Statistical Discovery.™ From SAS.

Version 13

Consumer Research

"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. 2016. *JMP® 13 Consumer Research*. Cary, NC: SAS Institute Inc.

JMP® 13 Consumer Research

ISBN 978-1-62960-472-5 (Hardcopy)
ISBN 978-1-62960-568-5 (EPUB)
ISBN 978-1-62960-569-2 (MOBI)

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

September 2016

Technology License Notices

- 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 licence 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

Contents

Consumer Research

Chapter 1

Learn about JMP

Documentation and Additional Resources

This chapter includes the following information:

- book conventions
- JMP documentation
- JMP Help
- additional resources, such as the following:
  - other JMP documentation
  - tutorials
  - indexes
  - Web resources
  - technical support options

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
  - script output
- Features that are for JMP Pro only are noted with the JMP Pro icon JMP PRO. For an overview of JMP Pro features, visit http://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:

  http://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.
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.
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.

Document Title	Document Purpose	Document Content
Essential Graphing	Find the ideal graph for your data.	Describes these Graph menu platforms: • Graph Builder • Overlay Plot • Scatterplot 3D • Contour Plot • Bubble Plot • Parallel Plot • Cell Plot • Treemap • Scatterplot Matrix • Ternary Plot • Chart 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.
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.

Document Title	Document Purpose	Document Content
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 • Gaussian Process • Time Series • Matched Pairs Describes these Analyze > Screening menu platforms: • Response Screening • Process Screening • Predictor Screening • Association Analysis 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 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

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 • Multiple Correspondence Analysis • Multidimensional Scaling • Factor Analysis • Choice • MaxDiff • Uplift • Item 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.

Document Title	Document Purpose	Document Content
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 http://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 (see "Tutorials" on page 24)
- Sample data (see "Sample Data Tables" on page 24)
- Indexes (see "Learn about Statistical and JSL Terms" on page 24)
- Tip of the Day (see "Learn JMP Tips and Tricks" on page 24)
- Web resources (see "JMP User Community" on page 25)
- JMPer Cable technical publication (see "JMPer Cable" on page 25)
- Books about JMP (see "JMP Books by Users" on page 26)
- JMP Starter (see "The JMP Starter Window" on page 26)

- Teaching Resources (see "Sample Data Tables" on page 24)

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\13\Samples\Data

On Macintosh: \Library\Application Support\JMP\13\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. See the *Using JMP* book for details.

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/.

JMPer Cable

The JMPer Cable is a yearly technical publication targeted to users of JMP. The JMPer Cable is available on the JMP website:

http://www.jmp.com/about/newsletters/jmpercable/

JMP Books by Users

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

http://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 http://www.jmp.com/support, including the technical support phone number.

Chapter 2

Introduction to Consumer Research

Overview of Customer and Behavioral Research Methods

You already collect information about how customers use a product or service or how satisfied they are with your offerings. The resulting insight lets you create better products and services, happier customers, and more revenue for your organization.

JMP now includes a full suite of tools for performing customer and consumer research. In the past, you might have had to use one product for consumer research work and JMP for design of experiments. Now you can do both types of analyses using a single product, for a more efficient use of your most precious resource: your time. Tools for performing these statistical analyses are now located in one convenient place: the Consumer Research menu. Use the following platforms to analyze your data:

- The Categorical platform supports survey analysis with questions in multiple formats, allowing for both detailed and compact reporting. You can also analyze multiple response questions, where your survey includes questions for which respondents can choose more than one answer. You can output the results in crosstab report tables, use share and frequency charts, view mean scores across responses, and perform tests and comparisons. And when you are finished, you can easily output the completed analysis tables. For more information, see Chapter 3, "Categorical Response Analysis".
- The Multiple Correspondence Analysis (MCA) platform takes multiple categorical variables and seeks to identify associations between levels of those variables. MCA is frequently used in the social sciences particularly in France and Japan. It can be used in survey analysis to identify question agreement. For more information, see Chapter 4, "Multiple Correspondence Analysis".
- The Multidimensional Scaling (MDS) platform enables you to create a visual representation of the pattern of proximities (similarities, dissimilarities, or distances) among a set of objects. For more information, see Chapter 5, "Multidimensional Scaling".
- The Factor Analysis platform enables you to discover simple arrangements in the pattern of relationships among variables. It seeks to discover if the observed variables can be explained in terms of a much smaller number of variables or factors. By using factor analysis, you can determine the number of factors that influence a set of measured, observed variables, and the strength of the relationship between each factor and each variable. For more information, see Chapter 6, "Factor Analysis".
- The Choice platform is designed for use in market research experiments, where the ultimate goal is to discover the preference structure of consumers. Then, this information

is used to design products or services that have the attributes most desired by consumers. For more information, see Chapter 7, “Choice Models”.

- The MaxDiff platform is an alternative to using standard preference scales to determine the relative importance of items being rated. A MaxDiff model forces respondents to report their most and least preferred options, thereby forcing respondents to rank options in terms of preference. For more information, see Chapter 8, “MaxDiff”.
- The Uplift platform enables you to maximize the impact of your marketing budget by sending offers only to individuals who are likely to respond favorably, even when you have large data sets and many possible behavioral or demographic predictors. You can use uplift models to make such predictions. This method has been developed to help optimize marketing decisions, define personalized medicine protocols, or, more generally, to identify characteristics of individuals who are likely to respond to some action. For more information, see Chapter 9, “Uplift Models”.
- The Item Analysis platform provides a method of scoring tests. Based on Item Response Theory (IRT), the platform helps analyze the design, analysis, and scoring of tests, questionnaires, and other tools that measure abilities, attitudes, and other variables. Although classical test theory methods have been widely used for a century, IRT provides a better and more scientifically based scoring procedure. For more information, see Chapter 10, “Item Analysis”.

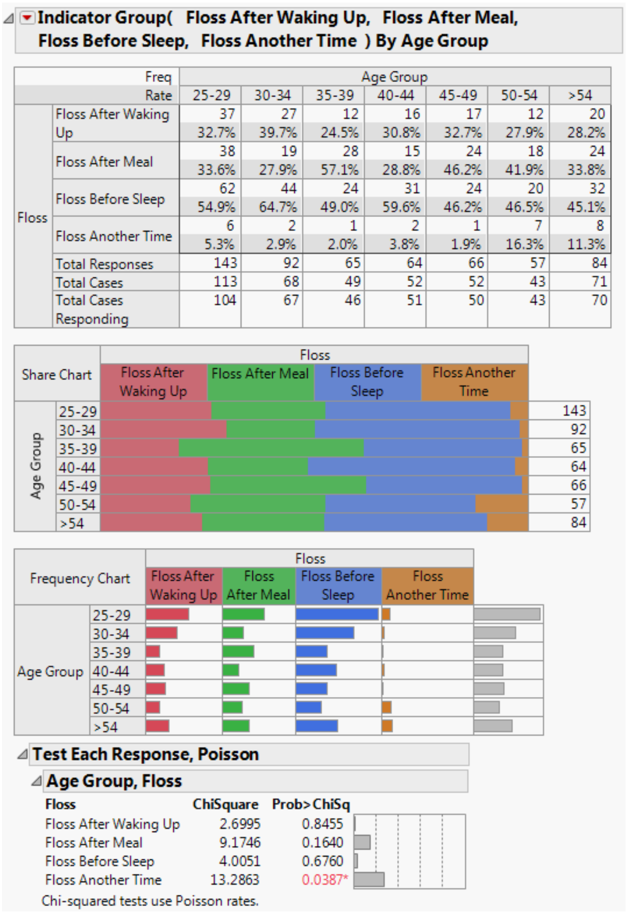
Chapter 3

Categorical Response Analysis

Analyzing Survey and Other Counting Data

The Categorical platform tabulates and summarizes categorical response data, including multiple response data, and calculates test statistics. The strength of the Categorical platform is that it can handle responses in a wide variety of formats without needing to reshape the data. It is designed to handle survey and other categorical response data, such as defect records, side effects, and so on.

Figure 3.1 Categorical Analysis Example



Floss	Freq / Rate	25-29	30-34	35-39	40-44	45-49	50-54	>54
Floss After Waking Up	Freq	37	27	12	16	17	12	20
	Rate	32.7%	39.7%	24.5%	30.8%	32.7%	27.9%	28.2%
Floss After Meal	Freq	38	19	28	15	24	18	24
	Rate	33.6%	27.9%	57.1%	28.8%	46.2%	41.9%	33.8%
Floss Before Sleep	Freq	62	44	24	31	24	20	32
	Rate	54.9%	64.7%	49.0%	59.6%	46.2%	46.5%	45.1%
Floss Another Time	Freq	6	2	1	2	1	7	8
	Rate	5.3%	2.9%	2.0%	3.8%	1.9%	16.3%	11.3%
Total Responses		143	92	65	64	66	57	84
Total Cases		113	68	49	52	52	43	71
Total Cases Responding		104	67	46	51	50	43	70

Floss	ChiSquare	Prob>ChiSq
Floss After Waking Up	2.6995	0.8455
Floss After Meal	9.1746	0.1640
Floss Before Sleep	4.0051	0.6760
Floss Another Time	13.2863	0.0387*

Chi-squared tests use Poisson rates.

Categorical Platform Overview

The Categorical platform can produce results from a rich variety of organizations of data, as reflected in the tabbed panels that enable you to specify the analyses that you want. The Categorical platform has capabilities similar to other platforms. The choice of platform depends on your focus, the shape of your data, and the desired level of detail. The strength of the Categorical platform is that it can handle responses in a wide variety of formats without needing to reshape the data. Table 3.1 shows several of JMP's analysis platforms and their strengths.

Table 3.1 Comparing JMP's Categorical Analyses

Platform	Specialty
Distribution	Separate, ungrouped categorical responses.
Fit Y By X: Contingency	Two-way situations, including chi-square tests, correspondence analysis, agreement.
Pareto Plot	Graphical analysis of multiple-response data, especially multiple-response defect data, with more rate tests than Fit Y By X.
Variability Chart: Attribute	Attribute gauge studies, with more detail on rater agreement.
Fit Model	Logistic categorical responses and generalized linear models.
Partition, Neural Net	Specific categorical response models.

Example of the Categorical Platform

This example uses the Consumer Preferences.jmp sample data table, which contains survey data on people's attitudes and opinions, and some questions concerning oral hygiene (source: Rob Reul, Isometric Solutions).

1. Select **Help > Sample Data Library** and open Consumer Preferences.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select I am working on my career and click **Responses** on the Simple tab.
4. Select Age Group and click **X, Grouping Category.**
5. Click **OK.**
6. Select **Crosstab Transposed** from the Categorical red triangle menu.

7. Select **Test Response Homogeneity** from the Categorical red triangle menu.

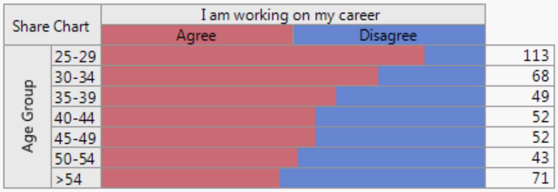
Figure 3.2 details the responses indicating that a respondent is currently working on his or her career and the age group. Of those responding positively, the highest majority working on their career were in the age group 25-29 at 84.1%. The highest majority of those responding oppositely were in the age group > 54 at 53.5%.

Figure 3.2 Survey Results by Age Group

Categorical

Responses(I am working on my career) By Age Group

	Freq / Share	Age Group						
		25-29	30-34	35-39	40-44	45-49	50-54	>54
I am working on my career	Agree	95 84.1%	49 72.1%	30 61.2%	29 55.8%	29 55.8%	22 51.2%	33 46.5%
	Disagree	18 15.9%	19 27.9%	19 38.8%	23 44.2%	23 44.2%	21 48.8%	38 53.5%
	Total Responses	113	68	49	52	52	43	71



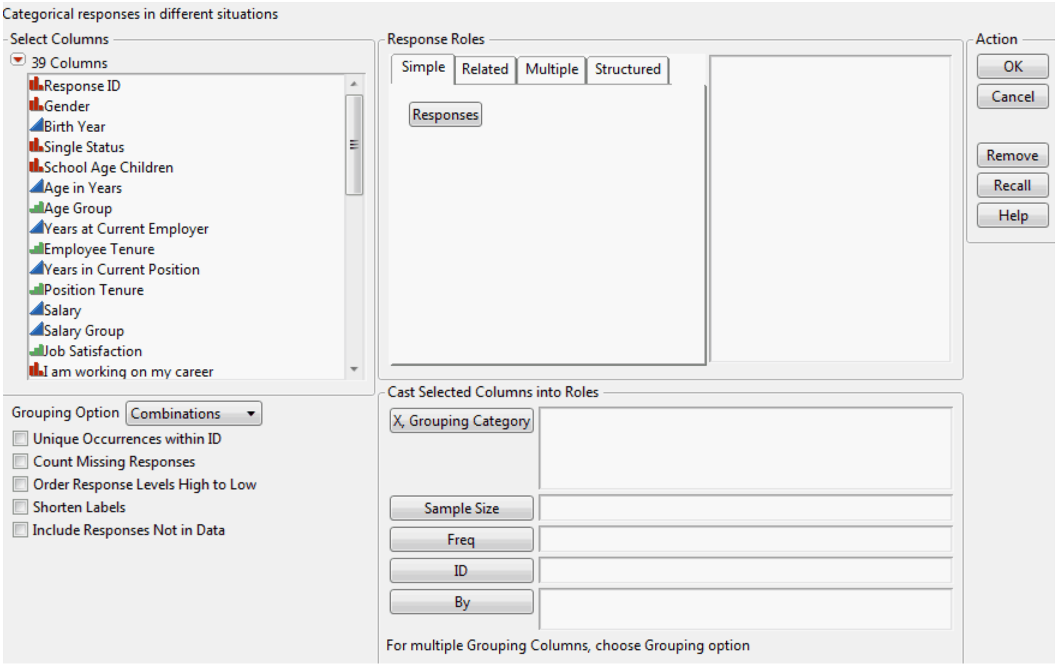
Test Response Homogeneity

Response Dimension Label	Sample Dimension Label	LR Chisq	LR PValue	Pearson Chisq	Pearson PValue
I am working on my career	Age Group	39.5874	<.0001*	37.4593	<.0001*

Launch the Categorical Platform

Launch the Categorical Platform by selecting **Analyze** > **Consumer Research** > **Categorical**.

Figure 3.3 Categorical Platform Launch Window



The launch window includes tabs for a variety of response roles (Simple, Related, and Multiple) and a Structured tab where you can create your own structured responses. The following sections describe the different response types and effects.

Response Roles

Use the response roles buttons within the tabs to choose selected columns as responses with specified roles. You can also drag column names to the response list. The response roles are summarized in Table .

Simple Tab

The default tab, Simple, contains a single button, Responses. This is appropriate for all basic analyses that do not have a special structure. You can drag column names from the Select Columns list to the Response list, or you can select columns and then click **Responses**. If a column has a Multiple Response column property or Multiple Response modeling type, JMP automatically changes the handling of the column to recognize this property.

Related Tab

The Related tab contains a set of response columns that all have the same type of categories in them:

Aligned Responses Performs the analysis like the default analysis, but shows the results more compactly by aligning the analyses side-by-side into one larger table.

Repeated Measures Indicates that the columns reflect responses made by the same individual at different times, and you are interested in the changes between the times. The Kish correction is used when there are overlapping samples. See Kish (1965, section 12.4).

Rater Agreement Is useful when each column is a rating for the same question, but by different individuals (raters) and you want to study how much the raters agree on their responses.

Multiple Tab

The Multiple tab is for multiple responses; when a question can involve checking off more than one choice. There are a variety of ways of storing multiple response data, so there are several buttons in the tab to accommodate the various means:

Multiple Response Means that you have several columns acting like fill-in-the-blank columns to specify the multiple responses.

Multiple Response by ID Indicates that you have several rows in a table corresponding to the multiple responses in one column, and the individuals are identified by an ID column.

Multiple Delimited Signifies that you have one column that has several responses in it separated by a comma.

Indicator Group Denotes that there is a column for each possible response, and each column is an indicator (for example, it has only two values, like 0 or 1).

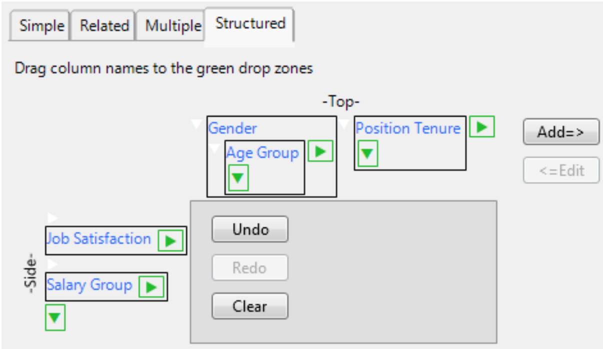
Response Frequencies Also has a column for each possible response, but has frequency counts instead of an indicator.

Free Text Is used for comment fields where the analysis counts the frequency of each word used. The Free Text option launches a Text Explorer report inside the Categorical report window. See the Text Explorer chapter in the *Basic Analysis* book.

Structured Tab

The Structured tab enables you to construct complex tables of descriptive statistics by dragging column names into green icon drop zones to create side-by-side and nested results. You can nest a variable within or beside another variable according to the structure that you want for the top and side of the table. Continue to drag columns, either beside or nested within another column, to specify the structure. For more information about structured reports, refer to “Structured Report Options” on page 59.

Figure 3.4 Structured Tab



To create a structured table:

1. Drag a column name to the green drop zone at the **Top** or **Side** of the table. The drop zone is highlighted in pink.
2. Add more variables by dragging a column name the appropriate green drop zone. You can also drag a column name between two columns.

   To remove a selection, click the selection and then select **Undo**.

   To add a selection back that you just removed, select **Redo**.

   To clear all of the selections, select **Clear**.
3. When you are finished creating the table, click **Add=>** to add the variables to the response list.

   To make a revision once you have added your selection to the response list, click the selection and then select **<=Edit**.
4. Complete the remainder of the launch window as necessary and click **OK**.

   The Categorical report window appears.
5. Should you want to make a change to the table, select **Relaunch Dialog** from the Categorical red triangle menu. The launch window reappears where you can make edits to your selections.

A few guidelines with Structured effects:

- Structured always assumes that the innermost terms on the Side are responses, and that all other terms are sample level grouping factors.
- You can analyze multiple response terms in the form of delimited multiple response columns, but you must indicate that it is a multiple response by having a Multiple Response column property or Multiple Response modeling type. Use the **Col Info** window to add this property, as needed.

Response Roles

Simple Tab

Responses Separate responses are in each column, resulting in a separate analysis for each column.

ID	Drink	Entrée
John	Coffee	Chicken
Jane	Tea	Veggie

Related Tab

Aligned Responses Responses share common categories across columns, resulting in better-organized reports.

ID	Coffee	Tea
John	Like	Dislike
Jane	Dislike	Like

Repeated Measures Aligned responses from an individual across different times or situations.

ID	Morning	Noon	Night
John	Coffee	Coffee	Water
Jane	Tea	Water	Tea

Rater Agreement Aligned responses from different raters evaluating the same unit, to study agreement across raters.

Sample	John	Jane
Sample1	Accept	Accept
Sample2	Accept	Reject
Sample3	Reject	Reject

Multiple Tab

Multiple Response Aligned responses, where multiple responses are entered across several columns, but treated as one grouped response.

ID	Drink 1	Drink 2	Drink 3
John	Coffee	Milk	Water
Jane	Tea	Water	

Multiple Response by ID Multiple responses across rows that have the same ID values.

ID	Drinks
John	Coffee
John	Milk
John	Water
Jane	Tea
Jane	Water

Multiple Delimited Several responses in a single cell, separated by commas.

ID	Drinks
John	Coffee, Milk, Water
Jane	Tea, Water

Indicator Group Binary responses across columns, like selected or deselected, yes or no, but all in a related group.

ID	Coffee	Milk	Tea	Water
John	Y	Y	N	Y
Jane	N	N	Y	Y

Response Frequencies Columns containing frequency counts for each response level, all in a related group.

Group	Coffee	Milk	Tea	Water
A	12	15	8	19
B	9	20	6	22

Free Text Counts the frequency of each word used in a comment field.

ID	Comment
John	I liked the coffee.
Jane	The juice was too sweet.

Structured Tab

Drag variables to green drop zones to create your own structured table.

Cast Selected Columns into Roles

The lower right panel of the Launch window has the following options:

X, Grouping Category Defines sample level groups to break the counts into. By default, it tabulates each combination of X values, but uses the Grouping Option below for other combinations.

Sample Size Defines the number of individual units in the group for which that frequency is applicable to, for multiple response roles with summarized data. For example, a Freq column might indicate 50 defects, where the sample size variable would reflect the defects for a batch of 100 units.

Freq Specifies the column containing frequency counts for each row for presummarized data.

ID Only required and used when Multiple Response by ID is selected.

By Identifies a variable to produce a separate analysis for each value that appears in the column.

Other Launch Window Options

Several launch options are presented in the lower left panel of the window that can be specified before the analysis. The options can also be selected later from the Categorical red triangle menu, and have the effect of rerunning the platform with the new setting. The default settings for some of the launch options can be changed in the Categorical red triangle menu. For more information, refer to "Set Preferences" on page 58.

Grouping Option Specifies whether you want to use the X columns individually or in a combination. Use this option only to specify more than one X (Grouping) column, and denote whether you want to treat the Xs one at a time, or in a fully nested grouping, or both. For example, if the X Columns are Region and Age Group, you can get separate tables for Response by Region and Response by Age Group (each individually) or get a nested table with each age group within each region (combinations), or both.

Combinations gives frequency results for combinations of the X variables.

Each Individually gives frequency results for each X variable individually.

Both gives frequency results for combinations of the X variables, and individually.

Unique Occurrences within ID Allows duplicate response levels within a subject to be counted only once. An ID variable must be specified.

The following options can be specified on the launch window as well as from the Categorical report red triangle menu. They are also available as Preference settings. For more information, refer to "Statistical Options" on page 43 and "Set Preferences" on page 58.

Count Missing Responses Changes the behavior to tabulate missing values as categories, while still excluding them from statistical comparisons. When you have missing values, this specifies whether you want to see them tabulated beside the nonmissing data, or just excluded. Missing values can be either standard (numeric NAN or character empty) or a code declared as missing with the column property Missing Value Codes. Note that if a column contains only missing values, the missing values are counted regardless of the state of this option.

Order Response Levels High to Low Changes the response order but keeps the X order from low to high. The default ordering is low to high. You can control the ordering with a column property (**Value Ordering**), but if you always want to see the high values first, then select this option. Often, ordered categories are ratings, and you want to see the positive ratings first. In the red triangle menu, this option is under Category Options.

Shorten Labels Shortens labels by removing common prefixes and suffixes. Sometimes surveys code a lengthy label that contain a common prefix or suffix. For example, “Occurred 5 to 10 times in the last year” might be a level, but the phrase “in the last year” is repeated for each value label, and you do not need to see it repeated in the report. This option trims the common prefixes and suffixes. It also changes multiple blanks into single blanks. The option only applies to value labels, not column names.

Include Responses Not in Data Includes a count for values that were not in the data. Sometimes when you conduct a survey and give choices, one of the choices is not selected. If you still want to see the choices that are not in the data, use this option. The option requires that the missing categories be specified using the Value Labels column property in the response column.

Supercategories

The term *Supercategories* refers to the extra slots in a table to aggregate over groups of categories. For example, when ratings are involved in a data set (for example, a five point scale), you might want to know the percent of the responses in the top two or other subset of ratings. Such a group of ratings can be defined in the data through the column property, Supercategories.

The Supercategories property supports four keywords: Group, Mean, Std Dev, and All. Mean and Std Dev calculate statistics for value scores, and All aggregates across all levels. For example, a “Top Two” Group supercategory could aggregate the two top categories in a response, as specified. Although we support Mean, Std Dev, and All, we do not recommend using them because they are available as built-in statistics as well as supercategories. Supercategories appear in Crosstab tables and Frequency Charts. Share Charts, however, do not show supercategories and ignore the **Hide** option described below.

To create a supercategory:

1. Select a column in your data table that contains categories that you would like to aggregate.
2. Select **Cols** > **Column Info**.
3. Click **Column Properties** and select **Supercategories**.
4. (Optional) To change the default name of the supercategory, enter a Supercategory Name.
5. Select one or more categories from the Column’s Categories list.
6. Click **Add**.

The Supercategories Options red triangle menu item provides the following commands for a selected category:

Hide Omits columns for each category from the report. Only a column for the supercategory is included.

Net Prevents individuals from being counted twice when they appear in more than one supercategory. Net is available only for a multiple response column.

The following red triangle options are also provided:

Add Mean Includes mean statistics in the report.

Add Std Dev Includes standard deviation statistics in the report.

Add All Includes total responses in the report. By default, the Total Responses column is always included.

Note: Supercategories are supported for all response effects except Repeated Measures and Rater Agreement. Some response effects do not support Mean and Std Dev slots, because the effects do not have a natural score.

The Categorical Report

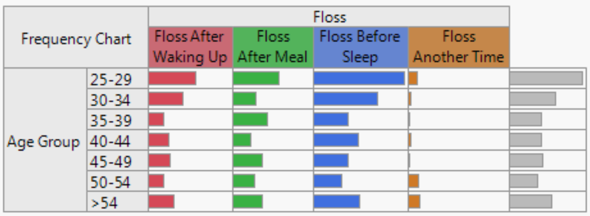
The Categorical platform produces a report with several tables and bar charts depending on your selections. You might or might not see all of the following options depending on which response type and options you selected. Frequencies, Share of Responses, and Rate Per Case appear in a single table by default. A Share Chart also appears by default (unless you used the Structured tab in the launch window). You can choose to view a Frequency Chart or Transposed Frequency Chart.

You can view or hide each option (**Frequencies**, **Share of Responses**, **Rate Per Case**, **Share Chart**, **Frequency Chart**, or **Transposed Freq Chart**) from the Categorical red triangle menu. Data from Consumer Preferences.jmp is displayed in Figure 3.5.

Figure 3.5 The Categorical Report

Indicator Group(Floss After Waking Up, Floss After Meal, Floss Before Sleep, Floss Another Time) By Age Group

	Freq / Share / Rate	Age Group: 25-29	30-34	35-39	40-44	45-49	50-54	>54
Floss	Floss After Waking Up	37	27	12	16	17	12	20
		25.9%	29.3%	18.5%	25.0%	25.8%	21.1%	23.8%
		32.7%	39.7%	24.5%	30.8%	32.7%	27.9%	28.2%
	Floss After Meal	38	19	28	15	24	18	24
		26.6%	20.7%	43.1%	23.4%	36.4%	31.6%	28.6%
		33.6%	27.9%	57.1%	28.8%	46.2%	41.9%	33.8%
	Floss Before Sleep	62	44	24	31	24	20	32
		43.4%	47.8%	36.9%	48.4%	36.4%	35.1%	38.1%
		54.9%	64.7%	49.0%	59.6%	46.2%	46.5%	45.1%
	Floss Another Time	6	2	1	2	1	7	8
		4.2%	2.2%	1.5%	3.1%	1.5%	12.3%	9.5%
		5.3%	2.9%	2.0%	3.8%	1.9%	16.3%	11.3%
	Total Responses	143	92	65	64	66	57	84
	Total Cases	113	68	49	52	52	43	71
	Total Cases Responding	104	67	46	51	50	43	70



Test Each Response, Poisson

Age Group, Floss

Floss	ChiSquare	Prob>ChiSq
Floss After Waking Up	2.6995	0.8455
Floss After Meal	9.1746	0.1640
Floss Before Sleep	4.0051	0.6760
Floss Another Time	13.2863	0.0387*

Chi-squared tests use Poisson rates.

The topmost item in the table is a Frequency count (Freq), showing the frequency counts for each category with the total frequency (Total Responses) and total units (Total Cases) at the bottom of the table.

In this example, the number of responses and cases for each age group by the 7 segments are displayed.

The Share of Responses (Share) is determined by dividing each count by the total number of responses. The number represents the percent of the response among all the responses in the sample (frequency divided by response total). This is either a column percentage or row percentage depending on whether your table has the responses on top or down the side (transposed).

For example, examine the second row of the table for Floss After Waking Up. The 37 responses who floss when they wake up were 25.9% of all responses (37/143*100).

The Rate Per Case (Rate) divides each count in the frequency table by the total number of cases. If you have multiple responses per case (subject), there are two types of percentages; the rate per case is frequency as a percent of total cases, whereas the share of responses is the frequency as a percent of the total responses. Rate is available only for multiple responses.

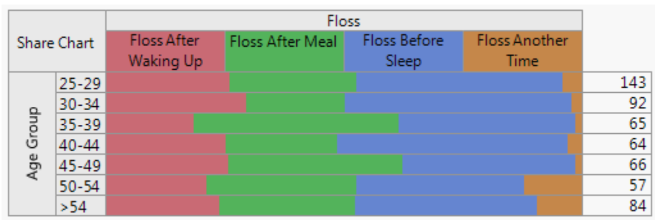
For example, in the third row of the table (Floss After Waking Up), the 37 respondents are from 113 cases, making the rate per respondent 32.7%.

To display the rate of response per case (excluding missing values), select Rate per Case Responding from the red triangle menu. The frequency of response divided by the total cases responding is displayed in the table.

Share Chart

The Share Chart presents a divided bar chart. The bar length is proportional to the percentage of responses for each type. The column on the right shows the number of responses.

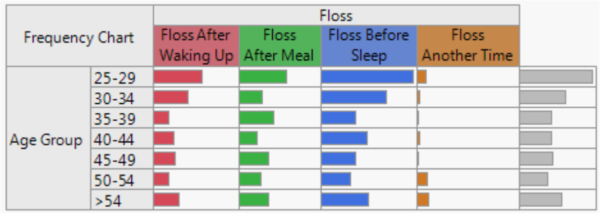
Figure 3.6 Share Chart



Frequency Chart

The Frequency Chart shows response frequencies. The bars reflect the frequency count on the same scale and the number of responses are displayed to the right. To view the frequency chart, select **Frequency Chart** from the Categorical red triangle menu.

Figure 3.7 Frequency Chart



The Transposed Freq Chart option produces a transposed version of the Frequency Chart. Marginal totals are given for each response, as opposed to each *X* variable.

Categorical Platform Options

The Categorical red triangle menu provides commands that customize the appearance of the report and provide the means to test and compare your results. The following options appear in the menu depending on response roles and options selected. You might or might not view all of the options depending on your selections.

Report Options

The default report format is the **Crosstab** format, which gathers all three statistics for each sample level and response together. The Crosstab format displays the responses on the top and the sample levels down the side, with multiple table elements together in each cell of the cross tabulation.

Figure 3.8 Crosstab Format

			Freq Share Rate	contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect	Total Responses	Total Cases	Total Cases Responding
clean	after	date	OCT 1	12 52.2% 24.0%	2 8.7% 4.0%	0 0.0% 0.0%	4 17.4% 8.0%	2 8.7% 4.0%	1 4.3% 2.0%	2 8.7% 4.0%	23	50	50
			OCT 2	10 43.5% 20.0%	1 4.3% 2.0%	1 4.3% 2.0%	5 21.7% 10.0%	1 4.3% 2.0%	2 8.7% 4.0%	3 13.0% 6.0%	23	50	50
			OCT 3	8 47.1% 16.0%	3 17.6% 6.0%	0 0.0% 0.0%	5 29.4% 10.0%	0 0.0% 0.0%	1 5.9% 2.0%	0 0.0% 0.0%	17	50	50
	before	date	OCT 1	14 45.2% 28.0%	2 6.5% 4.0%	1 3.2% 2.0%	2 6.5% 4.0%	3 9.7% 6.0%	8 25.8% 16.0%	1 3.2% 2.0%	31	50	50
			OCT 2	15 50.0% 30.0%	2 6.7% 4.0%	2 6.7% 4.0%	1 3.3% 2.0%	4 13.3% 8.0%	6 20.0% 12.0%	0 0.0% 0.0%	30	50	50
			OCT 3	22 61.1% 44.0%	2 5.6% 4.0%	3 8.3% 6.0%	4 11.1% 8.0%	0 0.0% 0.0%	3 8.3% 6.0%	2 5.6% 4.0%	36	50	50

The Crosstab format has a transposed version, **Crosstab Transposed**, which is useful when there are a lot of response categories but not a lot of sample levels. Crosstab Transposed displays the responses down the side and the sample levels across the top, with multiple table elements together in each cell.

The Structured analysis always uses the Crosstab Transposed form, but in a more complex arrangement. The Free Text analysis displays a Text Explorer report. See the Text Explorer chapter in the *Basic Analysis* book.

Selecting the **Legend** red triangle menu shows or hides the legend for the response column on the Share Chart.

Statistical Options

The main question of interest in any table is whether shares or rates vary from each group of sample levels, and specifically which groups are significantly different.

There are two families of tests and comparisons, which correspond to single category responses and multiple responses:

- With single responses, a given response is just one response category, and the question is whether the share of responses is different across sample levels.

  Single responses are tested with a chi-square test of homogeneity. However, there are two types of this test: the Likelihood Ratio Chi-square and the Pearson Chi-square. It is a matter of personal preference and training which one you prefer. The Chi-square Test Choices option in the Categorical red triangle menu enables you to show one or the other, or both. For more information, refer to "Test Options" on page 57.

- With multiple responses, each individual can select several categories, and the question is whether the rate is different across sample levels.

  For multiple responses, each response is treated in a separate account, with a probability of the count for each subject as a Poisson distribution (allowing for multiples of the same category). Each response is tested to determine whether the parameters are the same across sample levels.

  For overlap among the cells in a structured table, a first-order Rao-Scott correction is used to adjust the Chi-square tests. If there is no overlap, the Rao-Scott correction is 0.5. The first-order Rao-Scott correction is described in Lavassani et al. (2009).

Categorical Platform Commands

The following commands appear in the Categorical red triangle menu depending on the context.

Test Response Homogeneity

Identifies whether the probabilities across the response categories are the same across sample levels. Requires single response data.

The command is available for the following responses:

- Responses
- Aligned Responses
- Repeated Measures
- Response Frequencies with Sample Size
- Structured

The Test Response Homogeneity command shows the Marginal Homogeneity (Independence) Test, both Pearson and Chi-square likelihood ratio chi-square. See "Test Response Homogeneity" on page 48.

Test Multiple Response

Identifies whether the rates are the same across sample levels for each response category. Requires multiple response data.

The command is available for the following responses:

- Multiple Response
- Multiple Response by ID (with Sample Size)
- Multiple Delimited
- Response Frequencies with Sample Size
- Structured

The Test Multiple Responses command shows a Poisson regression or binomial test on the sample for each defect frequency. See "Test Multiple Response" on page 49.

Agreement Statistic

Identifies how closely raters agree and whether the lack of agreement symmetrical.

The command is available for Rater Agreement responses.

The Agreement Statistic command shows Kappa for agreement, and Bowker and McNemar for symmetry. See "Rater Agreement" on page 53.

Transition Report

Identifies how the categories changed across time.

The command is available for Repeated Measures responses.

The Transition Report command shows transition counts and rates matrices. See "Repeated Measures" on page 54.

Cell Chisq

Identifies how to further analyze the results to obtain more information.

The command is available for Responses. See "Cell Chisq" on page 51.

Compare Each Sample

Identifies whether levels of the response category differ significantly.

The command is available for the following responses:

- Responses
- Aligned Responses
- Repeated Measures
- Response Frequencies (if no Sample Size)
- Structured

See "Compare Each Sample" on page 55.

Compare Each Cell

Identifies whether pairs of levels within the two response categories differ significantly.

The command is available for the following responses:

- Single and Multiple Responses
- Structured

See "Compare Each Cell" on page 56.

Test Options

Identifies how to further analyze the results to obtain more information.

The command is available for the following responses:

- ChiSquare Test Choices
- Show Warnings
- Order by Significance
- Hide Nonsignificant

See "Test Options" on page 57.

Additional Categorical Platform Options

There are a series of options that add more detail for each group of sample levels. The options that appear depend on your selections and the details of your analysis:

Total Responses Shows the sum of the frequency counts for each group of sample levels.

Response Levels Shows the data levels for the response column.

Show Supercategories Shows extra slots for supercategories in the Crosstab table and the Frequency Chart to aggregate over groups of categories. For more information about supercategories, refer to "Supercategories" on page 38.

Total Cases For multiple response columns, shows the number of cases (subjects), which are different from the number of responses.

Total Cases Responding For multiple response columns used in Structured tables, counts each person who responded at least once. People who did not respond at all are not included.

Mean Score Calculates the response means, using the numeric categories, or value scores. This is enabled for columns that use numeric codes, or for categories that have a **Value Scores** property. To make the Mean Score interpretable, you can assign specific value scores in the Column Info window with the Value Scores column property.For more information and an example, refer to "Mean Score Example" on page 66.

Mean Score Comparison Compares the mean scores across groups of sample levels, showing which groups are significantly different. A pairwise multiple comparisons Student's t test is used for the mean score comparison, based on the specified comparison groups. The pairwise multiple comparison test does not pool variances across all samples. Instead, the test uses the unpooled Satterthwaite t test for pairwise comparisons. See SAS Institute (2009).

For more information about the letter codes, refer to "Comparisons with Letters" on page 56. For more information about specifying the comparison groups, refer to "Specify Comparison Groups" on page 60.

Std Dev Score Calculates the standard deviation of the value scores.

Order by Mean Score Orders the mean score calculations. The option only appears when there are no X columns in the analysis.

Save Tables Saves the report to a new data table. For more information, refer to "Save Tables" on page 57.

Filter Filters data to specific groups or ranges. Opens the Local Data Filter panel allowing you to identify varying subsets of data. The filtered rows do not appear in the reports. Sample levels with 0 values are always hidden. To show the filtered rows in reports, select **Include Responses Not in Data** in the launch window. You can also select the **Set Preferences** red triangle menu, and then select **Include Responses Not in Data**.

Contents Summary Collects all of the tests and mean scores into a summary at the top of the report with links to the associated item.

Show Columns Used in Report Shows or hides Columns Used in Report information. This option affects only columns that have an SPSS Name or SPSS Label column property.

Format Elements Enables you to specify formats for Frequencies, Shares and Rates, and how zeros are displayed. By default, Frequencies are Fixed Dec with 7 Width and 0 Decimals and Shares and Rates are Percent with 6 Width and 1 Decimal.

Arrange in Rows Arranges the reports across the page as opposed to down. Enter the number of reports that you want to view across the window.

Set Preferences Enables you to set preferences for future launches and sessions. For more information, refer to "Set Preferences" on page 58.

Category Options Contains options (Grouping Option, Count Missing Response, Order Response Levels High to Low, Shorten Labels, and Include Responses Not in Data) that are also presented on the launch window that could be specified before the analysis. The options can also be selected here and have the effect of rerunning the platform with the new option setting. For more information, refer to "Other Launch Window Options" on page 37.

Force Crosstab Shading Forces shading on crosstab reports even if the preference is set to no shading.

Relaunch Dialog Enables you to return to the launch window and edit the specifications for a structured table. For more information, refer to "Structured Tab" on page 33.

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.

Test Response Homogeneity

Test Response Homogeneity is the standard chi-square test (for single responses) across all sample levels. There is typically one categorical response variable and one categorical sample variable. Multiple sample variables are treated as a single variable.

The test is the chi-square test for marginal homogeneity of response patterns, testing that the response probabilities are the same across samples. This is equivalent to a test for independence when the sample level is like a response. There are two versions of this test, the Pearson form and the Likelihood Ratio form, both with chi-square statistics. The Test Options menu (ChiSquare Test Choices) is used to show or hide the Likelihood Ratio or Pearson tests. If Show Warnings is turned on, the report displays if the frequencies are too low to make good tests.

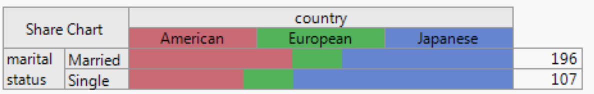
As an example:

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select country and click **Responses** on the Simple tab.
4. Select marital status and click **X, Grouping Category.**
5. Click **OK.**
6. Select **Test Response Homogeneity** from the Categorical red triangle menu.

Figure 3.9 Test Response Homogeneity

Responses(country) By marital status

	Freq / Share	country: American	European	Japanese	Total Responses
marital status	Married	83 42.3%	26 13.3%	87 44.4%	196
	Single	32 29.9%	14 13.1%	61 57.0%	107



Test Response Homogeneity

Response Dimension Label	Sample Dimension Label	LR Chisq	LR PValue	Pearson Chisq	Pearson PValue
country	marital status	5.14007	0.0765	5.08146	0.0788

The Share Chart indicates that the married group is more likely to buy American cars, and the single group is more likely to buy Japanese cars, but the statistical test only shows a significance of 0.08. Therefore, the difference in response probabilities across marital status is not statistically significant at an alpha level of 0.05.

Test Multiple Response

Test Multiple Response is the standard chi-square test for multiple responses, with one test statistic for each response category. When there are multiple responses, each response category can be modeled separately. The question is whether the response rates are the same across samples.

The Test Multiple Response red triangle menu provides the following options:

Count Test, Poisson For each response category, the frequency count has a random Poisson distribution. The rate test is obtained using a Poisson regression (through generalized linear models) of the frequency per unit modeled by the sample categorical variable. The result is a likelihood ratio chi-square test of whether the rates are different across sample levels. This test can also be done by the Pareto platform, as well as in the Generalized Linear Model personality of the Fit Model platform.

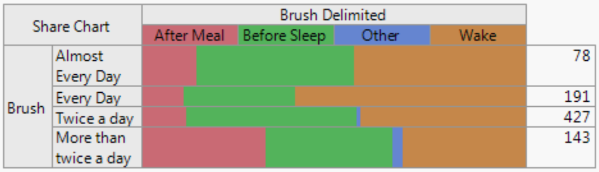
Homogeneity Test, Binomial For each response category, the frequency count has a random binomial distribution. Select this test when the response can be true only once for each respondent (for example, in a "check all that apply" questionnaire).

To test multiple responses, follow these steps:

1. Select **Help > Sample Data Library** and open Consumer Preferences.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select Brush Delimited and click **Multiple Delimited** on the Multiple tab.
4. Select brush and click **X, Grouping Category** to compare the sample levels across the brush treatment variable.
5. Click **OK.**
6. Select **Test Multiple Response** and then **Count Test, Poisson** from the Categorical red triangle menu.

Figure 3.10 Test Multiple Response, Poisson

	Freq / Share / Rate	Brush Delimited						
		After Meal	Before Sleep	Other	Wake	Total Responses	Total Cases	Total Cases Responding
Brush	Almost Every Day	11	32	0	35	78	48	48
		14.1%	41.0%	0.0%	44.9%			
		22.9%	66.7%	0.0%	72.9%			
	Every Day	21	55	0	115	191	142	142
		11.0%	28.8%	0.0%	60.2%			
		14.8%	38.7%	0.0%	81.0%			
	Twice a day	49	189	4	185	427	209	209
		11.5%	44.3%	0.9%	43.3%			
		23.4%	90.4%	1.9%	88.5%			
	More than twice a day	46	47	4	46	143	49	49
		32.2%	32.9%	2.8%	32.2%			
		93.9%	95.9%	8.2%	93.9%			



Test Each Response, Poisson

Brush, Brush Delimited

Brush Delimited	ChiSquare	Prob>ChiSq
After Meal	59.5436	<.0001*
Before Sleep	39.1097	<.0001*
Other	12.7131	0.0053*
Wake	1.8793	0.5978

Chi-squared tests use Poisson rates.

The *p*-values show that After Meal and Before Sleep are the most significantly different. Wake is not significantly different with this amount of data.

7. Select **Test Multiple Response** and then **Homogeneity Test, Binomial** from the Categorical red triangle menu.

Figure 3.11 Test Multiple Response, Binomial

Test Each Response, Poisson

Brush, Brush Delimited

Brush Delimited	ChiSquare	Prob>ChiSq
After Meal	59.5436	<.0001*
Before Sleep	39.1097	<.0001*
Other	12.7131	0.0053*
Wake	1.8793	0.5978

Chi-squared tests use Poisson rates.

Test Each Response, Binomial

Brush, Brush Delimited

Brush Delimited	ChiSquare	Prob>ChiSq
After Meal	113.315	<.0001*
Before Sleep	131.169	<.0001*
Other	12.9822	0.0047*
Wake	12.2461	0.0066*

Assuming multiple responses are Check All That Apply (CATA).

The Homogeneity Test, Binomial option always produces a larger test statistic (and therefore a smaller *p*-value) than the Count Test, Poisson option. The binomial distribution compares not only the rate at which the response occurred (the number of people who reported that they brush upon waking) but also the rate at which the response did not occur (the number of people who did not report that they brush upon waking).

Note: JMP detects a multiple response column by the Multiple Response modeling type or the Multiple Response column property.

Cell Chisq

For single responses, Cell Chisq displays the cell-by-cell composition of the Pearson chi-square overall, and also shows which cells have relatively more (red) or less (blue) than expected if they were the same across sample levels. The value shown is the p-value for the chi-square. The color is bright when they are significant, and grayer when less significant, denoting visually where the significant differences are. Note the JMP default colors are used in the following example.

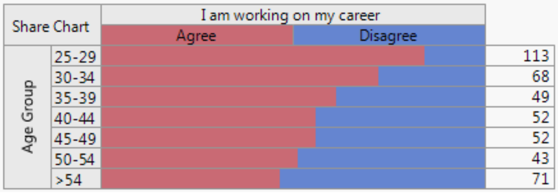
As an example:

1. Select **Help > Sample Data Library** and open Consumer Preferences.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select I am working on my career and click **Responses** on the Simple tab.
4. Select Age Group and click **X, Grouping Category**.
5. Click **OK.**
6. Select **Crosstab Transposed** from the red triangle menu.
7. Select **Cell Chisq** from the red triangle menu.

Figure 3.12 Cell Chisq

Responses(I am working on my career) By Age Group

I am working on my career	Freq / Share / Cell Chisq PVal	25-29	30-34	35-39	40-44	45-49	50-54	>54
Agree	Freq	95	49	30	29	29	22	33
	Share	84.1%	72.1%	61.2%	55.8%	55.8%	51.2%	46.5%
	Cell Chisq PVal	0.00788	0.41003	0.80398	0.45495	0.45495	0.29058	0.06415
Disagree	Freq	18	19	19	23	23	21	38
	Share	15.9%	27.9%	38.8%	44.2%	44.2%	48.8%	53.5%
	Cell Chisq PVal	0.00039	0.27136	0.74035	0.31848	0.31848	0.15823	0.01345
Total Responses		113	68	49	52	52	43	71



Relative Risk

The Relative Risk option is used to compute relative risks for different responses. The risk of responses is computed for each level of the X, Grouping variable. The risks are compared to get a relative risk. This option is available when the X, Grouping variable has two levels, and one of the following is true:

- The response variable has two levels.
- The response variable is a Multiple Response and the **Unique occurrences within ID** box is checked on the Categorical launch window.

A common application of this analysis is when the responses represent adverse events (side effects), and the X variable represents a treatment (drug versus placebo). The risk for getting each side effect is computed for both the drug and placebo. The relative risk is the ratio of the two risks.

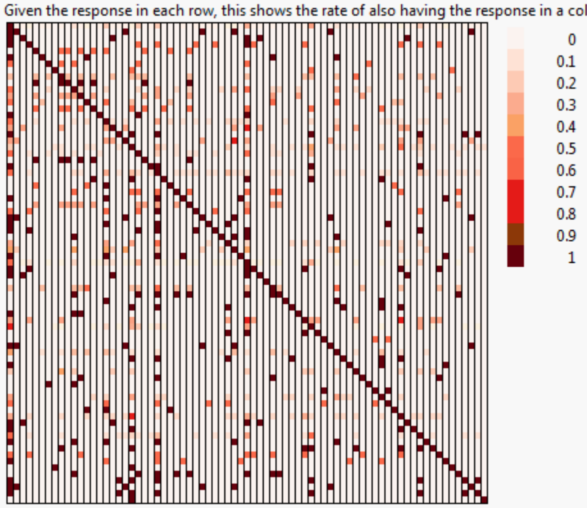
Conditional Association

The **Conditional Association** option is used to compute the conditional probability of one response given a different response. A table and color map of the conditional probabilities are given. This option is available only when the **Unique occurrences within ID** box is checked on the Categorical launch window. A common application of this analysis is when the responses represent adverse events (side effects) from a drug. The computations represent the conditional probability of one side effect given the presence of another side effect. For AdverseR.jmp, given the response in each row, Figure 3.13 shows the rate of also having the response in a column. Figure 3.13 only displays a few variables in the table due to size constraints.

