish Caught is desirable. (See the Profiler chapter in the *Profilers* book for more information about desirability functions.)

13. Click the Prediction Profiler red triangle menu and select **Optimization and Desirability > Maximize Desirability**.

Figure 7.7 Prediction Profiler with Fish Caught Maximized



You can vary the settings of the predictors to see the impact of the significant effects: Live Bait, Fishing Poles, Fishing Poles*Camper, and Fishing Poles*Children. For example, Live Bait is associated with more fish; a Camper tends to bring more fishing poles than someone who is not camping and therefore catches more fish.

14. Click the red triangle next to Adaptive Elastic Net with Validation Column and select **Save Columns > Save Prediction Formula** and **Save Columns > Save Variance Formula**.

    Two columns are added to the data table: Fish Caught Prediction Formula and Fish Caught Variance.

15. Right-click either column heading and select **Formula** to view the formula. Alternatively, click the plus sign to the right of the column name in the Columns panel. Note the appearance of the estimated zero-inflation parameter, 0.7819268, in both of these formulas.

Chapter 8

JMP PRO Mixed Models

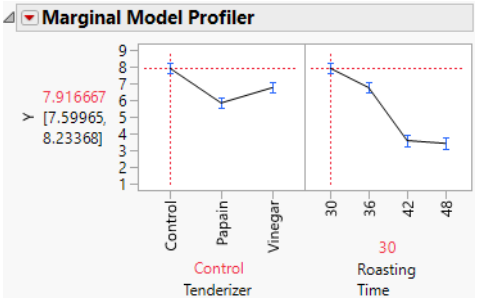
Jointly Model the Mean and Covariance

The Mixed Models personality of the Fit Model platform is available only in JMP Pro.

In JMP Pro, the Fit Model platform's Mixed Model personality fits a wide variety of linear models for continuous responses with complex covariance structures. These models include random coefficients, repeated measures, spatial data, and data with multiple correlated responses. Use the Mixed Model personality to specify linear mixed models and their covariance structures conveniently using an intuitive interface, and to fit these models using maximum likelihood methods.

Analytic results are supported by compelling dynamic visualization tools such as profilers, surface plots, and contour plots. These visual displays stimulate, complement, and support your understanding of the model. See the *Profilers* book for more information.

Figure 8.1 Marginal Model Profiler for a Split Plot Experiment



Contents

JMP PRO Overview of the Mixed Model Personality

In JMP Pro, the Mixed Model personality lets you analyze models with complex covariance structures. The situations that can be analyzed include:

- Split plot experiments
- Random coefficients models
- Repeated measures designs
- Spatial data
- Correlated response data

Split plot experiments are experiments with two or more levels, or sizes, of experimental units resulting in multiple error terms. Such designs are often necessary when some factors are easy to vary and others are more difficult to vary. (See the Custom Design chapter in *Design of Experiments Guide*.)

Random coefficients models are also known as hierarchical or multilevel models (Singer 1998; Sullivan et al. 1999). These models are used when batches or subjects are thought to differ randomly in intercept and slope. Drug stability trials in the pharmaceutical industry and individual growth studies in educational research often require random coefficient models.

Repeated measures designs, spatial data, and correlated response data share the property that observations are not independent, requiring that you model their correlation structure.

- Repeated measures designs, also known as within-subject designs, model changes in a response over time or space while allowing errors to be correlated.
- Spatial data are measurements made in two or more dimensions, typically latitude and longitude. Spatial measurements are often correlated as a function of their spatial proximity.
- Correlated response data result from making several measurements on the same experimental unit. For example, height, weight, and blood pressure readings taken on individuals in a medical study, or hardness, strength, and elasticity measured on a manufactured item, are likely to be correlated. Although these measurements can be studied individually, treating them as correlated responses can lead to useful insights.

Failure to account for correlation between observations can result in incorrect conclusions about treatment effects. However, estimating covariance structure parameters uses information in the data. The number of parameters being estimated impacts power and the Type I error rate. For this reason, you must choose covariance models judiciously. For more information, see "Repeated Measures Example" on page 378.

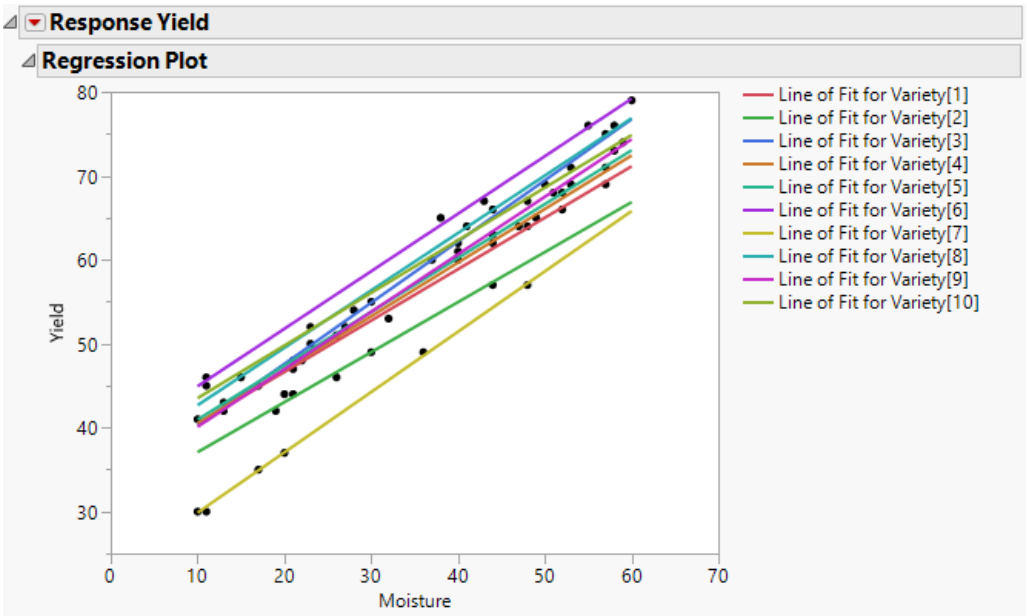
JMP PRO Example Using Mixed Model

In a study of wheat yield, 10 varieties of wheat are randomly selected from the population of varieties of hard red winter wheat adapted to dry climate conditions. These are randomly assigned to six one-acre plots of land. The preplanting moisture content of the plots could influence the germination rate and hence the eventual yield of the plots. Thus, the amount of preplanting moisture in the top 36 inches of soil is determined for each plot. You are interested in determining if the moisture content affects yield.

Because the varieties are randomly selected, the regression model for each variety is a random model selected from the population of variety models. The intercept and slope are random for each variety and might be correlated. The random coefficients are centered at the fixed effects. The fixed effects are the population intercept and the slope, which are the expected values of the population of the intercepts and slopes of the varieties. This example is taken from Littell et al. (2006, p. 320).

Fitting the model using REML in the Standard Least Squares personality lets you view the variation in intercepts and slopes (Figure 8.2). Note that the slopes do not have much variability, but the intercepts have quite a bit. The intercept and slope might be negatively correlated; varieties with lower intercepts seem to have higher slopes.

Figure 8.2 Standard Least Squares Regression

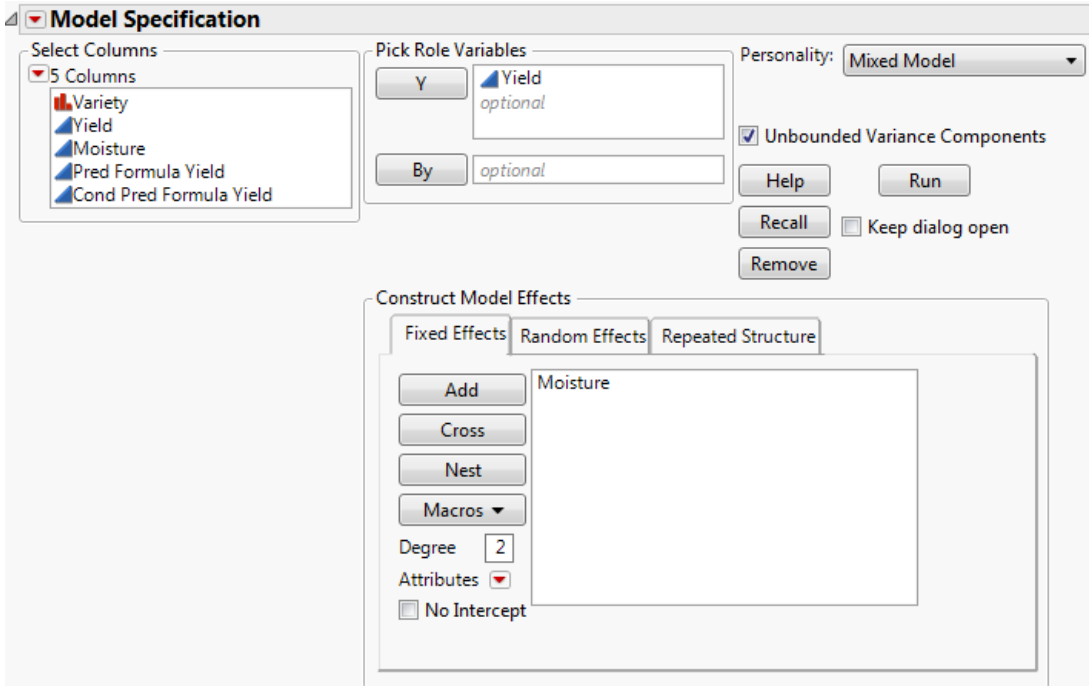


To model the correlation between the intercept and the slope, use the Mixed Models personality. You are interested in determining the population regression equation as well as variety-specific equations.

1. Select **Help > Sample Data Library** and open Wheat.jmp.
2. Select **Analyze > Fit Model**.
3. Select Yield and click **Y**.

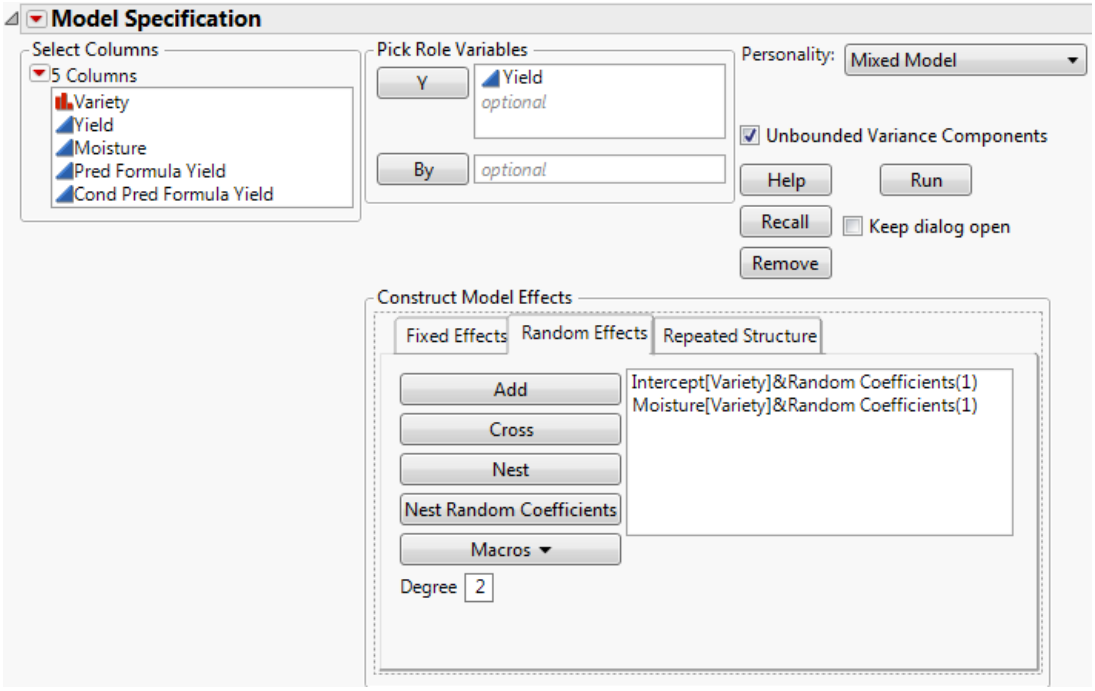
   When you add this column as Y, the fitting Personality becomes Standard Least Squares.
4. Select **Mixed Model** from the Personality list. Alternatively, you can select the Mixed Model personality first, and then click **Y** to add Yield.
5. Select Moisture and click **Add** on the Fixed Effects tab.

Figure 8.3 Completed Fit Model Launch Window Showing Fixed Effects



6. Select the **Random Effects** tab.
7. Select Moisture and click **Add**.
8. Select **Variety** from the Select Columns list, select Moisture from the Random Effects tab, and then click **Nest Random Coefficients**.

Figure 8.4 Completed Fit Model Launch Window Showing Random Effects Tab



Random effects are grouped by variety, and the intercept is included as a random component.

9. Click **Run**.

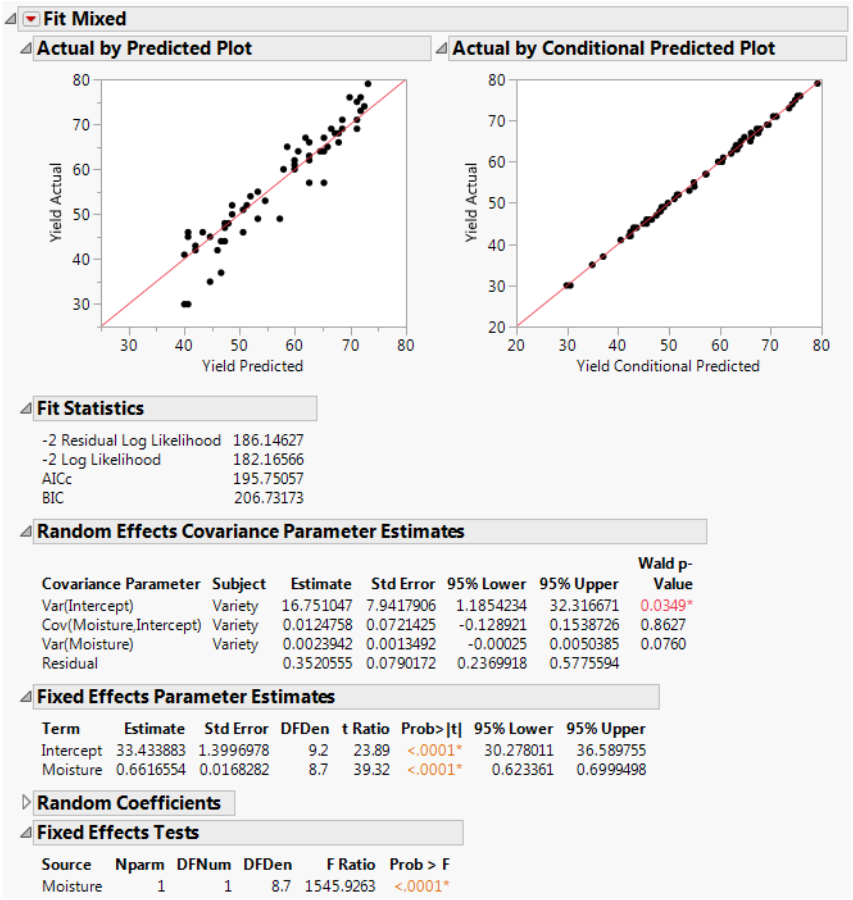
   The Fit Mixed report is shown in Figure 8.5. Note that some of the constituent reports are closed because of space considerations. The Actual by Predicted plot shows no discrepancy in terms of model fit and underlying assumptions.

   Because there are no apparent problems with the model fit, you can now interpret the statistical tests and obtain the regression equation. The effect of moisture upon yield is significant, as shown in the Fixed Effects Tests report. The estimates given in the Fixed Effects Parameter Estimates indicate that the estimated population regression equation is as follows:

   $Yield = 33.43 + 0.66 * Moisture$

   The Random Effects Covariance Parameter Estimates report gives estimates of the variance of the varieties' intercepts, Var(Intercept), and slopes, Var(Moisture), and their covariance, Cov(Moisture, Intercept). In this case, the intercept and slope are not significantly correlated, because the confidence interval for the estimate includes zero. The report also gives an estimate of the residual variance.

Figure 8.5 Fit Mixed Report



Although you have an estimate of the population regression equation, you are also interested in Variety 2's estimated yield.

10. Open the Random Coefficients report to see the estimates of the variety effects for Intercept and Moisture. These coefficients estimate how each variety differs from the population.

Figure 8.6 Random Coefficients Report

Random Coefficients

Variety

Moisture is centered by its mean at 35.5833.

Variety	Intercept	Moisture
1	-0.793305	-0.049211
2	-4.657588	-0.066697
3	1.983891	0.0672228
4	-0.13329	-0.023306
5	0.4076671	-0.019904
6	5.48919	0.0238888
7	-8.722307	0.0564236
8	3.1994336	0.0224337
9	0.6549	0.0233568
10	2.5714086	-0.034207

Covariance Matrix

Random Effect	Intercept	Moisture
Intercept	16.75108	0.012476
Moisture	0.012476	0.002394

In the Model Specification window, the Center Polynomials option is selected by default. Because of this, the Moisture effect is centered at its mean of 35.583, as stated in the note at the top of the Variety report. From the Fixed Effects Parameter Estimates and Random Coefficients reports, you obtain the following prediction equation for Variety 2:

$$\begin{aligned} \text{Yield} &= 33.433 + 0.662(\text{Moisture}) - 4.658 - 0.067(\text{Moisture} - 35.583) \\ &= 31.159 + 0.595(\text{Moisture}) \end{aligned}$$

Variety 2 starts with a lower yield than the population average and increases with Moisture at a slower rate than the population average.

JMP PRO Launch the Mixed Model Personality

Mixed Model is one of several personalities that you can select in the Fit Model launch window. This section describes how you select Mixed Model as your fitting methodology in the Fit Model launch window. Options that are specific to this selection are also covered.

JMP PRO Fit Model Launch Window

You can specify models with fixed effects, random effects, a repeated structure or a combination of those. The options differ based on the nature of the model that you specify.

To fit models using the mixed model personality, select **Analyze > Fit Model** and then select **Mixed Model** from the Personality list. Note that when you enter a continuous variable in the Y list *before* selecting a Personality, the Personality defaults to Standard Least Squares.

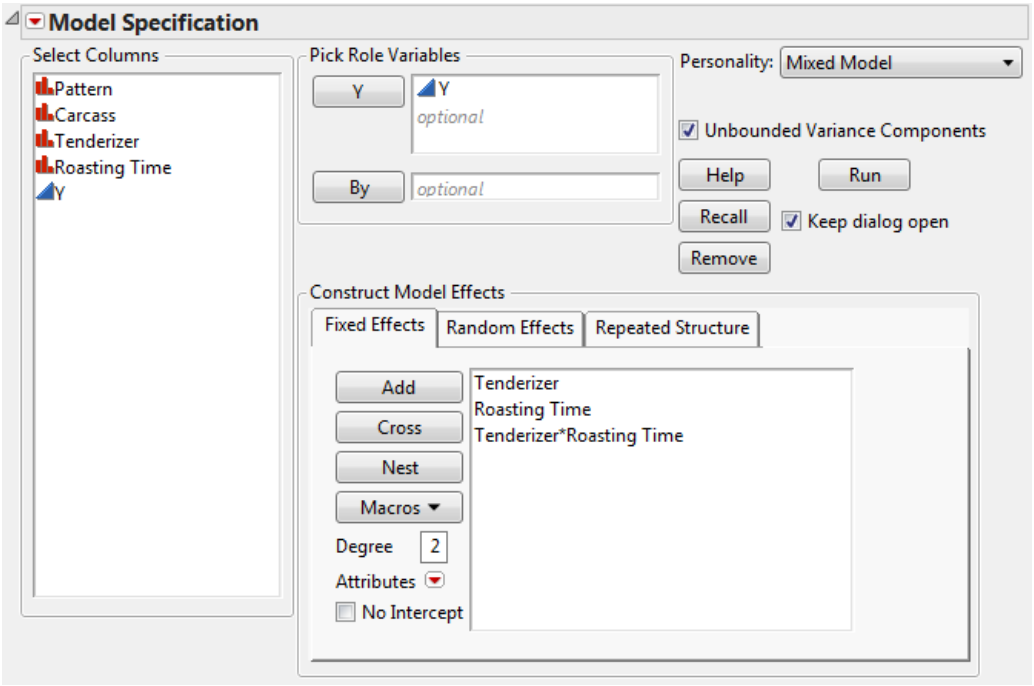
When fitting models using the Mixed Model personality, you can allow unbounded variance components. This means that variance components that have negative estimates are not reported as zero. This option is selected by default. It should remain selected if you are interested in fixed effects, because bounding the variance estimates at zero leads to bias in the tests for fixed effects. See "Negative Variances" on page 190 in the "Standard Least Squares Report and Options" chapter for more information about the Unbounded Variance Components option.

JMP PRO Fixed Effects Tab

Add all fixed effects on the Fixed Effects tab. Use the Add, Cross, Nest, Macros, and Attributes options as needed. For more information about these options, see the "Model Specification" chapter on page 33.

The fixed effects for analysis of the Split Plot.jmp sample data table appear in Figure 8.7. Note that it is possible to have no fixed effects in the model. For an example, see "Spatial Example: Uniformity Trial" on page 401.

Figure 8.7 Fit Model Launch Window Showing Completed Fixed Effects



JMP PRO Random Effects Tab

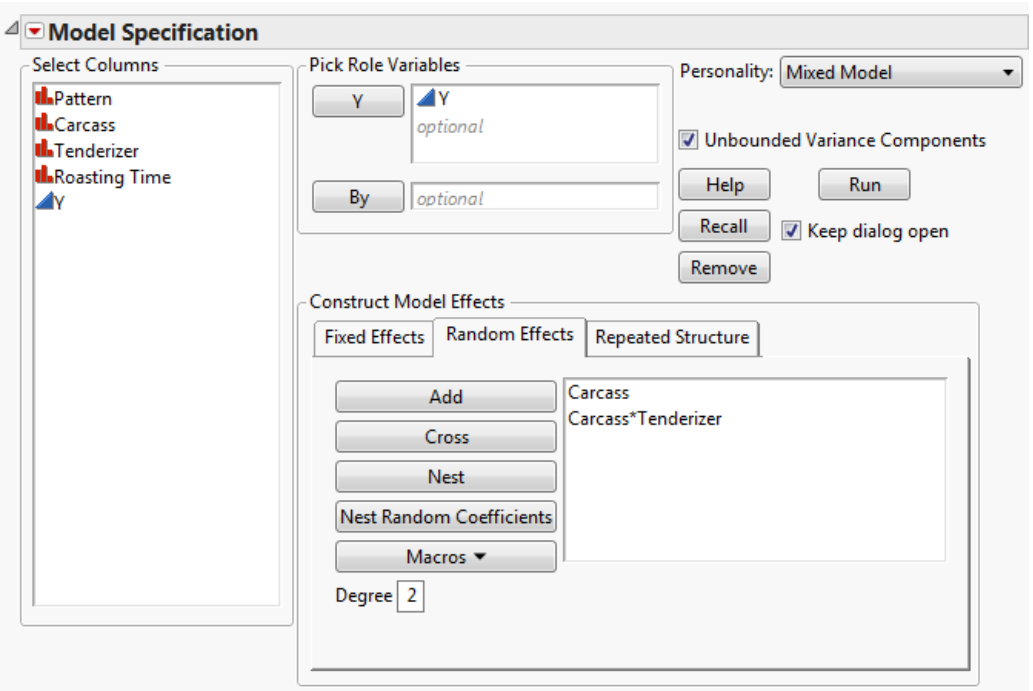
Specify traditional variance component models and random coefficients models using the Random Effects tab.

Variance Components

For a traditional variance component model, specify terms such as random blocks, whole plot error terms, and subplot error terms using the Add, Cross, or Nest options. For more information about these options, see the "Model Specification" chapter on page 33.

Figure 8.8 shows the random effects specification for the Split Plot.jmp sample data where Carcass is a random block. "Split Plot Example" on page 395 describes the example in detail.

Figure 8.8 Fit Model Launch Window Showing Completed Random Effects Tab



Random Coefficients

To construct random coefficients models, use the Nest Random Coefficients button to create groups of random coefficients.

1. Select the continuous columns from the Select Columns list that are predictors.
2. Select the **Random Effects** tab and then **Add**.

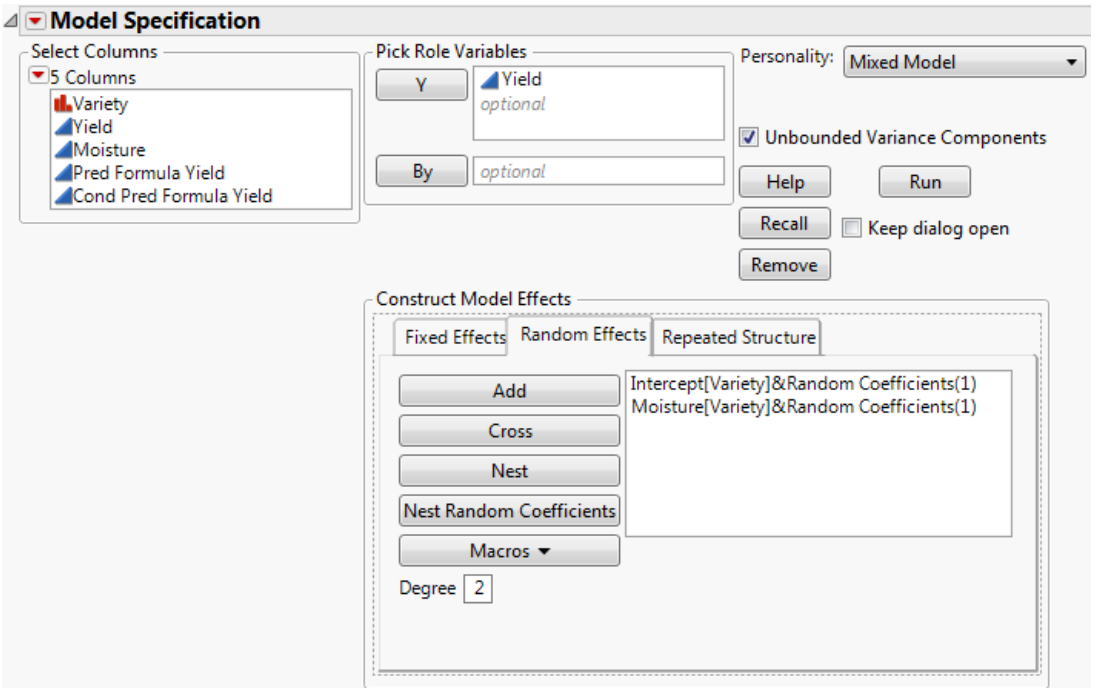
3. Select these effects in the Random Effects tab. Also select the column that contains the random effect whose levels define the individual regression models. This column is essentially the subject in a random statement in SAS PROC MIXED.
4. Click the **Nest Random Coefficients** button.

This last step creates random intercept and random slope effects that are correlated within the levels of the random effect. The subject is nested within the other effects due to the variability among subjects. If you believed that the intercept might be fixed for all groups, you would select Intercept[<*group*>]&Random Coefficients(1) and then click **Remove**.

You can define multiple groups of random coefficients in this fashion, as in hierarchical linear models. This might be necessary when you have both a random batch effect and a random batch by treatment effect on the slope and intercept coefficients. This might also be necessary in a hierarchical linear model: when you have a random student effect and random school effect on achievement scores and students are nested within school.

Random coefficients are modeled using an unstructured covariance structure. Figure 8.9 shows the random coefficients specification for the Wheat.jmp sample data. (See also "Example Using Mixed Model" on page 354.)

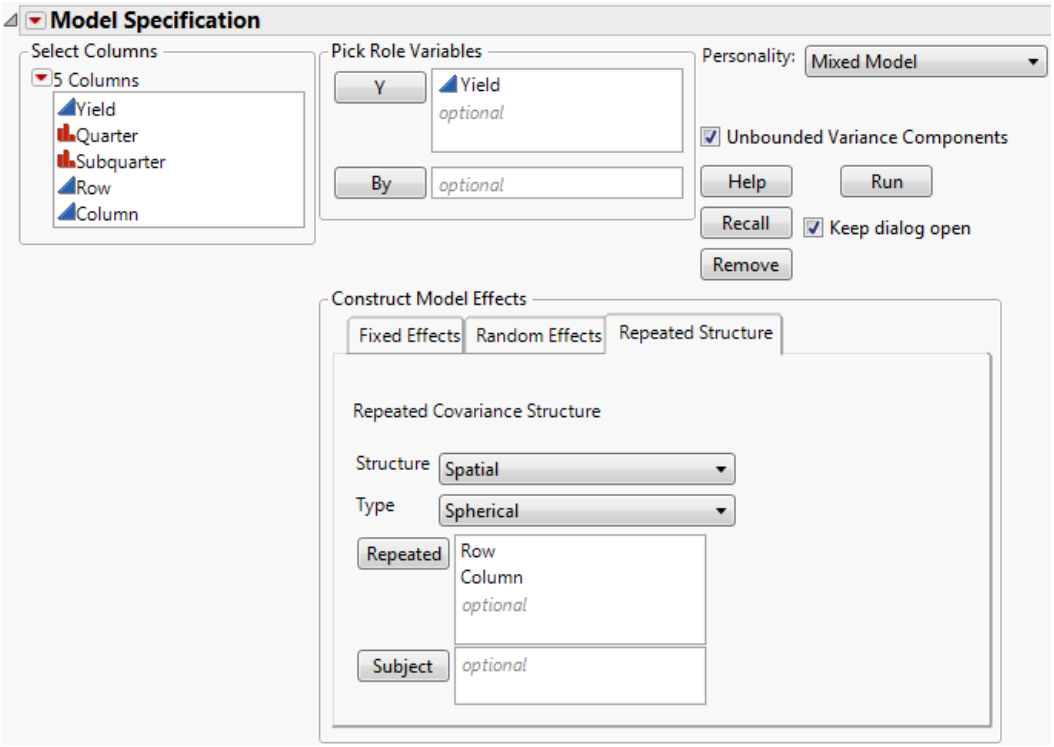
Figure 8.9 Completed Fit Model Launch Window Showing Random Coefficients



JMP PRO Repeated Structure Tab

Use the Repeated Structure tab to select a covariance structure for repeated effects in the model.

Figure 8.10 Completed Fit Model Launch Window Showing Repeated Structure Tab



Structure

The repeated structure is set to Residual by default. The Residual structure specifies that there is no covariance between observations, namely, the errors are independent. All other covariance structures model covariance between observations. For more information about the structures, see "Repeated Measures" on page 419 and "Antedependent Covariance Structure" on page 424 in the Statistical Details section.

Table 8.1 lists the covariance structures available, the requirements for using each structure, and the number of covariance parameters for the given structure. The number of observation times is denoted by *J*.

Table 8.1 Repeated Covariance Structure Requirements

Structure	Repeated Column Type	Required Number of Repeated Columns	Subject	Number of Parameters
Residual	not applicable	0	not applicable	0
Unequal Variances	categorical	1	optional	J
Unstructured	categorical	1	required	$J(J+1)/2$
AR(1)	continuous	1	optional	2
Compound Symmetry	categorical	1	required	2
Antedependent Equal Variance	categorical		required	J
Toeplitz	categorical	1	required	J
Compound Symmetry Unequal Variances	categorical	1	required	$J+1$
Antedependent	categorical		required	$2J-1$
Toeplitz Unequal Variances	categorical	1	required	$2J-1$
Spatial	continuous	2+	optional	
Spatial Anisotropic	continuous	2+	optional	
Spatial with Nugget	continuous	2+	optional	
Spatial Anisotropic with Nugget	continuous	2+	optional	

If you enter a Repeated or Subject column with the Residual structure, those columns are ignored. This alert appears: "Repeated columns and subject columns are ignored when the Residual covariance structure is selected."

Type

When you select one of the spatial covariance structures, a Type list appears from which you select a type of spatial structure. Four Types are available: Power, Exponential, Gaussian, and Spherical. Figure 8.10 shows the Spatial Spherical selection for the Uniformity Trial.jmp sample data.

Repeated

Enter columns that define the repeated measures structure. The modeling types of Repeated columns depend on the covariance structure. See Table 8.1 for more information.

Subject

Enter one or more columns that define the Subject. Subject columns must be categorical.

JMP PRO Data Format

The Mixed Models personality in the Fit Model platform requires that all response measurements be contained in one response column. Repeated measures data are sometimes recorded in multiple columns, where each row is a subject and the repeated measurements are recorded in separate response columns. Data that are in this format must be stacked before running the Mixed Models personality. The Cholesterol.jmp and Cholesterol Stacked.jmp sample data tables illustrate the wide format and the stacked format, respectively. Notice that each row in the wide table corresponds to one level of Patient in the stacked table.

JMP PRO The Fit Mixed Report

The Fit Mixed red triangle menu options enable you to customize reports.

Model Reports Produces reports that relate to the mixed model fit. These reports give estimates and tests for model parameters, as well as fit statistics. See "Model Reports" on page 365.

Multiple Comparisons Opens the Multiple Comparisons dialog window where you can select one or more effects and initial comparisons. This report is available for categorical fixed effects. See "Multiple Comparisons" on page 372.

Compare Slopes (Available only when there is one nominal term, one continuous term, and their interaction effect for the fixed effects.) Produces a report that enables you to compare the slopes of each level of the interaction effect in an analysis of covariance (ANCOVA) model. See "Compare Slopes" on page 372.

Inverse Prediction For one or more values of the response, predicts values of explanatory variables. See "Inverse Prediction" on page 143 in the "Standard Least Squares Report and Options" chapter.

Marginal Model Inference Produces plots based on marginal predicted values and marginal residuals. These plots display the variation due to random effects. See "Marginal Model Inference" on page 372.

Conditional Model Inference Produces plots based on conditional predicted values and conditional residuals. These plots display the variation that remains, once random effects are accounted for. See "Conditional Model Inference" on page 374.

Save Columns Contains options to save various model results as columns in the data table. See "Save Columns" on page 377.

Model Dialog Opens the completed Fit Model launch window for the current analysis. See "Fit Model Launch Window" on page 358.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Model Reports

The reports available under Model Reports are determined by the type of analysis that you conduct. Several of these reports are shown by default.

Fit Statistics Shows report for model fit statistics. See "Fit Statistics" on page 366.

Random Effects Covariance Parameter Estimates Shows report of random effects covariance parameter estimates. This report appears when you specify random effects in the launch window. See "Random Effects Covariance Parameter Estimates" on page 368.

Fixed Effects Parameter Estimates Shows report of fixed effects parameter estimates. See "Fixed Effects Parameter Estimates" on page 369.

Repeated Effects Covariance Parameter Estimates Shows report of repeated effects covariance parameter estimates. This report appears when you specify repeated effects in the launch window. See "Repeated Effects Covariance Parameter Estimates" on page 370.

Indicator Parameterization Estimates (Available only when there are nominal columns among the fixed effects.) Displays the Indicator Function Parameterization report. This report gives parameter estimates for the fixed effects based on a model where nominal fixed effect columns are coded using indicator (SAS GLM) parameterization and are treated as continuous. Ordinal columns remain coded using the usual JMP coding scheme. The SAS GLM and JMP coding schemes are described in "The Factor Models" on page 526 in the "Statistical Details" appendix.

Caution: Standard errors, t-ratios, and other results given in the Indicator Function Parameterization report will differ from those in the Parameter Estimates report. This is because the estimates are estimating different parameters.

Random Coefficients Shows report of random coefficients. This report appears when you specify random effects in the launch window. See "Random Coefficients" on page 371.

Random Effects Predictions Shows report of random effect predictions. This report appears when you specify random effects in the launch window. See "Random Effects Predictions" on page 371.

Fixed Effects Test Shows tests of fixed effects. This report appears when you specify fixed effects in the launch window. See "Fixed Effects Tests" on page 371.

JMP PRO Fit Statistics

The Fit Statistics report gives statistics used for model comparison. For all fit statistics, smaller is better. A likelihood ratio test between two models can be performed if one model is contained within the other. If not, a cautious comparison of likelihoods can be informative. For an example, see "Fit a Spatial Structure Model" on page 401.

"Description of the Fit Statistics Report" uses the following notation:

- Write the mixed model as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\gamma + \varepsilon$$

Here $\mathbf{y}$ is the $nx1$ vector of observations, β is a vector of fixed-effect parameters, γ is a vector of random-effect parameters, and ε is a vector of errors.

- The vectors γ and ε are assumed to have a multivariate normal distribution where

$$E\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

and

$$Var\begin{bmatrix}\gamma\\ \varepsilon\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

- With these assumptions, the variance of $\mathbf{y}$ is given as follows:

$$\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$$

Description of the Fit Statistics Report

-2 Residual Log Likelihood The final evaluation of twice the negative residual log likelihood, the objective function.

$$-2log\,likelihood_{\mathbf{R}}(\mathbf{G},\mathbf{R}) = \log|\mathbf{V}| + \log\left|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right| + \mathbf{r}'\mathbf{V}^{-1}\mathbf{r} + (n-p)\log(2\pi)$$

where

$$\mathbf{r} = \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$$

and p is the rank of X. Use the residual likelihood only for model comparisons where the fixed effects portion of the model is identical. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

-2 Log Likelihood The evaluation of twice the negative log likelihood function. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

Use the log-likelihood for model comparisons in which the fixed, random, and repeated effects differ in any of the models.

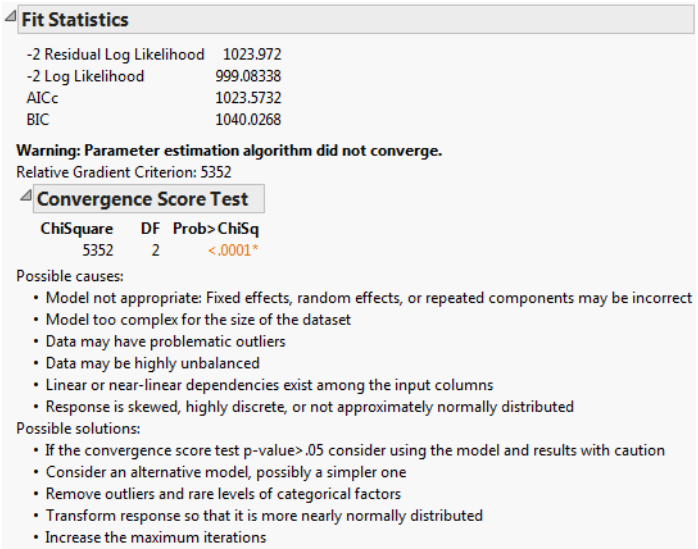
AICc Corrected Akaike's Information Criterion. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

BIC Bayesian Information Criterion. For more details, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

JMP PRO Convergence Score Test

If there are problems with model convergence, a warning message is displayed below the fit statistics. Figure 8.11 shows the warning that suggests the cause and possible solutions to the convergence issue. It also includes a test of the relative gradient at the final iteration. If this test is non-significant, the model might be correct but not fully reaching the convergence criteria. In this case, consider using the model and results with caution. For more information, see "Convergence Score Test" on page 417.

Figure 8.11 Convergence Score Test



JMP PRO Random Effects Covariance Parameter Estimates

The Random Effects Covariance Parameter Estimates report provides details for the covariance parameters of the random effects that you specified in the model.

Covariance Parameter Lists all the covariance parameters of the random effects that you specified in the model.

Note: This column is labeled Variance Component when the model contains only variance components.

Subject Lists the subject from which the block diagonal covariance matrix was formed.

Estimate Gives the estimated variance or covariance component for the effect.

Note: When the model is equivalent to a REML model, a row for a Total covariance parameter is added to the table. The estimate for the Total covariance component is the sum of the positive variance components only.

Std Error Gives the standard error for the covariance component estimate.

95% Lower Gives the lower 95% confidence limit for the covariance component. For more information, see "Confidence Intervals for Variance Components" on page 369.

95% Upper Gives the upper 95% confidence limit for the covariance component. For more information, see "Confidence Intervals for Variance Components" on page 369.

Wald p-Value Gives the *p*-value for the test that the covariance parameter is equal to zero. This column appears only when you have selected Unbounded Variance Components in the Fit Model launch window.

JMP PRO Confidence Intervals for Variance Components

The method used to calculate the confidence limits depends on whether you have selected Unbounded Variance Components in the Fit Model launch window. Note that the Unbounded Variance Components is selected by default.

- If Unbounded Variance Components is selected, Wald-based confidence intervals are computed. These intervals are valid asymptotically, but note that they can be unreliable with small samples. The intervals are wider, which might lead you to mistakenly believe that an estimate is not significantly different from zero.
- If Unbounded Variance Components is not selected, meaning that the parameters have a lower boundary constraint of zero, a Satterthwaite approximation is used (Satterthwaite 1946). The confidence intervals are also bounded at zero.

JMP PRO Fixed Effects Parameter Estimates

The Fixed Effects Parameter Estimates report provides details for the fixed effect parameters specified in the model. For each parameter, the report provides the following details:

- the estimate
- the standard error (Std Error)
- a *t* test for the hypothesis that the estimate equals zero
- a 95% confidence interval on the estimate

The Fixed Effects Parameter Estimates Report contains the following columns:

Term Gives the model term corresponding to the estimated parameter. The first term is always the intercept, unless you selected the No Intercept option in the Fit Model launch window. Continuous columns that are part of higher order terms are centered by default. Nominal or ordinal effects appear with values of levels in brackets. See "The Factor Models" on page 526 for information about the coding of nominal and ordinal terms.

Estimate Gives the parameter estimate for each term. This is the estimate of the term's coefficient in the model.

Std Error Gives an estimate of the standard error for the parameter estimate.

DFDen Gives the denominator degrees of freedom, that is, the degrees of freedom for error, for the effect test. DFDen is calculated using the Kenward-Roger first order approximation. For more information, see "The Kackar-Harville Correction" on page 428.

t Ratio Tests whether the true value of the parameter is zero. The *t* Ratio is the ratio of the estimate to its standard error. Given the usual assumptions about the model, the *t* Ratio has a Student's *t* distribution under the null hypothesis.

Prob>|t| Lists the *p*-value for a two-sided test of the *t* Ratio.

95% Lower Shows the lower 95% confidence limit for the parameter.

95% Upper Shows the upper 95% confidence limit for the parameter.

JMP PRO Repeated Effects Covariance Parameter Estimates

The Repeated Effects Covariance Parameter Estimates report provides details for the covariance parameters of the repeated effects that you specified in the model. It includes the Estimate, Standard Error, and 95% confidence bounds for each parameter. For isotropic spatial models, the covariance parameter estimates have interpretations in terms of range, nugget, and sill. See "Variogram" on page 426.

Note: Variances are covariances of effects with themselves.

The Repeated Effects Covariance Parameter Estimates Report contains the following columns:

Covariance Parameter Lists all the covariance parameters for the repeated effects in the model.

Estimate Gives the estimated variance or covariance component for the effect.

Std Error Gives the standard error for the variance or covariance component estimate.

95% Lower Gives the lower 95% confidence limit for the covariance component. For more information, see "Confidence Intervals for Variance Components" on page 369.

95% Upper Gives the upper 95% confidence limit for the covariance component. For more information, see "Confidence Intervals for Variance Components" on page 369.

JMP PRO Random Coefficients

For each random effect, the Mixed Models personality provides a report showing estimated coefficients and a report showing the matrix of covariance estimates. Each row of the coefficients report corresponds to one level of the random effect. The row shows all coefficient estimates associated with that level of the random effect. The random coefficient estimates are used in conjunction with fixed effect estimates to create predictions for any specific level of the random effect.

JMP PRO Random Effects Predictions

For each random effect in the model, this report gives an estimate known as the *best linear unbiased predictor* (BLUP), its standard error, and a Wald-based confidence interval.

Estimation of the standard errors requires calculation of the BLUP covariance matrix, which can be time-intensive. If the calculation time is noticeable, a progress bar appears.

JMP PRO Fixed Effects Tests

This report shows a significance test for each fixed effect in the model. The test for a given effect tests the null hypothesis that all parameters associated with that effect are zero. An effect might have only one parameter as for a single continuous explanatory variable. In this case, the test is equivalent to the *t* test for that term in the Fixed Effects Parameter Estimates report. A nominal or ordinal effect can have several associated parameters, based on its number of levels. The effect test for such an effect tests whether all of the associated parameters are zero.

The Fixed Effects Tests Report contains the following columns:

Source Lists the fixed effects in the model.

Nparm Shows the number of parameters associated with the effect. A continuous effect has one parameter. The number of parameters for a nominal or ordinal effect is one less than its number of levels. The number of parameters for a crossed effect is the product of the number of parameters for each individual effect.

DFNum Shows the numerator degrees of freedom for the effect test.

DFDen Shows the denominator degrees of freedom for the effect test (the degrees of freedom for error). DFDen is calculated using the Kenward-Roger first order approximation. For more information, see "The Kackar-Harville Correction" on page 428.

F Ratio Gives the computed *F* ratio for testing that the effect is zero.

Prob > F Gives the *p*-value for the effect test.

JMP PRO Multiple Comparisons

The Multiple Comparisons option provides various methods for comparing least squares means of main effects and interaction effects. For more information about the multiple comparisons options, see "Multiple Comparisons" on page 128 in the "Standard Least Squares Report and Options" chapter. For mixed model examples, see "Compare All Treatments in June" on page 391 and "Split Plot Example" on page 395.

Only the fixed effect portion of the model is used in the multiple comparisons. The Multiple Comparisons report shows estimates of the least squares means, standard error, a *t* test of no effect, and a 95% confidence interval. This report is followed by the multiple comparisons test that you select. The All Pairwise Comparisons report provides equivalence tests.

JMP PRO Compare Slopes

The Compare Slopes option appears when there is one nominal term, one continuous term, and their interaction effect for the fixed effects. This option produces a report that enables you to compare the slopes in an analysis of covariance (ANCOVA) model. The report compares the slopes of each level of the interaction effect to the overall slope. The comparison uses analysis of means (ANOM) with the overall average. For more information about the analysis of means (ANOM) report, see "Comparisons with Overall Average" on page 132 in the "Standard Least Squares Report and Options" chapter.

The overall average slope is a weighted average of the slopes, where the weights are inversely proportional to the variances of the slope estimates. These variances are the squared values of the Std Error column in the Differences from Overall Average Slope table.

JMP PRO Marginal Model Inference

The marginal model plots are based on marginal predicted values and marginal residuals.

Actual by Predicted Plot Plots actual values versus values predicted by the model, but without accounting for the random effects. The Actual by Predicted Plot appears by default. See "Actual by Predicted Plot" on page 373.

Residual Plots Provides residual plots that assess model fit, without accounting for the random effects. See "Residual Plots" on page 373.

Profiler, Contour Profiler, Mixture Profiler, Surface Profiler Provides profilers to examine the relationship between the response and the model terms, without accounting for random effects. See "Profilers" on page 374.

JMP PRO Actual by Predicted Plot

The Actual by Predicted plot appears by default. It provides a visual assessment of model fit that reflects variation due to random effects. It plots the observed values of Y against the marginal predicted values of Y. These are the predicted values obtained if you select Save Columns > Prediction Formula.

Denote the linear mixed model by $E[\mathbf{Y}|\gamma] = \mathbf{X}\beta + \mathbf{Z}\gamma$. Here β is the vector of fixed effect coefficients and γ is the vector of random effect coefficients. The *marginal predictions* are the predictions from the fixed effects part of the predictive model, given by $\mathbf{X}\hat{\beta}$.

JMP PRO Residual Plots

Marginal residuals reflect the prediction error based only on the fit of fixed effects. Marginal residuals are the differences between actual values and the predicted values obtained if you select Save Columns > Prediction Formula.

Denote the linear mixed model by $E[\mathbf{Y}|\gamma] = \mathbf{X}\beta + \mathbf{Z}\gamma$. Here β is the vector of fixed effect coefficients and γ is the vector of random effect coefficients. The *marginal residuals* are the residuals from the fixed effects part of the predictive model:

$$\mathbf{r} = \mathbf{Y} - \mathbf{X}\hat{\beta}$$

The Residual Plots option provides three visual methods to assess model fit:

Residual by Predicted Plot Shows the residuals plotted against the predicted values of Y. You typically want to see the residual values scattered randomly about zero.

Residual Quantile Plot Shows the quantiles of the residuals plotted against the quantiles of a standard normal distribution. Also shown is a bar chart of the residuals. If the residuals are normally distributed, the points on the normal quantile plot should approximately fall along the red diagonal line. This type of plot is also called a quantile-quantile plot, or Q-Q plot. The normal quantile plot also shows Lilliefors confidence bounds (Conover 1999).

Residual by Row Plot Shows residuals plotted against row numbers. This plot can help you detect patterns that result from the row ordering of the observations.

JMP PRO Profilers

The plots in the Marginal Model Profiler are based on marginal predicted values. These are the predicted values obtained if you select Save Columns > Prediction Formula.

Denote the linear mixed model by $E[\mathbf{Y}|\gamma] = \mathbf{X}\beta + \mathbf{Z}\gamma$. Here β is the vector of fixed effect coefficients and γ is the vector of random effect coefficients. The *marginal predictions* are the predictions from the fixed effects part of the predictive model, given by $\mathbf{X}\hat{\beta}$.

Note: Marginal model profiler plots show predictions for distinct settings of the fixed effects only. The Profiler shows only cells corresponding to fixed effects. In other profilers, where random effects can be displayed, only the settings of the fixed effects determine the predicted values.

Four types of profilers are provided:

- Profiler
- Contour Profiler
- Mixture Profiler
- Surface Profiler

Options that are appropriate for the model that you are fitting are enabled. See Figure 8.21 for an example of a profiler. See Figure 8.42 for an example of a Surface Profiler. For more details, see the Surface Plot chapter in the *Profilers* book.

JMP PRO Conditional Model Inference

The conditional model diagnostic plots are based on conditional residuals. Conditional residuals reflect the prediction error once both fixed and random effects have been fit.

Actual by Conditional Predicted Plot Plots actual values versus values predicted by the model, accounting for the random effects. When there are random effects, the Actual by Conditional Predicted Plot appears by default. See "Actual by Conditional Predicted Plot" on page 375.

Conditional Residual Plots Provides residual plots that assess model fit, accounting for the random effects. See "Conditional Residual Plots" on page 375.

Conditional Profiler, Conditional Contour Profiler, Conditional Mixture Profiler,

Conditional Surface Profiler Provides profilers to examine the relationship between the response and the model terms, accounting for random effects. See "Conditional Profilers" on page 376.

Variogram Provides a variogram plot that shows the change in covariance as the distance between observations increases. When the Residual structure is selected, you can select the columns to use as temporal or spatial coordinates. See "Variogram" on page 376.

JMP PRO Actual by Conditional Predicted Plot

The Actual by Conditional Predicted plot appears by default. It provides a visual assessment of model fit that accounts for variation due to random effects. It plots the observed values of Y against the conditional predicted values of Y. These are the predicted values obtained if you select Save Columns > Conditional Prediction Formula.

