



Version 17

Predictive and Specialized Modeling

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

Marcel Proust

JMP Statistical Discovery LLC
SAS Campus Drive
Cary, North Carolina 27513-2414

17.0

The correct bibliographic citation for this manual is as follows: JMP Statistical Discovery LLC 2022. *JMP® 17 Predictive and Specialized Modeling*. Cary, NC: JMP Statistical Discovery LLC

JMP® 17 Predictive and Specialized Modeling

Copyright © 2022, JMP Statistical Discovery LLC, Cary, NC, USA

All rights reserved. Produced in the United States of America.

JMP Statistical Discovery LLC, SAS Campus Drive, Cary, North Carolina 27513-2414.

October 2022

JMP® and all other JMP Statistical Discovery LLC product or service names are registered trademarks or trademarks of JMP Statistical Discovery LLC in the USA and other countries. ® indicates USA registration.

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

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

Get the Most from JMP

Whether you are a first-time or a long-time user, there is always something to learn about JMP.

Visit JMP.com to find the following:

- live and recorded webcasts about how to get started with JMP
- video demos and webcasts of new features and advanced techniques
- details on registering for JMP training
- schedules for seminars being held in your area
- success stories showing how others use JMP
- the JMP user community, resources for users including examples of add-ins and scripts, a forum, blogs, conference information, and so on

jmp.com/getstarted

Contents

Predictive and Specialized Modeling

1	Learn about JMP	17
	Documentation and Additional Resources	
	Formatting Conventions in JMP Documentation	19
	JMP Help	20
	JMP Documentation Library	20
	Additional Resources for Learning JMP	27
	Search JMP	27
	JMP Tutorials	27
	Sample Data Tables	28
	Learn about Statistical and JSL Terms	28
	Learn JMP Tips and Tricks	28
	JMP Tooltips	28
	JMP User Community	29
	Free Online Statistical Thinking Course	29
	JMP New User Welcome Kit	29
	Statistics Knowledge Portal	29
	JMP Training	30
	JMP Books by Users	30
	The JMP Starter Window	30
	JMP Technical Support	30
2	Introduction to Predictive and Specialized Modeling	31
	Overview of Modeling Techniques	
3	Neural Networks	35
	Fit Nonlinear Models Using Nodes and Layers	
	Overview of the Neural Platform	37
	Example of a Neural Network	37
	Launch the Neural Platform	40
	The Neural Launch Window	40
	The Model Launch Control Panel	42
	Validation Methods for Neural	43
	Hidden Layer Structure	44

	Boosting	45
	Neural Fitting Options	45
	Model Reports	47
	Training and Validation Measures of Fit	47
	Confusion Statistics	48
	Neural Platform Options	48
	Neural Model Options	49
4	Partition Models	51
	Use Decision Trees to Explore and Model Your Data	
	Overview of the Partition Platform	53
	Example of the Partition Platform	53
	Launch the Partition Platform	56
	Informative Missing	57
	The Partition Report	58
	Partition Plot	58
	Partition Buttons	60
	Summary of Fit Statistics	60
	Node Reports	61
	Partition Platform Options	63
	Show Fit Details	66
	Specify Profit Matrix	68
	ROC Curve	69
	Lift Curve	71
	Validation in Partition	72
	Additional Examples of the Partition Platform	73
	Example of a Continuous Response	73
	Example of Informative Missing	76
	Example of Profit Matrix and Decision Matrix Report	79
	Statistical Details for the Partition Platform	82
	Statistical Details for Responses and Factors	82
	Statistical Details for the Splitting Criterion	83
5	Bootstrap Forest	85
	Fit a Model By Averaging Many Trees	
	Overview of the Bootstrap Forest Platform	87
	Example of Bootstrap Forest with a Categorical Response	88
	Example of Bootstrap Forest with a Continuous Response	91
	Launch the Bootstrap Forest Platform	93
	Specification Window	95
	The Bootstrap Forest Report	97