Figure 3.13 Conditional Association

Conditional Association

Given the response in each row, this shows the rate of also having the response in a column.



Conditional Assoc

	ABDOMINAL PAIN	ABNORMAL VISION	ALOPECIA	AMBLYOPIA
ABDOMINAL PAIN	1.0000	0.0323	0.0323	0.0645
ABNORMAL VISION	1.0000	1.0000	0.0000	0.0000
ALOPECIA	1.0000	0.0000	1.0000	0.0000
AMBLYOPIA	1.0000	0.0000	0.0000	1.0000
ANEMIA	0.0000	0.0000	0.0000	0.0000
ANGINA PECTORIS	1.0000	0.0000	0.0000	0.0000
ANOREXIA	0.5000	0.0000	0.0000	0.0000
ANXIETY	0.0000	0.0000	0.0000	0.0000
ASTHENIA	0.2500	0.0000	0.0000	0.0000
ASTHMA	0.0000	0.0000	0.0000	0.0000
BACK PAIN	0.5000	0.0000	0.0000	0.0000
BRONCHITIS	0.3333	0.0000	0.0000	0.0000

Rater Agreement

The Rater Agreement analysis answers the questions of how closely raters agree with one another and if the lack of agreement is symmetrical. For example, open Attribute Gauge.jmp. The Attribute Chart script runs the Variability Chart platform, which has a test for agreement among raters.

Figure 3.14 Agreement Comparisons

Agreement Comparisons

Rater	Compared with Rater	Kappa	.2 .4 .6 .8	Standard Error
A	B	0.8629		0.0442
A	C	0.7761		0.0547
B	C	0.7880		0.0537

Launch the Categorical platform and designate the three raters (A, B, and C) as **Rater Agreement** responses on the Related tab on the launch window. In the resulting report, you have a similar test for agreement that is augmented by a symmetry test that the lack of agreement is symmetric.

Figure 3.15 Agreement Statistics

Rater Agreement Statistics

Degree of Agreement

Degree of Agreement

Response1	Response2	Kappa	Std Err	Bowker Symmetry	Bowker P-Value
A	B	0.862944	0.044198	1	0.3173
A	C	0.776119	0.054671	0.066667	0.7963
B	C	0.788007	0.053702	1.142857	0.2850

Two by two tables are the same as Bowkers test

Details

Rater Row	Rater Col	Level	0	1
B	A	0	44	3
B	A	1	6	97
C	A	0	43	8
C	A	1	7	92
C	B	0	42	9
C	B	1	5	94

Repeated Measures

Repeated Measures declares that the columns reflect responses made by the same individual at different times, and you are interested in the changes between the times. Individual reports are displayed for each item, with a transition report at the end demonstrating the transition counts and rate matrices. (This example uses the JMP default colors.)

Here is an example:

1. Select **Help > Sample Data Library** and open Presidential Elections.jmp.
2. Select **Analyze > Consumer Research > Categorical**.
3. Select 1980 Winner through 2012 Winner and click **Repeated Measures** on the Related tab.
4. Select State and click **X, Grouping Category**.
5. Click **OK.**
6. Scroll to the bottom of the report window and open the Transition Report outline.

Scroll through the responses to see how each State has voted over the years. Note that New Mexico has varied between Democratic and Republic over the years.

Figure 3.16 Repeated Measures

New Mexico	1980 Winner	1984 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**1984 Winner**
			Democrat	0	0	Democrat	.	0.0000	
			Republican	0	1	Republican	.	1.0000	
New Mexico	1984 Winner	1988 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**1988 Winner**
			Democrat	0	0	Democrat	.	0.0000	
			Republican	0	1	Republican	.	1.0000	
New Mexico	1988 Winner	1992 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**1992 Winner**
			Democrat	0	0	Democrat	0.0000	.	
			Republican	1	0	Republican	1.0000	.	
New Mexico	1992 Winner	1996 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**1996 Winner**
			Democrat	1	0	Democrat	1.0000	.	
			Republican	0	0	Republican	0.0000	.	
New Mexico	1996 Winner	2000 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**2000 Winner**
			Democrat	1	0	Democrat	1.0000	.	
			Republican	0	0	Republican	0.0000	.	
New Mexico	2000 Winner	2004 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**2004 Winner**
			Democrat	0	1	Democrat	.	1.0000	
			Republican	0	0	Republican	.	0.0000	
New Mexico	2004 Winner	2008 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**2008 Winner**
			Democrat	0	0	Democrat	0.0000	.	
			Republican	1	0	Republican	1.0000	.	
New Mexico	2008 Winner	2012 Winner		**Democrat**	**Republican**		**Democrat**	**Republican**	**2012 Winner**
			Democrat	1	0	Democrat	1.0000	.	
			Republican	0	0	Republican	0.0000	.	

Compare Each Sample

For a given response, Compare Each Sample tests whether the response probability for each of its levels differs from the response probabilities for its other levels. In simple situations, the Compare Each Sample report consists of symmetric matrices of *p*-values, as shown in Figure 3.17.

In addition, a new row or column, entitled Compare, appears in the Crosstabs table. The Compare row is placed at the bottom of the table, or the Compare column is placed at the far right. (Whether a row or column is appended depends on whether Crosstab or Crosstab Transposed is specified.) The Compare row or column contains letter codes showing which sample levels differ significantly. For more information about the letter codes, refer to "Comparisons with Letters" on page 56.

Figure 3.17 Compare Each Sample

Compare Each Sample

Letter comparisons use Pearson Chisq

Age Group, I am working on my career

LR PValues

LR Chi-square p-value on pairs

	A	B	C	D	E	F	G
A	1.0000	0.0552	0.0020	0.0001	0.0001	<.0001	<.0001
B	0.0552	1.0000	0.2183	0.0641	0.0641	0.0261	0.0020
C	0.0020	0.2183	1.0000	0.5781	0.5781	0.3312	0.1108
D	0.0001	0.0641	0.5781	1.0000	1.0000	0.6540	0.3083
E	0.0001	0.0641	0.5781	1.0000	1.0000	0.6540	0.3083
F	<.0001	0.0261	0.3312	0.6540	0.6540	1.0000	0.6276
G	<.0001	0.0020	0.1108	0.3083	0.3083	0.6276	1.0000

Pearson PValues

Pearson Chi-square p-value on pairs

	A	B	C	D	E	F	G
A	1.0000	0.0523	0.0015	<.0001	<.0001	<.0001	<.0001
B	0.0523	1.0000	0.2170	0.0638	0.0638	0.0255	0.0022
C	0.0015	0.2170	1.0000	0.5783	0.5783	0.3314	0.1119
D	<.0001	0.0638	0.5783	1.0000	1.0000	0.6540	0.3087
E	<.0001	0.0638	0.5783	1.0000	1.0000	0.6540	0.3087
F	<.0001	0.0255	0.3314	0.6540	0.6540	1.0000	0.6276
G	<.0001	0.0022	0.1119	0.3087	0.3087	0.6276	1.0000

Compare Each Cell

For a given response and a given X variable, Compare Each Cell tests, for each level of the X variable, whether the response probabilities differ across the levels of the response. In other words, Compare Each Cell tests response probabilities across the cells in a given row of the Crosstabs table. The Compare Each Cell report gives *p*-values in a tabular format. The letters across the top indicate the response levels tested for the given level of the X variable. An example is shown in Figure 3.18.

In addition, when a cell differs significantly from other cells, a letter code is inserted into the appropriate cell in the Crosstabs table. For details on the letter codes and on their placement in cells, refer to "Comparisons with Letters" on page 56.

Figure 3.18 Compare Each Cell (cut off after column AE)

Compare Each Cell - Details

Letter comparisons use Fisher's Exact Test

Age Group, I am working on my career

LR Pairs

		AB	BB	AC	BC	CC	AD	BD	CD	DD	AE
Agree	1.0000	0.0552	1.0000	0.0020	0.2183	1.0000	0.0001	0.0641	0.5781	1.0000	0.0001
Disagree	1.0000	0.0552	1.0000	0.0020	0.2183	1.0000	0.0001	0.0641	0.5781	1.0000	0.0001

Pearson Pairs

		AB	BB	AC	BC	CC	AD	BD	CD	DD	AE
Agree	1.0000	0.0523	1.0000	0.0015	0.2170	1.0000	0.0001	0.0638	0.5783	1.0000	0.0001
Disagree	1.0000	0.0523	1.0000	0.0015	0.2170	1.0000	0.0001	0.0638	0.5783	1.0000	0.0001

Fisher Exact Pairs

		AB	BB	AC	BC	CC	AD	BD	CD	DD	AE
Agree	1.0000	0.0591	1.0000	0.0022	0.2357	1.0000	0.0002	0.0825	0.6869	1.0000	0.0002
Disagree	1.0000	0.0591	1.0000	0.0022	0.2357	1.0000	0.0002	0.0825	0.6869	1.0000	0.0002

Comparisons with Letters

The Compare Each Cell, Compare Each Sample, and Mean Score Comparisons commands use a system of letters to identify sample levels. The first sample level is "A", the second "B". For more than 26 sample levels, numbers are appended after the letters. The letters are shown in the sample level headings when a comparison command is turned on.

If two sample levels are significantly different, the letter of the sample level with lesser share is placed into the comparison cell of the other category that is significantly different. To find out if a given group of sample levels, for example "B", is significantly different, then you have to look both in the comparison cell for column "B" for other letters, and also in all the other cells across the groups of sample levels for a "B".

Lowercase letters are also used for comparisons that are slightly less significant, according to Table 3.2. These comparisons suffer when the count for that sample level group (the Base Count) is small, and asterisks start to appear in the comparison cells to warn you.

The comparison features are controlled by four options set in Preferences or through a script. For more information, refer to "Set Preferences" on page 58.

Table 3.2 Letter Comparisons

Uppercase alpha level	0.05	The significance level for which uppercase letters show differences.
Lowercase alpha level	0.10	The significance level for which lowercase letters show differences.
Base Count minimum	≤ 29	The count for a sample level that leads to a ** warning.
Base Count warning	30 to 99	The count for a sample level that leads to a * warning.

Test Options

The Test Options menu on the Categorical red triangle menu has the following options depending on your selections:

ChiSquare Test Choices Single responses are tested with a chi-square test of homogeneity; either the Likelihood Ratio Chi-square or the Pearson Chi-square, or both. Options are: **Both LR and Pearson**, **LR Only**, or **Pearson Only**. You can set an option in Preferences.

Show Warnings Shows warnings for chi-square tests related to small sample sizes.

Order by Significance Reorders the reports so that the most significant reports are at the top. This option only applies to reports with one homogeneity test.

Hide Nonsignificant Suppresses reports that are deemed non-significant. This option only applies to reports with one homogeneity test.

Save Tables

The Save Tables menu on the Categorical red triangle menu has the following options depending on your selections:

Save Frequencies Saves the Frequency report to a new data table, without the marginal totals or supercategories.

Save Share of Responses Saves the Share of Responses report to a new data table, without the marginal totals.

Save Rate Per Case Saves the Rate Per Case report to a new data table, without the marginal totals.

Save Transposed Frequencies Saves a transposed version of the Frequency report to a new data table, without the marginal totals.

Save Transposed Share of Responses Saves a transposed version of the Share of Responses report to a new data table, without the marginal totals.

Save Transposed Rate Per Case Saves a transposed version of the Rate Per Case report to a new data table.

Save Test Rates Saves the results of the Test Multiple Response option to a new data table.

Save Test Homogeneity Saves the results of the Test Response Homogeneity option to a new data table.

Save Mean Scores Saves the mean scores for each sample group in a new data table.

Save tTests and pValues Save t-tests and p-values from the Mean Score Comparisons report in a new data table.

Save Excel File Creates a Microsoft Excel spreadsheet with the structure of the crosstab-format report. The option maps all of the tables to one sheet, with the response categories as rows, the sample levels as columns, sharing the headings for sample levels across multiple tables. When there are multiple elements in each table cell, you have the option to make them multiple or single cells in Microsoft Excel.

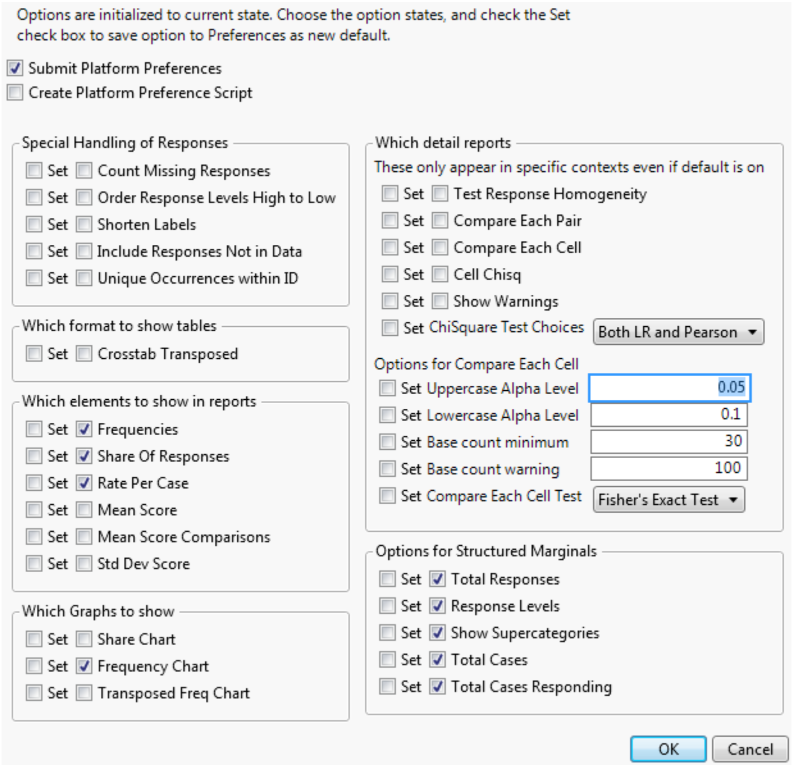
Set Preferences

You can specify settings and set preferences within the Categorical platform. Several options are available on the launch window and can be specified before the analysis. Some of the options can also be selected from the Categorical red triangle menu, and have the effect of rerunning the analysis with the new setting.

The options are initialized to the current state. Select the appropriate options and select either **Submit Platform Preferences** or **Create Platform Preference Script** to submit the options to your preferences as the new default. When the Categorical platform is launched, the preferences associated with the current preference set are enacted.

Preferences can be administered and shared through a script. The best way to share a preference set widely is to create an add-in, so that if the preference settings are reset to the initial state, the add-in could restore the preferred set.

Figure 3.19 Set Preferences Window

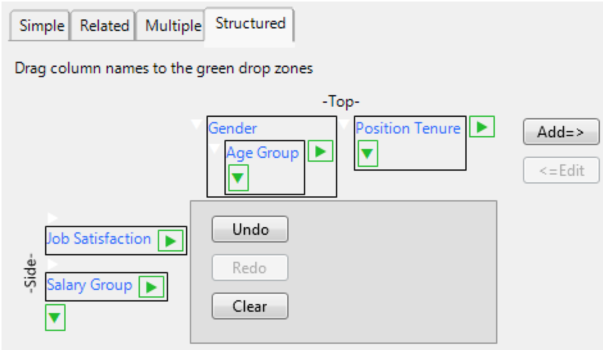


Structured Report Options

The Structured tab enables you to construct complex tables of descriptive statistics by dragging column names into green icon drop zones to create side-by-side and nested results. The following example uses the Consumer Preferences.jmp sample data table. From this data, suppose that you wanted to compare job satisfaction and salary against gender by age group and position tenure.

1. Select **Help > Sample Data Library** and open Consumer Preferences.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select the Structured tab.
4. Drag Gender to the green drop zone at the **Top** of the table on the Structured tab.
5. Drag Age Group to the green drop zone just below Gender.
6. Drag Position Tenure to the green drop zone at the **Top** of the table next to Gender.
7. Drag Job Satisfaction to the green drop zone at the **Side** of the table.
8. Drag Salary Group to the green drop zone at the **Side** of the table under Job Satisfaction.

Figure 3.20 Structured Tab Report Setup



9. Click **Add=>**.

10. Click **OK**.

Figure 3.21 Structured Tab Report Example

Job Satisfaction+Salary Group By Gender*Age Group+Position Tenure

	Freq	Gender																	
	Share	M							F										
		Age Group							Age Group							Position Tenure			
		25-29	30-34	35-39	40-44	45-49	50-54	>54	25-29	30-34	35-39	40-44	45-49	50-54	>54	less than 5 years	5 to 10 years	10 to 20 years	more than 20 years
Job Satisfaction	Not at all satisfied	3	3	1	3	3	2	2	3	1	0	5	2	1	3	15	11	6	0
		5.5%	10.3%	3.3%	8.8%	7.9%	8.3%	4.1%	5.2%	2.6%	0.0%	27.8%	14.3%	5.3%	13.6%	7.2%	8.2%	6.8%	0.0%
	Somewhat satisfied	32	14	15	19	19	10	26	34	30	12	6	5	11	9	122	70	40	10
		58.2%	48.3%	50.0%	55.9%	50.0%	41.7%	53.1%	58.6%	76.9%	63.2%	33.3%	35.7%	57.9%	40.9%	58.7%	52.2%	45.5%	55.6%
	Extremely satisfied	20	12	14	12	16	12	21	21	8	7	7	7	7	10	71	53	42	8
		36.4%	41.4%	46.7%	35.3%	42.1%	50.0%	42.9%	36.2%	20.5%	36.8%	38.9%	50.0%	36.8%	45.5%	34.1%	39.6%	47.7%	44.4%
	Total Responses	55	29	30	34	38	24	49	58	39	19	18	14	19	22	208	134	88	18
Salary Group	less than 40000	20	6	4	8	6	5	10	35	19	6	6	2	7	4	80	40	15	3
		36.4%	20.7%	13.3%	23.5%	15.8%	20.8%	20.4%	60.3%	48.7%	31.6%	33.3%	14.3%	36.8%	18.2%	38.5%	29.9%	17.0%	16.7%
	40000 to60000	16	14	11	12	14	7	13	15	10	8	6	7	6	12	67	38	37	9
		29.1%	48.3%	36.7%	35.3%	36.8%	29.2%	26.5%	25.9%	25.6%	42.1%	33.3%	50.0%	31.6%	54.5%	32.2%	28.4%	42.0%	50.0%
	60000 to80000	8	5	7	7	8	8	10	5	5	3	3	3	4	5	32	25	19	5
		14.5%	17.2%	23.3%	20.6%	21.1%	33.3%	20.4%	8.6%	12.8%	15.8%	16.7%	21.4%	21.1%	22.7%	15.4%	18.7%	21.6%	27.8%
	80000 to120000	9	3	6	4	7	3	7	1	2	1	2	1	1	0	19	17	10	1
		16.4%	10.3%	20.0%	11.8%	18.4%	12.5%	14.3%	1.7%	5.1%	5.3%	11.1%	7.1%	5.3%	0.0%	9.1%	12.7%	11.4%	5.6%
	greater than 120000	2	1	2	3	3	1	9	2	3	1	1	1	1	1	10	14	7	0
		3.6%	3.4%	6.7%	8.8%	7.9%	4.2%	18.4%	3.4%	7.7%	5.3%	5.6%	7.1%	5.3%	4.5%	4.8%	10.4%	8.0%	0.0%
	Total Responses	55	29	30	34	38	24	49	58	39	19	18	14	19	22	208	134	88	18

Figure 3.21 shows that the majority of both the male and female respondents were somewhat satisfied with their jobs, with the highest percentage of males being in the 25-29 age group, while the females were in the 30-34 age group. Most of those who were somewhat satisfied had been in their current position for less than 5 years.

The following options are available from the structured report's red triangle menu:

Show Letters Forces the table to display the column letter IDs, which usually come out automatically when you do a compare command.

Specify Comparison Groups Enables you to specify groups when the group of sample levels that you want to test and compare are not the same as the innermost term's structure. To use this option, you must look at the letter IDs, and then enter sets of letter IDs, separated by a slash, representing each group, separating multiple groups from each other by commas. For example, the default grouping might be "A/B/C, E/D/F", but you want to test

A with E, B with D and C with F, so you specify the groups as “A/E, B/D, C/F”. This determines which letters appear in the comparison fields. In addition, a summary report shows the overall tests for each column group.

Remove Removes the table from the report.

Additional Examples of the Categorical Platform

The following examples come from testing a fabrication line on three different occasions under two different conditions. Each set of operating conditions yielded 50 data points. Inspectors recorded the following types of defects:

- contamination
- corrosion
- doping
- metallization
- miscellaneous
- oxide defect
- silicon defect

Each unit could have several defects or even several defects of the same kind. We illustrate the data in a variety of different examples all within the Categorical platform.

Multiple Response

Suppose that the defects for each unit are entered via a web page, but because each unit rarely has more than three defect types, the form has three fields to enter any of the defect types for a unit, as in Failure3MultipleField.jmp.

1. Select **Help > Sample Data Library** and open Quality Control/Failure3MultipleField.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select Failure1, Failure2, and Failure3 and click **Multiple Response** on the Multiple tab.

   These columns contain defect types and are the variables that you want to inspect.
4. Select clean and date and click **X, Grouping Category.**
5. Click **OK.**

Figure 3.22 lists failure types and counts for each failure type from the **Multiple Response** analysis.

Figure 3.22 Multiple Response Failures from Failure3MultipleField.jmp

Multiple Response(Failure1, Failure2, Failure3) By clean*date

				Freq Share Rate	Failure									
					contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect	Total Responses	Total Cases	Total Cases Responding
clean	after	date	OCT 1		12	2	0	4	2	1	2	23	50	23
					52.2%	8.7%	0.0%	17.4%	8.7%	4.3%	8.7%			
					24.0%	4.0%	0.0%	8.0%	4.0%	2.0%	4.0%			
			OCT 2		10	1	1	5	1	2	3	23	50	23
					43.5%	4.3%	4.3%	21.7%	4.3%	8.7%	13.0%			
					20.0%	2.0%	2.0%	10.0%	2.0%	4.0%	6.0%			
			OCT 3		8	3	0	5	0	1	0	17	50	17
					47.1%	17.6%	0.0%	29.4%	0.0%	5.9%	0.0%			
					16.0%	6.0%	0.0%	10.0%	0.0%	2.0%	0.0%			
	before	date	OCT 1		14	2	1	2	3	8	1	31	50	28
					45.2%	6.5%	3.2%	6.5%	9.7%	25.8%	3.2%			
					28.0%	4.0%	2.0%	4.0%	6.0%	16.0%	2.0%			
			OCT 2		15	2	2	1	4	6	0	30	50	30
					50.0%	6.7%	6.7%	3.3%	13.3%	20.0%	0.0%			
					30.0%	4.0%	4.0%	2.0%	8.0%	12.0%	0.0%			
			OCT 3		22	2	3	4	0	3	2	36	50	36
					61.1%	5.6%	8.3%	11.1%	0.0%	8.3%	5.6%			
					44.0%	4.0%	6.0%	8.0%	0.0%	6.0%	4.0%			

Response Frequencies

Suppose the data have columns containing frequency counts for each batch and a column showing the total number of units of the batch, as in Failure3Freq.jmp.

Figure 3.23 Failure3Freq.jmp Data Table

	clean	date	contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect	SampleSize
1	after	OCT 1	12	2	0	4	2	1	2	50
2	after	OCT 2	10	1	1	5	1	2	3	50
3	after	OCT 3	8	3	0	5	0	1	0	50
4	before	OCT 1	14	2	1	2	3	8	1	50
5	before	OCT 2	15	2	2	1	4	6	0	50
6	before	OCT 3	22	2	3	4	0	3	2	50

1. Select **Help > Sample Data Library** and open Quality Control/Failure3Freq.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select the frequency variables (contamination, corrosion, doping, metallization, miscellaneous, oxide defect, silicon defect) and click **Response Frequencies** on the Multiple tab.
4. Select clean and date and click **X, Grouping Category.**
5. Select Sample Size and click **Sample Size.**
6. Click **OK.**

The resulting output in Figure 3.24 shows a frequency count table, with a separate column for each of the seven batches. The last two columns show the total number of defects (**Total Responses**) and cases (**Total Cases**).

Figure 3.24 Defect Rate Output

Response Frequencies(contamination, corrosion, doping, metallization, miscellaneous, oxide defect, silicon defect) By clean*date

				Freq / Share / Rate	Response									
					contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect	Total Responses	Total Cases	Total Cases Responding
clean	after	date	OCT 1	Freq	12	2	0	4	2	1	2	23	50	50
				Share	52.2%	8.7%	0.0%	17.4%	8.7%	4.3%	8.7%			
				Rate	24.0%	4.0%	0.0%	8.0%	4.0%	2.0%	4.0%			
			OCT 2	Freq	10	1	1	5	1	2	3	23	50	50
				Share	43.5%	4.3%	4.3%	21.7%	4.3%	8.7%	13.0%			
				Rate	20.0%	2.0%	2.0%	10.0%	2.0%	4.0%	6.0%			
			OCT 3	Freq	8	3	0	5	0	1	0	17	50	50
				Share	47.1%	17.6%	0.0%	29.4%	0.0%	5.9%	0.0%			
				Rate	16.0%	6.0%	0.0%	10.0%	0.0%	2.0%	0.0%			
	before	date	OCT 1	Freq	14	2	1	2	3	8	1	31	50	50
				Share	45.2%	6.5%	3.2%	6.5%	9.7%	25.8%	3.2%			
				Rate	28.0%	4.0%	2.0%	4.0%	6.0%	16.0%	2.0%			
			OCT 2	Freq	15	2	2	1	4	6	0	30	50	50
				Share	50.0%	6.7%	6.7%	3.3%	13.3%	20.0%	0.0%			
				Rate	30.0%	4.0%	4.0%	2.0%	8.0%	12.0%	0.0%			
			OCT 3	Freq	22	2	3	4	0	3	2	36	50	50
				Share	61.1%	5.6%	8.3%	11.1%	0.0%	8.3%	5.6%			
				Rate	44.0%	4.0%	6.0%	8.0%	0.0%	6.0%	4.0%			

Each Frequency Group contains the following information:

- The total number of defects for each defect type. For example, after cleaning on Oct 1st, there were 12 contamination defects.
- The share of responses. For example, after cleaning on Oct 1st, the 12 contamination defects were (12/23) accounting for 52.2% of all defects.
- The rate per case. For example, after cleaning on Oct 1st, the 12 contamination defects are from 50 units (12/50) making the rate per unit 24%.

Indicator Group

In some cases, the data is not yet summarized, so there are individual records for each unit. We illustrate this situation with the data table, Failures3Indicators.jmp.

Figure 3.25 Failure3Indicators.jmp Data Table

	clean	date	ID	ID Label	contamination	corrosion	doping	metallization	miscellaneous	oxide defect	silicon defect
1	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
2	before	OCT 1	1	OCT 1 before	0	0	0	0	0	1	0
3	before	OCT 1	1	OCT 1 before	1	0	0	0	0	1	0
4	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
5	before	OCT 1	1	OCT 1 before	1	0	0	0	0	0	0
6	before	OCT 1	1	OCT 1 before	0	0	0	0	0	1	0
7	before	OCT 1	1	OCT 1 before	1	0	0	0	0	0	0
8	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
9	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
10	before	OCT 1	1	OCT 1 before	1	0	0	1	0	0	0
11	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
12	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
13	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
14	before	OCT 1	1	OCT 1 before	1	0	0	0	0	0	0
15	before	OCT 1	1	OCT 1 before	0	0	0	0	0	0	0
16	before	OCT 1	1	OCT 1 before	0	0	0	0	1	0	0

1. Select **Help > Sample Data Library** and open Quality Control/Failures3Indicators.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select the defect columns (contamination, corrosion, doping, metallization, miscellaneous, oxide defect, silicon defect) and click **Indicator Group** on the Multiple tab.
4. Select clean and date and click **X, Grouping Category.**
5. Click **OK.**

When you click **OK**, you get the same output as in the Response Group example (Figure 3.24).

Multiple Delimited

Suppose that an inspector entered the observed defects for each unit. The defects are listed in a single column, delimited by a comma, as in Failures3Delimited.jmp. Note in the partial data table, shown below, that some units did not have any observed defects, so the failureS column is empty.

Figure 3.26 Failure3Delimited.jmp Data Table

	failureS	clean	date	ID	ID Label
1		before	OCT 1	1	OCT 1 before
2	oxide defect	before	OCT 1	1	OCT 1 before
3	contamination,oxide defect	before	OCT 1	1	OCT 1 before
4		before	OCT 1	1	OCT 1 before
5	contamination	before	OCT 1	1	OCT 1 before
6	oxide defect	before	OCT 1	1	OCT 1 before
7	contamination	before	OCT 1	1	OCT 1 before
8		before	OCT 1	1	OCT 1 before
9		before	OCT 1	1	OCT 1 before
10	metallization,contamination	before	OCT 1	1	OCT 1 before
11		before	OCT 1	1	OCT 1 before
12		before	OCT 1	1	OCT 1 before
13		before	OCT 1	1	OCT 1 before
14	contamination	before	OCT 1	1	OCT 1 before

1. Select **Help > Sample Data Library** and open Quality Control/ Failures3Delimited.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select failureS and click **Multiple Delimited** on the Multiple tab.
4. Select clean and date and click **X, Grouping Category.**
5. Select ID and click **ID.**
6. Click **OK.**

When you click **OK**, you get the same output as in Figure 3.24.

Note: If more than one delimited column is specified, separate analyses are produced for each column.

Multiple Response by ID

Suppose each failure type is a separate record, with an ID column that can be used to link together different defect types for each unit, as in Failure3ID.jmp.

Figure 3.27 Failure3ID.jmp Data Table

	failure	N	clean	date	SampleSize	ID
1	contamination	14	before	OCT 1	50	OCT 1 before
2	corrosion	2	before	OCT 1	50	OCT 1 before
3	doping	1	before	OCT 1	50	OCT 1 before
4	metallization	2	before	OCT 1	50	OCT 1 before
5	miscellaneous	3	before	OCT 1	50	OCT 1 before
6	oxide defect	8	before	OCT 1	50	OCT 1 before
7	silicon defect	1	before	OCT 1	50	OCT 1 before
8	doping	0	after	OCT 1	50	OCT 1 after
9	corrosion	2	after	OCT 1	50	OCT 1 after
10	metallization	4	after	OCT 1	50	OCT 1 after

1. Select **Help > Sample Data Library** and open Quality Control/Failure3ID.jmp.
2. Select **Analyze** > **Consumer Research > Categorical**.
3. Select failure and click **Multiple Response by ID** on the Multiple tab.
4. Select clean and date and click **X, Grouping Category.**
5. Select SampleSize and click **Sample Size.**
6. Select N and click **Freq.**
7. Select ID and click **ID.**
8. Click **OK.**

When you click **OK**, you get the same output as in Figure 3.24.

Mean Score Example

You can calculate response means in your data using Value Scores. To make the Mean Score interpretable, you can assign specific value scores in the Column Info window with the Value Scores column property. For more information about column properties, refer to The Column Info Window chapter in the *Using JMP* book.

In this example, you can assign Value Scores to calculate the Net Promoter Score (Reichheld, HBR 2003), which summarizes an 11-level rating with a favorability score between -100 and 100. Anything with a value of 6 or below is regarded as a detractor.

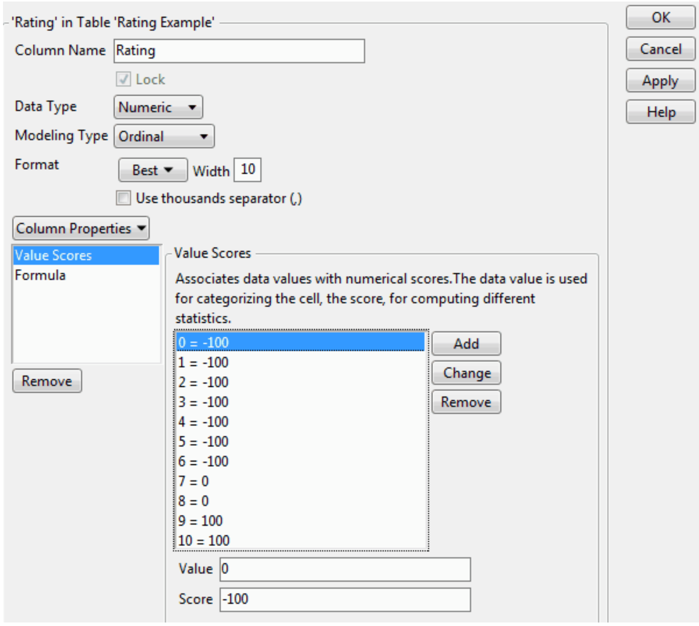
1. Run the following script:

```
New Table( "Rating Example",
   Add Rows( 300 ),
   New Script(
      "Categorical",
      Categorical( Responses( :Rating ), Mean Score( 1 ) )
   ),
   New Column( "Rating",
      Numeric,
      Ordinal,
      Set Property(
         "Value Scores",
         {0 = -100, 1 = -100, 2 = -100, 3 = -100, 4 = -100, 5 = -100, 6
          = -100, 7 = 0, 8 = 0, 9 = 100, 10 = 100}
      ),
      Formula(
         Random Category(
            0.05,
            0,
            0.05,
            1,
```

```
                0.05,
                2,
                0.05,
                3,
                0.05,
                4,
                0.05,
                5,
                0.05,
                6,
                0.05,
                7,
                0.05,
                8,
                0.3,
                9,
                0.25,
                10
            )
        ),
        Set Selected
    )
);
```

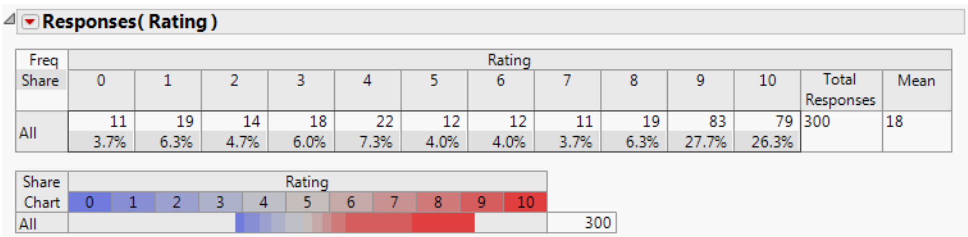
2. A data table with 300 rows of random rating data is created. Value scores were also defined for the Rating column. To view the scores, right-click the Rating column and select **Column Properties > Value Scores**.

Figure 3.28 Column Properties - Value Scores



3. Select **Analyze** > **Consumer Research > Categorical**.
4. Select Rating and click **Responses** on the Simple tab.
5. Click **OK.**
6. Select **Mean Score** from the Categorical red triangle menu.

Figure 3.29 Rating Example Report



Freq Share	Rating 0	1	2	3	4	5	6	7	8	9	10	Total Responses	Mean
All	11 3.7%	19 6.3%	14 4.7%	18 6.0%	22 7.3%	12 4.0%	12 4.0%	11 3.7%	19 6.3%	83 27.7%	79 26.3%	300	18

Based on the defined value scores, a mean score of 18 was determined. Your results might be different as the Rating column values are random. (Note that the colors shown are JMP default colors.)

Chapter 4

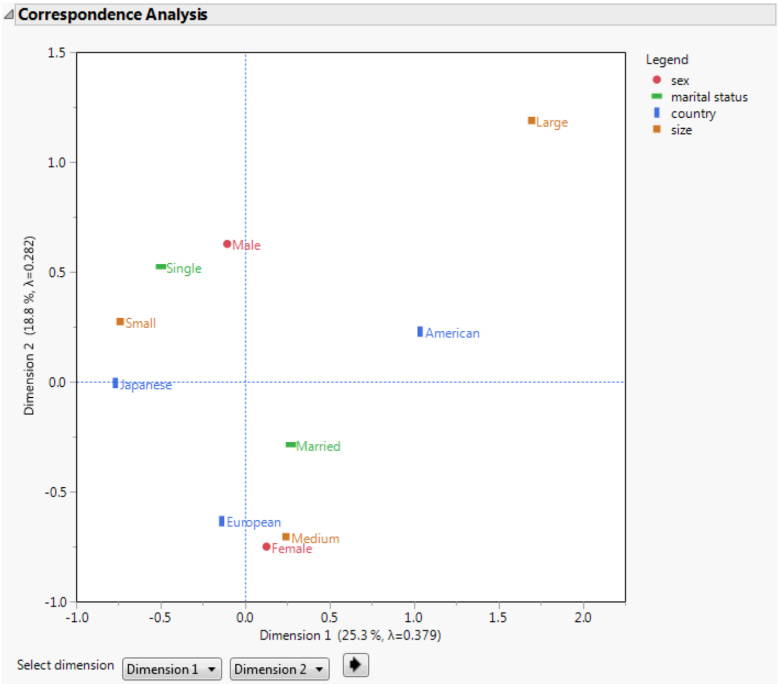
Multiple Correspondence Analysis

Identify Associations between Levels of Categorical Variables

Multiple Correspondence Analysis (MCA) takes multiple categorical variables and seeks to identify associations between levels of those variables. MCA extends correspondence analysis from two variables to many. It can be thought of as analogous to principal component analysis for quantitative variables. Similar to other multivariate methods, it is a dimension reducing method; it represents the data as points in 2- or 3-dimensional space.

Multiple correspondence analysis is frequently used in the social sciences particularly in France and Japan. It can be used in survey analysis to identify question agreement. It is also used in consumer research to identify potential markets for products. Microarray studies in genetics also use MCA to identify potential relationships between genes.

Figure 4.1 Multiple Correspondence Analysis



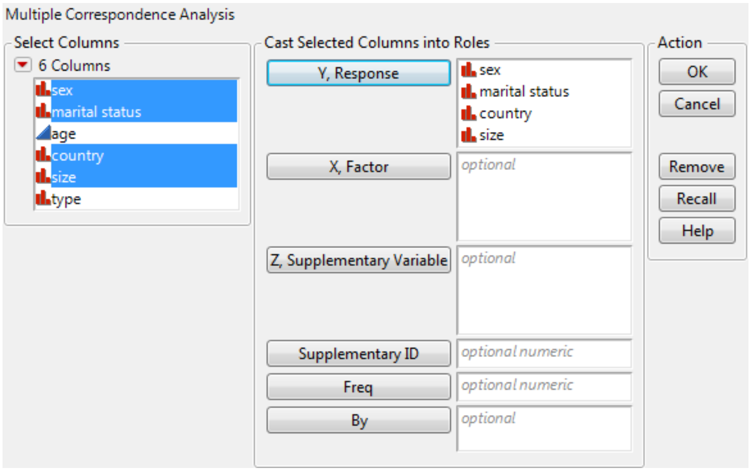
Example of Multiple Correspondence Analysis

This example uses the Car Poll.jmp sample data table, which contains data collected from car polls. The data include aspects about the individuals polled, such as sex, marital status, and age. The data also include aspects about the car that they own, such as the country of origin, the size, and the type. You want to explore relationships between sex, marital status, country and size of car to identify consumer preferences.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Consumer Research > Multiple Correspondence Analysis**.
3. Select sex, marital status, country, and size and click **Y, Response**.

   In MCA, usually all columns are considered responses rather than some being responses and others explanatory.
4. Click **OK**.

Figure 4.2 Completed Multiple Correspondence Analysis Launch Window

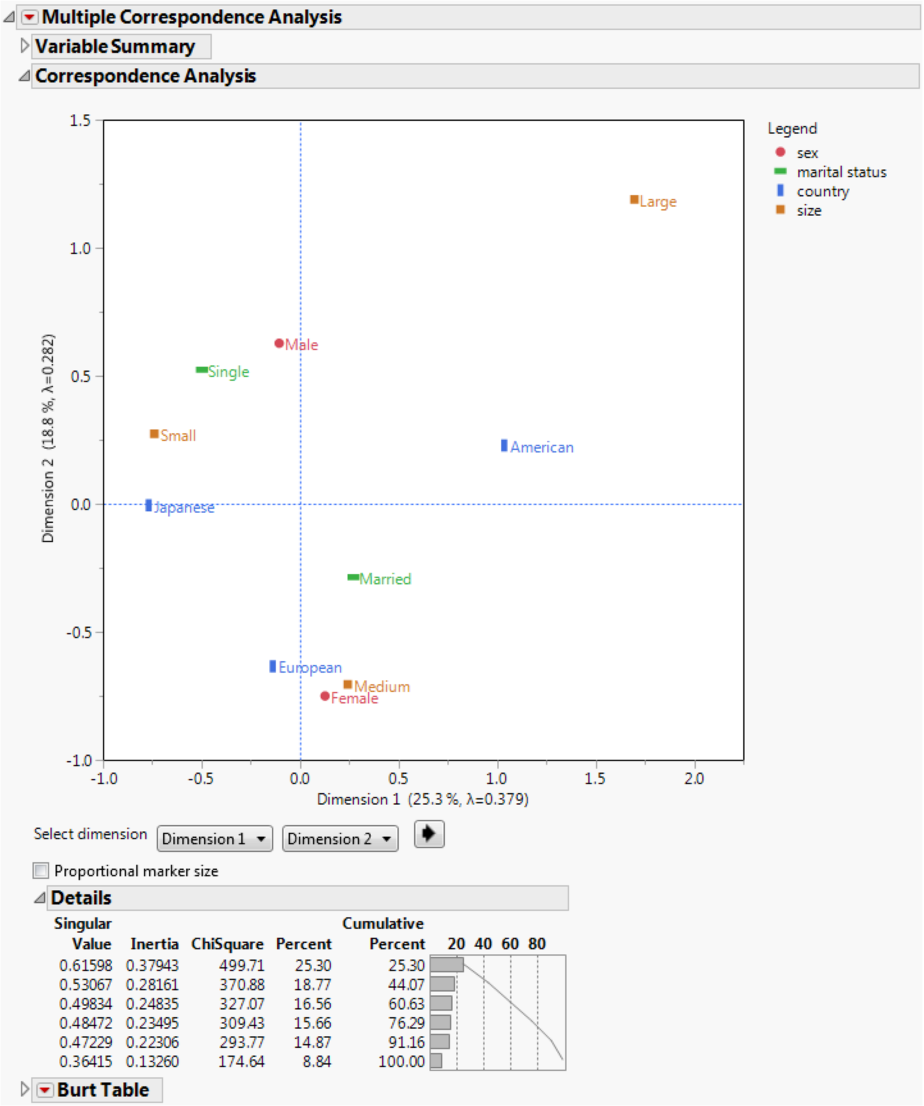


The Multiple Correspondence Analysis report is shown in Figure 4.3. Note that some of the outlines are closed because of space considerations.

The Variable Summary report provides a concise view of the analysis completed.

The Correspondence Analysis report shows the cloud of categories of the four variables as projected onto the two principal axes. From this cloud, you can see that Americans have a strong association with the large car size while Japanese are highly associated with the small car size. Also, males are strongly associated with the small car type and females are associated with the medium car size. This information could be used in market research to identify target audiences for advertisements.

Figure 4.3 Multiple Correspondence Analysis Report

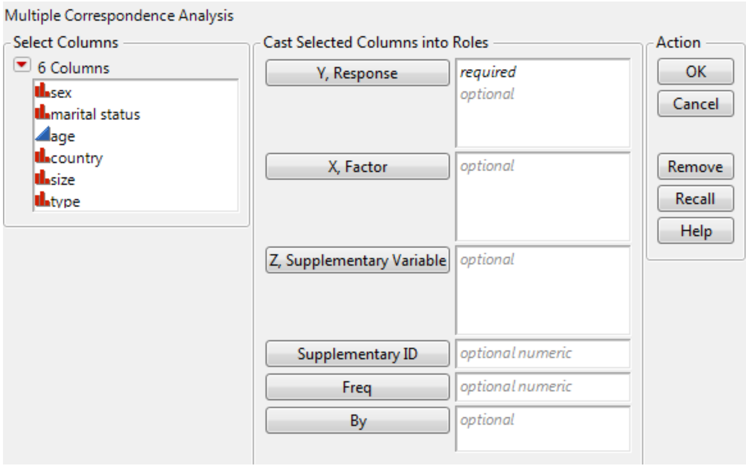


Singular Value	Inertia	ChiSquare	Percent	Cumulative Percent
0.61598	0.37943	499.71	25.30	25.30
0.53067	0.28161	370.88	18.77	44.07
0.49834	0.24835	327.07	16.56	60.63
0.48472	0.23495	309.43	15.66	76.29
0.47229	0.22306	293.77	14.87	91.16
0.36415	0.13260	174.64	8.84	100.00

Launch the Multiple Correspondence Analysis Platform

Launch the Multiple Correspondence Analysis platform by selecting **Analyze > Consumer Research > Multiple Correspondence Analysis**.

Figure 4.4 Multiple Correspondence Analysis Launch Window



Y, Response Assigns the categorical columns to be analyzed. In MCA, you are generally interested in the associations between variables, but there are not explicit "explanatory" and "response" variables.

X, Factor Assigns the categorical columns to be used as factor, or explanatory, variables.

Z, Supplementary Variable Assigns the columns to be used as supplementary variables. These variables are those you are interested in identifying associations with but not include in the calculations.

Supplementary ID Assigns the column that identifies rows to be used as supplementary. A supplementary ID column usually has 1s and 0s. The rows associated with ID 0 are treated as supplementary rows. The Supplementary ID column is ignored if there are levels of the X or Y variables present in the supplementary rows that are not present in the non-supplementary rows.

Note: Only one of the Supplementary ID and Z, Supplementary Variable roles can be specified at one time.

Freq Assigns a frequency variable to this role. This is useful if your data are summarized.

By Produces a separate report for each level of the By variable. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

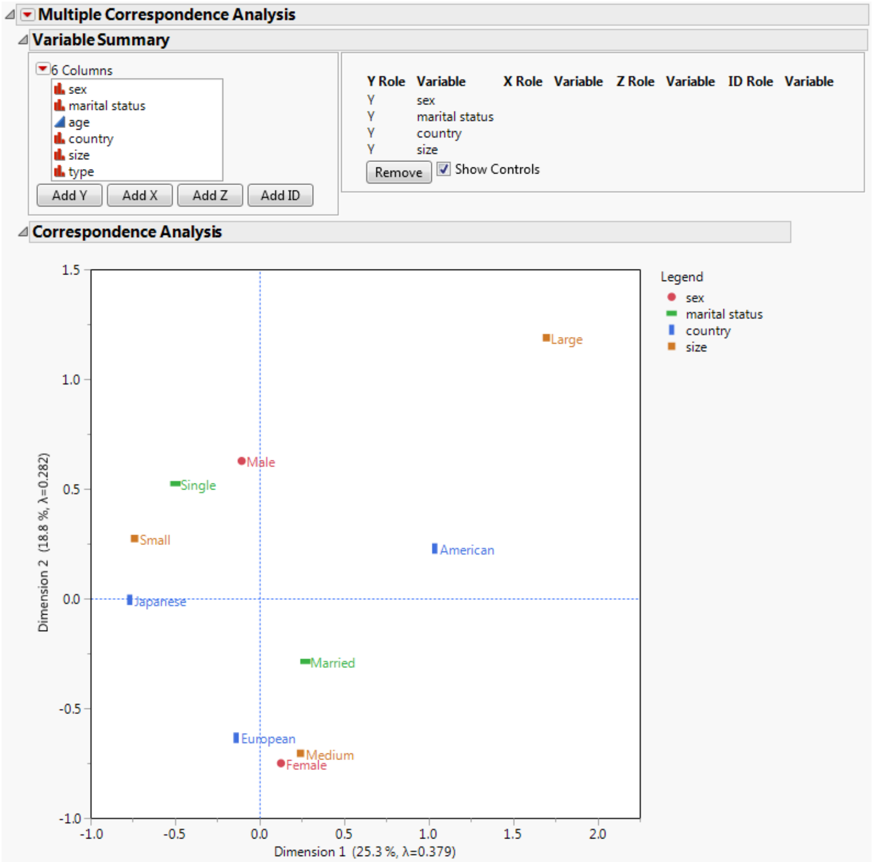
Note: The Multiple Correspondence Analysis platform handles missing values differently than many other JMP platforms. The analysis uses all nonmissing pairs of cells in a row. It does not remove entire rows from the computation.