Denote the linear mixed model by $E[\mathbf{Y}|\gamma] = \mathbf{X}\beta + \mathbf{Z}\gamma$. Here β is the vector of fixed effect coefficients and γ is the vector of random effect coefficients. The *conditional predictions* are the predictions obtained from the model given by $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma}$.

JMP PRO Conditional Residual Plots

Conditional residuals reflect the prediction error based on fitting both fixed and random effects. Conditional residuals are the differences between actual values and the conditional predicted values obtained if you select Save Columns > Conditional Prediction Formula.

Denote the linear mixed model by $E[\mathbf{Y}|\gamma] = \mathbf{X}\beta + \mathbf{Z}\gamma$. Here β is the vector of fixed effect coefficients and γ is the vector of random effect coefficients. The *conditional residuals* are given as follows:

$$\mathbf{r} = \mathbf{Y} - (\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma})$$

The Conditional Residual Plots option provides three visual methods to assess model fit:

Conditional Residual by Predicted Plot Shows the conditional residuals plotted against the conditional predicted values of Y. You typically want to see the conditional residual scattered randomly about zero.

Conditional Residual Quantile Plot Shows the quantiles of the conditional residuals plotted against the quantiles of a standard normal distribution. Also shown is a bar chart of the conditional residuals. If the conditional residuals are normally distributed, the points on the normal quantile plot should approximately fall along the red diagonal line. This type of plot is also called a quantile-quantile plot, or Q-Q plot. The normal quantile plot also shows Lilliefors confidence bounds (Conover 1999).

Conditional Residual by Row Plot Shows conditional residuals plotted against row numbers. This plot can help you detect patterns that result from the row ordering of the observations.

JMP PRO Conditional Profilers

The conditional model profiler plots are based on conditional predicted values. These are the predicted values obtained if you select Save Columns > Conditional Prediction Formula.

Denote the linear mixed model by $E[\mathbf{Y}|\gamma] = \mathbf{X}\beta + \mathbf{Z}\gamma$. Here β is the vector of fixed effect coefficients and γ is the vector of random effect coefficients. The *conditional predictions* are the predictions obtained from the model given by $\mathbf{X}\hat{\beta} + \mathbf{Z}\hat{\gamma}$.

Four types of profilers are provided:

- Conditional Profiler
- Conditional Contour Profiler
- Conditional Mixture Profiler
- Conditional Surface Profiler

Options that are appropriate for the model that you are fitting are enabled. See Figure 8.21 for an example of a Profiler. See Figure 8.42 for an example of a Surface Profiler. For more information about the profiler, see the Surface Plot chapter in the *Profilers* book.

JMP PRO Variogram

A Variogram plot describes the spatial or temporal correlation of observations in terms of their distance. The plot shows the semivariance as a function of distance or time. The theoretical semivariance is one half of the variance of the difference between response values at locations that are a given distance apart. Note that semivariance and correlation at a given distance are inversely related. If the correlation between values at a given distance is small, the semivariance between observations at that distance is large.

When you specify any isotropic Repeated Structure (AR(1), Spatial, or Spatial with Nugget) in the Fit Model window, a Variogram plot is shown by default. If you specify the Residual structure, selecting the Variogram option in the red triangle menu enables you to select the continuous columns to be used in calculating the variogram. You can include any number of columns that describe the spatial or temporal structure of your data.

The initial Variogram report shows a plot of the empirical semivariance against distance. For additional background and details, see "Antedependent Covariance Structure" on page 424.

Semivariance Curves for Isotropic Structures

Semivariance curves are provided for the isotropic covariance structures: AR(1), Power, Exponential, Gaussian, and Spherical. For the spatial structures, curves are provided for models with and without nuggets.

The curves for the theoretical models are fit using the covariance parameter estimates. For the underlying formulas, see Chilès and Delfiner (2012) and Cressie (1993).

Use the theoretical models to determine whether your data conform to your selected isotropic structure. If you have selected the Residual structure, you can use the empirical variogram to determine whether your data exhibit some temporal or spatial structure. If the points appear to follow a horizontal line, this suggests that the correlation does not change with distance and that the Residual structure is appropriate. If the points show a pattern, fitting various isotropic models might suggest an appropriate Repeated structure with which to refit your model.

JMP PRO Nugget

The *nugget* is the vertical jump from the value of 0 at the origin of the variogram to the value of the semivariance at a very small separation distance. A variogram model with a nugget has a discontinuity at the origin. The value of the theoretical curve for distances just above 0 is the nugget.

JMP PRO Variogram Options

AR(1) Plots a variogram for an AR(1) covariance structure.

Spatial Plots a variogram for an Exponential, Gaussian, Power, or Spherical covariance structure.

Spatial with Nugget Plots a variogram for an Exponential, Gaussian, Power, or Spherical covariance structure with nugget.

JMP PRO Save Columns

Each option in the Save Columns menu adds a new column to the data table.

Prediction Formula Creates a new column called Pred Formula <colname> that contains both the formula and the marginal mean predicted values. A Predicting column property is added, noting the source of the prediction. See "Marginal Model Inference" on page 372.

Standard Error of Predicted Creates a new column called StdErr Pred <colname> that contains standard errors for the predicted marginal mean responses.

Residuals Creates a new column called Residual <colname> that contains the observed response values minus their marginal mean predicted values. See "Marginal Model Inference" on page 372.

Conditional Prediction Formula Creates a new column called Cond Pred Formula <colname> that contains both the formula and the conditional mean predicted values. A Predicting column property is added, noting the source of the prediction. See "Conditional Profilers" on page 376.

Standard Error of Conditional Predicted Creates a new column called StdErr Cond Pred <colname> that contains standard errors for the predicted conditional mean responses.

Conditional Residuals Creates a new column called Cond Residual <colname> that contains the observed response values minus their conditional mean predicted values. See "Conditional Model Inference" on page 374.

JMP PRO Additional Examples of the Mixed Models Personality

JMP PRO Repeated Measures Example

Consider the Cholesterol Stacked.jmp sample data table. A study was performed to test two new cholesterol drugs against a control drug. Twenty patients with high cholesterol were randomly assigned to each of four treatments (the two experimental drugs, the control, and a placebo). Each patient's total cholesterol was measured at six times during the study: the first day in April, May, and June in the morning and afternoon. You are interested in whether either of the new drugs is effective at lowering cholesterol and in whether time and treatment interact.

JMP PRO Background

Two methods have historically been used to analyze such a design:

- Multivariate analysis of variance (MANOVA)

- A split-plot in time univariate analysis of variance (ANOVA) with either the Huynh-Feldt (1976) or Greenhouse-Geisser (1959) correction

Both of these options are available using the MANOVA personality in Fit Model. These two options are the two extremes for modeling the covariance structure. The MANOVA analysis assumes an unstructured covariance structure where all variances and covariances are estimated individually. The independent split-plot in time analysis assumes that all errors are independent. In the Gaussian data case, this is equivalent to assuming a compound symmetry covariance structure.

These two models can result in vastly different conclusions about the treatment effects. When you assume a complex covariance structure, information in the data is used to estimate the covariance parameters. If you fit too many covariance parameters, you run the risk of overfitting your model. When you model repeated measures data, you must find a covariance structure that balances these issues.

- When the model is overfit, the power to detect differences is smaller than if you were to assume a less complex covariance structure.
- When the model is underfit, Type I error control is lost. In some cases, this leads to inflated rejection rates. In other cases, decreased rejection rates occur due to inflated variance.

JMP PRO Covariance Structures

The Mixed Model personality fits a variety of covariance structures. For repeated measures in time, both the Toeplitz covariance structure and the first-order autoregressive (AR(1)) covariance structures often provide appropriate correlation structures. These structures allow for correlated observations without overfitting the model. The AR(1) assumes a common variance parameter, whereas the Toeplitz covariance matrix with unequal variances estimates unique variances for each unit of the repeated measure variable. See "Repeated Covariance Structure Requirements" on page 363.

In this example, you fit the four covariance structures. The number of observation times, J, is equal to six.

- "Covariance Structure: Unstructured" on page 380. The Unstructured model fits all covariance parameters, $J(J+1)/2$ in total. In this example, the model fits 21 variances.
- "Covariance Structure: Residual" on page 383. The Residual model is equivalent to the usual variance components structure. In this example, the model fits two variances.
- "Covariance Structure: Toeplitz" on page 385. The Toeplitz model fits $2J$-1 covariance parameters. In this example, the model fits 11 variances.
- "Covariance Structure: AR(1)" on page 387. This model fits two covariance parameters. One parameter determines the variance and the other determines how the covariance changes with time.

You use AICc to evaluate model fits. The BIC criterion can also be used. In this case, the same model is chosen by both criteria. You select a best covariance structure and then continue to do additional analysis:

- "Further Analysis Using AR(1) Structure" on page 389
- "Regression Model for AR(1) Model Example" on page 394

Tip: Leave the Fit Model launch window open as you work through this example.

JMP PRO Data Structure

The Cholesterol.jmp data table is in a format that is typically used for recording repeated measures data. To use the Mixed Model personality to analyze these data, each cholesterol measurement needs to be in its own row, as in Cholesterol Stacked.jmp. To construct Cholesterol Stacked.jmp, the data in Cholesterol.jmp were stacked using Tables > Stack.

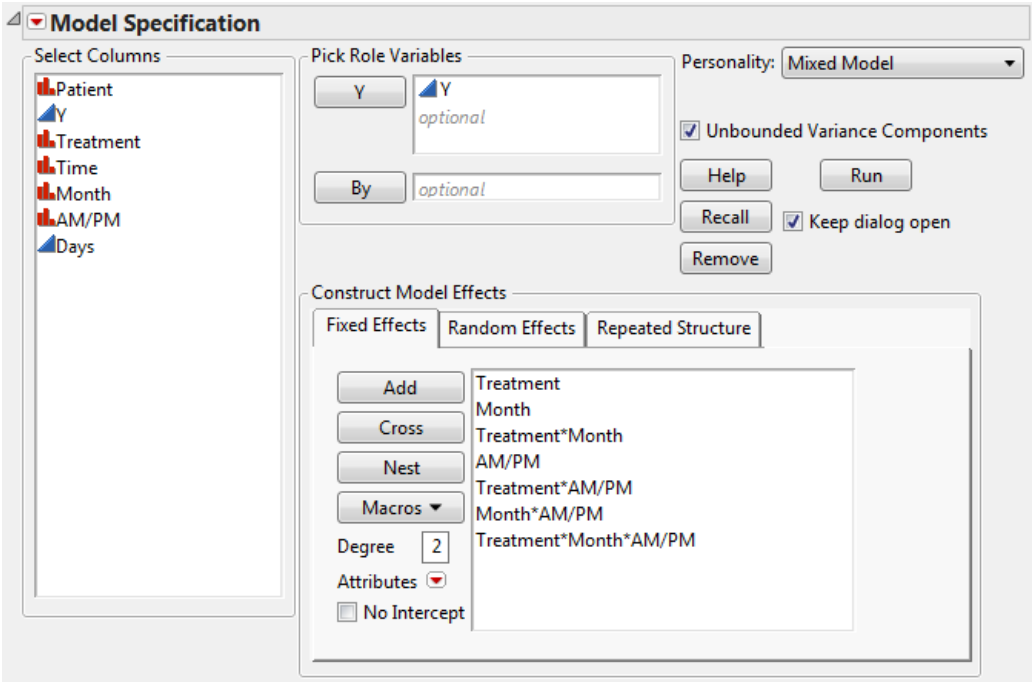
The Days column in the stacked table was constructed using a formula. The Days column gives the number days into the study when the cholesterol measurement was taken. Its modeling type is continuous. This is necessary because the AR(1) covariance structure requires the repeated effect be continuous.

JMP PRO Covariance Structure: Unstructured

Begin by fitting a model using an Unstructured covariance structure.

1. Select **Help > Sample Data Library** and open Cholesterol Stacked.jmp.
2. Select **Analyze > Fit Model**.
3. Select **Keep dialog open** so that you can return to the launch window in the next example.
4. Select Y and click **Y**.
5. Select **Mixed Model** from the Personality list.
6. Select Treatment, Month, and AM/PM, and then select **Macros > Full Factorial**.

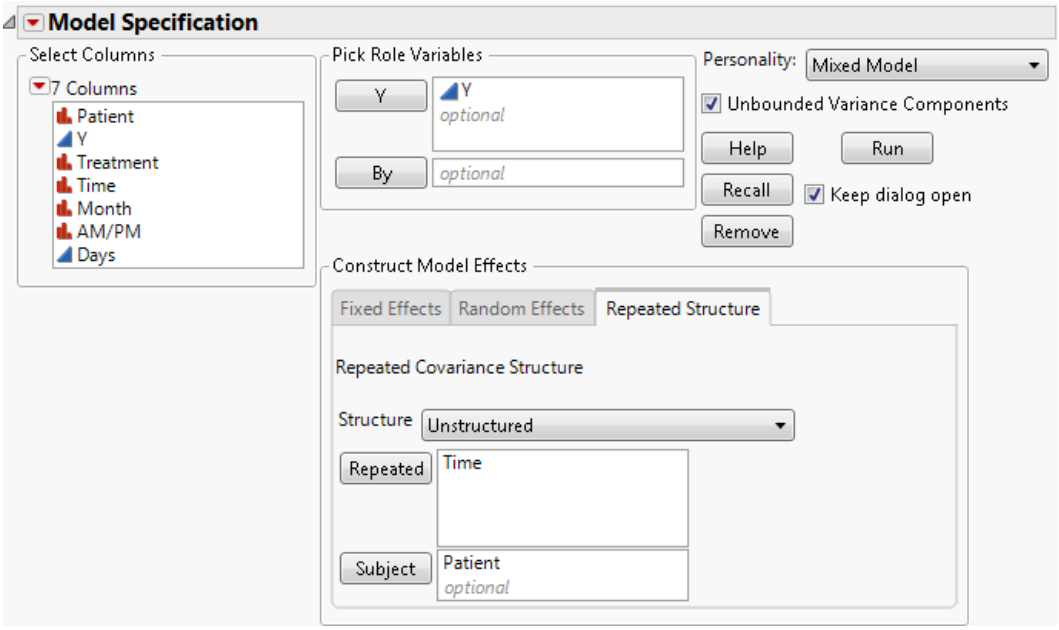
Figure 8.12 Fit Model Launch Window Showing Completed Fixed Effects Tab



7. Select the **Repeated Structure** tab.
8. Select **Unstructured** from the Structure list.
9. Select Time and click **Repeated**. The **Repeated** column defines the repeated measures within a subject.
10. Select Patient and click **Subject**.

Note: The Unstructured covariance model does not allow the repeated structure variables to assume duplicate values. Suppose that, in this example, the subject was nested within treatment, and that the patients had been numbered using the values 1, 2, 3, 4, and 5 within each treatment. A warning would be given when you run this analysis. You would need to renumber the patients to have different identifiers for each value of the Repeated variable. Or you would need to create a column in the data table that represents nesting within treatment and enter this effect as Subject.

Figure 8.13 Fit Model Launch Window Showing Completed Repeated Structure Tab



11. Click **Run**.

    The Fit Mixed report is shown in Figure 8.14. Because you want to compare your three models using AICc or BIC, you are interested in the Fit Statistics report. The AICc for the unstructured model is 703.84.

    The Repeated Effects Covariance Parameter Estimates report shows estimates of all 21 covariance parameters. As you would expect, observations taken closer in time have higher covariance than those farther apart. Also, variance increases with time.

Figure 8.14 Fit Mixed Report - Unstructured Covariance Structure

Fit Mixed

Actual by Predicted Plot

Fit Statistics

-2 Residual Log Likelihood	554.98818
-2 Log Likelihood	557.89101
AICc	703.83696
BIC	773.32814

Repeated Effects Covariance Parameter Estimates

Repeated Effect: Time
Subject: Patient

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
Var(April AM)	18.725	6.620287	5.7494755	31.700524
Cov(April PM,April AM)	18.354883	6.6035619	5.41214	31.297627
Var(April PM)	19.268932	6.8125961	5.9164889	32.621375
Cov(May AM,April AM)	9.2756707	8.4629179	-7.311344	25.862685
Cov(May AM,April PM)	5.5074057	8.3703999	-10.89828	21.913088
Var(May AM)	56.603346	20.012305	17.37995	95.826742
Cov(May PM,April AM)	9.4147225	8.5038582	-7.252533	26.081978
Cov(May PM,April PM)	6.6230804	8.4532301	-9.944946	23.191107
Cov(May PM,May AM)	55.365522	19.83529	16.489068	94.241976
Var(May PM)	57.058089	20.17308	17.519578	96.5966
Cov(June AM,April AM)	1.1945478	8.6266095	-15.7133	18.102392
Cov(June AM,April PM)	0.3183447	8.7461243	-16.82374	17.460433
Cov(June AM,May AM)	1.106725	14.992148	-28.27735	30.490795
Cov(June AM,May PM)	0.6455101	15.050552	-28.85303	30.144049
Var(June AM)	63.512276	22.45498	19.501324	107.52323
Cov(June PM,April AM)	1.5810194	8.768799	-15.60551	18.76755
Cov(June PM,April PM)	0.7647113	8.8882624	-16.65596	18.185386
Cov(June PM,May AM)	0.9262595	15.232066	-28.92804	30.78056
Cov(June PM,May PM)	0.6543895	15.292237	-29.31784	30.626624
Cov(June PM,June AM)	63.878685	22.700344	19.386829	108.37054
Var(June PM)	65.568481	23.181958	20.132677	111.00428

Fixed Effects Parameter Estimates

Fixed Effects Tests

Source	Nparm	DFNum	DFDen	F Ratio	Prob > F
Treatment	3	3	16.0	274.96713	<.0001*
Month	2	2	15.0	340.48166	<.0001*
Treatment*Month	6	6	18.3	123.47461	<.0001*
AM/PM	1	1	16.0	360.93594	<.0001*
Treatment*AM/PM	3	3	16.0	0.6339843	0.6038
Month*AM/PM	2	2	15.0	1.1988248	0.3289
Treatment*Month*AM/PM	6	6	18.3	1.1642781	0.3671

JMP PRO Covariance Structure: Residual

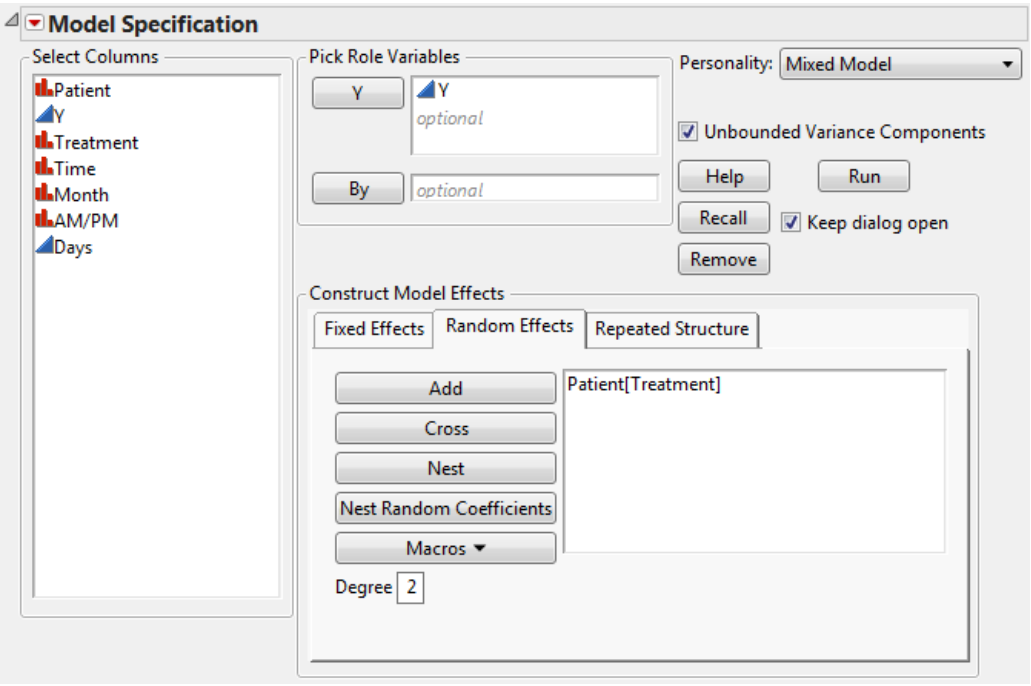
The Residual covariance structure is appropriate when you fit a split-plot model.

1. Complete step 1 through step 7 in "Covariance Structure: Unstructured" on page 380.

2. On the **Repeated Structure** tab, select **Residual** from the Structure list.
3. If you are continuing from the previous example, remove Time and Patient.

   Otherwise, a warning appears: "Repeated columns and subject columns are ignored when the Residual covariance structure is selected." You are given the option to click **OK** to continue the analysis.
4. Select the **Random Effects** tab.
5. Select Patient and click **Add**.
6. Select Patient in the Random Effects area, select the Treatment column, and then click **Nest**.

Figure 8.15 Fit Model Launch Window Showing Completed Random Effects Tab

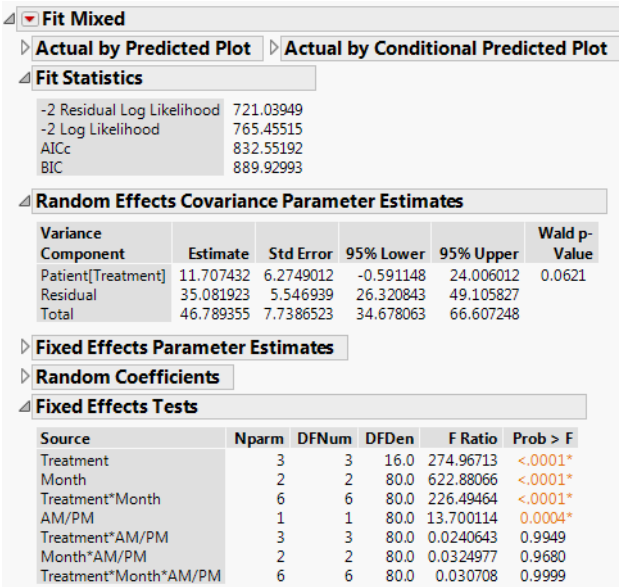


7. Click **Run**.

   The Fit Mixed report is shown in Figure 8.16. The Fit Statistics report shows that the AICc for the Residual model is 832.55, as compared to 703.84 for the Unstructured model.

   The estimates of the two covariance parameters are shown in the Random Effects Covariance Parameter Estimates report. These are estimates of the variance of Patient within Treatment, and of the Residual variance.

Figure 8.16 Fit Mixed Report - Residual Error Covariance Structure



Fit Mixed

Actual by Predicted Plot | Actual by Conditional Predicted Plot

Fit Statistics

-2 Residual Log Likelihood	721.03949
-2 Log Likelihood	765.45515
AICc	832.55192
BIC	889.92993

Random Effects Covariance Parameter Estimates

Variance Component	Estimate	Std Error	95% Lower	95% Upper	Wald p-Value
Patient[Treatment]	11.707432	6.2749012	-0.591148	24.006012	0.0621
Residual	35.081923	5.546939	26.320843	49.105827	
Total	46.789355	7.7386523	34.678063	66.607248	

Fixed Effects Parameter Estimates

Random Coefficients

Fixed Effects Tests

Source	Nparm	DFNum	DFDen	F Ratio	Prob > F
Treatment	3	3	16.0	274.96713	<.0001*
Month	2	2	80.0	622.88066	<.0001*
Treatment*Month	6	6	80.0	226.49464	<.0001*
AM/PM	1	1	80.0	13.700114	0.0004*
Treatment*AM/PM	3	3	80.0	0.0240643	0.9949
Month*AM/PM	2	2	80.0	0.0324977	0.9680
Treatment*Month*AM/PM	6	6	80.0	0.030708	0.9999

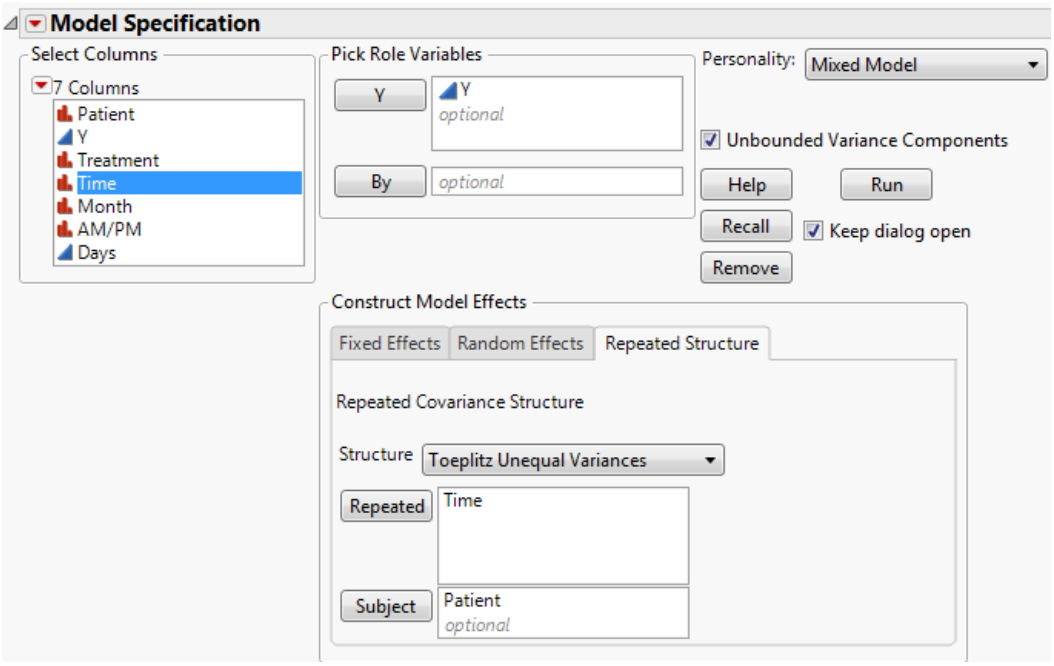
JMP PRO Covariance Structure: Toeplitz

Fit the model using the Toeplitz Unequal Variances structure.

1. Complete step 1 through step 6 in "Covariance Structure: Unstructured" on page 380.
2. If you are continuing from the previous example, select Patient[Treatment] on the **Random Effects** tab and then click **Remove**.

   If you include both random effects and repeated effects, there is often insufficient data to estimate both effects.
3. Select the **Repeated Structure** tab.
4. Select **Toeplitz Unequal Variances** from the Structure list.
5. Select Time and click **Repeated**.
6. Select Patient and click **Subject**.

Figure 8.17 Fit Model Launch Window Showing Completed Repeated Structure Tab



7. Click **Run**.

Figure 8.18 Fit Mixed Report - Toeplitz Unequal Variances Structure

Fit Mixed

Actual by Predicted Plot

Fit Statistics

-2 Residual Log Likelihood	659.09915
-2 Log Likelihood	688.03015
AICc	788.03015
BIC	855.59236

Repeated Effects Covariance Parameter Estimates

Subject: Patient

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
Toeplitz Correlation(1)	0.6229833	0.0856427	0.4551266	0.79084
Toeplitz Correlation(2)	0.2555088	0.1463162	-0.031266	0.5422833
Toeplitz Correlation(3)	0.0961411	0.1887658	-0.273833	0.4661153
Toeplitz Correlation(4)	-0.111291	0.2248465	-0.551982	0.3293997
Toeplitz Correlation(5)	0.2426279	0.1955585	-0.14066	0.6259155
Variance(April AM)	17.550867	6.9769855	3.8762269	31.225508
Variance(April PM)	24.652202	8.3212859	8.3427817	40.961623
Variance(May AM)	73.9439	25.926103	23.129672	124.75813
Variance(May PM)	63.670311	21.769686	21.00251	106.33811
Variance(June AM)	69.454988	22.577319	25.204256	113.70572
Variance(June PM)	52.479116	19.167505	14.911497	90.046735

Fixed Effects Parameter Estimates

Fixed Effects Tests

Source	Nparm	DFNum	DFDen	F Ratio	Prob > F
Treatment	3	3	14.1	230.28909	<.0001*
Month	2	2	23.7	363.85606	<.0001*
Treatment*Month	6	6	26.2	134.01683	<.0001*
AM/PM	1	1	5.2	106.74501	0.0001*
Treatment*AM/PM	3	3	5.2	0.1874977	0.9007
Month*AM/PM	2	2	26.9	0.0367022	0.9640
Treatment*Month*AM/PM	6	6	34.6	0.034727	0.9998

Note: The Mixed Models personality in JMP reports correlations, whereas PROC MIXED in SAS reports covariances.

The Fit Statistics report shows that the AICc for the Toeplitz with Unequal Variances model is 788.03. Compare this number to 832.55 for the Residual Model and 703.84 for the Unstructured model.

The Toeplitz Unequal Variances structure requires the estimation of eleven covariance parameters. These estimates are shown in the Repeated Effects Covariance Parameter Estimates report. The Toeplitz correlation estimates are shown, followed by the variance estimates for each time point. See "Repeated Measures" on page 419 for information about how this matrix is parameterized.

JMP PRO Covariance Structure: AR(1)

Finally, fit the AR(1) structure.

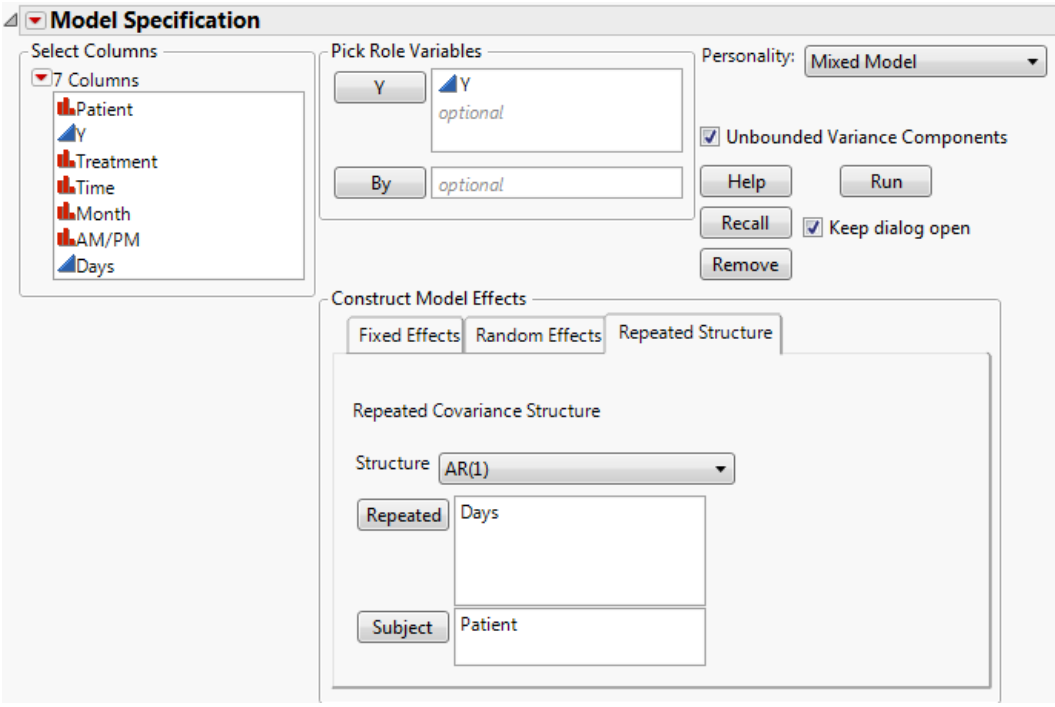
1. Complete step 1 through step 6 in "Covariance Structure: Unstructured" on page 380.

2. If you are continuing from the previous example, select Time in the **Repeated** box and then click **Remove**.

   AR(1) requires a continuous variable for the repeated value.

3. Select **AR(1)** from the Structure list.
4. Select Days and click **Repeated**.

Figure 8.19 Fit Model Launch Window Showing Completed Repeated Structure Tab



5. Click **Run**.

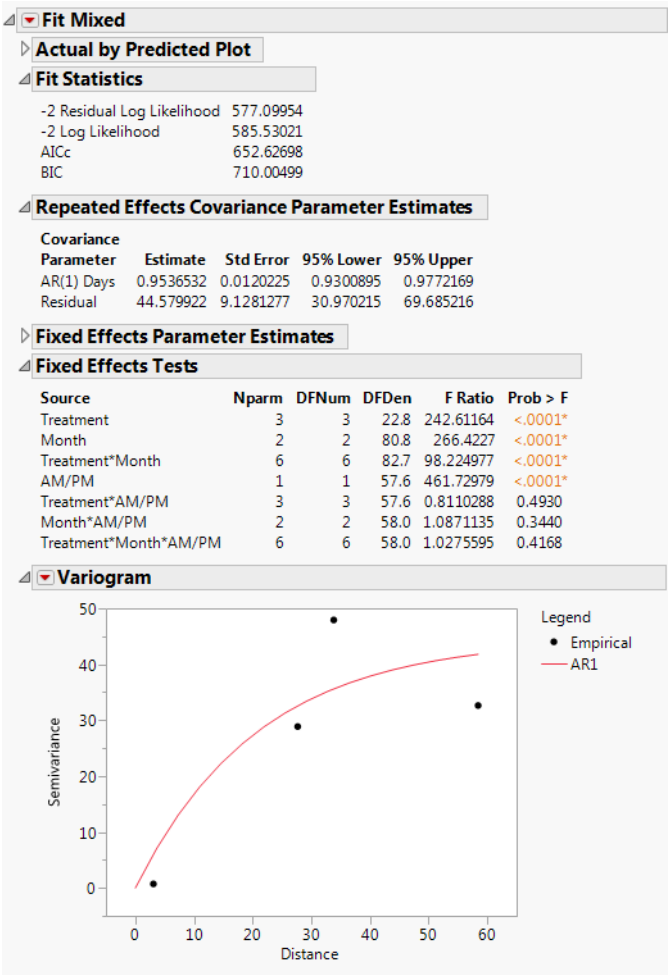
   The Fit Mixed report is shown in Figure 8.20. The Fit Statistics report shows that the AICc for the AR(1) model is 652.63. Compare this number to 832.55 for the Residual model, 703.84 for the Unstructured model, and 788.03 for the Toeplitz Unequal Variances model. Based on the AICc criterion, the AR(1) model is the best of the four models.

   The AR(1) structure requires the estimation of two covariance parameters. These estimates are shown in the Repeated Effects Covariance Parameter Estimates report. The AR(1) Days parameter estimate is an estimate of ρ, the correlation parameter in the AR(1) structure.

   The Variogram plot shows the empirical semivariances and the curve for the AR(1) model. Since there are only five nonzero values for Days, only four distance classes are possible and only four points are shown. The AR(1) structure seems appropriate. To explore other

structures, select options from the red triangle menu next to Variogram. For more information about Variogram options, see "Variogram" on page 376.

Figure 8.20 Fit Mixed Report - AR(1) Covariance Structure



JMP PRO Further Analysis Using AR(1) Structure

Because the AR(1) model gives the best fit, you adopt it as your model and proceed with your analysis. The Fixed Effects Tests report indicates that there is a significant interaction between Treatment and Month as well as a main effect of AM/PM. Here, we explore these significant effects.

1. Click the red triangle next to Fit Mixed and select **Marginal Model Inference > Profiler**.

The Marginal Model Profiler report (Figure 8.21) enables you to see the effect on cholesterol levels (Y) for various settings of Treatment, Month, and AM/PM.

2. In the plot for Month, drag the vertical dotted red line from April to May and then to June.

   Notice that the predicted AM measurements for Y decrease over the three months from a mean of 277.4 in April to a mean of 177.7 in June.

3. In the plot for Treatment, drag the vertical dotted red line from A to B.

   By dragging the line in the plot for Month from April to June, you see that, for Treatment B, the predicted AM mean for Y decreases from 276.8 in April to 191.2 in June.

4. In the plot for Treatment, drag the vertical dotted line to Control and then to Placebo.

   Notice that when you set Treatment to Control or Placebo, you see virtually no change over the three months (Figure 8.22).

   Next, you explore the effect of AM/PM.

5. Set Treatment and Month to all twelve combinations of their levels by dragging the vertical red lines.

   For all twelve combinations, the predicted cholesterol level is consistently higher in the afternoon than in the morning, demonstrating the main effect.

Note that Treatment A seems to result in lower cholesterol readings in May than Treatment B does. If this effect is significant, it might indicate that Treatment A acts more quickly than B. The next section, "Compare All Treatments in June" on page 391, shows you how to evaluate the treatments.

Figure 8.21 Marginal Profiler Plot for Treatment A

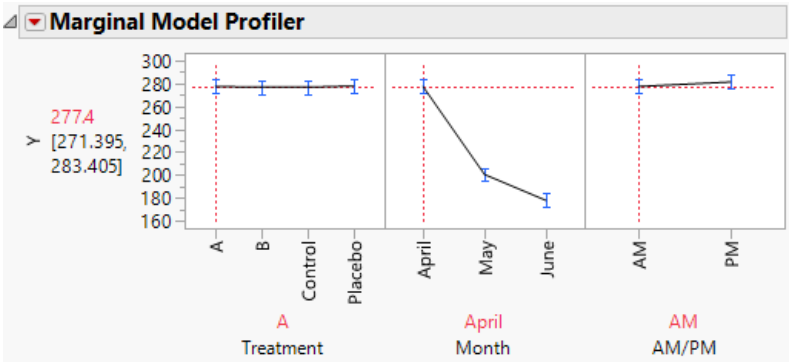
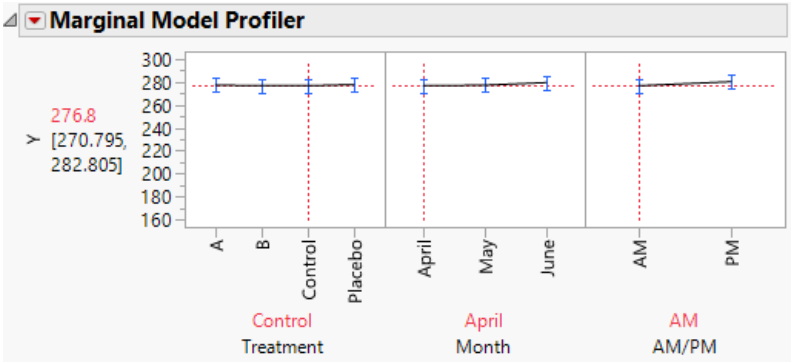


Figure 8.22 Marginal Profiler Plot for Control



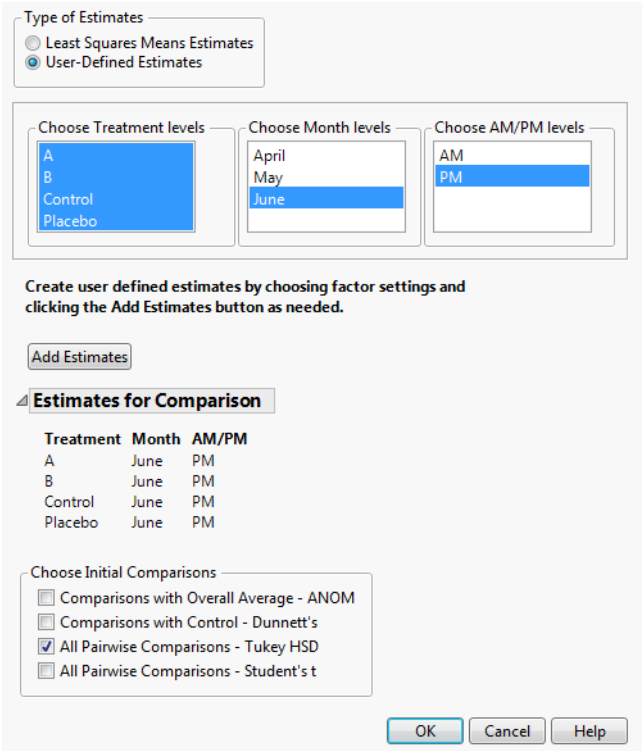
Compare All Treatments in June

The study is conducted over the months of April, May, and June. You are interested in which treatments differ on the PM measurement in June.

1. Click the red triangle next to Fit Mixed and select **Multiple Comparisons**.
2. Under Types of Estimates, select **User-Defined Estimates**.
3. From the Choose Treatment levels panel, select all four treatment types.
4. From the Choose Month levels panel, select June.
5. From the Choose AM/PM levels panel, select PM.
6. Click Add Estimates.
7. From the Choose Initial Comparisons list, select **All Pairwise Comparisons - Tukey HSD**.

   The Multiple Comparisons window should appear as shown in Figure 8.23.

Figure 8.23 Completed Multiple Comparisons Window



8. Click **OK**.

Figure 8.24 Tukey HSD All Pairwise Comparisons Report for All Treatments for June PM

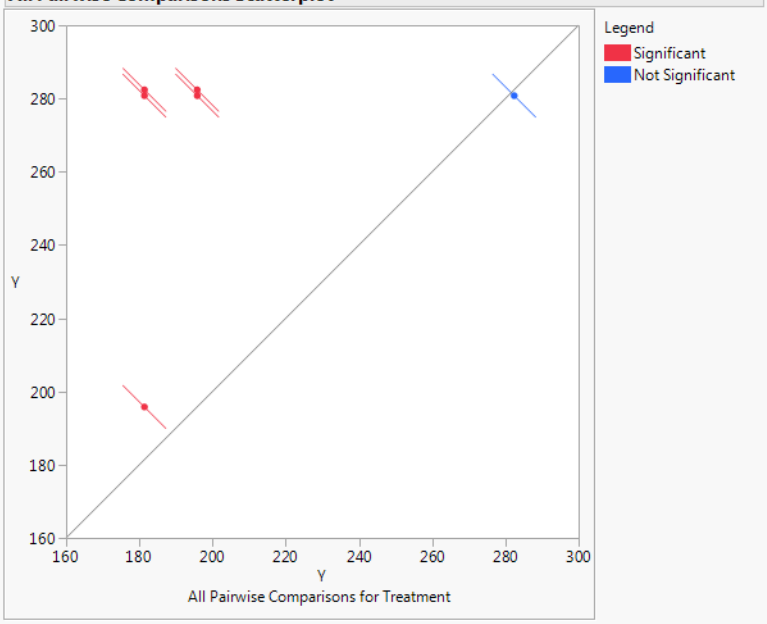
Tukey HSD All Pairwise Comparisons

Month = June, AM/PM = PM
Quantile = 2.76931, Adjusted DF = 22.8, Adjustment = Tukey

All Pairwise Differences

Treatment	-Treatment	Difference	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
A	B	-14.414	4.222792	-3.41	0.0120*	-26.108	-2.7196
A	Control	-101.014	4.222792	-23.92	<.0001*	-112.708	-89.3196
A	Placebo	-99.414	4.222792	-23.54	<.0001*	-111.108	-87.7196
B	Control	-86.600	4.222792	-20.51	<.0001*	-98.294	-74.9058
B	Placebo	-85.000	4.222792	-20.13	<.0001*	-96.694	-73.3058
Control	Placebo	1.600	4.222792	0.38	0.9810	-10.094	13.2942

All Pairwise Comparisons Scatterplot



The Tukey HSD All Pairwise Comparisons report shows an All Pairwise Differences report and an All Pairwise Comparisons Scatterplot. All treatments other than the Control and Placebo differ significantly on the June PM measurements.

Consider the difference between treatments A and B. The difference in means is -14.414 and the confidence interval ranges from -26.108 to -2.7196. You conclude that the reduction in cholesterol measurements due to treatment A exceeds the reduction by treatment B by somewhere between 2.7 and 26.1 points. Both treatments A and B are highly effective compared to the Control and Placebo.

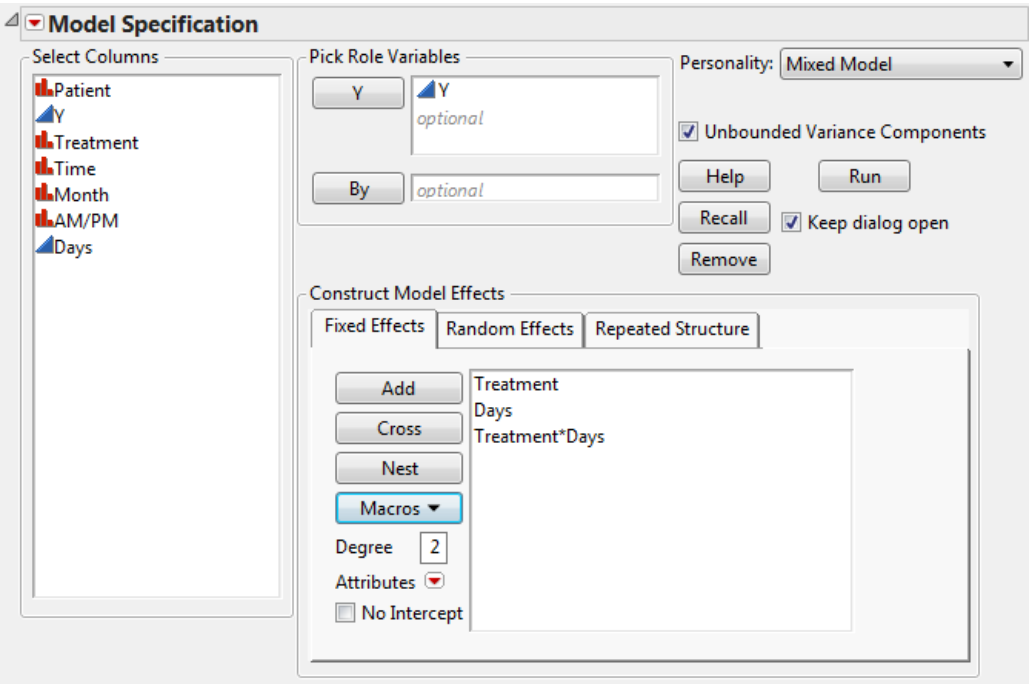
JMP PRO Regression Model for AR(1) Model Example

Using the Month and AM/PM categorical effects, you have compared four covariance structures for the cholesterol data. (Note that a categorical effect was required for the Unstructured fit.) You have decided to use an AR(1) covariance structure.

Suppose now that you want to model the effect of treatment in terms of the continuous effect Days instead of the categorical effects. You can then predict cholesterol levels at arbitrary time during treatment.

1. After following step 1 to step 4 in "Covariance Structure: AR(1)" on page 387, return to the Fit Model launch window.
2. On the Fixed Effects tab, select the existing fixed effects and click **Remove**.
3. Select Treatment and Days then select **Macros > Full Factorial**.

Figure 8.25 Fit Model Launch Window Showing Fixed Effects Tab

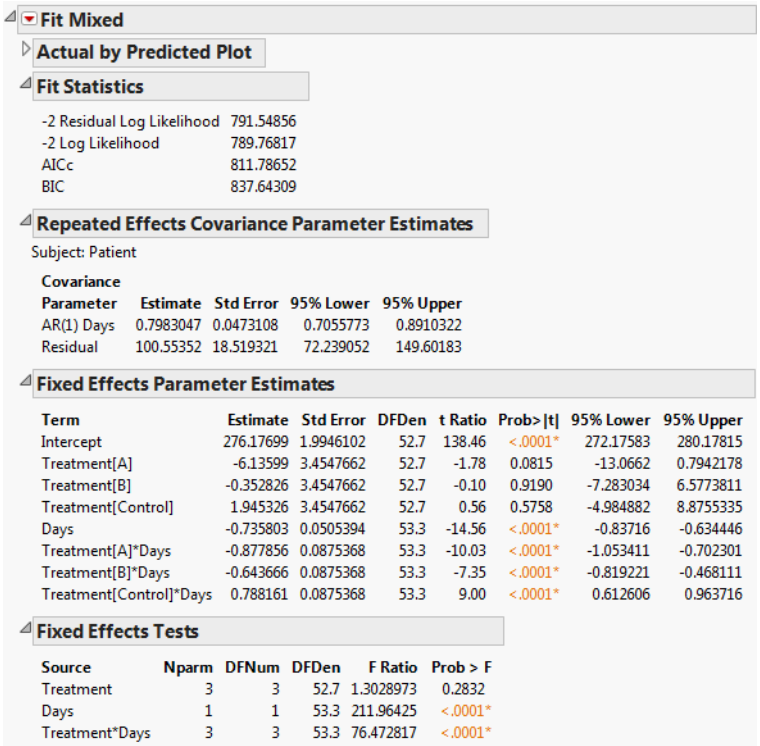


4. Click **Run**.

   The Fit Mixed report is shown in Figure 8.26. You see that the interaction of Treatment and Days is highly significant indicating different regressions for the drugs.

Note: To predict outcomes for the drugs at different days, use the profiler. See the Profiler chapter in the *Profilers* book for more information.

Figure 8.26 Fit Mixed Report - AR(1) Covariance Structure with Continuous Fixed Effect



Fit Mixed

Actual by Predicted Plot

Fit Statistics

-2 Residual Log Likelihood	791.54856
-2 Log Likelihood	789.76817
AICc	811.78652
BIC	837.64309

Repeated Effects Covariance Parameter Estimates

Subject: Patient

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
AR(1) Days	0.7983047	0.0473108	0.7055773	0.8910322
Residual	100.55352	18.519321	72.239052	149.60183

Fixed Effects Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	276.17699	1.9946102	52.7	138.46	<.0001*	272.17583	280.17815
Treatment[A]	-6.13599	3.4547662	52.7	-1.78	0.0815	-13.0662	0.7942178
Treatment[B]	-0.352826	3.4547662	52.7	-0.10	0.9190	-7.283034	6.5773811
Treatment[Control]	1.945326	3.4547662	52.7	0.56	0.5758	-4.984882	8.8755335
Days	-0.735803	0.0505394	53.3	-14.56	<.0001*	-0.83716	-0.634446
Treatment[A]*Days	-0.877856	0.0875368	53.3	-10.03	<.0001*	-1.053411	-0.702301
Treatment[B]*Days	-0.643666	0.0875368	53.3	-7.35	<.0001*	-0.819221	-0.468111
Treatment[Control]*Days	0.788161	0.0875368	53.3	9.00	<.0001*	0.612606	0.963716

Fixed Effects Tests

Source	Nparm	DFNum	DFDen	F Ratio	Prob > F
Treatment	3	3	52.7	1.3028973	0.2832
Days	1	1	53.3	211.96425	<.0001*
Treatment*Days	3	3	53.3	76.472817	<.0001*

JMP PRO Split Plot Example

The Mixed Model personality offers a straightforward approach to specifying and analyzing split-plot experiments. Mixed Model provides tabs for specifying effects. The resulting analysis is targeted to random effects. Note, however, that split-plot experiments can also be analyzed using the Standard Least Squares personality.