	Model Validation Set Summaries	99
	Specifications	99
	Overall Statistics	99
	Cumulative Validation	101
	Per Tree Summaries	102
	Bootstrap Forest Platform Options	102
6	Boosted Tree	105
	Fit Many Layers of Trees, Each Based on the Previous Layer	
	Overview of the Boosted Tree Platform	107
	Example of Boosted Tree with a Categorical Response	107
	Example of Boosted Tree with a Continuous Response	109
	Launch the Boosted Tree Platform	111
	Specification Window	113
	The Boosted Tree Report	115
	Model Validation Set Summaries	116
	Specifications	117
	Overall Statistics	117
	Cumulative Validation	118
	Boosted Tree Platform Options	119
	Statistical Details for the Boosted Tree Platform	121
	Statistical Details for the Overfit Penalty	121
7	K Nearest Neighbors	123
	Predict Response Values Using Nearby Observations	
	Overview of the K Nearest Neighbors Platform	125
	Example of K Nearest Neighbors with a Categorical Response	126
	Example of K Nearest Neighbors with a Continuous Response	128
	Launch the K Nearest Neighbors Platform	130
	The K Nearest Neighbors Report	132
	Continuous Responses	132
	Categorical Responses	133
	K Nearest Neighbors Platform Options	134
	Model Fit Options	135
8	NaiveBayes	137
	Classify Observations Using Probabilistic Assumptions	
	Overview of the Naive Bayes Platform	139
	Example of Naive Bayes	140
	Launch the Naive Bayes Platform	142
	The Naive Bayes Report	143

	Fit Details Report	144
	Response Column Report	144
	Confusion Matrix Report	144
	ROC Curves	144
	Naive Bayes Platform Options	145
	Additional Example of Naive Bayes	146
	Statistical Details for the Naive Bayes Platform	147
	Statistical Details for the Naive Bayes Algorithm	147
	Statistical Details for Saved Probability Formulas	149
9	Support Vector Machines	151
	Predict Response Categories Using Support Vectors	
	Overview of Support Vector Machines	153
	Example of Support Vector Machines	153
	Launch the Support Vector Machines Platform	156
	Model Launch Control Panel	157
	The Support Vector Machine Report	160
	Model Comparison Report	160
	Support Vector Machine Model Report	161
	Support Vector Machines Platform Options	163
	Support Vector Machine Model Report Options	164
	Additional Example of Support Vector Machines	166
10	Model Screening	169
	Fit and Compare Multiple Prediction Models	
	Example of Model Screening	171
	Launch the Model Screening Platform	173
	The Model Screening Report	176
	Summary Across the Folds Report	177
	Training, Validation, and Test Measures of Fit	178
	Model Screening Platform Options	179
	Decision Thresholds Report	182
11	Model Comparison	185
	Compare the Predictive Ability of Fitted Models	
	Example of Model Comparison	187
	Launch the Model Comparison Platform	191
	The Model Comparison Report	192
	Model Comparison Platform Options	193
	Continuous Responses	193
	Categorical Responses	194

Additional Example of Model Comparison	195
12 Make Validation Column	199
Divide Data into Training, Validation, and Test Sets	
Overview of the Make Validation Column Platform	201
Example of the Make Validation Column Platform	201
Launch the Make Validation Column Platform	203
Make Validation Column Report	206
Specify Rates or Relative Rates	206
Set Cutpoints	206
Set Number of Folds	207
Options	207
Go	208
Make Validation Column Platform Options	208
Additional Example of the Make Validation Column Platform	208
13 Formula Depot	211
Manage Models and Generate Scoring Code	
Overview of the Formula Depot Platform	213
Example of the Formula Depot	213
Launch the Formula Depot Platform	214
Platforms That Publish Prediction Formulas to the Formula Depot	215
Formula Depot Platform Options	215
Formula Depot Model Options	217
Generate Scoring Code from the Formula Depot Platform	218
14 Fit Curve	221
Fit Built-In Nonlinear Models to Your Data	
Overview of the Fit Curve Platform	223
Example of the Fit Curve Platform	224
Launch the Fit Curve Platform	227
The Fit Curve Report	229
Model Comparison Report	229
Model Fit Report	231
Fit Curve Options	231
Model-Free Comparisons Reports	234
Model Fit Options	236
Test Parallelism	238
Compare Parameter Estimates	241
Equivalence Test	242
Additional Example of the Fit Curve Platform	243