The Multiple Correspondence Analysis Report

The initial Multiple Correspondence Analysis report shows the variable summary, correspondence analysis plot, and details of the dimensions of the data in order of importance. From the plot of the cloud of categories or individuals, you can identify associations that exist within the data. The details provide information about whether the two dimensions shown in the plot are sufficient to understand the relationships within the table.

The Variable Summary shows the columns used in the analysis and the roles that you selected in the launch window. If you select the Show Controls check box, a list of the columns in the data table appears to the left. You can change the columns in the analysis either by selecting a column and clicking Add Y, Add X, Add Z, or Add ID. Or you can drag the column to the header in the variable summary table. This enables you to modify the analysis without returning to the launch window.

Figure 4.5 Multiple Correspondence Analysis Report with Show Controls Selected



Multiple Correspondence Analysis Platform Options

The Multiple Correspondence Analysis red triangle menu options give you the ability to customize reports according to your needs. The reports available are determined by the type of analysis that you conduct.

Correspondence Analysis Provides correspondence analysis reports. These reports give the plots, details, coordinates, and summary statistics. See "Additional Examples of the Multiple Correspondence Analysis Platform" on page 80.

Cross Table Provides the Burt table or contingency table as appropriate for variable roles selected. See "Cross Table" on page 79.

Cross Table of Supplementary Rows Provides a contingency table of the supplementary variable(s) versus the response variable(s). This table appears by default only if a supplementary variable has been specified in the launch window.

Cross Table of Supplementary Columns Provides a contingency table of the X, Factor variable(s) versus the supplementary variable(s). This table appears by default only if a factor variable and a supplementary variable have been specified in the launch window.

Mosaic Plot Displays a mosaic bar chart for each nominal or ordinal response variable. A mosaic plot is a stacked bar chart where each segment is proportional to its group's frequency count. This option is available if only one Y and only one X variable are selected.

Tests for Independence Provides the tests for independence whether there is association between the row and column variables. There are two versions of this test, the Pearson form and the Likelihood Ratio form, both with chi-square statistics. This option is available only when there is one Y variable and one X variable.

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.

Correspondence Analysis Options

The reports available under Correspondence Analysis are determined by the type of analysis that you conduct. Several of these reports are shown by default.

Show Plot Shows the two-dimensional cloud of categories in the plane described by the first two principal axes. This plot appears by default.

Show Detail Provides the details of the analysis including the singular values, inertias, ChiSquare statistics, percent, and cumulative percent. This report appears by default. See "Show Detail" on page 76.

Show Adjusted Inertia Provides reports of the Benzecri and Greenacre adjusted inertia. See Benzecri (1979) and Greenacre (1984). This option is not available when there are one or more X variables. See "Show Adjusted Inertia" on page 76.

Show Coordinates Provides a report of up to the first three principal coordinates for the categories in the analysis, as appropriate. See "Show Coordinates" on page 77.

Show Summary Statistics Provides a report of the summary statistics, Quality, Mass, and Inertia, for each category in the analysis. See "Show Summary Statistics" on page 77.

Show Partial Contributions to Inertia Provides a report of the contribution of each category to the inertia for each of up to the first three dimensions. See "Show Partial Contributions to Inertia" on page 78.

Show Squared Cosines Provides a report of the squared cosines of each category for each of up to the first three dimensions. The report includes a bar chart that shows, for each level of each Y variable, the squared cosine values for each of up to the first three dimensions. See "Show Squared Cosines" on page 78.

Cochran's Q Test (Available only when all of the Y variables have the same set of only two levels and the X variable has a unique value for each row.) Provides Cochran's *Q* statistic, which tests that the marginal probability of a specific response is unchanged across the Y variables. Cochran's *Q* statistic is a generalization of McNemar's statistic for more than two response variables. See Agresti (2002).

3D Correspondence Analysis Shows the three-dimensional cloud of categories of the Y, X, and Z variables in the space described by the first three principal axes. This option is not available if there are less than three dimensions.

Save Coordinates Saves the principal coordinates to one or more JMP data tables. Column coordinates, row coordinates, supplementary column coordinates, and supplementary row coordinates are saved to separate JMP data tables. You can choose how many columns to save.

Save Coordinate Formula Saves formula columns to the data table for the principal coordinates in multiple dimensions. The value for each observation is the average of the coordinates for the Y variables scaled by the singular value for each dimension. You can choose how many columns to save.

Show Plot

The plot displays a projection of the cloud of categories or individuals onto the plane described by the first two principal axes. The distance scale is the same in all directions. You can toggle the dimensions shown in the plot using the Select Dimension controls below the plot. The first control defines the horizontal axis of the plot, and the second control defines the vertical axis of the plot. Click the arrow button to cycle through the dimensions shown in the plot. Below the Select Dimension controls, you can specify if the size of the points in the plot should be proportional to the count of observations corresponding to each point.

Note: Selecting a point in the correspondence analysis plot also selects the corresponding rows in other tables in the report window. However, rows in the data table are not selected. To select all of the points in the plot associated with a particular variable, select the name of the variable in the plot legend.

Show Detail

Shows the table of singular values.

Singular Value Shows the singular values in the singular value decomposition of the contingency table or Burt table. For the formula, see "Details Report" on page 84.

Inertia Lists the square of the singular values, reflecting the relative variation accounted for in the canonical dimensions.

ChiSquare Lists the portion of the overall Chi-square for the Burt or contingency table represented by the dimension.

Percent Portion of inertia with respect to the total inertia.

Cumulative Percent Shows the cumulative portion of inertia. If the first two singular values capture the bulk of the inertia, then the 2-D correspondence analysis plot is sufficient to show the relationships in the table.

Show Adjusted Inertia

The principal inertias of a Burt table in MCA are the eigenvalues. The problem with these inertias is that they provide a pessimistic indication of fit. Benzécri proposed an inertia adjustment. Greenacre argued that the Benzécri adjustment overestimates the quality of fit and proposed an alternate adjustment. Both adjustments are calculated for your reference. See "Adjusted Inertia" on page 84.

Inertia Lists the square of the singular values, reflecting the relative variation accounted for in the canonical dimensions.

Adjusted Inertia Lists the adjusted inertia according to either the Benzécri or Greenacre adjustment.

Percent Portion of adjusted inertia with respect to the total inertia.

Cumulative Percent Shows the cumulative portion of adjusted inertia. If the first two singular values capture the bulk of the inertia, then the 2-D correspondence analysis plot is sufficient to show the relationships in the table.

Show Coordinates

Shows the Column Coordinates table or the Row and Column Coordinates tables.

X Lists the columns specified as X, Factor variables.

Y Lists the columns specified as Y, Response variables.

Z Lists the columns specified as Z, Supplementary Variables.

Category Lists the levels of the X, Y, or Z variables.

Dimension 1, Dimension 2, Dimension 3 For each level or each response, lists its coordinate along the indicated principal axis. By default, the tables show coordinates for up to the first three dimensions. Coordinate columns for additional dimensions are hidden. To show these optional columns, right-click in a table and select the dimension columns from the **Columns** submenu.

Note: If there are columns specified as X, Factor variables, the Coordinates report displays tables of both X and Y with the same report headings. If a Z, Supplementary Variable is specified, the coordinates are listed below the X and Y coordinates as applicable.

Show Summary Statistics

Shows the Summary Statistics for the Column Points table or the Summary Statistics for the Row and Column Points tables. The Y table gives Quality, Mass, and Inertia for each level of each response, called a *column point*. The X table gives Quality, Mass, and Inertia for each level of the X, Factor variables. See “Summary Statistics” on page 85.

X Lists the columns specified as X, Factor variables.

Y Lists the columns specified as Y, Response variables.

Category Lists the levels of the X or Y variables.

Quality(dim=2) Lists the quality of the representation of the level by the solution.

Mass Lists the row frequency for the level of the response divided by the total frequency. In the Burt table, this is the Total % for each row.

Inertia Lists the proportion of the total inertia accounted for by the level of the response. The inertia values sum to one across the levels and their responses.

Note: If there are columns specified as X, Factor variables, the Summary Statistics report displays tables of both X and Y with the same report headings.

Show Partial Contributions to Inertia

Shows the Partial Contributions to Inertia for the Column Points table or the Partial Contributions to Inertia for the Row and Column Points tables. Also shows the Plot of Partial Contributions to Inertia for the Column Points. This is a bar chart that shows, for each level of each Y variable, its partial contributions to each of the dimensions shown in the table.

X Lists the columns specified as X, Factor variables.

Y Lists the columns specified as Y, Response variables.

Category Lists the levels of the X or Y variables.

Dimension 1, Dimension 2, Dimension 3 Lists the contribution of the response or factor level to the inertia of the indicated dimension. By default, the tables show columns for up to the first three dimensions. Additional columns are hidden. To show these optional columns, right-click on a table and select the dimension columns from the **Columns** submenu.

Each level of each response contributes to the inertia of each dimension. The partial contributions within each dimension sum to 1.

Note: If there are columns specified as X, Factor variables, the Partial Contributions to Inertia report displays tables of both X and Y with the same report headings. See "Partial Contributions to Inertia" on page 85.

Show Squared Cosines

Shows the Squared Cosines for the Column Points table or the Squared Cosines for the Row and Column Points. Also shows the Plot of Squared Cosines for the Column Points. This is a bar chart that shows, for each level of each Y variable, the squared cosine values for each of up to the first three dimensions shown.

X Lists the columns specified as X, Factor variables.

Y Lists the columns specified as Y, Response variables.

Category Lists the levels of the X or Y variables.

Dimension 1, Dimension 2, Dimension 3 Lists the *quality* of the representation of the level by the indicated dimension. By default, the tables show results for up to the first three dimensions. Additional columns are hidden. To show these optional columns, right-click on a table and select the dimension columns from the **Columns** submenu.

The values indicate the quality of each point for the indicated dimension. The squared cosine can be interpreted as the squared correlation of the point with the dimension. The sum of the squared cosines of the first two dimensions equalsQuality(dim=2) in the Summary Statistics report. See "Summary Statistics" on page 85.

Note: If there are columns specified as X, Factor variables, the Squared Cosines report displays tables of both X and Y with the same report headings.

Cross Table

The Burt table is the basis of the multiple correspondence analysis. It is a partitioned symmetric table of all pairs of categorical variables. The diagonal partitions are diagonal matrices (a cross-table of a variable with itself). The off-diagonal partitions are ordinary contingency tables. When you select multiple Y, Response columns with no X, Factor columns, the Burt table is created. If you select any X, Factor columns, a traditional contingency table is created instead of a Burt table.

The red triangle menu for the Burt or contingency table contains options of statistics to display in the table.

Count Cell frequency, margin total frequencies, and grand total (total sample size). This appears by default.

Total % Percent of cell counts and margin totals to the grand total. This appears by default.

Cell Chi Square Chi-square values computed for each cell as $(O - E)^2 / E$.

Col % Percent of each cell count to its column total.

Row % Percent of each cell count to its row total.

Expected Expected frequency (E) of each cell under the assumption of independence. Computed as the product of the corresponding row total and column total divided by the grand total.

Deviation Observed cell frequency (O) minus the expected cell frequency (E).

Col Cum Cumulative column total.

Col Cum % Cumulative column percentage.

Row Cum Cumulative row total.

Row Cum % Cumulative row percentage.

Make Into Data Table Creates one data table for each statistic shown in the table.

Cross Table of Supplementary Rows

When a Z, Supplementary column is selected, a contingency table with the supplementary column levels as the rows and the response column levels as the columns is created. The red triangle menu contains the same options as the Burt Table.

Cross Table of Supplementary Columns

When an X, Factor column and a Z, Supplementary column are selected, a contingency table with the X, Factor levels as rows and the Supplementary levels as columns is created. The red triangle menu contains the same options as the Burt Table.

Additional Examples of the Multiple Correspondence Analysis Platform

Example Using a Supplementary Variable

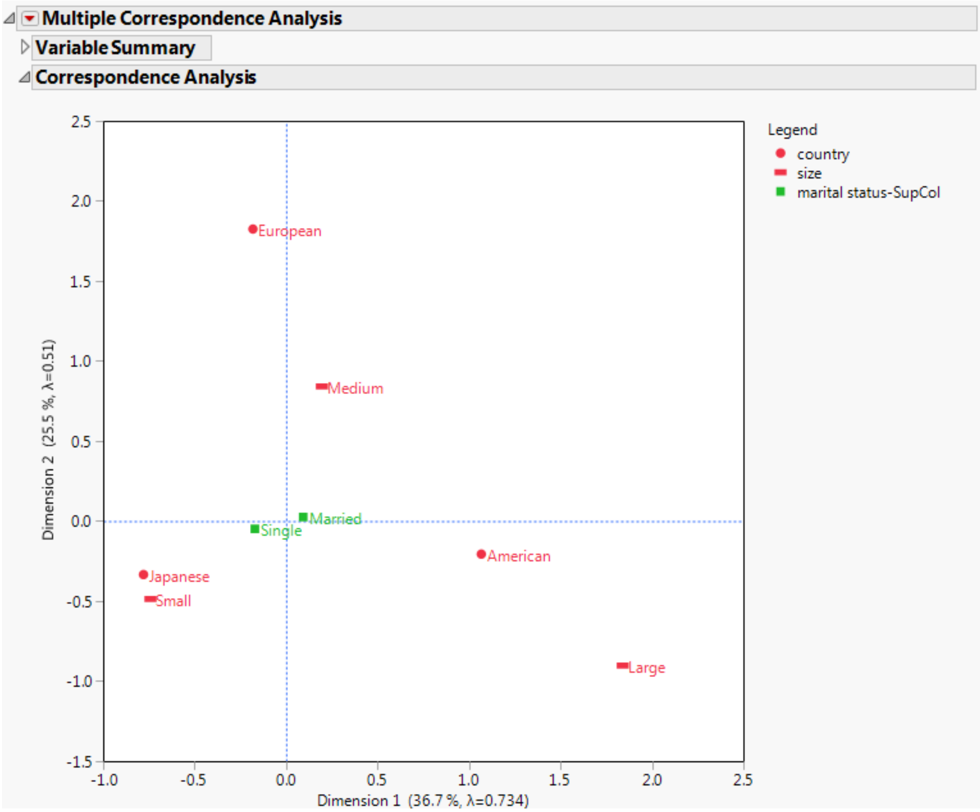
This example uses the Car Poll.jmp sample data table, which contains data collected from car polls. The data include aspects about the individuals polled, such as sex, marital status, and age. The data also include aspects about the car that they own, such as the country of origin, the size, and the type. You want to explore relationships between sex, country, and size of car to identify consumer preferences.

1. Select **Help > Sample Data Library** and open Car Poll.jmp.
2. Select **Analyze > Consumer Research > Multiple Correspondence Analysis**.
3. Select country and size and click **Y, Response**.
4. Select marital status and click **Z, Supplementary Variable**.
5. Click **OK**.

Unlike in the first example, this analysis does not use marital status in the calculations. Marital status is plotted after the calculations are complete.

You see from the plot strong relationships between Japanese and Small cars as well as American and Large cars. The two marital statuses are plotted in a different color. Single people seem to prefer smaller cars a bit more than married people.

Figure 4.6 MCA with Supplementary Variable Report



Example Using a Supplementary ID

The United States census allows for examining population growth over the last century. The US Regional Population.jmp sample data table contains populations of the 50 US states grouped into regions for each of the census years from 1920 to 2010. Alaska and Hawaii are treated as supplementary regions because they were not states during the entire time, and they are not part of the contiguous United States. You are interested in whether the population growth in these two states differs from the rest of the US.

1. Select **Help > Sample Data Library** and open US Regional Population.jmp.
2. Select **Analyze > Consumer Research > Multiple Correspondence Analysis**.
3. Select Year and click **Y, Response**.
4. Select Region and click **X, Factor**.
5. Select ID and click **Supplementary ID**.
6. Select Population and click **Freq**.

7. Click **OK**.

The Details report shows that the association between years and regions is almost entirely explained by the first dimension. The plot shows that years are in the correct order on the first dimension. This ordering occurs naturally through the correspondence analysis; there is no information about the order provided to the analysis.

Notice that the ordering of the regions reflects the population shift from the Midwest to the Northeast to the South and finally to the Mountain and West.

Alaska and Hawaii were not used in the computation of the analysis but are plotted based on the results. Their growth pattern is most similar to the Pacific states. Alaska's growth is even more extreme than the Pacific region.

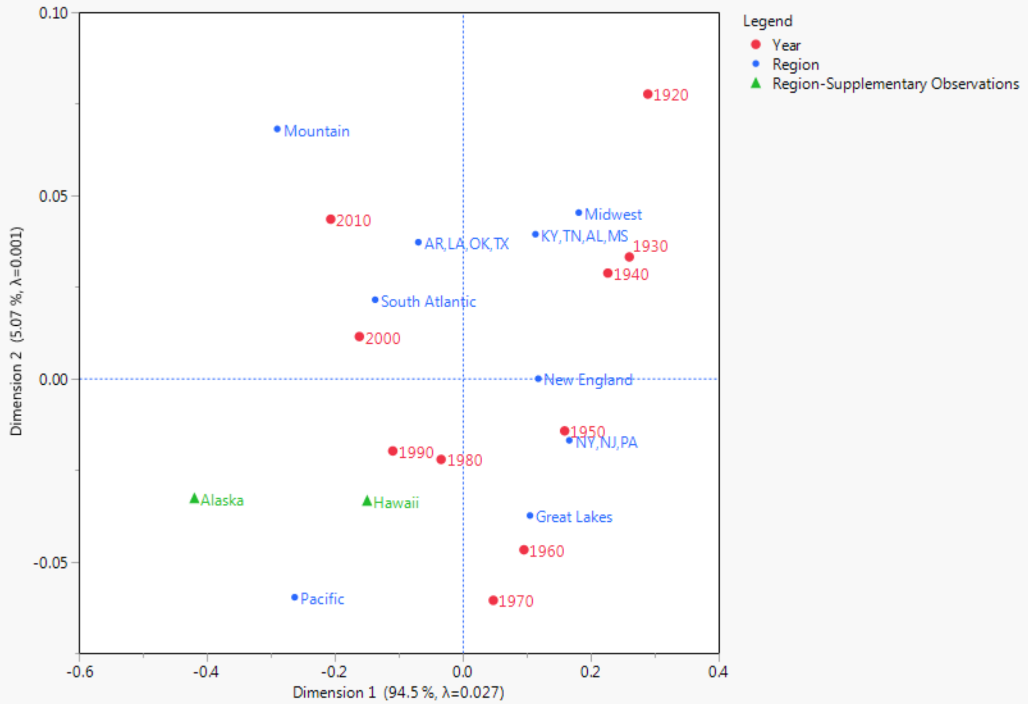
Figure 4.7 MCA with Supplementary ID Report

Multiple Correspondence Analysis

Variable Summary

Supplementary ID: ID
Freq: Population

Correspondence Analysis



Select dimension Dimension 1 Dimension 2

Proportional marker size

Details

Singular Value	Inertia	ChiSquare	Percent	Cumulative Percent
0.16553	0.02740	53425	94.46	94.46
0.03836	0.00147	2869	5.07	99.53
0.00738	5.45e-5	106	0.19	99.72
0.00704	0.00005	97	0.17	99.89
0.00438	0.00002	37	0.07	99.95
0.00316	0.00001	19	0.03	99.99
0.00169	2.87e-6	5.58669	0.01	100.00
0.00086	7.46e-7	1.45505	0.00	100.00

Statistical Details for the Multiple Correspondence Analysis Platform

This section contains statistical details for the Multiple Correspondence Analysis.

Details Report

When a simple Correspondence Analysis is performed, the report lists the singular values of the following equation:

$$\mathbf{D}_r^{-0.5}(\mathbf{P} - rc')\mathbf{D}_c^{-0.5} = \mathbf{U}\mathbf{D}_u\mathbf{V}'$$

where:

- $\mathbf{P}$ is the matrix of counts divided by the total frequency
- r and c are the row and column sums of $\mathbf{P}$
- the $\mathbf{D}$ matrices are diagonal matrices of the values of r and c

When Multiple Correspondence Analysis is performed, the singular value decomposition extends to:

$$\mathbf{D}^{-0.5}\left(\frac{\mathbf{C}}{Q^2 n} - \mathbf{D11'D}\right)\mathbf{D}^{-0.5} = \mathbf{U}\mathbf{D}_u\mathbf{V}'$$

where:

$$\mathbf{D} = \left(\frac{1}{m}\right)\mathrm{diag}(\mathbf{D}_1, \mathbf{D}_2, \ldots, \mathbf{D}_Q)$$

- $\mathbf{C}$ is the Burt table.
- Q is the number of categorical variables
- n is the number of observations
- $\mathbf{1}$ is a column vector of ones

Adjusted Inertia

The usual principal inertias of a Burt table constructed from m categorical variables in MCA are the eigenvalues u_k from $\mathbf{D}_u^2$. These inertias provide a pessimistic indication of fit. Benzécri (1979) proposed the following inertia adjustment; it is also described by Greenacre (1984, p. 145):

$$\left(\frac{m}{m-1}\right)^2 \times \left(u_k - \frac{1}{m}\right)^2 \text{ for } u_k > \frac{1}{m}$$

This adjustment computes the percent of adjusted inertia relative to the sum of the adjusted inertias for all inertias greater than $1/m$.

Greenacre (1994, p. 156) argues that the Benzécri adjustment overestimates the quality of fit. Greenacre proposes instead to compute the percentage of adjusted inertia relative to:

$$\frac{m}{m-1}\left(\mathrm{trace}(\mathbf{D}_u^4) - \frac{n_c - m}{m^2}\right)$$

for all inertias greater than $1/m$, where $\mathrm{trace}(\mathbf{D}_u^4)$ is the sum of squared inertias and n_c is the total number of categories across the m variables.

Summary Statistics

Quality is the ratio of the squared distance of a point from the origin in the space defined by the specified number of dimensions to the distance from the origin in the space with the maximum number of dimensions. For the Chi-Square metric, a point's quality in a given dimension can be obtained from the cosine that its vector makes with the vector that defines the dimension. Quality is also equal to the ratio of the sum of inertias in the specified dimensions to the sum of the inertias in all dimensions. Quality indicates how well the point is represented in the lower-dimensional space.

Mass is the proportion of row or column total frequency to the total frequency.

Inertia is analogous to variance in principal component analysis. The overall inertia is the total Pearson Chi-square for a two-way frequency table divided by the sum of all observations in the table.

Relative inertia is the proportion of the contribution of the point to the overall inertia. In the summary statistics table, the relative inertia is listed in the column labeled Inertia.

Partial Contributions to Inertia

The contribution of a row or column to the inertia of a dimension is calculated as:

$$\mathrm{contribution} = \frac{\mathrm{mass} \times \mathrm{coordinate}^2}{\mathrm{dimensioninertia}}$$

Chapter 5

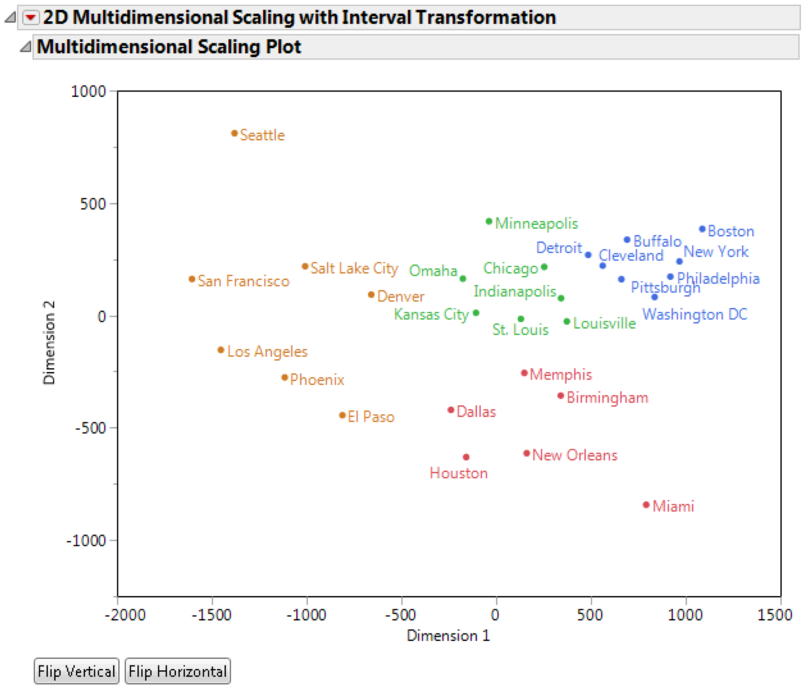
Multidimensional Scaling

Visualize Proximities among a Set of Objects

Multidimensional Scaling (MDS) is a technique that is used to create a visual representation of the pattern of proximities (similarities, dissimilarities, or distances) among a set of objects. For example, given a matrix of distances between cities, MDS can be used to generate a map of the cities in two dimensions.

Multidimensional Scaling is frequently used in consumer research where researchers have measures of perceptions about brands, tastes, or other product attributes. MDS is applicable to many other areas where one is interested in visualizing the proximity of objects based on a set of attributes or proximities.

Figure 5.1 Multidimensional Scaling Example



Multidimensional Scaling Platform Overview

The Multidimensional Scaling platform generates a plot of proximities among a set of objects. This plot can be used to visually explore structure in a data set. MDS is a multivariate technique that is used to visualize the patterns of proximities (distances, similarities) among a set of objects in a small number of dimensions. MDS is applied to a distance matrix. The coordinates for the MDS plot are obtained by minimizing a stress function (the difference between the actual and predicted proximities).

The term distance can refer to a measure of physical distance, such as between cities. More often distance is a subjective assessment rather than a precise measurement. Proximities can measure perceived similarities between brands of a product, correlations of crime rates, or economic similarities for a sample of countries. Distance can also be called proximity or similarity (dissimilarity). If the data are given as an attribute list, then a distance matrix is first constructed from the correlation structure of the attribute list.

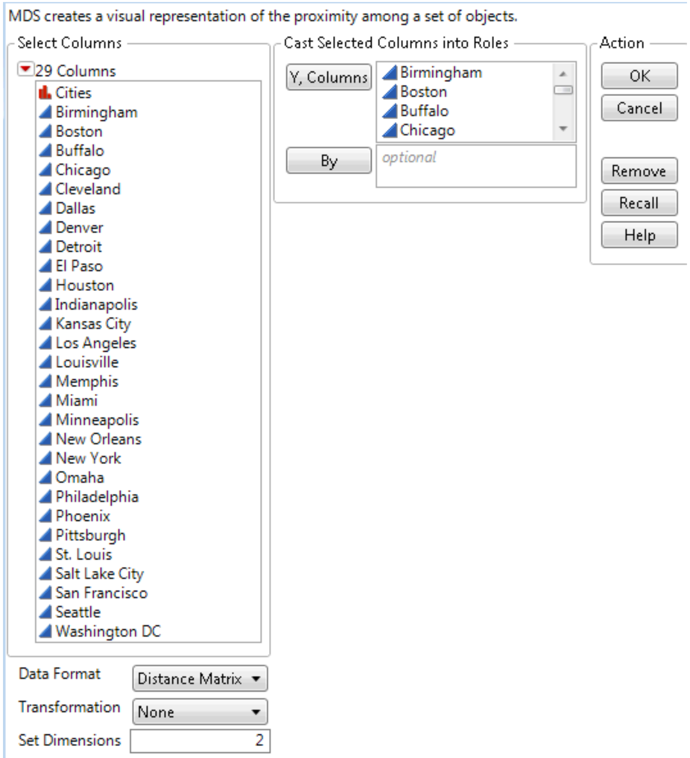
For more information about multidimensional scaling see Borg and Groenen(2005) or Jackson(2003).

Example of Multidimensional Scaling

This example uses the Flight Distances.jmp sample data table, which is a distance matrix of flight distances between 28 US cities. You can use MDS to construct a map of the cities in two dimensions that is based on the pairwise distances in the data table.

1. Select **Help > Sample Data Library** and open Flight Distances.jmp.
2. Select **Analyze > Consumer Research > Multidimensional Scaling.**
3. Select Birmingham through Washington DC and click **Y, Columns.**

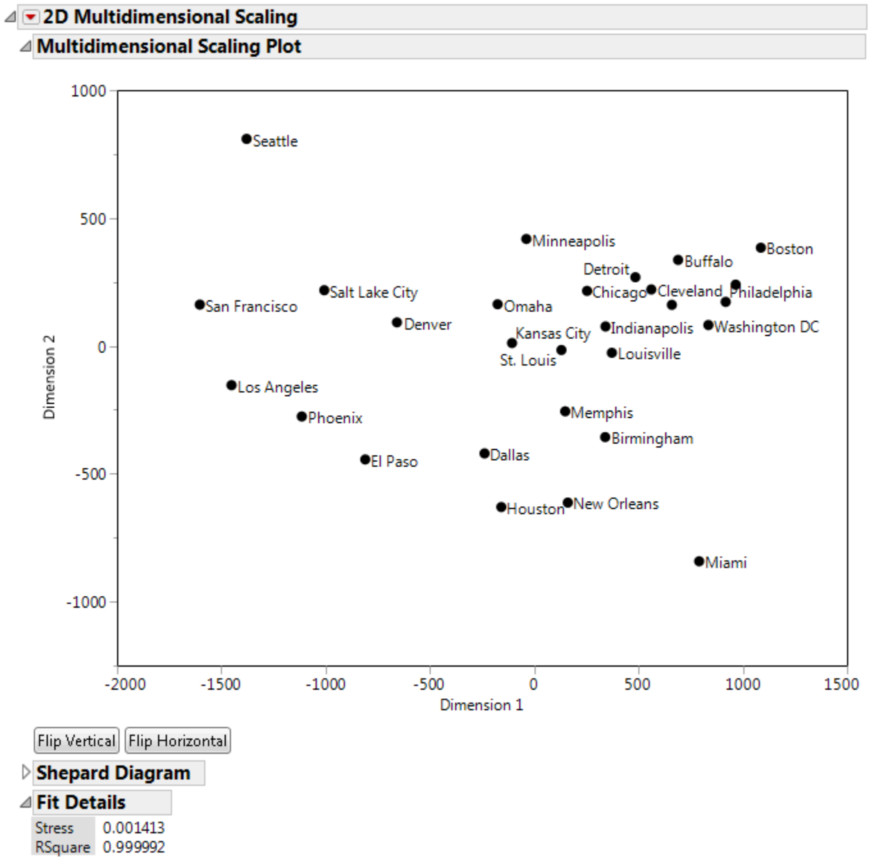
Figure 5.2 Completed Multidimensional Scaling Launch Window



4. Click **OK.**
5. Select the Flight Distances data table.
6. Right-click on the column Cities and select **Label/Unlable**.
7. Select **Rows > Row Selection> Select all Rows.**
8. Select **Rows > Lable/Unlabel.**
9. Select the Multidimensional Scaling Plot.
10. Click on the**Flip Vertical** button.
11. Click on the**Flip Horizontal** button.

    The Flip Vertical and Flip Horizontal buttons enable you to change the orientation of the MDS Plot. The MDS results are invariant to orientation. When the results have a known orientation, such as physical locations, then you might want to rotate or flip your plot.

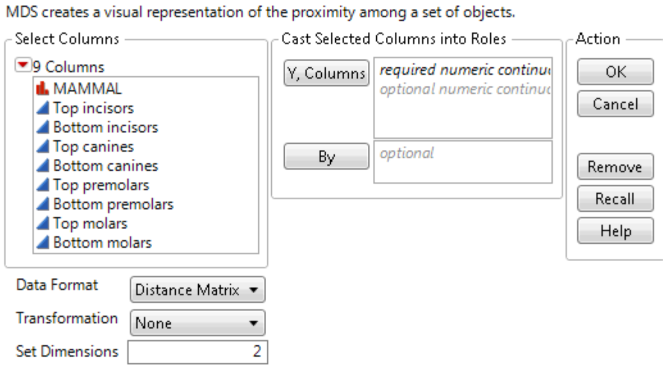
Figure 5.3 Multidimensional Scaling Plot



Launch the Multidimensional Scaling Platform

Launch the Multidimensional Scaling Platform by selecting **Analyze > Consumer Research > Multidimensional Scaling.**

Figure 5.4 Multidimensional Scaling Launch Window



Y, Columns The columns to be analyzed. These must have a Numeric data type.

By A column or columns whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed using the other variables that you have specified. The results are presented in separate reports. If more than one By variable is assigned, a separate report is produced for each possible combination of the levels of the By variables.

Note: When using a distance matrix, the By variable requires a full matrix for each level of the By variable.

Data Format MDS supports two data formats:

Distance Matrix A full symmetric or lower triangular matrix where the number of rows equals the number of columns).The diagonal entries can either be zeros or missing.

Attribute List A set of columns that contain measures of a quality or characteristic of an object. The objects are typically named in a column. The object column is not used in the analysis but rather is used as a label for the data points on the MDS plot.

Transformation Supported transformations are Ratio, Interval, and Ordinal.

None No transformation used.

Ratio Data has an ordering from smallest to largest, the differences between values have meaning, and the scale has a true zero. Used to scale the MDS plot.

Interval Data has an ordering from smallest to largest and the differences between values have meaning. Used to scale and shift the MDS plot.

Ordinal Data has an ordering from smallest to largest. Used for ordinal data.

Set Dimensions The number of dimensions for the visual representation of the proximities among your objects. Typically, 2 or 3 dimensions are used. With greater than 3 dimensions the visualization becomes complex.

Note: The dimension selected can be between 1 to n - 1 where n = the number of objects, otherwise the dimension is set to 2.

The Multidimensional Scaling Report

The initial Multidimensional Scaling report shows these reports: Multidimensional Scaling Plot, the Shepard Diagram, and the Fit Details. If you specify three or more dimensions for the fit in the launch window, then the Multidimensional Scaling Plot provides controls for selecting the dimensions that you view.

Objects that are close together on the MDS plot share similar attributes. Adding labels and colors to the plot can help in the identification of similar groups. The Shepard diagram and summary of fit statistics provide measures of how well the MDS plot represents the proximities of the objects.

Multidimensional Scaling Plot

The MDS plot displays the multidimensional scaling in two dimensions. Below the plot are two buttons to flip the axis either in the vertical or horizontal direction. The MDS solution can be reflected, rotated, or translated without changing the inter-point proximities. The rotating or reflection of the axes is most common when working with geographical objects that have a known map orientation.

If more than two dimensions were used in the analysis, then you can toggle the dimensions shown in the plot using the **Select Dimension** controls below the plot. The first control defines the horizontal axis of the plot, and the second control defines the vertical axis of the plot.

Shepard Diagram

The Shepard plot is a plot of the actual or transformed proximities versus the predicted proximities. The plot indicates how well the Multidimensional Scaling Plot reflects the actual proximities. The Shepard is analogous to an Actual by Predicted plot. Ideally the points fall on the Y = X line, which is shown in red.

Fit Details

The Fit Details provides statistics that summarize how well the MDS proximities match the actual proximities as well as details about transformations when used.

Stress The value of the stress function (Stress1) that was minimized in the fitting procedure. Stress can be between 0 and 1 with lower values indicating a better fit.

RSquare The R^2 value for linear fit of the actual or transformed proximities versus the predicted proximities.

Slope If a ratio or interval transformation was used, the slope for the transformation is provided. It is the slope of the linear regression of the actual against transformed proximities.

Intercept If an interval transformation was used, the intercept for the transformation is provided. It is the intercept of the linear regression of the actual against transformed proximities.

Multidimensional Scaling Platform Options

The Multidimensional Scaling red triangle menu options give you the ability to customize reports according to your needs. The options available are determined by the type of data and the number of dimensions that you use for your analysis.

MDS Plot Shows or hides the MDS Plot. See “Waern Links”.

Diagnostics Provides diagnostics for the MDS.

Shepard Diagram Shows a plot of actual proximity (or transformed proximity if a transformation is used) versus the predicted proximity. This report appears by default. See “Shepard Diagram”.

Waern Links Displays the Waern links on the MDS plot. Controls for the portion (smallest or largest) are available when this option is selected. See “Waern Links”.

Show Coordinates Provides a report of the solution coordinates. These are the coordinates of the points on the Multidimensional Scaling Plot. The report shows the coordinates of up to three dimensions. Right-click in the report and select **columns** to add additional dimensions to the report. The maximum number of dimensions is the number of dimensions set in the launch dialog.

Show Proximity Provides a report of the proximities. The original and derived proximities (distances) are provided between each pair of objects. The pairs are identified in the From and To object columns. If a transformation was used, the transformed proximities are also included in the table.

Save Proximity (Available only if Attribute List is the data format.) Saves the distance matrix to the data table.

3D Plot (Available only if three or more dimensions are specified for Set Dimensions in the launch window.) Shows a 3-D plot of the first three dimensions.

Save Coordinates Saves the solution coordinates to the data table in separate columns.

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.

Waern Links

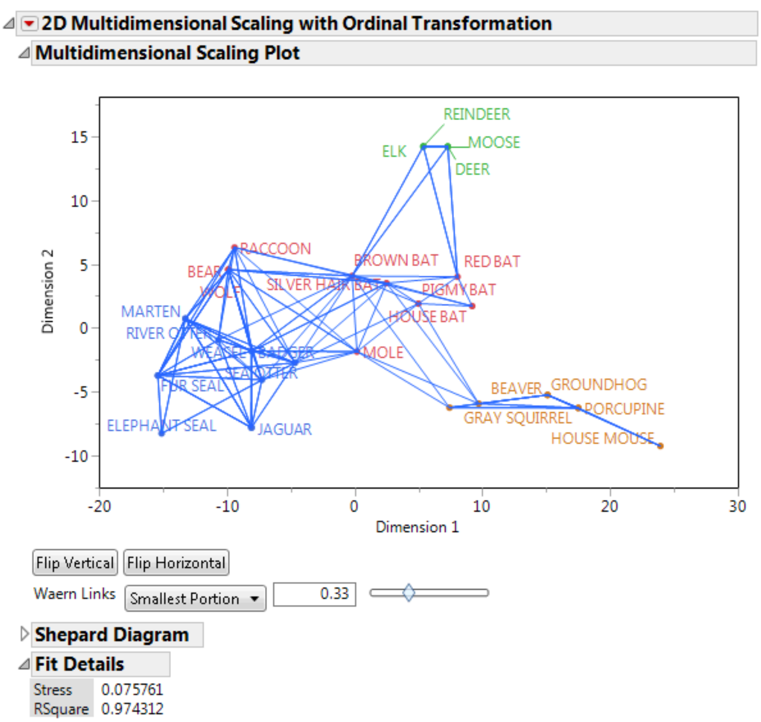
Waern links provide a visual check of the MDS results by comparing actual proximities to predicted proximities. The links join points on the Multidimensional Scaling Plot based on their actual proximities. The objects with the smallest (largest) proximities are connected. A typical scenario to consider is the smallest 33% of the proximities between objects. If the MDS Plot is a good representation of the proximities, then the links for the smallest actual proximities should connect the closest objects in the plot. If a link for a small proximity stretches across the plot connecting distant objects, then the MDS fit would be questioned.

Waern Link Controls

There is a list from which you can choose to show the Smallest Portion or the Largest Portion of links on the plot. The portion of links shown is controlled by entering a value in the box or by using the slider. Figure 5.5 shows Waern links for the Teeth.jmp data table for the 33% smallest portion.

For more information about Waern links see Waern (1972).

Figure 5.5 MDS Plot with Waern Links

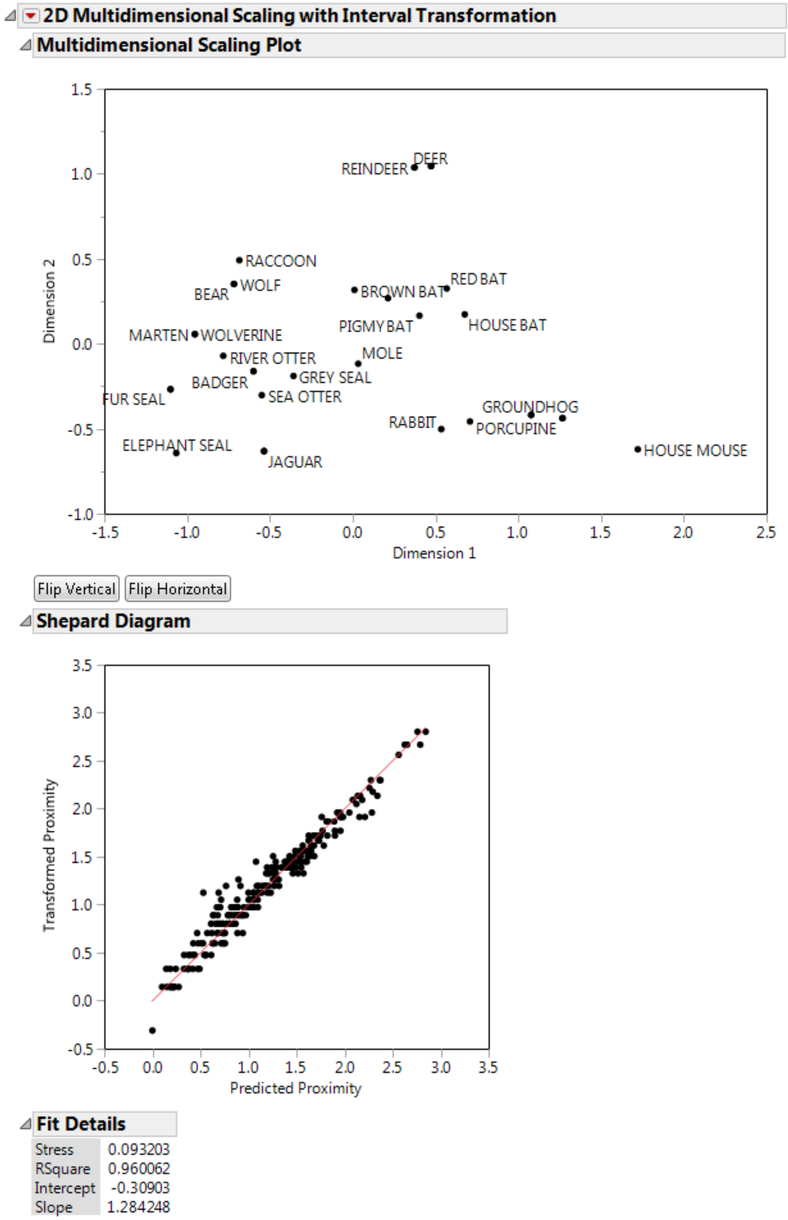


Additional Example of the Multidimensional Scaling Platform

This example uses the Teeth.jmp sample data table, which is an attribute list of the counts of eight teeth types in 32 mammals. You can use MDS to explore the similarities of mammals based on their teeth. An interval transformation is used to illustrate the output from that transformation. The data do have an ordering that has a meaning (2 teeth are twice as many as 4 teeth). One might explore other transformations such as the ordinal transformation.

1. Select **Help > Sample Data Library** and open Teeth.jmp.
2. Right-click on the column MAMMAL and select **Label/Unlabel**.
3. Select **Rows > Row Selection> Select all Rows.**
4. Select **Analyze > Consumer Research > Multidimensional Scaling**.
5. Select Top incisors through Bottom molars and click **Y, Columns**.
6. Select **Data Format > Attribute List**.
7. Select **Transformation > Interval**.
8. Click **OK**.

Figure 5.6 Multidimensional Scaling Report



The Shepard Diagram and the Fit Details indicate that the MDS Plot is a good representation of similarities of animals due to similarities in their teeth. The Stress statistics of 0.075 is low and the R^2 fit of the transformed versus predicted proximities is high at 0.97. In addition, the Fit Details provides the intercept and slope for the transformation of the actual proximities.

Statistical Details for the Multidimensional Scaling Platform

JMP uses a quasi Newton optimization method to minimize the Stress function to determine the MDS coordinates. This minimization leads to a set of coordinates in a predetermined number of dimensions that minimize the derived proximity measures for each pairwise set of the dimensions. When the data is ordinal then monotonic regression is used. Otherwise, standard least squares regression is used.

Stress

The following notation is used to define Stress;

- I - The number of dimensions specified for the fit
- i, j - indexes for the number of objects
- d_{ij} - the distance between objects i and j
- δ_{ij} - the derived distance between objects i and j
- $f(\delta_{rs})$ - transformation function for the distance

The Stress function is given by

$$\text{Stress} = \left[\frac{\sum_{i<j}[f(\delta_{ij}) - d_{ij}]^2}{\sum_{i<j} d_{ij}^2}\right]^{\frac{1}{2}}$$

This measure of stress is also known as Kruskal's Stress, Type I, or simply Stress1.

Transformations

The section uses the notation described in "Stress" on page 97. Transformations are used to scale the actual proximities. Transformations would be considered to improve the MDS representation of the actual proximities by taking into account specific structures in the data. The parameters in the transformation functions become additional parameters in the minimization algorithm.

Ratio Transformation

For ratio data:

$f(\delta_{rs}) = b\delta_{rs}$

Interval Transformation

For interval data:

$$f(\delta_{rs}) = a + b\delta_{rs}$$

Ordinal Transformation

For ordinal data the data is not transformed, rather the algorithm uses monotone regression rather than least squares regression.

Attributes List Format

When the data is in the attributes list format, it is converted to a distance matrix and then MDS is applied. The distance matrix is determined by the correlation structure of the data.

For an advanced example of MDS see the San Francisco Crime Distances.jmp sample data table and the source script for that table.

Chapter 6

Factor Analysis

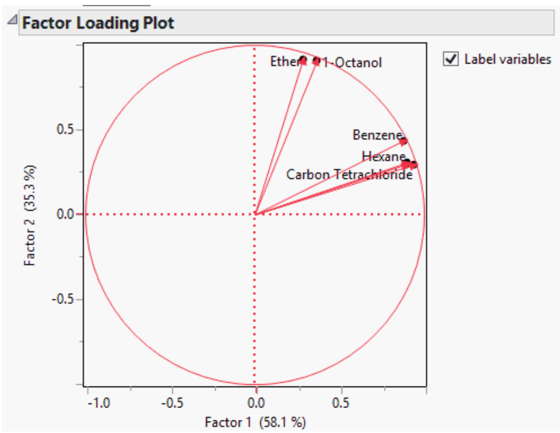
Identify Factors within Variables

Factor analysis seeks to describe a collection of observed variables in terms of a smaller collection of (unobservable) latent variables, or factors. Factor analysis is also known as common factor analysis and exploratory factor analysis. These factors are defined as linear combinations of the observed variables. They are constructed to explain variation that is *common* to the observed variables. A primary goal of factor analysis is to achieve a meaningful interpretation of the observed variables through the factors. Another goal is to reduce the number of variables.

Factor analysis is used in many areas, and is of particular value in psychology, sociology, and education. In these areas, factor analysis is used to understand how manifest behavior can be interpreted in terms of underlying patterns and structures. For example, measures of participation in outdoor activities, hobbies, exercise, and travel, might all relate to a factor that can be described as “active versus inactive personality type”. Factor analysis attempts to explain correlations among the observed variables in terms of the factor. In particular, it enables you to determine how much of the variance in each observable variable is accounted for by the factors that you have identified. It also tells you how much of the variance in all the variables is accounted for by each factor.

Use factor analysis when you need to explore or interpret underlying patterns and structure in your data. Also consider using it to summarize the information in your variables using a smaller number of latent variables.

Figure 6.1 Rotated Factor Loading



Factor Analysis Platform Overview

Factor analysis models a set of observable variables in terms of a smaller number of unobservable factors. These factors account for the correlation or covariance between the observed variables. Once the factors are extracted, you perform factor rotation in order to obtain a meaningful interpretation of the factors.