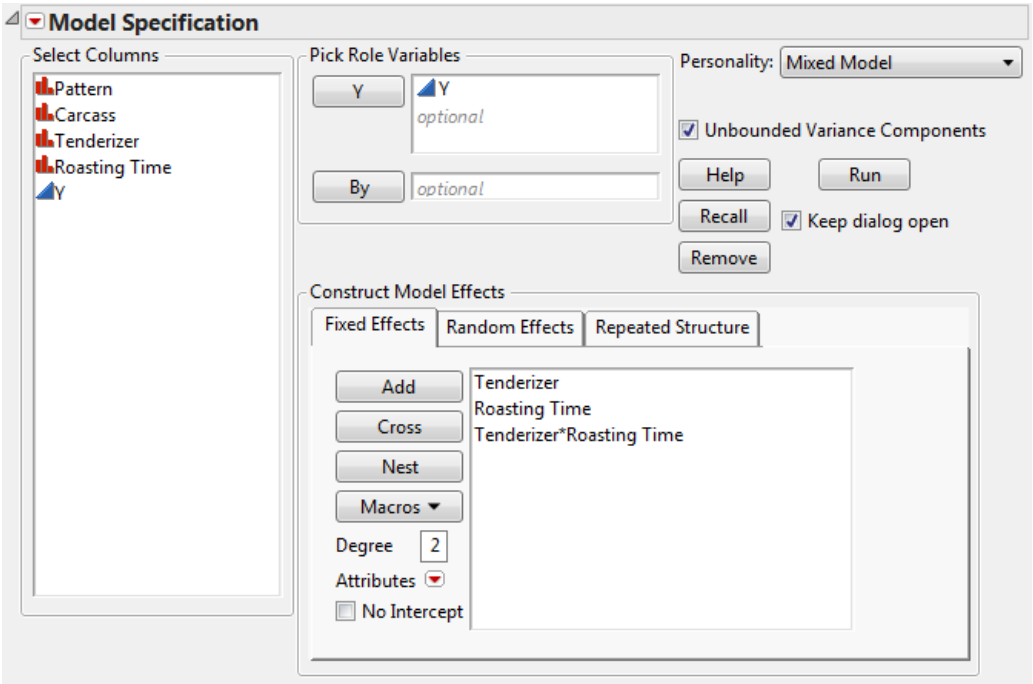
The data in the Split Plot.jmp sample data table come from a study of the effects of tenderizer and length of cooking time on meat. Six beef carcasses were randomly selected from carcasses at a meat packaging plant. From the right rib-eye muscle of each carcass, three rolled roasts were prepared under uniform conditions. Each of these three roasts was assigned a tenderizing treatment at random. After treatment, a coring device was used to mark four cores of meat near the center of each.

The three roasts from the same carcass were placed together in a preheated oven and allowed to cook. After 30 minutes, one of the cores was taken at random from each roast. Cores were removed in this fashion again after 36 minutes, 42 minutes, and 48 minutes. As each set cooled to serving temperature, the cores were measured for tenderness using the Warner-Bratzler device. Larger measurements indicate tougher meat.

Your interest centers on the effects of tenderizer, roasting time, and especially whether there is an interaction between tenderizer and roasting time. This design addresses that goal.

1. Select **Help > Sample Data Library** and open Split Plot.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Select **Mixed Model** from the Personality list.
5. Select Tenderizer and Roasting Time, and then select **Macros > Full Factorial**.

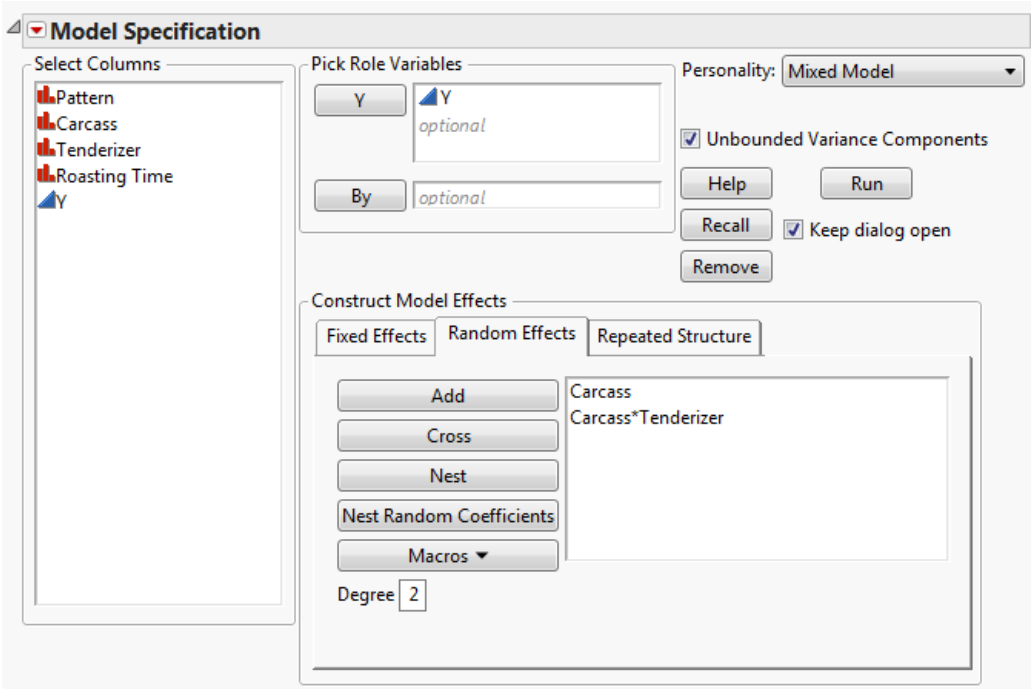
Figure 8.27 Fit Model Launch Window Showing Completed Fixed Effects Tab



6. Select the **Random Effects** tab.
7. Select Carcass and click **Add** to create the random carcass effect.
8. Select Carcass and Tenderizer and click **Cross**.

The Carcass*Tenderizer interaction is the error term for the whole plot factor, Tenderizer. This is equivalent to the Carcass*Tenderizer&Random term in Standard Least Squares.

Figure 8.28 Fit Model Launch Window Showing Completed Random Effects Tab

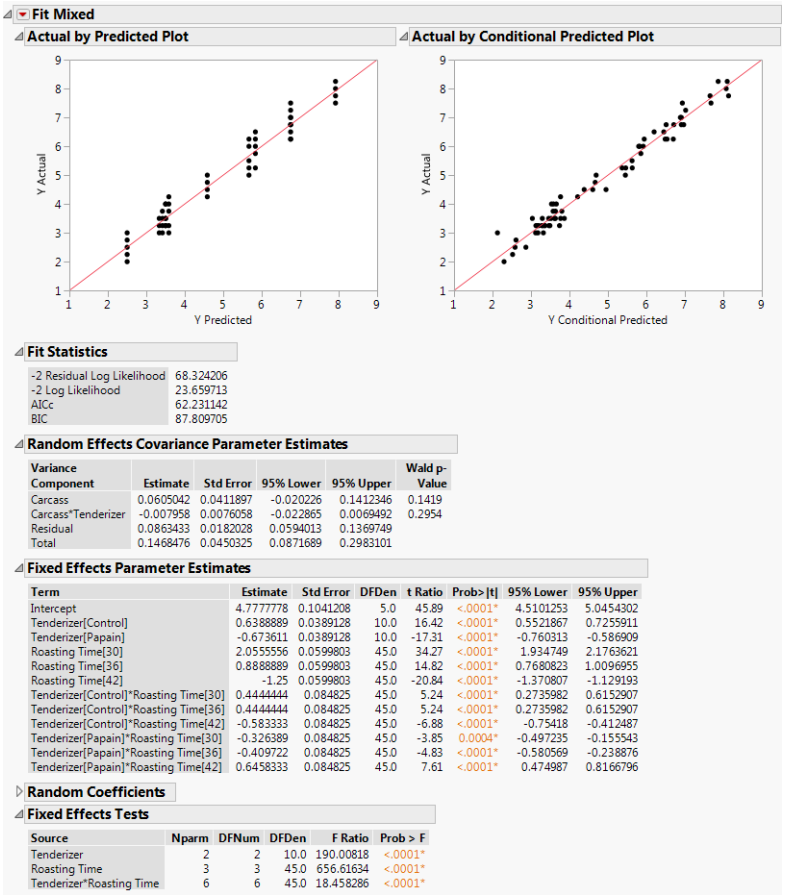


9. Click **Run**.

   The Fit Mixed report is shown in Figure 8.29.

   The Actual by Predicted Plot and the Actual by Conditional Predicted Plot show no issues with model fit, so you can proceed to interpret the results. The Fixed Effects Tests report indicates that there is a significant interaction between tenderizer and roasting time.

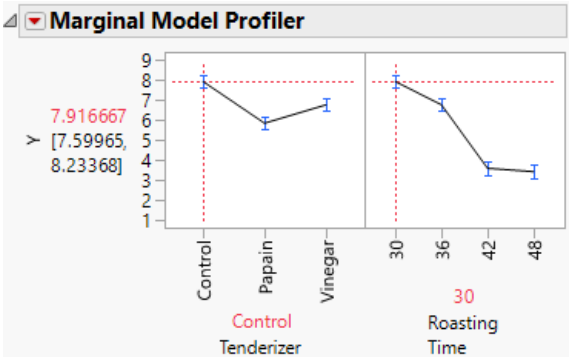
Figure 8.29 Fit Mixed Report



Explore the Interaction between Tenderizer and Roasting Time

1. Select **Marginal Model Inference > Profiler** from the Fit Mixed report's red triangle menu.

Figure 8.30 Marginal Model Profiler with Roasting Time Set to 30 Minutes



2. Move the red dashed vertical line in the Roasting Time panel to 36, 42, and 48.

   In Figure 8.30, notice that both the papain and vinegar tenderizers result in significantly lower tenderness scores than the control when roasting time is either 30 or 36 minutes. However, at 42 minutes, there are no significant differences. At 48 minutes, papain gives a value lower than the control, but vinegar does not. Papain gives lower tenderness scores than does vinegar at all times except 42 minutes.

3. Select **Multiple Comparisons** from the Fit Mixed red triangle menu.
4. Select **Tenderizer*Roasting Time**.
5. Select **All Pairwise Comparisons - Tukey HSD** and then click **OK**.

   Figure 8.31 shows a partial list of pairwise comparisons. Most of the differences between papain and vinegar that you observed in the profiler are statistically significant. Therefore, it appears that papain is the better tenderizer.

Figure 8.31 Multiple Comparisons, Partial View

Multiple Comparisons for Tenderizer*Roasting Time

Least Squares Means Estimates

Tenderizer	Roasting Time	Estimate	Std Error	DF	Lower 95%	Upper 95%
Control	30	7.9166667	0.15214568	20.355	7.5996506	8.2336827
Control	36	6.7500000	0.15214568	20.355	6.4329840	7.0670160
Control	42	3.5833333	0.15214568	20.355	3.2663173	3.9003494
Control	48	3.4166667	0.15214568	20.355	3.0996506	3.7336827
Papain	30	5.8333333	0.15214568	20.355	5.5163173	6.1503494
Papain	36	4.5833333	0.15214568	20.355	4.2663173	4.9003494
Papain	42	3.5000000	0.15214568	20.355	3.1829840	3.8170160
Papain	48	2.5000000	0.15214568	20.355	2.1829840	2.8170160
Vinegar	30	6.7500000	0.15214568	20.355	6.4329840	7.0670160
Vinegar	36	5.6666667	0.15214568	20.355	5.3496506	5.9836827
Vinegar	42	3.5000000	0.15214568	20.355	3.1829840	3.8170160
Vinegar	48	3.3333333	0.15214568	20.355	3.0163173	3.6503494

Tukey HSD All Pairwise Comparisons

Quantile = 3.44501, Adjusted DF = 45.0, Adjustment = Tukey-Kramer

All Pairwise Differences

Tenderizer	Roasting Time	-Tenderizer	-Roasting Time	Difference	Std Error	t Ratio	Prob>|t|	Lower 95%	Upper 95%
Control	30	Control	36	1.16667	0.1696485	6.88	<.0001*	0.58223	1.75111
Control	30	Control	42	4.33333	0.1696485	25.54	<.0001*	3.74889	4.91777
Control	30	Control	48	4.50000	0.1696485	26.53	<.0001*	3.91556	5.08444
Control	30	Papain	30	2.08333	0.1616429	12.89	<.0001*	1.52647	2.64019
Control	30	Papain	36	3.33333	0.1616429	20.62	<.0001*	2.77647	3.89019
Control	30	Papain	42	4.41667	0.1616429	27.32	<.0001*	3.85981	4.97353
Control	30	Papain	48	5.41667	0.1616429	33.51	<.0001*	4.85981	5.97353
Control	30	Vinegar	30	1.16667	0.1616429	7.22	<.0001*	0.60981	1.72353
Control	30	Vinegar	36	2.25000	0.1616429	13.92	<.0001*	1.69314	2.80686
Control	30	Vinegar	42	4.41667	0.1616429	27.32	<.0001*	3.85981	4.97353
Control	30	Vinegar	48	4.58333	0.1616429	28.35	<.0001*	4.02647	5.14019
Control	36	Control	42	3.16667	0.1696485	18.67	<.0001*	2.58223	3.75111
Control	36	Control	48	3.33333	0.1696485	19.65	<.0001*	2.74889	3.91777
Control	36	Papain	30	0.91667	0.1616429	5.67	<.0001*	0.35981	1.47353
Control	36	Papain	36	2.16667	0.1616429	13.40	<.0001*	1.60981	2.72353
Control	36	Papain	42	3.25000	0.1616429	20.11	<.0001*	2.69314	3.80686
Control	36	Papain	48	4.25000	0.1616429	26.29	<.0001*	3.69314	4.80686
Control	36	Vinegar	30	0.00000	0.1616429	0.00	1.0000	-0.55686	0.55686
Control	36	Vinegar	36	1.08333	0.1616429	6.70	<.0001*	0.52647	1.64019
Control	36	Vinegar	42	3.25000	0.1616429	20.11	<.0001*	2.69314	3.80686
Control	36	Vinegar	48	3.41667	0.1616429	21.14	<.0001*	2.85981	3.97353
Control	42	Control	48	0.16667	0.1696485	0.98	0.9974	-0.41777	0.75111
Control	42	Papain	30	-2.25000	0.1616429	-13.92	<.0001*	-2.80686	-1.69314
Control	42	Papain	36	-1.00000	0.1616429	-6.19	<.0001*	-1.55686	-0.44314
Control	42	Papain	42	0.08333	0.1616429	0.52	1.0000	-0.47353	0.64019
Control	42	Papain	48	1.08333	0.1616429	6.70	<.0001*	0.52647	1.64019
Control	42	Vinegar	30	-3.16667	0.1616429	-19.59	<.0001*	-3.72353	-2.60981
Control	42	Vinegar	36	-2.08333	0.1616429	-12.89	<.0001*	-2.64019	-1.52647
Control	42	Vinegar	42	0.08333	0.1616429	0.52	1.0000	-0.47353	0.64019
Control	42	Vinegar	48	0.25000	0.1616429	1.55	0.9187	-0.30686	0.80686
Control	48	Papain	30	-2.41667	0.1616429	-14.95	<.0001*	-2.97353	-1.85981
Control	48	Papain	36	-1.16667	0.1616429	-7.22	<.0001*	-1.72353	-0.60981
Control	48	Papain	42	-0.08333	0.1616429	-0.52	1.0000	-0.64019	0.47353
Control	48	Papain	48	0.91667	0.1616429	5.67	<.0001*	0.35981	1.47353
Control	48	Vinegar	30	-3.33333	0.1616429	-20.62	<.0001*	-3.89019	-2.77647
Control	48	Vinegar	36	-2.25000	0.1616429	-13.92	<.0001*	-2.80686	-1.69314
Control	48	Vinegar	42	-0.08333	0.1616429	-0.52	1.0000	-0.64019	0.47353
Control	48	Vinegar	48	0.08333	0.1616429	0.52	1.0000	-0.47353	0.64019

JMP PRO Spatial Example: Uniformity Trial

Consider the Uniformity Trial.jmp sample data table. An agronomic uniformity trial was conducted on an 8x8 grid of plots. In a uniformity trial, a test crop is grown on a field with no experimental treatments applied. The response variable, often yield, is measured. The idea is to characterize variability in the field as background for planning a designed experiment to be conducted on that field. (See Littell et al. 2006, p. 447.)

Your objective is to use the information from these data to design a yield trial with 16 treatments. Specifically, you want to decide whether to conduct future experiments on the field as follows:

- a complete block design with 4 blocks (denoted Quarter in the data)
- an incomplete block design with 16 blocks (denoted Subquarter in the data)
- a completely randomized design with spatially correlated errors

With this objective, spatial data can be treated as repeated measures with two or more dimensions as repeated effects. So, you can compare and choose an appropriate model using the values in the Fit Statistics report. You start by determining if there is significant spatial variability, then you determine whether there is a nugget effect.

Once you have established whether there is a nugget effect, you determine the best fitting spatial covariance structure. Finally, you fit the blocking models and compare these to the best spatial structure. In this example, both AICc and BIC are used to select a best model. "Spatial Correlation Structure" on page 426 provides more information about nugget effects and other spatial terminology.

Tip: This section walks you through many aspects of fitting spatial data (from fitting the model to deciding on the best covariance structure). Leave the Fit Model launch window open as you work through each example.

JMP PRO Fit a Spatial Structure Model

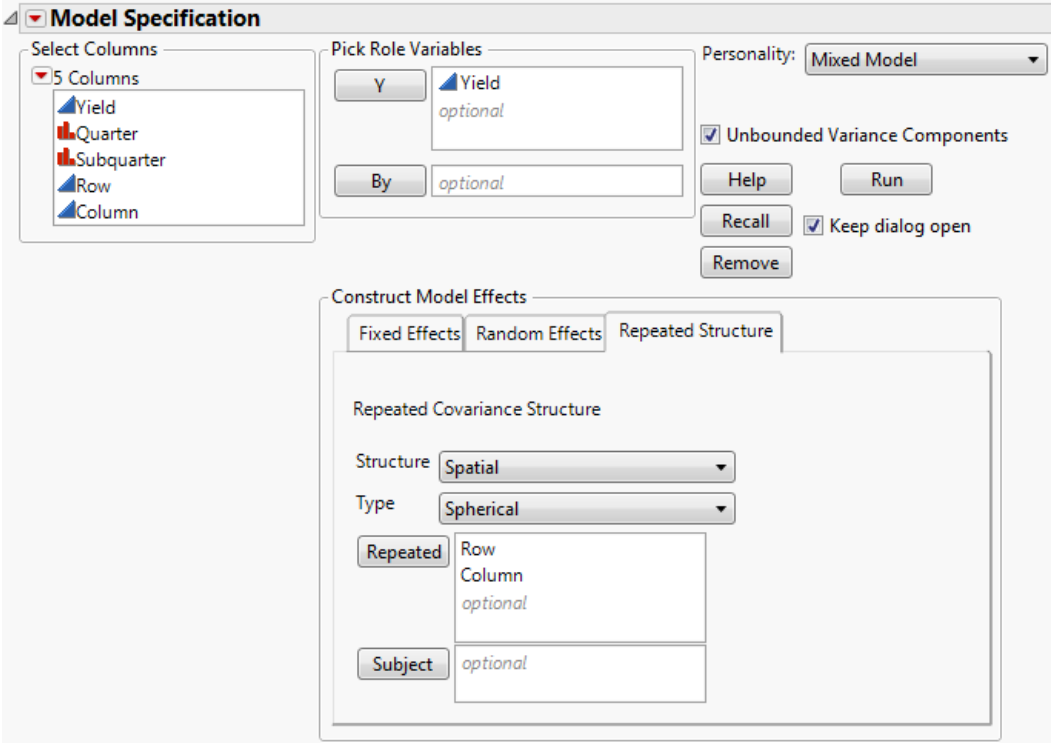
To determine whether there is significant spatial variability, you can fit a model that accounts for spatial variability. Then you can compare the likelihood for this spatial model to the likelihood for a model that does not account for spatial variability. You can do this because the independent errors model is nested within the spatial model family: The independent errors model is a spatial model with spatial correlation, ρ, equal to 0. This means that you can perform a formal likelihood ratio test of the two models.

First, fit the model that accounts for spatial structure.

1. Select **Help > Sample Data Library** and open Uniformity Trial.jmp.
2. Select **Analyze > Fit Model**.

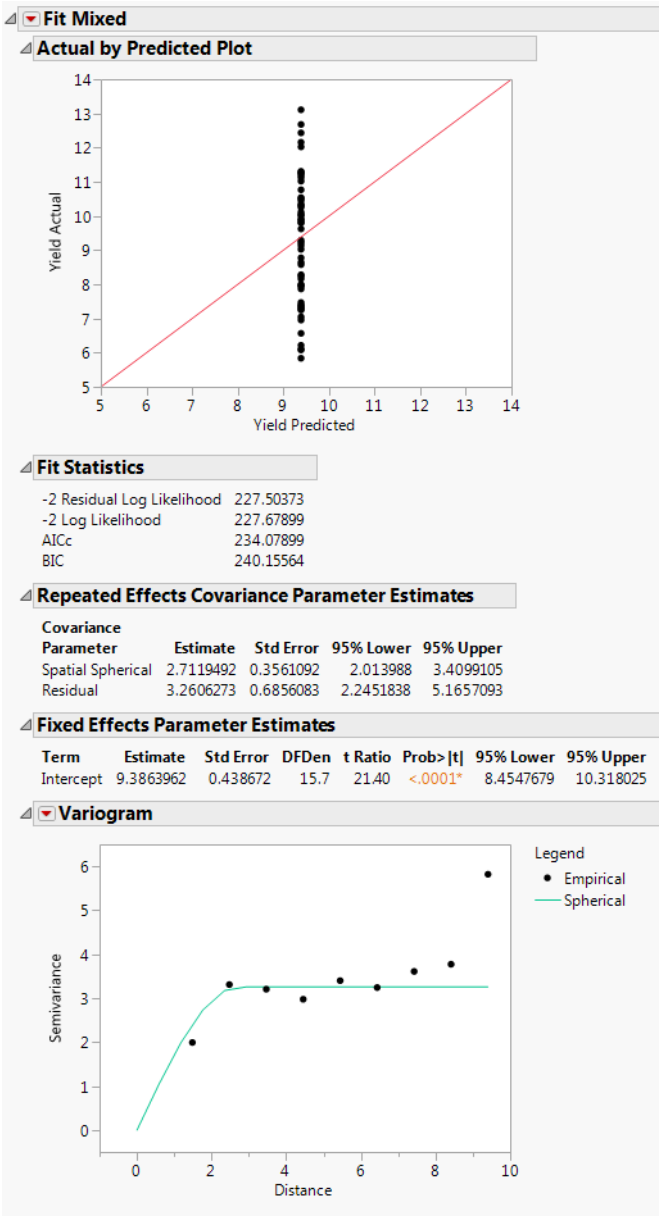
3. Select **Keep dialog open** so that you can return to the launch window in the next example.
4. Select Yield and click **Y**.
5. Select **Mixed Model** from the Personality list.
6. Select the **Repeated Structure** tab.
7. Choose **Spatial** from the list next to Structure.
8. Choose **Spherical** from the list next to Type.
9. Select Row and Column and click **Repeated**.

Figure 8.32 Completed Fit Model Launch Window Showing Repeated Structure Tab



10. Click **Run**.

Figure 8.33 Fit Mixed Report - Spatial Spherical



The Fit Mixed report is shown in Figure 8.33. The Actual by Predicted Plot shows that the predicted yield is a single value. This is because only spatial covariance was fit. The Fit Statistics report shows that -2 Log Likelihood is 227.68, and the AICc is 234.08.

Because an isotropic spatial structure was fit, a Variogram plot is shown. Because the trials are laid out in an 8 by 8 grid, there are more pairs of points at small distances than at very large

distances. See Figure 8.37 for the layout. The Variogram shows that a spherical spatial structure is an excellent fit for distances up to about 8.4. The distance class for the final distance consists of only the two diagonal pairs of points.

The Repeated Effects Covariance Parameter Estimates report gives estimates of the range (Spatial Spherical = 2.71) and the sill (Residual = 3.26). See "Variogram" on page 426.

JMP PRO Fit the Independent Errors Model

Next, fit the independent errors model.

1. Return to the Fit Model Launch Window.
2. Select **Repeated Structure** tab.
3. Select **Residual** from the Structure list.
4. Remove Row and Column from the Repeated effects list.

   Otherwise, a warning appears: "Repeated columns and subject columns are ignored when the Residual covariance structure is selected."
5. Click **Run**.

   The fit statistics for the independent errors model are: -2 Log Likelihood = 254.22, and AICc = 258.41. Each of these exceeds the corresponding value for the spatial correlation model, where -2 Log Likelihood is 227.68 and the AICc is 234.08. Because smaller values of these statistics indicate a better fit, the spatial model might provide a better fit.

JMP PRO Conduct a Likelihood Ratio Test (Optional)

A formal likelihood ratio test shows whether the spatial correlation model explains significant variation. One model must be nested in another model to create valid likelihood ratio tests.

Typically, spatial models are compared using AICc or BIC rather than through formal likelihood ratio testing. Evaluating the AICc or BIC is faster, and many spatial models are not nested.

You can conduct a likelihood ratio test in this example, because the independent errors model is nested within the spatial model family. The independent errors model is a spatial model with spatial correlation, ρ, equal to 0. This means that you can perform a formal likelihood ratio test of the two models.

In this example, the likelihood ratio test statistic is 254.22-227.68=26.54. Comparing this to a Chi-square distribution on one degree of freedom, the null hypothesis of no spatial correlation is rejected with a p-value < 0.0001. You can conclude that these data contain significant spatial variability.

JMP PRO Select the Type of Spatial Covariance

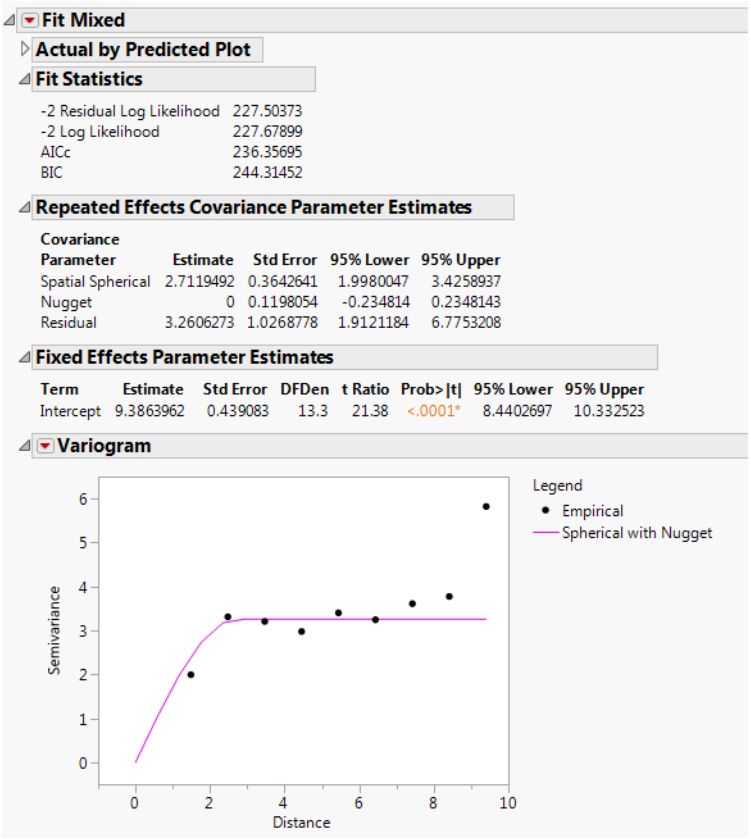
Next, you determine which spatial covariance structure best fits the data:

- with or without a nugget effect (variation over relatively small distances)
- isotropic (spatial correlation is equal in all directions) or anisotropic (spatial correlation differs in the two directions)
- type of structure, spherical, Gaussian, exponential, or power.

1. Return to the Fit Model launch window.
2. Select the **Repeated Structure** tab.
3. Select Row and Column and click **Repeated**.
4. Select **Spatial with Nugget** from the Structure list.
5. Select **Spherical** from the Type list.
6. Click **Run**.

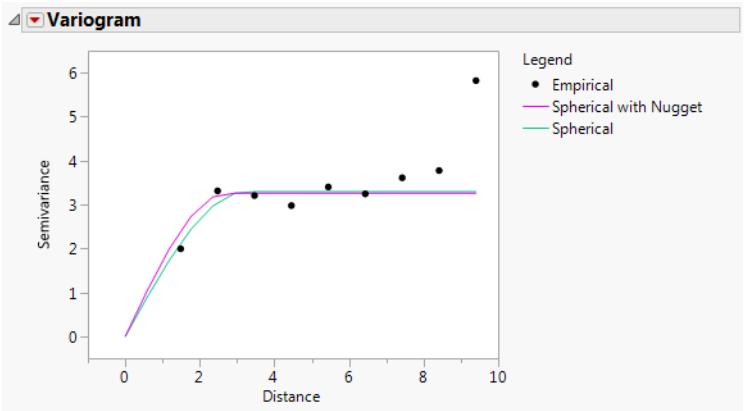
   The Fit Mixed report is shown in Figure 8.34. Notice that the log-likelihoods are essentially equal to those of the spherical with no nugget model, and the AICc is slightly higher (236.36 compared to 234.08). The Repeated Effects Covariance Parameter Estimates report shows that the Nugget covariance parameter has an estimate of zero. There does not appear to be any evidence for a nugget effect.

Figure 8.34 Fit Mixed Report - Spatial Spherical with Nugget



7. From the red triangle menu next to Variogram, select **Spatial > Spherical**.

Figure 8.35 Fit Mixed Report - Variogram



The two variograms are virtually identical. This also suggests that there is no evidence of a nugget effect.

8. Return to the Fit Model launch window.
9. Select **Repeated Structure** tab.
10. To test anisotropicity, select **Spatial Anisotropic** from the Structure list.
11. Select **Spherical** from the Type list.
12. Click **Run**.

    The Fit Mixed report is shown in Figure 8.36. The fit statistics indicate not as good a fit as the isotropic (spatial structure) spherical model (AICc 240.54 compared to 234.08). The Repeated Effects Covariance Parameter Estimates report shows that the estimates for the Row (Spatial Spherical Row) and Column (Spatial Spherical Column) covariances are very close. There is no evidence to suggest that spatial correlations within rows and columns of the grid differ.

Figure 8.36 Fit Mixed Report - Spatial Anisotropic Spherical

Fit Mixed

Actual by Predicted Plot

Fit Statistics

-2 Residual Log Likelihood	232.09009
-2 Log Likelihood	231.86637
AICc	240.54434
BIC	248.50191

Repeated Effects Covariance Parameter Estimates

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
Spatial Spherical Row	2.2453436	0.4114847	1.4388484	3.0518389
Spatial Spherical Column	2.2285839	0.3316819	1.5784993	2.8786684
Residual	3.0050357	0.6112678	2.0921736	4.6820351

Fixed Effects Parameter Estimates

Term	Estimate	Std Error	DFDen	t Ratio	Prob>|t|	95% Lower	95% Upper
Intercept	9.4039777	0.3585799	21.3	26.23	<.0001*	8.6589792	10.148976

JMP PRO Determine the Type of the Spatial Structure

An isotropic spatial structure with no nugget is appropriate. To determine the type of the spatial structure, you can compare the available types.

1. Return to the Fit Model launch window.
2. Select the **Repeated Structure** tab.
3. Select Row and Column and click **Repeated**.
4. Select **Spatial** from the Structure list.

5. Select **Power** from the Type list.
6. Click **Run**.

   Note the AICc and BIC values for the model fit.
7. Return to the Fit Model launch window.
8. Select **Exponential** from the Type list.
9. Click **Run**.

   Note the AICc and BIC values for the model fit.
10. Return to the Fit Model launch window.
11. Select **Gaussian** from the Type list.
12. Click **Run**.

    Note the AICc and BIC values for the model fit.

The observed AICc values for these types and the other fits that you performed are summarized in Table 8.2.

Table 8.2 Fit Statistics for Spatial Models Fit

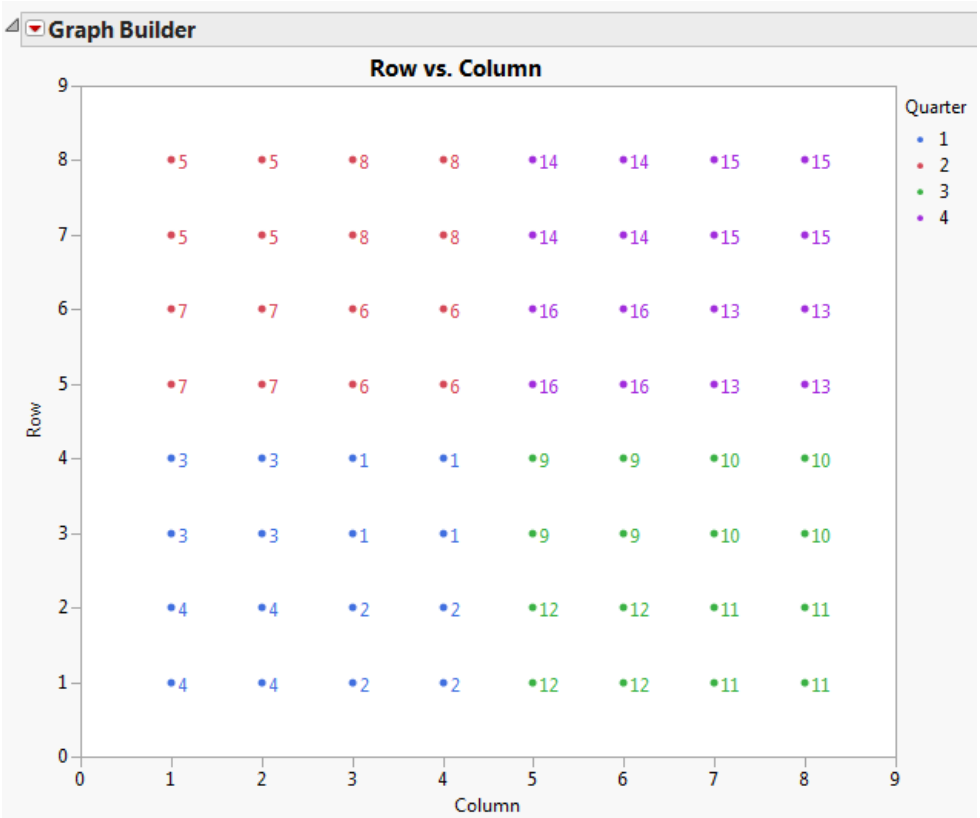
Structure	Type	AICc	BIC
Spatial	Spherical	234.08	240.16
Residual		258.41	262.53
Spatial with Nugget	Spherical	236.36	244.31
Spatial Anisotropic	Spherical	240.54	248.50
Spatial	Power	240.24	246.32
Spatial	Exponential	240.24	246.32
Spatial	Gaussian	238.37	244.44

The best fitting model (using the smallest AICc value) is the Spatial structure model with Spherical covariance structure. Now you will compare this model to the complete and incomplete block models to complete the objectives of the uniformity trial.

JMP PRO Compare the Model to Block Designs

1. Return to the Uniformity Trial data table.
2. In the Tables panel, run the Graph Builder script.

Figure 8.37 Graph Builder Plot of Proposed Complete and Incomplete Block Designs



This plot shows the proposed complete and incomplete block designs for the field. The color indicates the quarter fields that would serve as complete blocks. The numbered points represent the sub-quarter fields that would serve as incomplete blocks.

To fit the complete block model, follow these steps:

1. Return to the Fit Model launch window.
2. Select **Repeated Structure** tab.
3. Select **Residual** from the Structure list.
4. Remove Row and Column from the effect.

   Otherwise, a pop-up window appears, stating, “Repeated columns and subject columns are ignored when the Residual covariance structure is selected.”
5. Select the **Random Effects** tab.
6. Select Quarter and click **Add**.
7. Click **Run**.

To fit the incomplete block model, follow these steps:

1. Return to the Fit Model launch window.
2. Select the **Random Effects** tab.
3. Select Quarter and click **Remove**.
4. Select Subquarter and click **Add**.
5. Click **Run**.

The following list shows both AICc and BIC values for the competing models. The spherical covariance structure results in the best model fit. This indicates that, for future studies using this field, a completely randomized design with spatially correlated errors is preferred.

- Spherical model (see "Determine the Type of the Spatial Structure" on page 407)
  - AICc: 234.08
  - BICc: 240.16
- Complete block (RCBD) model
  - AICc: 259.90
  - BICc: 265.97
- Incomplete block model
  - AICc: 248.77
  - BICc: 254.85

JMP PRO Correlated Response Example

In this example, the effect of two layouts dealing with wafer production is studied for a characteristic of interest. Each of 50 wafers is partitioned into four quadrants and the characteristic is measured on each of these quadrants. Data of this type are usually presented in a format where each row contains all of the repeated measurements for one of the units of interest. Data of this type are often analyzed using separate models for each response. However, when repeated measurements are taken on a single unit, it is likely that there is within-unit correlation. Failure to account for this correlation can result in poor decisions and predictions. You can use the Mixed Model personality to account for and model the possible correlation.

For Mixed Model analysis of repeated measures data, each repeated measurement needs to be in its own row. If your data are in the typical format where all repeated measurements are in the same row, you can construct an appropriate data table for Mixed Model analysis by using Tables > Stack. See the Reshape Data chapter in the *Using JMP* for details.

In this example, you first fit univariate models using the usual data table format. Then you use the Mixed Model personality to fit models for the four responses while simultaneously accounting for possible correlation among the responses.

JMP PRO Fit Univariate Models

Use Standard Least Squares to fit a univariate model for each of the four quadrants.

1. Select **Help > Sample Data Library** and open Wafer Quadrants.jmp.

   This data table is structured for Mixed Model analysis, with one row for each Y measurement on each Quadrant. To conduct univariate analyses, you split the table using a saved script.

2. In the Wafer Quadrants.jmp data table, click the green triangle next to the **Split Y by Quadrant** script.

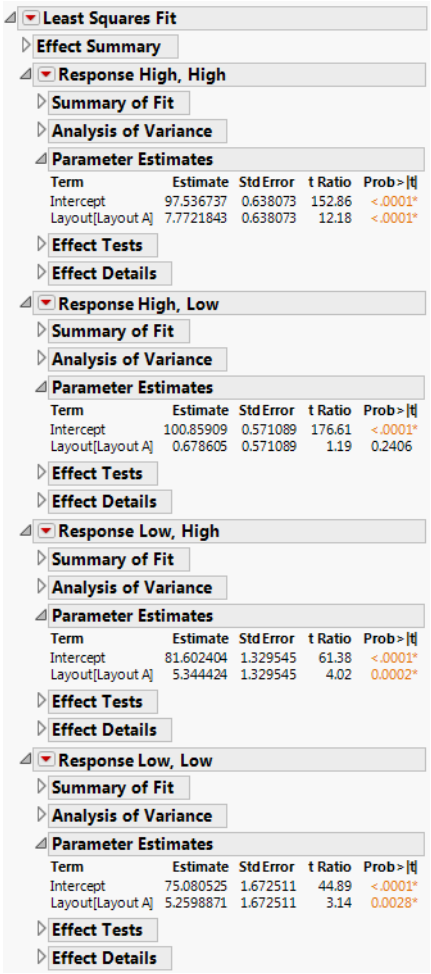
   The new data table is in the format often used to record repeated measures data. Each value of Wafer ID defines a row and the four measurements for that wafer are given in the single row.

3. Select **Analyze > Fit Model**.

   Since there is a Model script in the data table, the Model Specification window is already filled in. Note that the columns High, High through Low, Low are entered as Y and that Layout is the single model effect.

4. Click **Run**.

Figure 8.38 Four Univariate Models



Least Squares Fit

Effect Summary

Response High, High

Summary of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	97.536737	0.638073	152.86	<.0001*
Layout[Layout A]	7.7721843	0.638073	12.18	<.0001*

Effect Tests

Effect Details

Response High, Low

Summary of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	100.85909	0.571089	176.61	<.0001*
Layout[Layout A]	0.678605	0.571089	1.19	0.2406

Effect Tests

Effect Details

Response Low, High

Summary of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	81.602404	1.329545	61.38	<.0001*
Layout[Layout A]	5.344424	1.329545	4.02	0.0002*

Effect Tests

Effect Details

Response Low, Low

Summary of Fit

Analysis of Variance

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	75.080525	1.672511	44.89	<.0001*
Layout[Layout A]	5.2598871	1.672511	3.14	0.0028*

Effect Tests

Effect Details

The report indicates that Layout has a statistically significant effect for all quadrants except the High, Low quadrant.

JMP PRO Perform Mixed Model Analysis

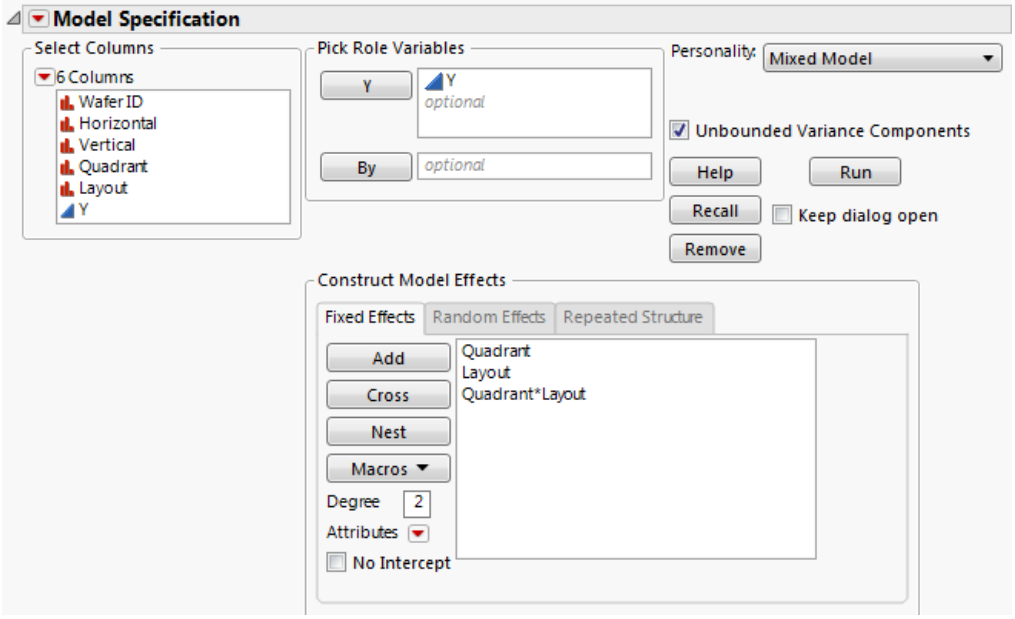
Using the Mixed Model analysis, you can obtain more information.

1. Return to Wafer Quadrants.jmp. Or, if you have closed it, select **Help > Sample Data Library** and open Wafer Quadrants.jmp.
2. Select **Analyze > Fit Model**.
3. Select Y and click **Y**.
4. Select **Mixed Model** from the **Personality** list.

5. Select Quadrant and Layout from the **Select Columns** list and click **Macros > Full Factorial**.

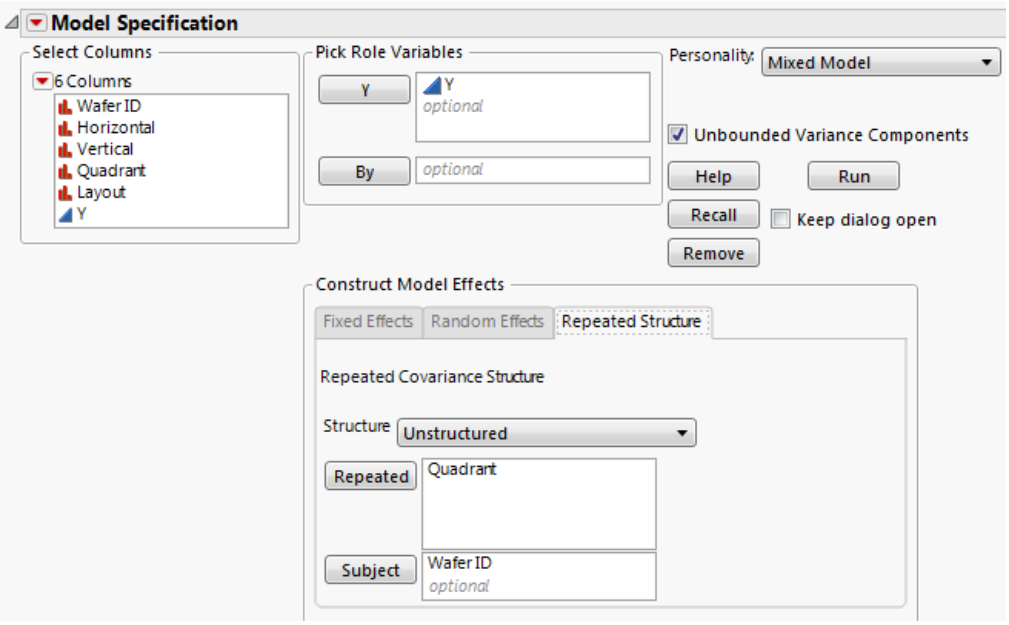
   This model specification enables you to explore the effect of Layout on the repeated measurements as well as the possible interaction of Layout with Quadrant.

Figure 8.39 Fit Model Launch Window Showing Completed Fixed Effects Tab



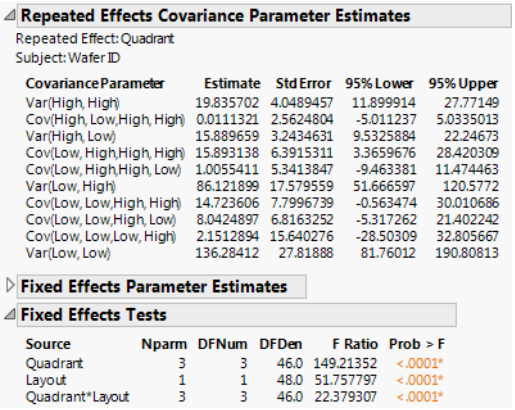
6. Select the **Repeated Structure** tab.
7. Select **Unstructured** from the Structure list.
8. Select Quadrant and click **Repeated**.
9. Select Wafer ID and click **Subject**.

Figure 8.40 Fit Model Launch Window Showing Repeated Structure Tab



10. Click **Run**.

Figure 8.41 Fit Mixed Report with Fixed Effects Parameter Estimates Report Closed



Repeated Effects Covariance Parameter Estimates

Repeated Effect: Quadrant
Subject: Wafer ID

Covariance Parameter	Estimate	Std Error	95% Lower	95% Upper
Var(High, High)	19.835702	4.0489457	11.899914	27.77149
Cov(High, Low,High, High)	0.0111321	2.5624804	-5.011237	5.0335013
Var(High, Low)	15.889659	3.2434631	9.5325884	22.24673
Cov(Low, High,High, High)	15.893138	6.3915311	3.3659676	28.420309
Cov(Low, High,High, Low)	1.0055411	5.3413847	-9.463381	11.474463
Var(Low, High)	86.121899	17.579559	51.666597	120.5772
Cov(Low, Low,High, High)	14.723606	7.7996739	-0.563474	30.010686
Cov(Low, Low,High, Low)	8.0424897	6.8163252	-5.317262	21.402242
Cov(Low, Low,Low, High)	2.1512894	15.640276	-28.50309	32.805667
Var(Low, Low)	136.28412	27.81888	81.76012	190.80813

Fixed Effects Parameter Estimates

Fixed Effects Tests

Source	Nparm	DFNum	DFDen	F Ratio	Prob > F
Quadrant	3	3	46.0	149.21352	<.0001*
Layout	1	1	48.0	51.757797	<.0001*
Quadrant*Layout	3	3	46.0	22.379307	<.0001*

The Repeated Effects Covariance Parameter Estimates report gives the estimated variances and covariances for the four responses. Note that the confidence interval for the covariance of Low, High with High, High does not include zero. This suggests that there is a positive covariance between measurements in these two quadrants. This is information that the Mixed Model analysis uses and that is not available when the responses are modeled independently.

The Fixed Effects Tests report indicates that there is a significant Layout by Characteristic interaction.

JMP PRO Explore the Layout by Characteristic Interaction with the Profiler

1. Click the Fit Mixed red triangle and select **Marginal Model Inference > Profiler**.
2. In the Profiler plot, compare the predicted values for Y across the quadrants by first setting the vertical red dotted line at Layout A and then at Layout B.

Figure 8.42 Profile for Quadrant for Layout A

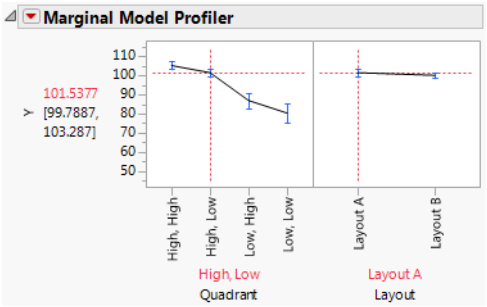
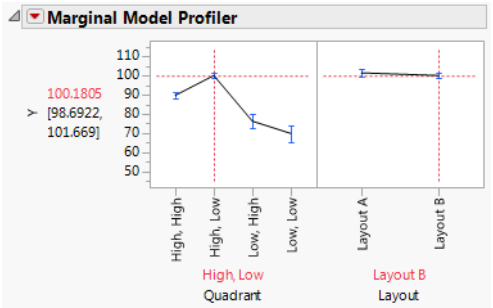


Figure 8.43 Profile for Quadrant for Layout B



The differences in the profiles at each setting of Layout give you insight into the significant interaction. It appears that the interaction is partially driven by the differences for the High, High quadrant.

JMP PRO Plot of Y by Layout and by Quadrant

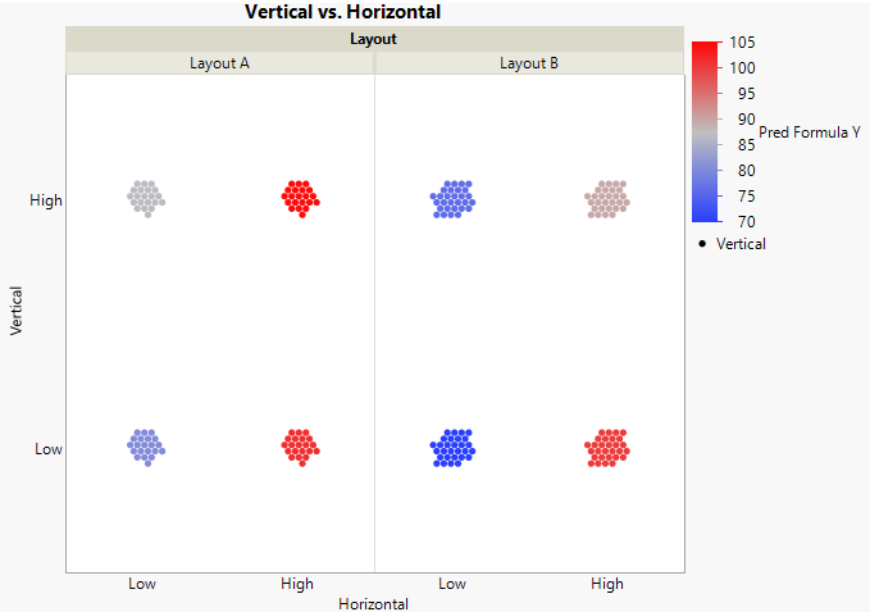
Use Graph Builder to explore the nature of the interaction.

1. Click the Fit Mixed red triangle and select **Save Columns > Prediction Formula**.

   The prediction formula is saved to the data table in the column Pred Formula Y.

2. Select **Graph > Graph Builder**.
3. Drag Horizontal to the **X** zone.
4. Drag Vertical to the **Y** zone.
5. Drag Pred Formula Y to the **Color** zone.
6. Drag Layout to the **Wrap** zone.
7. Click **Done** to hide the control panel.

Figure 8.44 Completed Graph Builder Plot



Note: Because the points in Graph Builder are randomly jittered, your plot might not match Figure 8.44 exactly.

The plot shows the predicted differences for the eight Layout and Quadrant combinations using a color intensity scale. The predicted values for the High, Low quadrant are in the lower right. The color gradient shows relatively little difference for these predicted values. Other differences are clearly indicated.

JMP PRO Statistical Details for the Mixed Models Personality

JMP PRO Convergence Score Test

The convergence failure warning shows the score test for the following hypothesis: that the unknown maximum likelihood estimate (MLE) is consistent with the parameter given in the final iteration of the model-fitting algorithm. This hypothesis test is possible because the relative gradient criterion is algebraically equivalent to the score test statistic. Remarkably, the score test does not require knowledge of the true MLE.

JMP PRO Score Test

Consider first the case of a single parameter, θ. Let l be the log-likelihood function for θ and let $\mathbf{x}$ be the data. The score is the derivative of the log-likelihood function with respect to θ:

$$U(\theta) = \frac{\partial l(\theta|\mathbf{x})}{\partial \theta}$$

The observed information is:

$$I(\theta) = -\frac{\partial^2}{\partial \theta^2} l(\theta|\mathbf{x})$$

The statistic for the score test of H_0: $\theta = \theta_0$ is:

$$S(\theta_0) = \frac{U(\theta_0)^2}{I(\theta_0)}$$

This statistic has an asymptotic Chi-square distribution with 1 degree of freedom under the null hypothesis.

The score test can be generalized to multiple parameters. Consider the vector of parameters θ. Then the test statistic for the score test of H_0: $\theta = \theta_0$ is:

$$S(\theta_0) = \mathbf{U}'(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{U}(\theta_0)$$

where

$$\mathbf{U}(\theta) = \frac{\partial l(\theta|\mathbf{x})}{\partial \theta}$$

and

$$\mathbf{I}(\theta) = -\frac{\partial^2 l(\theta|\mathbf{x})}{\partial \theta(\partial \theta')}$$

and $\mathbf{U}'$ denotes the transpose of the matrix $\mathbf{U}$.

The test statistic is asymptotically Chi-square distribution with k degrees of freedom. Here k is the number of unbounded parameters.

JMP PRO Relative Gradient

The convergence criterion for the Mixed Model fitting procedure is based on the relative gradient $\mathbf{g}'\mathbf{H}^{-1}\mathbf{g}$. Here, $\mathbf{g}(\theta) = \mathbf{U}(\theta)$ is the gradient of the log-likelihood function and $\mathbf{H}(\theta) = -\mathbf{I}(\theta)$ is its Hessian.

Let θ_0 be the value of θ where the algorithm terminates. Note that the relative gradient evaluated at θ_0 is the score test statistic. A p-value is calculated using a Chi-square distribution with k degrees of freedom. This p-value gives an indication of whether the value of the unknown MLE is consistent with θ_0. The number of unbounded parameters listed in the Random Effects Covariance Parameter Estimates report equals k.

JMP PRO Random Coefficient Model

The standard random coefficient model specifies a random intercept and slope for each subject. Let y_{ij} denote the measurement of the j^{th} observation on the i^{th} subject. Then the random coefficient model can be written as follows:

$$y_{ij} = a_i + b_i x_{ij} + e_{ij}$$

where

$$i = 1, 2, \ldots, t$$

$$j = 1, 2, \ldots, n_i$$

$$\begin{bmatrix} a_i \\ b_i \end{bmatrix} \sim iid\ N\left(\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \mathbf{G} \right)$$

$$\mathbf{G} = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}$$

and

$$e_{ij} \sim iid\ N(0, \sigma^2)$$

You can reformulate the model to reflect the fixed and random components that are estimated by JMP as follows.

$$y_{ij} = (\alpha + a_i^*) + (\beta + b_i^*)x_{ij} + e_{ij}$$

where

$$a_i^* = \alpha_i - \alpha$$

$$b_i^* = \beta_i - \beta$$

and

$$\begin{bmatrix} a_i^* \\ b_i^* \end{bmatrix} \sim iid\ N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \mathbf{G} \right)$$

with **G** and e_{ij} defined as above.

JMP PRO Repeated Measures

The form of the repeated measures model is $y_{ijk} = \alpha_{ij} + s_{ik} + e_{ijk}$, where

α_{ij} can be written as a treatment and time factorial

s_{ik} is the random effect of the k^{th} subject assigned to the i^{th} treatment

$j = 1,\ldots,m$ denotes the repeated measurements over time.

Assume that the s_{ik} are independent and identically distributed $N(0, \sigma_s^2)$ variables. Denote the number of treatment factors by t and the number of subjects by s. Then the distribution of e_{ijk} is $N(0, \Sigma)$, where

$$\Sigma = \mathbf{I}_{ts} \otimes Var(\mathbf{y}_{ik} | s_{ik})$$

and

$$\mathbf{y}'_{ik} | s_{ik} = \left(\left[y_{i1k} \; y_{i2k} \; \cdots \; y_{imk} \right] \middle| s_{ik} \right)$$

Denote the block diagonal component of the covariance matrix Σ corresponding to the ik^{th} subject within treatment by Σ_{ik}. In other words, $\Sigma_{ik} = Var(\mathbf{y}_{ik} | s_{ik})$. Because observations over time within a subject are not typically independent, it is necessary to estimate the variance of $y_{ijk} | s_{ik}$. Failure to account for the correlation leads to distorted inference.

See "Repeated Covariance Structures" on page 420 and "Spatial and Temporal Variability" on page 425 for more information about the covariance structures available for Σ_{ik}.

JMP PRO Repeated Covariance Structures

This section gives the parameterizations for the following covariance structures:

JMP PRO Unequal Variances Covariance Structure

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \cdots & 0 \\ & \sigma_2^2 & 0 & \cdots & 0 \\ & & \sigma_3^2 & \cdots & 0 \\ & & & \ddots & \vdots \\ & & & & \sigma_m^2 \end{bmatrix}$$

Here, the variance among observations taken at time j is:

$$\sigma_j^2 = Var(y_{ijk} | s_{ik})$$

The variances are allowed to differ across the levels of the repeated column. The covariances between observations at different levels are zero.

JMP PRO Unstructured Covariance Structure

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \cdots & \sigma_{1[m-1]} & \sigma_{1m} \\ & \sigma_2^2 & \sigma_{23} & \cdots & \sigma_{2[m-1]} & \sigma_{2m} \\ & & \sigma_3^2 & \cdots & \sigma_{3[m-1]} & \sigma_{3m} \\ & & & \ddots & \vdots & \vdots \\ & & & & \sigma_{m-1}^2 & \sigma_{[m-1]m} \\ & & & & & \sigma_m^2 \end{bmatrix}$$

Here, the variance among observations taken at time j is:

$$\sigma_j^2 = Var(y_{ijk} | s_{ik})$$

The covariance between observations taken at times j and j' is:

$$\sigma_{jj'} = Cov(y_{ijk}, y_{ij'k} | s_{ik})$$

The variances are allowed to differ across the levels of the repeated column. The covariances between observations at different levels is unique.

JMP PRO Compound Symmetry Covariance Structure

In JMP, a compound symmetry covariance structure is implemented using a mixed model with independent errors approach. Random effects are classified into two categories: G-side or R-side. See Searle et al. (1992) for more information.

The G-side random effects are associated with the design matrix for random effects. The R-side random effects are associated with residual error. Within-subject variance is part of the design structure and is modeled on the G-side. Between-subject variance falls into the residual structure and is modeled R-side. In the independent structure:

- The random effects G-side variance is modeled by $s_{ik} \sim$ iid N(0, σ_s^2).
- The R-side variance is modeled by $e_{ijk} \sim$ iid N(0, σ^2).

It follows that the covariance matrix is given as follows:

$$\Sigma_{ik} = \sigma_s^2\mathbf{J} + \sigma^2\mathbf{I} = \begin{bmatrix} \sigma_s^2+\sigma^2 & \sigma_s^2 & \cdots & \sigma_s^2 \\ & \sigma_s^2+\sigma^2 & \cdots & \sigma_s^2 \\ & & \ddots & \vdots \\ & & & \sigma_s^2+\sigma^2 \end{bmatrix}$$

where **J** is a matrix consisting of *1*s and **I** is an identity matrix.

Alternatively, all variance could be modeled on the R-side. Under the Gaussian assumption, this compound-symmetry covariance structure is equivalent to the independence model (Type=CS in SAS). This structure is available in JMP by using the Compound Symmetry structure in the repeated structure tab. Here, the correlation between pairs of repeated observations is the same regardless of the time difference between the observations. Thus, the correlation matrix can be written as follows:

$$C = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ & 1 & \cdots & \rho \\ & & \ddots & \vdots \\ & & & 1 \end{bmatrix}$$

Using the Compound Symmetry structure in JMP also assumes a common variance, σ_e^2, among observations taken at any time point. The covariance structure is then $\Sigma = \sigma_e^2 C$ where

$$\sigma_e^2 = \sigma_s^2 + \sigma^2$$

and

$$\rho = \frac{\sigma_s^2}{\sigma_s^2+\sigma^2}.$$

Here, ρ is the intra-class correlation coefficient and σ_e^2 is the residual variance. Another option is to use the Compound Symmetry Unequal Variances structure in JMP, which allows the variance to vary across time points. This leads to a covariance matrix as follows:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 & \rho\sigma_1\sigma_3 & \cdots & \rho\sigma_1\sigma_{m-1} & \rho\sigma_1\sigma_m \\ & \sigma_2^2 & \rho\sigma_2\sigma_3 & \cdots & \rho\sigma_2\sigma_{m-1} & \rho\sigma_2\sigma_m \\ & & \sigma_3^2 & \cdots & \rho\sigma_3\sigma_{m-1} & \rho\sigma_3\sigma_m \\ & & & \ddots & \vdots & \vdots \\ & & & & \sigma_{m-1}^2 & \rho\sigma_{m-1}\sigma_m \\ & & & & & \sigma_m^2 \end{bmatrix}$$

JMP PRO AR(1) Covariance Structure

$$\Sigma = \begin{bmatrix} \sigma^2 & \rho^{|t_1-t_2|}\sigma^2 & \rho^{|t_1-t_3|}\sigma^2 & \cdots & \rho^{|t_1-t_{m-1}|}\sigma^2 & \rho^{|t_1-t_m|}\sigma^2 \\ & \sigma^2 & \rho^{|t_2-t_3|}\sigma^2 & \cdots & \rho^{|t_2-t_{m-1}|}\sigma^2 & \rho^{|t_2-t_m|}\sigma^2 \\ & & \sigma^2 & \cdots & \rho^{|t_3-t_{m-1}|}\sigma^2 & \rho^{|t_3-t_m|}\sigma^2 \\ & & & \ddots & \vdots & \vdots \\ & & & & \sigma^2 & \rho^{|t_{m-1}-t_m|}\sigma^2 \\ & & & & & \sigma^2 \end{bmatrix}$$

Here t_j is the time of observation j. In this structure, observations taken at any given time have the same variance, σ^2. The parameter ρ, where $-1 < \rho < 1$, is the correlation between two observations that are one unit of time apart. As the time difference between observations increases, their covariance decreases because ρ is raised to a higher power. In many applications, AR(1) provides an adequate model of the within subject correlation, providing more power without sacrificing Type I error control.

JMP PRO Toeplitz Covariance Structure

In the Toeplitz structure, observations that are separated by a fixed number of time units have the same correlation. In contrast to the AR(1) correlation structure, the Toeplitz correlations at a fixed time difference are arbitrary. Denote the correlation for observations d units apart by ρ_d. The correlation matrix is as follows:

$$C = \begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{m-1} & \rho_m \\ & 1 & \rho_1 & \cdots & \rho_{m-2} & \rho_{m-1} \\ & & 1 & \cdots & \rho_{m-3} & \rho_{m-2} \\ & & & \ddots & \vdots & \vdots \\ & & & & 1 & \rho_1 \\ & & & & & 1 \end{bmatrix}$$

Two options in JMP use this correlation matrix:

- The Toeplitz structure assumes a common variance, σ^2, for observations from any time point. The covariance structure is $\Sigma = \sigma^2 C$.
- Alternatively, the Toeplitz Unequal Variances structure allows the variance to vary across time points:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_1\sigma_1\sigma_2 & \rho_2\sigma_1\sigma_3 & \cdots & \rho_{m-1}\sigma_1\sigma_{m-1} & \rho_m\sigma_1\sigma_m \\ & \sigma_2^2 & \rho_1\sigma_2\sigma_3 & \cdots & \rho_{m-2}\sigma_2\sigma_{m-1} & \rho_{m-1}\sigma_2\sigma_m \\ & & \sigma_3^2 & \cdots & \rho_{m-3}\sigma_3\sigma_{m-1} & \rho_{m-2}\sigma_3\sigma_m \\ & & & \ddots & \vdots & \vdots \\ & & & & \sigma_{m-1}^2 & \rho_1\sigma_{m-1}\sigma_m \\ & & & & & \sigma_m^2 \end{bmatrix}$$

JMP PRO Antedependent Covariance Structure

The antedependence model is a general model that is flexible and allows the correlation structure to change over time. In this model, the correlation between two observations at adjacent time points *j - 1* and *j* is unique and is denoted $\rho_{j[j-1]}$.

The correlation between pairs of observations at non-adjacent time points *j* and *j'* is the product of all the adjacent correlations in between. This is written as follows:

$$Corr(y_{ijk}, y_{ij'k} | s_{ik}) = \prod_{k=j}^{j'-1} \rho_{k[k-1]}$$

For example, the correlation between the pair of observations at time points *j*=2 and *j'*=6 would be $\rho_{21}\rho_{32}\rho_{43}\rho_{54}$.

The correlation matrix is given as follows:

$$C = \begin{bmatrix} 1 & \rho_{10} & \rho_{10}\rho_{21} & \cdots & \rho_{10}\cdots\rho_{[m-1][m-2]} & \rho_{10}\cdots\rho_{m[m-1]} \\ & 1 & \rho_{21} & \cdots & \rho_{21}\cdots\rho_{[m-1][m-2]} & \rho_{21}\cdots\rho_{m[m-1]} \\ & & 1 & \cdots & \rho_{32}\cdots\rho_{[m-1][m-2]} & \rho_{32}\cdots\rho_{m[m-1]} \\ & & & \ddots & \vdots & \vdots \\ & & & & 1 & \rho_{m[m-1]} \\ & & & & & 1 \end{bmatrix}$$

Two options in JMP use this correlation matrix:

- The Antedependent Equal Variance structure assumes equal variances across observation times while still allowing for the correlations to change. The variance among observations at any time is σ^2 and the covariance matrix is $\Sigma = \sigma^2 C$.
- The Antedependent structure allows the variance among observations at any given time to vary. Denote the variance among observations taken at time j is σ_j^2. Then the covariance matrix is as follows:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{10}\sigma_1\sigma_2 & \rho_{10}\rho_{21}\sigma_1\sigma_3 & \cdots & \rho_{10}\cdots\rho_{[m-1][m-2]}\sigma_1\sigma_{m-1} & \rho_{10}\cdots\rho_{m[m-1]}\sigma_1\sigma_m \\ & \sigma_2^2 & \rho_{21}\sigma_2\sigma_3 & \cdots & \rho_{21}\cdots\rho_{[m-1][m-2]}\sigma_2\sigma_{m-1} & \rho_{21}\cdots\rho_{m[m-1]}\sigma_2\sigma_m \\ & & \sigma_3^2 & \cdots & \rho_{32}\cdots\rho_{[m-1][m-2]}\sigma_3\sigma_{m-1} & \rho_{32}\cdots\rho_{m[m-1]}\sigma_3\sigma_m \\ & & & \ddots & \vdots & \vdots \\ & & & & \sigma_{m-1}^2 & \rho_{m[m-1]}\sigma_{m-1}\sigma_m \\ & & & & & \sigma_m^2 \end{bmatrix}$$

JMP PRO Spatial and Temporal Variability

Consider the simple model $y_i = \mu + e_i$. The spatial or temporal structure is modeled through the error term, e_i. In general, the spatial correlation model can be defined as $Var(e_i) = \sigma_i^2$ and $Cov(e_i, e_j) = \sigma_{ij}$.

Let $\mathbf{s}_i$ denote the location of y_i, where $\mathbf{s}_i$ is specified by coordinates reflecting space or time. The spatial or temporal structure is typically restricted by assuming that the covariance is a function of the Euclidean distance, d_{ij}, between $\mathbf{s}_i$ and $\mathbf{s}_j$. The covariance can be written as $Cov(e_i, e_j) = \sigma^2[f(d_{ij})]$, where $f(d_{ij})$ represents the correlation between observations y_i and y_j.

In the case of two or more location coordinates, if $f(d_{ij})$ does not depend on direction, then the covariance structure is *isotropic*. If it does, then the structure is *anisotropic*.

JMP PRO Spatial Correlation Structure

The correlation structures for spatial models available in JMP are shown below. These are parametrized by ρ, which is positive unless it is otherwise constrained.

- Spherical

$$f(d_{ij}) = [1 - 1.5(d_{ij}/\rho) + 0.5(d_{ij}/\rho)^3] \times 1_{\{d_{ij} < \rho\}}$$

$$\text{where } 1_{\{d_{ij} < \rho\}} = \begin{cases} 1, \text{ if } d_{ij} < \rho \\ 0, \text{ if } d_{ij} \geq \rho \end{cases}$$

- Exponential

$$f(d_{ij}) = \exp(-d_{ij}/\rho)$$

- Gaussian

$$f(d_{ij}) = \exp(-d_{ij}^2/\rho^2)$$

- Power

$$f(d_{ij}) = \rho^{d_{ij}}$$

For an anisotropic model, the correlation function contains a parameter, ρ_κ, for each direction.

JMP PRO Variogram

When the spatial process is second-order stationary, the structures listed in "Spatial Correlation Structure" on page 426 define *variograms*. Borrowed from geostatistics, the variogram is the standard tool for describing and estimating spatial variability. It measures spatial variability as a function of the distance, d_{ij}, between observations using the semivariance.

Let $Z(\mathbf{s})$ denote the value of the response at a location $\mathbf{s}$. The *semivariance* between observations at $\mathbf{s}_i$ and $\mathbf{s}_j$ is given as follows:

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) = (Var(Z(\mathbf{s}_i) - Z(\mathbf{s}_j)))/2$$

If the response has a constant mean, then the expression can be simplified to the following:

$$\gamma(\mathbf{s}_i, \mathbf{s}_j) = E[(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2]/2$$

If the process is isotropic, the semivariance depends only on the distance h between points and the function can be written as follows:

$$\gamma(h) = E[(Z(\mathbf{s}_i) - Z((\mathbf{s}_i + h)))^2]/2$$

The following terms are associated with variograms:

Nugget Defined as the intercept. This represents a jump discontinuity at $h = 0$.

Sill Defined as the value of the semivariogram at the plateau reached for larger distances. It corresponds to the variance of an observation. In models with no nugget effect, the sill is σ^2. In models with a nugget effect, the sill is $\sigma^2 + c_1$, where c_1 represents the nugget. The *partial sill* is defined as σ^2.

Range Defined as the distance at which the semivariogram reaches the sill. At distances less than the range, observations are spatially correlated. For distances greater than or equal to the range, spatial correlation is effectively zero. In spherical models, ρ is the range. In exponential models, 3ρ is the practical range. In Gaussian models, $\rho\sqrt{3}$ is the practical range. The practical range is defined as the distance where covariance is reduced to 95% of the sill.

In Figure 8.34 on page 406, the repeated effects covariance parameter estimates represent the various semivariogram features:

Spatial Spherical An estimate of the range, ρ.

Nugget A scaled estimate of c_1. The Residual times the Nugget is c_1.

Residual The partial sill or the sill in no nugget models.

JMP PRO Variogram Estimate

For a given isotropic spatial structure, the estimated variogram is obtained using a nonlinear least squares fit of the observed data to the appropriate function in "Spatial Correlation Structure" on page 426.

JMP PRO Empirical Semivariance

To compute the *empirical semivariance*, the distances between all pairs of points for the variables selected for the variogram covariance are computed. The range of the distances is divided into 10 equal intervals. If the data do not allow for 10 intervals, then as many intervals as possible are constructed.

Distance classes consisting of pairs of points are constructed. The h^{th} distance class consists of all pairs of points whose distances fall in the h^{th} interval.

Consider the following notation:

n total number of pairs of points

C_h distance class consisting of points whose distance falls into the h^{th} largest interval

$Z(\mathbf{x})$ value of the response at $\mathbf{x}$, where $\mathbf{x}$ is a vector of temporal or spatial coordinates

$\gamma(h)$ semivariance for distance class C_h

The semivariance function, γ, is defined as follows:

$$\gamma(h) = \begin{cases} \frac{1}{2n}\left\{ \sum_{(x,y) \text{ in } C_h} [Z(\mathbf{x}) - Z(\mathbf{y})]^2 \right\} & \text{for } h > 0 \\ \hat{c}_1 & \text{for } h = 0 \end{cases}$$

Here $\hat{c}_1$ is an estimate of the nugget effect.

JMP PRO The Kackar-Harville Correction

The variance matrix of the fixed effects is always modified to include a Kackar-Harville correction. The variance matrix of the BLUPs, and the covariances between the BLUPs and the fixed effects, are not Kackar-Harville corrected. The rationale for this approach is that corrections for BLUPs can be computationally and memory intensive when the random effects have many levels. In SAS, the Kackar-Harville correction is done for both fixed effects and BLUPs only when the DDFM=KENWARDROGER is set.

For covariance structures that have nonzero second derivatives with respect to the covariance parameters, the Kenward-Roger covariance matrix adjustment includes a second-order term. This term can result in standard error shrinkage. Also, the resulting adjusted covariance matrix can then be indefinite and is not invariant under reparameterization. The first-order Kenward-Roger covariance matrix adjustment eliminates the second derivatives from the calculation. All spatial structures and the AR(1) structure are covariance structures that generally lead to nonzero second derivatives.

Because JMP implements the Kenward-Roger first-order adjustment

- Standard errors for linear combinations involving only fixed effects parameters match PROC MIXED DDFM=KENWARDROGER(FIRSTORDER). This presumes that one has taken care to transform between the different parameterizations used by PROC MIXED and JMP.

- Standard errors for linear combinations involving only BLUP parameters match PROC MIXED DDFM=SATTERTHWAITE.
- Standard errors for linear combinations involving both fixed effects and BLUPS do not match PROC MIXED for any DDFM option if the data are unbalanced. However, these standard errors are between those obtained using the DDFM=SATTERTHWAITE and DDFM=KENWARDROGER options. If the data are balanced, JMP matches SAS regardless of the DDFM option, because the Kackar-Harville correction is null.

JMP PRO Degrees of Freedom

The degrees of freedom for tests involving only linear combinations of fixed effect parameters are calculated using the first-order Kenward-Roger correction. So JMP's results for these tests match PROC MIXED using the DDFM=KENWARDROGER(FIRSTORDER) option. If there are BLUPs in the linear combination, JMP uses a Satterthwaite approximation to get the degrees of freedom. The results then follow a pattern similar to what is described for standard errors in the preceding paragraph.

For more details about the Kackar-Harville correction and the Kenward-Roger DF approach, see Kenward and Roger (1997). The Satterthwaite method is described in detail in the MIXED procedure chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. 2017).

Chapter 9

Multivariate Response Models

Fit Relationships Using MANOVA

Multivariate models fit several responses (Y variables) to a set of effects. The functions across the Y variables can be tested with appropriate response designs.

In addition to creating standard MANOVA (Multivariate Analysis of Variance) models, you can use the following techniques:

- Repeated measures analysis when repeated measurements are taken on each subject and you want to analyze effects both between subjects and within subjects across the measurements. This multivariate approach is especially important when the correlation structure across the measurements is arbitrary.
- Canonical correlation to find the linear combination of the X and Y variables that has the highest correlation.
- Discriminant analysis to find distance formulas between points and the multivariate means of various groups so that points can be classified into the groups that they are most likely to be in. A more complete implementation of discriminant analysis is in the **Discriminant** platform.

The multivariate fit begins with a rudimentary preliminary analysis that shows parameter estimates and least squares means. You can then specify a response design across the Y variables and multivariate tests are performed.

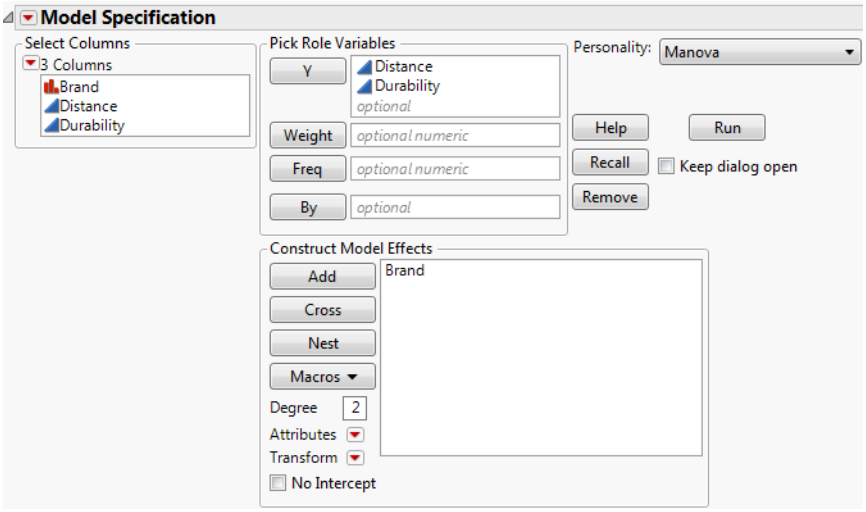
Contents

Example of a Multiple Response Model

This example uses the Golf Balls.jmp sample data table (McClave and Dietrich 1988). The data are a comparison of distances traveled and a measure of durability for three brands of golf balls. A robotic golfer hit a random sample of ten balls for each brand in a random sequence. The hypothesis to test is that distance and durability are the same for the three golf ball brands.

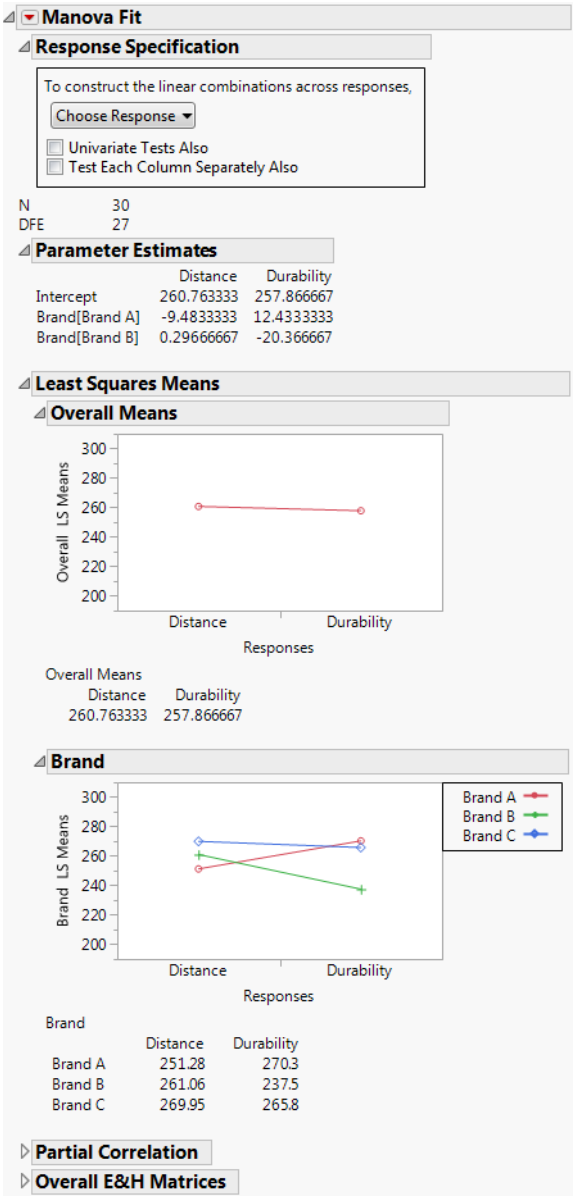
1. Select **Help > Sample Data Library** and open Golf Balls.jmp.
2. Select **Analyze > Fit Model**.
3. Select Distance and Durability and click **Y**.
4. Select Brand and click **Add**.
5. For **Personality**, select **Manova**.

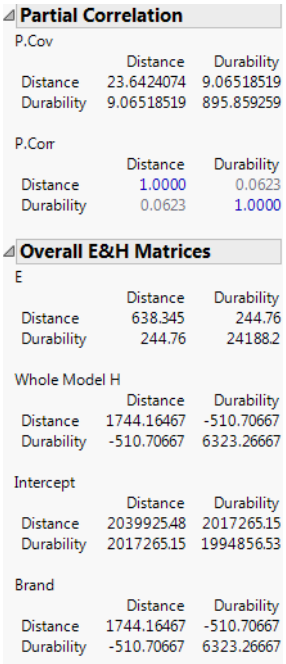
Figure 9.1 Manova Setup



6. Click **Run**.

Figure 9.2 Manova Report Window





The initial results might not be very interesting in themselves, because no response design has been specified yet. After you specify a response design, the multivariate platform displays tables of multivariate estimates and tests. For details about specifying a response design, see "Response Specification" on page 436.

The Manova Report

The Manova report window contains the following elements:

Manova Fit red triangle menu Contains save options. See "The Manova Fit Options" on page 435.

Response Specification Enables you to specify the response designs for various tests. See "Response Specification" on page 436.

Parameter Estimates Contains the parameter estimates for each response variable (without details like standard errors or *t*-tests). There is a column for each response variable.

Least Squares Means Reports the overall least squares means of all of the response columns, least squares means of each nominal level, and least squares means plots of the means.

Partial Correlation Shows the covariance matrix and the partial correlation matrix of residuals from the initial fit, adjusted for the *X* effects.

Overall E&H Matrices Shows the E and H matrices:

- The elements of the **E** matrix are the cross products of the residuals.
- The **H** matrices correspond to hypothesis sums of squares and cross products.

There is an **H** matrix for the whole model and for each effect in the model. Diagonal elements of the **E** and **H** matrices correspond to the hypothesis (numerator) and error (denominator) sum of squares for the univariate *F* tests. New **E** and **H** matrices for any given response design are formed from these initial matrices, and the multivariate test statistics are computed from them.

The Manova Fit Options

The following Manova Fit options are available:

Save Discrim Performs a discriminant analysis and saves the results to the data table. For more details, see "Discriminant Analysis" on page 450.

Save Predicted Saves the predicted responses to the data table.

Save Residuals Saves the residuals to the data table.

Model Dialog Shows the completed launch window for the current analysis.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

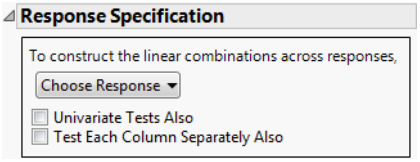
Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Response Specification

Specify the response designs for various tests using the Response Specification panel.

Figure 9.3 Response Specification Panel



Choose Response Provides choices for the M matrix. See "Choose Response Options" on page 437.

Univariate Tests Also Obtains adjusted and unadjusted univariate repeated measures tests and multivariate tests. Use in repeated measures models.

Test Each Column Separately Also Obtains univariate ANOVA tests and multivariate tests on each response.

The following buttons are available only after you have chosen a response option:

Run Performs the analysis and shows the multivariate estimates and tests. See "Multivariate Tests" on page 441.

Help Shows the help for the Response Specification panel.

Orthogonalize Orthonormalizes the matrix. Orthonormalization is done after the column contrasts (sum to zero) for all response types except Sum.

Delete Last Column Reduces the dimensionality of the transformation.

Choose Response Options

The response design forms the M matrix. The columns of an M matrix define a set of transformation variables for the multivariate analysis. The **Choose Response** button contains the following options for the M matrix:

Repeated Measures Constructs and runs both Sum and Contrast responses.

Sum Sum of the responses that gives a single value.

Identity Uses each separate response, the identity matrix.

Contrast Compares each response and the first response.

Polynomial Constructs a matrix of orthogonal polynomials.

Helmert Compares each response with the combined responses listed below it.

Profile Compares each response with the following response.

Mean Compares each response with the mean of the others.

Compound Creates and runs several response functions that are appropriate if the responses are compounded from two effects.

Custom Uses any custom M matrix that you enter.

The most typical response designs are **Repeated Measures** and **Identity** for multivariate regression. There is little difference in the tests given by the **Contrast**, **Helmert**, **Profile**, and **Mean** options, since they span the same space. However, the tests and details in the Least Squares means and Parameter Estimates tables for them show correspondingly different highlights.

The **Repeated Measures** and the **Compound** options display dialogs to specify response effect names. They then fit several response functions without waiting for further user input. Otherwise, selections expand the control panel and give you more opportunities to refine the specification.

Custom Test Option

Set up custom tests of effect levels using the **Custom Test** option.

Note: For instructions on how to create custom tests, see "Custom Test" on page 125 in the "Standard Least Squares Report and Options" chapter.

The menu icon beside each effect name gives you the following options to request additional information about the multivariate fit:

Test Details Displays the eigenvalues and eigenvectors of the $E^{-1}H$ matrix used to construct multivariate test statistics. See "Test Details" on page 438.

Centroid Plot Plots the centroids (multivariate least squares means) on the first two canonical variables formed from the test space. See "Centroid Plot" on page 439.

Save Canonical Scores Saves variables called Canon[1], Canon[2], and so on, as columns in the current data table. These columns have both the values and their formulas. For an example, see "Save Canonical Scores" on page 440. For technical details, see "Canonical Details" on page 454.

Contrast Performs the statistical contrasts of treatment levels that you specify in the contrasts dialog.

Note: The **Contrast** command is the same as for regression with a single response. See the "LSMeans Contrast" on page 104 in the "Standard Least Squares Report and Options" chapter, for a description and examples of the **LSMeans Contrast** commands.

Test Details

The Test Details report gives canonical details about the test for the whole model or the specified effect.

Eigenvalue The eigenvalues of the $E^{-1}H$ matrix used in computing the multivariate test statistics.

Canonical Corr The canonical correlations associated with each eigenvalue. This is the canonical correlation of the transformed responses with the effects, corrected for all other effects in the model.

Eigvec The eigenvectors of the $E^{-1}H$ matrix, or equivalently of $(E + \mathbf{H})^{-1}H$.

Example of Test Details

1. Select **Help > Sample Data Library** and open Iris.jmp.

   The Iris data (Mardia et al. 1979) have three levels of Species named Virginica, Setosa, and Versicolor. There are four measures (Petal length, Petal width, Sepal length, and Sepal width) taken on each sample.

2. Select **Analyze > Fit Model**.
3. Select Petal length, Petal width, Sepal length, and Sepal width and click **Y**.
4. Select Species and click **Add**.
5. For **Personality**, select **Manova**.
6. Click **Run**.
7. Click on the **Choose Response** button and select **Identity**.
8. Click **Run**.
9. From the red triangle menu next to Species, select **Test Details**.

The eigenvalues, eigenvectors, and canonical correlations appear. See Figure 9.4.

Figure 9.4 Test Details

Species

Test	Value	Approx. F	NumDF	DenDF	Prob>F
Wilks' Lambda	0.0234386	199.1453	8	288	<.0001*
Pillai's Trace	1.1918988	53.4665	8	290	<.0001*
Hotelling-Lawley	32.47732	582.1970	8	203.4	<.0001*
Roy's Max Root	32.191929	1166.9574	4	145	<.0001*

Eigenvalue	Canonical Corr
32.1919292	0.98482089
0.28539104	0.47119702
1.235e-15	0
-6.174e-16	0

Eigvec

Sepal length	-0.0684059	0.00198791	-0.2350196	0.1176771
Sepal width	-0.1265612	0.1785267	0.21657608	0.04510419
Petal length	0.18155288	-0.0768636	0.23964446	0.06563465
Petal width	0.23180286	0.23417227	-0.2865277	-0.2438953

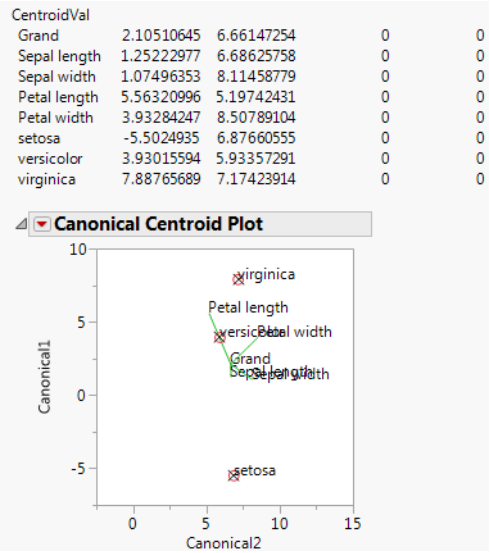
Centroid Plot

The **Centroid Plot** command (accessed from the red triangle next to Species) plots the centroids (multivariate least squares means) on the first two canonical variables formed from the test space, as in Figure 9.5. The first canonical axis is the vertical axis so that if the test space is only one dimensional the centroids align on a vertical axis. The centroid points appear with a circle corresponding to the 95% confidence region (Mardia et al. 1979). When centroid plots are created under effect tests, circles corresponding to the effect being tested appear in red. Other circles appear blue. Biplot rays show the directions of the original response variables in the test space. See "Details for Centroid Plot Option" on page 454.

Click the **Centroid Val** disclosure icon to show additional information, shown in Figure 9.5.

The first canonical axis with an eigenvalue accounts for much more separation than does the second axis. The means are well separated (discriminated), with the first group farther apart than the other two. The first canonical variable seems to load the petal length variables against the petal width variables. Relationships among groups of variables can be verified with Biplot Rays and the associated eigenvectors.

Figure 9.5 Centroid Plot and Centroid Values



Save Canonical Scores

Saves columns called Canon[i] to the data table, where *i* refers to the i^{th} canonical score for the Y variables. The canonical scores are computed based on the $E^{-1}H$ matrix used to construct the multivariate test statistic. Canonical scores are saved for eigenvectors corresponding to nonzero eigenvalues.

Canonical Correlation

Canonical correlation analysis is not a specific command, but it can be done by a sequence of commands in the multivariate fitting platform, as follows:

1. Follow step 1 through step 8 in "Example of Test Details" on page 438.
2. From the red triangle menu next to Whole Model, select **Test Details**.
3. From the red triangle menu next to Whole Model, select **Save Canonical Scores**.

The details list the canonical correlations (Canonical Corr) next to the eigenvalues. The saved variables are called Canon[1], Canon[2], and so on. These columns contain both the values and their formulas.

To obtain the canonical variables for the *X* side, repeat the same steps, but interchange the *X* and *Y* variables. If you already have the columns Canon[n] appended to the data table, the new columns are called Canon[n] 2 (or another number) that makes the name unique.

For another example, proceed as follows:

1. Select **Help > Sample Data Library** and open Exercise.jmp.

2. Select **Analyze > Fit Model**.
3. Select chins, situps, and jumps and click Y.
4. Select weight, waist, and pulse and click **Add**.
5. For **Personality**, select **Manova**.
6. Click **Run**.
7. Click on the **Choose Response** button and select **Identity**.
8. Click **Run**.
9. From the red triangle menu next to Whole Model, select **Test Details**.
10. From the red triangle menu next to Whole Model, select **Save Canonical Scores**.

Figure 9.6 Canonical Correlations

Whole Model

Test	Value	Approx. F	NumDF	DenDF	Prob>F
Wilks' Lambda	0.3503905	2.0482	9	34.223	0.0635
Pillai's Trace	0.6784815	1.5587	9	48	0.1551
Hotelling-Lawley	1.7719415	2.6397	9	19.053	0.0357*
Roy's Max Root	1.7247387	9.1986	3	16	0.0009*

Eigenvalue	Canonical Corr
1.72473874	0.79560815
0.0419084	0.20055604
0.00529433	0.07257029

Eigvec

chins	0.02503681	-0.016636	0.05641878
situps	0.00637953	0.0004622	-0.004547
jumps	-0.0052909	0.0048507	0.0018787

The output canonical variables use the eigenvectors shown as the linear combination of the Y variables. For example, the formula for Canon[1] is as follows:

```
0.02503681*chins + 0.00637953*situps + -0.0052909*jumps
```

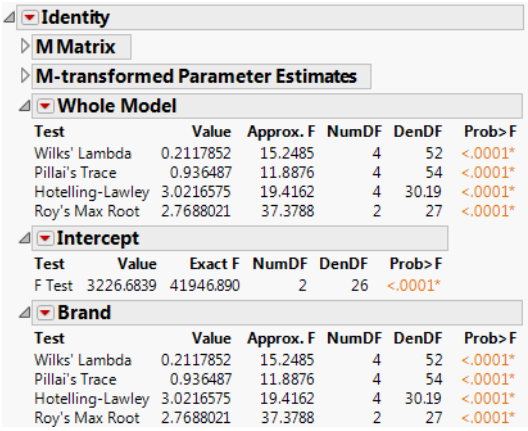
This canonical analysis does not produce a standardized variable with mean 0 and standard deviation 1, but it is easy to define a new standardized variable with the calculator that has these features.

Multivariate Tests

Example Choosing a Response

1. Complete step 1 through step 6 in "Example of a Multiple Response Model" on page 433.
2. Click on the **Choose Response** button and select **Identity**.
3. Click **Run**.

Figure 9.7 Multivariate Test Reports



Identity

M Matrix

M-transformed Parameter Estimates

Whole Model

Test	Value	Approx. F	NumDF	DenDF	Prob>F
Wilks' Lambda	0.2117852	15.2485	4	52	<.0001*
Pillai's Trace	0.936487	11.8876	4	54	<.0001*
Hotelling-Lawley	3.0216575	19.4162	4	30.19	<.0001*
Roy's Max Root	2.7688021	37.3788	2	27	<.0001*

Intercept

Test	Value	Exact F	NumDF	DenDF	Prob>F
F Test	3226.6839	41946.890	2	26	<.0001*

Brand

Test	Value	Approx. F	NumDF	DenDF	Prob>F
Wilks' Lambda	0.2117852	15.2485	4	52	<.0001*
Pillai's Trace	0.936487	11.8876	4	54	<.0001*
Hotelling-Lawley	3.0216575	19.4162	4	30.19	<.0001*
Roy's Max Root	2.7688021	37.3788	2	27	<.0001*

The **M Matrix** report gives the response design that you specified. The **M-transformed Parameter Estimates** report gives the original parameter estimates matrix multiplied by the transpose of the M matrix.

Note: Initially in this chapter, the matrix names **E** and **H** refer to the error and hypothesis cross products. After specification of a response design, **E** and **H** refer to those matrices transformed by the response design, which are actually **M′EM** and **M′HM**.

The Extended Multivariate Report

In multivariate fits, the sums of squares due to hypothesis and error are matrices of squares and cross products instead of single numbers. And there are lots of ways to measure how large a value the matrix for the hypothesis sums of squares and cross products (called **H** or **SSCP**) is compared to that matrix for the residual (called **E**). JMP reports the four multivariate tests that are commonly described in the literature. If you are looking for a test at an exact significance level, you may need to go hunting for tables in reference books. Fortunately, all four tests can be transformed into an approximate *F* test. If the response design yields a single value, or if the hypothesis is a single degree of freedom, the multivariate tests are equivalent and yield the same exact *F* test. JMP labels the test Exact F; otherwise, JMP labels it Approx. F.

In the golf balls example, there is only one effect so the Whole Model test and the test for Brand are the same, which show the four multivariate tests with approximate F tests. There is only a single intercept with two DF (one for each response), so the *F* test for it is exact and is labeled Exact F.

The red triangle menus on the Whole Model, Intercept, and Brand reports contain options to generate additional information, which includes eigenvalues, canonical correlations, a list of centroid values, a centroid plot, and a **Save** command that lets you save canonical variates.

The effect (Brand in this example) popup menu also includes the option to specify contrasts.

The custom test and contrast features are the same as those for regression with a single response. See the "Standard Least Squares Report and Options" chapter on page 75.

To see formulas for the MANOVA table tests, see "Multivariate Tests" on page 452.

The extended Multivariate Report contains the following columns:

Test Labels each statistical test in the table. If the number of response function values (columns specified in the **M** matrix) is 1 or if an effect has only one degree of freedom per response function, the exact *F* test is presented. Otherwise, the standard four multivariate test statistics are given with approximate F tests: Wilks' Lambda (Λ), Pillai's Trace, the Hotelling-Lawley Trace, and Roy's Maximum Root.

Value Value of each multivariate statistical test in the report.

Approx. F (or Exact F) *F*-values corresponding to the multivariate tests. If the response design yields a single value or if the test is one degree of freedom, this is an exact *F* test.

NumDF Numerator degrees of freedom.

DenDF Denominator degrees of freedom.

Prob>F Significance probability corresponding to the *F*-value.

Note: For details about the Sphericity Test table, see "Univariate Tests and the Test for Sphericity" on page 444).

Comparison of Multivariate Tests

Although the four standard multivariate tests often give similar results, there are situations where they differ, and one might have advantages over another. Unfortunately, there is no clear winner. In general, here is the order of preference in terms of power:

1. Pillai's Trace
2. Wilks' Lambda
3. Hotelling-Lawley Trace
4. Roy's Maximum Root

When there is a large deviation from the null hypothesis and the eigenvalues differ widely, the order of preference is the reverse (Seber 1984).

Univariate Tests and the Test for Sphericity

There are cases, such as a repeated measures model, that allow transformation of a multivariate problem into a univariate problem (Huynh and Feldt 1970). Using univariate tests in a multivariate context is valid in the following situations:

- If the response design matrix **M** is orthonormal (**M′M** = Identity).
- If **M** yields more than one response the coefficients of each transformation sum to zero.
- If the *sphericity* condition is met. The sphericity condition means that the M-transformed responses are uncorrelated and have the same variance. **M′ΣM** is proportional to an identity matrix, where Σ is the covariance of the *Y* variables.

If these conditions hold, the diagonal elements of the **E** and **H** test matrices sum to make a univariate sums of squares for the denominator and numerator of an *F* test. Note that if the above conditions do not hold, then an error message appears. In the case of Golf Balls.jmp, an identity matrix is specified as the **M**-matrix. Identity matrices cannot be transformed to a full rank matrix after centralization of column vectors and orthonormalization. So the univariate request is ignored.

Example of Univariate and Sphericity Test

1. Select **Help > Sample Data Library** and open Dogs.jmp.
2. Select **Analyze > Fit Model**.
3. Select LogHist0, LogHist1, LogHist3, and LogHist5 and click **Y**.
4. Select drug and dep1 and click **Add**.
5. In the Construct Model Effects panel, select drug. In the Select Columns panel, select dep1. Click **Cross**.
6. For Personality, select **Manova**.
7. Click **Run**.
8. Select the check box next to **Univariate Tests Also**.
9. In the **Choose Response** menu, select **Repeated Measures**.

   Time should be entered for YName, and **Univariate Tests Also** should be selected.
10. Click **OK**.

Figure 9.8 Sphericity Test

Sphericity Test	
Mauchly Criterion	0.1752641
ChiSquare	16.930873
DF	5
Prob >Chisq	0.0046328

The sphericity test checks the appropriateness of an unadjusted univariate *F* test for the within-subject effects using the Mauchly criterion to test the sphericity assumption (Anderson 1958). The sphericity test and the univariate tests are always done using an orthonormalized **M** matrix. You interpret the sphericity test as follows:

- If the true covariance structure is spherical, you can use the unadjusted univariate *F*-tests.
- If the sphericity test is significant, the test suggests that the true covariance structure is not spherical. Therefore, you can use the multivariate or the adjusted univariate tests.

The univariate *F* statistic has an approximate *F*-distribution even without sphericity, but the degrees of freedom for numerator and denominator are reduced by some fraction epsilon (ε). Box (1954), Greenhouse and Geisser (1959), and Huynh-Feldt (1976) offer techniques for estimating the epsilon degrees-of-freedom adjustment. Muller and Barton (1989) recommend the Greenhouse-Geisser version, based on a study of power.

The epsilon adjusted tests in the multivariate report are labeled G-G (Greenhouse-Geisser) or H-F (Huynh-Feldt), with the epsilon adjustment shown in the value column.

Multivariate Model with Repeated Measures

One common use of multivariate fitting is to analyze data with repeated measures, also called *longitudinal data*. A subject is measured repeatedly across time, and the data are arranged so that each of the time measurements form a variable. Because of correlation between the measurements, data should not be stacked into a single column and analyzed as a univariate model unless the correlations form a pattern termed *sphericity*. See the previous section, "Univariate Tests and the Test for Sphericity" on page 444, for more details about this topic.

With repeated measures, the analysis is divided into two layers:

- Between-subject (or across-subject) effects are modeled by fitting the sum of the repeated measures columns to the model effects. This corresponds to using the **Sum** response function, which is an **M**-matrix that is a single vector of 1s.
- Within-subjects effects (repeated effects, or time effects) are modeled with a response function that fits differences in the repeated measures columns. This analysis can be done using the **Contrast** response function or any of the other similar differencing functions: **Polynomial**, **Helmert**, **Profile**, or **Mean**. When you model differences across the repeated measures, think of the differences as being a new within-subjects effect, usually time. When you fit effects in the model, interpret them as the interaction with the within-subjects effect. For example, the effect for Intercept becomes the Time (within-subject) effect, showing overall differences across the repeated measures. If you have an effect A, the within-subjects tests are interpreted to be the tests for the A*Time interaction, which model how the differences across repeated measures vary across the A effect.

Table 9.1 on page 447 shows the relationship between the response function and the model effects compared with what a univariate model specification would be. Using both the **Sum** (between-subjects) and **Contrast** (within-subjects) models, you should be able to reconstruct the tests that would have resulted from stacking the responses into a single column and obtaining a standard univariate fit.

There is a direct and an indirect way to perform the repeated measures analyses:

- The direct way is to use the popup menu item Repeated Measures. This prompts you to name the effect that represents the within-subject effect across the repeated measures. Then it fits both the **Contrast** and the **Sum** response functions. An advantage of this way is that the effects are labeled appropriately with the within-subjects effect name.
- The indirect way is to specify the two response functions individually. First, do the **Sum** response function and second, do either **Contrast** or one of the other functions that model differences. You need to remember to associate the within-subjects effect with the model effects in the contrast fit.