Consider a situation where you have ten observed variables, X_1, X_2, …, X_{10}. Suppose that you want to model these ten variables in terms of two latent factors, F_1 and F_2. For convenience, it is assumed that the factors are uncorrelated and that each has mean zero and variance one. The model that you want to derive is of the form:

$$X_i = \beta_{i0} + \beta_{i1}F_1 + \beta_{i2}F_2 + \varepsilon_i$$

It follows that $Var(X_i) = \beta_{i1}^2 + \beta_{i2}^2 + Var(\varepsilon_i)$. The portion of the variance of X_i that is attributable to the factors, the common variance or *communality*, is $\beta_{i1}^2 + \beta_{i2}^2$. The remaining variance, $Var(\varepsilon_i)$, is the specific variance, and is considered to be unique to X_i.

The Factor Analysis platform provides a Scree Plot for the eigenvalues of the correlation or covariance matrix. You can use this as a guide in determining the number of factors to extract. Alternatively, you can accept the platform's suggestion of setting the number of factors equal to the number of eigenvalues that exceed one.

The platform provides two factoring methods for estimating the parameters of this model: Principal Components and Maximum Likelihood.

JMP provides two options for estimating the proportion of variance contributed by common factors for each variable. These Prior Communality options impose assumptions on the diagonal of the correlation (or covariance) matrix. The Principal Components option treats the correlation matrix, which has ones on its diagonal (or the covariance matrix with variances on its diagonal), as the structure to be analyzed. The Common Factor Analysis option sets the diagonal entries to values that reflect the proportion of the variation that is shared with other variables.

To support interpretability of the extracted factors, you rotate the factor structure. The Factor Analysis platform provides a variety of rotation methods that encompass both orthogonal and oblique rotations.

In contrast with factor analysis, which looks at common variance, principal component analysis accounts for the total variance of the observed variables. See the Principal Components chapter in the *Multivariate Methods* book.

Example of the Factor Analysis Platform

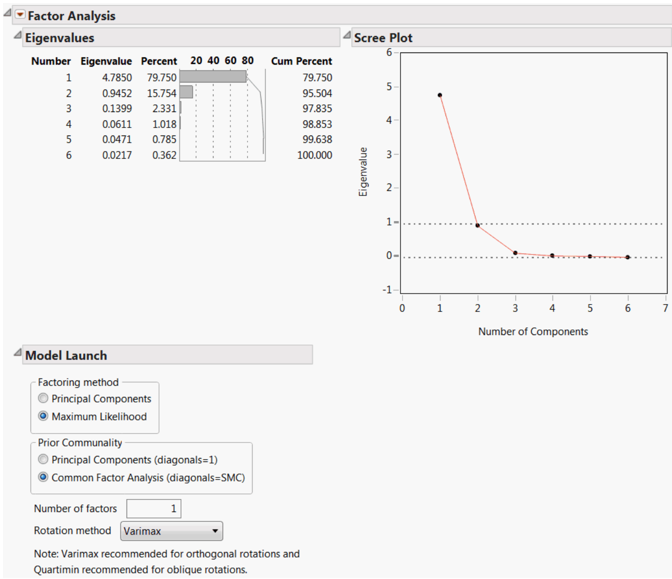
To view an example Factor Analysis report for a data table for two factors:

1. Select **Help > Sample Data Library** and open Solubility.jmp.
2. Select **Analyze > Consumer Research > Factor Analysis.**

   The Factor Analysis launch window appears.
3. Select all of the continuous columns and click **Y, Columns**.
4. Keep the default **Estimation Method** and **Variance Scaling**.
5. Click **OK**.

   The initial Factor Analysis report appears.

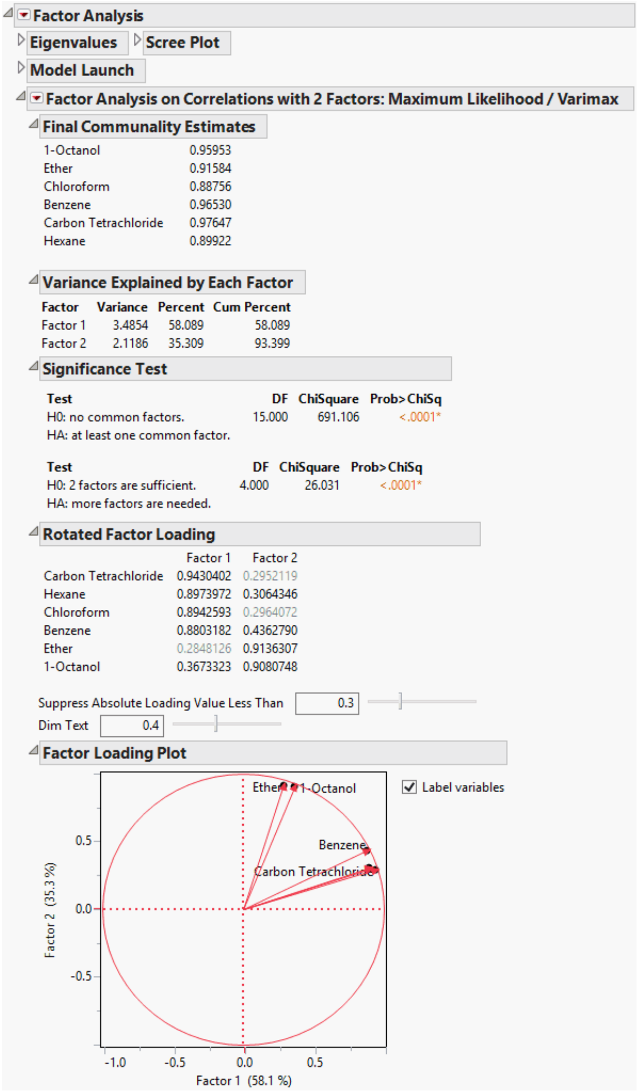
Figure 6.2 Initial Factor Analysis Report



6. For the Model Launch, select the following options:
   – Factoring Method as **Maximum Likelihood**
   – Prior Communality as **Common Factor Analysis**
   – Number of factors = 2
   – Rotation Method as **Varimax**
7. After all selections are made, click **Go**.

   The Factor Analysis report appears.

Figure 6.3 Example Factor Analysis Report

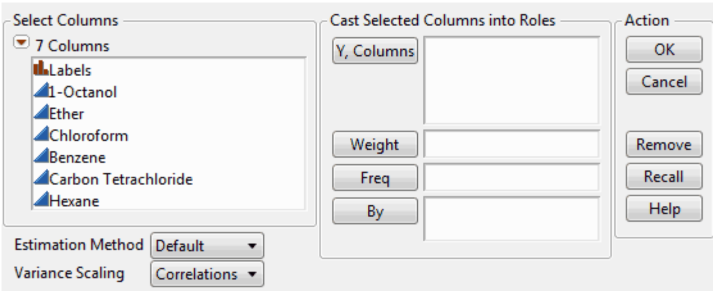


The report lists the communality estimates, variance, significance tests, rotated factor loadings, and a factor loading plot. Note that in the Factor Loading Plot, Factor 1 relates to the Carbon Tetrachloride-Chloroform-Benzene-Hexane cluster of variables, and Factor 2 relates to the Ether–1-Octanol cluster of variables. See "Factor Analysis Model Fit Options" on page 109 for details of the information shown in the report.

Launch the Factor Analysis Platform

Launch the Factor Analysis platform by selecting **Analyze > Consumer Research > Factor Analysis**.

Figure 6.4 Factor Analysis Launch Window



Y, Columns Lists the continuous columns to be analyzed.

Weight Enables you to weight the analysis to account for pre-summarized data.

Freq Identifies a column whose numeric values assign a frequency to each row in the analysis.

By Creates a Factor Analysis report for each value specified by the By column so that you can perform separate analyses for each group.

Estimation Method Lists different methods for fitting the model. For details about the methods, see the Multivariate chapter in the *Multivariate Methods* book.

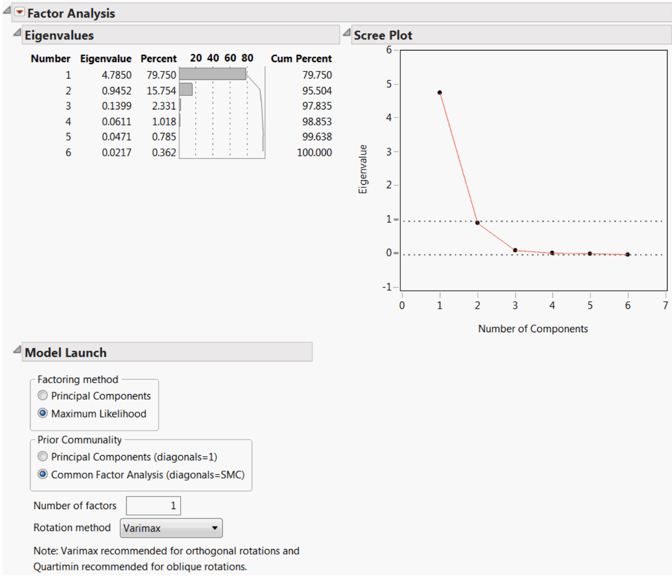
Variance Scaling Lists the scaling methods for performing the factor analysis based on **Correlations** (the same as Principal Components), **Covariances**, or **Unscaled**.

The Factor Analysis Report

The initial Factor Analysis report shows Eigenvalues and the Scree Plot. The Eigenvalues are obtained from a principal components analysis. The Scree Plot graphs these eigenvalues. The number of factors that JMP suggests in the Model Launch equals the number of eigenvalues that exceed 1.0.

Alternatively, you can use the scree plot to guide your initial choice for number of factors. The number of eigenvalues that appear before the scree plot levels out can provide an upper bound on the number of factors.

Figure 6.5 Factor Analysis Report

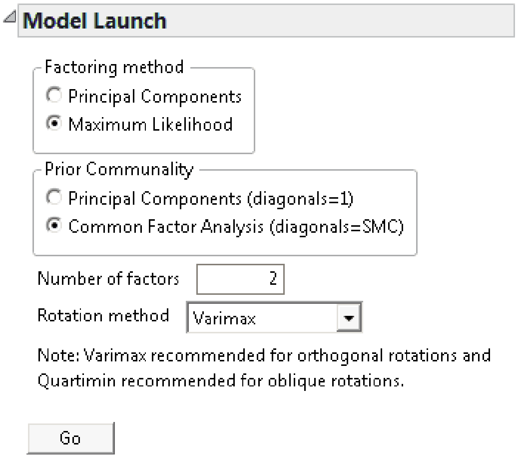


In the example shown in Figure 6.5, the Scree Plot begins to level out after the second eigenvalue. The Eigenvalues table indicates that the first eigenvalue accounts for 79.75% of the variation and the second eigenvalue accounts for 15.75%. Therefore, the first two eigenvalues account for 95.50% of the total variation. The third eigenvalue only explains 2.33% of the variation, and the contributions from the remaining eigenvalues are negligible. Although the **Number of factors** box is initially set to 1, this analysis suggests that it is appropriate to extract 2 factors.

Model Launch

To configure the Factor Analysis model, use the Model Launch section at the bottom of the Factor Analysis Report (Figure 6.6).

Figure 6.6 Model Launch



The Model Launch section enables you to configure the following options:

1. **Factoring method -** the method for extracting factors.
   - The **Principal Components** method is a computationally efficient method, but it does not allow for hypothesis testing.
   - The **Maximum Likelihood** method has desirable properties and enables you to test hypotheses about the number of common factors.

   Note: The **Maximum Likelihood** method requires a positive definite correlation matrix. If your correlation matrix is not positive definite, select the **Principal Components** method.

2. **Prior Communality -** the method for estimating the proportion of variance contributed by common factors for each variable.
   - **Principal Components (diagonals = 1)** sets all communalities equal to 1, indicating that 100% of each variable's variance is shared with the other variables. Using this option with Factoring Method set to **Principal Components** results in principal component analysis.
   - **Common Factor Analysis (diagonals = SMC)** sets the communalities equal to squared multiple correlation (SMC) coefficients. For a given variable, the SMC is the RSquare for a regression of that variable on all other variables.
3. The **Number of factors** (or principal components) determined by eigenvalues greater than or equal to 1.0 or from the scree plot where the graph begins to level out.

   Note: Alternatively, the *Kaiser criterion* retains those factors with eigenvalues greater than 1.0. In our example, only factor 1 would be retained for analysis.

4. The **Rotation method** to align the factor directions with the original variables for ease of interpretation. The default value is **Varimax**. See "Rotation Methods" on page 106 for a description of the available selections.
5. Click **Go** to generate the Factor Analysis report.

   Depending on the selected Variance Scaling, the appropriate factor analysis results appear. See "Factor Analysis Model Fit Options" on page 109 for details about the contents of the report. The Factor Analysis on Correlations and Factor Analysis on Unscaled reports show the same information.

Rotation Methods

Rotations align the directions of the factors with the original variables so that the factors are more interpretable. You hope for clusters of variables that are highly correlated to define the rotated factors.

After the initial extraction, the factors are uncorrelated with each other. If the factors are rotated by an orthogonal transformation, the rotated factors are also uncorrelated. If the factors are rotated by an oblique transformation, the rotated factors become correlated. Oblique rotations often produce more useful patterns than do orthogonal rotations. However, a consequence of correlated factors is that there is no single unambiguous measure of the importance of a factor in explaining a variable.

Orthogonal Rotation Methods

Table 6.1 lists the available orthogonal (that is, uncorrelated) rotation methods.

Table 6.1 Orthogonal Rotation Methods

Method	SAS PROC FACTOR Equivalent
Varimax	ROTATE=ORTHOMAX with GAMMA = 1 **Note:** This is the default selection.
Biquartimax	ROTATE=ORTHOMAX with GAMMA = 0.5
Equamax	ROTATE=ORTHOMAX with GAMMA = number of factors/2
Factorparsimax	ROTATE=ORTHOMAX with GAMMA = number of variables

Table 6.1 Orthogonal Rotation Methods *(Continued)*

Method	SAS PROC FACTOR Equivalent
Orthomax	ROTATE=ORTHOMAX Or ROTATE=ORTHOMAX(*p*), where *p* is the orthomax weight or the GAMMA = value. **Note:** The default *p* value is 1 unless specified otherwise in the GAMMA = option. For additional information about orthomax weight, see the SAS documentation, "Simplicity Functions for Rotations."
Parsimax	ROTATE=ORTHOMAX with GAMMA = $\left[\frac{(nvar(nfact-1))}{(nvar+nfact-2)}\right]$ where *nvar* is the number of variables, and *nfact* is the number of factors.
Quartimax	ROTATE=ORTHOMAX with GAMMA=0

Oblique Rotation Methods

Table 6.2 lists the available oblique (that is, correlated) rotation methods.

Table 6.2 Oblique Rotation Methods

Method	SAS PROC FACTOR Equivalent
Biquartimin	ROTATE=OBLIMIN(.5) Or ROTATE=OBLIMIN with TAU=.5
Covarimin	ROTATE=OBLIMIN(1) Or ROTATE=OBLIMIN with TAU=1
Obbiquartimax	ROTATE=OBBIQUARTIMAX
Obequamax	ROTATE=OBEQUAMAX
Obfactorparsimax	ROTATE=OBFACTORPARSIMAX

Table 6.2 Oblique Rotation Methods *(Continued)*

Method	SAS PROC FACTOR Equivalent
Oblimin	ROTATE=OBLIMIN, where the default *p* value is zero, unless specified otherwise in the TAU= option. ROTATE=OBLIMIN(*p*) specifies *p* as the oblimin weight or the TAU= value. **Note:** For additional information about oblimin weight, see the SAS documentation, "Simplicity Functions for Rotations."
Obparsimax	ROTATE=OBPARSIMAX
Obquartimax	ROTATE=OBQUARTIMAX
Obvarimax	ROTATE=OBVARIMAX
Quartimin	ROTATE=OBLIMIN(0) or ROTATE=OBLIMIN with TAU=0
Promax	ROTATE=PROMAX

Factor Analysis Platform Options

The following options are accessed by clicking the Factor Analysis red triangle menu in the report window:

Eigenvalues A table that indicates the total number of factors extracted based on the eigenvalues (that is, the amount of variance contributed by each factor). The table includes the percent of the total variance contributed by that factor, a bar chart illustrating the percent contribution, and the cumulative percent contributed by each successive factor. The number of eigenvalues greater than or equal to 1.0 can be taken as the number of sufficient factors for analysis.

Scree Plot A plot of the eigenvalues versus the number of components (or factors). The plot can be used to determine the number of factors that contribute to the maximum amount of variance. The point at which the plotted line levels out can be taken as the number of sufficient factors for analysis. See Figure 6.2 on page 101 for an example of scree plots.

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.

Factor Analysis Model Fit Options

After submitting the Model Launch, the model results appear. The following options are available from the Factor Analysis report's red triangle menu.

Prior Communality An initial estimate of the communality for each variable. For a given variable, this estimate is the squared multiple correlation coefficient (SMC), or RSquare, for a regression of that variable on all other variables.

Note: The Prior Communality Estimates table only appears if the **Common Factor Analysis (diagonals = SMC)** option is selected.

Figure 6.7 Prior Communality Estimates

Prior Communality Estimates:SMC	
1-Octanol	0.89679
Ether	0.88297
Chloroform	0.90228
Benzene	0.96385
Carbon Tetrachloride	0.96040
Hexane	0.90635

Eigenvalues Shows the eigenvalues of the reduced correlation matrix and the percent of the common variance for which they account. The reduced correlation matrix is the correlation matrix with its diagonal entries replaced by the communality estimates. The eigenvalues indicate the common variance explained by the factors. The Cum Percent can exceed 100% because the reduced correlation matrix is not necessarily positive definite and can have negative eigenvalues.

Note that the table indicates the number of factors retained for analysis.

The Eigenvalues option is available only when the Prior Communality option is set to Common Factor Analysis (diagonals = SMC). The communality estimates are the SMC (square multiple correlation) values.

Figure 6.8 indicates that the first two factors account for 100% of the common variance. This pattern suggests that you might not need more than two factors to model your data.

Figure 6.8 Eigenvalues of the Reduced Correlation Matrix

Eigenvalues of the Reduced Correlation Matrix

Number	Eigenvalue	Percent	Cum Percent
1	104.1143	86.233	86.233
2	16.6212	13.767	100.000
3	0.7181	0.595	100.595
4	0.0236	0.020	100.614
5	-0.2677	-0.222	100.393
6	-0.4740	-0.393	100.000

2 factors will be retained by the number of factor criterion.

Unrotated Factor Loading Shows the factor loading matrix before rotation. Factor loadings measure the influence of a common factor on a variable. Because the unrotated factors are orthogonal, the factor loading matrix is the matrix of correlations between the variables and the factors. The closer the absolute value of a loading is to 1, the stronger the effect of the factor on the variable.

Use the slider and value to **Suppress Absolute Loading Values Less Than** the specified value in the table. Suppressed values appear dimmed according to the setting specified by **Dim Text**.

Use the **Dim Text** slider and value to control the table's font transparency gradient for factor values less in absolute value than the specified **Suppress Absolute Loading Values Less Than** value.

Note: The **Suppress Absolute Loading Values Less Than** value and **Dim Text** value are the same values used in the Rotated Factor Loading table. Changes to one loading table's settings changes the settings in the other loading table.

Figure 6.9 Unrotated Factor Loading

Unrotated Factor Loading

	Factor 1	Factor 2
Benzene	0.977736	-0.096606
Carbon Tetrachloride	0.956158	-0.249473
Hexane	0.923401	-0.215765
Chloroform	0.915425	-0.222605
1-Octanol	0.792777	0.575359
Ether	0.725745	0.623805

Suppress Absolute Loading Value Less Than 0.3
Dim Text 0.4

Note: The Unrotated Factor Loading matrix is re-ordered so that variables associated with the same factor appear next to each other.

Rotation Matrix Shows the calculations used for rotating the factor loading plot and the factor loading matrix.

Figure 6.10 Rotation Matrix

Rotation Matrix

0.84800	0.53000
-0.53000	0.84800

Target Matrix Shows the matrix to which the varimax factor pattern is rotated. This option is available only for the Promax rotation.

Figure 6.11 Target Matrix

Target Matrix

	Factor 1	Factor 2
1-Octanol	0.060672	0.915571
Ether	0.030328	1.000000
Chloroform	0.984004	0.035792
Benzene	0.827612	0.100628
Carbon Tetrachloride	1.000000	0.030643
Hexane	0.975109	0.038782

Factor Structure Shows the matrix of correlations between variables and common factors. This option is available only for oblique rotations.

Figure 6.12 Factor Structure

Factor Structure

	Factor 1	Factor 2
1-Octanol	0.6261049	0.9767863
Ether	0.5491838	0.9569717
Chloroform	0.9420450	0.5544378
Benzene	0.9713036	0.6834367
Carbon Tetrachloride	0.9881517	0.5681409
Hexane	0.9480839	0.5649445

Final Communality Estimates Estimates of the communalities after the factor model has been fit. When the factors are orthogonal, the final communality estimate for a variable equals the sum of the squared loadings for that variable.

Figure 6.13 Final Communality Estimates

Final Communality Estimates

1-Octanol	0.95953
Ether	0.91584
Chloroform	0.88756
Benzene	0.96530
Carbon Tetrachloride	0.97647
Hexane	0.89922

Standard Score Coefficients Lists the multipliers used to convert factor values when saving rotated components as factors to the source data table.

Figure 6.14 Standard Score Coefficients

Standard Score Coefficients

	Factor 1	Factor 2
1-Octanol	-0.269592	0.782996
Ether	-0.153368	0.400176
Chloroform	0.125223	-0.054222
Benzene	0.311045	0.008094
Carbon Tetrachloride	0.646847	-0.305395
Hexane	0.138317	-0.056834

Variance Explained by Each Factor Gives the variance, percent, and cumulative percent, of common variance explained by each rotated factor.

Figure 6.15 Variance Explained by Each Factor

Variance Explained by Each Factor

Factor	Variance	Percent	Cum Percent
Factor 1	3.4854	58.089	58.089
Factor 2	2.1186	35.309	93.399

Significance Test If you select **Maximum Likelihood** as the factoring method, the results of two Chi-square tests are provided.

The first test is for H_0: No common factors. This null hypothesis indicates that none of the common factors are sufficient to explain the intercorrelations among the variables. This test is Bartlett's Test for Sphericity, whose null hypothesis is that the correlation matrix of the factors is an identity matrix (Bartlett, 1954).

The second test is for H_0: N factors are sufficient, where N is the specified number of factors. Rejection of this null hypothesis indicates that more factors might be required to explain the intercorrelations among the variables (Bartlett, 1954).

The tests in Figure 6.16 indicate that the common factors already included in the model explain some of the intercorrelations, but that more factors are needed.

Note: The Significance Test table only appears if the **Maximum Likelihood** factoring method option is selected.

Figure 6.16 Significance Test

Significance Test

Test	DF	ChiSquare	Prob>ChiSq
H0: no common factors. HA: at least one common factor.	15.000	691.106	<.0001*

Test	DF	ChiSquare	Prob>ChiSq
H0: 2 factors are sufficient. HA: more factors are needed.	4.000	26.031	<.0001*

Rotated Factor Loading Shows the factor loading matrix after rotation. If the rotation is orthogonal, these values are the correlations between the variables and the rotated factors.

Use the slider and value to **Suppress Absolute Loading Values Less Than** the specified value in the table. Suppressed values appear dimmed according to the setting specified by **Dim Text**.

Use the **Dim Text** slider and value to control the table's font transparency gradient for factor values less in absolute value than the specified **Suppress Absolute Loading Values Less Than** value.

Note: The **Suppress Absolute Loading Values Less Than** value and **Dim Text** value are the same values used in the Unrotated Factor Loading table. Changes to one loading table's settings changes the settings in the other loading table.

Figure 6.17 Rotated Factor Loading

Rotated Factor Loading

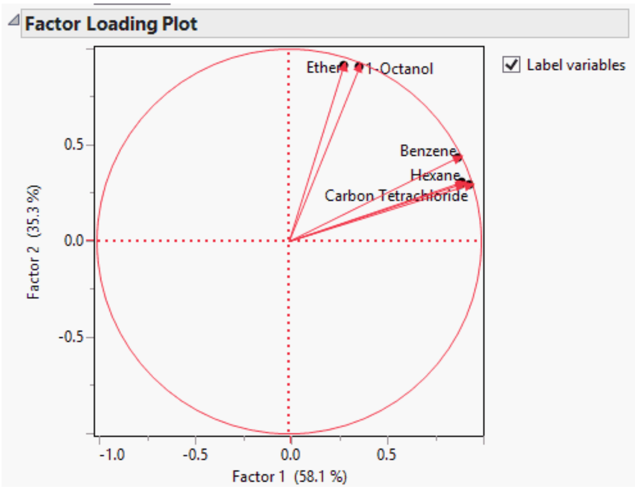
	Factor 1	Factor 2
Carbon Tetrachloride	0.9430402	0.2952119
Hexane	0.8973972	0.3064346
Chloroform	0.8942593	0.2964072
Benzene	0.8803182	0.4362790
Ether	0.2848126	0.9136307
1-Octanol	0.3673323	0.9080748

Suppress Absolute Loading Value Less Than 0.3
Dim Text 0.4

Note: The Rotated Factor Loading matrix is re-ordered so that variables associated with the same factor appear next to each other.

Factor Loading Plot The plot of the rotated loading factors.

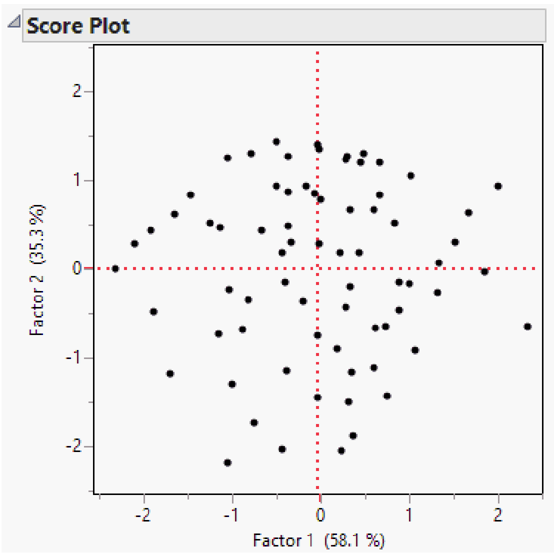
Figure 6.18 Factor Loading Plot



Note that in the Factor Loading Plot, Factor 1 relates to the Carbon Tetrachloride-Chloroform-Benzene-Hexane cluster of variables, and Factor 2 relates to the Ether–1-Octanol cluster of variables. See the matrix of "Rotated Factor Loading" on page 113 for details.

Score Plot The Score Plot graphs each factor's calculated values in relation to the other adjusting each value for the mean and standard deviation.

Figure 6.19 Score Plot



Score Plot with Imputation Imputes any missing values and creates a score plot. This option is available only if there are missing values.

Display Options Enables you to show or hide arrows on all plots that can display arrows.

Save Rotated Components Saves the rotated components to the data table, with a formula for computing the components. The formula cannot evaluate rows with missing values.

Save Rotated Components with Imputation Imputes missing values, and saves the rotated components to the data table. The column contains a formula for doing the imputation, and computing the rotated components. This option appears after the Factor Analysis option is used, and if there are missing values.

Remove Fit Removes the fit model results from the Factor Analysis Fit Model report. This option enables you to change the Model Launch configuration for a new report.

Chapter 7

Choice Models

Fit Models for Choice Experiments

Use the Choice platform to analyze the results of choice experiments conducted in the course of market research, in order to discover which product or service attributes your potential customers prefer. You can use this information to design products or services that have the attributes that your customers most desire.

The Choice platform enables you to do the following:

- Use information about subject traits as well as product attributes.
- Analyze choice experiments where respondents were allowed to select “none of these”.
- Integrate data from one, two, or three sources.
- Use the integrated profiler to understand, visualize, and optimize the response (utility) surface.
- Obtain subject-level scores for segmenting or clustering your data.
- JMP PRO Estimate subject-specific coefficients using a Bayesian approach.
- Use bias-corrected maximum likelihood estimators (Firth, 1993).

Figure 7.1 Choice Platform Utility Profiler

Choice Modeling Platform Overview

Choice modeling, pioneered by McFadden (1974), is a powerful analytic method used to estimate the probability of individuals making a particular choice from presented alternatives. Choice modeling is also called conjoint choice modeling, discrete choice analysis, and conditional logistic regression.

A choice experiment studies customer preferences for a set of product or process (in the case of a service) attributes. Respondents are presented sets of product attributes, called *profiles*. Each respondent is shown a small set of profiles, called a *choice set*, and asked to select the preference that he or she most prefers. Each respondent is usually presented with several choice sets. Use the Choice platform to analyze the results of a choice experiment.

Note: You can design your choice experiment using the Choice Design platform. See the Discrete Choice Design chapter in *Design of Experiments Guide*.

Because customers vary in how they value attributes, many market researchers view market segmentation as an important step in analyzing choice experiments. Otherwise, you risk designing a product or process that pleases the “average” customer, who does not actually exist, and ignoring the preferences of market segments that *do* exist.

For background on choice modeling, see Louviere et al. (2015), Train (2009), and Rossi et al. (2006).

The Choice Platform

The Choice Modeling platform uses a form of conditional logistic regression to estimate the probability that a configuration is preferred. Unlike simple logistic regression, choice modeling uses a linear model to model choices based on response attributes and not solely upon subject characteristics. In choice modeling, a respondent might choose between two cars that are described by a combination of ten attributes, such as price, passenger load, number of cup holders, color, GPS device, gas mileage, anti-theft system, removable-seats, number of safety features, and insurance cost.

The Choice platform allows respondents to *not* make a choice from among a set of profiles. The *no choice* option is treated as a product with a single attribute (“Select none of these”) that respondents are allowed to select. The parameter estimate for the No Choice attribute can then be interpreted in many ways, depending on the assumptions of the model. The Choice platform also enables you to obtain subject-level information, which can be useful in segmenting preference patterns.

You can obtain bias-corrected maximum likelihood estimators as described by Firth (1993). This method has been shown to produce better estimates and tests than MLEs without bias correction. In addition, bias-corrected MLEs improve separation problems that tend to occur in logistic-type models. Refer to Heinze and Schemper (2002) for a discussion of the separation problem in logistic regression.

Note: The Choice platform is not appropriate to use for fitting models that involve ranking, scoring, or nested hierarchical choices. You can use PROC MDC in SAS/ETS for these analyses.

Choice Designs in Developing Products and Services

Although customer satisfaction surveys can disclose what is wrong with a product or service, they fail to identify consumer preferences with regard to specific product attributes. When engineers design a product, they routinely make hundreds or thousands of small design decisions. If customer testing is feasible and test subjects (prospective customers) are available, you can use choice experiments to guide some design decisions.

Decreases in survey deployment, modeling, and prototyping costs facilitate the customer evaluation of many attributes and alternatives as a product is designed. Choice modeling can be used in Six Sigma programs to improve consumer products, or, more generally, to make the products that people want. Choice experiments obtain data on customer preferences, and choice modeling analysis reveals such preferences.

Segmentation

Market researchers sometimes want to analyze the preference structure for each subject separately in order to see whether there are groups of subjects that behave differently. However, there are usually not enough data to do this with ordinary estimates. If there are sufficient data, you can specify the subject identifier as a "By groups" in the Response Data or you could introduce a subject identifier as a subject-side model term. This approach, however, is costly if the number of subjects is large.

If there are not sufficient data to specify "By groups," you can segment in JMP by clustering subjects using the Save Gradients by Subject option. The option creates a new data table containing the average Hessian-scaled gradient on each parameter for each subject. For an example, see "Example of Segmentation" on page 154. For details about the gradient values, see "Gradients" on page 172.

JMP PRO In JMP Pro, you can request that the Choice platform use a Hierarchical Bayes approach in order to facilitate market segmentation. Bayesian modeling provides subject-specific estimates of model parameters (also called part-worths) in choice models that can be analyzed through hierarchical clustering or some other type of cluster analysis to reveal market segments.

Examples of the Choice Platform

One Table Format with No Choice

In a study of pizza preferences, each respondent is presented with four choice sets, each containing two profiles. Some respondents do not express a preference for either profile. The data are presented in a one-table format. When a respondent does not express a preference, the respondent's choice indicator is entered as missing.

1. Select **Help > Sample Data Library** and open Pizza Combined No Choice.jmp.

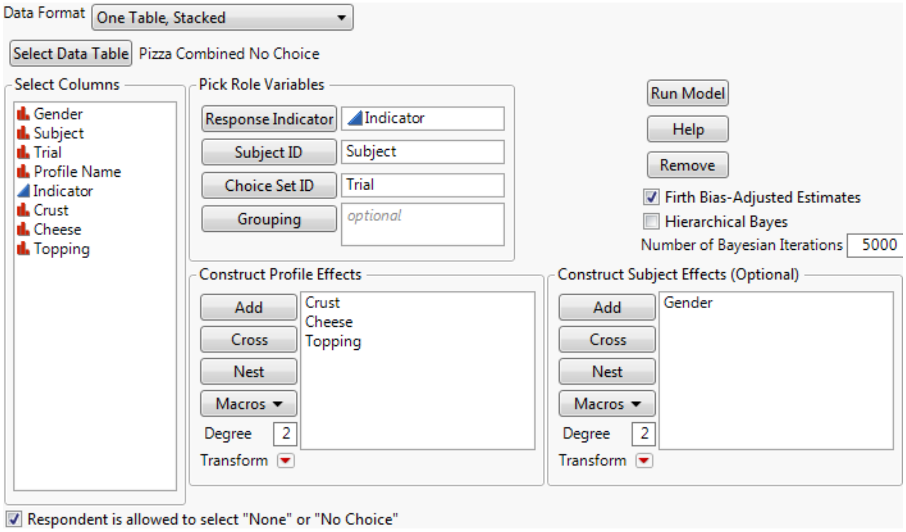
   Choice sets are defined by the combination of Subject and Trial. Notice that there are missing values in the Indicator column for some choice sets.

2. Select **Analyze** > **Consumer Research** > **Choice**.

   The One Table, Stacked data format is the default.

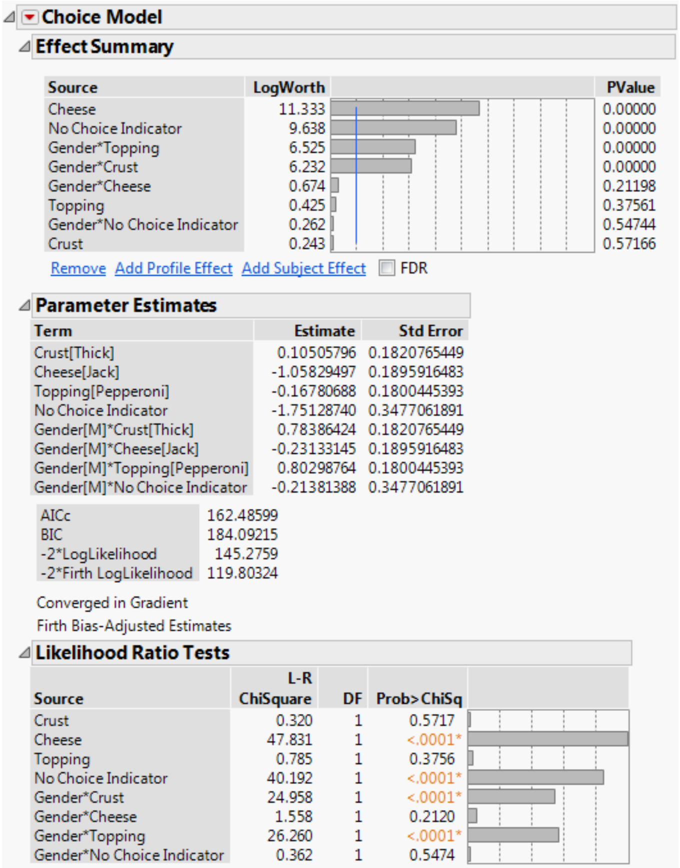
3. Click **Select Data Table**.
4. Select **Pizza Combined No Choice** and click **OK**.
5. Complete the launch window as follows:
   – Select Indicator and click **Response Indicator**.
   – Select Subject and click **Subject ID**.
   – Select Trial and click **Choice Set ID**.
   – Select Crust, Cheese, and Topping and click **Add** in the Construct Profile Effects panel.
   – Select Gender and click **Add** in the Construct Subject Effects (Optional) panel.

Figure 7.2 Completed Launch Window



6. Check the box next to **Respondent is allowed to select “None” or “No Choice”**.
7. Click **Run Model**.

Figure 7.3 Report Showing No Choice as an Effect



Choice Model

Effect Summary

Source	LogWorth	PValue
Cheese	11.333	0.00000
No Choice Indicator	9.638	0.00000
Gender*Topping	6.525	0.00000
Gender*Crust	6.232	0.00000
Gender*Cheese	0.674	0.21198
Topping	0.425	0.37561
Gender*No Choice Indicator	0.262	0.54744
Crust	0.243	0.57166

Remove Add Profile Effect Add Subject Effect FDR

Parameter Estimates

Term	Estimate	Std Error
Crust[Thick]	0.10505796	0.1820765449
Cheese[Jack]	-1.05829497	0.1895916483
Topping[Pepperoni]	-0.16780688	0.1800445393
No Choice Indicator	-1.75128740	0.3477061891
Gender[M]*Crust[Thick]	0.78386424	0.1820765449
Gender[M]*Cheese[Jack]	-0.23133145	0.1895916483
Gender[M]*Topping[Pepperoni]	0.80298764	0.1800445393
Gender[M]*No Choice Indicator	-0.21381388	0.3477061891

AICc	162.48599
BIC	184.09215
-2*LogLikelihood	145.2759
-2*Firth LogLikelihood	119.80324

Converged in Gradient
Firth Bias-Adjusted Estimates

Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Crust	0.320	1	0.5717
Cheese	47.831	1	<.0001*
Topping	0.785	1	0.3756
No Choice Indicator	40.192	1	<.0001*
Gender*Crust	24.958	1	<.0001*
Gender*Cheese	1.558	1	0.2120
Gender*Topping	26.260	1	<.0001*
Gender*No Choice Indicator	0.362	1	0.5474

The Effect Summary report shows the effects in order of significance. Cheese is the most significant effect, followed by the No Choice Indicator, which is treated as a model effect. The subject effect interactions Gender*Topping and Gender*Crust are also significant, indicating that preferences for Topping and Crust depend on Gender market segments.

To get some insight on the nature of the No Choice responses, select and view those choice sets that resulted in No Choice.

8. In the data table, right-click in a cell in the Indicator column where the response is missing and select **Select Matching Cells**.

9. In the Rows panel, right-click **Selected** and select **Data View**.

Figure 7.4 Choice Sets with No Choice Responses

	Gender	Subject	Trial	Profile Name	Indicator	Crust	Cheese	Topping
1	F	2	4	TrimPepperjack	•	Thin	Jack	Pepperoni
2	F	2	4	TrimOni	•	Thin	Mozzarella	Pepperoni
3	M	7	2	Trimella	•	Thin	Mozzarella	None
4	M	7	2	TrimJack	•	Thin	Jack	None
5	M	7	3	Trimella	•	Thin	Mozzarella	None
6	M	7	3	TrimJack	•	Thin	Jack	None
7	F	8	2	ThickElla	•	Thick	Mozzarella	None
8	F	8	2	ThickJack	•	Thick	Jack	None
9	M	11	4	ThickOni	•	Thick	Mozzarella	Pepperoni
10	M	11	4	ThickJackoni	•	Thick	Jack	Pepperoni
11	F	14	2	TrimOni	•	Thin	Mozzarella	Pepperoni
12	F	14	2	TrimPepperjack	•	Thin	Jack	Pepperoni
13	F	18	3	ThickJack	•	Thick	Jack	None
14	F	18	3	ThickElla	•	Thick	Mozzarella	None
15	F	24	3	ThickJack	•	Thick	Jack	None
16	F	24	3	TrimOni	•	Thin	Mozzarella	Pepperoni
17	M	29	1	TrimPepperjack	•	Thin	Jack	Pepperoni
18	M	29	1	ThickJackoni	•	Thick	Jack	Pepperoni

In the table in Figure 7.4, consider the profiles in the first seven choice sets, which are defined by the Subject and Trial combinations in rows 1 to 14. The only difference within each choice set is the Cheese. There is an indication that some respondents might not be able to detect the difference in cheeses. However, the analysis takes the No Choice Indicator into account and concludes that, despite this behavior, Cheese is significant.

To see how to further analyze data of this type, see "Find Optimal Profiles" on page 127.

Multiple Table Format

In this example, you examine pizza choices where three attributes, with two levels each, are presented to the subjects.

This example uses three data tables: Pizza Profiles.jmp, Pizza Responses.jmp, and Pizza Subjects.jmp.

1. Select **Help > Sample Data Library** and open Pizza Profiles.jmp, Pizza Responses.jmp, and Pizza Subjects.jmp.
   - The profile data table, Pizza Profiles.jmp, lists all the pizza choice combinations that you want to present to the subjects. Each choice combination is given an ID.
   - The responses data table, Pizza Responses.jmp, contains the design and results. For the experiment, each subject is given four choice sets, where each choice set consists of two choice profiles (Choice1 and Choice2). The subject selects a preference (Choice) for each choice set. For information about how to construct a choice design, see the Discrete Choice Designs chapter in the *Design of Experiments Guide*. Notice that each value in the Choice column is an ID value in the Profile data table that contains the attribute information.

- The subjects data table, Pizza Subjects.jmp, includes a Subject ID column and a single characteristic of the subject, Gender. Each value of Subject in the Pizza Subjects.jmp data table corresponds to values in the Subject column in the Pizza Responses.jmp data table.

2. Select **Analyze > Consumer Research > Choice** to open the launch window.

Note: This can be done from any of the three open data tables.

3. From the Data Format menu, select **Multiple Tables, Cross-Referenced**.

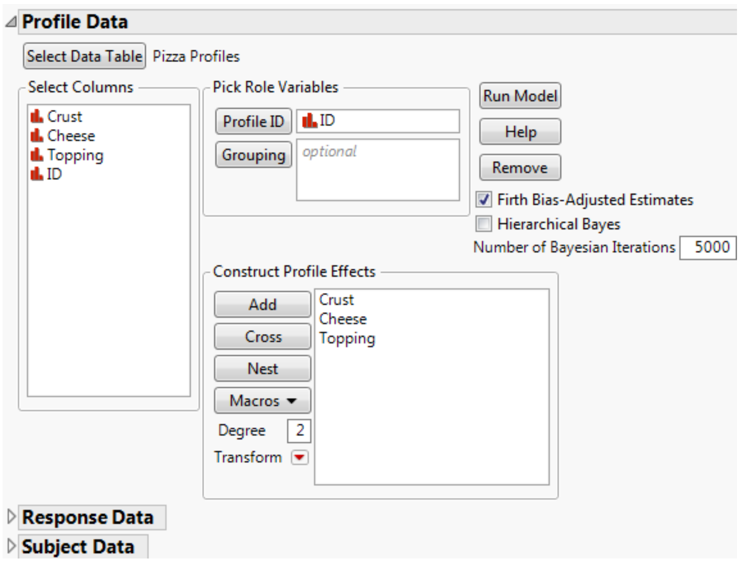
   There are three separate sections, one for each of the data sources.

4. Click **Select Data Table** under Profile Data.

   A Profile Data Table window appears, which prompts you to specify the data table for the profile data.

5. Select Pizza Profiles.jmp and click **OK**.
6. Select ID and click **Profile ID**.
7. Select Crust, Cheese, and Topping and click **Add**.

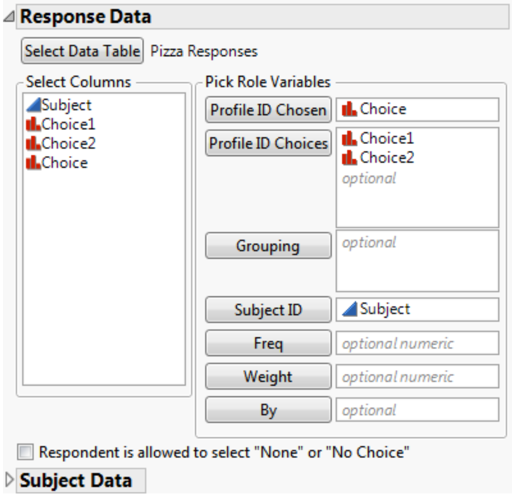
Figure 7.5 Profile Data



8. Click the disclosure icon next to Response Data to open the outline and click **Select Data Table**.
9. Select Pizza Responses.jmp and click **OK**.
10. Do the following:
    - Select Choice and click **Profile ID Chosen**.

- Select Choice1 and Choice2 and click **Profile ID Choices**.
- Select Subject and select **Subject ID**.

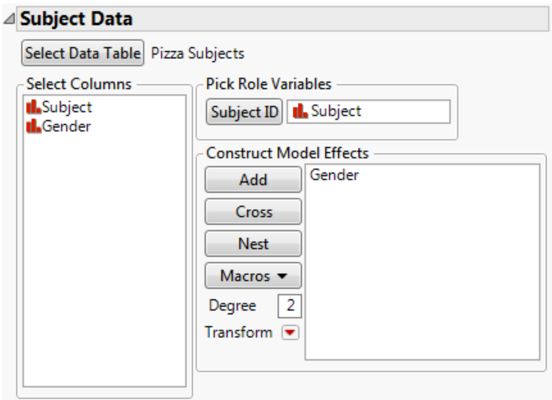
Figure 7.6 Response Data Window



Choice1 and Choice2 are the profiles presented to a subject in each of four choice sets. The Choice column contains the chosen preference between Choice1 and Choice2.

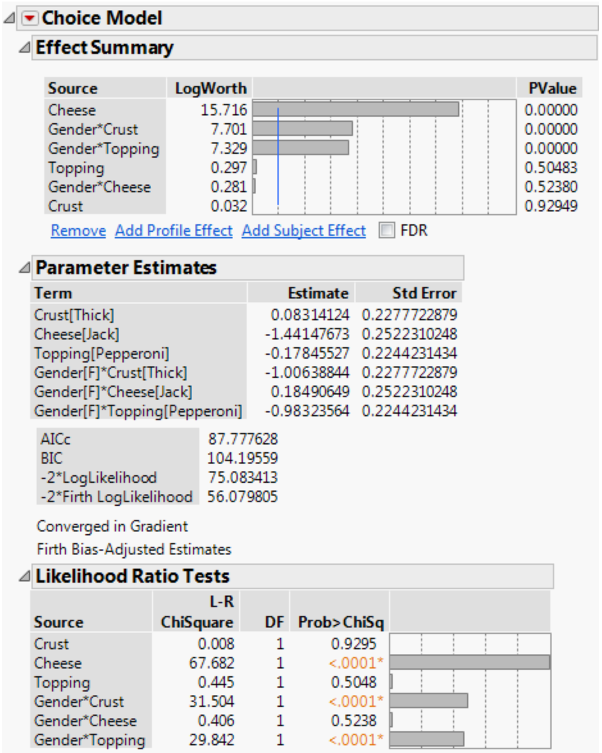
11. Click the disclosure icon next to Subject Data to open the outline and click **Select Data Table**.
12. Select Pizza Subjects.jmp and click **OK**.
13. Select Subject and click **Subject ID**.
14. Select Gender and click **Add**.