Repeated Measures Example

For example, consider a study by Cole and Grizzle (1966). The results are in the Dogs.jmp table in the sample data folder. Sixteen dogs are assigned to four groups defined by variables drug and dep1, each having two levels. The dependent variable is the blood concentration of histamine at 0, 1, 3, and 5 minutes after injection of the drug. The log of the concentration is used to minimize the correlation between the mean and variance of the data.

1. Select **Help > Sample Data Library** and open Dogs.jmp.
2. Select **Analyze > Fit Model**.
3. Select LogHist0, LogHist1, LogHist3, and LogHist5 and click **Y**.
4. Select drug and dep1 and select **Full Factorial** from the **Macros** menu.
5. For Personality, select **Manova**.
6. Click **Run**.
7. In the **Choose Response** menu, select **Repeated Measures**.

   Time should be entered for YName. If you check the **Univariate Tests Also** check box, the report includes univariate tests, which are calculated as if the responses were stacked into a single column.
8. Click **OK**.

Figure 9.9 Repeated Measures Window

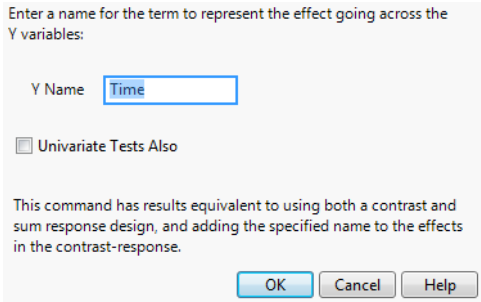


Table 9.1 shows how the multivariate tests for a **Sum** and **Contrast** response designs correspond to how univariate tests would be labeled if the data for columns LogHist0, LogHist1, LogHist3, and LogHist5 were stacked into a single *Y* column, with the new rows identified with a nominal grouping variable, Time.

Table 9.1 Corresponding Multivariate and Univariate Tests

Sum M-Matrix Between Subjects		Contrast M-Matrix Within Subjects	
Multivariate Test	**Univariate Test**	**Multivariate Test**	**Univariate Test**
intercept	intercept	intercept	time
drug	drug	drug	time*drug
depl	depl	depl	time*depl

The between-subjects analysis is produced first. This analysis is the same (except titling) as it would have been if **Sum** had been selected on the popup menu.

The within-subjects analysis is produced next. This analysis is the same (except titling) as it would have been if **Contrast** had been selected on the popup menu, though the within-subject effect name (Time) has been added to the effect names in the report. Note that the position formerly occupied by Intercept is Time, because the intercept term is estimating overall differences across the repeated measurements.

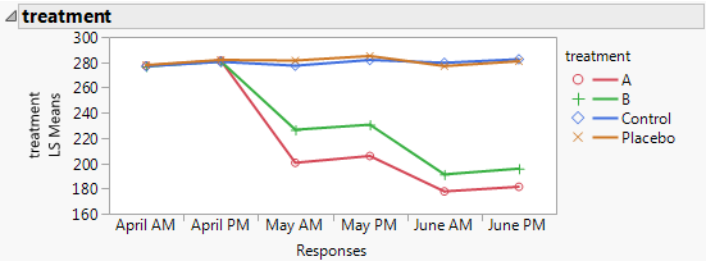
Example of a Compound Multivariate Model

JMP can handle data with layers of repeated measures. For example, see the Cholesterol.jmp data table. Groups of five participants belong to one of four treatment groups called A, B,

Control, and Placebo. Cholesterol was measured in the morning and again in the afternoon once a month for three months (the data are fictional). In this example, the response columns are arranged chronologically with time of day within month.

1. Select **Help > Sample Data Library** and open Cholesterol.jmp.
2. Select **Analyze > Fit Model**.
3. Select April AM, April PM, May AM, May PM, June AM, and June PM and click **Y**.
4. Select treatment and click **Add**.
5. Next to Personality, select **Manova**.
6. Click **Run**.

Figure 9.10 Treatment Graph



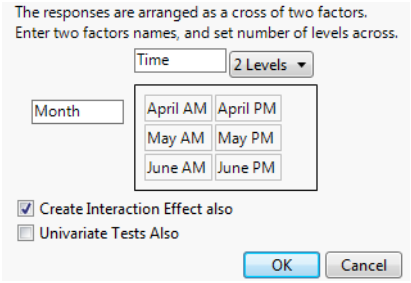
In the treatment graph, you can see that the four treatment groups began the study with very similar mean cholesterol values. The A and B treatment groups appear to have lower cholesterol values at the end of the trial period. The control and placebo groups remain unchanged.

7. Click on the **Choose Response** menu and select **Compound**.

   Complete this window to tell JMP how the responses are arranged in the data table and the number of levels of each response. In the cholesterol example, the time of day columns are arranged within month. Therefore, you name time of day as one factor and the month effect as the other factor. Testing the interaction effect is optional.

8. Use the options in Figure 9.11 to complete the window.

Figure 9.11 Compound Window



9. Click **OK**.

   The tests for each effect appear. Parts of the report are shown in Figure 9.12. Note the following:

   – The report for Time shows a *p*-value of 0.6038 for the interaction between Time and treatment, indicating that the interaction is not significant. This means that there is no evidence of a difference in treatment between AM and PM. Since Time has two levels (AM and PM) the exact *F*-test appears.

   – The report for Month shows *p*-values of <.0001 for the interaction between Month and treatment, indicating that the interaction is significant. This suggests that the differences between treatment groups change depending on the month. The treatment graph in Figure 9.10 indicates no difference among the groups in April, but the difference between treatment types (A, B, Control, and Placebo) becomes large in May and even larger in June.

   – The report for Time*Month shows no significant *p*-values for treatment. This indicates that the three-way interaction effect involving Time, Month, and treatment is not statistically significant.

Figure 9.12 Cholesterol Study Results

Compound

Time

M Matrix

M-transformed Parameter Estimates

Whole Model

Intercept

treatment

Test	Value	Exact F	NumDF	DenDF	Prob>F
F Test	0.1188721	0.6340	3	16	0.6038

Compound

Month

M Matrix

M-transformed Parameter Estimates

Whole Model

Intercept

treatment

Test	Value	Approx. F	NumDF	DenDF	Prob>F
Wilks' Lambda	0.013025	38.8109	6	30	<.0001*
Pillai's Trace	1.3128917	10.1907	6	32	<.0001*
Hotelling-Lawley	50.753209	123.4368	6	18.326	<.0001*
Roy's Max Root	50.255302	268.0283	3	16	<.0001*

Compound

Time*Month

M Matrix

M-transformed Parameter Estimates

Whole Model

Intercept

treatment

Test	Value	Approx. F	NumDF	DenDF	Prob>F
Wilks' Lambda	0.6623742	1.1435	6	30	0.3619
Pillai's Trace	0.3582613	1.1638	6	32	0.3498
Hotelling-Lawley	0.4785668	1.1639	6	18.326	0.3671
Roy's Max Root	0.4008469	2.1378	3	16	0.1355

Discriminant Analysis

Discriminant analysis is a method of predicting some level of a one-way classification based on known values of the responses. The technique is based on how close a set of measurement variables are to the multivariate means of the levels being predicted. Discriminant analysis is more fully implemented using the Discriminant Platform (see the Discriminant Analysis chapter in the *Multivariate Methods* book).

In JMP you specify the measurement variables as *Y* effects and the classification variable as a single *X* effect. The multivariate fitting platform gives estimates of the means and the covariance matrix for the data, assuming that the covariances are the same for each group. You obtain discriminant information with the **Save Discrim** option in the popup menu next to the MANOVA platform name. This command saves distances and probabilities as columns in the current data table using the initial **E** and **H** matrices.

For a classification variable with k levels, JMP adds k distance columns, k classification probability columns, the predicted classification column, and two columns of other computational information to the current data table.

Example of the Save Discrim Option

Examine Fisher's Iris data as found in Mardia et al. (1979). There are $k = 3$ levels of species and four measures on each sample.

1. Select **Help > Sample Data Library** and open Iris.jmp.
2. Select **Analyze > Fit Model**.
3. Select Sepal length, Sepal width, Petal length, and Petal width and click **Y**.
4. Select Species and click **Add**.
5. Next to Personality, select **Manova**.
6. Click **Run**.
7. From the red triangle menu next to Manova Fit, select **Save Discrim**.

The following columns are added to the Iris.jmp sample data table:

SqDist[0] Quadratic form needed in the Mahalanobis distance calculations.

SqDist[setosa] Mahalanobis distance of the observation from the Setosa centroid.

SqDist[versicolor] Mahalanobis distance of the observation from the Versicolor centroid.

SqDist[virginica] Mahalanobis distance of the observation from the Virginica centroid.

Prob[0] Sum of the negative exponentials of the Mahalanobis distances, used below.

Prob[setosa] Probability of being in the Setosa category.

Prob[versicolor] Probability of being in the Versicolor category.

Prob[virginica] Probability of being in the Virginica category.

Pred Species Species that is most likely from the probabilities.

Now you can use the new columns in the data table with other JMP platforms to summarize the discriminant analysis with reports and graphs. For example:

1. From the updated Iris.jmp data table (that contains the new columns) select **Analyze > Fit Y by X**.
2. Select Species and click **Y, Response**.
3. Select Pred Species and click **X, Factor**.

4. Click **OK**.

The Contingency Table summarizes the discriminant classifications. Three misclassifications are identified.

Figure 9.13 Contingency Table of Predicted and Actual Species

Contingency Table

Pred Species	Count / Total % / Col % / Row %	setosa	versicolor	virginica	Total
		Species			
setosa		50	0	0	50
		33.33	0.00	0.00	33.33
		100.00	0.00	0.00	
		100.00	0.00	0.00	
versicolor		0	48	1	49
		0.00	32.00	0.67	32.67
		0.00	96.00	2.00	
		0.00	97.96	2.04	
virginica		0	2	49	51
		0.00	1.33	32.67	34.00
		0.00	4.00	98.00	
		0.00	3.92	96.08	
Total		50	50	50	150
		33.33	33.33	33.33	

Statistical Details for the Manova Personality

Multivariate Tests

In the following, **E** is the residual cross product matrix and **H** is the model cross product matrix. Diagonal elements of **E** are the residual sums of squares for each variable. Diagonal elements of **H** are the sums of squares for the model for each variable. In the discriminant analysis literature, **E** is often called **W**, where **W** stands for *within*.

Test statistics in the multivariate results tables are functions of the eigenvalues λ of $\mathbf{E}^{-1}\mathbf{H}$. The following list describes the computation of each test statistic.

Note: After specification of a response design, the initial **E** and **H** matrices are premultiplied by **M'** and postmultiplied by **M**.

- Wilks' Lambda

$$\Lambda = \frac{\det(\mathbf{E})}{\det(\mathbf{H}+\mathbf{E})} = \prod_{i=1}^{n}\left(\frac{1}{1+\lambda_i}\right)$$

- Pillai's Trace

$$V = \text{Trace}[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}] = \sum_{i=1}^{n}\frac{\lambda_i}{1+\lambda_i}$$

- Hotelling-Lawley Trace

$$U = \text{Trace}(\mathbf{E}^{-1}\mathbf{H}) = \sum_{i=1}^{n}\lambda_i$$

- Roy's Max Root

  $\Theta = \lambda_1$, the maximum eigenvalue of $\mathbf{E}^{-1}\mathbf{H}$.

E and **H** are defined as follows:

$$\mathbf{E} = \mathbf{Y}'\mathbf{Y} - \mathbf{b}'(\mathbf{X}'\mathbf{X})\mathbf{b}$$

$$\mathbf{H} = (\mathbf{L}\mathbf{b})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}')^{-1}(\mathbf{L}\mathbf{b})$$

where **b** is the estimated vector for the model coefficients and $\mathbf{A}^{-}$ denotes the generalized inverse of a matrix **A**.

The whole model **L** is a column of zeros (for the intercept) concatenated with an identity matrix having the number of rows and columns equal to the number of parameters in the model. **L** matrices for effects are subsets of rows from the whole model **L** matrix.

Approximate F-Tests

To compute *F*-values and degrees of freedom, let p be the rank of $\mathbf{H}+\mathbf{E}$. Let q be the rank of $\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}'$, where the **L** matrix identifies elements of $\mathbf{X}'\mathbf{X}$ associated with the effect being tested. Let v be the error degrees of freedom and s be the minimum of p and q. Also let $m = 0.5(|p-q|-1)$ and $n = 0.5(v-p-1)$.

Table 9.2 on page 454, gives the computation of each approximate F from the corresponding test statistic.

Table 9.2 Approximate *F*-statistics

Test	Approximate F	Numerator DF	Denominator DF
Wilks' Lambda	$F = \left(\frac{1-\Lambda^{1/t}}{\Lambda^{1/t}}\right)\left(\frac{rt-2u}{pq}\right)$	pq	$rt-2u$
Pillai's Trace	$F = \left(\frac{V}{s-V}\right)\left(\frac{2n+s+1}{2m+s+1}\right)$	$s(2m+s+1)$	$s(2n+s+1)$
Hotelling-Lawley Trace	$F = \frac{2(sn+1)U}{s^2(2m+s+1)}$	$s(2m+s+1)$	$2(sn+1)$
Roy's Max Root	$F = \frac{\Theta(v-\max(p,q)+q)}{\max(p,q)}$	$\max(p,q)$	$v-\max(p,q)+q$

Canonical Details

Details for the Test Details Option

When you select the Test Details option for a given test, eigenvalues, canonical correlations, and eigenvectors are shown in the report.

The canonical correlations produced by the Test Details option are computed as follows:

$$\rho_i = \sqrt{\frac{\lambda_i}{1+\lambda_i}}$$

where λ_i is the i^{th} eigenvalue of the $\mathbf{E}^{-1}\mathbf{H}$ matrix used in computing the multivariate test statistics

The matrix labeled Eigvec is the **V** matrix, which is the matrix of eigenvectors of $\mathbf{E}^{-1}\mathbf{H}$ for the given test.

Note: The **E** and **H** matrices for the given test refer to **M'EM** and **M'HM** in terms of the original **E** and **H** matrices. The **M** matrix is defined by the response design. The **E** and **H** used in this section are defined in "Multivariate Tests" on page 452.

Details for Centroid Plot Option

The total sample centroid and centroid values for effects are computed as follows:

$$\text{Grand} = (c'_1\bar{y}, c'_2\bar{y}, \ldots, c'_g\bar{y})$$

$$\text{Effect}_j = (c'_1\bar{x}_j, c'_2\bar{x}_j, \ldots, c'_g\bar{x}_j)$$

where

$$c_i = \left(v'_i\left(\frac{E}{N-r}\right)v_i\right)^{-1/2} v_i$$

N is the number of observations

v_i is the i^{th} column of **V**, the eigenvector matrix of $\mathbf{E}^{-1}\mathbf{H}$ for the given test

$\bar{x}_j$ is the multivariate least squares mean for the j^{th} effect

$\bar{y}$ is the overall mean of the responses

g is the number of eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ greater than 0

r is the rank of the **X** matrix

Note: The **E** and **H** matrices for the given test refer to **M'EM** and **M'HM** in terms of the original **E** and **H** matrices. The **M** matrix is defined by the response design. The **E** and **H** used in this section are defined in "Multivariate Tests" on page 452.

The centroid radii for effects are calculated as follows:

$$d = \sqrt{\frac{\chi^2_g(0.95)}{L(X'X)^{-1}L'}}$$

where g is the number of eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ greater than 0 and the **L** matrices in the denominator are from the multivariate least squares means calculations.

Details for the Save Canonical Scores Option

The canonical Y values are calculated as follows:

$$\tilde{Y} = \mathbf{YM'V}$$

where

Y is the matrix of response variables

M' is the transpose of the response design matrix

V is the matrix of eigenvectors of $\mathbf{E}^{-1}\mathbf{H}$ for the given test

Note: The **E** and **H** matrices for the given test refer to **M'EM** and **M'HM** in terms of the original **E** and **H** matrices. The **M** matrix is defined by the response design. The **E** and **H** used in this section are defined in "Multivariate Tests" on page 452.

Canonical Y values are saved for eigenvectors corresponding to eigenvalues larger than zero.

Chapter 10

Loglinear Variance Models

Model the Variance and the Mean of the Response

The Loglinear Variance personality of the Fit Model platform enables you to model both the expected value and the variance of a response using regression models. The log of the variance is fit to one linear model and the expected response is fit to a different linear model simultaneously.

Note: The estimates are demanding in their need for a lot of well-designed, well-fitting data. You need more data to fit variances than you do means.

For many engineers, the goal of an experiment is not to maximize or minimize the response itself, but to aim at a target response and achieve minimum variability. The loglinear variance model provides a very general and effective way to model variances, and can be used for unreplicated data, as well as data with replications.

Contents

Overview of the Loglinear Variance Model

The loglinear-variance model provides a way to model the variance simply through a linear model. See Harvey (1976), Cook and Weisberg (1983), Aitken (1987), and Carroll and Ruppert (1988). In addition to having regressor terms to model the mean response, there are regressor terms in a linear model to model the log of the variance:

mean model: $E(y) = \mathbf{X}\beta$

variance model: $\log(Variance(y)) = \mathbf{Z}\,\lambda$,

or equivalently

$Variance(y) = \exp(\mathbf{Z}\,\lambda)$

where the columns of X are the regressors for the mean of the response, and the columns of Z are the regressors for the variance of the response. The regular linear model parameters are represented by β, and λ represents the parameters of the variance model.

Log-linear variance models are estimated using REML.

A *dispersion* or *log-variance* effect can model changes in the variance of the response. This is implemented in the Fit Model platform by a fitting personality called the Loglinear Variance personality.

Dispersion Effects

Modeling dispersion effects is not very widely covered in textbooks, with the exception of the Taguchi framework. In a Taguchi-style experiment, this is handled by taking multiple measurements across settings of an outer array, constructing a new response which measures the variability off-target across this outer array, and then fitting the model to find out the factors that produce minimum variability. This kind of modeling requires a specialized design that is a complete Cartesian product of two designs. The method of this chapter models variances in a more flexible, model-based approach. The particular performance statistic that Taguchi recommends for variability modeling is $STD = -\log(s)$. In JMP's methodology, the $\log(s^2)$ is modeled and combined with a model that has a mean. The two are basically equivalent, since $\log(s^2)=2\log(s)$.

Model Specification

Log-linear variance effects are specified in the Fit Model dialog by highlighting them and selecting **LogVariance Effect** from the **Attributes** drop-down menu. **&LogVariance** appears at the end of the effect. When you use this attribute, it also changes the fitting **Personality** at the top to **LogLinear Variance**. If you want an effect to be used for both the mean and variance of the response, then you must specify it twice, once with the **LogVariance** option.

The effects you specify with the log-variance attribute become the effects that generate the Z's in the model, and the other effects become the X's in the model.

Notes

Every time another parameter is estimated for the mean model, at least one more observation is needed, and preferably more. But with variance parameters, several more observations for each variance parameter are needed to obtain reasonable estimates. It takes more data to estimate variances than it does means.

The log-linear variance model is a very flexible way to fit dispersion effects, and the method deserves much more attention than it has received so far in the literature.

Example Using Loglinear Variance

The data table InjectionMolding.jmp contains the experimental results from a 7-factor 2^{7-3} fractional factorial design with four added centerpoints (Montgomery 1991). Preliminary investigation determined that the mean response only seemed to vary with the first two factors, Mold Temperature, and Screw Speed, and the variance seemed to be affected by Holding Time.

Figure 10.1 Injection Molding Data

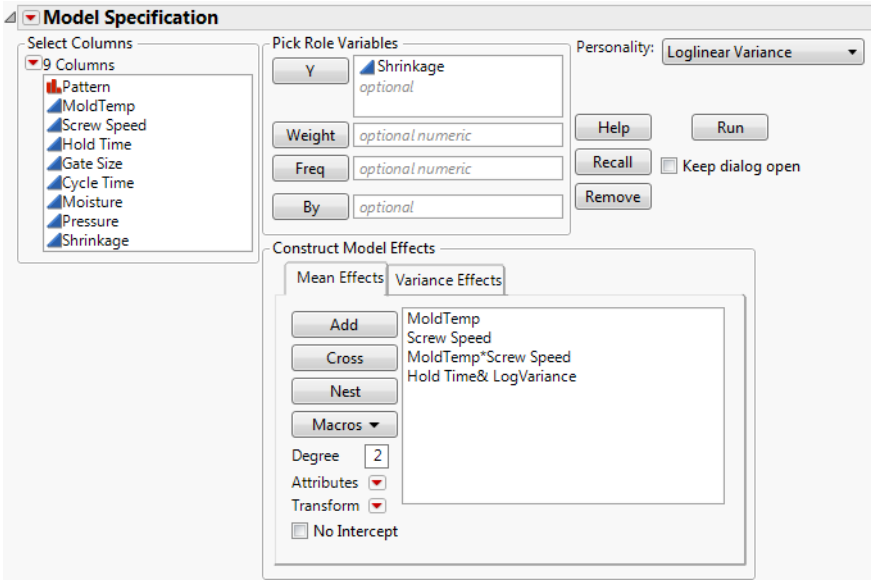
InjectionMolding
Model
Columns (9/0)
Pattern
MoldTemp *
Screw Speed *
Hold Time *
Gate Size *
Cycle Time +*
Moisture +*
Pressure +*
Shrinkage *
Rows
All rows 20
Selected 0
Excluded 0
Hidden 0
Labelled 0

	Pattern	MoldTemp	Screw Speed	Hold Time	Gate Size	Cycle Time	Moisture	Pressure	Shrinkage
1	-------	-1	-1	-1	-1	-1	-1	-1	6
2	+-----++	1	-1	-1	-1	1	-1	1	10
3	-+---+-	-1	1	-1	-1	1	1	-1	32
4	+*---++-	1	1	-1	-1	-1	1	1	60
5	--+-++-	-1	-1	1	-1	1	1	1	4
6	+-+-+-+	1	-1	1	-1	-1	1	-1	15
7	-++--++	-1	1	1	-1	-1	-1	1	26
8	+++----	1	1	1	-1	1	-1	-1	60
9	---++++	-1	-1	-1	1	-1	1	1	8
10	+--++--	1	-1	-1	1	1	1	-1	12
11	-+-+-+-	-1	1	-1	1	1	-1	1	34
12	++-+--+	1	1	-1	1	-1	-1	-1	60
13	--++--+	-1	-1	1	1	1	-1	-1	16
14	+-++-+-	1	-1	1	1	-1	-1	1	5
15	-++++--	-1	1	1	1	-1	1	-1	37
16	+++++...	1	1	1	1	1	1	1	52
17	0000000	0	0	0	0	0	0	0	25
18	0000000	0	0	0	0	0	0	0	29
19	0000000	0	0	0	0	0	0	0	24
20	0000000	0	0	0	0	0	0	0	27

1. Select **Help > Sample Data Library** and open InjectionMolding.jmp.
2. Select **Analyze > Fit Model**.

Since the variables in the data table have preselected roles assigned to them, the launch window is already filled out.

Figure 10.2 Fit Model Dialog



3. Click **Run**.

Figure 10.3 Loglinear Variance Report Window

Loglinear Variance Fit

Fit Statistics

-2 Residual Log Likelihood	90.712809
-2 Log Likelihood	87.777035
AICc	106.23857
BIC	105.75143

Mean Model for Shrinkage

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	27.582516	0.380697	72.45	<.0001*
MoldTemp	7.683713	0.392907	19.56	<.0001*
Screw Speed	18.673515	0.392907	47.53	<.0001*
MoldTemp*Screw Speed	5.765297	0.392907	14.67	<.0001*

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
MoldTemp	1	1	5.758	382.4387	<.0001*
Screw Speed	1	1	5.758	2258.768	<.0001*
MoldTemp*Screw Speed	1	1	5.758	215.3094	<.0001*

Variance Model for Shrinkage

Likelihood Ratio Test for Equal Variance

Source	-2 Residual Log Likelihood
Equal Variances Initially	101.76083063
After Fitted Variances	90.712808548
Difference: Chi-sq	11.048022086
Degrees of Freedom	1
Prob > ChiSquare	0.0008878186

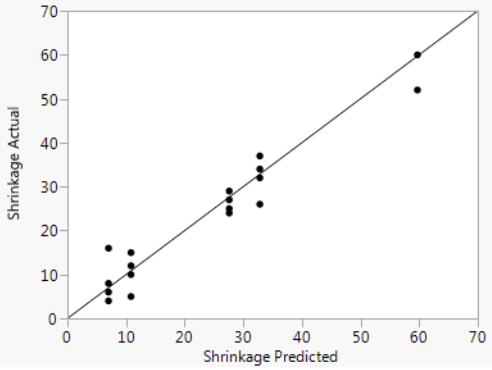
Variance Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95% Limit	Upper 95% Limit
Hold Time	1.5788108	0.412603	11.0480	0.0009*	0.7057476	2.37325
Residual	6.6912718	2.445524	7.4864	0.0062*	3.2689762	13.696373

Variance Effect Likelihood Ratio Tests

Source	Test Type	DF	ChiSquare	Prob>ChiSq
Hold Time	Likelihood	1	11.0480	0.0009*

Actual by Predicted Plot

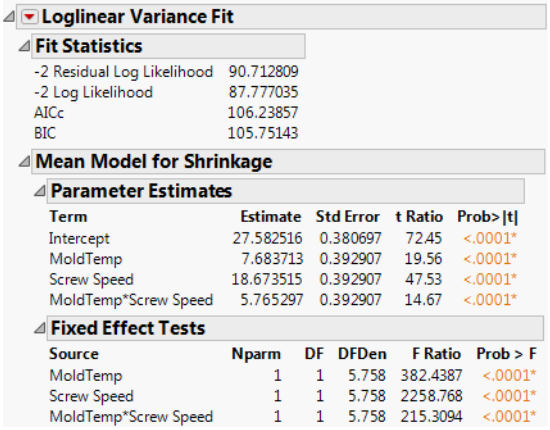


The Mean Model for Shrinkage report gives the parameters for the mean model, and the Variance Model for Shrinkage report gives the parameters for the variance model.

The Loglinear Report

The top portion of the resulting report shows the fitting of the Expected response. Table 8.2 on page 408 describes how the fit statistics are calculated. The Parameter Estimates and Fixed Effect Tests are similar to reports found in the standard least squares personality, though they are derived from restricted maximum likelihood (REML).

Figure 10.4 Mean Model Output



Loglinear Variance Fit

Fit Statistics

-2 Residual Log Likelihood	90.712809
-2 Log Likelihood	87.777035
AICc	106.23857
BIC	105.75143

Mean Model for Shrinkage

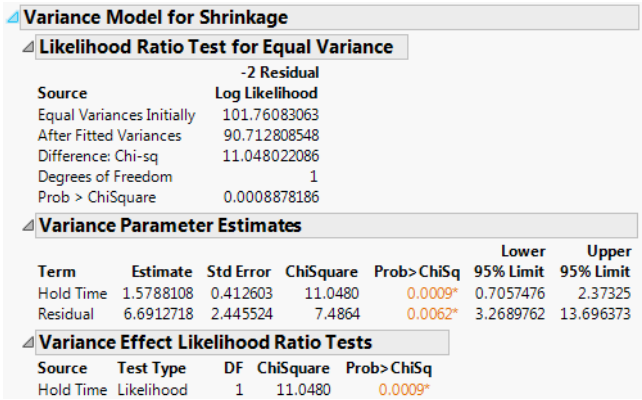
Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	27.582516	0.380697	72.45	<.0001*
MoldTemp	7.683713	0.392907	19.56	<.0001*
Screw Speed	18.673515	0.392907	47.53	<.0001*
MoldTemp*Screw Speed	5.765297	0.392907	14.67	<.0001*

Fixed Effect Tests

Source	Nparm	DF	DFDen	F Ratio	Prob > F
MoldTemp	1	1	5.758	382.4387	<.0001*
Screw Speed	1	1	5.758	2258.768	<.0001*
MoldTemp*Screw Speed	1	1	5.758	215.3094	<.0001*

Figure 10.5 Variance Model Output



Variance Model for Shrinkage

Likelihood Ratio Test for Equal Variance

Source	-2 Residual Log Likelihood
Equal Variances Initially	101.76083063
After Fitted Variances	90.712808548
Difference: Chi-sq	11.048022086
Degrees of Freedom	1
Prob > ChiSquare	0.0008878186

Variance Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq	Lower 95% Limit	Upper 95% Limit
Hold Time	1.5788108	0.412603	11.0480	0.0009*	0.7057476	2.37325
Residual	6.6912718	2.445524	7.4864	0.0062*	3.2689762	13.696373

Variance Effect Likelihood Ratio Tests

Source	Test Type	DF	ChiSquare	Prob>ChiSq
Hold Time	Likelihood	1	11.0480	0.0009*

The second portion of the report shows the fit of the variance model. The **Variance Parameter Estimates** report shows the estimates and relevant statistics. Two hidden columns are provided:

- The hidden column **exp(Estimate)** is the exponential of the estimate. So, if the factors are coded to have +1 and -1 values, then the +1 level for a factor would have the variance

multiplied by the **exp(Estimate)** value and the -1 level would have the variance multiplied by the reciprocal of this column. To see a hidden column, right-click on the report and select the name of the column from the **Columns** menu that appears.

- The hidden column labeled **exp(2|Estimate|)**is the ratio of the higher to the lower variance if the regressor has the range -1 to +1.

The report also shows the standard error, chi-square, *p*-value, and profile likelihood confidence limits of each estimate. The residual parameter is the overall estimate of the variance, given all other regressors are zero.

Does the variance model fit significantly better than the original model? The likelihood ratio test for this question compares the fitted model with the model where all parameters are zero except the intercept, the model of equal-variance. In this case the *p*-value is highly significant. Changes in Hold Time change the variance.

The **Variance Effect Likelihood Ratio Tests** refit the model without each term in turn to create the likelihood ratio tests. These are generally more trusted than Wald tests.

Loglinear Platform Options

The Loglinear Variance Fit menu contains the following options:

Save Columns Creates one or more columns in the data table. See "Save Columns" on page 465.

Row Diagnostics Plots row diagnostics. See "Row Diagnostics" on page 465.

Profilers Opens the Profiler, Contour Profiler, or Surface Profiler. See "Factor Profiling" on page 164 in the "Standard Least Squares Report and Options" chapter.

Model Dialog Shows the completed launch window for the current analysis.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Save Columns

The following Save Columns options are available:

Prediction Formula Creates a new column called Mean. The new column contains the predicted values for the mean, as computed by the specified model.

Variance Formula Creates a new column called Variance. The new column contains the predicted values for the variance, as computed by the specified model.

Std Dev Formula Creates a new column called Std Dev. The new column contains the predicted values for the standard deviation, as computed by the specified model.

Residuals Creates a new column called Residual that contains the residuals, which are the observed response values minus predicted values. See "Examining the Residuals" on page 466.

Studentized Residuals Creates a new column called Studentized Resid. The new column values are the residuals divided by their standard error.

Std Error of Predicted Creates a new column called Std Err Pred. The new column contains the standard errors of the predicted values.

Std Error of Individual Creates a new column called Std Err Indiv. The new column contains the standard errors of the individual predicted values.

Mean Confidence Interval Creates two new columns, Lower 95% Mean and Upper 95% Mean. The new columns contain the bounds for a confidence interval for the prediction mean.

Indiv Confidence Interval Creates two new columns, Lower 95% Indiv and Upper 95% Indiv. The new columns contain confidence limits for individual response values.

Row Diagnostics

The following Row Diagnostics options are available:

Plot Actual by Predicted Plots the observed values by the predicted values of *Y*. This is the leverage plot for the whole model.

Plot Studentized Residual by Predicted Plots the Studentized residuals by the predicted values of *Y*.

Plot Studentized Residual by Row Plots the Studentized residuals by row.

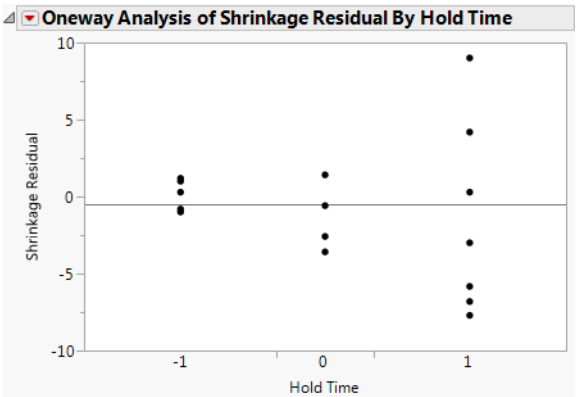
Examining the Residuals

To see the dispersion effect, follow these steps:

1. Select **Help > Sample Data Library** and open InjectionMolding.jmp.
2. Select **Analyze > Fit Model**.

   Since the variables in the data table have preselected roles assigned to them, the launch window is already filled out.
3. Click **Run**.
4. From the red triangle menu next to Loglinear Variance Fit, select **Save Columns > Residuals**.
5. In the InjectionMolding.jmp sample data table, right-click on the continuous icon next to Hold Time in the Columns panel, and select **Nominal**.
6. Select **Analyze > Fit Y by X**.
7. Select Shrinkage Residual and click **Y, Response**.
8. Select Hold Time and click **X, Factor**.
9. Click **OK**.

Figure 10.6 Residual by Dispersion Effect



In this plot it is easy to see the variance go up as the Hold Time increases. This is done by treating Hold Time as a nominal factor.

Profiling the Fitted Model

Use the **Profiler**, **Contour Profiler**, or **Surface Profiler** to gain further insight into the fitted model. To select a profiler option, click on the red triangle menu next to Loglinear Variance Fit and select one of the options under the **Profilers** menu.

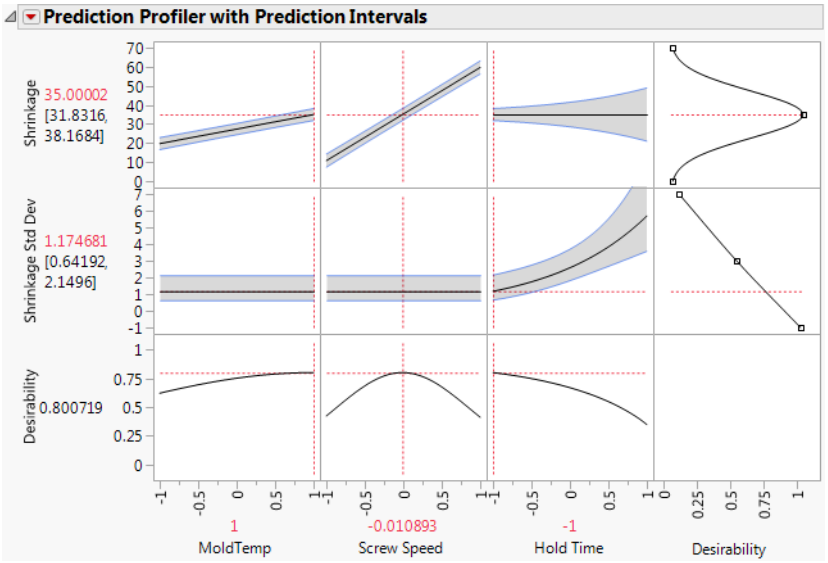
Example of Profiling the Fitted Model

For example, suppose that the goal was to find the factor settings that achieved a target of 35 for the response, but at the smallest variance. Fit the models and choose Profiler from the report menu. For example, Figure 10.7 shows the Profiler set up to match a target value for a mean and to minimize variance.

1. Select **Help > Sample Data Library** and open InjectionMolding.jmp.
2. Select **Analyze > Fit Model**.

   Since the variables in the data table have preselected roles assigned to them, the launch window is already filled out.
3. Click **Run**.
4. From the red triangle menu next to Loglinear Variance Fit, select **Profilers > Profiler**.
5. From the red triangle menu next to Prediction Profiler, select **Optimization and Desirability > Set Desirabilities**.
6. In the Response Goal window that appears, change Maximize to Match Target.
7. Click **OK**.
8. In the second Response Goal window, click **OK**.
9. From the red triangle menu next to Prediction Profiler select **Optimization and Desirability > Maximize Desirability**.
10. From the red triangle menu next to Prediction Profiler, select **Prediction Intervals**.

Figure 10.7 Profiler to Match Target and Minimize Variance with Prediction Intervals



One of the best ways to see the relationship between the mean and the variance (both modeled with the LogVariance personality) is through looking at the individual prediction confidence intervals about the mean. Regular confidence intervals (those shown by default in the Profiler) do not show information about the variance model as well as individual prediction confidence intervals do. Prediction intervals show both the mean and variance model in one graph.

If Y is the modeled response, and you want a prediction interval for a new observation at x_n, then:

$$s^2\left|x_n = s^2_Y\right|x_n + s^2_{\hat{Y}}\left|x_n\right.$$

where:

$s^2|x_n$ is the variance for the individual prediction at x_n

$s^2_Y|x_n$ is the variance of the distribution of Y at x_n

$s^2_{\hat{Y}}\left|x_n\right.$ is the variance of the sampling distribution of $\hat{Y}$, and is also the variance for the mean.

Because the variance of the individual prediction contains the variance of the distribution of Y, the effects of the changing variance for Y can be seen. Not only are the individual prediction intervals wider, but they can change shape with a change in the variance effects.

Chapter 11

Logistic Regression Models

Fit Regression Models for Nominal or Ordinal Responses

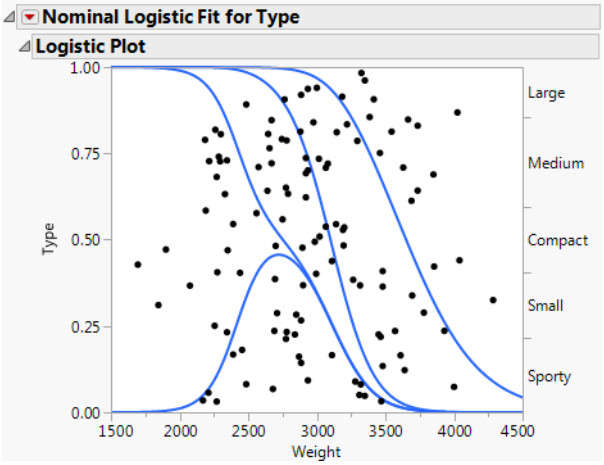
When your response variable has discrete values, you can use the Fit Model platform to fit a logistic regression model. The Fit Model platform provides two personalities for fitting logistic regression models. The personality that you use depends on the modeling type (Nominal or Ordinal) of your response column.

For nominal response variables, the Nominal Logistic personality fits a linear model to a multi-level logistic response function.

For ordinal response variables, the Ordinal Logistic personality fits the cumulative response probabilities to the logistic distribution function of a linear model.

Both personalities provide likelihood ratio tests for the model, a confusion matrix, and ROC and lift curves. When the response is binary, the Nominal Logistic personality also provides odds ratios (with corresponding confidence intervals).

Figure 11.1 Logistic Plot for a Nominal Logistic Regression Model



Contents

Overview of the Nominal and Ordinal Logistic Personalities

Logistic regression models the probabilities of the levels of a categorical *Y* response variable as a function of one or more *X* effects. The Fit Model platform provides two personalities for fitting logistic regression models. The personality that you use depends on the modeling type (Nominal or Ordinal) of your response column.

For more information about fitting logistic regression models, see Walker and Duncan (1967), Nelson (1976), Harrell (1986), and McCullagh and Nelder (1989).

For more information about the parameterization of the logistic regression model, see "Logistic Regression Model" on page 494.

Nominal Logistic Regression

When the response variable has a nominal modeling type, the platform fits a linear model to a multi-level logistic response function using maximum likelihood. Therefore, all but one response level is modeled by a logistic curve that represents the probability of the response level given the value of the *X* effects. The probability of the final response level is 1 minus the sum of the other fitted probabilities. As a result, at all values of the *X* effects, the fitted probabilities for the response levels sum to 1.

If the response variable is binary, you can set the Target Level in the Fit Model window to specify the level whose probability you want to model. By default, the model estimates the probability of the first level of the response variable.

For more information about fitting models for nominal response variables, see "Nominal Responses" on page 524 in the "Statistical Details" chapter.

Ordinal Logistic Regression

When the response variable has an ordinal modeling type, the platform fits the cumulative response probabilities to the logistic function of a linear model using maximum likelihood. Therefore, the cumulative probability of being at or below each response level is modeled by a curve. The curves are the same for each level except that they are shifted to the right or left.

Tip: If there are many response levels, the ordinal model is much faster to fit and uses less memory than the nominal model.

For more information about fitting models with ordinal response variables, see "Ordinal Responses" on page 525 in the "Statistical Details" chapter.

Other JMP Platforms That Fit Logistic Regression Models

There are many other places in JMP where you can fit logistic regression models:

- To fit logistic regression models with a single continuous main effect, you can use the Fit Y by X platform to see a cumulative logistic probability plot for each effect. See the Logistic Analysis chapter in *Basic Analysis*.
- To perform variable selection in logistic regression models, you can use the Stepwise personality of the Fit Model platform. See the "Stepwise Regression Models" chapter on page 249.
- To fit logistic regression models that use a link function other than the Logit link, you can use the Generalized Linear Model personality of the Fit Model platform. See the "Generalized Linear Models" chapter on page 499.
- JMP PRO To perform variable selection in logistic regression models and fit penalized logistic regression models, you can use the Generalized Regression personality of the Fit Model platform. See the "Generalized Regression Models" chapter on page 285.

Examples of Logistic Regression

This section contains two examples, one for each of the logistic regression personalities in the Fit Model platform (Nominal and Ordinal):

Example of Nominal Logistic Regression

An experiment was performed on metal ingots that were prepared with different heating and soaking times and then tested for readiness to roll. See Cox and Snell (1989). The data are contained in the Ingots.jmp sample data table. In this example, the Fit Model platform fits the probability of the Ready response using a logistic regression model with regressors heat and soak.

1. Select **Help > Sample Data Library** and open Ingots.jmp.

   The values of the categorical variable ready, Ready and Not Ready, indicate whether an ingot is ready to roll.
2. Select **Analyze > Fit Model**.
3. Select ready and click **Y**.

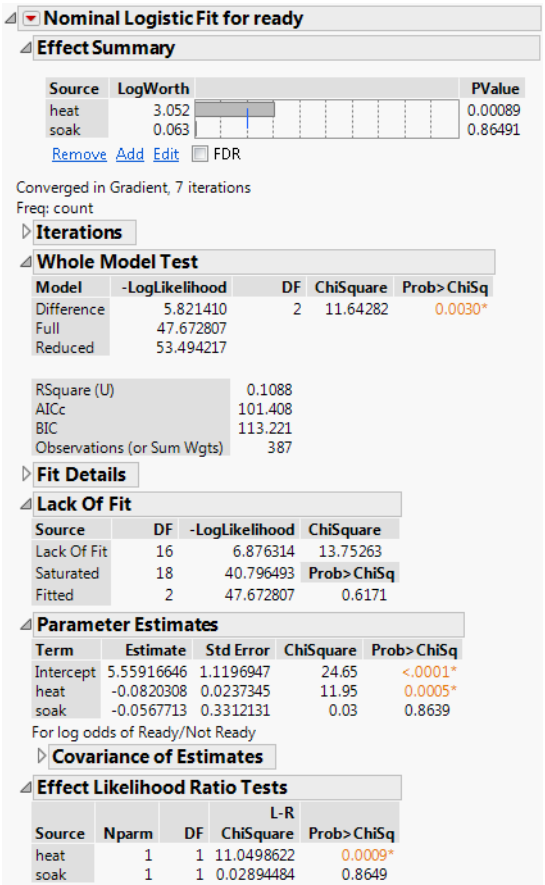
   Because you selected a column with the Nominal modeling type, the Fit Model Personality updates to Nominal Logistic.

Because ready is a Nominal column with only two levels, the Target Level option appears. This option enables you to specify the response level whose probability you want to model. In this model, the Target Level is Ready, so you are modeling the probability of the Ready response.

4. Select heat and soak and click **Add**.
5. Select count and click **Freq**.
6. Click **Run**.

   When the fitting process converges, the nominal regression report appears.

Figure 11.2 Nominal Logistic Fit Report



Nominal Logistic Fit for ready

Effect Summary

Source	LogWorth	PValue
heat	3.052	0.00089
soak	0.063	0.86491

Remove Add Edit FDR

Converged in Gradient, 7 iterations
Freq: count

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	5.821410	2	11.64282	0.0030*
Full	47.672807			
Reduced	53.494217			

RSquare (U)	0.1088
AICc	101.408
BIC	113.221
Observations (or Sum Wgts)	387

Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	16	6.876314	13.75263
Saturated	18	40.796493	Prob>ChiSq
Fitted	2	47.672807	0.6171

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	5.55916646	1.1196947	24.65	<.0001*
heat	-0.0820308	0.0237345	11.95	0.0005*
soak	-0.0567713	0.3312131	0.03	0.8639

For log odds of Ready/Not Ready

Covariance of Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
heat	1	1	11.0498622	0.0009*
soak	1	1	0.02894484	0.8649

In the Whole Model Test report, the chi-square statistic (11.64) has a small *p*-value (0.0030), which indicates that the overall model is significant. However, the parameter estimate for soak has a *p*-value of 0.8639, which indicates that soaking time might not be important.

At this point, you might also be interested in using inverse prediction to find the heating time at a specific soaking time and given a particular probability of readiness to roll. See "Example of Inverse Prediction" on page 487 for a continuation of this example.

Example of Ordinal Logistic Regression

An experiment was conducted to test whether various cheese additives (A to D) had an effect on cheese taste. Taste was measured by a tasting panel and recorded on an ordinal scale from 1 (strong dislike) to 9 (excellent taste). See McCullagh and Nelder (1989). The data are in the Cheese.jmp sample data table.

1. Select **Help > Sample Data Library** and open Cheese.jmp.
2. Select **Analyze > Fit Model**.
3. Select Response and click **Y**.

   Because you selected a column with the Ordinal modeling type, the Fit Model Personality updates to Ordinal Logistic.
4. Select Cheese and click **Add**.
5. Select Count and click **Freq**.
6. Click **Run**.

Figure 11.3 Ordinal Logistic Fit Report

Ordinal Logistic Fit for Response

Freq: Count

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	74.22695	3	148.4539	<.0001*
Full	355.67395			
Reduced	429.90090			

RSquare (U)	0.1727
AICc	734.695
BIC	770.061
Observations (or Sum Wgts)	208

Fit Details

Lack Of Fit

Source	DF	-LogLikelihood	ChiSquare
Lack Of Fit	21	10.15410	20.30819
Saturated	24	345.51986	Prob>ChiSq
Fitted	3	355.67395	0.5018

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept[1]	-4.6051335	0.4267239	116.46	<.0001*
Intercept[2]	-3.5499488	0.3073868	133.37	<.0001*
Intercept[3]	-2.4503882	0.2398702	104.36	<.0001*
Intercept[4]	-1.381774	0.2001699	47.65	<.0001*
Intercept[5]	-0.0455233	0.1803299	0.06	0.8007
Intercept[6]	0.90648953	0.1886609	23.09	<.0001*
Intercept[7]	2.408151	0.2380517	102.34	<.0001*
Intercept[8]	3.96800057	0.3466551	131.02	<.0001*
Cheese[A]	-0.8622328	0.2289236	14.19	0.0002*
Cheese[B]	2.48960592	0.2703151	84.82	<.0001*
Cheese[C]	0.8476542	0.2280359	13.82	0.0002*

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Cheese	3	3	148.453899	<.0001*

The model fit in this example reduces the –LogLikelihood of 429.9 for the intercept-only model to 355.67 for the full model. This reduction yields a likelihood ratio chi-square statistic for the whole model of 148.45 with 3 degrees of freedom. Therefore, the difference in perceived cheese taste is highly significant.

The most preferred cheese additive is the one with the most negative parameter estimate. Cheese[D] does not appear in the Parameter Estimates report, because it does not have its own column of the design matrix. However, Cheese D's effect can be computed as the negative sum of the others, and is shown in Table 11.1.

Table 11.1 Preferences for Cheese Additives in Cheese.jmp

Cheese	Estimate	Preference
A	–0.8622	2nd place
B	2.4896	least liked
C	0.8477	3rd place

Table 11.1 Preferences for Cheese Additives in Cheese.jmp *(Continued)*

D	–2.4750	most liked

Comparison to Nominal Logistic Model

The Lack of Fit report shows a test of whether the model fits the data well.

As an ordinal problem, each of the first eight response levels has an intercept, but there are only three parameters for the four levels of Cheese. As a result, there are 3 degrees of freedom in the ordinal model. The ordinal model is the Fitted model in the Lack of Fit test.

There are 28 rows with a nonzero value of Count in the data table, so there are 28 - 4 = 24 replicated points with respect to the levels of Cheese. Therefore, the Saturated model in the Lack of Fit test has 24 degrees of freedom.

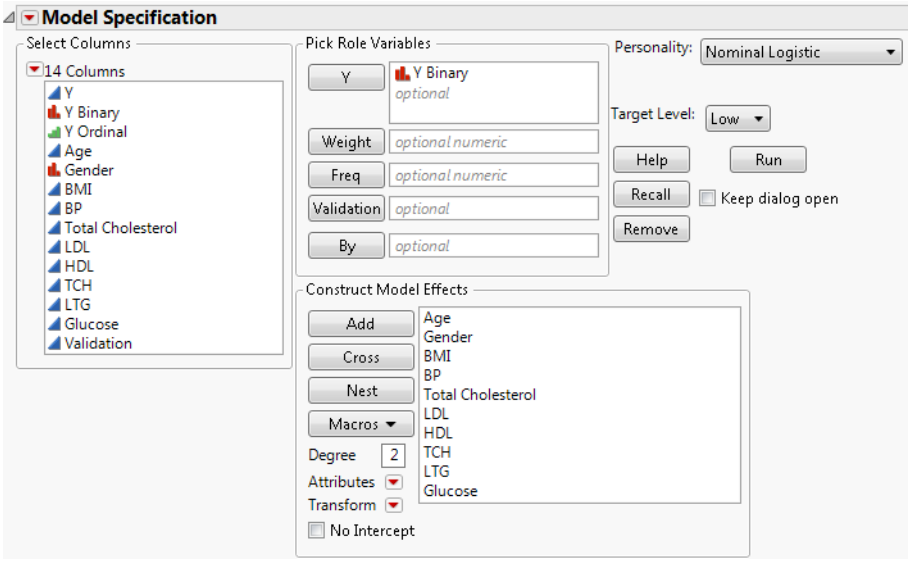
As a nominal problem, each of the first eight response levels has an intercept as well as three parameters for the four levels of Cheese. As a result, there are $8 \times 3 = 24$ $8 \times 3 = 24$ degrees of freedom in the nominal model. Therefore, the nominal model is the Saturated model in the Lack of Fit test.

In this example, the Lack of Fit test for the ordinal model happens to be testing the ordinal response model against the nominal model. The nonsignificance of Lack of Fit leads one to believe that the ordinal model is reasonable.

Launch the Nominal and Ordinal Logistic Personalities

Launch the Nominal Logistic and Ordinal Logistic personalities by selecting **Analyze > Fit Model** and entering one or more non-continuous columns for **Y**. If multiple columns are entered for **Y**, a model for each response is fit.

Figure 11.4 Fit Model Launch Window with Nominal Logistic Selected



For more information about aspects of the Fit Model window that are common to all personalities, see the"Model Specification" chapter on page 33. Information specific to the Nominal Logistic and Ordinal Logistic personalities is presented here.

If your model effects have missing values, you can treat these missing values as informative categories. Select the Informative Missing option from the Model Specification red triangle menu.

To specify a model without an intercept term, select the No Intercept option in the Construct Model Effects panel of the Fit Model window. The No Intercept option is not available for the Ordinal Logistic personality.

The event of interest in the logistic regression model is defined by the Target Level option in the Fit Model window. This option is available only when you specify a binary response column in the Nominal Logistic personality.

Note: The Logistic personalities in the Fit Model platform require that your data be in a stacked format such that all of the responses are in one column. Sometimes, your data are formatted in multiple columns. See "Example of Stacking Counts in Multiple Columns" on page 493 for an example of converting responses in multiple columns into a single column response.

JMP PRO Validation

Validation is the process of using part of a data set to estimate model parameters, and using the other part to assess the predictive ability of the model.

- The *training* set is used to estimate model parameters.
- The *validation* set is used to assess or validate the predictive ability of the model.
- The *test* set is a final, independent assessment of the model's predictive ability. The test set is available only when using a validation column.

The training, validation, and test sets are created as subsets of the original data. This is done through the use of a validation column in the Fit Model launch window.

The validation column's values determine how the data is split, and what method is used for validation:

- If the column has two distinct values, then training and validation sets are created.
- If the column has three distinct values, then training, validation, and test sets are created.
- If the column has more than three distinct values, or only one, then no validation is performed.

When a validation column is used, model fit statistics are given for the training, validation, and test sets in the Fit Details report. There is also a separate ROC curve, lift curve, and confusion matrix for each of the Training, Validation, and Test sets.

The Logistic Fit Report

When you fit a model using the Nominal Logistic or Ordinal Logistic personality, you obtain a Nominal Logistic Fit report or an Ordinal Logistic Fit report, respectively. By default, these reports contain the following reports:

Effect Summary An interactive report that enables you to add or remove effects from the model. See "Effect Summary Report" on page 182 in the "Standard Least Squares Report and Options" chapter.

Logistic Plot (Available only if the model consists of a single continuous effect.) The logistic probability plot provides a picture of what the logistic model is fitting. At each x value, the probability scale in the y direction is divided up (partitioned) into probabilities for each response category. The probabilities are measured as the vertical distance between the curves, with the total across all Y category probabilities summing to 1.

Iterations (Available only in the Nominal Logistic personality.) After launching Fit Model, an iterative estimation process begins and is reported iteration by iteration. After the

fitting process completes, you can open the Iteration History report and see the iteration steps.

Whole Model Test Shows tests that compare the whole-model fit to the model that omits all the regressor effects except the intercept parameters. The test is analogous to the Analysis of Variance table for continuous responses. For more information about the Whole Model Test report, see "Whole Model Test" on page 479.

Fit Details Shows various measures of fit for the model. See "Fit Details" on page 480.

Lack of Fit (Available only when there are replicated points with respect to the *X* effects and the model is not saturated.) Shows a lack of fit test, also called a goodness of fit test, that addresses whether more terms are needed in the model. See "Lack of Fit Test" on page 481.

Parameter Estimates Shows the parameter estimates, standard errors, and associated hypothesis tests. The Covariance of Estimates report gives the variances and covariances of the parameter estimates.

Note: The Covariance of Estimates report appears only for nominal response variables, not for ordinal response variables.

Singularity Details (Appears only when there are linear dependencies among the model terms.) Shows a report that contains the linear functions that the model terms satisfy.

Effect Likelihood Ratio Tests The likelihood ratio chi-square tests are calculated as twice the difference of the log-likelihoods between the full model and the model constrained by the hypothesis to be tested. The constrained model is the model that does not contain the effect. These tests can take time to do because each test requires a separate set of iterations.

Note: Likelihood ratio tests are the platform default if they are projected to take less than 20 seconds to complete. Otherwise, the default effect tests are Wald tests.

Whole Model Test

The Whole Model Test table shows tests that compare the whole-model fit to the model that omits all the regression parameters except the intercept parameters. The test is analogous to the Analysis of Variance table for continuous responses. The negative log-likelihood corresponds to the sums of squares, and the chi-square test corresponds to the *F* test.

The Whole Model Test table shows these quantities:

Model Lists the model labels.

Difference The difference between the Full model and the Reduced model. This model is used to measure the significance of the regressors as a whole to the fit.

Full The complete model that includes the intercepts and all effects.

Reduced The model that includes only the intercept parameters.

–LogLikelihood The negative log-likelihood for the respective models.

DF The degrees of freedom (DF) for the Difference between the Full and Reduced model.

Chi-Square The likelihood ratio chi-square test statistic for the hypothesis that all regression parameters are zero. The test statistic is computed by taking twice the difference in negative log-likelihoods between the fitted model and the reduced model that has only intercepts.

Prob>ChiSq The probability of obtaining a greater chi-square value by chance alone if the specified model fits no better than the model that includes only intercepts.

RSquare (U) The proportion of the total uncertainty that is attributed to the model fit, defined as the ratio of the Difference to the Reduced negative log-likelihood values. RSquare ranges from zero for no improvement in fit to 1 for a perfect fit. An RSquare (U) value of 1 indicates that the predicted probabilities for events that occur are equal to one: There is no uncertainty in predicted probabilities. Because certainty in the predicted probabilities is rare for logistic models, RSquare (U) tends to be small.

RSquare (U) is sometimes referred to as *U*, the uncertainty coefficient, or as *McFadden's pseudo* R^2.

AICc The corrected Akaike Information Criterion. For more information, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

BIC Bayesian Information Criterion. For more information, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

Observations (or Sum Weights) Total number of observations in the sample. If a Freq or Weight column is specified in the Fit Model window, this value is the sum of the values of a column assigned to the Freq or Weight role.

Fit Details

The Fit Details report contains the following statistics:

Entropy RSquare Equivalent to RSquare (U). See "Whole Model Test" on page 479.

Generalized RSquare A measure that can be applied to general regression models. It is based on the likelihood function L and is scaled to have a maximum value of 1. The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the Nagelkerke or Craig and Uhler R^2, which is a normalized version of Cox and Snell's pseudo R^2. See Nagelkerke (1991).

Mean -Log p The average of -log(p), where p is the fitted probability associated with the event that occurred.

RMSE The root mean square error, where the differences are between the response and p (the fitted probability for the event that actually occurred).

Mean Abs Dev The average of the absolute values of the differences between the response and p (the fitted probability for the event that actually occurred).

Misclassification Rate The rate for which the response category with the highest fitted probability is not the observed category.

For Entropy RSquare and Generalized RSquare, values closer to 1 indicate a better fit. For Mean -Log p, RMSE, Mean Abs Dev, and Misclassification Rate, smaller values indicate a better fit.

To test that the effects as a whole are significant (the Whole Model test), a chi-square statistic is computed by taking twice the difference in negative log-likelihoods between the fitted model and the reduced model that has only intercepts.

Lack of Fit Test

The Lack of Fit test addresses whether there is enough information in the current model or whether more complex terms are needed. This test is sometimes called a goodness-of-fit test. The lack of fit test calculates a pure-error negative log-likelihood by constructing categories for every combination of the model effect values in the data. The Saturated row in the Lack Of Fit table contains this log-likelihood. The Lack of Fit report also contains a test of whether the Saturated log-likelihood is significantly better than the Fitted model.

The Saturated degrees of freedom is m–1, where m is the number of unique populations. The Fitted degrees of freedom is the number of parameters not including the intercept.

The Lack of Fit table contains the negative log-likelihood for error due to Lack of Fit, error in a Saturated model (pure error), and the total error in the Fitted model. The chi-square statistic tests for lack of fit.

Logistic Fit Platform Options

Options for Nominal and Ordinal Fits

The following options are available in both the Nominal Logistic Fit and Ordinal Logistic Fit red triangle menus:

Logistic Plot (Available only if the model consists of a single continuous effect.) Shows or hides the Logistic Plot report. See "The Logistic Fit Report" on page 478.

Likelihood Ratio Tests Shows or hides the Effect Likelihood Ratio Tests report. The likelihood ratio chi-square tests are calculated as twice the difference of the log-likelihoods between the full model and the model constrained by the hypothesis to be tested. The constrained model is the model that does not contain the effect). These tests can take time to do because each test requires a separate set of iterations. Therefore, they could take a long time for large problems.

Note: Likelihood ratio tests are the platform default if they are projected to take less than 20 seconds to complete. This default option is highly recommended.

Wald Tests Shows or hides the Effect Wald Tests report. The Wald chi-square is a quadratic approximation to the likelihood ratio test, and it is a by-product of the calculations. Though Wald tests are considered less trustworthy, they do provide an adequate significance indicator for screening effects. Each parameter estimate and effect is shown with a Wald test. This is the default test if the likelihood ratio tests are projected to take more than 20 seconds to complete.

Confidence Intervals Shows or hides profile-likelihood confidence intervals for the model parameters. You can change the confidence level by selecting Set Alpha Level in the Model Specification red triangle menu in the Fit Model window. Each confidence limit requires a set of iterations in the model fit and can take a long time to compute. Furthermore, the effort does not always succeed in finding limits.

ROC Curve Shows or hides an ROC curve for the model. Receiver Operating Characteristic (ROC) curves measure the sorting efficiency of the model's fitted probabilities to sort the response levels. ROC curves can also aid in setting criterion points in diagnostic tests. The

higher the curve from the diagonal, the better the fit. An introduction to ROC curves is found in the Logistic Analysis chapter in the *Basic Analysis* book.

If the logistic fit has more than two response levels, it produces a generalized ROC curve (identical to the one in the Partition platform). In such a plot, there is a curve for each response level, which is the ROC curve of that level versus all other levels. See the Partition chapter in the *Predictive and Specialized Modeling* book.

If you specified a validation column, an ROC curve is shown for each of the Training, Validation, and Test sets.

Lift Curve Shows or hides a lift curve for the model. A lift curve shows the same information as an ROC curve, but in a way to dramatize the richness of the ordering at the beginning. The vertical axis shows the ratio of how rich that portion of the population is in the chosen response level compared to the rate of that response level as a whole. If you specified a validation column, a lift curve is shown for each of the Training, Validation, and Test sets. See the Partition chapter in the *Predictive and Specialized Modeling* book for more information about lift curves.

Confusion Matrix Shows or hides a confusion matrix, which is a two-way classification of the actual response levels and the predicted response levels. The predicted response level is the level with the highest response probability. For a good model, predicted response levels should be the same as the actual response levels. The confusion matrix provides a way to assess how the predicted responses align with the actual responses. If you specified a validation column, a confusion matrix is shown for each of the Training, Validation, and Test sets.

If the response is nominal and has a Profit Matrix column property, a Decision Matrix report also appears when this option is selected. For more information about the Decision Matrix report, see the Partition chapter in the *Predictive and Specialized Modeling* book.

Profiler Shows or hides the prediction profiler, showing the fitted values for a specified response probability as the values of the factors in the model are changed. This feature is available for both nominal and ordinal responses. For detailed information about profiling features, refer to the Profiler chapter in the *Profilers* book.

Model Dialog Shows the completed launch window for the current analysis.

Effect Summary Shows or hides the Effect Summary report, which enables you to interactively update the effects in the model. See "The Logistic Fit Report" on page 478.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Local Data Filter Shows or hides the local data filter that enables you to filter the data used in a specific report.

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Options for Nominal Fits

The following options are available in the Nominal Logistic Fit red triangle menu:

Plot Options (Available only if the model consists of a single continuous effect.) The Plot Options menu contains the following options:

Show Points Toggles the points in the Logistic Plot on or off.

Show Rate Curve Enables you to compare the rate at various values of the effect variable with the fitted logistic curve. This option is useful only if you have several points for each value of the effect. In these cases, you get reasonable estimates of the rate at each value, and compare this rate with the fitted logistic curve. To prevent too many degenerate points, usually at zero or one, JMP shows only the rate value if there are at least three points at the *x*-value.

Line Color Specifies the color of the plot curves.

Odds Ratios Shows or hides an Odds Ratios report that contains Unit Odds Ratios and Range Odds Ratios, as shown in Figure 11.5.

Figure 11.5 Odds Ratios

Odds Ratios

For ready odds of Ready versus Not Ready

Unit Odds Ratios

Per unit change in regressor

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
heat	0.921244	0.87937	0.965111	1.0854892
soak	0.94481	0.493646	1.808313	1.0584137

Range Odds Ratios

Per change in regressor over entire range

Term	Odds Ratio	Lower 95%	Upper 95%	Reciprocal
heat	0.027069	0.003496	0.209605	36.942229
soak	0.8434	0.120295	5.913179	1.185677

Tests and confidence intervals on odds ratios are Wald based.

Inverse Prediction (Available only for two-level nominal responses.) Finds the *x* value that results in a specified probability. See the appendix "Statistical Details" on page 521 for more information about inverse prediction.

Save Probability Formula Creates columns in the current data table that contain formulas for linear combinations of the response levels, prediction formulas for the response levels, and a prediction formula giving the most likely response.

For a nominal response model with *r* levels, JMP creates the following columns:

- columns called Lin[j] that contain a linear combination of the regressors for response levels $j = 1, 2, \ldots r - 1$
- a column called Prob[r], with a formula for the fit to the last level, *r*
- columns called Prob[j] for $j < r$ with a formula for the fit to level *j*
- a column called Most Likely response name that selects the most likely level of each row based on the computed probabilities.

JMP PRO **Publish Probability Formulas** Creates probability formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

Indicator Parameterization Estimates (Available only when there are nominal columns among the model effects.) Shows or hides the Indicator Function Parameterization report, which gives parameter estimates for the model where nominal columns are coded using indicator (SAS GLM) parameterization and are treated as continuous. Ordinal columns remain coded using the usual JMP coding scheme. The SAS GLM and JMP coding schemes are described in "The Factor Models" on page 526 in the "Statistical Details" appendix.

Caution: Standard errors and chi-square values given in the Indicator Function Parameterization report differ from those in the Parameter Estimates report. This is because the estimates are estimating different model parameters.

Options for Ordinal Fits

The following options are available in the Ordinal Logistic Fit red triangle menu:

Save Contains the following Save options:

Save Probability Formula Creates columns in the current data table that contain formulas for linear combinations of the response levels, prediction formulas for the response levels, and a prediction formula giving the most likely response.

For an ordinal response model with *r* levels, JMP creates the following columns:

- a column called Linear that contains the formula for a linear combination of the regressors without an intercept term
- columns called Cum[j] that each contain a formula for the cumulative probability that the response is less than or equal to level *j*, for levels *j* = 1, 2, ... *r* - 1. There is no Cum[j = 1, 2, ... r - 1] that is 1 for all rows
- columns called Prob[j = 1, 2, ... r - 1], for $1 < j < r$ that each contain the formula for the probability that the response is level *j*. Prob[j] is the difference between Cum[*j*] and Cum[j –1]. Prob[1] is Cum[1], and Prob[r] is 1–Cum[r –1].
- a column called Most Likely response name that selects the most likely level of each row based on the computed probabilities.

JMP PRO **Publish Probability Formulas** Creates probability formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See the Formula Depot chapter in the *Predictive and Specialized Modeling* book.

Save Quantiles (Available only when the response is numeric and has the ordinal modeling type.) Creates columns in the current data table named OrdQ.05, OrdQ.50, and OrdQ.95 that fit the quantiles for these three probabilities.

Save Expected Value (Available only when the response is numeric and has the ordinal modeling type.) Creates a column in the current data table called Ord Expected. This column contains the linear combination of the response values with the fitted response probabilities for each row and gives the expected value.

Additional Examples of Logistic Regression

Example of Inverse Prediction

Inverse prediction enables you to predict the value for one independent variable at a specified probability of the response. If there is more than one independent variable, you must specify values for all of them except the one you are predicting.

An experiment was performed on metal ingots. The ingots were prepared with different heating and soaking times and then tested for readiness to roll. See Cox and Snell (1989). In JMP, the data are contained in the Ingots.jmp sample data table.

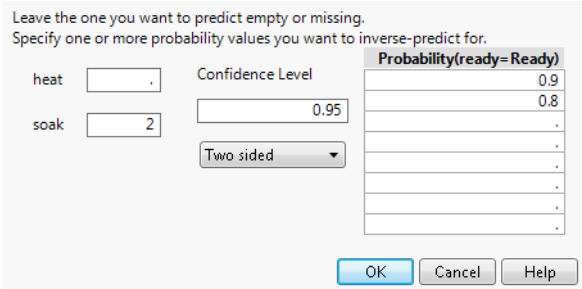
You are interested in predicting the heating time at a soaking time of 2 for probabilities of readiness to roll of 0.8 and 0.9.

1. For the analysis, follow the steps in "Example of Nominal Logistic Regression" on page 472.
2. From the red triangle next to Nominal Logistic Fit for ready, select **Inverse Prediction**.
3. Delete the value for heat.

   You want to find the predicted value of heat, so you leave the heat value empty.
4. Enter 2 for soak.

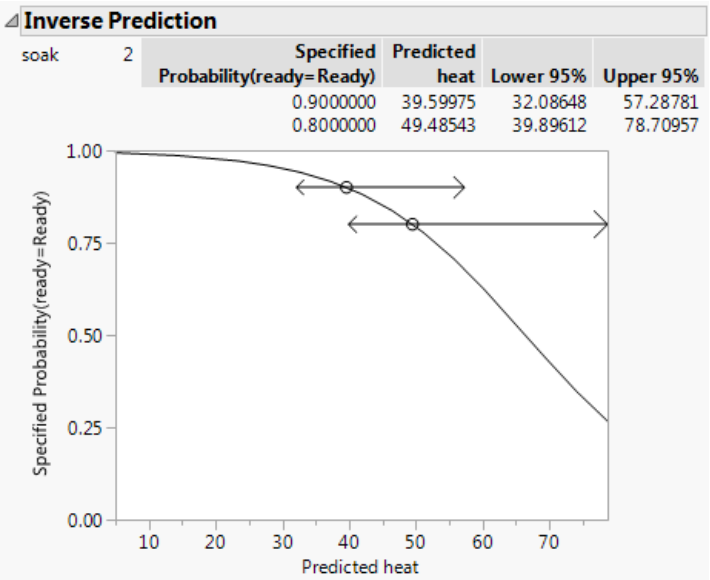
   You want to predict heating time when soak is 2.
5. Under Probability, enter 0.9 and 0.8 in the first two rows.

Figure 11.6 The Inverse Prediction Specification Window



6. Click **OK**.

Figure 11.7 Inverse Prediction Report



When soak is 2, the predicted value of heat for which there is a 90% chance of an ingot being ready to roll is 39.60 with a 95% confidence interval from 32.09 to 57.29. When soak is 2, the predicted value for heat for which there is an 80% chance of an ingot being ready to roll is 49.49 with a 95% confidence interval from 39.90 to 78.71.

Example of Using Effect Summary for a Nominal Logistic Model

A market research study was undertaken to evaluate preference for a brand of detergent. See Ries and Smith (1963). The results of the study are in the Detergent.jmp sample data table. The model is defined by the following:

- the response variable, brand, with values m and x
- an effect called softness (water softness) with values soft, medium, and hard
- an effect called previous use with values yes and no
- an effect called temperature with values high and low
- a count variable, count, which gives the frequency counts for each combination of effect categories

The study begins by specifying the full three-factor factorial model, as follows:

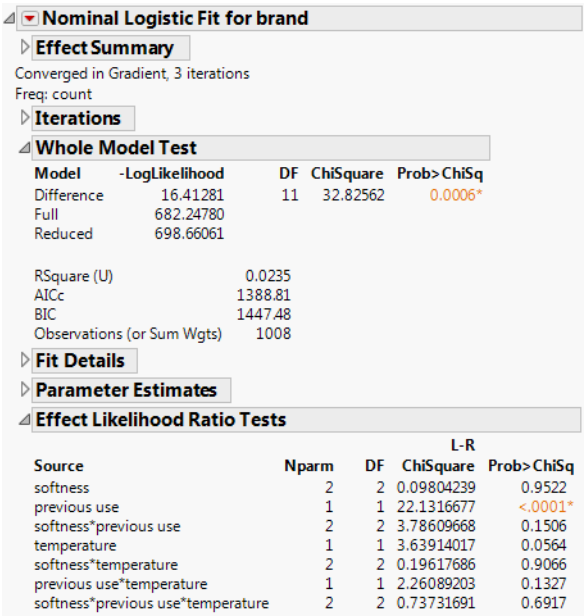
1. Select **Help > Sample Data Library** and open Detergent.jmp.
2. Select **Analyze > Fit Model**.
3. Select brand from the Select Columns list and click **Y**.

Because you specified a nominal response variable, the Personality changes to Nominal Logistic.

Because brand is a Nominal column with only two levels, the Target Level option appears. This option enables you to specify the response level whose probability you want to model.

4. Select count and click **Freq**.
5. Select softness through temperature and click **Macros > Full Factorial**.
6. Click **Run**.

Figure 11.8 Nominal Logistic Fit for Three-Factor Factorial Model



Nominal Logistic Fit for brand

Effect Summary

Converged in Gradient, 3 iterations
Freq: count

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	16.41281	11	32.82562	0.0006*
Full	682.24780			
Reduced	698.66061			

RSquare (U)	0.0235
AICc	1388.81
BIC	1447.48
Observations (or Sum Wgts)	1008

Fit Details

Parameter Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
softness	2	2	0.09804239	0.9522
previous use	1	1	22.1316677	<.0001*
softness*previous use	2	2	3.78609668	0.1506
temperature	1	1	3.63914017	0.0564
softness*temperature	2	2	0.19617686	0.9066
previous use*temperature	1	1	2.26089203	0.1327
softness*previous use*temperature	2	2	0.73731691	0.6917

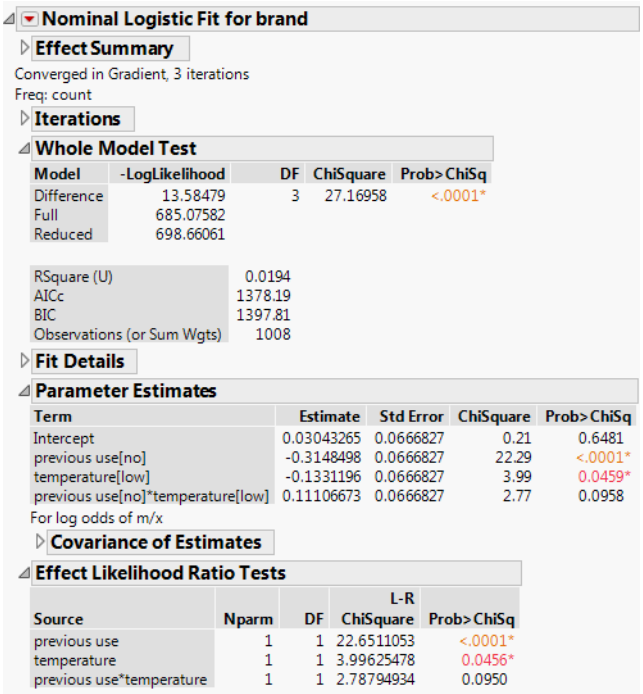
The Whole Model Test report shows that the three-factor full factorial model as a whole is significant (Prob>ChiSq = 0.0006).

The Effect Likelihood Ratio Tests report shows that the effects that include softness do not contribute significantly to the model fit. This leads you to consider removing softness from the model. You can do this from the Effect Summary report.

7. In the Effect Summary report, select softness*previous use through softness under Source and click **Remove**.

The report updates to show the two-factor factorial model (Figure 11.9). The Whole Model Test report shows that the two-factor model is also significant as a whole.

Figure 11.9 Nominal Logistic Fit for Two-Factor Factorial Model



Nominal Logistic Fit for brand

Effect Summary

Converged in Gradient, 3 iterations
Freq: count

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	13.58479	3	27.16958	<.0001*
Full	685.07582			
Reduced	698.66061			

RSquare (U)	0.0194
AICc	1378.19
BIC	1397.81
Observations (or Sum Wgts)	1008

Fit Details

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	0.03043265	0.0666827	0.21	0.6481
previous use[no]	-0.3148498	0.0666827	22.29	<.0001*
temperature[low]	-0.1331196	0.0666827	3.99	0.0459*
previous use[no]*temperature[low]	0.11106673	0.0666827	2.77	0.0958

For log odds of m/x

Covariance of Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
previous use	1	1	22.6511053	<.0001*
temperature	1	1	3.99625478	0.0456*
previous use*temperature	1	1	2.78794934	0.0950

From the report shown in Figure 11.9, you conclude that previous use of a detergent brand and water temperature have an effect on detergent preference. You also note that the interaction between temperature and previous use is not statistically significant, so you conclude that the effect of temperature does not depend on previous use.

Example of a Quadratic Ordinal Logistic Model

The Ordinal Logistic Personality can fit a quadratic surface to optimize the probabilities of higher or lower levels of an ordinal response.

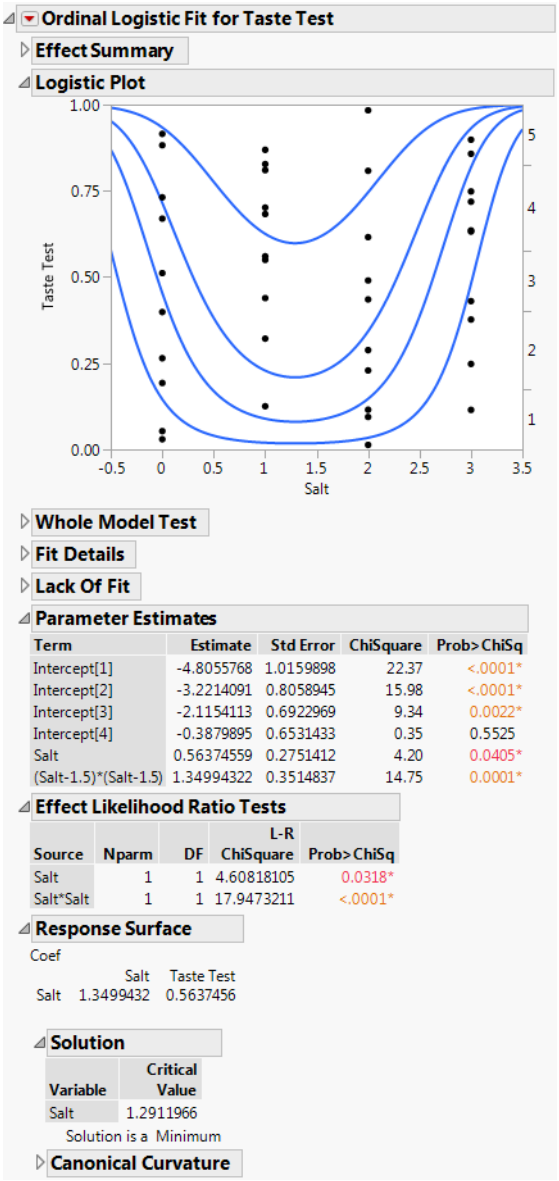
In this example, a microwave popcorn manufacturer wants to find out how much salt consumers like in their popcorn. To do this, the manufacturer looks for the maximum probability of a favorable response as a function of how much salt is added to the popcorn package. In this experiment, the salt amounts are controlled at 0, 1, 2, and 3 teaspoons. Respondents rate the taste on a scale of 1 (low) to 5 (high). The optimal amount of salt is the amount that maximizes the probability of more favorable responses. The ten observations for each of the salt levels are shown in Table 11.2.

Table 11.2 Salt in Popcorn

Salt Amount	Salt Rating Responses
no salt	1, 3, 2, 4, 2, 2, 1, 4, 3, 4
1 tsp.	4, 5, 3, 4, 5, 4, 5, 5, 4, 5
2 tsp.	4, 3, 5, 1, 4, 2, 5, 4, 3, 2
3 tsp.	3, 1, 2, 3, 1, 2, 1, 2, 1, 2

1. Select **Help > Sample Data Library** and open Salt in Popcorn.jmp.
2. Select **Analyze > Fit Model**.
3. Select Taste Test from the Select Columns list and click **Y**.
4. Select Salt from the Select Columns list and click **Macros > Response Surface**.
5. Click **Run**.
6. Click the disclosure icon next to Response Surface to open the report.

Figure 11.10 Ordinal Logistic Fit for Salt in Popcorn.jmp



The report shows how the quadratic model fits the response probabilities. The curves, instead of being shifted logistic curves, become stacked U-shaped curves where each curve achieves its minimum at the same point. The critical value is at Mean(X)–0.5 * (b1/b2) where b1 is the linear coefficient and b2 is the quadratic coefficient. This formula is for centered X. From the Parameter Estimates table, you can compute that the optimal amount of salt is 1.5 - 0.5 * (0.5637/1.3499) = 1.29 teaspoons.

The vertical distance at a specific amount of salt between each curve measures the probability of each of the five response levels for the specific amount of salt. The probability for the highest response level is the distance from the top curve to the top of the plot rectangle. This distance reaches a maximum when the amount of salt is about 1.3 teaspoons. All curves share the same critical point.

The parameter estimates for Salt and Salt*Salt are the coefficients used to find the critical value. Although the critical value appears on the logistic plot as a minimum, it is a maximum in the sense of maximizing the probability of the highest response. The Solution portion of the report is shown under Response Surface in Figure 11.10, where 1.29 is shown under Critical Value.

Example of Stacking Counts in Multiple Columns

When data that are frequencies (counts) are listed in several columns of your data table, you must transform the data into the form that you need for logistic regression. For example, the Ingots2.jmp data table (see Figure 11.11) has columns Nready and Nnotready. These columns give the number of ingots that are ready to roll and ingots that are not ready to roll for each combination of Heat and Soak values.

Figure 11.11 Ingots2.jmp Sample Data Table

	Heat	Soak	Nready	Nnotready	Ntotal	P	Loss
1	7	1	10	0	10	0.5000	6.9315
2	7	1.7	17	0	17	0.5000	11.7835
3	7	2.2	7	0	7	0.5000	4.8520
4	7	2.8	12	0	12	0.5000	8.3178
5	7	4	9	0	9	0.5000	6.2383
6	14	1	31	0	31	0.5000	21.4876
7	14	1.7	43	0	43	0.5000	29.8053
8	14	2.2	31	2	33	0.5000	22.8739
9	14	2.8	31	0	31	0.5000	21.4876
10	14	4	19	0	19	0.5000	13.1698
11	27	1	55	1	56	0.5000	38.8162
12	27	1.7	40	4	44	0.5000	30.4985
13	27	2.2	21	0	21	0.5000	14.5561
14	27	2.8	21	1	22	0.5000	15.2492
15	27	4	15	1	16	0.5000	11.0904

Before fitting a logistic regression model, use the following steps to stack the Nready and Nnotready columns into a single column:

1. Select **Help > Sample Data Library** and open Ingots2.jmp.
2. Select **Tables > Stack**.
3. Select Nready and NNotReady from the Select Columns list and click **Stack Columns**.

4. Click **OK**.

   This creates the new table in Figure 11.12. Label is the response (*Y*) column and Data is the frequency column.

This stacked data table is equivalent to the Ingots.jmp sample data table used in "Example of Nominal Logistic Regression" on page 472.

Figure 11.12 Stacked Data Table

	Heat	Soak	Ntotal	P	Loss	Label	Data
1	7	1	10	0.5000	6.9315	Nready	10
2	7	1	10	0.5000	6.9315	Nnotready	0
3	7	1.7	17	0.5000	11.7835	Nready	17
4	7	1.7	17	0.5000	11.7835	Nnotready	0
5	7	2.2	7	0.5000	4.8520	Nready	7
6	7	2.2	7	0.5000	4.8520	Nnotready	0
7	7	2.8	12	0.5000	8.3178	Nready	12
8	7	2.8	12	0.5000	8.3178	Nnotready	0
9	7	4	9	0.5000	6.2383	Nready	9
10	7	4	9	0.5000	6.2383	Nnotready	0
11	14	1	31	0.5000	21.4876	Nready	31
12	14	1	31	0.5000	21.4876	Nnotready	0
13	14	1.7	43	0.5000	29.8053	Nready	43
14	14	1.7	43	0.5000	29.8053	Nnotready	0
15	14	2.2	33	0.5000	22.8739	Nready	31
16	14	2.2	33	0.5000	22.8739	Nnotready	2
17	14	2.8	31	0.5000	21.4876	Nready	31
18	14	2.8	31	0.5000	21.4876	Nnotready	0

Statistical Details for the Nominal and Ordinal Logistic Personalities

Logistic Regression Model

Logistic regression fits nominal *Y* responses to a linear model of *X* terms. To be more precise, it fits probabilities for the response levels using a logistic function. For two response levels, the function is:

$P(Y = r_1) = (1 + e^{-Xb})^{-1}$ where r_1 is the first response level

or equivalently:

$$\log\left(\frac{P(Y = r_1)}{P(Y = r_2)}\right) = Xb$$ where r_1 and r_2 are the two responses levels, respectively

Note: When Y is binary and has a nominal modeling type, you can set the Target Level in the Fit Model window to specify the level whose probability you want to model. In this section, the target level is designated as r_1.

For *r* nominal response levels, where $r > 2$, the model is defined by $r - 1$ linear model parameters of the following form:

$$\log\left(\frac{P(Y = j)}{P(Y = r)}\right) = Xb_j$$

The fitting principal of maximum likelihood means that the βs are chosen to maximize the joint probability attributed by the model to the responses that did occur. This fitting principal is equivalent to minimizing the negative log-likelihood (–LogLikelihood):

$$\text{Loss} = -\text{logLikelihood} = \sum_{i=1}^{n} -\log(\text{Prob}(i\text{th row has the } y_j\text{th response}))$$

Odds Ratios

For two response levels, the logistic regression model is as follows:

$$\log\left(\frac{\text{Prob}(Y = r_1)}{\text{Prob}(Y = r_2)}\right) = Xb$$ where r_1 and r_1 are the two response levels

Therefore, the *odds* are defined as follows:

$$\frac{\text{Prob}(Y = r_1)}{\text{Prob}(Y = r_2)} = \exp(X\beta) = \exp(\beta_0) \cdot \exp(\beta_1 X_1) \cdots \exp(\beta_i X_i)$$

Note that $\exp(\beta_i(X_i + 1)) = \exp(\beta_i X_i)\exp(\beta_i)$. This shows that if X_i changes by a unit amount, the odds is multiplied by $\exp(\beta_i)$, which we label the *unit odds ratio*. As X_i changes over its whole range, the odds are multiplied by $\exp((X_{\text{high}} - X_{\text{low}})\beta_i)$, which we label the *range odds ratio*. For binary responses, the log odds ratio for flipped response levels involves only changing the sign of the parameter. Therefore, you might want the reciprocal of the reported value to focus on the last response level instead of the first.

Two-level nominal effects are coded 1 and -1 for the first and second levels, so range odds ratios or their reciprocals would be of interest.

If no confidence intervals for the parameter estimates have been calculated when you select the Odds Ratio option, selecting the Odds Ratio option produces Wald-based confidence intervals for the odds ratio.

Selecting the Odds Ratio option produces profile likelihood-based confidence intervals for the odds ratios when all of the following conditions are met:

- the Likelihood Ratio Tests option has been selected
- the number of parameters for each response is less than 8
- the number of rows is less than 1000.

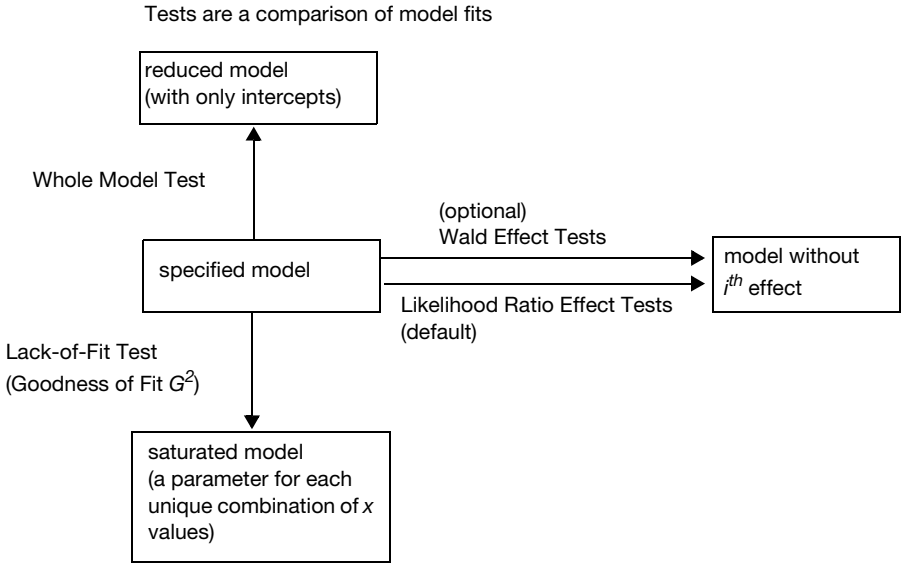
The method used for computing confidence intervals for the odds ratios is noted at the bottom of the Odds Ratios report.

Relationship of Statistical Tests

All of the statistical tests in the Logistic Regression reports compare the fit of the specified model with subset or superset models, as illustrated in Figure 11.13. If a test shows significance, then the higher order model is justified.

- Whole model tests: if the specified model is significantly better than a reduced model without any effects except the intercepts.
- Lack of Fit tests: if a saturated model is significantly better than the specified model.
- Effect tests: if the specified model is significantly better than a model without a given effect.

Figure 11.13 Relationship of Statistical Tests



Chapter 12

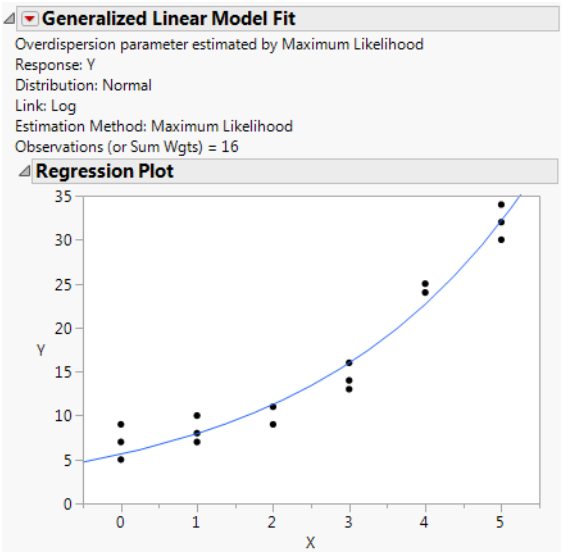
Generalized Linear Models

Fit Models for Nonnormal Response Distributions

Generalized linear models provide a unified way to fit responses that do not fit the usual requirements of traditional linear models. For example, frequency counts are often characterized as having a Poisson distribution and fit using a generalized linear model.

The Generalized Linear Model personality of the Fit Model platform enables you to fit generalized linear models for responses with binomial, normal, Poisson, or exponential distributions. The platform provides reports similar to those that are provided for traditional linear models. The platform also accommodates separation in logistic regression models using the Firth correction.

Figure 12.1 Example of a Generalized Linear Model Fit



Contents

Overview of the Generalized Linear Model Personality

Traditional linear models are used extensively in statistical data analysis. However, there are situations that violate the assumptions of traditional linear models. In these situations, traditional linear models are not appropriate. Traditional linear models assume that the responses are continuous and normally distributed with constant variance across all observations. These assumptions might not be reasonable. For example, these assumptions are not reasonable if you want to model counts, or if the variance of the observed responses increases as the response increases. Another example of violating the assumptions of traditional linear models is when the mean of the response is restricted to a specific range of values, such as proportions that fall between 0 and 1.

For situations such as these that fall into a wider range of data analysis problems, generalized linear models can be applied. Generalized linear models are an extension of traditional linear models. A generalized linear model consists of a linear component, a link function, and a variance function. The link function, $g(\mu_i) = x'_i\beta$, is a monotonic and differentiable function that describes how the expected value of Y_i is related to the linear predictors. An example of generalized linear regression is Poisson regression, where $\log(\mu_i)$ is the link function. For a complete list of the generalized linear regression models available using the Generalized Linear Models personality of the Fit Model platform, see "Statistical Details for the Generalized Linear Model Personality" on page 516.

Fitted generalized linear models can be summarized and evaluated using the same statistics as traditional linear models. The Fit Model platform provides parameter estimates, standard errors, goodness-of-fit statistics, confidence intervals, and hypothesis tests for generalized linear models. It should be noted that exact distribution theory is not always available or practical for generalized linear models. Therefore, some inference procedures are based on asymptotic results.

An important aspect of fitting generalized linear models is the selection of explanatory variables in the model. Changes in goodness-of-fit statistics are often used to evaluate the contribution of subsets of explanatory variables to a particular model. The *deviance* is defined as twice the difference between the maximum attainable value of the log-likelihood function and the value of the log-likelihood function at the maximum likelihood estimates of the regression parameters. The deviance is often used as a measure of goodness of fit. The maximum attainable log-likelihood is achieved with a model that has a parameter for every observation.

JMP PRO For variable selection and penalized methods in generalized linear modeling, you can use the Generalized Regression personality of the Fit Model platform in JMP Pro. See "Generalized Regression Models" chapter on page 285.

Example of a Generalized Linear Model

This example uses Poisson regression to model count data from a study of nesting horseshoe crabs. Each female crab had a male crab resident in her nest. The study investigated whether there were other males, called satellites, residing nearby. The data table CrabSatellites.jmp contains a response variable listing the number of male satellites, as well as variables that describe the color, spine condition, weight, and carapace width of the female crab. You are interested in the relationship between the number of satellites and the variables that describe the female crabs.

To fit the Poisson regression model, follow these steps:

1. Select **Help > Sample Data Library** and open CrabSatellites.jmp.
2. Select **Analyze > Fit Model**.
3. Select satell and click **Y**.
4. Select color, spine, width, and weight and click **Add**.
5. From the Personality list, select **Generalized Linear Model**.
6. From the Distribution list, select **Poisson**.

   In the Link Function list, **Log** should be selected for you automatically.
7. Click **Run**.

Figure 12.2 Crab Satellite Results

Generalized Linear Model Fit

Response: satell
Distribution: Poisson
Link: Log
Estimation Method: Maximum Likelihood
Observations (or Sum Wgts) = 173

Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	41.6030357	83.2061	7	<.0001*
Full	452.44163			
Reduced	494.044666			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>ChiSq
Pearson	533.8165	165	<.0001*
Deviance	549.5856	165	<.0001*

AICc
921.7613

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
color	3	9.2405455	0.0263*
spine	2	1.794722	0.4076
width	1	0.1168819	0.7324
weight	1	9.0438942	0.0026*

Parameter Estimates

Term	Estimate	Std Error	L-R ChiSquare	Prob>ChiSq	Lower CL	Upper CL
Intercept	-0.710184	0.9339406	0.5761557	0.4478	-2.532002	1.1287428
color[Light Med]	0.3273541	0.132167	5.7773064	0.0162*	0.0619978	0.5810457
color[Medium]	0.0625029	0.0737267	0.7238945	0.3949	-0.080871	0.2083795
color[Dark Med]	-0.186351	0.0963747	3.8096887	0.0510	-0.377505	0.0007691
spine[Both Good]	0.0210298	0.0947416	0.049346	0.8242	-0.163737	0.2084429
spine[One Worn/Broken]	-0.129342	0.1313469	1.0163566	0.3134	-0.399791	0.1168764
width	0.0167487	0.0489197	0.1168819	0.7324	-0.07991	0.1118218
weight	0.0004965	0.0001663	9.0438942	0.0026*	0.000172	0.0008234

The Whole Model Test shows that the difference between log-likelihoods for the full and reduced models is 41.6. The Prob>ChiSq value is equivalent to the *p*-value for a whole model *F* test. A *p*-value less than 0.0001 indicates that the model as a whole is significant. The report also contains the corrected Akaike Information Criterion (AICc), 921.7613. This value can be compared with other models to determine the best-fitting model for the data. Smaller AICc values indicate better fitting models.

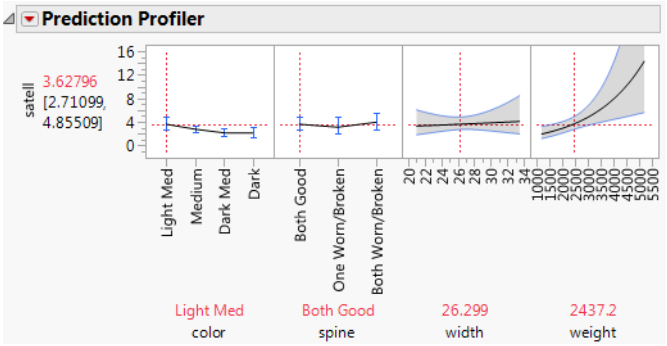
The Effects Tests report shows that weight is a significant factor in the model. Note that the *p*-value for weight, 0.0026, is the same in the Parameter Estimates table as well, since weight is a continuous variable.

The Effect Tests report also shows that the categorical variable color is significant. Color has four levels, Light Med, Medium, Dark Med, and Dark. The parameter estimates for the first three levels are shown in the Parameter Estimates table. The parameter estimate for Dark is the negative of the sum of the parameter estimates for the first three levels.

You can use the Prediction Profiler to further explore the model.

8. From the Generalized Linear Model Fit red triangle menu, select **Profilers > Profiler**.

Figure 12.3 Prediction Profiler for Satell

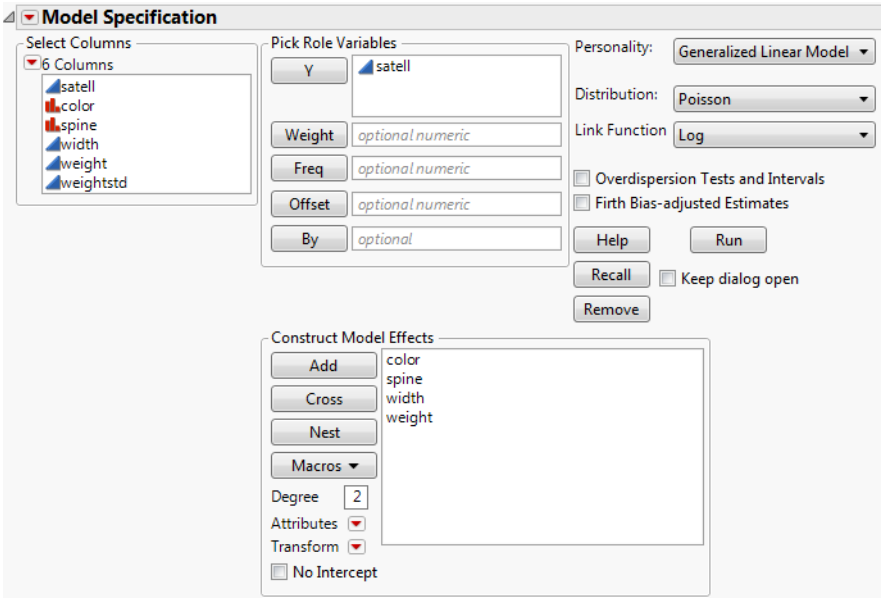


The confidence band on weight indicates that there is less variability in the model for smaller weight values than there is for larger weight values. The profiler enables you to easily explore the levels of the categorical variables. From the profiler, you can see that the predicted number of satellite crabs decreases as the color of the crab changes from light to dark.

Launch the Generalized Linear Model Personality

Launch the Generalized Linear Model personality by selecting **Analyze > Fit Model**, entering one or more columns for **Y**, and selecting **Generalized Linear Model** from the **Personality** menu.

Figure 12.4 Fit Model Launch Window with Generalized Linear Model Selected



For more information about aspects of the Fit Model window that are common to all personalities, see the "Model Specification" chapter on page 33. Information specific to the Generalized Linear Model personality is presented here.

If your model effects have missing values, you can treat these missing values as informative categories. Select the Informative Missing option from the Model Specification red triangle menu.

Tip: The No Intercept option is not available in the Generalized Linear Model personality of the Fit Model platform.

When you select Generalized Linear Model for Personality, the Fit Model launch window changes to include additional options. The following additional options are available in the Generalized Linear Model personality:

Distribution Specifies a probability distribution for the response variable.

Link Function Specifies a link function that relates the linear model to the response variable.

Overdispersion Tests and Intervals Specifies that an overdispersion parameter should be included in the model. Overdispersion occurs when the variance of the response is greater than would be expected by the theoretical variance of the response distribution. This can arise in Poisson and binomial response models. McCullagh and Nelder (1989) state that overdispersion is not uncommon in practice.

Note: This option adds a column for Overdispersion to the Goodness-of-Fit Statistics table in the Whole Model Test report.

Firth Bias-Adjusted Estimates Specifies that the Firth bias-adjusted method is used to fit the model. This maximum likelihood-based method has been shown to produce better estimates and tests than maximum likelihood-based models that do not use bias correction. In addition, bias-corrected MLEs ameliorate separation problems that tend to occur in logistic-type models. For more information about the separation problem in logistic regression, see Firth (1993) and Heinze and Schemper (2002).

Offset (Appears as a Pick Role Variables button.) Specifies a variable for the offset. An offset variable is one that is treated like a regression covariate whose parameter is fixed to be 1.0. Offset variables are most often used to scale the modeling of the mean in Poisson regression situations with a log link.

Response Specification for the Binomial Response Distribution

When you select Binomial as the Distribution, the response variable must be specified in one of the following ways:

- If your data are not summarized as frequencies of events, specify a single binary column as the response. The response column must be nominal.
- If your data are summarized as frequencies of events, specify a single binary column as the response and a frequency variable in the Freq role. The response column must be nominal, and the frequency variable contains the count of each response level.
- If your data are summarized as frequencies of events and number of trials, specify two continuous columns in this order: a count of the number of successes, and a count of the number of trials. Alternatively, you can specify the number of failures instead of the number of successes.

Generalized Linear Model Fit Report

By default, the Generalized Linear Model Fit report contains details about the model specification as well as the following reports:

Singularity Details (Appears only when there are linear dependencies among the model terms.) Shows a report that contains the linear functions that the model terms satisfy.

Regression Plot (Appears only when there is one continuous predictor and no more than one categorical predictor.) Shows a plot of the response on the vertical axis and the continuous predictor on the horizontal axis. A regression line is shown over the points. If

there is a categorical predictor in the model, each level of the categorical predictor has a separate regression line and a legend appears next to the plot.

Whole Model Test Shows tests that compare the whole-model fit to the model that omits all the effects except the intercept parameters. This report also contains goodness-of-fit statistics and the corrected Akaike's Information Criterion (AICc) value. See "Whole Model Test" on page 507.