Figure 7.7 Subject Data Window



15. Click **Run Model**.

Figure 7.8 Choice Model Results



Choice Model

Effect Summary

Source	LogWorth	PValue
Cheese	15.716	0.00000
Gender*Crust	7.701	0.00000
Gender*Topping	7.329	0.00000
Topping	0.297	0.50483
Gender*Cheese	0.281	0.52380
Crust	0.032	0.92949

Remove Add Profile Effect Add Subject Effect FDR

Parameter Estimates

Term	Estimate	Std Error
Crust[Thick]	0.08314124	0.2277722879
Cheese[Jack]	-1.44147673	0.2522310248
Topping[Pepperoni]	-0.17845527	0.2244231434
Gender[F]*Crust[Thick]	-1.00638844	0.2277722879
Gender[F]*Cheese[Jack]	0.18490649	0.2522310248
Gender[F]*Topping[Pepperoni]	-0.98323564	0.2244231434

AICc	87.777628
BIC	104.19559
-2*LogLikelihood	75.083413
-2*Firth LogLikelihood	56.079805

Converged in Gradient
Firth Bias-Adjusted Estimates

Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Crust	0.008	1	0.9295
Cheese	67.682	1	<.0001*
Topping	0.445	1	0.5048
Gender*Crust	31.504	1	<.0001*
Gender*Cheese	0.406	1	0.5238
Gender*Topping	29.842	1	<.0001*

Six effects are entered into the model. The effects Crust, Cheese, and Topping are product attributes. The interaction effects, Gender*Crust, Gender*Cheese, and Gender*Topping are subject-effect interactions with the attributes. These interaction effects enable you to construct products that meet market-segment preferences.

Note: For Choice models, subject effects cannot be entered as main effects. They only appear as interaction terms.

The Effect Summary and Likelihood Ratio Tests reports show strong interactions between Gender and Crust and between Gender and Topping. Notice that the main effects of Crust and Topping are not significant. If you had not included subject-level effects, you might have overlooked important information relative to market segmentation.

Find Optimal Profiles

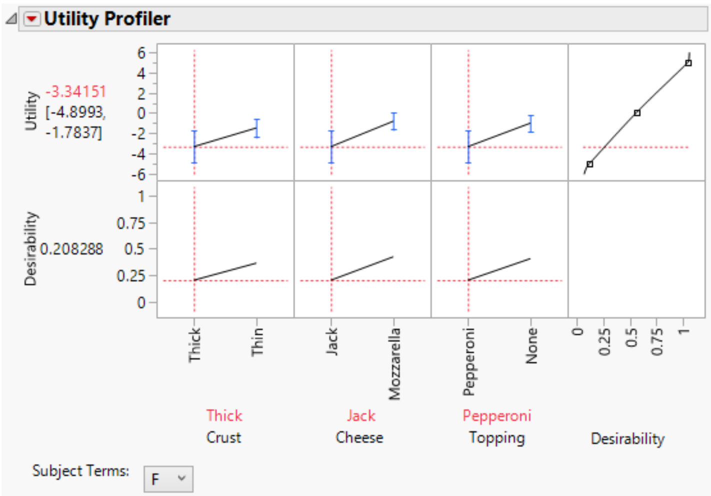
Next, you use the Utility Profiler to explore your results and find optimal settings for the attributes.

1. Click the Choice Model red triangle and select **Utility Profiler**.

   The Subject Terms menu beneath the profiler indicates that it is showing results for females.

2. Click the red triangle next to Utility Profiler and select **Optimization and Desirability > Desirability Functions**.

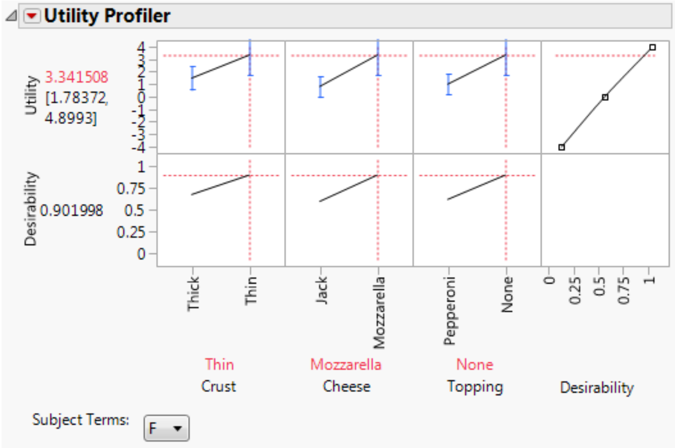
Figure 7.9 Utility Profiler with Desirability Function



A desirability function that maximizes utility is added to the profiler. See the Profiler chapter in the *Profilers* book.

3. Click the red triangle next to Utility Profiler and select **Optimization and Desirability > Maximize Desirability**.

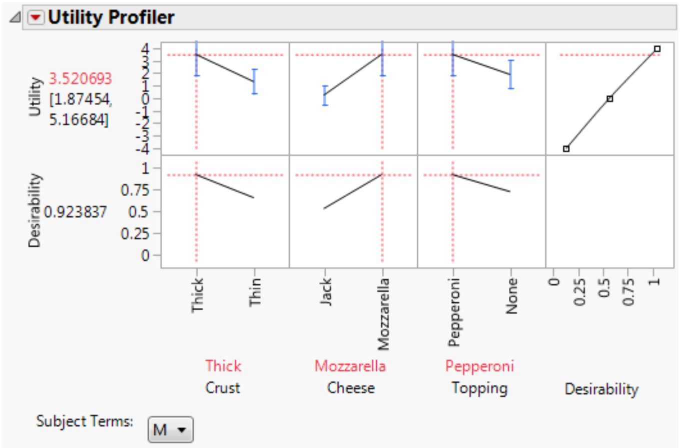
Figure 7.10 Utility Profiler with Optimal Settings for Females



The optimal settings for females are a thin crust, Mozzarella cheese, and no topping.

4. From the Subject Terms menu, select **M**.

Figure 7.11 Utility Profiler with Male Level Factor Setting



The optimal settings for males are a thick crust, Mozzarella cheese, and a Pepperoni topping.

In this example, understanding the preferences of gender-defined market segments enables you to provide two pizza choices that appeal to two segments of customers.

Launch the Choice Platform

Launch the Choice platform by selecting **Analyze > Consumer Research > Choice**.

Your data for the Choice platform can be combined in a single data table or it can reside in two or three separate data tables. When the Choice window opens, the first menu item asks you to specify the Data Format.

One Table, Stacked

For this format, the data are combined into a single data table. There is a row for every profile presented to a subject and an indicator of whether that profile was selected.

For an example of data in the one-table format, see “One Table Format with No Choice” on page 120. For details, see “Launch Window for One Table, Stacked” on page 130.

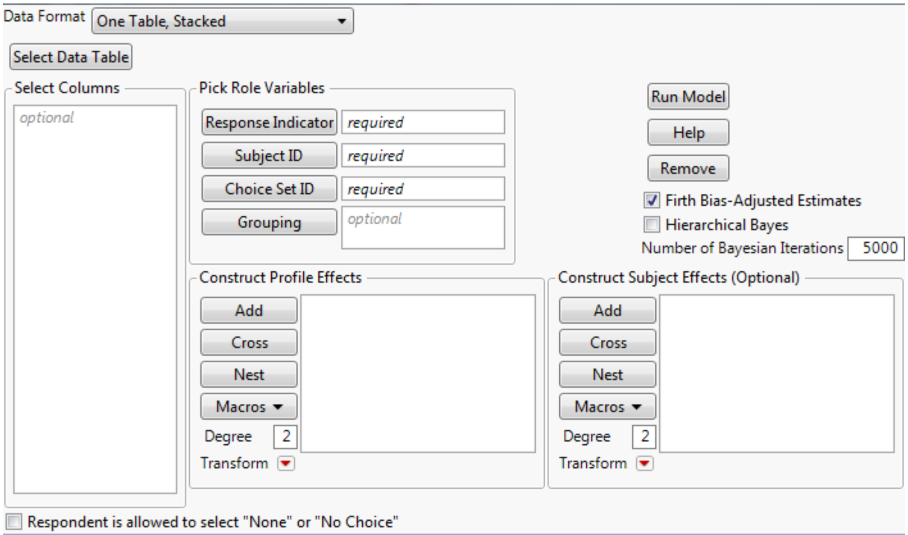
Multiple Tables, Cross-referenced

For this format, the data are stored in two or three separate tables: a Profile Data and Response Data table are required and a Subject Data table is optional. The Choice Launch Window contains three sections, each corresponding to a different data table. You can expand or collapse each section of the launch window, as needed.

For an example of data in the multiple-tables format, see “Multiple Table Format” on page 123. For details, see “Launch Window for Multiple Tables, Cross-Referenced” on page 131.

Launch Window for One Table, Stacked

Figure 7.12 Launch Window for One Table, Stacked Data Format



Select Data Table Select or open the data table that contains the combined data. Select Other to open a file that is not already open.

Response Indicator A column that contains values that indicate the preferred choice. A 1 indicates the preferred profile and a 0 indicates the other profiles. If respondents are given an option to select no preference, enter missing values for choice sets where no preference is indicated. See "Respondent is allowed to select "None" or "No Choice"" on page 131.

Subject ID An identifier for the study participant.

Choice Set ID An identifier for the choice set presented to the subject for a given preference determination.

Grouping A column which, when used with the Choice Set ID column, uniquely designates each choice set. For example, if a choice set has Choice Set ID = 1 for Survey = A, and another choice set has Choice Set ID = 1 for Survey = B, then Survey should be used as a Grouping column.

Construct Profile Effects Add effects constructed from the attributes in the profiles.

For information about the Construct Profile Effects panel, see the Construct Model Effects section in the Model Specification chapter of the *Fitting Linear Models* book.

Construct Subject Effects (Optional) Add effects constructed from subject-related factors.

For information about the Construct Subject Effects panel, see the Construct Model Effects section in the Model Specification chapter of the *Fitting Linear Models* book.

Firth Bias-adjusted Estimates Computes bias-corrected MLEs that produce better estimates and tests than MLEs without bias correction. These estimates also improve separation problems that tend to occur in logistic-type models. Refer to Heinze and Schemper (2002) for a discussion of the separation problem in logistic regression.

Hierarchical Bayes Uses a Bayesian approach to estimate subject-specific parameters. See "Bayesian Parameter Estimates" on page 138.

Number of Bayesian Iterations (Applicable only if Hierarchical Bayes is selected.) The total number of iterations of the adaptive Bayes algorithm used to estimate subject-specific parameters. This number includes a burn-in period of iterations that are discarded. The number of burn-in iterations is equal to half of the Number of Bayesian Iterations specified on the launch window.

Respondent is allowed to select "None" or "No Choice" Enters a No Choice Indicator into the model for response rows containing missing values. For the One Table, Stacked data format, the No Choice rows must contain (numeric) missing values in the Response Indicator column. The option appears at the bottom of the launch window.

Launch Window for Multiple Tables, Cross-Referenced

Figure 7.13 Launch Window for Multiple Tables, Cross-Referenced Data Format

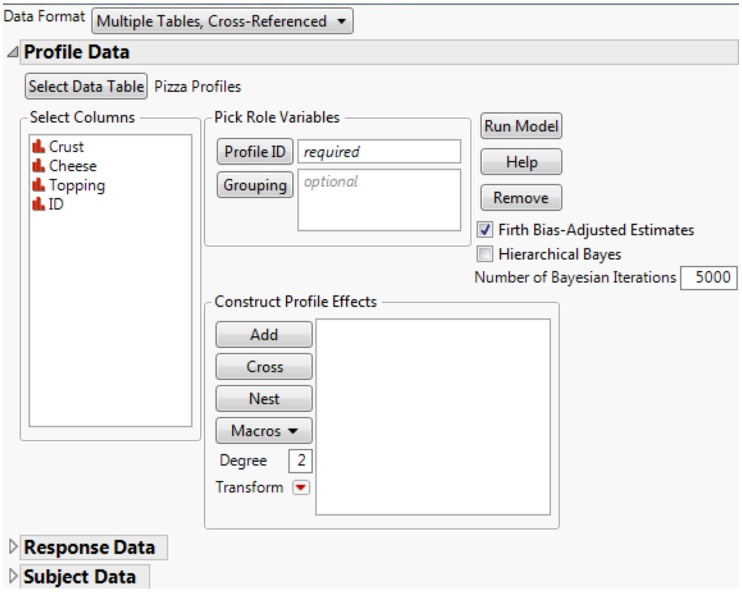


Figure 7.13 shows the launch window for Multiple Tables, using Pizza Profiles.jmp as the Profile table.

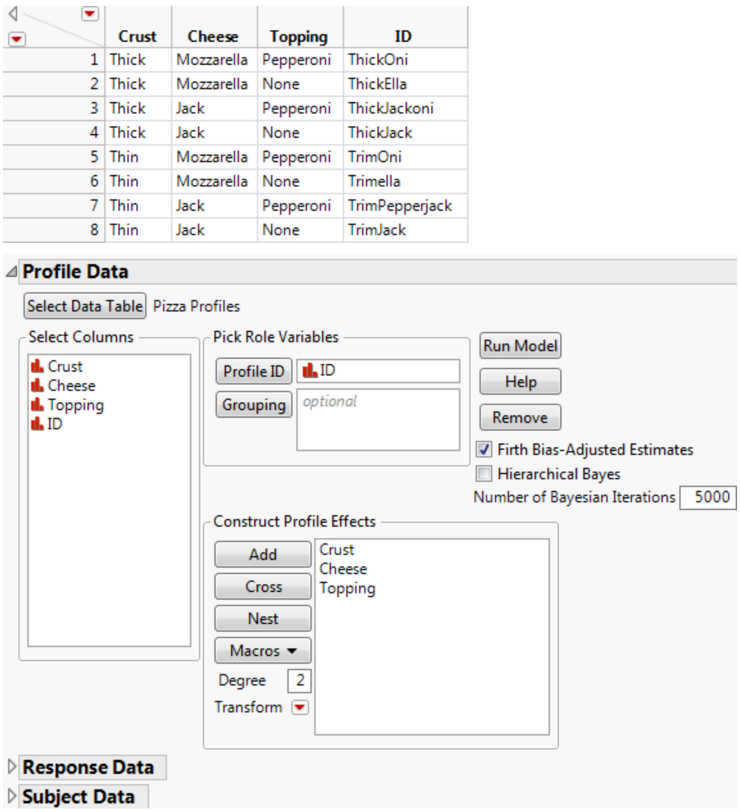
In the case of Multiple Tables, Cross-referenced, the launch window has three sections:

- "Profile Data" on page 132
- "Response Data" on page 133
- "Subject Data" on page 135

Profile Data

The profile data table describes the attributes associated with each choice. Each attribute defines a column in the data table. There is a row for each profile. A column in the table contains a unique identifier for each profile. Figure 7.14 shows the Pizza Profiles.jmp data table and a completed Profile Data panel.

Figure 7.14 Profile Data Table and Completed Profile Data Outline



	Crust	Cheese	Topping	ID
1	Thick	Mozzarella	Pepperoni	ThickOni
2	Thick	Mozzarella	None	ThickElla
3	Thick	Jack	Pepperoni	ThickJackoni
4	Thick	Jack	None	ThickJack
5	Thin	Mozzarella	Pepperoni	TrimOni
6	Thin	Mozzarella	None	Trimella
7	Thin	Jack	Pepperoni	TrimPepperjack
8	Thin	Jack	None	TrimJack

Select Data Table Select or open the data table that contains the profile data. Select Other to open a file that is not already open.

Profile ID Identifier for each row of attribute combinations (profile). If the **Profile ID** column does not uniquely identify each row in the profile data table, you need to add **Grouping**

columns. Add **Grouping** columns until the combination of **Grouping** and **Profile ID** columns uniquely identify the row, or profile.

Grouping A column which, when used with the Choice Set ID column, uniquely designates each choice set. For example, if Profile ID = 1 for Survey = A, and a different Profile ID = 1 for Survey = B, then Survey would be used as a **Grouping** column.

Construct Profile Effects Add effects constructed from the attributes in the profiles.

For information about the Construct Profile Effects panel, see the Construct Model Effects section in the Model Specification chapter of the *Fitting Linear Models* book.

Firth Bias-adjusted Estimates Computes bias-corrected MLEs that produce better estimates and tests than MLEs without bias correction. These estimates also improve separation problems that tend to occur in logistic-type models. Refer to Heinze and Schemper (2002) for a discussion of the separation problem in logistic regression.

JMP PRO **Hierarchical Bayes** Uses a Bayesian approach to estimate subject-specific parameters. See "Bayesian Parameter Estimates" on page 138.

JMP PRO **Number of Bayesian Iterations** (Applicable only if Hierarchical Bayes is selected.) The total number of iterations of the adaptive Bayes algorithm used to estimate subject effects. This number includes a burn-in period of iterations that are discarded. The number of burn-in iterations is equal to half of the Number of Bayesian Iterations specified on the launch window.

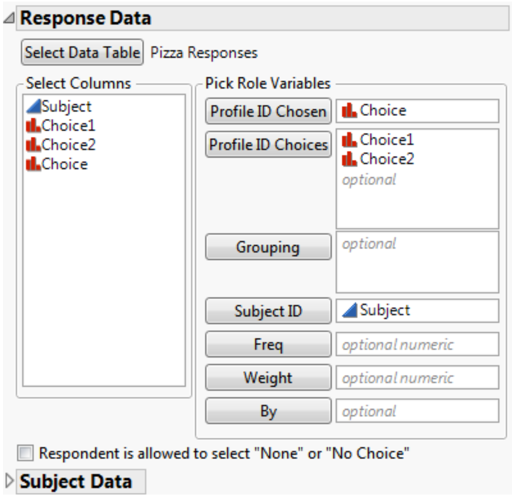
Response Data

The response data table includes a subject identifier column, columns that list the profile identifiers for the profiles in each choice set, and a column containing the preferred profile identifier. There is a row for each subject and choice set. Grouping variables can be used to distinguish choice sets when the data contain more than one group of choice sets. Figure 7.15 shows the Pizza Responses.jmp data table and a completed Response Data panel.

Grouping variables can be used to align choice indices when more than one group is contained within the data.

Figure 7.15 Response Data Table and Completed Responses Data Outline

	Subject	Choice1	Choice2	Choice
1	1	ThickJack	TrimPepperjack	TrimPepperjack
2	1	TrimPepperjack	ThickElla	ThickElla
3	1	TrimOni	Trimella	TrimOni
4	1	ThickElla	ThickJack	ThickElla
5	2	Trimella	ThickJackoni	Trimella
6	2	TrimJack	ThickElla	ThickElla
7	2	Trimella	TrimPepperjack	Trimella
8	2	TrimPepperjack	TrimOni	TrimOni
9	3	TrimOni	ThickJackoni	TrimOni
10	3	TrimPepperjack	ThickElla	ThickElla
11	3	ThickJackoni	TrimPepperjack	ThickJackoni
12	3	ThickOni	Trimella	ThickOni
13	4	ThickElla	ThickOni	ThickElla
14	4	TrimPepperjack	ThickJack	ThickJack



Select Data Table Select or open the data table that contains the profile data. Select Other to open a file that is not already open.

Profile ID Chosen The Profile ID from the Profile data table that represents the subject's selected profile.

Grouping A column which, when used with the **Profile ID Chosen** column, uniquely designates each choice set.

Profile ID Choices The Profile IDs of the set of possible profiles.

Subject ID An identifier for the study participant.

Freq A column containing frequencies. If *n* is the value of the Freq variable for a given row, then that row is used in computations n times. If it is less than 1 or missing, then JMP does not use it to calculate any analyses.

Weight A column containing a weight for each observation in the data table. The weight is included in analyses only when its value is greater than zero.

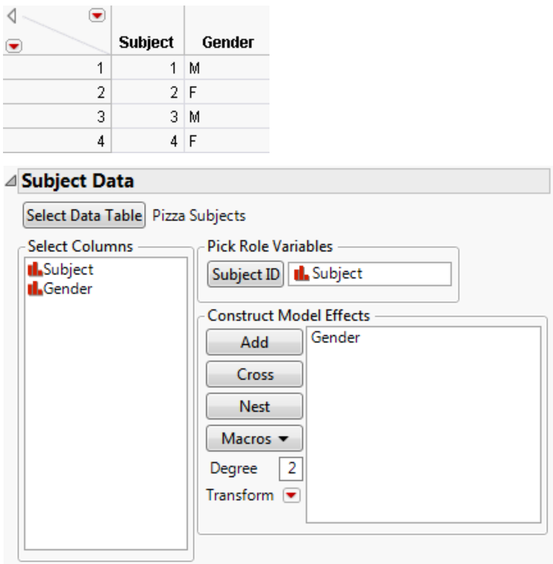
By A column whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed as a separate analysis on a separate table. The results are presented in separate reports. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

Respondent is allowed to select "None" or "No Choice" Enters a No Choice Indicator into the model for response rows containing missing values. For the Multiple Tables, Cross-Referenced data format, the No Choice rows must contain (categorical) missing values in the Profile ID Chosen column in the Response Data table. The option appears at the bottom of the Response Data panel.

Subject Data

The subject data table is optional and depends on whether you want to model subject effects. The table contains a column with the subject identifier used in the response table, and columns for attributes or characteristics of the subjects. You can put subject data in the response data table, but you should specify the subject effects in the Subject Data outline. Figure 7.16 shows the Pizza Subjects.jmp data table and a completed Subject Data panel.

Figure 7.16 Subject Data Table and Completed Subject Data Outline



Select Data Table Select or open the data table that contains the subject data. Select Other to open a file that is not already open.

Subject ID Unique identifier for the subject.

Construct Model Effects Add effects constructed from columns in the subject data table.

For information about the Construct Model Effects panel, see the Construct Model Effects section in the Model Specification chapter of the *Fitting Linear Models* book.

Choice Model Report

Unless Hierarchical Bayes is selected on the launch window, the Choice Model report consists of the following:

- "Effect Summary" on page 136
- "Parameter Estimates" on page 137
- "Likelihood Ratio Tests" on page 138

Note: The Effect Summary and Likelihood Ratio Tests reports appear by default only if the data set is small enough for them to be calculated in a reasonable amount of time. If they do not appear, select Likelihood Ratio Tests from the red triangle menu to make both appear.

JMP PRO If Hierarchical Bayes is selected on the launch window, the following report appears:

- "Bayesian Parameter Estimates" on page 138

Effect Summary

The Effect Summary report appears if your model contains more than one effect and if it can be calculated quickly. (If the report does not appear, select Likelihood Ratio Tests from the red triangle menu to make both reports appear.) It lists the effects estimated by the model and gives a plot of the LogWorth (or FDR LogWorth) values for these effects. 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. For details, see the Effect Summary Report section in the Standard Least Squares Report and Options chapter in the *Fitting Linear Models* book.

The Effect Summary report does not appear when Bayesian Subject Effects is checked in the launch window. This is because likelihood ratio tests are not conducted in this case.

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 $-\log_{10}(p\text{-value})$. 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. Select the **FDR** check box to replace the LogWorth column with the **FDR LogWorth** column.

Bar Chart Shows a bar chart 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 the *p*-value corresponding to the significance test displayed in the Likelihood Ratio Tests 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 details about the FDR correction, see Benjamini and Hochberg, 1995. For details about the false discovery rate, see the Response Screening chapter in the *Predictive and Specialized Modeling* book or Westfall et al. (2011).

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 Profile Effect Opens a panel that contains a list of all columns in the data table for the OneTable, Stacked data format, and for the columns in the Profile Data table for the Multiple Tables, Cross-Referenced data format. 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.

Add Subject Effect Opens a panel that contains a list of all columns in the data table for the OneTable, Stacked data format, and for the columns in the Subject Data table for the Multiple Tables, Cross-Referenced data format. 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.

Parameter Estimates

The Parameter Estimates report gives estimates and standard errors of the coefficients of utility associated with the effects listed in the Term column. The coefficients associated with attributes are sometimes referred to as *part-worths*. When the Firth Bias-Adjusted Estimates option is selected in the launch window, the parameter estimates are based on the Firth bias-corrected maximum likelihood estimators. These estimates considered to be more accurate than MLEs without bias correction. For details about utility, see “Utility and Probabilities” on page 171.

Comparison Criteria

The following fit statistics are shown as part of the report and can be used to compare models: AICc (corrected Akaike's Information Criterion), BIC (Bayesian Information Criterion), – 2*LogLikelihood, and – 2*Firth LogLikelihood. For details and formulas, see the section Likelihood, AICc, and BIC in the Statistical Details appendix of the *Fitting Linear Models* book.

The – 2*Firth LogLikelihood fit statistic is included in the report when the Firth Bias-Adjusted Estimates option is selected in the launch window. Note that this option is checked by default. The decision to use or not use the Firth Bias-Adjusted Estimates does not affect the AICc score or the – 2*LogLikelihood results.

Note: For each of these statistics, a smaller value indicates a better fit.

Likelihood Ratio Tests

The Likelihood Ratio Test report appears by default if the model is fit in less than five seconds. If the report does not appear, you can select the Likelihood Ratio Tests option from the Choice Model red triangle menu. The report gives the following:

Source Lists the effects in the model.

L-R ChiSquare The value of the likelihood ratio ChiSquare statistic for a test of the corresponding effect.

DF The degrees of freedom for the ChiSquare test.

Prob>ChiSq The p-value for the ChiSquare test.

Bar Graph Shows a bar chart of the L-R ChiSquare values.

JMP PRO Bayesian Parameter Estimates

(Appears only if Hierarchical Bayes is selected on the launch window.)The Bayesian Parameter Estimates report gives results for model effects. The estimates are based on a Hierarchical Bayes fit that integrates the subject-level covariates into the likelihood function and estimates their effects on the parameters directly. The subject-level covariates are estimated using a version of the algorithm described in Train (2001), which incorporates Adaptive Bayes and Metropolis-Hastings approaches. Posterior means and variances are calculated for each model effect. The algorithm also provides subject-specific estimates of the model effect parameters. See "Save Subject Estimates" on page 142.

During the estimation process, each individual is assigned his or her own vector of parameter estimates, essentially treating the estimates as random effects and covariates. The vector of coefficients for an individual is assumed to come from a multivariate normal distribution with arbitrary mean and covariance matrix. The likelihood function for the utility parameters for a given subject is based on a multinominal logit model for each subject's preference within a choice set, given the attributes in the choice set. The prior distribution for a given subject's vector of coefficients is normal with mean equal to zero and a diagonal covariance matrix with the same variance for each subject. The covariance matrix is assumed to come from an inverse Wishart distribution with a scale matrix that is diagonal with equal diagonal entries.

For each subject, a number of burn-in iterations at the beginning of the chain is discarded. By default, this number is equal to half of the Number of Bayesian Iterations specified on the launch window.

Figure 7.17 Bayesian Parameter Estimates Report

Choice Model

Bayesian Parameter Estimates

Term	Posterior Mean	Posterior Std Dev	Subject Std Dev
Crust[Thick]	0.13053926	0.2313481983	0.7783790503
Cheese[Jack]	-2.24826571	0.5771612629	0.5964736512
Topping[Pepperoni]	-0.39667788	0.2654992008	0.6675958456

Total Iterations	5000
Burn-In Iterations	2500
Number of Respondents	32
Avg Log Likelihood After Burn-In	-32.19285

Term The model term.

Posterior Mean The parameter estimate for the term's coefficient. For each iteration after the burn-in period, the mean of the subject-specific coefficient estimates is computed. The Posterior Mean is the average of these means.

Tip: Select the red-triangle option Save Bayes Chain to see the individual estimates for each iteration.

Posterior Std Dev The standard deviation of the means of the subject-specific estimates over the iterations after burn-in.

Subject Std Dev The standard deviation of the subject-specific estimates.

Tip: Select the red-triangle option Save Subject Estimates to see the individual estimates.

Total Iterations The total number of iterations performed, including the burn-in period.

Burn-In Iterations The number of burn-in iterations. This number is equal to half of the Number of Bayesian Iterations specified on the launch window.

Number of Respondents The number of subjects.

Avg Log Likelihood After Burn-In The average of the log-likelihood function, computed on values obtained after the burn-in period.

Choice Platform Options

The Choice Modeling platform has many available options. To access these options, select the Choice Model red triangle menu.

Note: When you use Hierarchical Bayes, the subject-level estimates are based on Monte Carlo sampling. For this reason, results obtained for the options below will vary from run to run.

Likelihood Ratio Tests See "Likelihood Ratio Tests" on page 138.

JMP PRO **Show MLE Parameter Estimates** (Available only if Hierarchical Bayes is selected on the launch window.) Shows non-Firth maximum likelihood estimates and standard errors for the coefficients of model terms. These estimates are used as starting values for the Hierarchical Bayes algorithm.

Joint Factor Tests (Not available if Hierarchical Bayes is selected on the launch window.) Tests each factor in the model by constructing a likelihood ratio test for all the effects involving that factor. For more information about Joint Factor Tests, see the Standard Least Squares Report and Options chapter in the *Fitting Linear Models* book.

Confidence Intervals If Hierarchical Bayes was not selected, shows a confidence interval for each parameter in the Parameter Estimates report.

If you selected Hierarchical Bayes, the confidence intervals appear in the Bayesian Parameter Estimates report. The intervals are constructed assuming a normal distribution and are based on the Posterior Mean and Posterior Std Dev.

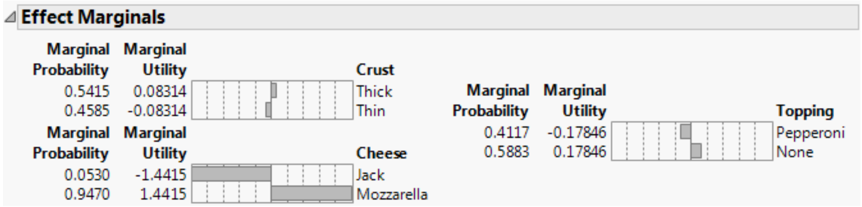
Correlation of Estimates If Hierarchical Bayes was not selected, shows the correlations between the maximum likelihood parameter estimates.

If you selected Hierarchical Bayes, shows the correlation matrix for the posterior means of the parameter estimates. The correlations are calculated from the iterations after burn-in. The posterior means from each iteration after burn-in are treated as if they are columns in a data table. The Correlation of Estimates table is obtained by calculating the correlation matrix for these columns.

Effect Marginals Shows marginal probabilities and marginal utilities for each main effect in the model. The marginal probability is the probability that an individual selects attribute A over B with all other attributes at their mean or default levels.

In Figure 7.18, the marginal probability of any subject choosing a pizza with mozzarella cheese, thick crust and pepperoni, over that same pizza with Monterey Jack cheese instead of mozzarella, is 0.9470.

Figure 7.18 Example of Marginal Effects



Utility Profiler Shows the predicted utility for different factor settings. The utility is the value predicted by the linear model. See "Find Optimal Profiles" on page 127 for an example of the Utility Profiler. For details about utility, see "Utility and Probabilities" on page 171. For details about the Utility Profiler options, see the Prediction Profiler Options section in the Profiler chapter of the *Profilers* book.

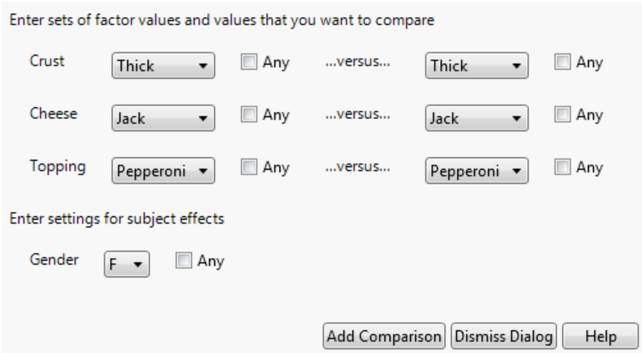
Probability Profiler Allows you to compare choice probabilities among a number of potential products. This predicted probability is defined as $(\exp(U))/(\exp(U) + \exp(U_b))$, where U is the utility for the current settings and U_b is the utility for the baseline settings. This implies that the probability for the baseline settings is 0.5. For details, see "Utility and Probabilities" on page 171.

See "Comparisons to Baseline" on page 150 for an example of using the Probability Profiler. For details about the Probability Profiler options, see the Prediction Profiler Options section in the Profiler chapter of the *Profilers* book.

Multiple Choice Profiler Provides the number of probability profilers that you specify. This enables you to set each profiler to the settings of a given profile so that you can compare the probabilities of choosing each profile relative to the others. See "Multiple Choice Comparisons" on page 152 for an example of using the Multiple Choice Profiler. For details about the Multiple Choice Profiler options, see the Prediction Profiler Options section in the Profiler chapter of the *Profilers* book.

Comparisons Performs comparisons between specific alternative choice profiles. Enables you to select the factors and the values that you want to compare. You can compare specific configurations, including comparing all settings on the left or right by selecting the **Any** check boxes. If you have subject effects, you can select the levels of the subject effects to compare. Using Any does not compare all combinations across features, but rather all combinations of comparisons, one feature at a time, using the left settings as the settings for the other factors.

Figure 7.19 Utility Comparisons Window



Willingness to Pay Requires that your data table contains a continuous price column. Calculates how much a price must change allowing for the new feature settings to produce the same predicted outcome. The result is calculated using the Baseline settings for each background setting.

Save Utility Formula Creates a new data table that contains a formula column for utility. The new data table contains a row for each subject and profile combination, and columns for the profiles and the subject effects.

Save Gradients by Subject Constructs a new table that has a row for each subject containing the average (Hessian-scaled-gradient) steps for the likelihood function on each parameter. This corresponds to using a Lagrangian multiplier test for separating that subject from the remaining subjects. These values can later be clustered, using the built-in-script, to indicate unique market segments represented in the data. For more details, see “Gradients” on page 172. For an example, see “Example of Segmentation” on page 154.

JMP PRO **Save Subject Estimates** (Available only if Hierarchical Bayes is selected.) Creates a table where each row contains the subject-specific parameter estimates for each effect. The distribution of subject-specific parameter effects for each effect is centered at the estimate for the term given in the Bayesian Parameter Estimates report. The Subject Acceptance Rate gives the rate of acceptance for draws of new parameter estimates during the Metropolis-Hastings step. Generally, an acceptance rate of 0.20 is considered to be good. See “Bayesian Parameter Estimates” on page 138.

JMP PRO **Save Bayes Chain** (Available only if Hierarchical Bayes is selected.) Creates a table that gives information about the chain of iterations used in computing subject-specific Bayesian estimates. See “Save Bayes Chain” on page 145.

Model Dialog Shows the Choice launch window, which can be used to modify and re-fit the model. You can specify new data sets, new IDs, and new model effects.

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.

Willingness to Pay

The term *willingness to pay* refers to the price that a customer is willing to pay for new features, calculated to match a customer's utility for baseline features. For example, suppose that a customer is willing to pay $1,000 for a computer with a 40 GB hard drive. Willingness to Pay for an 80 GB hard drive is calculated by setting the Hard drive feature to 80 GB and then solving for the price that delivers the same utility as the $1000 40 GB hard drive.

Willingness to Pay Launch Window Options

When you select the Willingness to Pay option, the Willingness to Pay launch window is shown. The launch window in Figure 7.20 is obtained by selecting the Willingness to Pay option in the report that results from running the **Choice** data table script in Laptop Profile.jmp.

Factor The variables from the analysis. These can be product features or subject-specific attributes.

Baseline The baseline setting for each factor. If the factor is categorical, select the baseline value from a list. If the factor is numeric, enter the baseline value.

Role The type of factor. You can choose from the following list:

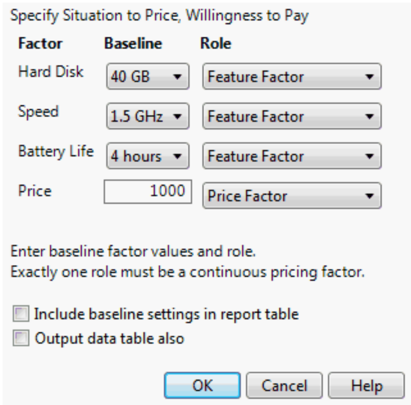
- **Feature Factor** A product or service feature from the experiment that you want to price.
- **Price Factor** A price factor in the experiment. The price factor must be continuous, and there can only be one specified price factor for each Willingness to Pay analysis.
- **Background Constant** A factor that you want to hold constant when comparing products. Generally, these are subject-specific variables.
- **Background Variable** A factor that you want to hold constant while solving for price in the Willingness to Pay calculation, but that should remain constant when comparing two different products. Generally, these are subject-level factors. Specifying a subject factor as a Background Variable rather than a Background Constant provides Willingness to Pay estimates for all levels of the variable.

Include baseline settings in report table Adds the baseline settings with a price change of zero to the Willingness to Pay report.

Tip: If you make an output table, use this option to display all the baseline settings as well as the attribute settings.

Output data table also Creates a data table containing the Willingness to Pay report.

Figure 7.20 Willingness to Pay Launch Window



Once you complete your first Willingness to Pay calculation, the platform remembers the baseline values and assigned roles that you selected. This enables you to do multiple Willingness to Pay comparisons without having to re-enter the baseline information. If there is no factor called Price, but there is a continuous factor used in the analysis, the continuous factor is automatically assigned as the Price factor in the Willingness to Pay window. Common cost variables that are not prices in the traditional sense include factors such as travel time or distance.

Willingness to Pay Report

The Willingness to Pay report displays the baseline value for each factor, as well as baseline utility values. For each factor, the report shows the feature setting, estimated price change, and new price. If there are no interaction or second-order effects, standard errors and confidence intervals are also shown. These are calculated using the delta method.

Figure 7.21 Willingness to Pay Report

Willingness to Pay

Factor	Baseline Value
Hard Disk	40 GB
Speed	1.5 GHz
Battery Life	4 hours
Price	1000

Baseline Utility
-3.4736

Factor	Feature Setting	Price Change	Std Error	Lower 95%	Upper 95%	New Price
Hard Disk	80 GB	$959.67	381.743	$211.47	$1,707.87	$1,959.67
Speed	2.0 GHz	$496.86	218.515	$68.58	$925.14	$1,496.86
Battery Life	6 hours	$387.88	191.088	$13.35	$762.40	$1,387.88

Standard deviations for Price Change calculated by Delta method.

Save Bayes Chain

You can use the Bayes Chain data table to determine whether your estimates have stabilized. The table that is created has a number of rows equal to the Number of Bayesian Iterations (specified on the launch window) plus one. The first row, Iteration 1, gives the starting values. The following rows show the results of the iterations, in order. The columns are arranged as follows:

Iteration Gives the iteration number, where the first row shows starting values.

Log Likelihood The log-likelihood of the model for that iteration. You can plot the Log Likelihood against Iteration to view behavior over the burn-in and tuning periods.

Adaptive Sigma for <model effect> Gives the estimate of the square root of the diagonal entries of the inverse Wishart distribution scale matrix for the corresponding effect.

Acceptance for <model effect> Gives the sampling acceptance rate for the corresponding effect.

Mean of <model effect> Gives the estimated mean for the corresponding effect.

Variance of <model effect> Gives the estimated variance for the corresponding effect.

Additional Examples

This section contains the following examples:

- "Example of Making Design Decisions" on page 146
- "Example of Segmentation" on page 154
- "Example of Logistic Regression Using the Choice Platform" on page 158
- "Example of Logistic Regression for Matched Case-Control Studies" on page 161

- "Example of Transforming Data to Two Analysis Tables" on page 163
- "Example of Transforming Data to One Analysis Table" on page 168

Example of Making Design Decisions

You can use the Choice Modeling platform to determine the relative importance of product attributes. Even if the attributes of a particular product that are important to the consumer are known, information about preference trade-offs with regard to these attributes might be unknown. By gaining such information, a market researcher or product designer is able to incorporate product features that represent the optimal trade-off from the perspective of the consumer. This example illustrates the advantages of this approach to product design.

It is already known that four attributes are important for laptop design: hard-disk size, processor speed, battery life, and selling price. The data gathered for this study are used to determine which of four laptop attributes (Hard Disk, Speed, Battery Life, and Price) are most important. It also assesses whether there are Gender or Job effects associated with these attributes.

This example has the following sections:

- "Complete the Launch Window" on page 146
- "Analyze the Model" on page 148
- "Comparisons to Baseline" on page 150
- "Multiple Choice Comparisons" on page 152

Complete the Launch Window

1. Select **Help > Sample Data Library** and open Laptop Runs.jmp.

Note: If you prefer not to follow the manual steps in this section, click the green triangle next to the script **Choice with Gender** to run the model, and go to "Analyze the Model" on page 148.

2. Click the green triangle next to the **Open Profile and Subject Tables** script.

   The script opens the Laptop Profile.jmp and Laptop Subjects.jmp data tables.

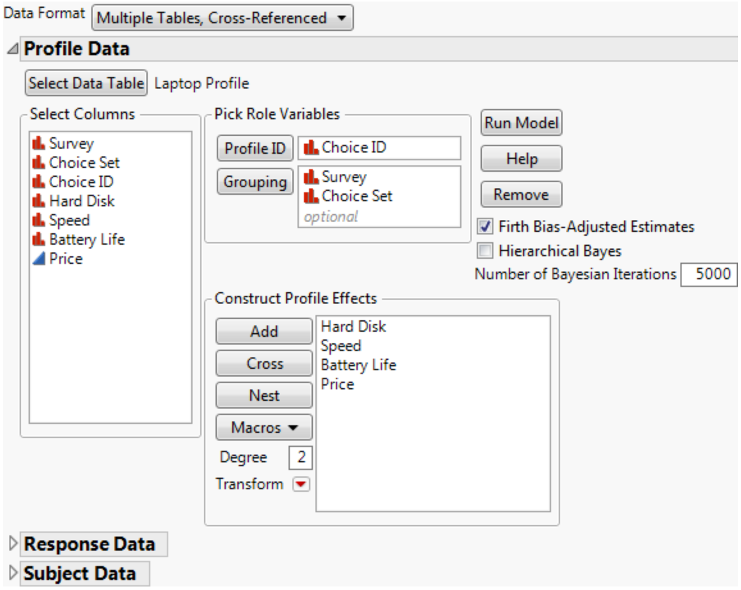
3. Select **Analyze** > **Consumer Research** > **Choice**.

   Note: This can be done from any of the three open data tables.

4. From the Data Format list, select **Multiple Tables, Cross-Referenced**.
5. Click **Select Data Table** under Profile Data and select Laptop Profile.jmp. Select Choice ID and click Profile ID.
6. Select Hard Disk, Speed, Battery Life, and Price and click **Add**.

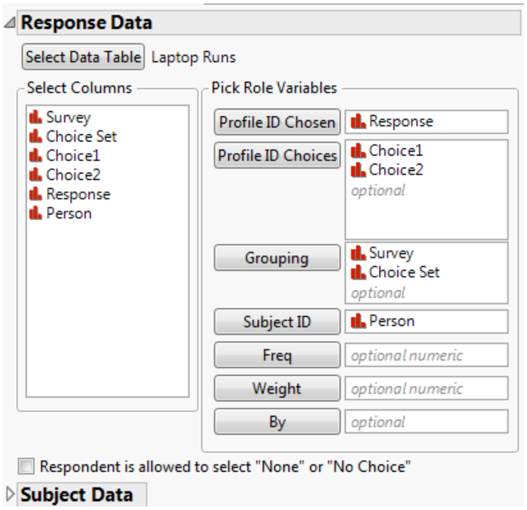
7. Select Survey and Choice Set and click **Grouping**.

Figure 7.22 Profile Data Window for Laptop Study



8. Open the **Response Data** outline.
9. From the **Select Data Table** list, select Laptop Runs.jmp.
10. Complete the Response Data table as follows:
    – Select Response and click **Profile ID Chosen**.
    – Select Choice1 and Choice2 and click **Profile ID Choices**.
    – Select Survey and Choice Set and click **Grouping**
    – Select Person and click **Subject ID**. The Response Data window is shown in Figure 7.23.

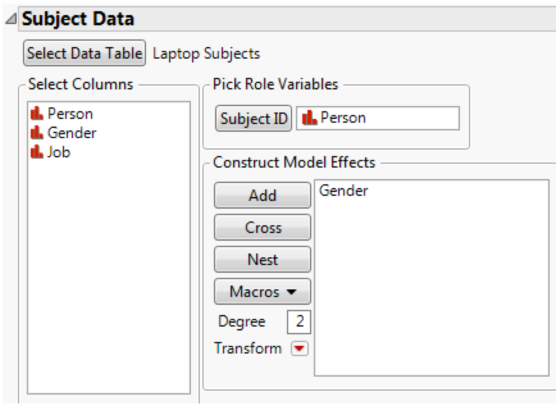
Figure 7.23 Response Data Window for Laptop Study



11. Open the **Subject Data** outline.
12. From the **Select Data Table** list, select Laptop Subjects.jmp.
13. Select Person and click **Subject ID**.
14. Select Gender click **Add**.

    The Subject Data window is shown in Figure 7.24.

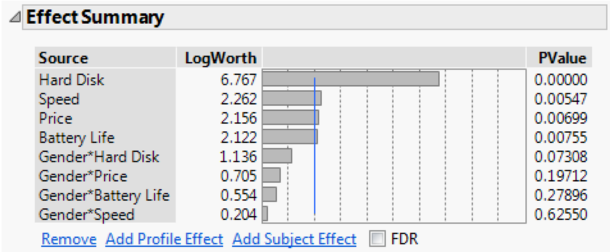
Figure 7.24 Subject Data Window for Laptop Study



Analyze the Model

1. Click **Run Model**.

Figure 7.25 Laptop Effect Summary



Effect Summary

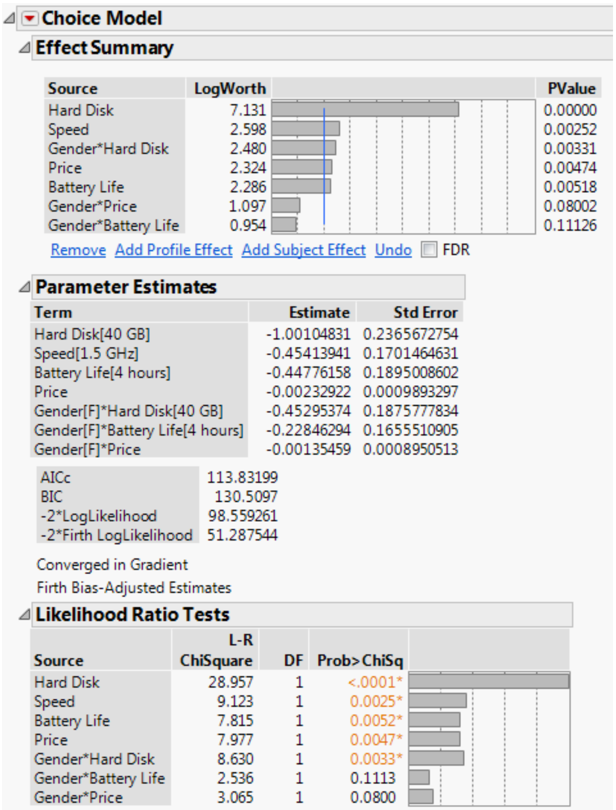
Source	LogWorth	PValue
Hard Disk	6.767	0.00000
Speed	2.262	0.00547
Price	2.156	0.00699
Battery Life	2.122	0.00755
Gender*Hard Disk	1.136	0.07308
Gender*Price	0.705	0.19712
Gender*Battery Life	0.554	0.27896
Gender*Speed	0.204	0.62550

Remove Add Profile Effect Add Subject Effect FDR

The Effect Summary report shows that Hard Disk is the most significant effect. You can reduce the model by removing terms with a p-value greater than 0.15. This process should be done one term at a time. Here, Gender*Speed is the least significant effect, with a p-value of 0.625.