Effect Summary An interactive report that enables you to add or remove effects from the model. See "Effect Summary Report" on page 182 in the "Standard Least Squares Report and Options" chapter.

Effect Tests The Effect Tests are joint tests that all the parameters for an individual effect are zero. If an effect has only one parameter, as with continuous effects, then the tests are the same as the tests in the Parameter Estimates table.

Note: Even if the Firth adjustment is used, the Effect Tests are based on the non-penalized likelihood function.

Parameter Estimates Shows the parameter estimates, standard errors, and associated hypothesis tests and confidence limits. Simple continuous effects have only one parameter. Models with complex classification effects have a parameter for each anticipated degree of freedom.

Note: If there are more than 1,000 observations, Wald-based confidence intervals are shown. Otherwise, profile-likelihood confidence intervals are shown.

Studentized Deviance Residual by Predicted Shows a plot of studentized deviance residuals on the vertical axis and the predicted response values on the horizontal axis.

Whole Model Test

The Whole Model Test table shows tests that compare the whole-model fit to the model that omits all the regression parameters except the intercept parameter. It also contains two goodness-of-fit statistics and the AICc value to assess model adequacy.

The Whole Model Test table shows these quantities:

Model Lists the model labels.

Difference The difference between the Full model and the Reduced model. This model is used to measure the significance of the regressors as a whole to the fit.

Full The complete model that includes the intercepts and all effects.

Reduced The model that includes only the intercept parameters.

–LogLikelihood The negative log-likelihood for the respective models.

L-R ChiSquare The likelihood ratio chi-square test statistic for the hypothesis that all regression parameters are zero. The test statistic is computed by taking twice the difference in negative log-likelihoods between the fitted model and the reduced model that has only an intercept.

DF The degrees of freedom (DF) for the Difference between the Full and Reduced model.

Prob>ChiSq The probability of obtaining a greater chi-square value by chance alone if the specified model fits no better than the model that includes only an intercept.

Goodness of Fit Statistic Lists the two goodness-of-fit statistics: Pearson and Deviance.

ChiSquare The chi-square test statistic for the respective goodness-of-fit statistics.

DF The degrees of freedom for the respective goodness-of-fit statistics.

Prob>ChiSq The *p*-value for the respective goodness-of-fit statistics.

Overdispersion (Appears only when the Overdispersion Tests and Intervals option is selected in the launch window.) An estimate of the overdispersion parameter. See "Statistical Details for the Generalized Linear Model Personality" on page 516.

AICc The corrected Akaike Information Criterion. For more information, see "Likelihood, AICc, and BIC" on page 550 in the "Statistical Details" appendix.

Generalized Linear Model Fit Report Options

The Generalized Linear Model Fit red triangle menu contains the following options:

Custom Test Enables you to test a custom hypothesis. For more information about custom tests, see "Custom Test" on page 125 in the "Standard Least Squares Report and Options" chapter.

Contrast Enables you to test for differences in levels within a variable. If a contrast involves a covariate, you can specify the value of the covariate at which to test the contrast. For an example of the Contrast option, see "Using Contrasts to Compare Differences in the Levels of a Variable" on page 511.

Inverse Prediction (Available only for continuous *X* variables.) Enables you to predict an *X* value, given specific values for *Y* and the other *X* variables. For more information about

the Inverse Prediction option, see "Inverse Prediction" on page 143 in the "Standard Least Squares Report and Options" chapter.

Covariance of Estimates Shows or hides a covariance matrix for all the effects in a model. The estimated covariance matrix of the parameter estimator is defined as follows:

$$\Sigma = -\mathbf{H}^{-1}$$

where **H** is the Hessian (or second derivative) matrix evaluated using the parameter estimates on the last iteration. Note that the dispersion parameter, whether estimated or specified, is incorporated into **H**. Rows and columns corresponding to aliased parameters are not included in Σ.

Correlation of Estimates Shows or hides a correlation matrix for all the effects in a model. The correlation matrix is the normalized covariance matrix. For each σ_{ij} element of Σ, the corresponding element of the correlation matrix is $\sigma_{ij}/\sigma_i\sigma_j$, where $\sigma_i = \sqrt{\sigma_{ii}}$.

Profilers Shows a submenu of the following profilers:

Profiler Shows or hides a prediction profiler for examining prediction traces for each X variable. For more information about the prediction profiler, see "Profiler" on page 165 in the "Standard Least Squares Report and Options" chapter.

Contour Profiler Shows or hides an interactive contour profiler. For more information about the contour profiler, see the Contour Profiler chapter in the *Profilers* book.

Surface Profiler Shows or hides a 3-D surface profiler. For more information about the surface profiler, see the Surface Plot chapter in the *Profilers* book.

Diagnostic Plots Shows a submenu of plots of residuals, predicted values, and actual values. These plots enable you to search for outliers and determine the adequacy of your model. For more information about deviance, see "Model Selection and Deviance" on page 518. The following plots are available:

Studentized Deviance Residuals by Predicted Shows or hides a plot of studentized deviance residuals on the vertical axis and the predicted response values on the horizontal axis.

Studentized Pearson Residuals by Predicted Shows or hides a plot of the studentized Pearson residuals on the vertical axis and the predicted response values on the horizontal axis.

Deviance Residuals by Predicted Shows or hides a plot of the deviance residuals on the vertical axis and the predicted response values on the horizontal axis.

Pearson Residuals by Predicted Shows or hides a plot of the Pearson residuals on the vertical axis and the predicted response values on the horizontal axis.

Regression Plot (Available only when there is one continuous predictor and no more than one categorical predictor.) Shows or hides a plot of the response on the vertical axis and the continuous predictor on the horizontal axis. A regression line is shown over the points. If there is a categorical predictor in the model, each level of the categorical predictor has a separate regression line and a legend appears next to the plot.

Linear Predictor Plot (Available only when there is one continuous predictor and no more than one categorical predictor.) Shows or hides a plot of responses transformed by the inverse link function on the vertical axis and the continuous predictor on the horizontal axis. A transformed regression line is shown over the points. If there is a categorical predictor in the model, each level of the categorical predictor has a separate transformed regression line and a legend appears next to the plot.

Save Columns Shows a submenu of options that enable you to save certain quantities as new columns in the current data table. For more information about the residual formulas, see "Residual Formulas" on page 520.

Prediction Formula Creates a formula column in the current data table that predicts the model.

Predicted Values Creates a column in the current data table that contains the values predicted by the model.

Mean Confidence Interval Creates columns in the current data table that contain the 95% confidence limits for the prediction equation for the model. These confidence limits reflect the variation in the parameter estimates.

Note: You can change the α level for the confidence limits in the Fit Model window by selecting Set Alpha Level from the red triangle menu.

Save Indiv Confid Limits Creates columns in the current data table that contain the 95% confidence limits for a given individual value for the model. These confidence limits reflect variation in the error and variation in the parameter estimates.

Note: You can change the α level for the confidence limits in the Fit Model window by selecting Set Alpha Level from the red triangle menu.

Deviance Residuals Creates a column in the current data table that contains the deviance residuals.

Pearson Residuals Creates a column in the current data table that contains the Pearson residuals.

Studentized Deviance Residuals Creates a column in the current data table that contains the studentized deviance residuals.

Studentized Pearson Residuals Creates a column in the current data table that contains the studentized Pearson residuals.

Model Dialog Shows the completed launch window for the current analysis.

Effect Summary Shows or hides the Effect Summary report, which enables you to interactively update the effects in the model. See "Effect Summary Report" on page 182 in the "Standard Least Squares Report and Options" chapter.

See the JMP Reports chapter in the *Using JMP* book for more information about the following options:

Redo Contains options that enable you to repeat or relaunch the analysis. In platforms that support the feature, the Automatic Recalc option immediately reflects the changes that you make to the data table in the corresponding report window.

Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

Save By-Group Script Contains options that enable you to save a script that reproduces the platform report for all levels of a By variable to several destinations. Available only when a By variable is specified in the launch window.

Additional Examples of the Generalized Linear Models Personality

Using Contrasts to Compare Differences in the Levels of a Variable

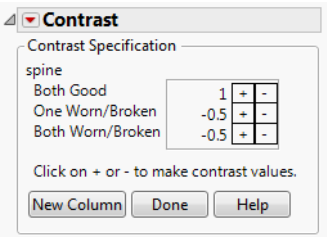
This example continues the crab satellite example in "Example of a Generalized Linear Model" on page 502. Suppose that you want to test whether female crabs with good spines attracted a different number of male crabs (satellites) than female crabs with worn or broken spines.

1. Complete step 1 through step 7 of "Example of a Generalized Linear Model" on page 502.

2. Click the red triangle next to Generalized Linear Model Fit and select **Contrast**. The Select Contrast Effect options appear at the bottom of the report window.
3. Select spine, the variable of interest, and click **Go**.
4. To compare the crabs with good spines to crabs with worn or broken spines, click the **+** button beside Both Good and the **-** button beside both One Worn/Broken and Both Worn/Broken.

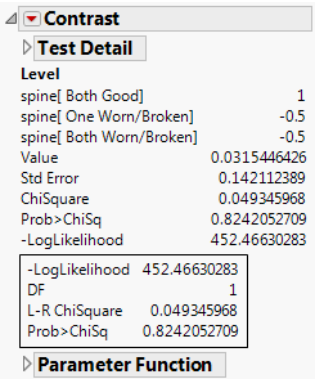
   This creates a contrast specification that compares the female crabs with good spines against the female crabs with worn or broken spines.

Figure 12.5 Contrast Specification Window



5. Click **Done**.

Figure 12.6 Contrast Report



The Prob>Chisq, 0.8242, is much greater than 0.05, so we cannot conclude that there is a difference in satellite attraction based on spine condition.

Poisson Regression with Offset

The sample data table Ship Damage.JMP is adapted from data found in McCullagh and Nelder (1989). The data table contains information about a certain type of damage caused by waves to the forward section of the hull. Hull construction engineers are interested in the risk

of damage associated with three variables: ship type (Type), the year in which the ship was constructed (Yr Made), and the block of years the ship was in service (Yr Used).

In this analysis we use the variable Service, the log of the aggregate months of service, as an *offset variable*. An offset variable is one that is treated like a regression covariate whose parameter is fixed to be 1.0. Offset variables are most often used to scale the modeling of the mean in Poisson regression situations with a log link. In this example, we use log(months of service) since one would expect that the number of repairs be proportional to the number of months in service.

To see how an offset variable is used, assume the linear component of the GLM is called η. Then, with a log link function, the model for the mean with the offset included is as follows:

$$\exp[\text{Log}(\text{months of service}) + \eta] = [(\text{months of service}) * \exp(\eta)].$$

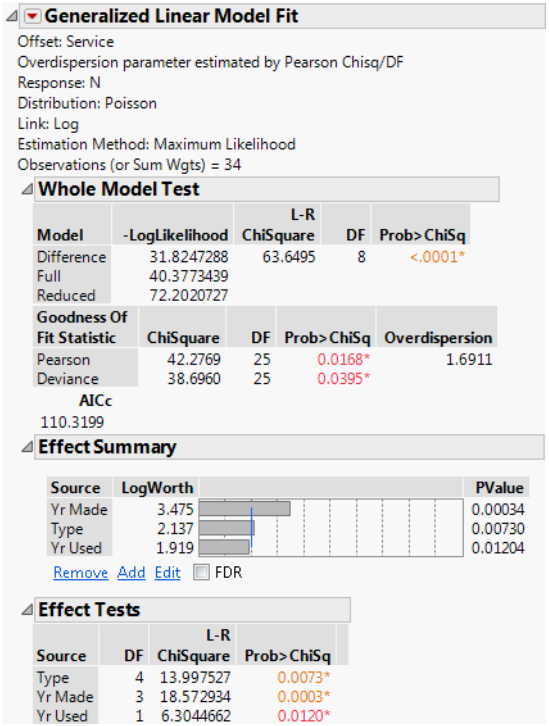
To run this example, follow these steps:

1. Select **Help > Sample Data Library** and open Ship Damage.jmp.
2. Select **Analyze > Fit Model**.
3. From the Personality list, select **Generalized Linear Model**.
4. From the Distribution list, select **Poisson**.

   In the Link Function list, **Log** should be selected for you automatically.
5. Select N and click **Y**.
6. Select Service and click **Offset**.
7. Select Type, Yr Made, Yr Used and click **Add**.
8. Click the check mark box for **Overdispersion Tests and Intervals**.
9. Click **Run**.

From the report shown in Figure 12.7, notice that all three effects (Type, Yr Made, Yr Used) are significant.

Figure 12.7 Partial Report for a Poisson with Offset Model



Generalized Linear Model Fit

Offset: Service
Overdispersion parameter estimated by Pearson Chisq/DF
Response: N
Distribution: Poisson
Link: Log
Estimation Method: Maximum Likelihood
Observations (or Sum Wgts) = 34

Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	31.8247288	63.6495	8	<.0001*
Full	40.3773439			
Reduced	72.2020727			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>ChiSq	Overdispersion
Pearson	42.2769	25	0.0168*	1.6911
Deviance	38.6960	25	0.0395*	

AICc
110.3199

Effect Summary

Source	LogWorth	PValue
Yr Made	3.475	0.00034
Type	2.137	0.00730
Yr Used	1.919	0.01204

Remove Add Edit FDR

Effect Tests

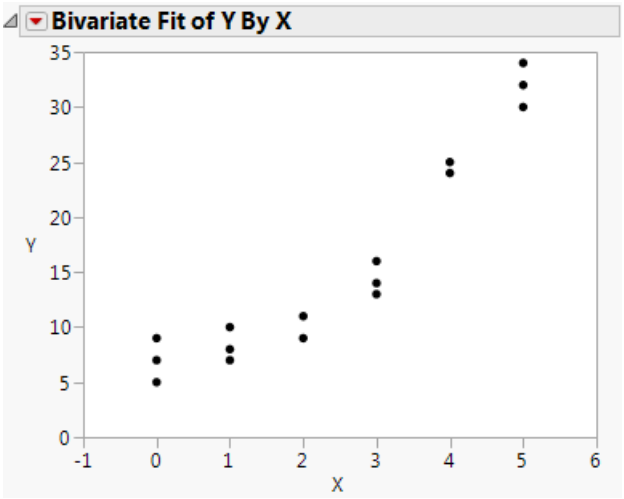
Source	DF	L-R ChiSquare	Prob>ChiSq
Type	4	13.997527	0.0073*
Yr Made	3	18.572934	0.0003*
Yr Used	1	6.3044662	0.0120*

Normal Regression with a Log Link

This example uses the Nor.jmp data table, where X is an explanatory variable and Y is the response variable. First, you explore the relationship between X and Y to determine the appropriate link function to use in the Generalized Linear Model personality of the Fit Model platform.

1. Select **Help > Sample Data Library** and open Nor.jmp.
2. Select **Analyze > Fit Y By X**.
3. Select Y and click **Y, Response**.
4. Select X, click **X, Factor**, and then click **OK**.

Figure 12.8 Y by X Results for Nor.jmp



You can see that Y varies nonlinearly with X and that the variance is approximately constant. Therefore, a normal distribution with a log link function is appropriate to model these data; that is, $\log(\mu_i) = \mathbf{x}_i'\beta$ so that $\mu_i = \exp(\mathbf{x}_i'\beta)$.

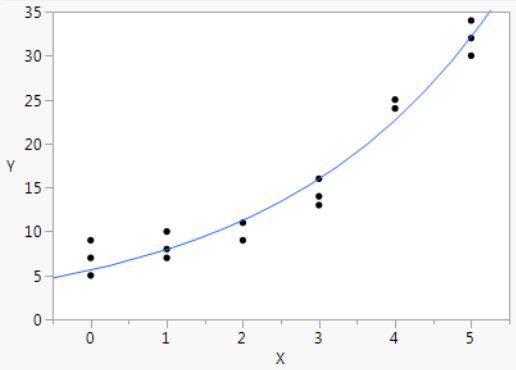
5. Select **Analyze > Fit Model**.
6. In the Personality list, select the **Generalized Linear Model**.
7. In the Distribution list, select **Normal**.
8. In the Link Function list, select **Log**.
9. Select Y and click **Y**.
10. Select X and click **Add**.
11. Click **Run**.

Figure 12.9 Nor Results

Generalized Linear Model Fit

Overdispersion parameter estimated by Maximum Likelihood
Response: Y
Distribution: Normal
Link: Log
Estimation Method: Maximum Likelihood
Observations (or Sum Wgts) = 16

Regression Plot



Whole Model Test

Model	-LogLikelihood	L-R ChiSquare	DF	Prob>ChiSq
Difference	26.5218281	53.0437	1	<.0001*
Full	32.1782768			
Reduced	58.7001049			

Goodness Of Fit Statistic	ChiSquare	DF	Prob>ChiSq	Overdispersion
Pearson	52.3000	14	<.0001*	3.2687
Deviance	52.3000	14	<.0001*	

AICc
72.3566

Effect Tests

Source	DF	L-R ChiSquare	Prob>ChiSq
X	1	53.043656	<.0001*

Parameter Estimates

Term	Estimate	Std Error	L-R ChiSquare	Prob>ChiSq	Lower CL	Upper CL
Intercept	1.7213577	0.0893972	37.973777	<.0001*	1.5239809	1.8978651
X	0.3496015	0.0206493	53.043656	<.0001*	0.3081539	0.3944836

Statistical Details for the Generalized Linear Model Personality

To construct a generalized linear model, you must select response and explanatory variables for your data. You then must choose an appropriate link function and probability distribution for your response. Explanatory variables can be any combination of continuous variables, classification variables, and interactions. Some common examples of generalized linear models are listed in Table 12.1.

Table 12.1 Examples of Generalized Linear Models

Model	Response Variable	Distribution	Default Link Function
Traditional Linear Model	continuous	Normal	identity, $g(\mu) = \mu$
Logistic Regression	a count or a binary random variable	Binomial	logit, $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
Poisson Regression in Log Linear Model	a count	Poisson	log, $g(\mu) = \log(\mu)$
Exponential Regression	positive continuous	Exponential	$\frac{1}{\mu}$

The platform fits a generalized linear model to the data by maximum likelihood estimation of the parameter vector. In general, there is no closed-form solution for the maximum likelihood estimates of the parameters. Therefore, the platform estimates the parameters of the model numerically through an iterative fitting process using a technique pioneered by Nelder and Wedderburn (1972). The overdispersion parameter ϕ is estimated by dividing the Pearson goodness-of-fit statistic by its degrees of freedom. Covariances, standard errors, and confidence limits are computed for the estimated parameters based on the asymptotic normality of maximum likelihood estimators.

A number of link functions and probability distributions are available in the Generalized Linear Model personality of the Fit Model platform. Table 12.2 lists the built-in link functions.

Table 12.2 Built-in Link Functions

Link Function Name	Link Function Formula
Identity	$g(\mu) = \mu$
Logit	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
Probit	$g(\mu) = \Phi^{-1}(\mu)$, where Φ is the standard normal cumulative distribution function

Table 12.2 Built-in Link Functions *(Continued)*

Link Function Name	Link Function Formula
Log	$g(\mu) = \log(\mu)$
Reciprocal	$g(\mu) = \frac{1}{\mu}$
Power	$g(\mu) = \begin{cases} \mu^{\lambda} & \text{if } (\lambda \neq 0) \\ \log(\mu) & \text{if } \lambda = 0 \end{cases}$
Comp LogLog	$g(m) = \log(-\log(1 - \mu))$

When you select the Power link function, a number box appears that enables you to enter the desired power.

Table 12.3 lists the variance functions associated with the available distributions for the response variable.

Table 12.3 Variance Functions for Response Distributions

Distribution	Variance Function
Normal	$V(\mu) = 1$
Binomial	$V(\mu) = \mu(1 - \mu)$
Poisson	$V(\mu) = \mu$
Exponential	$V(\mu) = \mu^2$

Model Selection and Deviance

An important aspect of generalized linear modeling is the selection of explanatory variables in the model. Changes in goodness-of-fit statistics are often used to evaluate the contribution of subsets of explanatory variables to a particular model. The *deviance* is defined to be twice the difference between the maximum attainable log-likelihood and the log-likelihood at the maximum likelihood estimates of the regression parameters. The deviance is often used as a measure of goodness of fit. The maximum attainable log-likelihood is achieved with a model that has a parameter for every observation. Table 12.4 lists the deviance formula for each of the available distributions for the response variable.

Table 12.4 Deviance Formulas for Response Distributions

Distribution	Deviance Formula
Normal	$\sum_i w_i(y_i - \mu_i)^2$
Binomial	$2\sum_i w_i m_i\left[y_i \log\left(\frac{y_i}{\mu_i}\right) + (1-y_i)\log\left(\frac{1-y_i}{1-\mu_i}\right)\right]$
Poisson	$2\sum_i w_i\left[y_i \log\left(\frac{y_i}{\mu_i}\right) - (y_i - \mu_i)\right]$
Exponential	$2\sum_i w_i\left[-\log\left(\frac{y_i}{\mu_i}\right) + \left(\frac{y_i - \mu_i}{\mu_i}\right)\right]$

The Pearson chi-square statistic is defined as follows:

$$X^2 = \sum_i \frac{w_i(y_i - \mu_i)^2}{V(\mu_i)}$$

where

y_i is the i^{th} response

μ_i is the corresponding predicted mean

$V(\mu_i)$ is the variance function

w_i is a known weight for the i^{th} observation

Note: If no weight is specified, $w_i = 1$ for all observations.

One strategy for variable selection is to fit a sequence of models. You start with a simple model that contains only an intercept term, and then include one additional explanatory variable in each successive model. You can measure the importance of the additional explanatory variable by the difference in deviance or fitted log-likelihood values between successive models. Asymptotic tests enable you to assess the statistical significance of the additional term.

When the distribution is non-normal, a normal critical value is used instead of a *t*-distribution critical value in inverse prediction.

Residual Formulas

Deviance

$$r_{Di} = \sqrt{d_i}(sign(y_i - \mu_i))$$

Studentized Deviance

$$r_{Di} = \frac{sign(y_i - \mu_i)\sqrt{d_i}}{\sqrt{\phi(1 - h_i)}}$$

Pearson

$$r_{Pi} = \frac{(y_i - \mu_i)}{\sqrt{V(\mu_i)}}$$

Studentized Pearson

$$r_{Pi} = \frac{y_i - \mu_i}{\sqrt{V(\mu_i)(1 - h_i)}}$$

where

$(y_i - \mu_i)$ is the raw residual

$sign(y_i - \mu_i)$ is 1 if $(y_i - \mu_i)$ is positive and -1 if $(y_i - \mu_i)$ is negative

d_i is the contribution to the total deviance from observation i

ϕ is the dispersion parameter

$V(\mu_i)$ is the variance function

h_i is the i^{th} diagonal element of the matrix $W_e^{(1/2)}X(X'W_eX)^{-1}X'W_e^{(1/2)}$, where W_e is the weight matrix used in computing the expected information matrix.

For more information about residuals and generalized linear models, see the GENMOD Procedure chapter in the *SAS/STAT 14.3 User's Guide* (SAS Institute Inc. 2017, ch. 46).

Appendix A

Statistical Details

Fitting Linear Models

This appendix discusses the different types of response models, their factors, their design coding, and parameterization. It also includes many other details of methods described in the main text.

The JMP system fits linear models to three different types of response models that are labeled continuous, ordinal, and nominal. Many details about the factor side are the same for the different response models, but JMP supports graphics and marginal profiles only for continuous responses—not for ordinal and nominal.

Different computer programs use different design-matrix codings, and thus parameterizations, to fit effects and construct hypothesis tests. JMP uses a different coding than the GLM procedure in the SAS system, although in most cases JMP and SAS GLM procedure produce the same results. The following sections describe the details of JMP coding and highlight those cases when it differs from that of the SAS GLM procedure.

Contents

The Response Models

JMP fits linear models to three different types of responses: continuous, nominal, and ordinal. The models and methods available in JMP are practical, are widely used, and suit the need for a general approach in a statistical software tool. As with all statistical software, you are responsible for learning the assumptions of the models that you choose to use, and the consequences if the assumptions are not met. For more information see "The Usual Assumptions" on page 545 in this chapter.

Continuous Responses

When the response column (column assigned the Y role) is continuous, JMP fits the value of the response directly. The basic model is that for each observation,

Y = (some function of the X's and parameters) + error

Statistical tests are based on the assumption that the error term in the model is normally distributed.

Fitting Principle for Continuous Response

The Fitting principle is called *least squares*. The least squares method estimates the parameters in the model to minimize the sum of squared errors. The errors in the fitted model, called *residuals*, are the difference between the actual value of each observation and the value predicted by the fitted model.

The least squares method is equivalent to the maximum likelihood method of estimation if the errors have a normal distribution. This means that the analysis estimates the model that gives the most likely residuals. The log-likelihood is a scale multiple of the sum of squared errors for the normal distribution.

Base Model for Continuous Responses

The simplest model for continuous measurement fits just one value to predict all the response values. This value is the estimate of the *mean*. The mean is just the arithmetic average of the response values. All other models are compared to this base model.

Nominal Responses

Nominal responses are analyzed with a straightforward extension of the logit model. For a binary (two-level) response, a logit response model is as follows:

$$\log\left(\frac{P(y=1)}{P(y=2)}\right) = X\beta$$

The above can also be written as follows:

$$P(y=1) = F(X\beta)$$

where $F(x)$ is the cumulative distribution function of the standard logistic distribution:

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

For r response levels, JMP fits the probabilities that the response is one of r different response levels given by the data values. The probability estimates must all be positive. For a given configuration of Xs, the probability estimates must sum to 1 over the response levels. The function that JMP uses to predict probabilities is a composition of a linear model and a multi-response logistic function. This is sometimes called a *log-linear* model because the logs of ratios of probabilities are linear models. JMP relates each response probability to the r^{th} probability and fit a separate set of design parameters to these r - 1 models.

$$\log\left(\frac{P(y=j)}{P(y=r)}\right) = X\beta_{(j)} \text{ for } j = 1, \ldots, r-1$$

Fitting Principle for Nominal Response

The fitting principle is called *maximum likelihood*. It estimates the parameters such that the joint probability for all the responses given by the data is the greatest obtainable by the model. Rather than reporting the joint probability (likelihood) directly, it is more manageable to report the total of the negative logs of the likelihood.

The uncertainty (negative log-likelihood) is the sum of the negative logs of the probabilities attributed by the model to the responses that actually occurred in the sample data. For a sample of size n, it is often denoted as H and written

$$H = \sum_{i=1}^{n} -\log(P(y = y_i))$$

If you attribute a probability of 1 to each event that did occur, then the sum of the negative logs is zero for a perfect fit.

The nominal model can take a lot of time and memory to fit, especially if there are many response levels. JMP tracks the progress of its calculations with an *iteration history*, which shows the negative log-likelihood values becoming smaller as they converge to the estimates.

Base Model for Nominal Responses

The simplest model for a nominal response is a set of constant response probabilities fitted as the occurrence rates for each response level across the whole data table. In other words, the probability that y is response level j is estimated by dividing the total sample count n into the total of each response level nj. This probability is written as follows:

$$p_j = \frac{n_j}{n}$$

All other models are compared to this base model. The base model serves the same role for a nominal response as the sample mean does for continuous models.

The R^2 statistic measures the portion of the uncertainty accounted for by the model, which is

$$1 - \frac{H(\text{full model})}{H(\text{base model})}$$

However, it is rare in practice to get an R^2 near 1 for categorical models.

Ordinal Responses

With an ordinal response (Y), as with nominal responses, JMP fits probabilities that the response is one of r different response levels given by the data.

Ordinal data have an order like continuous data. The order is used in the analysis but the spacing or distance between the ordered levels is not used. If you have a numeric response but want your model to ignore the spacing of the values, you can assign the ordinal level to that response column. If you have a classification variable and the levels are in some natural order such as low, medium, and high, you can use the ordinal modeling type.

Ordinal responses are modeled by fitting a series of parallel logistic curves to the cumulative probabilities. Each curve has the same design parameters but a different intercept and is written as follows:

$$\mathrm{P}(y \le j) = F(\alpha_j + \mathrm{X}\beta) \text{ for } j = 1, \ldots, r-1$$

where r response levels are present and $F(x)$ is the standard logistic cumulative distribution function

$$F(x) = \frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

Another way to write this is in terms of an unobserved continuous variable, z, that causes the ordinal response to change as it crosses various thresholds

$$y = \begin{cases} r & \alpha_{r-1} \le z \\ j & \alpha_{j-1} \le z < \alpha_j \\ 1 & z \le \alpha_1 \end{cases}$$

where z is an unobservable function of the linear model and error

$$z = X\beta + \varepsilon$$

and ε has the logistic distribution.

These models are attractive because they recognize the ordinal character of the response, they need far fewer parameters than nominal models, and the computations are fast.

A different but mathematically equivalent way to envision an ordinal model is to think of a nominal model where, instead of modeling the odds, you model the cumulative probability. Instead of fitting functions for all but the last level, you fit only one function and slide it to fit each cumulative response probability.

Fitting Principle for Ordinal Response

The maximum likelihood fitting principle for an ordinal response model is the same as for a nominal response model. It estimates the parameters such that the joint probability for all the responses that occur is the greatest obtainable by the model. It uses an iterative method that is faster and uses less memory than nominal fitting.

Base Model

The simplest model for an ordinal response, like a nominal response, is a set of response probabilities fitted as the occurrence rates of the response in the whole data table.

The Factor Models

The way the x-variables (factors) are modeled to predict an expected value or probability is the subject of the factor side of the model.

The factors enter the prediction equation as a linear combination of x values and the parameters to be estimated. For a continuous response model, where i indexes the

observations and j indexes the parameters, the assumed model for a typical observation, y_i, is written

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + \varepsilon_i \text{ where}$$

y_i is the response

x_{ij} are functions of the data

ε_i is an unobservable realization of the random error

b_j are unknown parameters to be estimated.

The way the x's in the linear model are formed from the factor terms is different for each modeling type. The linear model x's can also be complex effects such as interactions or nested effects. Complex effects are discussed in detail later.

Continuous Factors

Continuous factors are placed directly into the design matrix as regressors. If a column is a linear function of other columns, then the parameter for this column is marked *zeroed* or *nonestimable*. Continuous factors are centered by their mean when they are crossed with other factors (interactions and polynomial terms). Centering is suppressed if the factor has a Column Property of **Mixture** or **Coding**, or if the centered polynomials option is turned off when specifying the model. If there is a coding column property, the factor is coded before fitting.

Nominal Factors

Nominal factors are transformed into indicator variables for the design matrix. SAS GLM constructs an indicator column for each nominal level. JMP constructs the same indicator columns for each nominal level except the last level. When the last nominal level occurs, a one is subtracted from all the other columns of the factor. For example, consider a nominal factor A with three levels coded for GLM and for JMP as shown below.

Table A.1 Nominal Factor A

	GLM			JMP	
A	A1	A2	A3	A13	A23
A1	1	0	0	1	0
A2	0	1	0	0	1
A3	0	0	1	–1	–1

In GLM, the linear model design matrix has linear dependencies among the columns, and the least squares solution uses a generalized inverse. The solution chosen happens to be such that the A3 parameter is set to zero.

In JMP, the linear model design matrix is coded so that it achieves full rank unless there are missing cells or other incidental collinearity. The parameter for the A effect for the last level is the negative sum of the other levels, which makes the parameters sum to zero over all the effect levels.

Interpretation of Parameters

Note: The parameter for a nominal level is interpreted as the differences in the predicted response for that level from the average predicted response over all levels.

The design column for a factor level is constructed as the zero-one indicator of that factor level minus the indicator of the last level. This is the coding that leads to the parameter interpretation above.

Table A.2 Interpreting Parameters

JMP Parameter Report	How to Interpret	Design Column Coding
Intercept	mean over all levels	$\mathbf{1}'$
A[1]	$\alpha_1 - 1/3(\alpha_1 + \alpha_2 + \alpha_3)$	(A==1) – (A==3)
A[2]	$\alpha_2 - 1/3(\alpha_1 + \alpha_2 + \alpha_3)$	(A==2) – (A==3)

Interactions and Crossed Effects

Interaction effects with both GLM and JMP are constructed by taking a direct product over the rows of the design columns of the factors being crossed. For example, the GLM code

```
PROC GLM;
  CLASS A B;
  MODEL A B A*B;
```

yields this design matrix:

Table A.3 Design Matrix

		A			B			AB								
A	B	1	2	3	1	2	3	11	12	13	21	22	23	31	32	33

Table A.3 Design Matrix *(Continued)*

A1	B1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0
A1	B2	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0
A1	B3	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0
A2	B1	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0
A2	B2	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0
A2	B3	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0
A3	B1	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0
A3	B2	0	0	1	0	1	0	0	0	0	0	0	0	0	1	0
A3	B3	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1

Using the JMP **Fit Model** command and requesting a factorial model for columns A and B produces the following design matrix. Note that A13 in this matrix is A1–A3 in the previous matrix. However, A13B13 is A13*B13 in the current matrix.

Table A.4 Current Matrix

		A		B					
A	B	13	23	13	23	A13 B13	A13 B23	A23 B13	A23 B23
A1	B1	1	0	1	0	1	0	0	0
A1	B2	1	0	0	1	0	1	0	0
A1	B3	1	0	–1	–1	–1	–1	0	0
A2	B1	0	1	1	0	0	0	1	0
A2	B2	0	1	0	1	0	0	0	1
A2	B3	0	1	–1	–1	0	0	–1	–1
A3	B1	–1	–1	1	0	–1	0	–1	0
A3	B2	–1	–1	0	1	0	–1	0	–1
A3	B3	–1	–1	–1	–1	1	1	1	1

The JMP coding saves memory and some computing time for problems with interactions of factors with few levels.

The expected values of the cells in terms of the parameters for a three-by-three crossed model are:

Table A.5 Three-by-Three Crossed Model

	B1	B2	B3
A1	$\mu + \alpha_1 + \beta_1 + \alpha\beta_{11}$	$\mu + \alpha_1 + \beta_2 + \alpha\beta_{12}$	$\mu + \alpha_1 - \beta_1 - \beta_2 - \alpha\beta_{11} - \alpha\beta_{12}$
A2	$\mu + \alpha_2 + \beta_1 + \alpha\beta_{21}$	$\mu + \alpha_2 + \beta_2 + \alpha\beta_{22}$	$\mu + \alpha_2 - \beta_1 - \beta_2 - \alpha\beta_{21} - \alpha\beta_{22}$
A3	$\mu - \alpha_1 - \alpha_2 + \beta_1 - \alpha\beta_{11} - \alpha\beta_{21}$	$\mu - \alpha_1 - \alpha_2 + \beta_2 - \alpha\beta_{12} - \alpha\beta_{22}$	$\mu - \alpha_1 - \alpha_2 - \beta_1 - \beta_2 + \alpha\beta_{11} + \alpha\beta_{12} + \alpha\beta_{21} + \alpha\beta_{22}$

Nested Effects

Nested effects in GLM are coded the same as interaction effects because GLM determines the right test by what is not in the model. Any effect not included in the model can have its effect soaked up by a containing interaction (or, equivalently, nested) effect.

Nested effects in JMP are coded differently. JMP uses the terms inside the parentheses as grouping terms for each group. For each combination of levels of the nesting terms, JMP constructs the effect on the outside of the parentheses. The levels of the outside term do not need to line up across the levels of the nesting terms. Each level of nest is considered separately with regard to the construction of design columns and parameters.

Table A.6 Nested Effects

				B(A)					
				A1	A1	A2	A2	A3	A3
A	B	A13	A23	B13	B23	B13	B23	B13	B23
A1	B1	1	0	1	0	0	0	0	0
A1	B2	1	0	0	1	0	0	0	0
A1	B3	1	0	–1	–1	0	0	0	0
A2	B1	0	1	0	0	1	0	0	0
A2	B2	0	1	0	0	0	1	0	0
A2	B3	0	1	0	0	–1	–1	0	0

Table A.6 Nested Effects *(Continued)*

A3	B1	–1	–1	0	0	0	0	1	0
A3	B2	–1	–1	0	0	0	0	0	1
A3	B3	–1	–1	0	0	0	0	–1	–1

Least Squares Means across Nominal Factors

Least squares means are the predicted values corresponding to some combination of levels, after setting all the other factors to some neutral value. The neutral value for direct continuous regressors is defined as the sample mean. The neutral value for an effect with uninvolved nominal factors is defined as the average effect taken over the levels (which happens to result in all zeros in our coding). Ordinal factors use a different neutral value in "Ordinal Least Squares Means" on page 541. The least squares means might not be estimable, and if not, they are marked nonestimable. The least squares means in JMP agree with those in SAS PROC GLM (Goodnight and Harvey 1978) in all cases except when a weight is used. When a weight variable is used, JMP uses a weighted mean and SAS PROC GLM uses an unweighted mean for its neutral values.

Effective Hypothesis Tests

Generally, the hypothesis tests produced by JMP agree with the hypothesis tests of most other trusted programs, such as SAS PROC GLM (Hypothesis types III and IV). The following two sections describe where there are differences.

In SAS PROC GLM, the hypothesis tests for Types III and IV are constructed using the general form of estimable functions and finding functions that involve only the effects of interest and effects contained by the effects of interest (Goodnight 1978).

The same tests are constructed in JMP. However, because there is a different parameterization, an effect can be tested (assuming full rank for now) by doing a joint test on all the parameters for that effect. The tests do not involve containing interaction parameters because the coding has made them uninvolved with the tests on their contained effects.

If there are missing cells or other singularities, the JMP tests are different from GLM tests. There are several ways to describe them:

- JMP tests are equivalent to testing that the least squares means are different, at least for main effects. If the least squares means are nonestimable, then the test cannot include some comparisons and therefore loses degrees of freedom. For interactions, JMP is testing that the least squares means differ by more than just the marginal pattern described by the containing effects in the model.

- JMP tests an effect by comparing the SSE for the model with that effect to the SSE for the model without that effect. This logic applies if there are no nested terms, which complicate the logic slightly. JMP parameterizes the model so that this method makes sense.
- JMP implements the *effective hypothesis tests* described by Hocking (1985, pp. 80–89, 163–166), although JMP uses structural rather than cell-means parameterization. Effective hypothesis tests start with the hypothesis desired for the effect and include "as much as possible" of that test. Of course, if there are containing effects with missing cells, then this test has to drop part of the hypothesis because the complete hypothesis would not be estimable. The effective hypothesis drops as little of the complete hypothesis as possible.
- The differences among hypothesis tests in JMP and GLM (and other programs) that relate to the presence of missing cells are not considered interesting tests anyway. If an interaction is significant, the test for the contained main effects is not interesting. If the interaction is not significant, then it can always be dropped from the model. Some tests are not even unique. If you relabel the levels in a missing cell design, then the GLM Type IV tests can change.

The following section continues this topic in finer detail.

Singularities and Missing Cells in Nominal Effects

Consider the case of linear dependencies among the design columns. With JMP coding, this does not occur unless there is insufficient data to fill out the combinations that need estimating, or unless there is some type of confounding or collinearity of the effects.

With linear dependencies, a least squares solution for the parameters might not be unique and some tests of hypotheses cannot be tested. The strategy chosen for JMP is to set parameter estimates to zero in sequence as their design columns are found to be linearly dependent on previous effects in the model. A special column in the report shows what parameter estimates are zeroed and which parameter estimates are estimable. A separate *singularities* report shows what the linear dependencies are.

In cases of singularities the hypotheses tested by JMP can differ from those selected by GLM. Generally, JMP finds fewer degrees of freedom to test than GLM because it holds its tests to a higher standard of marginality. In other words, JMP tests always correspond to tests across least squares means for that effect, but GLM tests do not always have this property.

For example, consider a two-way model with interaction and one missing cell where A has three levels, B has two levels, and the A3B2 cell is missing.

Table A.7 Two-Way Model with Interaction

A B	A1	A2	B1	A1B1	A2B1
A1 B1	1	0	1	1	0

Table A.7 Two-Way Model with Interaction *(Continued)*

A B	A1	A2	B1	A1B1	A2B1	
A2 B1	0	1	1	0	1	
A3 B1	–1	–1	1	–1	–1	
A1 B2	1	0	–1	–1	0	
A2 B2	0	1	–1	0	–1	
A3 B2	–1	–1	–1	1	1	Suppose this interaction is missing.

The expected values for each cell are:

Table A.8 Expected Values

	B1	B2
A1	$\mu + \alpha_1 + \beta_1 + \alpha\beta_{11}$	$\mu + \alpha_1 - \beta_1 - \alpha\beta_{11}$
A2	$\mu + \alpha_2 + \beta_1 + \alpha\beta_{21}$	$\mu + \alpha_2 - \beta_1 - \alpha\beta_{21}$
A3	$\mu - \alpha_1 - \alpha_2 + \beta_1 - \alpha\beta_{11} - \alpha\beta_{21}$	$\mu - \alpha_1 - \alpha_2 - \beta_1 + \alpha\beta_{11} + \alpha\beta_{21}$

Obviously, any cell with data has an expectation that is estimable. The cell that is missing has an expectation that is nonestimable. In fact, its expectation is precisely that linear combination of the design columns that is in the singularity report

$$\mu - \alpha_1 - \alpha_2 - \beta_1 + \alpha\beta_{11} + \alpha\beta_{21}$$

Suppose that you want to construct a test that compares the least squares means of B1 and B2. In this example, the average of the rows in the above table give these least squares means.

$$\begin{aligned} \text{LSM(B1)} = {} & (1/3)(\mu + \alpha_1 + \beta_1 + \alpha\beta_{11} + \\ & \mu + \alpha_2 + \beta_1 + \alpha\beta_{21} + \\ & \mu - \alpha_1 - \alpha_2 + \beta_1 - \alpha\beta_{11} - \alpha\beta_{21}) \\ = {} & \mu + \beta_1 \end{aligned}$$

$$\begin{aligned} \text{LSM(B2)} = {} & (1/3)(\mu + \alpha_1 + -\beta_1 - \alpha\beta_{11} + \\ & \mu + \alpha_2 + -\beta_1 - \alpha\beta_{21} + \\ & \mu - \alpha_1 - \alpha_2 - \beta_1 + \alpha\beta_{11} + \alpha\beta_{21}) \\ = {} & \mu - \beta_1 \end{aligned}$$

$$\text{LSM(B1)} - \text{LSM(B2)} = 2\beta_1$$

Note that this shows that a test on the β_1 parameter is equivalent to testing that the least squares means are the same. But because β_1 is not estimable, the test is not testable, meaning there are no degrees of freedom for it.

Now, construct the test for the least squares means across the A levels.

$$\begin{aligned}\text{LSM(A1)} &= (1/2)(\mu + \alpha_1 + \beta_1 + \alpha\beta_{11} + \mu + \alpha_1 - \beta_1 - \alpha\beta_{11}) \\ &= \mu + \alpha_1\end{aligned}$$

$$\begin{aligned}\text{LSM(A2)} &= (1/2)(\mu + \alpha_2 + \beta_1 + \alpha\beta_{21} + \mu + \alpha_2 - \beta_1 - \alpha\beta_{21}) \\ &= \mu + \alpha_2\end{aligned}$$

$$\begin{aligned}\text{LSM(A3)} &= (1/2)(\mu - \alpha_1 - \alpha_2 + \beta_1 - \alpha\beta_{11} - \alpha\beta_{21} + \\ &\quad \mu - \alpha_1 - \alpha_2 - \beta_1 + \alpha\beta_{11} + \alpha\beta_{21}) \\ &= \mu - \alpha_1 - \alpha_2\end{aligned}$$

$$\text{LSM(A1)} - \text{LSM(A3)} = 2\alpha_1 + \alpha_2$$

$$\text{LSM(A2)} - \text{LSM(A3)} = 2\alpha_2 + \alpha_1$$

Neither of these turn out to be estimable, but there is another comparison that is estimable; namely comparing the two A columns that have no missing cells.

$$\text{LSM(A1)} - \text{LSM(A2)} = \alpha_1 - \alpha_2$$

This combination is indeed tested by JMP using a test with 1 degree of freedom, although there are two parameters in the effect.

The estimability can be verified by taking its inner product with the singularity combination, and checking that it is zero:

Table A.9 Verification

	singularity	combination
parameters	combination	to be tested
m	1	0
a_1	–1	1
a_2	–1	–1
b_1	–1	0

Table A.9 Verification *(Continued)*

ab_{11}	1	0
ab_{21}	1	0

It turns out that the design columns for missing cells for any interaction always knocks out degrees of freedom for the main effect (for nominal factors). Thus, there is a direct relation between the non-estimability of least squares means and the loss of degrees of freedom for testing the effect corresponding to these least squares means.

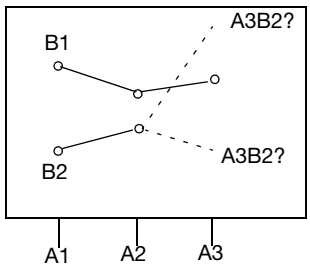
How does this compare with what GLM does? GLM and JMP do the same test when there are no missing cells. That is, they effectively test that the least squares means are equal. But when GLM encounters singularities, it focuses out these cells in different ways, depending on whether they are Type III or Type IV. For Type IV, it looks for estimable combinations that it can find. These might not be unique, and if you reorder the levels, you might get a different result. For Type III, it does some orthogonalization of the estimable functions to obtain a unique test. But the test might not be very interpretable in terms of the cell means.

The JMP approach has several points in its favor, although at first it might seem distressing that you might lose more degrees of freedom than with GLM:

1. The tests are philosophically linked to LSMs.
2. The tests are easy computationally, using reduction sum of squares for reparameterized models.
3. The tests agree with Hocking's "Effective Hypothesis Tests".
4. The tests are *whole marginal tests*, meaning they always go completely across other effects in interactions.

The last point needs some elaboration: Consider a graph of the expected values of the cell means in the previous example with a missing cell for A3B2.

Figure A.1 Expected Values of the Cell Means



The graph shows expected cell means with a missing cell. The means of the A1 and A2 cells are profiled across the B levels. The JMP approach says you cannot test the B main effect with

a missing A3B2 cell, because the mean of the missing cell could be anything, as allowed by the interaction term. If the mean of the missing cell were the higher value shown, the B effect would likely test significant. If it were the lower, it would likely test nonsignificant. The point is that you do not know. That is what the least squares means are saying when they are declared nonestimable. That is what the hypotheses for the effects should be saying too—that you do not know.

If you want to test hypotheses involving margins for subsets of cells, then that is what GLM Type IV does. In JMP, you would have to construct these tests yourself by partitioning the effects with a lot of calculations or by using contrasts.

JMP and GLM Hypotheses

GLM works differently than JMP and produces different hypothesis tests in situations where there are missing cells. In particular, GLM does not recognize any difference between a nesting and a crossing in an effect, but JMP does. Suppose that you have a three-layer nesting of A, B(A), and C(A B) with different numbers of levels as you go down the nested design.

Table A.10 on page 536, shows the test of the main effect A in terms of the GLM parameters. The first set of columns is the test done by JMP. The second set of columns is the test done by GLM Type IV. The third set of columns is the test equivalent to that by JMP; it is the first two columns that have been multiplied by a matrix:

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

to be comparable to the GLM test. The last set of columns is the GLM Type III test. The difference is in how the test distributes across the containing effects. In JMP, it seems more top-down hierarchical. In GLM Type IV, the test seems more bottom-up. In practice, the test statistics are often similar.

Table A.10 Comparison of GLM and JMP Hypotheses

Parameter	JMP Test for A		GLM-IV Test for A		JMP Rotated Test		GLM-III Test for A	
u	0	0	0	0	0	0	0	0
a1	0.6667	-0.3333	1	0	1	0	1	0
a2	–0.3333	0.6667	0	1	0	1	0	1
a3	–0.3333	-0.3333	-1	-1	-1	-1	-1	-1
a1b1	0.1667	-0.0833	0.2222	0	0.25	0	0.2424	0

Table A.10 Comparison of GLM and JMP Hypotheses *(Continued)*

a1b2	0.1667	-0.0833	0.3333	0	0.25	0	0.2727	0
a1b3	0.1667	-0.0833	0.2222	0	0.25	0	0.2424	0
a1b4	0.1667	-0.0833	0.2222	0	0.25	0	0.2424	0
a2b1	-0.1667	0.3333	0	0.5	0	0.5	0	.5
a2b2	-0.1667	0.3333	0	0.5	0	0.5	0	.5
a3b1	-0.1111	-0.1111	-0.3333	-0.3333	-0.3333	-0.3333	-0.3333	-0.3333
a3b2	-0.1111	-0.1111	-0.3333	-0.3333	-0.3333	-0.3333	-0.3333	-0.3333
a3b3	-0.1111	-0.1111	-0.3333	-0.3333	-0.3333	-0.3333	-0.3333	-0.3333
a1b1c1	0.0833	-0.0417	0.1111	0	0.125	0	0.1212	0
a1b1c2	0.0833	-0.0417	0.1111	0	0.125	0	0.1212	0
a1b2c1	0.0556	-0.0278	0.1111	0	0.0833	0	0.0909	0
a1b2c2	0.0556	-0.0278	0.1111	0	0.0833	0	0.0909	0
a1b2c3	0.0556	-0.0278	0.1111	0	0.0833	0	0.0909	0
a1b3c1	0.0833	-0.0417	0.1111	0	0.125	0	0.1212	0
a1b3c2	0.0833	-0.0417	0.1111	0	0.125	0	0.1212	0
a1b4c1	0.0833	-0.0417	0.1111	0	0.125	0	0.1212	0
a1b4c2	0.0833	-0.0417	0.1111	0	0.125	0	0.1212	0
a2b1c1	-0.0833	0.1667	0	0.25	0	0.25	0	0.25
a2b1c2	-0.0833	0.1667	0	0.25	0	0.25	0	0.25
a2b2c1	-0.0833	0.1667	0	0.25	0	0.25	0	0.25
a2b2c2	-0.0833	0.1667	0	0.25	0	0.25	0	0.25
a3b1c1	-0.0556	-0.0556	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667

Table A.10 Comparison of GLM and JMP Hypotheses *(Continued)*

a3b1c2	-0.0556	-0.0556	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667
a3b2c1	-0.0556	-0.0556	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667
a3b2c2	-0.0556	-0.0556	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667
a3b3c1	-0.0556	-0.0556	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667
a3b3c2	-0.0556	-0.0556	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667	-0.1667

Ordinal Factors

Factors marked with the ordinal modeling type are coded differently than nominal factors. The parameters estimates are interpreted differently, the tests are different, and the least squares means are different.