2. In the Effect Summary report, select Gender*Speed and click **Remove**.

Figure 7.26 Laptop Results



Choice Model

Effect Summary

Source	LogWorth	PValue
Hard Disk	7.131	0.00000
Speed	2.598	0.00252
Gender*Hard Disk	2.480	0.00331
Price	2.324	0.00474
Battery Life	2.286	0.00518
Gender*Price	1.097	0.08002
Gender*Battery Life	0.954	0.11126

Remove Add Profile Effect Add Subject Effect Undo FDR

Parameter Estimates

Term	Estimate	Std Error
Hard Disk[40 GB]	-1.00104831	0.2365672754
Speed[1.5 GHz]	-0.45413941	0.1701464631
Battery Life[4 hours]	-0.44776158	0.1895008602
Price	-0.00232922	0.0009893297
Gender[F]*Hard Disk[40 GB]	-0.45295374	0.1875777834
Gender[F]*Battery Life[4 hours]	-0.22846294	0.1655510905
Gender[F]*Price	-0.00135459	0.0008950513

AICc	113.83199
BIC	130.5097
-2*LogLikelihood	98.559261
-2*Firth LogLikelihood	51.287544

Converged in Gradient

Firth Bias-Adjusted Estimates

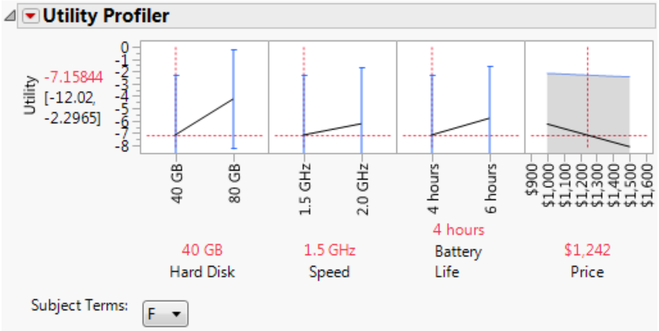
Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Hard Disk	28.957	1	<.0001*
Speed	9.123	1	0.0025*
Battery Life	7.815	1	0.0052*
Price	7.977	1	0.0047*
Gender*Hard Disk	8.630	1	0.0033*
Gender*Battery Life	2.536	1	0.1113
Gender*Price	3.065	1	0.0800

Once Gender*Speed is removed from the model, all effects have a *p*-value of 0.15 or less. Therefore, you use this as your final model.

3. Click the Choice Model red triangle and select **Utility Profiler**.

Figure 7.27 Laptop Profiler Results for Females

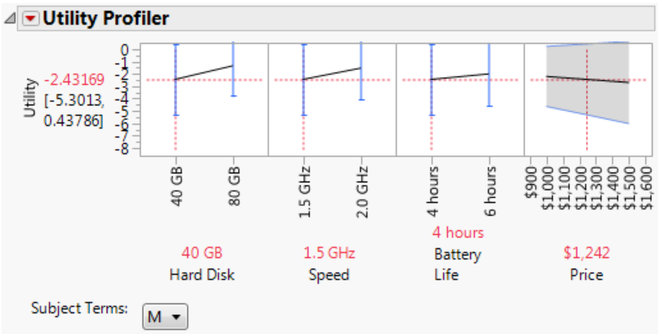


Tip: If your utility profiler does not look like Figure 7.27, click the red triangle next to Utility Profiler and select **Appearance > Adapt Y Axis**.

---

4. From the list next to Subject Terms, select **M**.

Figure 7.28 Laptop Profiler Results for Males in Development



The interaction effect between Gender and Hard Disk is highly significant, with a *p*-value of 0.0033. See Figure 7.26 on page 149. In the Utility Profilers, check the slope for Hard Disk for both levels of Gender. You see that the slope is steeper for females than for males.

Comparisons to Baseline

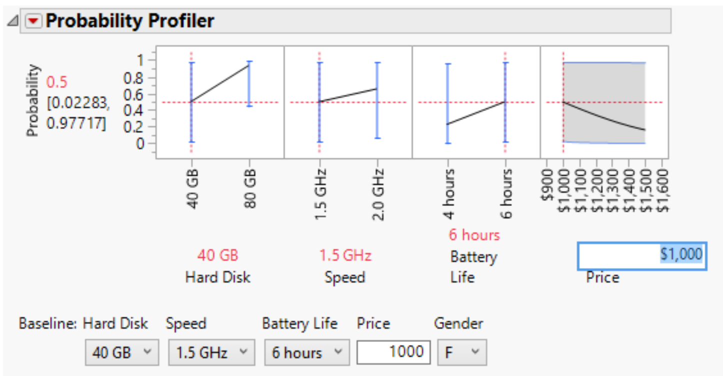
Suppose you are developing a new product. You want to explore the likelihood that a customer selects the new product over the old product, or over a competitor's product. Use the Probability Profiler to compare profiles to a baseline profile.

In this example, your company is currently producing laptops with 40 GB hard drives, 1.5 GHz processors, and 6-hour battery life, that cost $1,000. You are looking for a way to make your product more desirable by changing as few factors as possible. You set the current product configuration as the baseline. JMP adjusts the probabilities so that the probability of preference for the baseline configuration is 0.5. Then you compare the probabilities of other configurations to the baseline probability.

1. Do one of the following:
   – Follow the steps in “Complete the Launch Window” on page 146. Then complete step 1 and step 2 in “Analyze the Model” on page 148.
   – In the Laptop Runs. jmp sample data table, click the green triangle next to the **Choice Reduced Model** script.
2. Click the Choice Model red triangle and select **Probability Profiler**.

   Note that the Probability Profiler is for Gender = F. You can change this later.
3. Using the menus and text box below the profiler, in the Baseline area, specify the Baseline settings as 40 GB, 1.5 GHz, 6 hours, and 1000.
4. Now set these as the values in the Probability Profiler. To set the Price at $1000, click $1252 above Price under the rightmost profiler cell, and type 1000. Then click outside the text box.

Figure 7.29 Probability Profiler with Text Entry Area for Price

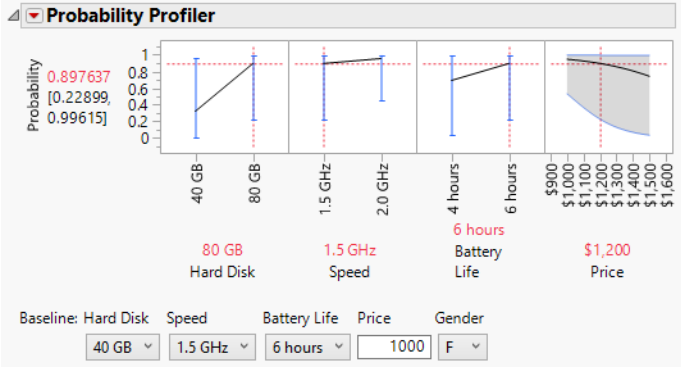


This configuration has probability 0.5.

5. In the Probability Profiler, move the slider for HardDisk to 80 GB.

   Notice that, with this change, the probability is relatively insensitive to increases in Price.
6. Click the $1000 label above the Price cell in the profiler, type **$1,200**, and click outside the text box.

Figure 7.30 Laptop Probability Profiler Results with Baseline Effects



An increase in Hard Disk size from 40 GB to 80 GB and an increase in price to $1200 coincides with an increased probability of preference, from 0.50 to 0.90 for females. Change the Gender effect in the Baseline to **M**. The probability of preference is 0.71.

Multiple Choice Comparisons

Use the Multiple Choice Profiler to compare product profiles.

- You currently produce a low-end laptop with a small hard drive, a slow processor, and low battery life. You charge $1000.
- Company A produces a product with a fast processor speed and high battery life at a reasonable price of $1200.
- Company B makes the biggest hard drives with the fastest speed, but at a high price of $1500 and low battery life.

You want to gain market share by increasing only one area of performance, and price.

1. Do one of the following:
   – Follow the steps in "Complete the Launch Window" on page 146. Then complete step 1 and step 2 in "Analyze the Model" on page 148.
   – In the Laptop Runs. jmp sample data table, click the green triangle next to the **Choice Reduced Model** script.
2. Click the Choice Model red triangle and select **Multiple Choice Profiler**.

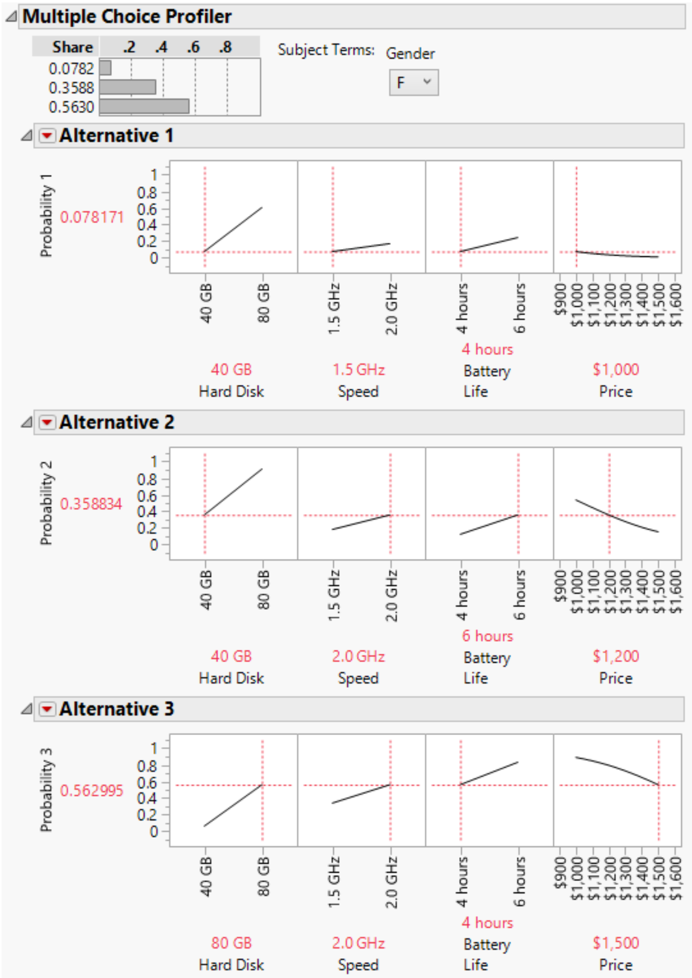
   A window appears, asking for the number of alternative choices to profile. Accept the default number of 3.
3. Click **OK**.

   Three Alternative profilers appear. Notice that the profilers are set for Gender = F.

Each factor in each profiler is set to its default values. Alternative 1 indicates the product that you want to develop. Alternative 2 indicates Company A's product. Alternative 3 indicates Company B's product.

4. For Alternative 1, set Hard Disk to 40 GB, Speed to 1.5 GHz, Battery Life to 4hours, and Price to $1,000.
5. For Alternative 2, set Hard Disk to 40 GB, Speed to 2.0 GHz, Battery Life to 6 hours, and Price to $1,200.
6. For Alternative 3, set Hard Disk to 80 GB, Speed to 2.0 GHz, Battery Life to 4 hours, and Price to $1,500.

Figure 7.31 Multiple Choice Profiler for Females

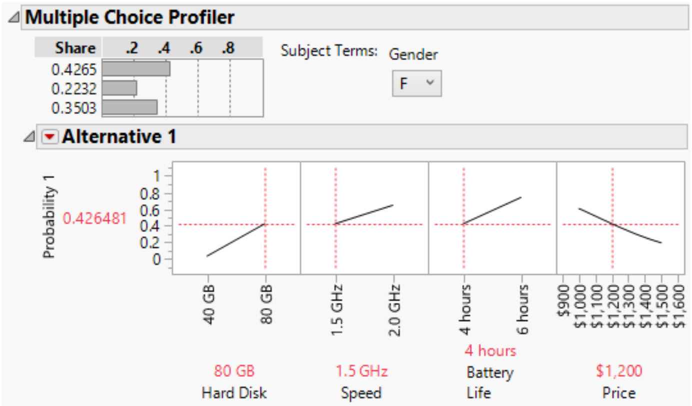


You can see that Company B has the greatest Share of 0.5630. It is obvious that with your company's settings, very few females buy your product.

You want to increase your market share by upgrading your company's laptop in one of the performance areas while increasing price. The slope of the line in Alternative 1's Hard Disk profile suggests increasing hard disk space increases market share the most.

7. For Alternative 1, set Hard Disk to **80 GB** and Price to **$1,200**.

Figure 7.32 Multiple Choice Profiler with Improved Laptop



By increasing hard disk space, you can increase the price of your laptop and expect a market share among females of about 43%. This share exceeds that of Company B's high-performance laptop and is much better than the market share with the initial low-end settings seen in Figure 7.31.

Explore the settings that increase your market share for males. If you increase both Hard Disk size and Speed, you can capture a 44% market share among males.

Example of Segmentation

In this example, you attempt to identify market segments for pizza preferences.

To see how to complete the launch window for this example, see step 1 to step 15 in the example "Multiple Table Format" on page 123. Otherwise, follow the instructions below.

Define Clusters

1. Select **Help > Sample Data Library** and open Pizza Responses.jmp.
2. Click the green triangle next to the **Choice** script.
3. Click the Choice Model red triangle and select **Save Gradients by Subject**.

   A data table appears with gradient forces saved for each main effect and subject interaction.

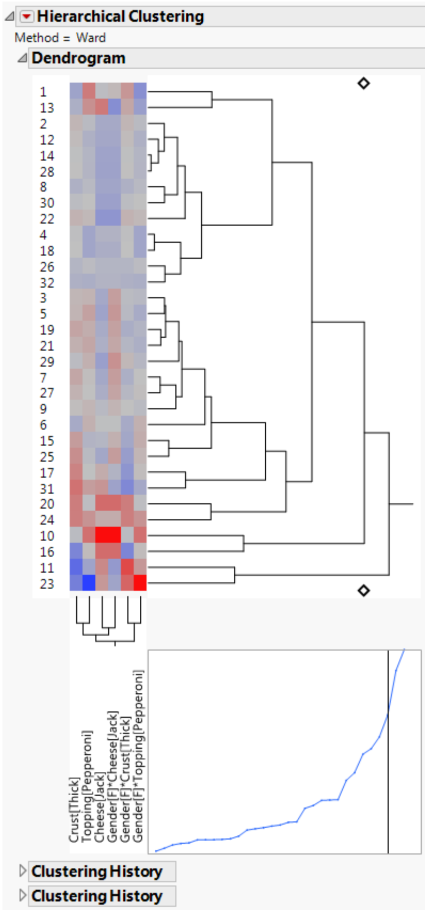
Figure 7.33 Gradients by Subject for Pizza Data, Partial View

Untitled
Hierarchical Cluster
Merge Data Back

	Subject	Crust[Thick]	Cheese[Jack]	Topping[Pepperoni]	Gender[F]*Crust [Thick]	Gender[F]*Cheese [Jack]	Gender[F]*Topping [Pepperoni]
1	1	-0.00959	-0.00168	0.014876	0.009585	0.001685	-0.01488
2	2	0.002373	-0.00758	-0.00239	0.002373	-0.00758	-0.00239
3	3	0.002129	-0.0079	0.003031	-0.00213	0.007899	-0.00303
4	4	-0.00106	-0.00485	-0.00901	-0.00106	-0.00485	-0.00901
5	5	0.002828	-0.00945	0.00725	-0.00283	0.009453	-0.00725
6	6	-0.0073	-0.00089	0.003761	-0.0073	-0.00089	0.003761

4. Click the green triangle next to the **Hierarchical Cluster** script.

Figure 7.34 Dendrogram of Subject Clusters for Pizza Data



The script runs a hierarchical cluster analysis on all columns in the gradient table, except for Subject. Click on either diamond to see that the rows have been placed into three clusters.

5. Click the red triangle next to Hierarchical Clustering and select **Save Clusters**.

A new column called Cluster is added to the data table containing the gradients. Each subject has been assigned a Cluster value that is associated with other subjects having similar gradient forces. See the Hierarchical Cluster platform chapter in the *Multivariate Methods* book for a discussion of other Hierarchical Clustering options.

You can delete the gradient columns because they were used only to obtain the clusters.

6. Select all columns except Subject and Cluster. Right-click on the selected columns and select **Delete Columns**.
7. Click the green triangle next to the **Merge Data Back** script(Figure 7.33).

   The cluster information is merged into the Subject data table. The columns in the Subject data table are now Subject, Gender, and Cluster, as shown in Figure 7.35.

Figure 7.35 Subject Data with Cluster Column

	Subject	Gender	Cluster
1	1	M	1
2	2	F	1
3	3	M	1
4	4	F	1
5	5	M	1
6	6	F	1

This table can now be used for further analysis.

Explore the Clusters

1. Click the icon to the left of the Cluster variable in the columns panel and select **Ordinal**.
2. Select **Analyze** > **Fit Y by X**.
3. Select Gender and click **Y, Response**.
4. Select Cluster and click **X, Factor**.
5. Click **OK**.

Figure 7.36 Contingency Analysis of Gender by Cluster

Contingency Analysis of Gender By Cluster

Mosaic Plot

Contingency Table

Gender

Cluster	Count / Total % / Col % / Row %	F	M	
1	Count	14	14	28
	Total %	43.75	43.75	87.50
	Col %	87.50	87.50	
	Row %	50.00	50.00	
2	Count	2	0	2
	Total %	6.25	0.00	6.25
	Col %	12.50	0.00	
	Row %	100.00	0.00	
3	Count	0	2	2
	Total %	0.00	6.25	6.25
	Col %	0.00	12.50	
	Row %	0.00	100.00	
		16	16	32
		50.00	50.00	

Tests

N	DF	-LogLike	RSquare (U)
32	2	2.7725887	0.1250

Test	ChiSquare	Prob>ChiSq
Likelihood Ratio	5.545	0.0625
Pearson	4.000	0.1353

Warning: 20% of cells have expected count less than 5, ChiSquare suspect.

You see the following:

– Cluster 1 is evenly divided between males and females

– Cluster 2 consists of only females

– Cluster 3 consists of only males

If desired, you could now refit and analyze the model with the addition of the Cluster variable.

Example of Logistic Regression Using the Choice Platform

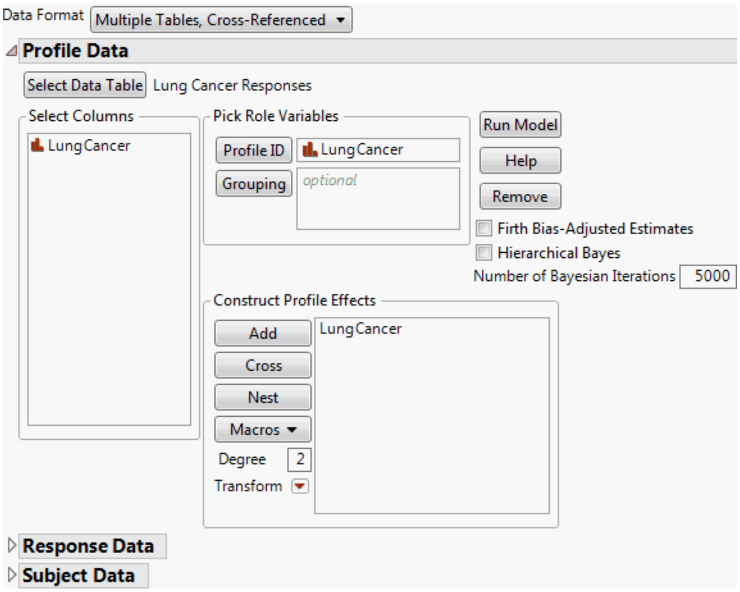
Use the Choice Platform

1. Select **Help > Sample Data Library** and open Lung Cancer Responses.jmp.

   Notice this data table has only one column (Lung Cancer) with two rows (Cancer and NoCancer).

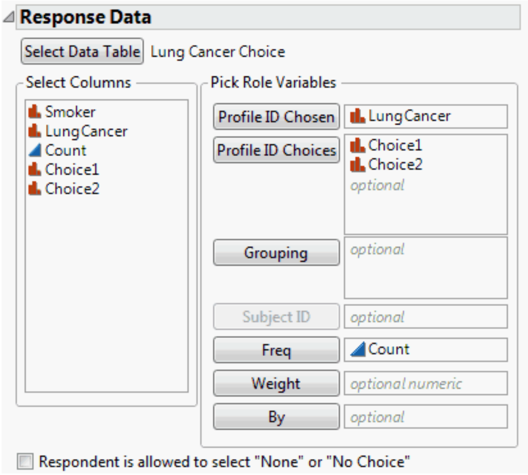
2. Select **Analyze** > **Consumer Research** > **Choice**
3. Select **Multiple Tables, Cross-Referenced** from the list next to Data Format.
4. Click **Select Data Table**, select Lung Cancer Responses.jmp,and click **OK**.
5. Select Lung Cancer and click **Profile** ID.
6. Select Lung Cancer and click **Add**.
7. Uncheck the Firth Bias-Adjusted Estimates box.

Figure 7.37 Completed Profile Data Panel



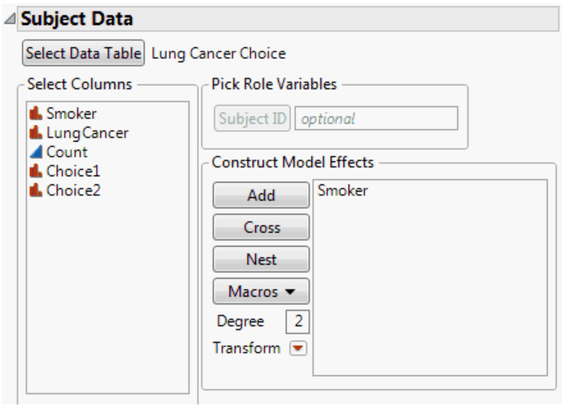
8. Open the **Response Data** outline.
9. Click **Select Data Table**, select Lung Cancer Choice.jmp, and click **OK**.
10. Do the following:
    – Select Lung Cancer and click **Profile ID Chosen**.
    – Select Choice1 and Choice2 and click **Profile ID Choices**.
    – Select Count and click **Freq**.

Figure 7.38 Completed Response Data Panel



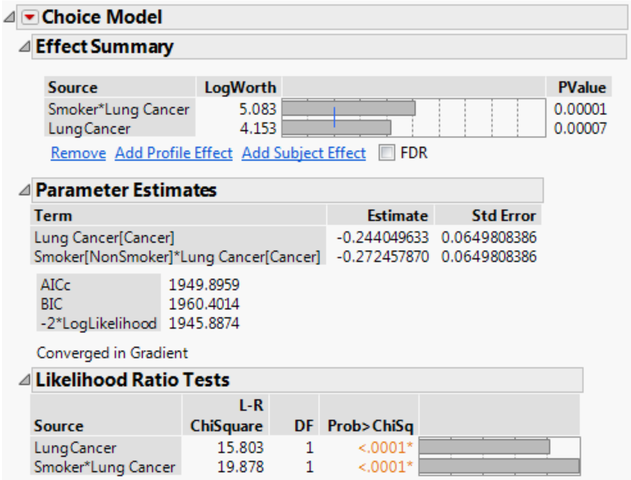
11. Open the **Subject Data** outline.
12. Click **Select Data Table**, select Lung Cancer Choice.jmp, and click **OK**.
13. Select Smoker and click **Add**.

Figure 7.39 Completed Subject Data Panel



14. Click **Run Model**.

Figure 7.40 Choice Modeling Logistic Regression Results


Choice Model

Effect Summary

Source	LogWorth	PValue
Smoker*Lung Cancer	5.083	0.00001
LungCancer	4.153	0.00007

Remove Add Profile Effect Add Subject Effect FDR

Parameter Estimates

Term	Estimate	Std Error
Lung Cancer[Cancer]	-0.244049633	0.0649808386
Smoker[NonSmoker]*Lung Cancer[Cancer]	-0.272457870	0.0649808386

AICc	1949.8959
BIC	1960.4014
-2*LogLikelihood	1945.8874

Converged in Gradient

Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
LungCancer	15.803	1	<.0001*
Smoker*Lung Cancer	19.878	1	<.0001*

Use the Fit Model Platform

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

   Because the data table contains a model script, the Model Specification window is automatically completed. The **Nominal Logistic** personality is selected.
3. Click **Run**.

Figure 7.41 Fit Model Nominal Logistic Regression Results

Nominal Logistic Fit for Lung Cancer

Effect Summary

Converged in Gradient, 4 iterations
Freq: Count

Iterations

Whole Model Test

Model	-LogLikelihood	DF	ChiSquare	Prob>ChiSq
Difference	9.93901	1	19.87802	<.0001*
Full	972.94369			
Reduced	982.88270			

RSquare (U)	0.0101
AICc	1949.9
BIC	1960.4
Observations (or Sum Wgts)	1418

Fit Details

Parameter Estimates

Term	Estimate	Std Error	ChiSquare	Prob>ChiSq
Intercept	-0.4880993	0.1299617	14.11	0.0002*
Smoker[NonSmoker]	-0.5449157	0.1299617	17.58	<.0001*

For log odds of Cancer/NoCancer

Covariance of Estimates

Effect Likelihood Ratio Tests

Source	Nparm	DF	L-R ChiSquare	Prob>ChiSq
Smoker	1	1	19.878016	<.0001*

Notice that the likelihood ratio chi-square test for Smoker*Lung Cancer in the Choice model matches the likelihood ratio chi-square test for Smoker in the Logistic model. The reports shown in Figure 7.40 and Figure 7.41 support the conclusion that smoking has a strong effect on developing lung cancer. See the Logistic Regression chapter in the *Fitting Linear Models* book for more details.

Example of Logistic Regression for Matched Case-Control Studies

This section provides an example using the Choice platform to perform logistic regression on the results of a study of endometrial cancer with 63 matched pairs. The data are from the Los Angeles Study of the Endometrial Cancer Data in Breslow and Day (1980) and the SAS/STAT(R) 9.2 User's Guide, Second Edition (2006). The goal of the case-control analysis was to determine the relative risk for gallbladder disease, controlling for the effect of hypertension. The Outcome of 1 indicates the presence of endometrial cancer, and 0 indicates the control. Gallbladder and Hypertension data indicators are also 0 or 1.

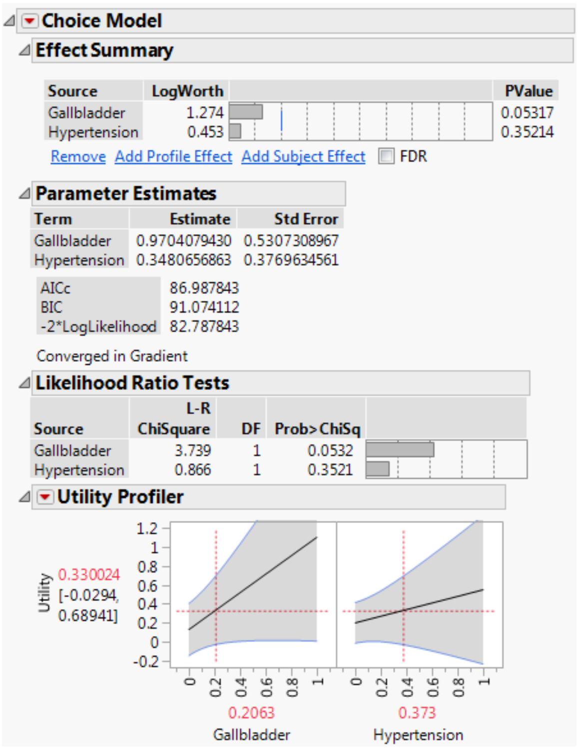
For details about performing logistic regression using the Choice platform, see "Logistic Regression" on page 171.

1. Select **Help > Sample Data Library** and open Endometrial Cancer.jmp.
2. Select **Analyze > Consumer Research > Choice**.
3. Check that the Data Format selected is **One-Table, Stacked**.
4. Click the **Select Data Table** button.

5. Select Endometrial Cancer as the profile data table. Click **OK**.
6. Select Outcome and click **Response Indicator**.
7. Select Pair and click **Grouping**.
8. Select Gallbladder and Hypertension and click **Add** in the Construct Profile Effects window.
9. Deselect the **Firth Bias-Adjusted Estimates** check box.
10. Click **Run Model**.
11. Click the Choice Model red triangle and select **Utility Profiler**.

    The report is shown in Figure 7.42.

Figure 7.42 Logistic Regression on Endometrial Cancer Data



Choice Model

Effect Summary

Source	LogWorth	PValue
Gallbladder	1.274	0.05317
Hypertension	0.453	0.35214

Remove Add Profile Effect Add Subject Effect FDR

Parameter Estimates

Term	Estimate	Std Error
Gallbladder	0.9704079430	0.5307308967
Hypertension	0.3480656863	0.3769634561

AICc	86.987843
BIC	91.074112
-2*LogLikelihood	82.787843

Converged in Gradient

Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Gallbladder	3.739	1	0.0532
Hypertension	0.866	1	0.3521

Utility Profiler

Likelihood Ratio tests are given for each factor. Note that Gallbladder is nearly significant at the 0.05 level (p-value = 0.0532). Use the Utility Profiler to visualize the impact of the factors on the response.

Example of Transforming Data to Two Analysis Tables

Consider the data from Daganzo, found in Daganzo Trip.jmp. This data set contains the travel time for three transportation alternatives and the preferred transportation alternative for each subject.

Add Choice Mode and Subjects

1. Select **Help > Sample Data Library** and open the Daganzo Trip.jmp data table.

   A partial listing of the data set is shown in Figure 7.43.

Figure 7.43 Partial Daganzo Travel Time Table for Three Alternatives

	Subway	Bus	Car	Choice
1	16.481	16.196	23.89	2
2	15.123	11.373	14.182	2
3	19.469	8.822	20.819	2
4	18.847	15.649	21.28	2
5	12.578	10.671	18.335	2

   Each Choice number listed must first be converted to one of the travel mode names. This transformation is easily done by using the **Choose** function in the formula editor, as follows.

2. Select **Cols > New Columns**.
3. Specify the Column Name as Choice Mode and the modeling type as **Nominal**.
4. Click the **Column Properties** and select **Formula**.
5. Click **Conditional** in the functions list, select **Choose**, and press the comma key twice to obtain additional arguments for the function.
6. Click Choice for the Choose expression (expr), and double click each clause entry box to enter "Subway", "Bus", and "Car" (with the quotation marks) as shown in Figure 7.44.

Figure 7.44 Choose Function for Choice Mode Column of Daganzo Data

```
Choose( Choice ) ( 1    ⇒ "Subway"
                   2    ⇒ "Bus"
                   else ⇒ "Car" )
```

7. Click **OK** in the Formula Editor window.
8. Click **OK** in the New Column window.

   The new Choice Mode column appears in the data table. Because each row contains a choice made by each subject, another column containing a sequence of numbers should be created to identify the subjects.

9. Select **Cols > New Columns**.
10. Specify the Column Name as Subject.
11. Click **Missing/Empty** next to Initialize Data and select **Sequence Data**.
12. Click **OK**.

    A partial listing of the modified table is shown in Figure 7.45.

Figure 7.45 Daganzo Data with New Choice Mode and Subject Columns

	Subway	Bus	Car	Choice	Choice Mode	Subject
1	16.481	16.196	23.89	2	Bus	1
2	15.123	11.373	14.182	2	Bus	2
3	19.469	8.822	20.819	2	Bus	3
4	18.847	15.649	21.28	2	Bus	4
5	12.578	10.671	18.335	2	Bus	5
6	11.513	20.582	27.838	1	Subway	6

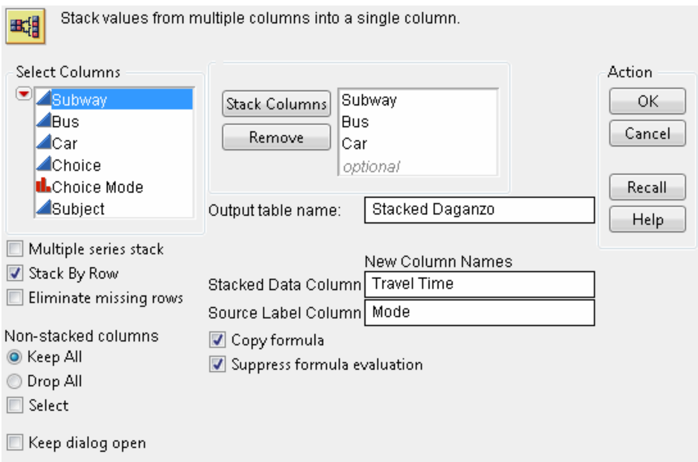
Stack the Data

In order to construct the Profile data, each alternative needs to be expressed in a separate row.

1. Select **Tables > Stack**.
2. Select Subway, Bus, and Car and click **Stack Columns**.
3. For the Output table name, type Stacked Daganzo. Type Travel Time for the Stacked Data Column and Mode for the Source Label Column.

   The resulting Stack window is shown in Figure 7.46.

Figure 7.46 Stack Operation for Daganzo Data



4. Click **OK**.

A partial view of the resulting table is shown in Figure 7.47.

Figure 7.47 Partial Stacked Daganzo Table

	Choice	Choice Mode	Subject	Mode	Travel Time
1	2	Bus	1	Subway	16.481
2	2	Bus	1	Bus	16.196
3	2	Bus	1	Car	23.89
4	2	Bus	2	Subway	15.123
5	2	Bus	2	Bus	11.373
6	2	Bus	2	Car	14.182
7	2	Bus	3	Subway	19.469

Make the Profile Data Table

For the Profile Data Table, you need the Subject, Mode, and Travel Time columns.

1. Select the Subject, Mode, and Travel Time columns and select **Tables** > **Subset**.
2. Select **All Rows** and **Selected Columns** and click **OK**.

   A partial data table is shown in Figure 7.48. Note the default table name is Subset of Stacked Daganzo.

Figure 7.48 Partial Subset Table of Stacked Daganzo Data

	Subject	Mode	Travel Time
1	1	Subway	16.481
2	1	Bus	16.196
3	1	Car	23.89
4	2	Subway	15.123
5	2	Bus	11.373
6	2	Car	14.182
7	3	Subway	19.469

Make the Response Data Table

For the Response Data Table, you need the Subject and Choice Mode columns, but you also need a column for each possible choice.

3. With the Daganzo Trip.jmp data table open, select the Subject and Choice Mode columns.
4. Select **Tables > Subset**.
5. Select **All Rows** and **Selected Columns** and click **OK**.

   Note that the default table name is Subset of Daganzo Trip.
6. Select **Cols > New Columns**.
7. For the Column prefix, type Choice.
8. Select **Character** and **Nominal**.
9. Type 3 next to Number of columns to add.

10. Click **OK**.

    The columns Choice 1, Choice 2, and Choice 3 have been added.

11. Type "Bus" (without quotation marks) in the first row of Choice 1. Right-click the cell and select **Fill > Fill to end of table**.
12. Type "Subway" (without quotation marks) in the first row of Choice 2. Right-click the cell and select **Fill > Fill to end of table**.
13. Type "Car" (without quotation marks) in the first row of Choice 3. Right-click the cell and select **Fill > Fill to end of table**.

    The resulting table is shown in Figure 7.49.

Figure 7.49 Partial Subset Table of Daganzo Data with Choice Set

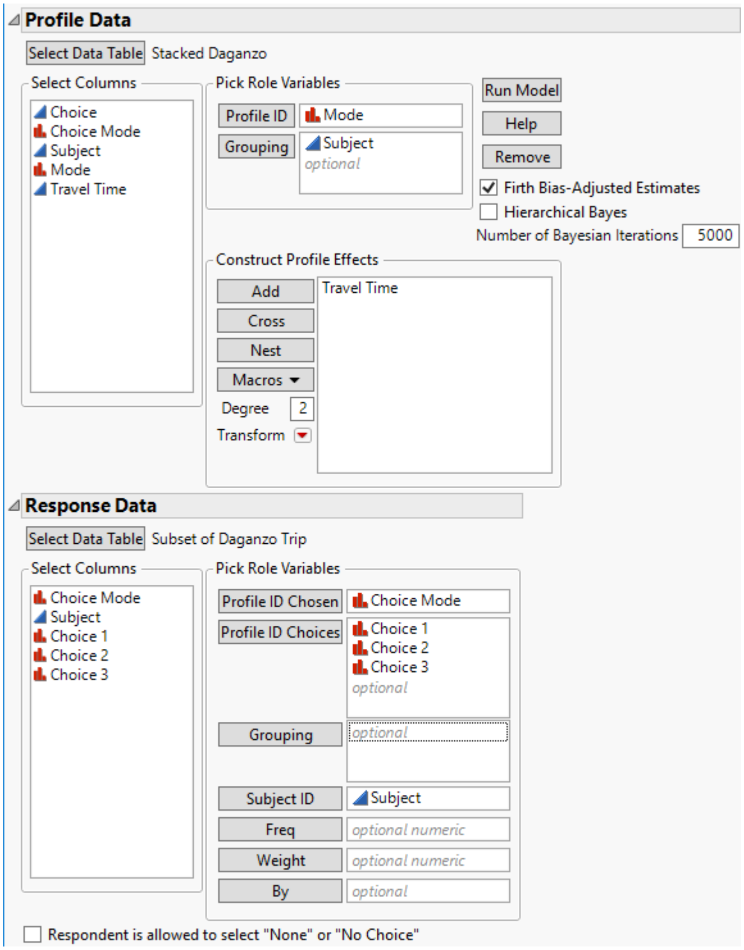
	Choice Mode	Subject	Choice 1	Choice 2	Choice 3
1	Bus	1	Bus	Subway	Car
2	Bus	2	Bus	Subway	Car
3	Bus	3	Bus	Subway	Car
4	Bus	4	Bus	Subway	Car
5	Bus	5	Bus	Subway	Car
6	Subway	6	Bus	Subway	Car
7	Subway	7	Bus	Subway	Car

Fit the Model

Now that you have separated the original Daganzo Trip.jmp table into two separate tables, you can run the Choice Platform.

1. Select **Analyze > Consumer Research > Choice**.
2. From the Data Format list, select **Multiple Tables, Cross-Referenced**.
3. Specify the model, as shown in Figure 7.50.

Figure 7.50 Choice Dialog Box for Subset of Daganzo Data



4. Click **Run Model**.

   The resulting parameter estimate now expresses the utility coefficient for Travel Time and is shown in Figure 7.51.

Figure 7.51 Parameter Estimate for Travel Time of Daganzo Data

Choice Model

Parameter Estimates

Term	Estimate	Std Error
Travel Time	-0.341768586	0.0745222259

AICc	68.766653
BIC	70.595342
-2*LogLikelihood	66.683319
-2*Firth LogLikelihood	61.490004

Converged in Gradient
Firth Bias-Adjusted Estimates

Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Travel Time	43.178	1	<.0001*

The negative coefficient implies that increased travel time has a negative effect on consumer utility or satisfaction. The likelihood ratio test result indicates that the Choice model with the effect of Travel Time is significant.

Example of Transforming Data to One Analysis Table

Rather than creating two or three tables, it can be more practical to transform the data so that only one table is used. For the one-table format, the subject effect is added as in the previous example. A response indicator column is added instead of using three different columns for the choice sets (Choice 1, Choice 2, Choice 3). The transformation for the one-table scenario includes the following steps.

1. Create or open Stacked Daganzo.jmp from the “Stack the Data” steps shown in “Example of Transforming Data to Two Analysis Tables” on page 163.
2. Select **Cols > New Columns**.
3. Type Response as the Column Name.
4. Click **Column Properties** and select **Formula**.
5. Select **Conditional** in the functions list and then select **If**.
6. Select the column Choice Mode for the expression (expr).
7. Enter “=” and select Mode.
8. Type 1 for the **Then Clause** and 0 for the **Else Clause**.
9. Click **OK** in the Formula Editor window. Click **OK** in the New Column window.

   The completed formula should look like Figure 7.52.

Figure 7.52 Formula for Response Indicator for Stacked Daganzo Data

If	*Choice Mode* == *Mode*	⇒ 1
	else	⇒ 0

10. Select the **Subject**, **Travel Time**, and **Response** columns and then select **Tables** > **Subset**.
11. Select **All Rows** and **Selected Columns** and click **OK**.

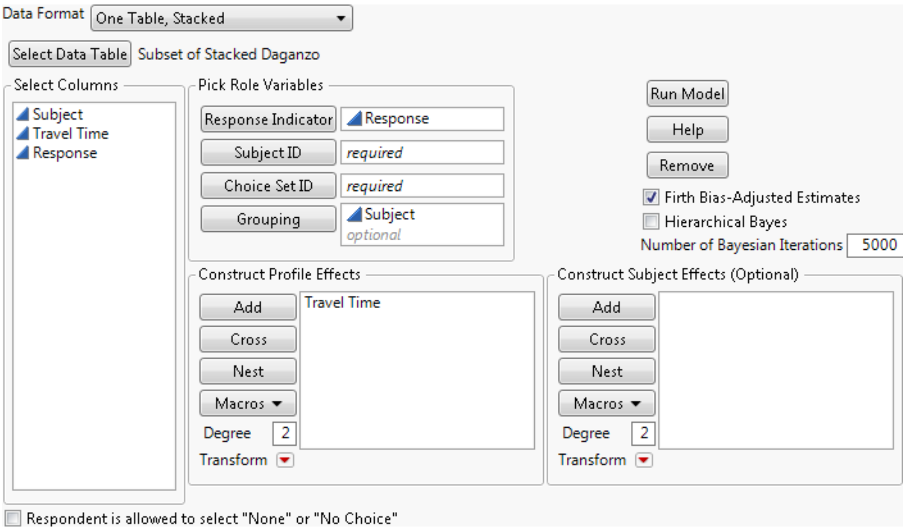
    A partial listing of the new data table is shown in Figure 7.53.

Figure 7.53 Partial Table of Stacked Daganzo Data Subset

	Subject	Travel Time	Response
1	1	16.481	0
2	1	16.196	1
3	1	23.89	0
4	2	15.123	0
5	2	11.373	1
6	2	14.182	0
7	3	19.469	0

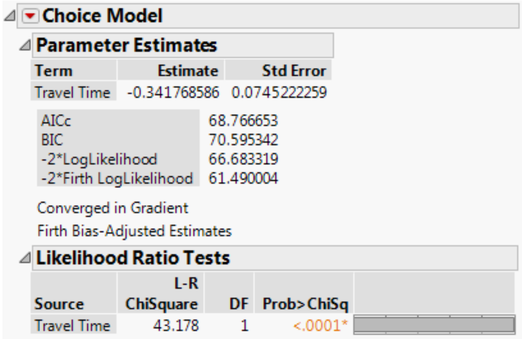
12. Select **Analyze** > **Consumer Research** > **Choice** to open the launch window and specify the model as shown in Figure 7.54.

Figure 7.54 Choice Dialog Box for Subset of Stacked Daganzo Data for One-Table Analysis



13. Click **Run Model.**

Figure 7.55 Parameter Estimate for Travel Time of Daganzo Data from One-Table Analysis



Choice Model

Parameter Estimates

Term	Estimate	Std Error
Travel Time	-0.341768586	0.0745222259

AICc	68.766653
BIC	70.595342
-2*LogLikelihood	66.683319
-2*Firth LogLikelihood	61.490004

Converged in Gradient
Firth Bias-Adjusted Estimates

Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Travel Time	43.178	1	<.0001*

Notice that the result is identical to that obtained for the two-table model, shown earlier in Figure 7.51.

This chapter illustrates the use of the Choice Modeling platform with simple examples. This platform can also be used for more complex models, such as those involving more complicated transformations and interaction terms.

Technical Details

This section contains information about the following topics:

- "Special Data Table Rules" on page 170
- "Utility and Probabilities" on page 171
- "Gradients" on page 172

Special Data Table Rules

Default Choice Set

If in every trial, you can choose any of the response profiles, you can omit the **Profile ID Choices** selection under **Pick Role Variables** in the Response Data section of the Choice launch window. The Choice Model platform then assumes that all choice profiles are available on each run.

Subject Data with Response Data

If you have subject data in the Response data table, select this table as the **Select Data Table** under the Subject Data. In this case, a **Subject ID** column does not need to be specified. In fact, it is not used. It is generally assumed that the subject data repeats consistently in multiple runs for each subject.

Logistic Regression

Ordinary logistic regression can be performed with the Choice Modeling platform.

Note: The Fit Y by X and Fit Model platforms are more convenient to use than the Choice Modeling platform for logistic regression modeling. This section is used only to demonstrate that the Choice Modeling platform can be used for logistic regression, if desired.

If your data are already in the choice-model format, you might want to use the steps given below for logistic regression analysis. However, three steps are needed:

- Create a trivial Profile data table with a row for each response level.
- Put the explanatory variables into the Response data.
- Specify the Response data table, again, for the Subject data table.

For examples of conducting Logistic Regression using the Choice Platform, see "Example of Logistic Regression Using the Choice Platform" on page 158 and "Example of Logistic Regression for Matched Case-Control Studies" on page 161.

Utility and Probabilities

Parameter estimates from the choice model identify consumer *utility*, or marginal utilities in the case of a linear utility function. Utility is the level of satisfaction consumers receive from products with specific attributes and is determined from the parameter estimates in the model.

The choice statistical model is expressed as follows:

Let $X[k]$ represent a subject attribute design row, with intercept

Let $Z[j]$ represent a choice attribute design row, without intercept

Then, the probability of a given choice for the $k'th$ subject to the $j'th$ choice of m choices is:

$$P_i[jk] = \frac{\exp(\beta'(X[k] \otimes Z[j]))}{\sum_{l=1}^{m} \exp(\beta'(X[k] \otimes Z[l]))}$$

where:

- $\otimes$ is the Kronecker rowwise product
- the numerator calculates for the $j'th$ alternative actually chosen
- the denominator sums over the m choices presented to the subject for that trial

Gradients

The gradient values that you obtain when you select the Save Gradients by Subject option are the subject-aggregated Newton-Raphson steps from the optimization used to produce the estimates. At the estimates, the total gradient is zero, and $\Delta = H^{-1}g = 0$, where g is the total gradient of the log-likelihood evaluated at the MLE, and H^{-1} is the inverse Hessian function or the inverse of the negative of the second partial derivative of the log-likelihood.

But, the disaggregation of Δ results in the following:

$$\Delta = \Sigma_{ij}\Delta_{ij} = \Sigma H^{-1}g_{ij} = 0,$$

Here i is the subject index, j is the choice response index for each subject, Δij are the partial Newton-Raphson steps for each run, and g_{ij} is the gradient of the log-likelihood by run.

The mean gradient step for each subject is then calculated as follows:

$$\bar{\Delta}_i = \Sigma_j \frac{\Delta_{ij}}{n_i},$$

where n_i is the number of runs per subject. These $\bar{\Delta}_i$ are related to the force that subject i is applying to the parameters. If groups of subjects have truly different preference structures, these forces are strong, and they can be used to cluster the subjects. The $\bar{\Delta}_i$ are the gradient forces that are saved. You can then cluster these values using the Clustering platform.

Chapter 8

MaxDiff

Use MaxDiff (maximum difference scaling) as an alternative to standard preference scales to determine the relative importance of items being rated. MaxDiff forces respondents to report their most and least preferred options. This often results in rankings that are more definitive than rankings obtained using standard preference scales.

The MaxDiff platform enables you to do the following:

- Use information about subject traits as well as product attributes.
- Integrate data from one, two, or three sources.
- Obtain subject-level scores for segmenting or clustering your data.
- JMP PRO Estimate subject-specific coefficients using a Bayesian approach.
- Use bias-corrected maximum likelihood estimators (Firth, 1993).

Figure 8.1 MaxDiff All Comparisons Report

All Levels Comparison Report

Difference (Row-Column) Standard Error of Difference Wald p-Value	Gyro	Sour Cream and Onion	Truffle Fries	Biscuits and Gravy	Reuben	Ketchup	All Dressed	Dill Pickle	Barbecue	Southern Barbecue
Gyro	0	-1.1747	-1.3082	-1.2688	-0.6992	-0.8229	-0.9467	-1.027	-2.4417	-1.9535
	0	0.33327	0.35477	0.34456	0.34758	0.32802	0.33571	0.35251	0.41499	0.38939
		0.00062	0.00035	0.00036	0.04668	0.01356	0.00569	0.00432	4.32e-8	2.01e-6
Sour Cream and Onion	1.17469	0	-0.1335	-0.0941	0.47548	0.35175	0.22804	0.14766	-1.267	-0.7788
	0.33327	0	0.31379	0.30716	0.30932	0.31915	0.28158	0.30707	0.37654	0.34896
	0.00062		0.67137	0.75983	0.12709	0.27279	0.41975	0.63156	0.00105	0.02764
Truffle Fries	1.30817	0.13348	0	0.03935	0.60896	0.48523	0.36152	0.28114	-1.1335	-0.6453
	0.35477	0.31379	0	0.31345	0.31817	0.31157	0.31136	0.30871	0.39213	0.34904
	0.00035	0.67137		0.90032	0.05821	0.12223	0.24809	0.36444	0.00462	0.06715
Biscuits and Gravy	1.26882	0.09413	-0.0393	0	0.56961	0.44588	0.32217	0.24179	-1.1729	-0.6847
	0.34456	0.30716	0.31345	0	0.30639	0.29752	0.29974	0.28249	0.36267	0.34038
	0.00036	0.75983	0.90032		0.06567	0.1368	0.28477	0.39388	0.00161	0.0467
Reuben	0.69921	-0.4755	-0.609	-0.5696	0	-0.1237	-0.2474	-0.3278	-1.7425	-1.2543
	0.34758	0.30932	0.31817	0.30639	0	0.30893	0.30148	0.30328	0.37117	0.35324
	0.04668	0.12709	0.05821	0.06567		0.68956	0.41355	0.28208	7.69e-6	0.00056
Ketchup	0.82294	-0.3517	-0.4852	-0.4459	0.12373	0	-0.1237	-0.2041	-1.6188	-1.1305
	0.32802	0.31915	0.31157	0.29752	0.30893	0	0.29728	0.27938	0.36911	0.34808
	0.01356	0.27279	0.12223	0.1368	0.68956		0.67812	0.46661	2.64e-5	0.00154
All Dressed	0.94665	-0.228	-0.3615	-0.3222	0.24744	0.12371	0	-0.0804	-1.4951	-1.0068
	0.33571	0.28158	0.31136	0.29974	0.30148	0.29728	0	0.29447	0.36838	0.3418
	0.00569	0.41975	0.24809	0.28477	0.41355	0.67812		0.78538	9.22e-5	0.00393
Dill Pickle	1.02703	-0.1477	-0.2811	-0.2418	0.32782	0.20409	0.08038	0	-1.4147	-0.9264
	0.35251	0.30707	0.30871	0.28249	0.30328	0.27938	0.29447	0	0.3539	0.35532
	0.00432	0.63156	0.36444	0.39388	0.28208	0.46661	0.78538		0.00012	0.01038
Barbecue	2.44172	1.26703	1.13355	1.1729	1.7425	1.61877	1.49507	1.41468	0	0.48824
	0.41499	0.37654	0.39213	0.36267	0.37117	0.36911	0.36838	0.3539	0	0.39164
	4.32e-8	0.00105	0.00462	0.00161	7.69e-6	2.64e-5	9.22e-5	0.00012		0.21514
Southern Barbecue	1.95348	0.77879	0.6453	0.68465	1.25426	1.13053	1.00683	0.92644	-0.4882	0
	0.38939	0.34896	0.34904	0.34038	0.35324	0.34808	0.3418	0.35532	0.39164	0
	2.01e-6	0.02764	0.06715	0.0467	0.00056	0.00154	0.00393	0.01038	0.21514	

Contents

Choice Modeling Platform Overview

MaxDiff, also known as *best-worst scaling* (BWS), is a choice-based measurement method. Rather than asking a respondent to report one favorite choice among several alternative profiles, MaxDiff asks a respondent to report both a *best* and a *worst* choice. The MaxDiff approach can provide more information about preferences than an approach where a respondent reports only a favorite choice. For background on MaxDiff studies, see Louviere et al. (2015). For background on choice modeling, see Louviere et al. (2015), Train (2009), and Rossi et al. (2006).

MaxDiff analysis uses the framework of random utility theory. A choice is assumed to have an underlying value, or *utility*, to respondents. The MaxDiff platform estimates these utilities. The MaxDiff platform also estimates the probabilities that a choice is preferred over other choices. This is done using conditional logistic regression. See McFadden (1974).

Note: One-factor MaxDiff studies can be designed using the MaxDiff Design platform. See the MaxDiff Design chapter in the *Design of Experiments Guide*.

Segmentation and Bayesian Subject-Level Effects

Market researchers sometimes want to analyze the preference structure for each subject separately in order to see whether there are groups of subjects that behave differently. If there are sufficient data, you can specify “By groups” in the Response Data or you could introduce a Subject identifier as a subject-side model term. This approach, however, is costly if the number of subjects is large. Other segmentation techniques discussed in the literature include Bayesian and mixture methods.

If there are not sufficient data to specify “By groups,” you can segment in JMP by clustering subjects using response data and the **Save Gradients by Subject** option. The option creates a new data table containing the average Hessian-scaled gradient on each parameter for each subject. For an example, see “Example of Segmentation” on page 154 in the “Choice Models” chapter. For details about the gradient values, see “Gradients” on page 172 in the “Choice Models” chapter.

MaxDiff also provides a Hierarchical Bayesian approach to estimating subject-level effects. This approach can be useful in market segmentation.

Examples of the MaxDiff Platform

Thirty respondents participated in a MaxDiff study to compare seven flavors of potato chips. Each choice set consisted of three profiles (potato chip flavors). For each choice set, a respondent’s favorite choice was recorded as 1 and his or her least favorite choice was recorded as -1. Intermediate choices were recorded as 0.

The MaxDiff platform can analyze data that is presented in a one-table format or in a multiple-table format. In the multiple table format, information about responses, choice sets, and subjects is saved in different data tables. In the one-table format, that information is contained in a single data table.

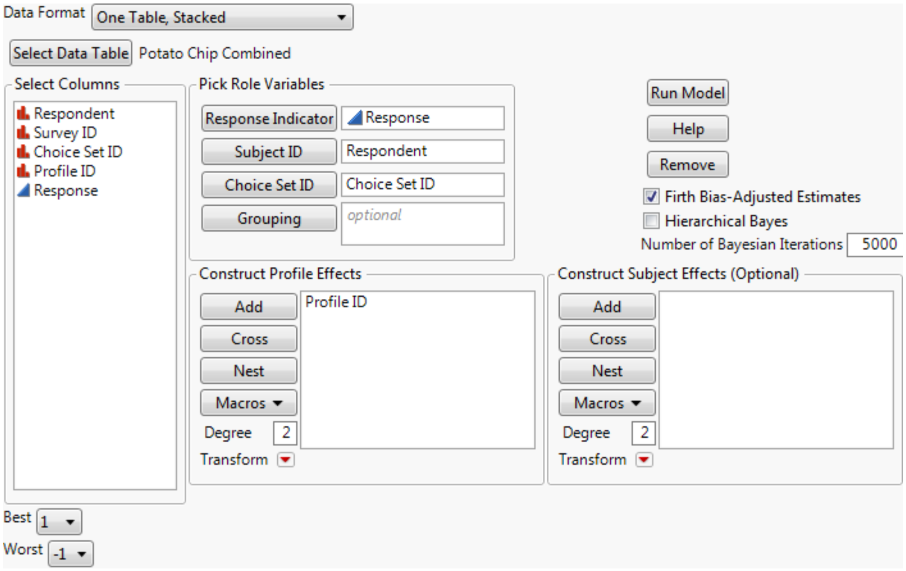
- "One Table Format" on page 176 shows how to analyze a subset of the available data in a one-table format. Note that you could add additional profile and subject data to the single table for a more complete analysis.
- "Multiple Table Format" on page 178 shows how to bring together information from different tables into one MaxDiff analysis.

One Table Format

1. Select **Help > Sample Data Library** and open Potato Chip Combined.jmp.
2. Select **Analyze > Consumer Research > MaxDiff**.

   Note that the default Data Format is set to One Table, Stacked.
3. Click **Select Data Table**.
4. Select Potato Chip Combined and click **OK**.
5. Assign roles to columns as follows. The completed launch dialog is shown in Figure 8.2.
   – Select Response and click **Response Indicator**.
   – Select Respondent and click **Subject ID**.
   – Select Choice Set ID and click **Choice Set ID**.
   – Select ProfileID and click **Add** in the Construct Profile Effects panel.

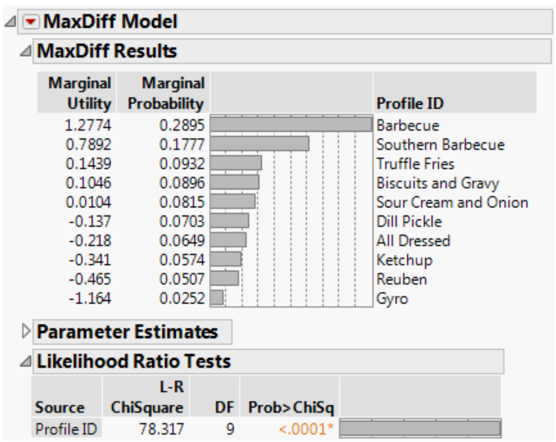
Figure 8.2 Completed MaxDiff Launch Window



Because you designated the Best choice as 1 and the Worst choice as -1, you make no change to the Best and Worst choice indicators at the bottom left of the launch window.

6. Click **Run Model**.

Figure 8.3 MaxDiff Report for Potato Chip Combined.jmp



The report indicates that **Profile ID** is significant, indicating that preferences for the various chip types differ significantly. The highest Marginal Utility is for Barbecue chips. The estimated probability that Barbecue chips are preferred to other chip types is 0.2895.

7. Click the red triangle next to MaxDiff Model and select **All Levels Comparison Report.**

Figure 8.4 All Comparisons Report

All Levels Comparison Report

Difference (Row-Column) Standard Error of Difference Wald p-Value	Gyro	Sour Cream and Onion	Truffle Fries	Biscuits and Gravy	Reuben	Ketchup	All Dressed	Dill Pickle	Barbecue	Southern Barbecue
Gyro	0	-1.1747	-1.3082	-1.2688	-0.6992	-0.8229	-0.9467	-1.027	-2.4417	-1.9535
	0	0.33327	0.35477	0.34456	0.34758	0.32802	0.33571	0.35251	0.41499	0.38939
		0.00062	0.00035	0.00036	0.04668	0.01356	0.00569	0.00432	4.32e-8	2.01e-6
Sour Cream and Onion	1.17469	0	-0.1335	-0.0941	0.47548	0.35175	0.22804	0.14766	-1.267	-0.7788
	0.33327	0	0.31379	0.30716	0.30932	0.31915	0.28158	0.30707	0.37654	0.34896
	0.00062		0.67137	0.75983	0.12709	0.27279	0.41975	0.63156	0.00105	0.02764
Truffle Fries	1.30817	0.13348	0	0.03935	0.60896	0.48523	0.36152	0.28114	-1.1335	-0.6453
	0.35477	0.31379	0	0.31345	0.31817	0.31157	0.31136	0.30871	0.39213	0.34904
	0.00035	0.67137		0.90032	0.05821	0.12223	0.24809	0.36444	0.00462	0.06715
Biscuits and Gravy	1.26882	0.09413	-0.0393	0	0.56961	0.44588	0.32217	0.24179	-1.1729	-0.6847
	0.34456	0.30716	0.31345	0	0.30639	0.29752	0.29974	0.28249	0.36267	0.34038
	0.00036	0.75983	0.90032		0.06567	0.1368	0.28477	0.39388	0.00161	0.0467
Reuben	0.69921	-0.4755	-0.609	-0.5696	0	-0.1237	-0.2474	-0.3278	-1.7425	-1.2543
	0.34758	0.30932	0.31817	0.30639	0	0.30893	0.30148	0.30328	0.37117	0.35324
	0.04668	0.12709	0.05821	0.06567		0.68956	0.41355	0.28208	7.69e-6	0.00056
Ketchup	0.82294	-0.3517	-0.4852	-0.4459	0.12373	0	-0.1237	-0.2041	-1.6188	-1.1305
	0.32802	0.31915	0.31157	0.29752	0.30893	0	0.29728	0.27938	0.36911	0.34808
	0.01356	0.27279	0.12223	0.1368	0.68956		0.67812	0.46661	2.64e-5	0.00154
All Dressed	0.94665	-0.228	-0.3615	-0.3222	0.24744	0.12371	0	-0.0804	-1.4951	-1.0068
	0.33571	0.28158	0.31136	0.29974	0.30148	0.29728	0	0.29447	0.36838	0.3418
	0.00569	0.41975	0.24809	0.28477	0.41355	0.67812		0.78538	9.22e-5	0.00393
Dill Pickle	1.02703	-0.1477	-0.2811	-0.2418	0.32782	0.20409	0.08038	0	-1.4147	-0.9264
	0.35251	0.30707	0.30871	0.28249	0.30328	0.27938	0.29447	0	0.3539	0.35532
	0.00432	0.63156	0.36444	0.39388	0.28208	0.46661	0.78538		0.00012	0.01038
Barbecue	2.44172	1.26703	1.13355	1.1729	1.7425	1.61877	1.49507	1.41468	0	0.48824
	0.41499	0.37654	0.39213	0.36267	0.37117	0.36911	0.36838	0.3539	0	0.39164
	4.32e-8	0.00105	0.00462	0.00161	7.69e-6	2.64e-5	9.22e-5	0.00012		0.21514
Southern Barbecue	1.95348	0.77879	0.6453	0.68465	1.25426	1.13053	1.00683	0.92644	-0.4882	0
	0.38939	0.34896	0.34904	0.34038	0.35324	0.34808	0.3418	0.35532	0.39164	0
	2.01e-6	0.02764	0.06715	0.0467	0.00056	0.00154	0.00393	0.01038	0.21514	

Each comparison is the difference in estimated utilities between the chip type labeling the row and the chip type labeling the column. Small *p*-values are colored with an intense blue or red color, depending on the sign of the difference. For example, based on the blue colors across the Gyro row, you can see that Gyro chips have significantly lower utility than all other chip types. Barbecue chips have higher utility than all other chip types, though they do not differ significantly from Southern Barbecue chips.

Note: Because the All Comparisons Report *p*-values are not corrected for multiple comparisons, use them as a guide.

Multiple Table Format

This version of the potato chip study uses three data tables: Potato Chip Profiles.jmp, Potato Chip Responses.jmp, and Potato Chip Subjects.jmp. Although you can always arrange your data into a single table, a multi-table approach can be more convenient than a one-table analysis when you have additional profile and subject variables that you want to include in your analysis.

Complete the Launch Window

1. Select **Help > Sample Data Library** and open the Potato Chip Responses.jmp sample data table.

Note: If you prefer not to follow the steps for completing the launch window, click the green triangle next to the **MaxDiff for Flavor** script. Then proceed to "Explore the Model" on page 181.

2. Click the green triangle next to the **Open Profile and Subject Tables** script.
   – The profile data table, Potato Chip Profiles.jmp, lists all the potato chip types in the study (Flavor) along with information on the country of origin (Product Of). Each choice has a Profile ID.
   – The subjects data table, Potato Chip Subjects.jmp, lists the respondents. It also gives additional information about each respondent: Citizenship and Gender.
   – The responses data table, Potato Chip Responses.jmp, lists the respondents. For each respondent, the Survey ID and Choice Set ID for each set of profiles is listed, along with the Profile ID values for each choice set. The table also contains response data in the Best Profile and Worst Profile columns.
3. From any of the three data tables, select **Analyze > Consumer Research > MaxDiff**.
4. From the Data Format list, select Multiple Tables, Cross-referenced.

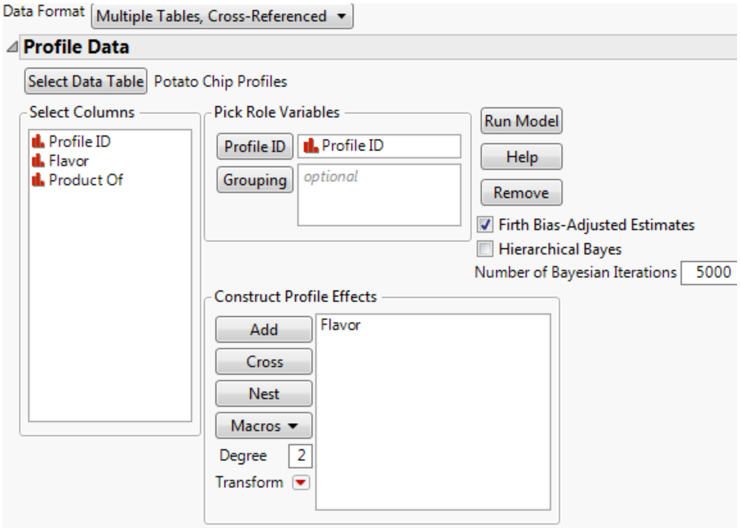
   There are three separate outlines, one for each of the data sources.
5. Click **Select Data Table** under Profile Data.

   A Profile Data Table window appears, which prompts you to specify the data table for the profile data.
6. Select Potato Chip Profiles.jmp and click **OK**.

   The columns from this table appear in the **Select Columns**.
7. Select Profile ID from the Select Columns list and click **Profile ID** under **Pick Role Variables**.
8. Select Flavor and click **Add** under **Construct Model Effects.**

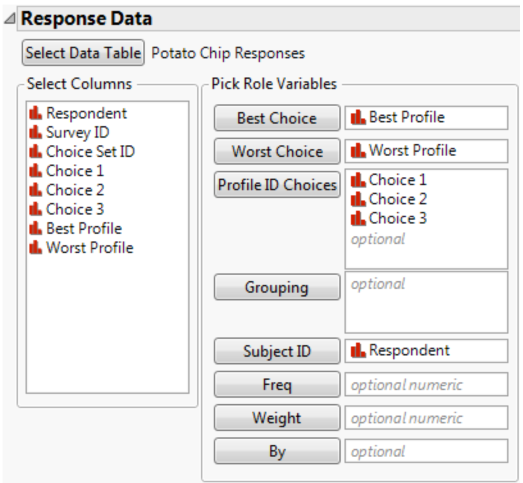
   Note that Product Of is another profile effect that you could add to the effects list.

Figure 8.5 Complete Profile Data Outline



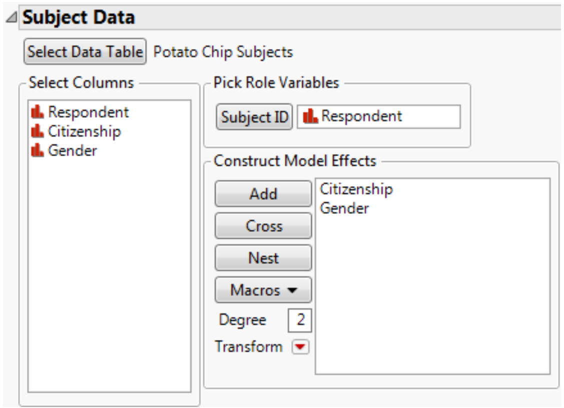
9. Open the Response Data outline. Click **Select Data Table**.
10. Select Potato Chip Responses.jmp and click **OK**.
11. Assign roles to columns as follows. The completed launch dialog is shown in Figure 8.6.
    – Select Best Profile and click **Best Choice**.
    – Select Worst Profile and click **Worst Choice**.
    – Select Choice 1, Choice 2, and Choice 3 and click **Profile ID Choices**.
    – Select Respondent and click **Subject ID**.

Figure 8.6 Completed Response Data Outline



12. Open the Subject Data outline. Click **Select Data Table**.
13. Select Potato Chip Subjects.jmp and click **OK**.
14. Select Respondent and click **Subject ID**.
15. Select Citizenship and Gender and click **Add** under **Construct Model Effects**.

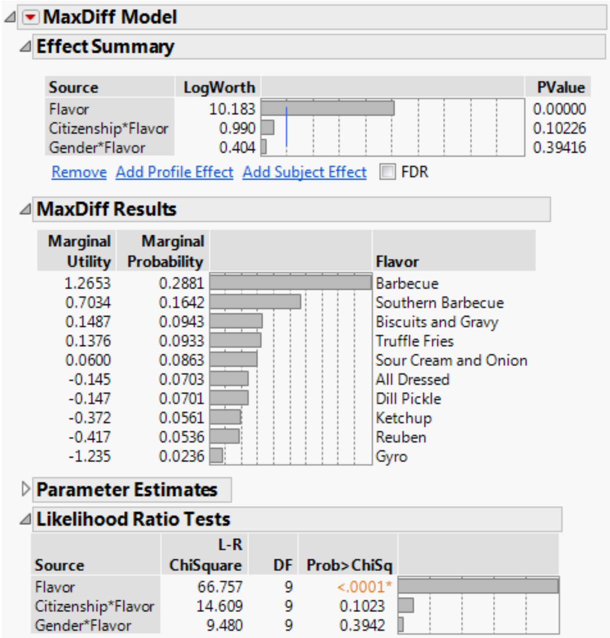
Figure 8.7 Completed Subject Data Outline



Explore the Model

1. Click **Run Model**.

Figure 8.8 MaxDiff Model Report



MaxDiff Model

Effect Summary

Source	LogWorth	PValue
Flavor	10.183	0.00000
Citizenship*Flavor	0.990	0.10226
Gender*Flavor	0.404	0.39416

Remove Add Profile Effect Add Subject Effect FDR

MaxDiff Results

Marginal Utility	Marginal Probability	Flavor
1.2653	0.2881	Barbecue
0.7034	0.1642	Southern Barbecue
0.1487	0.0943	Biscuits and Gravy
0.1376	0.0933	Truffle Fries
0.0600	0.0863	Sour Cream and Onion
-0.145	0.0703	All Dressed
-0.147	0.0701	Dill Pickle
-0.372	0.0561	Ketchup
-0.417	0.0536	Reuben
-1.235	0.0236	Gyro

Parameter Estimates

Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Flavor	66.757	9	<.0001*
Citizenship*Flavor	14.609	9	0.1023
Gender*Flavor	9.480	9	0.3942

The Effect Summary report in Figure 8.8 shows the terms in the model and gives *p*-values for their significance. Notice that Flavor is a profile effect, and that each of Citizenship*Flavor and Gender*Flavor is an interaction of a subject and a profile effect.

The Likelihood Ratio Tests report indicates that Flavor is significant.

2. Click the red triangle next to MaxDiff Model and select **All Levels Comparison Report.**

   Because you have included subject effects, you must specify the subject values for which you want to make comparisons.

3. From the menu next to Citizenship, select USA.
4. From the menu next to Gender, select Male.
5. Click **OK**.

Figure 8.9 All Comparisons Report

All Levels Comparison Report

Comparisons for: Citizenship=USA,Gender=Male

Difference (Row-Column) Standard Error of Difference Wald p-Value	All Dressed	Barbecue	Biscuits and Gravy	Dill Pickle	Gyro	Ketchup	Reuben	Sour Cream and Onion	Southern Barbecue	Truffle Fries
All Dressed	0	-0.828	0.14227	0.11778	1.30353	0.17663	0.52737	0.78624	-0.4758	-0.1792
	0	0.53818	0.45162	0.4627	0.52473	0.46308	0.46179	0.4522	0.5175	0.47513
		0.12676	0.75333	0.79954	0.01448	0.70362	0.25591	0.08486	0.35983	0.70675
Barbecue	0.82802	0	0.97029	0.9458	2.13155	1.00465	1.35539	1.61425	0.35217	0.6488
	0.53818	0	0.52699	0.51562	0.5951	0.51733	0.53789	0.55	0.55347	0.57013
	0.12676		0.06827	0.06929	0.00051	0.05467	0.01316	0.00405	0.52589	0.25758
Biscuits and Gravy	-0.1423	-0.9703	0	-0.0245	1.16126	0.03436	0.3851	0.64396	-0.6181	-0.3215
	0.45162	0.52699	0	0.43018	0.51066	0.44934	0.46776	0.46682	0.49636	0.46888
	0.75333	0.06827		0.9547	0.02489	0.93919	0.41212	0.17052	0.21564	0.49436
Dill Pickle	-0.1178	-0.9458	0.02449	0	1.18575	0.05885	0.40959	0.66845	-0.5936	-0.297
	0.4627	0.51562	0.43018	0	0.53387	0.42095	0.47062	0.47541	0.53139	0.46318
	0.79954	0.06929	0.9547		0.02838	0.88908	0.38601	0.1625	0.26635	0.5227
Gyro	-1.3035	-2.1316	-1.1613	-1.1858	0	-1.1269	-0.7762	-0.5173	-1.7794	-1.4827
	0.52473	0.5951	0.51066	0.53387	0	0.49057	0.51634	0.50769	0.57491	0.54826
	0.01448	0.00051	0.02489	0.02838		0.02349	0.13563	0.31045	0.00249	0.00792
Ketchup	-0.1766	-1.0046	-0.0344	-0.0588	1.1269	0	0.35074	0.60961	-0.6525	-0.3558
	0.46308	0.51733	0.44934	0.42095	0.49057	0	0.45998	0.48847	0.5162	0.4642
	0.70362	0.05467	0.93919	0.88908	0.02349		0.44737	0.21466	0.20888	0.44497
Reuben	-0.5274	-1.3554	-0.3851	-0.4096	0.77617	-0.3507	0	0.25887	-1.0032	-0.7066
	0.46179	0.53789	0.46776	0.47062	0.51634	0.45998	0	0.46668	0.53003	0.49592
	0.25591	0.01316	0.41212	0.38601	0.13563	0.44737		0.58022	0.061	0.15702
Sour Cream and Onion	-0.7862	-1.6143	-0.644	-0.6685	0.5173	-0.6096	-0.2589	0	-1.2621	-0.9654
	0.4522	0.55	0.46682	0.47541	0.50769	0.48847	0.46668	0	0.52839	0.50156
	0.08486	0.00405	0.17052	0.1625	0.31045	0.21466	0.58022		0.0186	0.0568
Southern Barbecue	0.47585	-0.3522	0.61812	0.59363	1.77938	0.65248	1.00322	1.26208	0	0.29663
	0.5175	0.55347	0.49636	0.53139	0.57491	0.5162	0.53003	0.52839	0	0.53083
	0.35983	0.52589	0.21564	0.26635	0.00249	0.20888	0.061	0.0186		0.57742
Truffle Fries	0.17921	-0.6488	0.32149	0.297	1.48275	0.35584	0.70658	0.96545	-0.2966	0
	0.47513	0.57013	0.46888	0.46318	0.54826	0.4642	0.49592	0.50156	0.53083	0
	0.70675	0.25758	0.49436	0.5227	0.00792	0.44497	0.15702	0.0568	0.57742	

The report shows, for example, that US males prefer Barbecue chips to Gyro, Reuben, and Sour Cream and Onion chips with a high degree of significance.

Launch the MaxDiff Platform

Launch the MaxDiff platform by selecting **Analyze > Consumer Research > MaxDiff**.

Your data for the MaxDiff platform can be combined in a single data table or it can reside in two or three separate data tables. When the Choice window opens, specify whether you are using one or several data tables by selecting from the Data Format list.

One Table, Stacked

For this format, the data are combined into a single data table with a row for every profile presented to a subject and an indicator for the best and worst choices in that profile.

For an example of data in the one-table format, see "One Table Format" on page 176. For details, see "Launch Window for One Table, Stacked" on page 184.

Multiple Tables, Cross-referenced

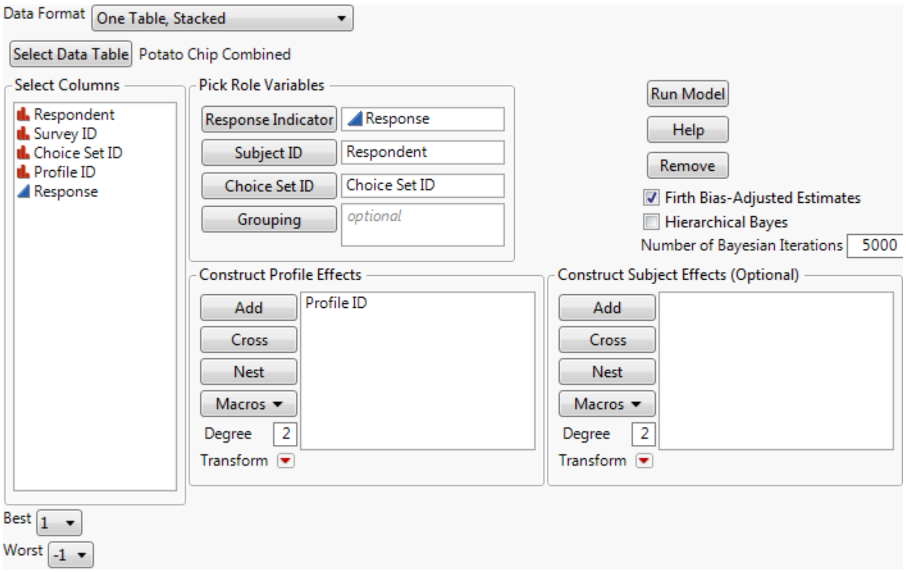
Your data can reside in two or three separate tables: a Profile Data and Response Data table are required, and a Subject Data table is optional. The MaxDiff Launch Window contains three sections, each corresponding to a different data table. You can expand or collapse each section of the launch window, as needed.

For an example of data in the multiple-tables format, see “Multiple Table Format” on page 178. For details, see “Launch Window for Multiple Tables, Cross-referenced” on page 185.

Launch Window for One Table, Stacked

Figure 8.10 shows the one-table launch window populated using Potato Chip Combined.jmp.

Figure 8.10 Launch Window for One Table, Stacked Data Format



Select Data Table Select or open the data table that contains the combined data. Select Other to open a file that is not already open.

Response Indicator A column containing the preference data. Use two of the values 1, -1, and 0 for the Best and Worst choices, and the third value for profiles that are not Best or Worst. Unless you specify a different coding using the menus next to Best and Worst in the lower left portion of the window, a 1 will indicate the Best choice and a -1 the Worst choice.

Subject ID An identifier for the study participant.

Choice Set ID An identifier for the set of profiles presented to the subject for a given preference determination.

Grouping A column which, when used with the Choice Set ID, uniquely designates each choice set. For example, if a choice set has Choice Set ID = 1 for Survey = A, and another choice set has Choice Set ID = 1 for Survey = B, then Survey should be used as a **Grouping** column.

Construct Profile Effects Add effects constructed from the attributes for the profiles.

For information about the Construct Profile Effects panel, see the Construct Model Effects section in the Model Specification chapter of the *Fitting Linear Models* book.

Construct Subject Effects (Optional) Add effects constructed from subject-related factors.

For information about the Construct Subject Effects panel, see the Construct Model Effects section in the Model Specification chapter of the *Fitting Linear Models* book.

Firth Bias-adjusted Estimates Computes bias-corrected MLEs that produce better estimates and tests than MLEs without bias correction. These estimates also improve separation problems that tend to occur in logistic-type models. Refer to Heinze and Schemper (2002) for a discussion of the separation problem in logistic regression.

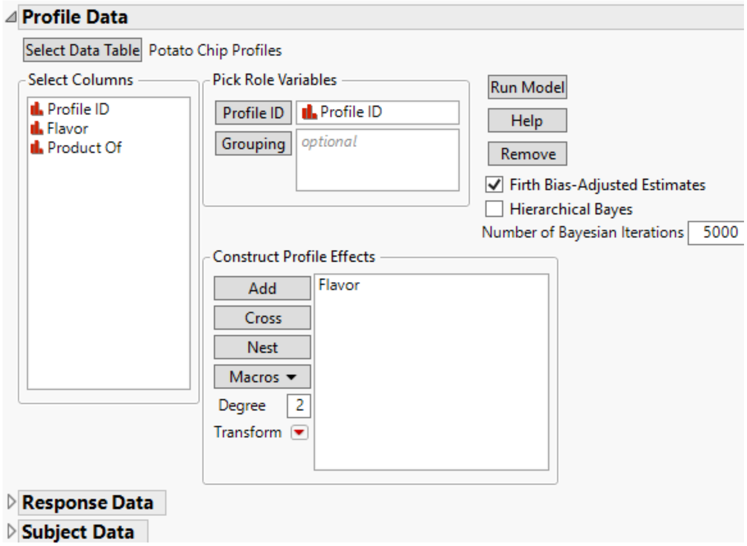
Hierarchical Bayes Uses a Bayesian approach to estimate subject-specific parameters. See "Bayesian Parameter Estimates" on page 193.

Number of Bayesian Iterations (Applicable only if Hierarchical Bayes is selected.) The total number of iterations of the adaptive Bayes algorithm used to estimate subject effects. This number includes a burn-in period of iterations that are discarded. The number of burn-in iterations is equal to half of the Number of Bayesian Iterations specified on the launch window.

Launch Window for Multiple Tables, Cross-referenced

Figure 8.11 shows the multiple-table launch window, with the Profile Data outline populated using Potato Chip Profile.jmp.

Figure 8.11 Launch Window for Multiple Tables, Cross-Referenced Data Format



In the case of Multiple Tables, Cross-referenced, the launch window has three sections:

- "Profile Data" on page 186
- "Response Data" on page 187
- "Subject Data" on page 188

Profile Data

The profile data table describes the attributes associated with each choice. Each choice can comprise many different attributes, and each attribute is listed as a column in the data table. There is a row for each possible choice, and each possible choice contains a unique ID.

Select Data Table Select or open the data table that contains the profile data. Select Other to open a file that is not already open.

Profile ID Identifier for each row of choice combinations. If the **Profile ID** column does not uniquely identify each row in the profile data table, you need to add **Grouping** columns. Add **Grouping** columns until the combination of **Grouping** and **Profile ID** columns uniquely identifies the row, or profile.

Grouping A column which, when used with the Choice Set ID column, uniquely designates each choice set. For example, if Profile ID = 1 for Survey = A, and a different Profile ID = 1 for Survey = B, then Survey would be used as a **Grouping** column.

Construct Profile Effects Add effects constructed from the attributes in the profiles.

For information about the Construct Profile Effects panel, see the Construct Model Effects section in the Model Specification chapter of the *Fitting Linear Models* book.

Firth Bias-adjusted Estimates Computes bias-corrected MLEs that produce better estimates and tests than MLEs without bias correction. These estimates also improve separation problems that tend to occur in logistic-type models. Refer to Heinze and Schemper (2002) for a discussion of the separation problem in logistic regression.

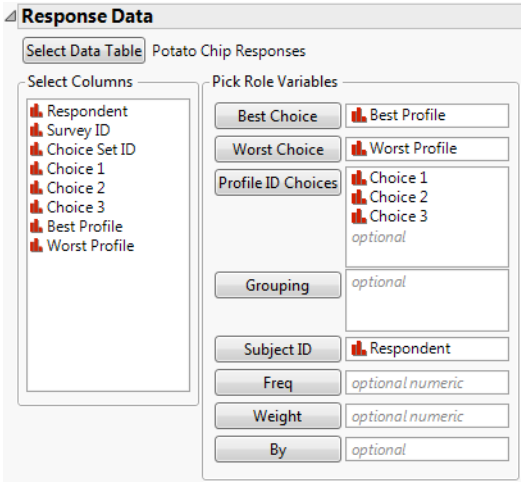
JMP PRO **Hierarchical Bayes** Uses a Bayesian approach to estimate subject-specific parameters. See "Bayesian Parameter Estimates" on page 193.

JMP PRO **Number of Bayesian Iterations** (Applicable only if Hierarchical Bayes is selected.) The total number of iterations of the adaptive Bayes algorithm used to estimate subject effects. This number includes a burn-in period of iterations that are discarded. The number of burn-in iterations is equal to half of the Number of Bayesian Iterations specified on the launch window.

Response Data

Figure 8.12 shows the Response Data outline populated using Potato Chip Responses.jmp.

Figure 8.12 Response Data Outline



The response data table contains the study results. It gives the choice set IDs for each trial as well as the profiles selected as best and worst by the subject. The Response data are linked to the Profile data through the choice set columns and the choice response column. Grouping variables can be used to align choice indices when more than one group is contained within the data.

Select Data Table Select or open the data table that contains the profile data. Select Other to open a file that is not already open.

Best Choice The Response table column containing the Profile ID of the profile that the subject designated as Best.

Worst Choice The Response table column containing the Profile ID of the profile that the subject designated as Worst.

Profile ID Choices The columns that contain the **Profile ID**s of the set of possible choices for each choice set.

Grouping A column which, when used with the Profile ID Chosen column, uniquely designates each choice set.

Subject ID A unique identifier for the study participant.

Freq A column containing frequencies. If *n* is the value of the Freq variable for a given row, then that row is used in computations n times. If it is less than 1 or missing, then JMP does not use it to calculate any analyses.

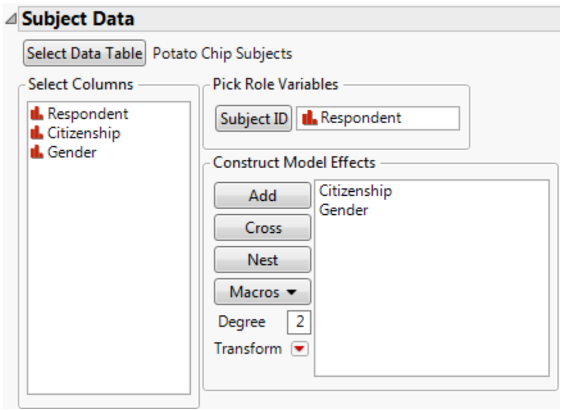
Weight A column containing a weight for each observation in the data table. The weight is included in analyses only when its value is greater than zero.

By A column whose levels define separate analyses. For each level of the specified column, the corresponding rows are analyzed as a separate analysis on a separate table. The results are presented in separate reports. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

Subject Data

Figure 8.13 shows the Subject Data outline populated using Potato Chip Subjects.jmp.

Figure 8.13 Subject Data Outline



Note: A subject data table is optional, depending on whether subject effects are to be modeled.

The subject data table contains the Subject ID and one or more columns of attributes or characteristics for each subject. The subject data table contains the same number of rows as subjects and has an identifier column that matches a similar column in the Response data table.

Note: You can include subject data in the response data table, but you need to specify subject effects in the Subject Data outline.

Select Data Table Select or open the data table that contains the subject data. Select Other to open a file that is not already open.

Subject ID Unique identifier for the subject.

Construct Model Effects Add effects constructed from columns in the subject data table.

For information about the Construct Model Effects panel, see the Construct Model Effects section in the Model Specification chapter of the *Fitting Linear Models* book.

MaxDiff Model Report

The MaxDiff Model window shows some of the following reports by default, depending on your selections in the launch window:

- "Effect Summary" on page 189
- "MaxDiff Results" on page 191
- "Parameter Estimates" on page 191
- "Bayesian Parameter Estimates" on page 193
- "Likelihood Ratio Tests" on page 194

For descriptions of the platform options, see "MaxDiff Platform Options" on page 194.

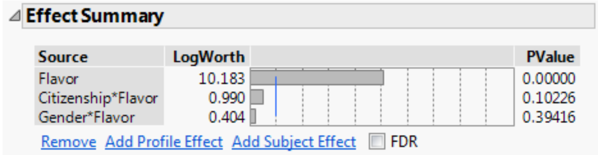
Effect Summary

The Effect Summary report appears if your model contains more than one effect. It lists the effects estimated by the model and gives a plot of the LogWorth (or FDR LogWorth) values for these effects. 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. For details, see the Effect Summary Report section in the Standard Least Squares Report and Options chapter in the *Fitting Linear Models* book.

The Effect Summary report does not appear when Bayesian Subject Effects is checked in the launch window. This is because likelihood ratio tests are not conducted in this case.

Figure 8.14 shows the Effect Summary report obtained by running the script **MaxDiff for Flavor** in Potato Chip Responses.jmp.

Figure 8.14 Effect Summary Report



Effect Summary

Source	LogWorth		PValue
Flavor	10.183		0.00000
Citizenship*Flavor	0.990		0.10226
Gender*Flavor	0.404		0.39416

Remove Add Profile Effect Add Subject Effect FDR

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. 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 the *p*-value corresponding to the significance test displayed in the Likelihood Ratio Tests 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 details about the FDR correction, see Benjamini and Hochberg, 1995. For details about the false discovery rate, see the Response Screening chapter in the *Predictive and Specialized Modeling* book or Westfall et al. (2011).

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.

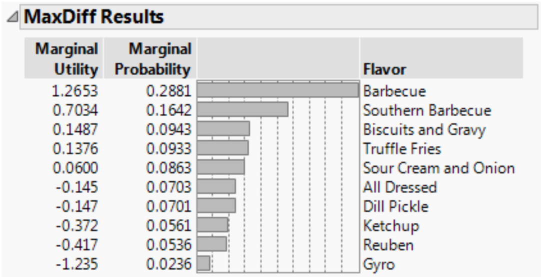
Add Profile Effect Opens a panel that contains a list of all columns in the data table for the OneTable, Stacked data format, and for the columns in the Profile Data table for the Multiple Tables, Cross-Referenced data format. Select columns that you want to add to the model, and then click Add below the column selection list to add the columns to the model. Click Close to close the panel.

Add Subject Effect Opens a panel that contains a list of all columns in the data table for the OneTable, Stacked data format, and for the columns in the Subject Data table for the Multiple Tables, Cross-Referenced data format. 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.

MaxDiff Results

Figure 8.15 shows the MaxDiff Results report obtained by running the script **MaxDiff for Flavor** in Potato Chip Responses.jmp.

Figure 8.15 MaxDiff Results Report



MaxDiff Results

Marginal Utility	Marginal Probability	Flavor
1.2653	0.2881	Barbecue
0.7034	0.1642	Southern Barbecue
0.1487	0.0943	Biscuits and Gravy
0.1376	0.0933	Truffle Fries
0.0600	0.0863	Sour Cream and Onion
-0.145	0.0703	All Dressed
-0.147	0.0701	Dill Pickle
-0.372	0.0561	Ketchup
-0.417	0.0536	Reuben
-1.235	0.0236	Gyro

For each Profile effect specified in the launch window, the following are displayed:

Marginal Utility An indicator of the perceived value of the corresponding level of the effect. Larger values suggest that the feature is of greater value.

Marginal Probability The estimated probability that a subject expresses a preference for the corresponding level of the effect over all other levels. For each effect, the marginal probabilities sum to one.

Bar Graph Shows a bar graph of the marginal probabilities.

Effect Column Gives the name of the effect and a list of its levels. The levels define the features to which the Marginal Utility and Marginal Probability estimates apply.

Parameter Estimates

This report gives details about parameter estimates, fit criteria, and the fitting algorithm.

Figure 8.16 shows the Parameter Estimates report obtained by running the script **MaxDiff for Flavor** in Potato Chip Responses.jmp.

Figure 8.16 Parameter Estimates Report

Parameter Estimates

Term	Estimate	Std Error
Flavor[All Dressed]	-0.15616234	0.2227718641
Flavor[Barbecue]	1.22210326	0.2978972284
Flavor[Biscuits and Gravy]	0.16398758	0.2249076601
Flavor[Dill Pickle]	-0.17356828	0.2151668801
Flavor[Gyro]	-1.11927509	0.2827016652
Flavor[Ketchup]	-0.47308393	0.2319900823
Flavor[Reuben]	-0.50927309	0.2294979936
Flavor[Sour Cream and Onion]	0.21115573	0.2450361368
Flavor[Southern Barbecue]	0.70149945	0.2695322771
Citizenship[Canadian]*Flavor[All Dressed]	-0.04368106	0.2239442590
Citizenship[Canadian]*Flavor[Barbecue]	-0.16180196	0.2978036763
Citizenship[Canadian]*Flavor[Biscuits and Gravy]	0.05734172	0.2233312021
Citizenship[Canadian]*Flavor[Dill Pickle]	-0.09824391	0.2188896544
Citizenship[Canadian]*Flavor[Gyro]	0.43257276	0.2907051874
Citizenship[Canadian]*Flavor[Ketchup]	-0.38035261	0.2349720398
Citizenship[Canadian]*Flavor[Reuben]	-0.34677939	0.2342643217
Citizenship[Canadian]*Flavor[Sour Cream and Onion]	0.56678250	0.2355493683
Citizenship[Canadian]*Flavor[Southern Barbecue]	-0.00712532	0.2720518372
Gender[Female]*Flavor[All Dressed]	-0.26955535	0.2106170508
Gender[Female]*Flavor[Barbecue]	0.39881410	0.2961368976
Gender[Female]*Flavor[Biscuits and Gravy]	0.09184462	0.2245258406
Gender[Female]*Flavor[Dill Pickle]	-0.11461679	0.2131955199
Gender[Female]*Flavor[Gyro]	-0.40538827	0.2950786212
Gender[Female]*Flavor[Ketchup]	-0.07317652	0.2135446256
Gender[Female]*Flavor[Reuben]	0.20780054	0.2227994637
Gender[Female]*Flavor[Sour Cream and Onion]	0.27353428	0.2337547848
Gender[Female]*Flavor[Southern Barbecue]	0.07570279	0.2627935035

AICc	397.44423
BIC	456.27172
-2*LogLikelihood	327.00945
-2*Firth LogLikelihood	244.39836

Converged in Gradient

Firth Bias-Adjusted Estimates

Term Lists the terms in the model.

Estimate An estimate of the parameter associated with the corresponding term. In discrete choice experiments, parameter estimates are sometimes referred to as *part-worths*. Each part-worth is the coefficient of utility associated with the given term. By default, these estimates are based on the Firth bias-corrected maximum likelihood estimators and therefore are considered to be more accurate than MLEs without bias correction.

Std Error An estimate of the standard deviation of the parameter estimate.

Comparison Criteria

The following fit statistics are shown as part of the report and can be used to compare models: AICc (corrected Akaike's Information Criterion), BIC (Bayesian Information Criterion), – 2*LogLikelihood, and – 2*Firth LogLikelihood. See the Statistical Details appendix in the *Fitting Linear Models* book for details on the first three of these measures.

The – 2*Firth LogLikelihood value is included in the report only when the Firth Bias-adjusted Estimates check box is checked in the launch window. This option is checked by default.

For each of these statistics, a smaller value indicates a better fit.

Bayesian Parameter Estimates

(Appears only if Hierarchical Bayes is selected on the launch window.)The Bayesian Parameter Estimates report gives results for model effects. The estimates are based on a Hierarchical Bayes fit that integrates the subject-level covariates into the likelihood function and estimates their effects on the parameters directly. The subject-level covariates are estimated using a version of the algorithm described in Train (2001), which incorporates Adaptive Bayes and Metropolis-Hastings approaches. Posterior means and variances are calculated for each model effect. The algorithm also provides subject-specific estimates of the model effect parameters. See "Save Subject Estimates" on page 196.

During the estimation process, each individual is assigned his or her own vector of parameter estimates, essentially treating the estimates as random effects and covariates. The vector of coefficients for an individual is assumed to come from a multivariate normal distribution with arbitrary mean and covariance matrix. The likelihood function for the utility parameters for a given subject is based on a multinominal logit model for each subject's preference within a choice set, given the attributes in the choice set. The prior distribution for a given subject's vector of coefficients is normal with mean equal to zero and a diagonal covariance matrix with the same variance for each subject. The covariance matrix is assumed to come from an inverse Wishart distribution with a scale matrix that is diagonal with equal diagonal entries.

For each subject, a number of burn-in iterations at the beginning of the chain is discarded. By default, this number is equal to half of the Number of Bayesian Iterations specified on the launch window.

Figure 8.17 Bayesian Parameter Estimates Report

Bayesian Parameter Estimates

Term	Posterior Mean	Posterior Std Dev	Subject Std Dev
Product Of[Canada]	-0.341776602	0.1656807233	0.2008820509
Citizenship[Canadian]*Product Of[Canada]	-0.021887377	0.1665704936	0.1752016641
Gender[Female]*Product Of[Canada]	-0.312611454	0.1699325105	0.1747635982

Total Iterations	5000
Burn-In Iterations	2500
Number of Respondents	30
Avg Log Likelihood After Burn-In	-197.0257

Term The model term.

Posterior Mean The parameter estimate for the term's coefficient. For each iteration after the burn-in period, the mean of the subject-specific coefficient estimates is computed. The Posterior Mean is the average of these means.

Tip: Select the red-triangle option Save Bayes Chain to see the individual estimates for each iteration.

Posterior Std Dev The standard deviation of the means of the subject-specific estimates over the iterations after burn-in.

Subject Std Dev The standard deviation of the subject-specific estimates around the posterior mean.

Tip: Select the red-triangle option Save Subject Estimates to see the individual estimates.

Total Iterations The total number of iterations performed, including the burn-in period.

Burn-In Iterations The number of burn-in iterations, which are discarded. This number is equal to half of the Number of Bayesian Iterations specified on the launch window.

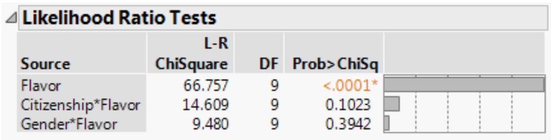
Number of Respondents The number of subjects

Avg Log Likelihood After Burn-In The average of the log-likelihood function, computed on values obtained after the burn-in period.

Likelihood Ratio Tests

Figure 8.18 shows the Likelihood Ratio Tests report obtained by running the script **MaxDiff for Flavor** in Potato Chip Responses.jmp.