The theme for ordinal factors is that the first level of the factor is a control or baseline level, and the parameters measure the effect on the response as the ordinal factor is set to each succeeding level. The coding is appropriate for factors with levels representing various doses, where the first dose is zero:

Table A.11 Ordinal Factors

Term	Coded Column		
A	a2	a3	
A1	0	0	control level, zero dose
A2	1	0	low dose
A3	1	1	higher dose

From the perspective of the JMP parameterization, the tests for A are:

Table A.12 Tests for A

parameter	GLM–IV test		JMP test	
m	0	0	0	0
a13	2	1	1	0
a23	1	2	0	1

Table A.12 Tests for A *(Continued)*

a1:b14	0	0	0	0
a1:b24	0.11111	0	0	0
a1:b34	0	0	0	0
a2:b12	0	0	0	0
a3:b13	0	0	0	0
a3:b23	0	0	0	0
a1b1:c12	0	0	0	0
a1b2:c13	0	0	0	0
a1b2:c23	0	0	0	0
a1b3:c12	0	0	0	0
a1b4:c12	0	0	0	0
a2b1:c13	0	0	0	0
a2b2:c12	0	0	0	0
a3b1:c12	0	0	0	0
a3b2:c12	0	0	0	0
a3b3:c12	0	0	0	0

So from JMP's perspective, the GLM test looks a little strange, putting a coefficient on the a1b24 parameter.

The pattern for the design is such that the lower triangle is ones with zeros elsewhere.

For a simple main-effects model, this can be written as follows:

$$y = \mu + \alpha_2 X_{(a \le 2)} + \alpha_3 X_{(a \le 3)} + \varepsilon$$

noting that μ is the expected response at $A = 1$, $\mu + \alpha_2$ is the expected response at $A = 2$, and $\mu + \alpha_2 + \alpha_3$ is the expected response at $A = 3$. Thus, α_2 estimates the effect moving from $A = 1$ to $A = 2$ and α_3 estimates the effect moving from $A = 2$ to $A = 3$.

If all the parameters for an ordinal main effect have the same sign, then the response effect is monotonic across the ordinal levels.

Ordinal Interactions

The ordinal interactions, as with nominal effects, are produced with a horizontal direct product of the columns of the factors. Consider an example with two ordinal factors A and B, where each factor has three levels. The ordinal coding in JMP produces the design matrix shown next. The pattern for the interaction is a block lower-triangular matrix of lower-triangular matrices of ones.

Table A.13 Ordinal Interactions

						A*B			
						A2		A3	
A	B	A2	A3	B2	B3	B2	B3	B2	B3
A1	B1	0	0	0	0	0	0	0	0
A1	B2	0	0	1	0	0	0	0	0
A1	B3	0	0	1	1	0	0	0	0
A2	B1	1	0	0	0	0	0	0	0
A2	B2	1	0	1	0	1	0	0	0
A2	B3	1	0	1	1	1	1	0	0
A3	B1	1	1	0	0	0	0	0	0
A3	B2	1	1	1	0	1	0	1	0
A3	B3	1	1	1	1	1	1	1	1

Note: When you test to see whether there is no effect, there is not much difference between nominal and ordinal factors for simple models. However, there are major differences when interactions are specified. We recommend that you use nominal rather than ordinal factors for most models.

Hypothesis Tests for Ordinal Crossed Models

To see what the parameters mean, examine this table of the expected cell means in terms of the parameters, where μ is the intercept, α_2 is the parameter for level A2, and so on.

Table A.14 Expected Cell Means

	B1	B2	B3
A1	μ	$\mu + \alpha\beta_2 + \alpha\beta_{12}$	$\mu + \beta_2 + \beta_3$
A2	$\mu + \alpha_2$	$\mu + \alpha_2 + \beta_2 + \alpha\beta_{22}$	$\mu + \alpha_2 + \beta_2 + \beta_3 + \alpha\beta_{22} + \alpha\beta_{23}$
A3	$\mu + \alpha_2 + \alpha_3$	$\mu + \alpha_2 + \alpha_3 + \beta_2 + \alpha\beta_{22} + \alpha\beta_{32} + \alpha\beta_{23} + \alpha\beta_{32} + \alpha\beta_{33}$	$\mu + \alpha_2 + \alpha_3 + \beta_2 + \beta_3 + \alpha\beta_{22} + \alpha\beta_{23} + \beta_{32} + \alpha\beta_{33}$

Note that the main effect test for A is really testing the A levels holding B at the first level. Similarly, the main effect test for B is testing across the top row for the various levels of B holding A at the first level. This is the appropriate test for an experiment where the two factors are both doses of different treatments. The main question is the efficacy of each treatment by itself, with fewer points devoted to looking for *drug interactions* when doses of both drugs are applied. In some cases it might even be dangerous to apply large doses of each drug.

Note that each cell's expectation can be obtained by adding all the parameters associated with each cell that is to the left and above it, inclusive of the current row and column. The expected value for the last cell is the sum of all the parameters.

Though the hypothesis tests for effects contained by other effects differs with ordinal and nominal codings, the test of effects not contained by other effects is the same. In the crossed design above, the test for the interaction would be the same no matter whether A and B were fit nominally or ordinally.

Ordinal Least Squares Means

As stated previously, least squares means are the predicted values corresponding to some combination of levels, after setting all the other factors to some neutral value. JMP defines the neutral value for an effect with uninvolved ordinal factors as the effect at the first level, meaning the control, or *baseline* level.

This definition of least squares means for ordinal factors maintains the idea that the hypothesis tests for contained effects are equivalent to tests that the least squares means are equal.

Singularities and Missing Cells in Ordinal Effects

With the ordinal coding, you are saying that the first level of the ordinal effect is the baseline. It is thus possible to get good tests on the main effects even when there are missing cells in the interactions—even if you have no data for the interaction.

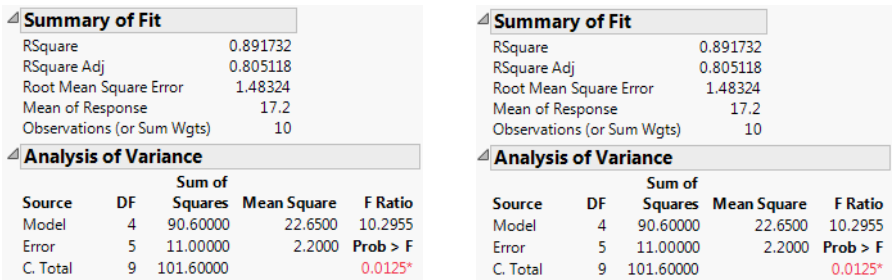
Example with Missing Cell

The example is the same as above, with two observations per cell except that the A3B2 cell has no data. You can now compare the results when the factors are coded nominally with results when they are coded ordinally. The model fit is the same, as seen in Figure A.2.

Table A.15 Observations

Y	A	B
12	1	1
14	1	1
15	1	2
16	1	2
17	2	1
17	2	1
18	2	2
19	2	2
20	3	1
24	3	1

Figure A.2 Summary Information for Nominal Fits (Left) and Ordinal Fits (Right)



Summary of Fit

RSquare	0.891732
RSquare Adj	0.805118
Root Mean Square Error	1.48324
Mean of Response	17.2
Observations (or Sum Wgts)	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	90.60000	22.6500	10.2955
Error	5	11.00000	2.2000	Prob > F
C. Total	9	101.60000		0.0125*

Summary of Fit

RSquare	0.891732
RSquare Adj	0.805118
Root Mean Square Error	1.48324
Mean of Response	17.2
Observations (or Sum Wgts)	10

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Ratio
Model	4	90.60000	22.6500	10.2955
Error	5	11.00000	2.2000	Prob > F
C. Total	9	101.60000		0.0125*

The parameter estimates are very different because of the different coding. Note that the missing cell affects estimability for some nominal parameters but for none of the ordinal parameters.

Figure A.3 Parameter Estimates for Nominal Fits (Left) and Ordinal Fits (Right)

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		17.333333	0.60553	28.63	<.0001*
A[1]		-4.333333	0.856349	-5.06	0.0039*
A[2]		-0.333333	0.856349	-0.39	0.7131
B[2-1]	Biased	1.5	1.48324	1.01	0.3583
A[1]*B[2-1]	Biased	1	2.097618	0.48	0.6537
A[2]*B[2-1]	Zeroed	0	0	.	.

Parameter Estimates

Term		Estimate	Std Error	t Ratio	Prob>|t|
Intercept		13	1.048809	12.40	<.0001*
A[2-1]		4	1.48324	2.70	0.0429*
A[3-2]		5	1.48324	3.37	0.0199*
B[2-1]		2.5	1.48324	1.69	0.1527
A[2-1]*B[2-1]		-1	2.097618	-0.48	0.6537
A[3-2]*B[2-1]	Zeroed	0	0	.	.

The singularity details show the linear dependencies (and also identify the missing cell by examining the values).

Figure A.4 Singularity Details for Nominal Fits (Left) and Ordinal Fits (Right)

Singularity Details

Term	Details
B[2-1]	=A[1]*B[2-1] + A[2]*B[2-1]

Singularity Details

Term	Details
A[3-2]*B[2-1]	= 0

The effect tests lose degrees of freedom for nominal. In the case of B, there is no test. For ordinal, there is no loss because there is no missing cell for the *base* first level.

Figure A.5 Effects Tests for Nominal Fits (Left) and Ordinal Fits (Right)

Effect Tests

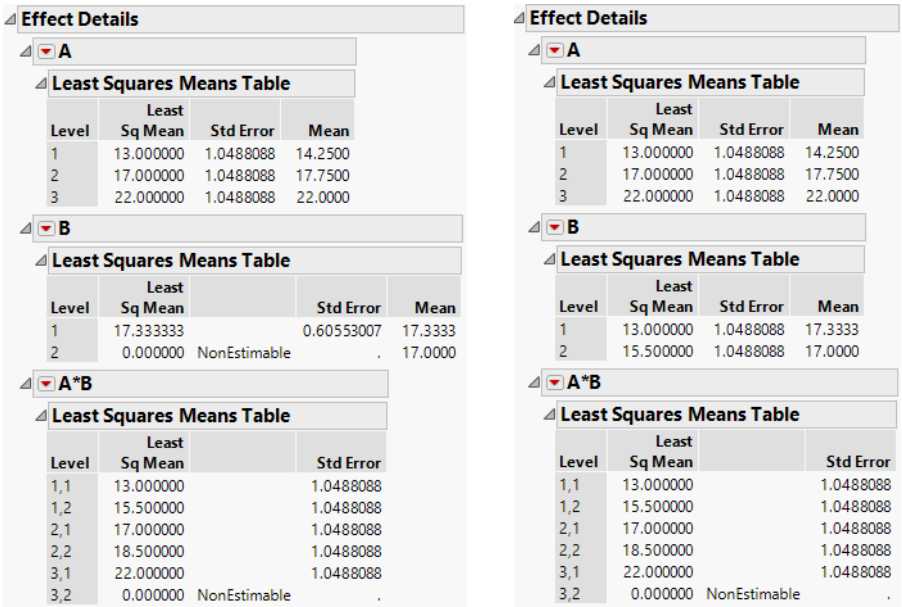
Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
A	2	2	81.333333	18.4848	0.0049*	
B	1	0	0.000000	.	.	LostDFs
A*B	2	1	0.500000	0.2273	0.6537	LostDFs

Effect Tests

Source	Nparm	DF	Sum of Squares	F Ratio	Prob > F	
A	2	2	81.333333	18.4848	0.0049*	
B	1	1	6.250000	2.8409	0.1527	
A*B	2	1	0.500000	0.2273	0.6537	LostDFs

The least squares means are also different. The nominal LSMs are not all estimable, but the ordinal LSMs are. You can verify the values by looking at the cell means. Note that the A*B LSMs are the same for the two. Figure A.6 shows least squares means for nominal and ordinal fits.

Figure A.6 Least Squares Means for Nominal Fits (Left) and Ordinal Fits (Right)



Effect Details

A

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
1	13.000000	1.0488088	14.2500
2	17.000000	1.0488088	17.7500
3	22.000000	1.0488088	22.0000

B

Least Squares Means Table

Level	Least Sq Mean		Std Error	Mean
1	17.333333		0.60553007	17.3333
2	0.000000	NonEstimable	.	17.0000

A*B

Least Squares Means Table

Level	Least Sq Mean		Std Error
1,1	13.000000		1.0488088
1,2	15.500000		1.0488088
2,1	17.000000		1.0488088
2,2	18.500000		1.0488088
3,1	22.000000		1.0488088
3,2	0.000000	NonEstimable	.

Effect Details

A

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
1	13.000000	1.0488088	14.2500
2	17.000000	1.0488088	17.7500
3	22.000000	1.0488088	22.0000

B

Least Squares Means Table

Level	Least Sq Mean	Std Error	Mean
1	13.000000	1.0488088	17.3333
2	15.500000	1.0488088	17.0000

A*B

Least Squares Means Table

Level	Least Sq Mean		Std Error
1,1	13.000000		1.0488088
1,2	15.500000		1.0488088
2,1	17.000000		1.0488088
2,2	18.500000		1.0488088
3,1	22.000000		1.0488088
3,2	0.000000	NonEstimable	.

Frequencies

The impact of frequencies, including those with non-integer values, on an analysis is explained by their effect on the loss function. Suppose that you want to estimate the parameter θ using response values y_i and predictors $x_{i1}, x_{i2},..., x_{in}$. Suppose that the loss function, assuming no frequency variable, is given by the following:

$$L(\theta|\underset{\sim}{y}) = \sum_{i=1}^{n} L(\theta|y_i, x_{i1}, x_{i2},\ldots,x_{in})$$

If frequencies f_i are defined, then the loss function is:

$$L(\theta|\underset{\sim}{y}, \underset{\sim}{f}) = \sum_{i=1}^{n} f_i L(\theta|y_i, x_{i1}, x_{i2},\ldots,x_{in})$$

Calculations for all inference-base quantities, such as parameter estimates, standard errors, hypothesis tests, and confidence intervals, are based on this form of the loss function.

The Usual Assumptions

Before you put your faith in statistics, reassure yourself that you know both the value and the limitations of the techniques that you use. Statistical methods are just tools—they cannot guard you from incorrect science (invalid statistical assumptions) or bad data.

Assumed Model

Most statistics are based on the assumption that the model is correct. To the extent that your model might not be correct, you must attenuate your credibility in the statistical reports that result from the model.

Relative Significance

Many statistical tests do not evaluate the model in an absolute sense. Significant test statistics might be saying only that the model fits better than some reduced model, such as the mean. The model can appear to fit the data but might not describe the underlying physical model well at all.

Multiple Inferences

Often the value of the statistical results is not that you believe in them directly, but rather that they provide a key to some discovery. To confirm the discovery, you might need to conduct further studies. Otherwise, you might just be sifting through the data.

For example, if you conduct enough analyses you can find 5% significant effects in 5% of your studies by chance alone, even if the factors have no predictive value. Similarly, to the extent that you use your data to shape your model (instead of testing the correct model for the data), you are corrupting the significance levels in your report. The random error then influences your model selection and leads you to believe that your model is better than it really is.

Validity Assessment

Some of the various techniques and patterns to look for in assessing the validity of the model are as follows:

- Model validity can be checked against a saturated version of the factors with Lack of Fit tests. The Fit Model platform presents these tests automatically if you have replicated x data in a non-saturated model.
- You can check the distribution assumptions for a continuous response by looking at plots of residuals and studentized residuals from the Fit Model platform. Or, use the **Save**

commands in the platform pop-up menu to save the residuals in data table columns. Then use the **Analyze > Distribution** on these columns to look at a histogram with its normal curve and the normal quantile plot. The residuals are not quite independent, but you can informally identify severely non-normal distributions.

- The best all-around diagnostic tool for continuous responses is the leverage plot because it shows the influence of each point on each hypothesis test. If you suspect that there is a mistaken value in your data, this plot helps determine whether a statistical test is heavily influenced by a single point.
- It is a good idea to scan your data for outlying values and examine them to see whether they are valid observations. You can spot univariate outliers in the Distribution platform reports and plots. Bivariate outliers appear in Fit Y by X scatterplots and in the Multivariate scatterplot matrix. You can see trivariate outliers in a three-dimensional plot produced by the **Graph > Scatterplot 3D**. Higher dimensional outliers can be found with Principal Components or Scatterplot 3D, and with Mahalanobis and jack-knifed distances computed and plotted in the Multivariate platform.

Alternative Methods

The statistical literature describes special nonparametric and robust methods, but JMP implements only a few of them at this time. These methods require fewer distributional assumptions (nonparametric), and then are more resistant to contamination (robust). However, they are less conducive to a general methodological approach, and the small sample probabilities on the test statistics can be time consuming to compute.

If you are interested in linear rank tests and need only normal large sample significance approximations, you can analyze the ranks of your data to perform the equivalent of a Wilcoxon rank-sum or Kruskal-Wallis one-way test.

If you are uncertain that a continuous response adequately meets normal assumptions, you can change the modeling type from continuous to ordinal and then analyze safely, even though this approach sacrifices some richness in the presentations and some statistical power as well.

Key Statistical Concepts

There are two key concepts that unify classical statistics and encapsulate statistical properties and fitting principles into forms that you can visualize:

- a unifying concept of uncertainty
- two basic fitting machines

These two ideas help unlock the understanding of statistics with intuitive concepts that are based on the foundation laid by mathematical statistics.

Statistics is to science what accounting is to business. It is the craft of weighing and balancing observational evidence. Statistical tests are like credibility audits. But statistical tools can do more than that. They are instruments of discovery that can show unexpected things about data and lead to interesting new ideas. Before using these powerful tools, you need to understand a bit about how they work.

Uncertainty, a Unifying Concept

When you do accounting, you total money amounts to get summaries. When you look at scientific observations in the presence of uncertainty or noise, you need some statistical measurement to summarize the data. Just as money is additive, uncertainty is additive if you choose the right measure for it.

The best measure is not the direct probability because to get a joint probability, you have to assume that the observations are independent and then multiply probabilities rather than add them. It is easier to take the log of each probability because then you can sum them and the total is the log of the joint probability.

However, the log of a probability is negative because it is the log of a number between 0 and 1. In order to keep the numbers positive, JMP uses the negative log of the probability. As the probability becomes smaller, its negative log becomes larger. This measure is called uncertainty, and it is measured in reverse fashion from probability.

In business, you want to maximize revenues and minimize costs. In science, you want to minimize uncertainty. Uncertainty in science plays the same role as cost plays in business. All statistical methods fit models such that uncertainty is minimized.

It is not difficult to visualize uncertainty. Just think of flipping a series of coins where each toss is independent. The probability of tossing a head is 0.5, and -log(0.5) is 1 for base 2 logarithms. The probability of tossing h heads in a row is as follows:

$$p = \left(\frac{1}{2}\right)^h$$

Solving for h produces the following:

$$h = -\log_2 p$$

You can think of the uncertainty of some event as the number of consecutive “head” tosses you have to flip to get an equally rare event.

Almost everything we do statistically has uncertainty, -log$p$, at the core. Statistical literature refers to uncertainty as *negative log-likelihood*.

The Two Basic Fitting Machines

An amazing fact about statistical fitting is that most of the classical methods reduce to using two simple machines, the spring and the pressure cylinder.

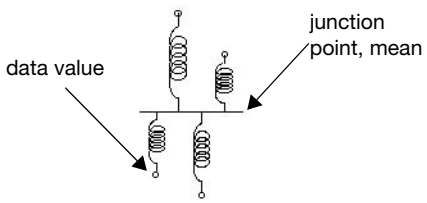
Springs

First, springs are the machine of fit for a continuous response model (Farebrother 1987). Suppose that you have *n* points and that you want to know the expected value (mean) of the points. Envision what happens when you lay the points out on a scale and connect them to a common junction with springs (see Figure A.7). When you let go, the springs wiggle the junction point up and down and then bring it to rest at the mean. This is what must happen according to physics.

If the data are normally distributed with a mean at the junction point where springs are attached, then the physical energy in each point's spring is proportional to the uncertainty of the data point. All you have to do to calculate the energy in the springs (the uncertainty) is to compute the sum of squared distances of each point to the mean.

To choose an estimate that attributes the least uncertainty to the observed data, the spring settling point is chosen as the estimate of the mean. That is the point that requires the least energy to stretch the springs and is equivalent to the least squares fit.

Figure A.7 Connect Springs to Data Points

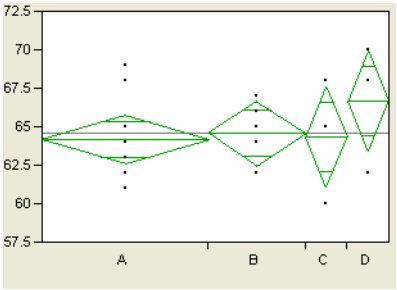


That is how you fit one mean or fit several means. That is how you fit a line, or a plane, or a hyperplane. That is how you fit almost any model to continuous data. You measure the energy or uncertainty by the sum of squares of the distances that you must stretch the springs.

Statisticians put faith in the normal distribution because it is the one that requires the least faith. It is, in a sense, the most random. It has the most non-informative shape for a distribution. It is the one distribution that has the most expected uncertainty for a given variance. It is the distribution whose uncertainty is measured in squared distance. In many cases it is the limiting distribution when you have a mixture of distributions or a sum of independent quantities. It is the distribution that leads to test statistics that can be measured fairly easily.

When the fit is constrained by hypotheses, you test the hypotheses by measuring this same spring energy. Suppose you have responses from four different treatments in an experiment, and you want to test if the means are significantly different. First, envision your data plotted in groups as shown in Figure A.8, but with springs connected to a separate mean for each treatment. Then exert pressure against the spring force to move the individual means to the common mean. Presto! The amount of energy that constrains the means to be the same is the test statistic that you need. That energy is the main ingredient in the *F* test for the hypothesis that tests whether the means are the same.

Figure A.8 A Oneway Plot for a Continuous Response Variable



Pressure Cylinders

What if your response is categorical instead of continuous? For example, suppose that the response is the country of origin for a sample of cars. For your sample, there are probabilities for the three response levels, American, European, and Japanese. You can set these probabilities for country of origin to some estimate and then evaluate the uncertainty in your data. This uncertainty is found by summing the negative logs of the probabilities of the responses given by the data. It is written as follows:

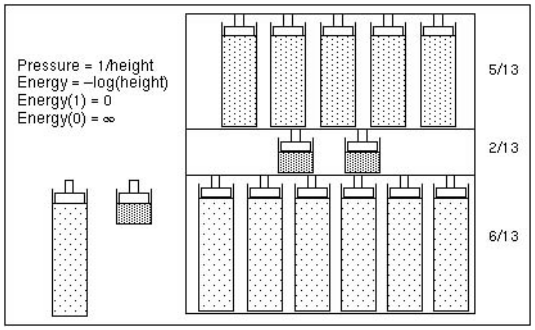
$$H = \sum h_{y(i)} = -\sum \log p_{y(i)}$$

The idea of springs illustrates how a mean is fit to continuous data. When the response is categorical, statistical methods estimate the response probabilities directly and choose the estimates that minimize the total uncertainty of the data. The probability estimates must be nonnegative and sum to 1. You can picture the response probabilities as the composition along a scale whose total length is 1. For each response observation, load into its response area a gas pressure cylinder, for example, a tire pump. Let the partitions between the response levels vary until an equilibrium of lowest potential energy is reached. The sizes of the partitions that result then estimate the response probabilities.

Figure A.9 shows what the situation looks like for a single category such as the medium size cars (see the mosaic column from Carpoll.jmp labeled medium in Figure A.10). Suppose there are thirteen responses (cars). The first level (American) has six responses, the next has two, and the last has five responses. The response probabilities become 6/13, 2/13, and 5/13,

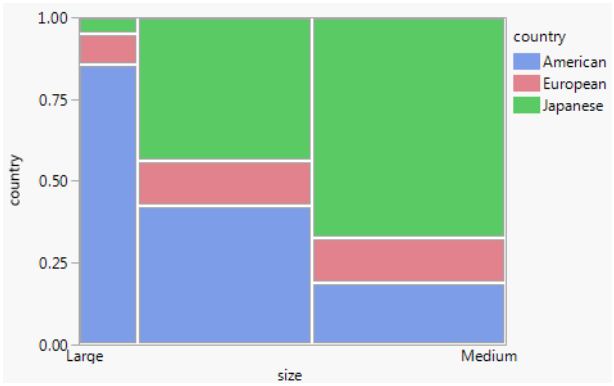
respectively, as the pressure against the response partitions balances out to minimize the total energy.

Figure A.9 Effect of Pressure Cylinders in Partitions



As with springs for continuous data, you can divide your sample by some factor and fit separate sets of partitions. Then test that the response rates are the same across the groups by measuring how much additional energy you need to push the partitions to be equal. Imagine the pressure cylinders for car origin probabilities grouped by the size of the car. The energy required to force the partitions in each group to align horizontally tests whether the variables have the same probabilities. Figure A.10 shows these partitions.

Figure A.10 A Mosaic Plot for Categorical Data



Likelihood, AICc, and BIC

Many statistical models in JMP are fit using a technique called *maximum likelihood*. This technique seeks to estimate the parameters of a model, which we denote generically by β, by maximizing the likelihood function. The likelihood function, denoted L(β), is the product of the probability density functions (or probability mass functions for discrete distributions)

evaluated at the observed data values. Given the observed data, maximum likelihood estimation seeks to find values for the parameters, β, that maximize L(β).

Rather than maximize the likelihood function L(β), it is more convenient to work with the negative of the natural logarithm of the likelihood function, *-Log* L(β). The problem of maximizing L(β) is reformulated as a minimization problem where you seek to minimize the negative log-likelihood (*-LogLikelihood* = *-Log* L(β)). Therefore, smaller values of the negative log-likelihood or twice the negative log-likelihood (*-2LogLikelihood*) indicate better model fits.

You can use the value of negative log-likelihood to choose between models and to conduct custom hypothesis tests that compare models fit using different platforms in JMP. This is done through the use of likelihood ratio tests. One reason that *-2LogLikelihood* is reported in many JMP platforms is that the distribution of the difference between the full and reduced model *-2LogLikelihood* values is asymptotically Chi-square. The degrees of freedom associated with this likelihood ratio test are equal to the difference between the numbers of parameters in the two models (Wilks 1938).

The corrected Akaike's Information Criterion (AICc) and the Bayesian Information Criterion (BIC) are information-based criteria that assess model fit. Both are based on *-2LogLikelihood*.

AICc is defined as follows:

$$\text{AICc} = -2LogLikelihood + 2k + 2k(k+1)/(n-k-1)$$

where k is the number of estimated parameters in the model and n is the number of observations used in the model. This value can be used to compare various models for the same data set to determine the best-fitting model. The model having the smallest value, as discussed in Akaike (1974), is usually the preferred model.

BIC is defined as follows:

$$\text{BIC} = -2LogLikelihood + k\ln(n)$$

where k is the number of estimated parameters in the model and n is the number of observations used in the model. When comparing the BIC values for two models, the model with the smaller BIC value is considered better.

In general, BIC penalizes models with more parameters more than AICc does. For this reason, it leads to choosing more parsimonious models, that is, models with fewer parameters, than does AICc. For a detailed comparison of AICc and BIC, see Burnham and Anderson (2004).

Power Calculations

The next sections give formulas for computing the least significant number (LSN), least significant value (LSV), power, and adjusted power. With the exception of LSV, these

computations are provided for each effect, and for a collection of user-specified contrasts (under Custom Test and LS Means Contrast). LSV is computed only for a single linear contrast. In the details below, the *hypothesis* refers to the collection of contrasts of interest.

Computations for the LSN

The LSN solves for N in the equation:

$$\alpha = 1 - FDist\left[\frac{N\delta^2}{df_{Hyp}\sigma^2}, df_{Hyp}, N - df_{Hyp} - 1\right]$$

where

FDist is the cumulative distribution function of the central F distribution

df_{Hyp} represents the degrees of freedom for the hypothesis

σ^2 is the error variance

δ^2 is the squared effect size

For retrospective analyses, δ^2 is estimated by the sum of squares for the hypothesis divided by n, the size of the current sample. If the test is for an effect, then $\delta^2$is estimated by the sum of squares for that effect divided by the number of observations in the current study. For retrospective studies, the error variance σ^2 is estimated by the mean square error. These estimates, along with an α value of 0.05, are entered into the Power Details window as default values.

When you are conducting a prospective analysis to plan a future study, consider determining the sample size that will achieve a specified power (see "Computations for the Power" on page 553.)

Computations for the LSV

The LSV is computed only for a single linear contrast.

Test of a Single Linear Contrast

Consider the one-degree-freedom test $L\beta = 0$, where L is a row vector of constants. The test statistic for a t test for this hypothesis is:

$$\frac{Lb}{s\sqrt{L(X'X)^{-1}L'}}$$

where s is the root mean square error. We reject the hypothesis at significance level α if the absolute value of the test statistic exceeds the 1 - $\alpha/2$ quantile of the t distribution, $t_{1-\alpha/2}$, with degrees of freedom equal to those for error.

To find the least significant value, denoted $(Lb)^{LSV}$, we solve for Lb:

$$(Lb)^{\mathrm{LSV}} = t_{1-\alpha/2}s\sqrt{L(X'X)^{-1}L'}$$

Test of a Single Parameter

In the special case where the linear contrast tests a hypothesis setting a single β_i equal to 0, this reduces to the following:

$$b_i^{\mathrm{LSV}} = t_{1-\alpha/2}s\sqrt{(X'X)_{ii}^{-1}} = t_{1-\alpha/2}StdError(b_i)$$

Test of a Difference in Means

In a situation where the test of interest is a comparison of two group means, the literature talks about the *least significant difference (LSD)*. In the special case where the model contains only one nominal variable, the formula for testing a single linear contrast reduces to the formula for the LSD:

$$\mathrm{LSD} = t_{1-\alpha/2}s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

However, in JMP, the parameter associated with a level for a nominal effect measures the difference between the mean of that level and the mean for all levels. So, the LSV for such a comparison is half the LSD for the differences of the means.

Note: If you are testing a contrast across the levels of a nominal effect, keep in mind how JMP codes nominal effects. Namely, the parameter associated with a given level measures the difference to the average for all levels.

Computations for the Power

Suppose that you are interested in computing the power of a test of a linear hypothesis, based on significance level α and a sample size of N. You want to detect an effect of size δ.

To calculate the power, begin by finding the critical value for an α-level F test of the linear hypothesis. This is given by solving for F_C in the equation

$$\alpha = 1\text{–}FDist[F_C, df_{Hyp}, N - df_{Model} - 1]$$

Here, df_{Hyp} represents the degrees of freedom for the hypothesis, df_{Model} represents the degrees of freedom for the model, and N is the proposed (or actual) sample size.

Then calculate the noncentrality parameter associated with the desired effect size. The noncentrality parameter is as follows:

$$\lambda = (N\delta^2)/\sigma^2$$

where σ^2 is a proposed (or estimated) value of the error variance.

Given an effect of size δ, the test statistic has a non-central F distribution, where the distribution function is denoted *FDist* below, with noncentrality parameter λ. To obtain the power of your test, calculate the probability that the test statistic exceeds the critical value:

$$\text{Power} = 1\text{–}FDist\left[F_C, df_{Hyp}, N - df_{Model} - 1, \frac{N\delta^2}{\sigma^2}\right]$$

In obtaining retrospective power for a study with n observations, JMP estimates the noncentrality parameter $\lambda = (n\delta^2)/\sigma^2$ by $\hat{\lambda} = SS_{Hyp}/\hat{\sigma}^2$, where SS_{Hyp} represents the sum of squares due to the hypothesis.

Computations for the Adjusted Power

The adjusted power calculation (Wright and O'Brien 1988) is relevant only for retrospective power analysis. Adjusted power calculates power using a noncentrality parameter estimate that has been adjusted to remove the positive bias that occurs when parameters are simply replaced by their sample estimates.

The estimate of the noncentrality parameter, λ, obtained by estimating δ and σ by their sample estimates, appears as follows:

$$\hat{\lambda} = SS_{Hyp}/MSE$$

Wright and O'Brien (1988) explain that an unbiased estimate of the noncentrality parameter is given by the following:

$$[\hat{\lambda}(df_{Error} - 2)/df_{Error}] - df_{Hyp} = \frac{\hat{\lambda}(N - df_{Model} - 1 - 2)}{N - df_{Model} - 1} - df_{Hyp}$$

The expression on the right illustrates the calculation of the unbiased noncentrality parameter when a sample size N, different from the study size n, is proposed for a retrospective power

analysis. Here, df_{Hyp} represents the degrees of freedom for the hypothesis and df_{Model} represents the degrees of freedom for the whole model.

Unfortunately, this adjustment to the noncentrality estimate can lead to negative values. Negative values are set to zero, reintroducing some slight bias. The adjusted noncentrality estimate is

$$\hat{\lambda}_{\text{adj}} = \text{Max}\left[0, \frac{\hat{\lambda}(N - df_{Model} - 1 - 2)}{N - df_{Model} - 1} - df_{Hyp}\right]$$

The adjusted power is

$$\text{Power}_{\text{adj}} = 1 - FDist[\text{F}_{\text{C}}, df_{Hyp}, N - df_{Model} - 1, \hat{\lambda}_{\text{adj}}]$$

Confidence limits for the noncentrality parameter are constructed as described in Dwass (1955):

$$\text{Lower CL for } \lambda = \text{Max}\left[0, [(\sqrt{SS_{\text{Hyp}}/MSE}) - \sqrt{df_{Hyp}\text{F}_{\text{C}}}]^2\right]$$

$$\text{Upper CL for } \lambda = [(\sqrt{SS_{\text{Hyp}}/MSE}) - \sqrt{df_{Hyp}\text{F}_{\text{C}}}]^2$$

Confidence limits for the power are obtained by substituting these confidence limits for λ into the following equation:

$$\text{Power} = 1 - FDist[\text{F}_{\text{C}}, df_{Hyp}, N - df_{Model} - 1, \lambda]$$

Inverse Prediction with Confidence Limits

Inverse prediction estimates a value of an independent variable from a response value. In bioassay problems, inverse prediction with confidence limits is especially useful. In JMP, you can request inverse prediction estimates for continuous and binary response models. If the response is continuous, you can request confidence limits for an individual response or an expected response.

The confidence limits are computed using Fieller's theorem (1954), which is based on the following logic. The goal is predicting the value of a single regressor and its confidence limits given the values of the other regressors and the response.

- Let **b** estimate the parameters β so that we have **b** distributed as $N(\beta,V)$.
- Let **x** be the regressor values of interest, with the i^{th} value to be estimated.
- Let **y** be the response value.

We desire a confidence region on the value of **x**[i] such that $\beta' x = y$ with all other values of **x** given.

The inverse prediction is

$$x[i] = \frac{y - \beta'_{(i)} x_{(i)}}{\beta[i]}$$

where the subscript (*i*) in parentheses indicates that the i^{th} component is omitted. A confidence interval can be formed from the relation

$$(y - b'x)^2 < t^2 x'Vx$$

where *t* is the *t* value for the specified confidence level.

The equation

$$(y - b'x)^2 - t^2 x'Vx = 0$$

can be written as a quadratic in terms of $z = x[i]$:

$$gz^2 + hz + f = 0$$

where

$$g = b[i]^2 - t^2 V[i, i]$$

$$h = -2yb[i] + 2b[i]b'_{(i)}x_{(i)} - 2t^2 V[i, (i)]'x_{(i)}$$

$$f = y^2 - 2yb'_{(i)}x_{(i)} + (b'_{(i)}x_{(i)})^2 - t^2 x_{(i)}'V_{(i)}x_{(i)}$$

Depending on the values of g, h, and f, the set of values satisfying the inequality, and hence the confidence interval for the inverse prediction, can have a number of forms:

- an interval of the form (ϕ_1, ϕ_2), where $\phi_1 < \phi_2$
- two disjoint intervals of the form $(-\infty, \phi_1) \cup (\phi_2, \infty)$, where $\phi_1 < \phi_2$
- the entire real line, $(-\infty, \infty)$
- only one of $-(\infty, \phi)$ or (ϕ, ∞)

In the case where the Fieller confidence interval is the entire real line, Wald intervals are presented.

Note: The Fit Y by X logistic platform and the Fit Model Nominal Logistic personalities use *t* values when computing confidence intervals for inverse prediction. The Fit Model Generalized Linear Model personality, as well as PROC PROBIT in SAS/STAT, use z values, which give different results.

Platforms That Support Validation

This appendix lists the types of crossvalidation available in each platform.

Parent Platform	Platform	Use Excluded Rows as Validation Holdback	Random Validation Holdback	K-Fold Cross-Validation	Validation Role Column
Fit Model	Fit Least Squares	No	No	No	Yes (only for model evaluation)
Fit Model	Forward Stepwise Regression	No	No	Yes (continuous response models only)	Yes
Fit Model	Logistic Regression	No	No	No	Yes (only for model evaluation)
Fit Model	Generalized Regression	No	Yes	Yes	Yes
Fit Model	Partial Least Squares	No	Yes	Yes	Yes
Partition	Decision Tree	Yes	Yes	Yes	Yes
Partition	Bootstrap Forest	Yes	Yes	No	Yes

Parent Platform	Platform	Use Excluded Rows as Validation Holdback	Random Validation Holdback	K-Fold Cross-Validation	Validation Role Column
Partition	Boosted Tree	Yes	Yes	No	Yes
Partition	K Nearest Neighbors	Yes	Yes	No	Yes
Partition	Naive Bayes	Yes	Yes	No	Yes
Neural	Neural	Yes	Yes	Yes	Yes

Appendix B

References

Aitken, M. (1987). "Modelling Variance Heterogeneity in Normal Regression Using GLIM." *Journal of the Royal Statistical Society, Series C* 36:332–339.

Akaike, H. (1974). "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* AC-19:716–723.

Anderson, T. W. (1958). *An Introduction to Multivariate Statistical Analysis*. New York: John Wiley & Sons.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley & Sons.

Benjamini, Y., and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57:289–300.

Box, G. E. P. (1954). "Some Theorems on Quadratic Forms Applied in the Study of Analysis of Variance Problems, Part 2: Effects of Inequality of Variance and of Correlation between Errors in the Two-Way Classification." *Annals of Mathematical Statistics* 25:484–498.

Box, G. E. P., and Cox, D. R. (1964). "An Analysis of Transformations." *Journal of the Royal Statistical Society, Series B* 26:211–243.

Box, G. E. P., and Meyer, R. D. (1986). "An Analysis for Unreplicated Fractional Factorials." *Technometrics* 28:11–18.

Box, G. E. P., and Meyer, R. D. (1993). "Finding the Active Factors in Fractionated Screening Experiments." *Journal of Quality Technology* 25:94–105.

Burnham, K. P., and Anderson, D. R. (2004). "Multimodel Inference: Understanding AIC and BIC in Model Selection." *Sociological Methods and Research* 33:261–304.

Burnham, K. P., Andersen, D. R., and Huyvaert, K. P. (2011). "AIC Model Selection and Multimodel Inference in Behavioral Ecology: Some Background, Observations, and Comparisons." *Behavioral Ecology and Sociobiology* 65:23–35.

Candes, E., and Tao, T. (2007). "The Dantzig Selector: Statistical Estimation when p is Much Larger than n." *The Annals of Statistics* 35:2313–2351.

Carroll, R. J., and Ruppert, D. (1988). *Transformation and Weighting in Regression*. London: Chapman & Hall.

Chilès, J.-P., and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. 2nd ed. New York: John Wiley & Sons.

Cobb, G. W. (1998). *Introduction to Design and Analysis of Experiments*. New York: Springer-Verlag.

Cohen, J. (1977). *Statistical Power Analysis for the Behavioral Sciences*. New York: Academic Press.

Cole, J. W. L., and Grizzle, J. E. (1966). "Applications of Multivariate Analysis of Variance to Repeated Measures Experiments." *Biometrics* 22:810–828.

Conover, W. J. (1999). *Practical Nonparametric Statistics*. 3rd ed. New York: John Wiley & Sons.

Cook, R. D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. New York: Chapman & Hall.

Cook, R. D., and Weisberg, S. (1983). "Diagnostics for Heteroscedasticity in Regression." *Biometrika* 70:1–10.

Cornell, J. A. (1990). *Experiments with Mixtures*. 2nd ed. New York: John Wiley & Sons.

Cox, D. R. (1972). "Regression Models and Life-Tables." *Journal of the Royal Statistical Society, Series B* 34:187–220.

Cox, D. R., and Snell, E. J. (1989). *The Analysis of Binary Data*. 2nd ed. London: Chapman & Hall.

Cressie, N. A. C. (1993). *Statistics for Spatial Data*. Rev. ed. New York: John Wiley & Sons.

Daniel, C. (1959). "Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments." *Technometrics* 1:311–341.

Dunnett, C. W. (1955). "A Multiple Comparisons Procedure for Comparing Several Treatments with a Control." *Journal of the American Statistical Association* 50:1096–1121.

Dwass, M. (1955). "A Note on Simultaneous Confidence Intervals." *Annals of Mathematical Statistics* 26:146–147.

Efron, B. (1977). "The Efficiency of Cox's Likelihood Function for Censored Data." *Journal of the American Statistical Association* 72:557–565.

Farebrother, R. W. (1987). "Mechanical Representations of the L1 and L2 Estimation Problems." In *Statistical Data Analysis Based on L_1 Norm and Related Methods*, edited by Y. Dodge, 455–464. Amsterdam: North-Holland.

Fieller, E. C. (1954). "Some Problems in Interval Estimation." *Journal of the Royal Statistical Society, Series B* 16:175–185.

Firth, D. (1993). "Bias Reduction of Maximum Likelihood Estimates." *Biometrika* 80:27–38.

Fleming, T. R., and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: John & Sons.

Goodnight, J. H. (1978). *Tests of Hypotheses in Fixed Effects Linear Models*. Technical Report R–101, SAS Institute Inc., Cary, NC.

Goodnight, J. H., and Harvey, W. R. (1978). *Least-Squares Means in the Fixed-Effects General Linear Models*. Technical Report R-103, SAS Institute Inc., Cary, NC.

Goos, P., and Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach*. Chichester, UK: John Wiley & Sons.

Greenhouse, S. W., and Geisser, S. (1959). "On Methods in the Analysis of Profile Data." *Psychometrika* 32:95–112.

Harrell, F. E. (1986). "The LOGIST Procedure." In *SUGI Supplemental Library Guide, Version 5 Edition*. Cary, NC: SAS Institute Inc.

Harvey, A. C. (1976). "Estimating Regression Models with Multiplicative Heteroscedasticity." *Econometrica* 44:461–465.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag.

Hayter, A. J. (1984). "A Proof of the Conjecture That the Tukey-Kramer Method Is Conservative." *Annals of Statistics* 12:61–75.

Heinze, G., and Schemper, M. (2002). "A Solution to the Problem of Separation in Logistic Regression." *Statistics in Medicine* 21:2409–2419.

Hocking, R. R. (1985). *The Analysis of Linear Models*. Monterey, CA: Brooks/Cole.

Hoenig, J. M., and Heisey, D. M. (2001). "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis." *American Statistician* 55:19–24.

Hoerl, A. (1962). "Application of Ridge Analysis to Regression Problems." *Chemical Engineering Progress* 58:54–59.

Hoerl, A., and Kennard, R. (1970). "Ridge Regression: Biased Estimation for Non-orthogonal Problems." *Technometrics* 12:55–67.

Hsu, J. C. (1992). "The Factor Analytic Approach to Simultaneous Inference in the General Linear Model." *Journal of Computational and Graphical Statistics* 1:151–168.

Hsu, J. C. (1996). Multiple Comparisons: Theory and Methods. London: Chapman & Hall.

Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed. John Wiley & Sons.

Hui, F., Warton, D., and Foster, S. (2015). "Tuning Parameter Selection for the Adaptive Lasso Using ERIC." *Journal of the American Statistical Association* 110:262–269.

Huynh, H., and Feldt, L. S. (1970). "Conditions Under Which Mean Square Ratios in Repeated Measurements Designs Have Exact F-Distributions." *Journal of the American Statistical Association* 65:1582–1589.

Huynh, H., and Feldt, L. S. (1976). "Estimation of the Box Correction for Degrees of Freedom from Sample Data in the Randomized Block and Split Plot Designs." *Journal of Educational Statistics* 1:69–82.

Kackar, R. N., and Harville, D. A. (1984). "Approximations for Standard Errors of Estimators of Fixed and Random Effects in Mixed Linear Models." *Journal of the American Statistical Association* 79:853–862.

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

Kenward, M. G., and Roger, J. H. (1997). "Small Sample Inference for Fixed Effects from Restricted Maximum Likelihood." *Biometrics* 53:983–997.

Koenker, R., and Hallock, K. (2001). "Quantile Regression: An Introduction." *Journal of Economic Perspectives* 15:143–156.

Kramer, C. Y. (1956). "Extension of Multiple Range Tests to Group Means with Unequal Numbers of Replications." *Biometrics* 12:307–310.

Lenth, R. V. (1989). "Quick and Easy Analysis of Unreplicated Factorials." *Technometrics* 31:469–473.

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., and Schabenberger, O. (2006). *SAS for Mixed Models*. 2nd ed. Cary, NC: SAS Institute Inc.

Mallows, C. L. (1973). "Some Comments on C_p." *Technometrics* 15:661–675.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate Analysis*. London: Academic Press.

McClave, J. T., and Dietrich, F. H. (1988). *Statistics*. San Francisco: Dellen.

McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.

McCulloch, C. E., Searle, S. R., and Neuhaus, J. M. (2008). *Generalized, Linear, and Mixed Models*. New York: John Wiley & Sons.

Meeker, W. Q., and Escobar, L. A. (1998). *Statistical Methods for Reliability Data*. New York: John Wiley & Sons.

Miller, A. J. (1990). *Subset Selection in Regression*. New York: Chapman & Hall.

Montgomery, D. C. (1991). "Using Fractional Factorial Designs for Robust Process Development." *Quality Engineering* 3:193–205.

Muller, K. E., and Barton, C. N. (1989). "Approximate Power for Repeated-Measures ANOVA Lacking Sphericity." *Journal of the American Statistical Association* 84:549–555. Also see "Correction to 'Approximate Power for Repeated-Measures ANOVA Lacking Sphericity'," *Journal of the American Statistical Association* 86 (1991): 255–256.

Nagelkerke, N. J. D. (1991). "A Note on a General Definition of the Coefficient of Determination." *Biometrika* 78:691–692.

Nelder, J. A., and Wedderburn, R. W. M. (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A* 135:370–384.

Nelson, F. D. (1976). "On a General Computer Algorithm for the Analysis of Models with Limited Dependent Variables." *Annals of Economic and Social Measurement* 5:493–509.

Nelson, P. R., Wludyka, P. S., and Copeland, K. A. F. (2005). *The Analysis of Means: A Graphical Method for Comparing Means, Rates, and Proportions*. Philadelphia: SIAM.

Patterson, H. D., and Thompson, R. (1974). "Maximum Likelihood Estimation of Components of Variance." In *Proceedings of the Eighth International Biometric Conference*, 197–207. Washington, DC: International Biometric Society.

Portnoy, S., and Koenker, R. (1997). "The Gaussian Hare and the Laplacian Tortoise: Computation of Squared-Error vs. Absolute-Error Estimators." *Statistical Science* 12:279–300.

Rawlings, J. O. (1988). *Applied Regression Analysis: A Research Tool*. Pacific Grove, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Ries, P. N., and Smith, H. (1963). "The Use of Chi-Square for Preference Testing in Multidimensional Problems." *Chemical Engineering Progress* 59:39–43.

Sall, J. P. (1990). "Leverage Plots for General Linear Hypotheses." *American Statistician* 44:308–315.

SAS Institute Inc. (2017a). "The GENMOD Procedure." In *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/onlinedoc/stat/143/genmod.pdf

SAS Institute Inc. (2017b). "The GLM Procedure." In *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/onlinedoc/stat/143/glm.pdf

SAS Institute Inc. (2017c). "Introduction to Statistical Modeling with SAS/STAT Software." In *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/onlinedoc/stat/143/intromod.pdf

SAS Institute Inc. (2017d). "The MIXED Procedure." In *SAS/STAT 14.3 User's Guide*. Cary, NC: SAS Institute Inc. http://support.sas.com/documentation/onlinedoc/stat/143/mixed.pdf

Satterthwaite, F. E. (1946). "An Approximate Distribution of Estimates of Variance Components." *Biometrics Bulletin* 2:110–114.

Scheffé, H. (1958). "Experiments with Mixtures." *Journal of the Royal Statistical Society, Series B* 20:344–360.

Schuirmann, D. J. (1987). "A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability." *Journal of Pharmacokinetics and Biopharmaceutics* 15:657–680.

Searle, S. R., Casella, G., and McCulloch, C. E. (1992). *Variance Components*. New York: John Wiley & Sons.

Seber, G. A. F. (1984). *Multivariate Observations*. New York: John Wiley & Sons.

Singer, J. D. (1998). "Using SAS PROC MIXED to Fit Multilevel Models, Hierarchical Models, and Individual Growth Models." *Journal of Educational and Behavioral Statistics* 23:323–355.

Snedecor, G. W., and Cochran, W. G. (1967). Statistical Methods. 6th ed. Ames: Iowa State University Press.

Spiller, S. A., Fitzsimons, G. J., Lynch, J. G., Jr., and McClelland, G. (2013). "Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression." *Journal of Marketing Research* 50:277–288.

Stone, C., and Koo, C. Y. (1985). "Additive Splines in Statistics." In *Proceedings of the Statistical Computing Section,* 45–48. Alexandria, VA: American Statistical Association.

Sullivan, L. M., Dukes, K. A., and Losina, E. (1999). "An Introduction to Hierarchical Linear Modelling." *Statistics in Medicine* 18:855–888.

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society, Series B* 58:267–288.

Tukey, J. W. (1953). "The Problem of Multiple Comparisons." In *Multiple Comparisons*, 1948–1983, edited by H. I. Braun, vol. 8 of *The Collected Works of John W. Tukey* (published 1994), 1–300. London: Chapman & Hall. Unpublished manuscript.

Walker, S. H., and Duncan, D. B. (1967). "Estimation of the Probability of an Event as a Function of Several Independent Variables." *Biometrika* 54:167–179.

Westfall, P. H., Tobias, R. D., and Wolfinger, R. D. (2011). *Multiple Comparisons and Multiple Tests Using SAS.* 2nd ed. Cary, NC: SAS Institute Inc.

Wilks, S. S. (1938). "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *Annals of Mathematical Statistics*. 9:60–62.

Wolfinger, R. D., Tobias, R. D., and Sall, J. (1994). "Computing Gaussian Likelihoods and Their Derivatives for General Linear Mixed Models." *SIAM Journal on Scientific Computing* 15:1294–1310.

Wright, S. P., and O'Brien, R. G. (1988). "Power Analysis in an Enhanced GLM Procedure: What it Might Look Like." In *Proceedings of the Thirteenth Annual SAS Users Group International Conference*, 1097–1102. Cary, NC: SAS Institute Inc. http://www.sascommunity.org/sugi/SUGI88/Sugi-13-220%20Wright%20OBrien.pdf

Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties." *Journal of the American Statistical Association* 101:1418–1429.

Zou, H., and Hastie, T. (2005). "Regularization and Variable Selection via the Elastic Net." *Journal of the Royal Statistical Society, Series B* 67:301–320.

Index

Fitting Linear Models

Symbols

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q

R

S

T

U

V

W-Z