Figure 8.18 Likelihood Ratio Tests



Likelihood Ratio Tests

Source	L-R ChiSquare	DF	Prob>ChiSq
Flavor	66.757	9	<.0001*
Citizenship*Flavor	14.609	9	0.1023
Gender*Flavor	9.480	9	0.3942

Source Lists the effects in the model.

L-R ChiSquare The value of the likelihood ratio ChiSquare statistic for a test of the corresponding effect.

DF The degrees of freedom for the ChiSquare test.

Prob>ChiSq The *p*-value for the ChiSquare test.

Bar Graph Shows a bar graph of the L-R ChiSquare values.

MaxDiff Platform Options

Show MLE Parameter Estimates (Available only if Hierarchical Bayes is selected on the launch window.) Shows non-Firth maximum likelihood estimates and standard errors for the coefficients of model terms. These estimates are used as starting values for the Hierarchical Bayes algorithm.

Joint Factor Tests (Not available if Hierarchical Bayes is selected on the launch window.) Tests each factor in the model by constructing a likelihood ratio test for all the effects involving that factor. For more information about Joint Factor Tests, see the Standard Least Squares Report and Options chapter in the *Fitting Linear Models* book.

Confidence Intervals If Hierarchical Bayes was not selected, shows a confidence interval for each parameter in the Parameter Estimates report.

If you selected Hierarchical Bayes, the confidence intervals appear in the Bayesian Parameter Estimates report. The intervals are constructed assuming a normal distribution and are based on the Posterior Mean and Posterior Std Dev.

Correlation of Estimates If Hierarchical Bayes was not selected, shows the correlations between the maximum likelihood parameter estimates.

If you selected Hierarchical Bayes, shows the correlation matrix for the posterior means of the parameter estimates. The correlations are calculated from the iterations after burn-in. The posterior means from each iteration after burn-in are treated as if they are columns in a data table. The Correlation of Estimates table is obtained by calculating the correlation matrix for these columns.

Comparisons Performs comparisons between specific alternative choice profiles. Enables you to select factor values and the values that you want to compare. You can compare specific configurations, including comparing all settings on the left or right by selecting the **Any** check boxes. Using **Any** does not compare all combinations across features, but rather all combinations of comparisons, one feature at a time, using the left settings as the settings for the other factors. See "See the JMP Reports chapter in the *Using JMP book for more information about the following options:*" on page 196.

All Levels Comparison Report Shows the All Levels Comparison Report, which contains a table with information on all pairwise comparisons of profiles. If you are modeling subject effects, you must specify a combination of subject effects and the table is specific to that combination of subject effects. Each cell of the table shows the difference in utilities for the row level and column level, the standard error of the difference, and a Wald p-value for a test of no difference.

Caution: The p-values are not corrected for multiple comparisons. Use these results as a guide.

The Wald p-values are colored. A saturated blue (respectively, red) color indicates that the Difference (Row - Column) is negative (respectively positive). The intensity of the red and blue coloring indicates the degree of significance, with a highly saturated red or blue meaning that the difference is highly significant.

Save Utility Formula Creates a new data table that contains a formula column for utility. The new data table contains a row for each subject and profile combination, and columns for the profiles and the subject effects.

Save Gradients by Subject Constructs a new table that has a row for each subject containing the average (Hessian-scaled-gradient) steps for the likelihood function on each parameter.

This corresponds to using a Lagrangian multiplier test for separating that subject from the remaining subjects. These values can later be clustered, using the built-in-script, to indicate unique market segments represented in the data. For more details, see “Example of Segmentation” on page 154 in the “Choice Models” chapter.

Save Subject Estimates (Available only if Hierarchical Bayes is selected.) Creates a table where each row contains the subject-specific parameter estimates for each effect. The distribution of subject-specific parameter effects for each effect is centered at the estimate for the term given in the Bayesian Parameter Estimates report. The Subject Acceptance Rate gives the rate of acceptance for draws of new parameter estimates during the Metropolis-Hastings step. Generally, an acceptance rate of 0.20 is considered to be good. See“Bayesian Parameter Estimates” on page 193.

Save Bayes Chain (Available only if Hierarchical Bayes is selected.) Creates a table that gives information on the chain of iterations used in computing subject-specific Bayesian estimates. See “Save Bayes Chain” on page 197.

Model Dialog Shows the MaxDiff launch window that resulted in the current analysis, which can be used to modify and re-fit the model. You can specify new data sets, new IDs, and new model effects.

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.

Comparisons Report

The Comparisons report is shown when you specify pairwise comparisons. It contains the following columns:

Factor Shows the levels of the subject factors that you specified.

Compared 1 Shows the factor and levels for the profile variables in the first component of the comparison.

Compared 2 Shows the factor and levels for the profile variables in the second component of the comparison.

Utility 1 Shows the estimated utility of the first component for the subjects specified in the Factor column.

Utility 2 Shows the estimated utility of the second component for the subjects specified in the Factor column.

Probability 1 Shows the predicted probability that the first component is preferred to the second for the subjects specified in the Factor column.

Probability 2 Shows the predicted probability that the second component is preferred to the first for the subjects specified in the Factor column.

Odds Ratio 1 Probability 1 divided by Probability 2.

Odds Ratio 2 Probability 2 divided by Probability 1.

Comparison Difference Utility 1 minus Utility 2.

Standard Deviation The sample standard error of the estimated Comparison Difference.

Save Bayes Chain

You can use the Bayes Chain data table to determine if your estimates have stabilized. The table that is created has a number of rows equal to the Number of Bayesian Iterations (specified on the launch window) plus one. The first row, Iteration 1, gives the starting values. The following rows show the results of the iterations, in order. The columns are arranged as follows:

Iteration Gives the iteration number, where the first row shows starting values.

Log Likelihood The log-likelihood of the model for that iteration. You can plot the Log Likelihood against Iteration to view behavior over the burn-in and tuning periods.

Adaptive Sigma for <model effect> Gives the estimate of the square root of the diagonal entries of the inverse Wishart distribution scale matrix for the corresponding effect.

Acceptance for <model effect> Gives the sampling acceptance rate for the corresponding effect.

Mean of <model effect> Gives the estimated mean for the corresponding effect.

Variance of <model effect> Gives the estimated variance for the corresponding effect.

Chapter **9**

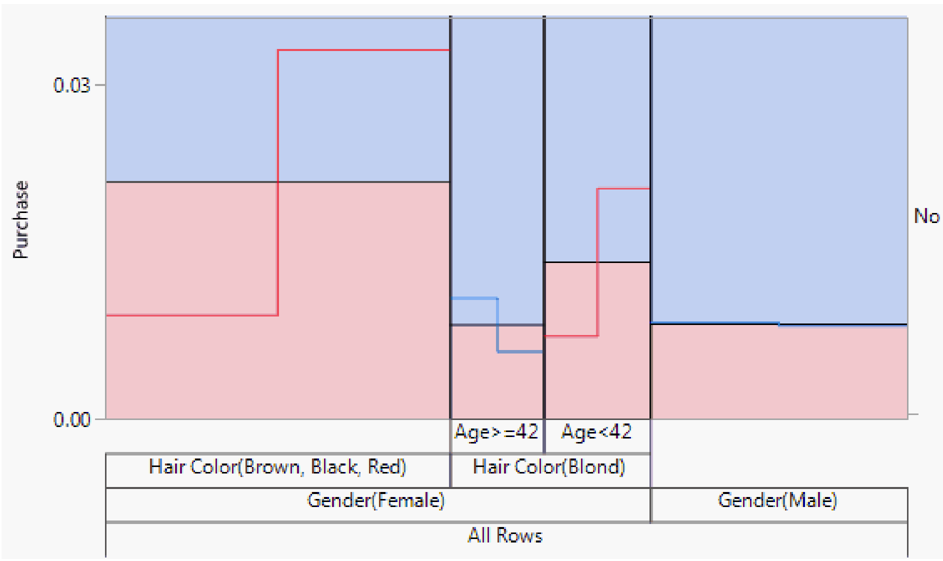
Uplift Models

Model the Incremental Impact of Actions on Consumer Behavior

Many features in this platform are available only in JMP Pro and noted with this icon.

Use uplift modeling to optimize marketing decisions, to define personalized medicine protocols, or, more generally, to identify characteristics of individuals who are likely to respond to an intervention. Also known as incremental modeling, true lift modeling, or net modeling, uplift modeling differs from traditional modeling techniques in that it finds the interactions between a treatment and other variables. It directs focus to individuals who are likely to react positively to an action or treatment.

Figure 9.1 Example of Uplift for a Hair Product Marketing Campaign



Term	Number of Splits	Contribution	Portion
Gender	1	44.1799575	0.4691
Age	1	25.474089	0.2705
Hair Color	1	24.5280291	0.2604
U.S. Region	0	0	0.0000
Residence	0	0	0.0000

JMP PRO Uplift Platform Overview

Use the Uplift platform to model the incremental impact of an action, or *treatment*, on individuals. An uplift model helps identify groups of individuals who are most likely to respond to the action. Identification of these groups leads to efficient and targeted decisions that optimize resource allocation and impact on the individual. (See Radcliffe and Surry, 2011.)

The Uplift platform fits partition models. While traditional partition models find splits to optimize a prediction, uplift models find splits to maximize a treatment difference.

The uplift partition model accounts for the fact that some individuals receive the treatment, while others do not. It does this by fitting a linear model to each possible (binary) split. A continuous response is modeled as a linear function of the split, the treatment, and the interaction of the split and treatment. A categorical response is expressed as a logistic function of the split, the treatment, and the interaction of the split and treatment. In both cases, the interaction term measures the difference in uplift between the groups of individuals in the two splits.

The criterion used by the Uplift platform in defining splits is the significance of the test for interaction over all possible splits. However, predictor selection based solely on *p*-values introduces bias favoring predictors with many levels. For this reason, JMP adjusts *p*-values to account for the number of levels. (See the paper “Monte Carlo Calibration of Distributions of Partition Statistics” on the JMP website.) The splits in the Uplift platform are determined by maximizing the adjusted *p*-values for *t* tests of the interaction effects. The logworth for each adjusted *p*-value, namely $-\log_{10}$(adj *p*-value), is reported.

Example of the Uplift Platform

The Hair Care Product.jmp sample data table results from a marketing campaign designed to increase purchases of a hair coloring product targeting both genders. For purposes of designing the study and tracking purchases, 126,184 "club card" members of a major beauty supply chain were identified. Approximately half of these members were randomly selected and sent a promotional offer for the product. Purchases of the product over a subsequent three-month period by all club card members were tracked.

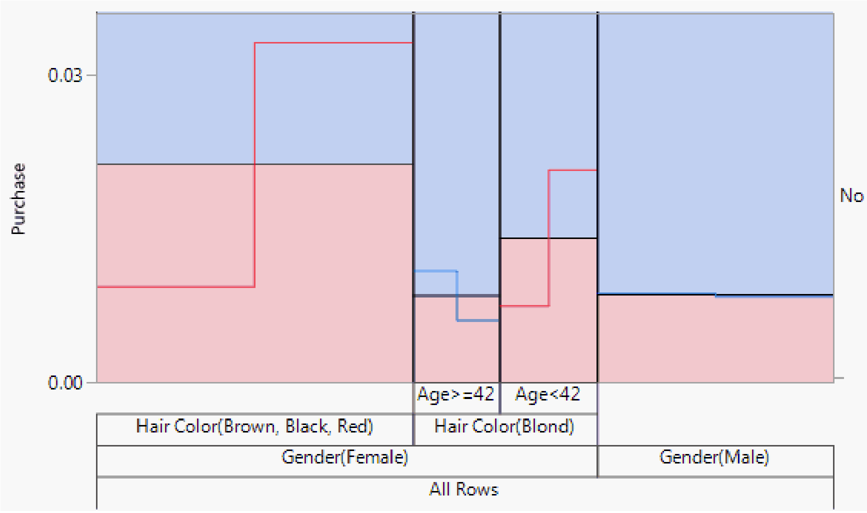
The data table shows a Promotion column, indicating whether the member received promotional material. The column Purchase indicates whether the member purchased the product over the test period. For each member, the following information was assembled: Gender, Age, Hair Color (natural), U.S. Region, and Residence (whether the member is located in an urban area). Also shown is a Validation column consisting of about 33% of the subjects.

For a categorical response, the Uplift platform interprets the first level in its value ordering as the response of interest. This is why the column Purchase has the Value Ordering column property. This property ensures that "Yes" responses are first in the ordering.

1. Select **Help > Sample Data Library** and open Hair Care Product.jmp.
2. Select **Analyze > Consumer Research > Uplift**.
3. From the Select Columns list:
   – Select Promotion and click **Treatment**.
   – Select Purchase and click **Y, Response**.
   – Select Gender, Age, Hair Color, U.S. Region, and Residence, and click **X, Factor**.
   – Select Validation and click **Validation**.
4. Click **OK**.
5. Below the Graph in the report that appears, click **Go**.

   Based on the validation set, the optimal Number of Splits is determined to be three. The Graph is shown in Figure 9.2. Note that the vertical scale has been modified in order to show the detail.

Figure 9.2 Graph after Three Splits

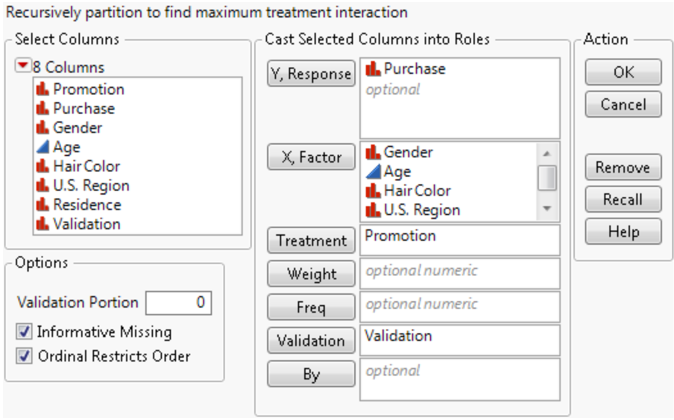


The graph indicates that uplift in purchases occurs for females with black, red, or brown hair and for younger females (Age < 42) with blond hair. For older blond-haired women (Age ≥ 42) and males, the promotion has a negative effect.

Launch the Uplift Platform

To launch the Uplift platform, select **Analyze > Consumer Research > Uplift**. Figure 9.3 shows a launch window for the Hair Care Product.jmp sample data table. The columns that you enter for Y, Response, and X, Factor can be continuous or categorical. In typical usage, the Treatment column is categorical, and often has only two levels. If your Treatment column contains more than two levels, the first level is treated as Treatment1 and the remaining levels are combined in Treatment2.

Figure 9.3 Launch Window for Uplift



You can specify your own Validation column, or designate a random portion of your data to be selected as a Validation Portion. 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.

The following options are also available:

Informative Missing If selected, enables missing value categorization for categorical predictors and informative treatment of missing values for continuous predictors.

Ordinal Restricts Order If selected, restricts consideration of splits to those that preserve the ordering.

The Uplift Model Report

The report opens by showing the Graph and the initial node of the Tree, as well as controls for splitting.

Uplift Model Graph

The graph represents the response on the vertical axis. The horizontal axis corresponds to observations, arranged by nodes. For each node, a black horizontal line shows the mean response. Within each split, there is a subsplit for treatment shown by a red or blue line. These lines indicate the mean responses for each of the two treatment groups within the split. The value ordering of the treatment column determines the placement order of these lines. As nodes are split, the graph updates to show the splits beneath the horizontal axis. Vertical lines divide the splits.

Beneath the graph are the control buttons: **Split**, **Prune**, and **Go**. The Go button only appears if there is a validation set. Also shown is the name of the Treatment column and its two levels, called Treatment1 and Treatment2. If more than two levels are specified for the Treatment column, all but the first level are treated as a single level and combined into Treatment2.

To the right of the Treatment column information is a report showing summary values relating to prediction. (Keep in mind that prediction is not the objective in uplift modeling.) The report updates as splitting occurs. If a validation set is used, values are shown for both the training and the validation sets.

RSquare The RSquare for the regression model associated with the tree. Note that the regression model includes interactions with the treatment column. An RSquare closer to 1 indicates a better fit to the data than does an RSquare closer to 0.

> **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 uplift model may have a low RSquare value.

RMSE The root mean square error (RMSE) for the regression model associated with the tree. RMSE is only given for continuous responses. For more details, see the *Fitting Linear Models* book.

N The number of observations.

Number of Splits The number of times splitting has occurred.

AICc The Corrected Akaike Information Criterion (AICc), computed using the associated regression model. AICc is only given for continuous responses. For more details, see the Statistical Details appendix in the *Fitting Linear Models* book.

JMP PRO Uplift Decision Tree

The decision tree shows the splits used to model uplift. See Figure 9.4 for an example using the Hair Care Product.jmp sample data table. Each node contains the following information:

Treatment The name of the treatment column is shown, with its two levels.

Rate Only appears for two-level categorical responses. For each treatment level, the proportion of subjects in this node who responded.

Mean Only appears for continuous responses. For each treatment level, the mean response for subjects in this node.

Count The number of subjects in this node in the specified treatment level.

t Ratio The *t* ratio for the test for a difference in response across the levels of Treatment for subjects in this node. If the response is categorical, it is treated as continuous (values 0 and 1) for this test.

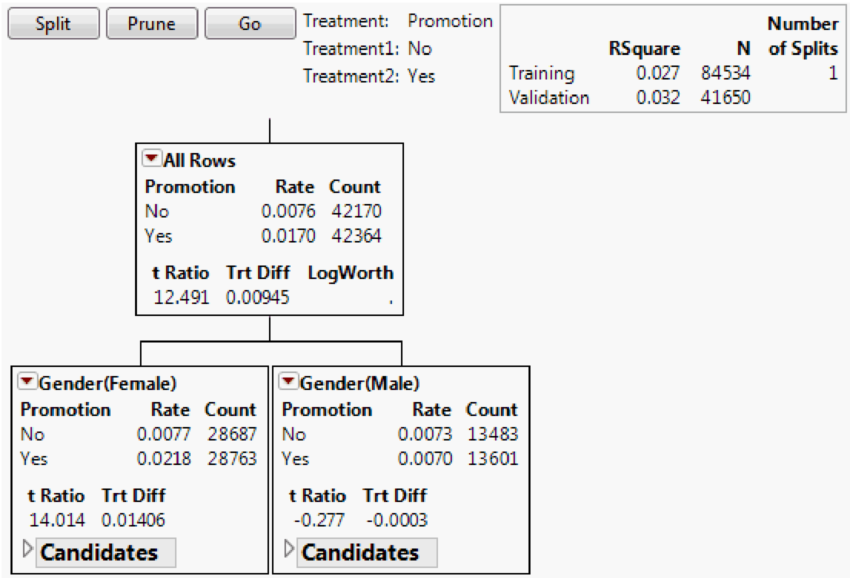
Trt Diff The difference in response means across the levels of Treatment. This is the uplift, assuming that:

– The first level in the treatment column's value ordering represents the treatment.

– The response is defined so that larger values reflect greater impact.

LogWorth The value of the logworth for the subsequent split based on the given node.

Figure 9.4 Nodes for First Split



Candidates Report

Each node also contains a Candidates report. This report gives the following information:

Term The model term.

LogWorth The maximum logworth over all possible splits for the given term. The logworth corresponding to a split is -$\log_{10}$ of the adjusted *p*-value.

F Ratio When the response is continuous, this is the F Ratio associated with the interaction term in a linear regression model. The regression model specifies the response as a linear function of the treatment, the binary split, and their interaction. When the response is categorical, this is the ChiSquare value for the interaction term in a nominal logistic model.

Gamma When the response is continuous, this is the coefficient of the interaction term in the linear regression model used in computing the *F* ratio. When the response is categorical, this is an estimate of the interaction constructed from Firth-adjusted log-odds ratios.

Cut Point If the term is continuous, this is the point that defines the split. If the term is categorical, this describes the first (left) node.

Uplift Report Options

With the exception of the options described below, all of the red triangle options for the Uplift report are described in the documentation for the Partition platform. For details about these options, see the Partition Models chapter in the *Predictive and Specialized Modeling* book.

Minimum Size Split

This option presents a window where you enter a number or a fractional portion of the total sample size to define the minimum size split allowed. To specify a number, enter a value greater than or equal to 1. To specify a fraction of the sample size, enter a value less than 1. The default value for the Uplift platform is set to 25 or the floor of the number of rows divided by 2,000, whichever value is greater.

Column Uplift Contributions

This table and plot address a column's contribution to the uplift tree structure. A column's contribution is computed as the sum of the F Ratio values associated with its splits. Recall that these values measure the significance of the treatment-by-split interaction term in the linear regression model.

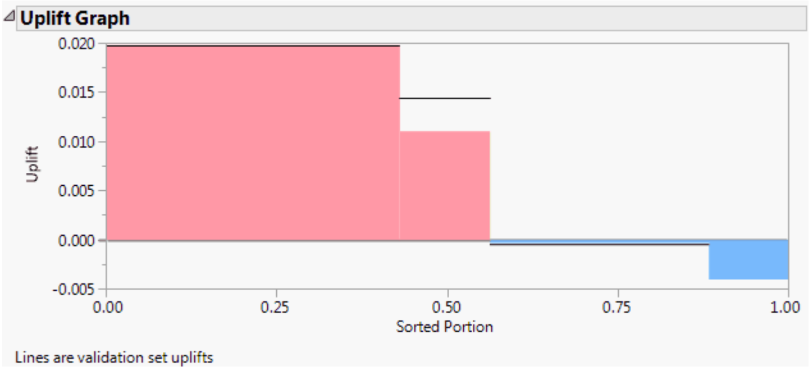
Uplift Graph

Consider the observations in the training set. Define uplift for an observation as the difference between the predicted probabilities or means across the levels of Treatment for the observation's terminal node. These uplift values are sorted in descending order. On its vertical axis, the Uplift Graph shows the uplift values. On its horizontal axis, the graph shows the proportion of observations with each uplift value.

See Figure 9.5 for an example of an Uplift Graph for the Hair Care Product.jmp sample data table after three splits. Note that, for two groups of subjects (males and non-blond women in the Age ≥ 42 group), the promotion has a negative effect.

The horizontal lines shown on the Uplift Graph delineate the graph for the validation set. Specifically, the decision tree is evaluated for the validation set and the Uplift Graph is constructed from the estimated uplifts.

Figure 9.5 Uplift Graph



Save Columns

Save Difference Saves the estimated difference in mean responses across levels of Treatment for the observation's node. This is the estimated uplift.

Save Difference Formula Saves the formula for the Difference, or uplift.

Publish Difference Formula Creates the difference 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.

Chapter 10

Item Analysis

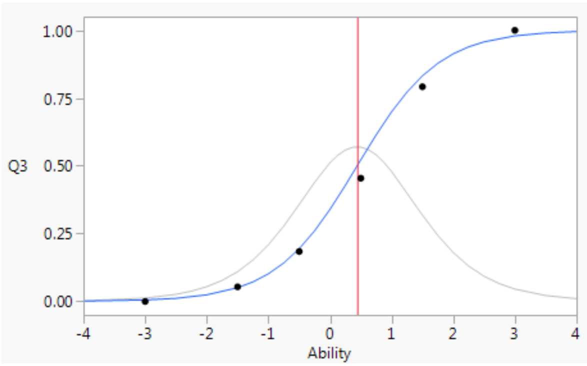
Analyze Test Results by Item and Subject

Item Response Theory (IRT) is a method of scoring tests. Although classical test theory methods have been widely used for a century, IRT provides a better and more scientifically based scoring procedure.

Its advantages include:

- Scoring tests at the item level, giving insight into the contributions of each item on the total test score.
- Producing scores of both the test takers and the test items on the same scale.
- Fitting nonlinear logistic curves, more representative of actual test performance than classical linear statistics.

Figure 10.1 Item Analysis Example



Item Analysis Platform Overview

Psychological measurement is the process of assigning quantitative values as representations of characteristics of individuals or objects, so-called *psychological constructs*. *Measurement theories* consist of the rules by which those quantitative values are assigned. Item Response Theory (IRT) is a measurement theory.

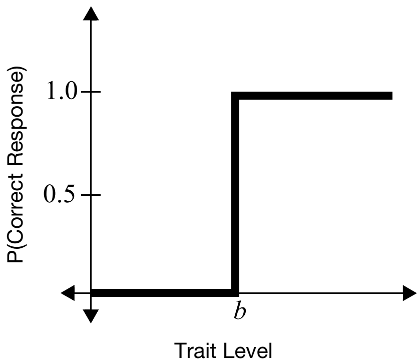
IRT uses a mathematical function to relate an individual's probability of correctly responding to an item to a trait of that individual. Frequently, this trait is not directly measurable and is therefore called a *latent trait*.

To see how IRT relates traits to probabilities, first examine a test question that follows the Guttman "perfect scale" as shown in Figure 10.2. The horizontal axis represents the amount of the theoretical trait that the examinee has. The vertical axis represents the probability of a correct response to the item by the examinee.

Note: A missing value for a test question is treated as an incorrect response.

The curve in Figure 10.2 is called an *item characteristic curve* (ICC).

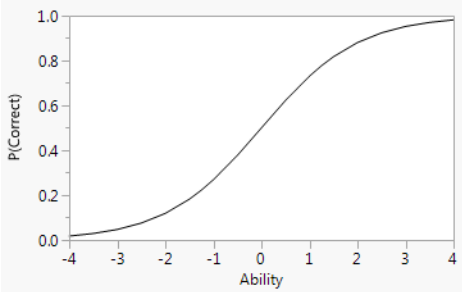
Figure 10.2 Item Characteristic Curve of a Perfect Scale Item



This figure shows that a person who has ability less than the value *b* has a 0% chance of getting the item correct. A person with trait level higher than *b* has a 100% chance of getting the item correct.

Of course, this is an unrealistic item, but it is illustrative in showing how a trait and a question probability relate to each other. More typical is a curve that allows probabilities that vary from zero to one. A typical curve found empirically is the S-shaped logistic function with a lower asymptote at zero and upper asymptote at one. It is markedly nonlinear. An example curve is shown in Figure 10.3.

Figure 10.3 Example Item Response Curve

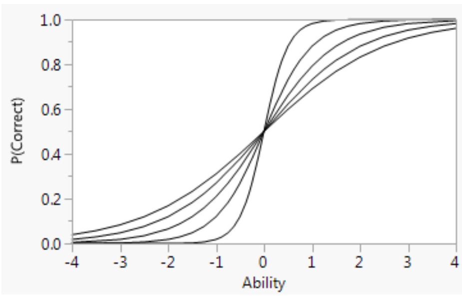


The logistic model is the best choice to model this curve. It has desirable asymptotic properties, yet it is easier to deal with computationally than other proposed models (such as the cumulative normal density function). The model itself is

$$P(\theta) \;=\; c + \frac{1-c}{1 + e^{-(a)(\theta - b)}}$$

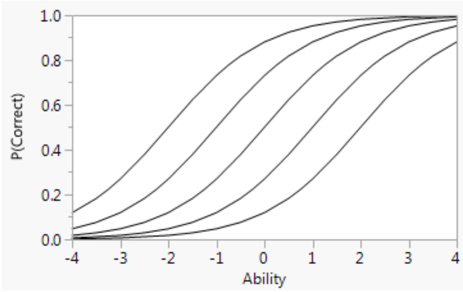
In this model, referred to as a Three-Parameter Logistic (*3PL*) model, the variable *a* represents the steepness of the curve at its inflection point. Curves with varying values of *a* are shown in Figure 10.4. The *a* parameter can be interpreted as a measure of the discrimination of an item. The discrimination of an item refers to how much more difficult the item is for people with high levels of the trait than for those with low levels of the trait. Very large values of *a* make the model practically the step function shown in Figure 10.2. It is generally assumed that an examinee has a higher probability of getting an item correct as their level of the trait increases. Therefore, *a* is assumed to be positive and the ICC is monotonically increasing. Some use this positive-increasing property of the curve as a test of the appropriateness of the item. Items whose curves do not have this shape should be considered as candidates to be dropped from the test.

Figure 10.4 Logistic Model for Several Values of *a*



Changing the value of *b* merely shifts the curve from left to right, as shown in Figure 10.5. It corresponds to the value of θ at the point where *P*(θ)=0.5. The parameter *b* can therefore be interpreted as item difficulty where (graphically), the more difficult items have their inflection points farther to the right along their x-coordinate.

Figure 10.5 Logistic Curve for Several Values of b

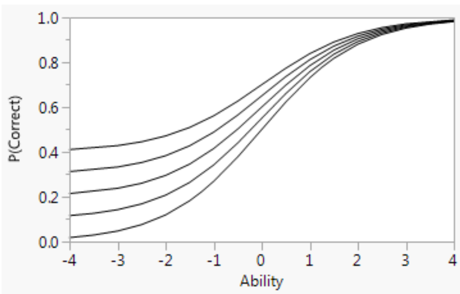


Notice that

$$\lim_{\theta \to -\infty} P(\theta) = c$$

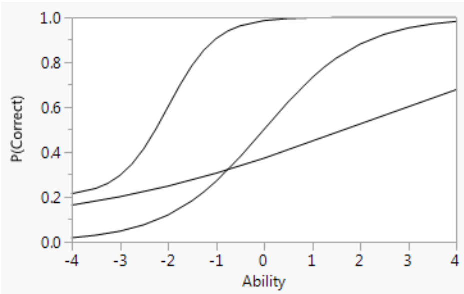
and therefore c represents the lower asymptote, which can be nonzero. ICCs for several values of c are shown graphically in Figure 10.6. The c parameter is theoretically pleasing, because a person with no ability of the trait might have a nonzero chance of getting an item right. Therefore, c is sometimes called the *pseudo-guessing parameter*.

Figure 10.6 Logistic Model for Several Values of c



By varying these three parameters, a wide variety of probability curves are available for modeling. A sample of three different ICCs is shown in Figure 10.7. The lower asymptote varies because of the assumption that there might be a lower guessing parameter. However, the upper asymptote does vary, because there is always a theoretical chance of 100% probability of correctly answering the item.

Figure 10.7 Three Item Characteristic Curves



Note, however, that the 3PL model might by unnecessarily complex for many situations. If, for example, the *c* parameter is restricted to be zero (in practice, a reasonable restriction), there are fewer parameters to predict. This simpler model, where only *a* and *b* parameters are estimated, is called the 2PL model.

The 2PL model has greater stability than the 3PL model. Another advantage of the 2PL model is that *b* can be interpreted as the point where an examinee has a 50% chance of getting an item correct. This interpretation is not true for 3PL models.

A further restriction can be imposed on the general model when a researcher can assume that test items have equal discriminating power. In these cases, the parameter *a* is set equal to 1, leaving a single parameter to be estimated, the *b* parameter. This *1PL* model is frequently called the *Rasch model*, named after Danish mathematician Georg Rasch, the developer of the model. The Rasch model is quite elegant, and is the least expensive to use computationally.

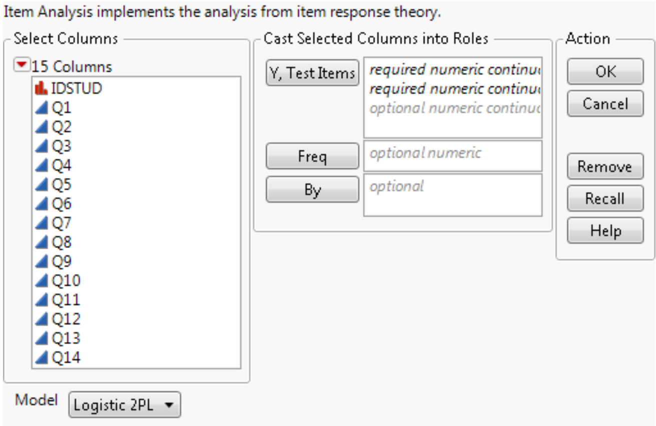
Caution: You must have a lot of data to produce stable parameter estimates using a 3PL model. 2PL models are frequently sufficient for tests that intuitively deserve a guessing parameter. Therefore, the 2PL model is the default and recommended model.

Launch the Item Analysis Platform

For example, open the sample data file MathScienceTest.jmp. These data are a subset of the data from the Third International Mathematics and Science Study (TIMMS) conducted in 1996.

To launch the Item Analysis platform, select **Analyze > Consumer Research > Item Analysis**. This shows the dialog in Figure 10.8.

Figure 10.8 Item Analysis Launch Window



Y, Test Items Are the questions from the test instrument.

Freq Specifies a variable used to specify the number of times each response pattern appears.

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

Specify the desired model (1PL, 2PL, or 3PL) by selecting it from the **Model** drop-down menu.

For this example, specify all fourteen continuous questions (Q1, Q2,..., Q14) as **Y, Test Items** and click **OK**. This accepts the default 2PL model.

Special Note on 3PL Models

If you select the 3PL model, a dialog pops up asking for a penalty for the *c* parameters (thresholds). This is not asking for the threshold itself. The penalty that it requests is similar to the type of penalty parameter that you would see in ridge regression, or in neural networks.

The penalty is on the sample variance of the estimated thresholds, so that large values of the penalty force the estimated thresholds' values to be closer together. This has the effect of speeding up the computations, and reducing the variability of the threshold (at the expense of some bias).

Sometimes, the items are questions on a multiple choice test where there are the same number of possible responses for each question. In this situation, there is often reason to believe (*a priori*) that the threshold parameters would be similar across items. For example, suppose you are analyzing the results of a 20-question multiple choice test where each question has four possible responses. It is reasonable to believe that the guessing, or threshold, parameters would all be near 0.25. So, in some cases, applying a penalty like this has some "physical intuition" to support it, in addition to its computational advantages.

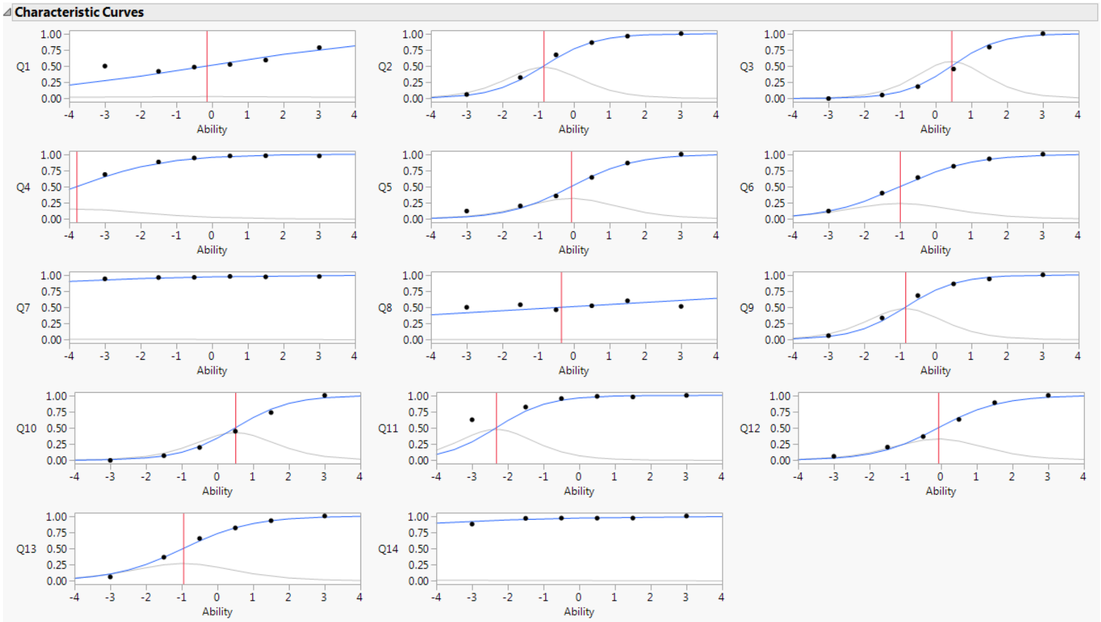
The Item Analysis Report

The following plots appear in Item Analysis reports.

Characteristic Curves

Item characteristic curves for each question appear in the top section of the output. Initially, all curves are shown stacked in a single column. They can be rearranged using the **Number of Plots Across** command in the Item Analysis red triangle menu. For Figure 10.9, four plots across are displayed.

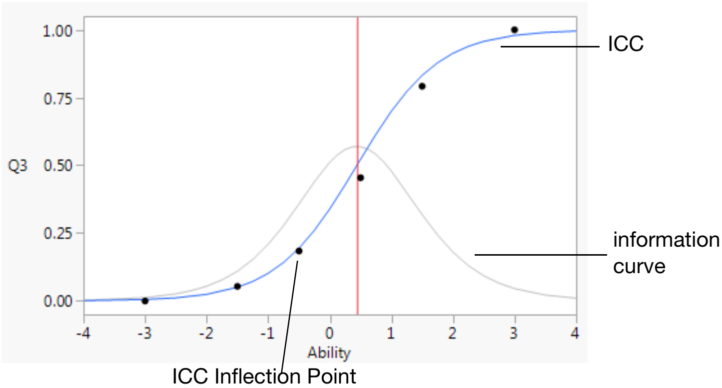
Figure 10.9 Component Curves



A vertical red line is drawn at the inflection point of each curve. In addition, dots are drawn at the actual proportion correct for each ability level, providing a graphical method of judging goodness-of-fit.

Gray information curves show the amount of information each question contributes to the overall information of the test. The information curve is the slope of the ICC curve, which is maximized at the inflection point.

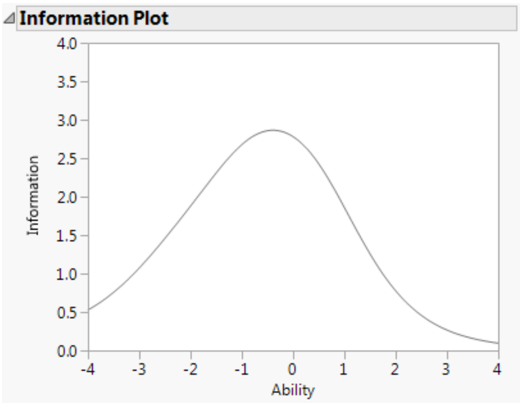
Figure 10.10 Elements of the ICC Display



Information Curves

Questions provide varying levels of information for different ability levels. The gray information curves for each item show the amount of information that each question contributes to the total information of the test. The total information of the test for the entire range of abilities is shown in the Information Plot section of the report (Figure 10.11).

Figure 10.11 Information Plot

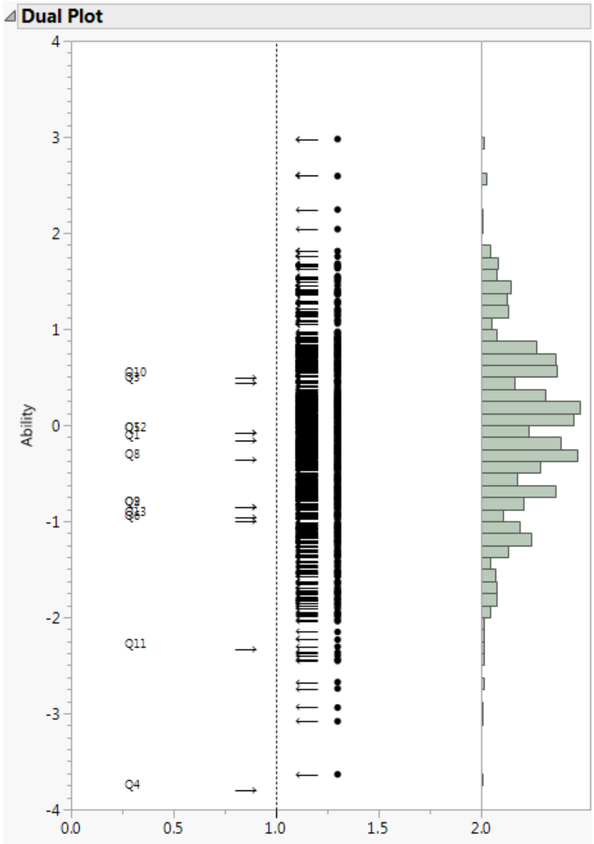


Dual Plots

The information gained from item difficulty parameters in IRT models can be used to construct an increasing scale of questions, from easiest to hardest, on the same scale as the examinees. This structure gives information about which items are associated with low levels of the trait, and which are associated with high levels of the trait.

JMP shows this correspondence with a *dual plot*. The dual plot for this example is shown in Figure 10.12.

Figure 10.12 Dual Plot



Questions are plotted to the left of the vertical dotted line, examinees on the right. In addition, a histogram of ability levels is appended to the right side of the plot.

This example shows a wide range of abilities. Q10 is rated as difficult; an examinee needs to be around half a standard deviation above the mean to have a 50% chance of correctly answering the question. Other questions are distributed at lower ability levels. Q11 and Q4 appear as easier than most of the other questions. There are some questions that are off the displayed scale (Q7 and Q14).

The estimated parameter estimates appear below the Dual Plot, as shown in Figure 10.13.

Figure 10.13 Parameter Estimates

Parameter Estimates

Item	Difficulty	Discrimination
Q1	-0.1561	0.3505
Q2	-0.8481	1.3857
Q3	0.4392	1.5106
Q4	-3.7953	0.7921
Q5	-0.0803	1.1318
Q6	-1.0003	0.9835
Q7	-11.111	0.3056
Q8	-0.3622	0.1300
Q9	-0.8527	1.3829
Q10	0.4977	1.3062
Q11	-2.3294	1.3821
Q12	-0.0796	1.1507
Q13	-0.9568	1.0374
Q14	-10.345	0.3285

Item Identifies the test item.

Difficulty Is the *b* parameter from the model. A histogram of the difficulty parameters is shown beside the difficulty estimates.

Discrimination Is the *a* parameter from the model, shown only for 2PL and 3PL models. A histogram of the discrimination parameters is shown beside the discrimination estimates.

Threshold Is the *c* parameter from the model, shown only for 3PL models.

Item Analysis Platform Options

The following three commands are available from the drop-down menu on the title bar of the report.

Number of Plots Across Brings up a dialog to specify how many plots should be grouped together on a single line. Initially, plots are stacked one-across. Figure 10.9 on page 215 shows four plots across.

Save Ability Formula Creates a new column in the data table containing a formula for calculating ability levels. Because the ability levels are stored as a formula, you can add rows to the data table and have them scored using the stored ability estimates. In addition, you can run several models and store several estimates of ability in the same data table.

The ability is computed using the `IRT Ability` function. The function has the following form:

```
IRT Ability (Q1, Q2,...,Qn, [a1, a2,..., an, b1, b2,..., bn, c1, c2, ...,
  cn]);
```

where `Q1`,...,`Qn` are columns from the data table containing items; `a1`,...,`an` are the corresponding discrimination parameters; `b1`,...,`bn` are the corresponding difficulty parameters for the items; `c1`,...,`cn` are the corresponding threshold parameters. Note that the parameters are entered as a matrix, enclosed in square brackets.

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.

Technical Details

Note that $P(\theta)$ does not necessarily represent the probability of a positive response from a *particular* individual. An individual examinee might use prior experiences and knowledge to definitely select an incorrect answer, or know an answer for sure, apart from the trait level. It is more correct to think of $P(\theta)$ as the probability of response for a set of individuals with ability level θ. Said another way, if a large group of individuals with equal trait levels answered the item, $P(\theta)$ predicts the proportion that would answer the item correctly. This implies that IRT models are item-invariant; theoretically, they would have the same parameters regardless of the group tested.

An assumption of these IRT models is that the underlying trait is unidimensional. That is to say, there is a single underlying trait that the questions measure that can be theoretically measured on a continuum. This continuum is the horizontal axis in the plots of the curves. If there are several traits that have complex interactions with each other being measured, then these unidimensional models are not appropriate.

Appendix A

References

Agresti, A. (2002), *Categorical Data Analysis*, Second Edition, New York: John Wiley and Sons, Inc.

Akaike, H. (1974), "Factor Analysis and AIC," *Pschychometrika*, 52, 317–332.

Akaike, H. (1987), "A new Look at the Statistical Identification Model," *IEEE Transactions on Automatic Control*, 19, 716–723.

Bartlett, M.S. (1954), "A Note on the Multiplying Factors for Various Chi Square Approximations," *Journal of the Royal Statistical Society, 16 (Series B)*, 296-298.

Benzécri, J. P. (1979), "Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire, addendum et erratum à [BIN. MULT.]," *Cahiers de l'Analyse des Données*, 4, 377–378.

Borg, I. and Groenen, P. J. F. (2005), *Modern Multidimensional Scaling: Theory and Applications,* Second Edition, New York: Springer.

Dwass, M. (1955), "A Note on Simultaneous Confidence Intervals," *Annals of Mathematical Statistics* 26: 146–147.

Farebrother, R.W. (1981), "Mechanical Representations of the L1 and L2 Estimation Problems," *Statistical Data Analysis*, 2nd Edition, Amsterdam, North Holland: edited by Y. Dodge.

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 80:1, 27–38.

Goodnight, J.H. (1978), "Tests of Hypotheses in Fixed Effects Linear Models," *SAS Technical Report R–101*, Cary: SAS Institute Inc, also in Communications in Statistics (1980), A9 167–180.

Goodnight, J.H. and W.R. Harvey (1978), "Least Square Means in the Fixed Effect General Linear Model," *SAS Technical Report R–103*, Cary NC: SAS Institute Inc.

Greenacre, M. J. (1984), *Theory and Applications of Correspondence Analysis*, London: Academic Press.

Heinze, G. and Schemper, M. (2002), "A Solution to the Problem of Separation in Logistic Regression," *Statistics in Medicine* 21:16, 2409–2419.

Hocking, R.R. (1985), *The Analysis of Linear Models*, Monterey: Brooks–Cole.

Hosmer, D.W. and Lemeshow, S. (2000), *Applied Logistic Regression*, Second Edition, New York: John Wiley and Sons.

Jackson, J.E. (2003), *A User's Guide to Principal Components,* New York: John Wiley and Sons, Inc.

Kaiser, H.F. (1958), “The varimax criterion for analytic rotation in factor analysis” *Psychometrika*, 23, 187–200.

Kish, L. (1965), *Survey Sampling*, New York: John Wiley and Sons, Inc.

Lavassani, K.M., Movahedi, B., and Kumar, V. (2009). “Developments in Analysis of Multiple Response Survey Data in Categorical Data Analysis: The Case of Enterprise System Implementation in Large North American Firms,” Journal of Applied Quantitative Methods, Vol. 4, No. 1, 45–53.

Louviere, J.J., Flynn, T.N., and Marley, A.A. (2015), *Best-Worst Scaling: Theory, Methods and Applications*, Cambridge University Press.

McFadden, D. (1974), “Conditional Logit Analysis of Qualitative Choice Behavior,” in P. Zarembka, ed., *Frontiers in Econometrics*, pp. 105–142. See http://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf, retrieved 4/25/2016.

Radcliffe, N. J., and Surry, P. D. (2011), “Real-World Uplift Modelling with Significance-Based Uplift Trees,” Stochastic Solutions White Paper, Portrait Technical Report TR-2011-1.

Reichheld, F. F. (2003), “The One Number You Need to Grow,” *Harvard Business Review*, Vol. 81 No. 12, 46–54.

Rossi, P.E., Allenby, G.M., and McCulloch, R. (2006), *Bayesian Statistics and Marketing*, Wiley & Sons, Ltd.

SAS Institute (2009), *SAS/STAT 9.2 User’s Guide, Second Edition*, Chapter 92, SAS Institute Inc., Cary NC. See https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_ttest_a0000000126.htm, retrieved 7/14/2016.

Train, K.E. (2009), *Discrete Choice Methods and Simulation*, Cambridge University Press.

Waern, Y. (1972) “Structure in Similarity Matrices: A graphic approach,” *Scandinavian Journal of Psychology*, Vol. 13, 5–16.

Wright, S.P. and R.G. O’Brien (1988), “Power Analysis in an Enhanced GLM Procedure: What it Might Look Like,” *SUGI 1988, Proceedings of the Thirteenth Annual Conference*, 1097–1102, Cary NC: SAS Institute Inc.

Index
Consumer Research

L

M

N

O

P

R

S

T

U

W-Z