Statistical Details for the Fit Curve Platform	245
Statistical Details for Fit Curve Models	246
Statistical Details for Dissolution Curve Analysis	256
15 Nonlinear Regression	259
Fit Custom Nonlinear Models to Your Data	
Example of the Nonlinear Platform	261
Launch the Nonlinear Platform	264
The Nonlinear Fit Report	265
Nonlinear Platform Options	269
Create a Formula Using the Model Library	273
Customize the Nonlinear Model Library	275
Additional Examples of the Nonlinear Platform	277
Example of Maximum Likelihood	277
Example of a Probit Model with Binomial Errors	280
Example of a Poisson Loss Function	281
Example of Setting Parameter Limits	283
Example of Analyzing Left-Censored Data	286
Example of Fitting a Weibull Loss Function	287
Example of Fitting Simple Survival Distributions	290
Statistical Details for the Nonlinear Platform	293
Statistical Details for Profile Likelihood Confidence Limits	293
Statistical Details for Custom Loss Functions	294
Statistical Details on Derivatives	295
Statistical Details on Effective Nonlinear Modeling	297
16 Functional Data Explorer	299
Perform Functional Data Analysis	
Overview of the Functional Data Explorer Platform	301
Example of Functional Data Explorer	301
Launch the Functional Data Explorer Platform	305
The Functional Data Explorer Report	308
Model Reports	309
Functional Data Explorer Platform Options	316
Functional Data Explorer Group Options	318
Data Processing Options	319
Model Report Options	323
Additional Examples of the Functional Data Explorer Platform	326
Example for Multiple Functional Processes	327
Example of Functional DOE Analysis	331
Statistical Details for the Functional Data Explorer Platform	333

Statistical Details for Functional Model Fits	334
Statistical Details for the Function Summaries Report	335
17 Gaussian Process	337
Fit Data Using Smoothing Models	
Example of the Gaussian Process Platform	339
Launch the Gaussian Process Platform	341
The Gaussian Process Report	342
Actual by Predicted Plot	342
Model Report	342
Marginal Model Plots	343
Gaussian Process Platform Options	343
Additional Examples of the Gaussian Process Platform	344
Example of a Gaussian Process Model	345
Example of a Gaussian Process Model with Categorical Predictors	346
Statistical Details for the Gaussian Process Platform	348
Statistical Details for Models with Continuous Predictors	348
Statistical Details for Models with Categorical Predictors	349
Statistical Details for Variance Formula Parameterization	350
Statistical Details for the Model Fit	350
18 Time Series Analysis	351
Fit Time Series Models and Transfer Functions	
Overview of the Time Series Platform	353
Example of the Time Series Platform	353
Launch the Time Series Platform	356
The Time Series Report	358
Time Series Graph	358
Time Series Basic Diagnostics Chart	359
Time Series Platform Options	361
Time Series Diagnostics	361
Differencing and Decomposition	362
Time Series Models	365
Additional Options	372
Additional Time Series Reports	373
Difference Report	374
Decomposition Reports	374
Model Comparison Report	376
Model Report	379
State Space Smoothing Report	383
Transfer Function Model Report	386

Spectral Density Report	387
Additional Examples of the Time Series Platform	388
Example of Creating a Time ID Column	388
Example of a Box-Cox Transformation	393
Example Using a Holdback Set	395
Statistical Details for the Time Series Platform	397
Statistical Details for Spectral Density	398
Statistical Details for X-11 Decomposition	398
Statistical Details for Exponential Smoothing Models	399
Statistical Details for ARIMA Models	402
Statistical Details for Transfer Functions	404
19 Time Series Forecast	405
Forecast Multiple Time Series	
Example of Time Series Forecast	407
Launch the Time Series Forecast Platform	409
Data Format	409
The Time Series Forecast Report	411
Analysis of Time Pattern	411
Group Summary of Time	411
Modeling Specifications	413
Time Series Forecast Platform Options	416
Time Series Forecast Model Reports	417
20 Matched Pairs Analysis	419
Compare Measurements on the Same Subject	
Overview of the Matched Pairs Platform	421
Example of Comparing Matched Pairs	421
Launch the Matched Pairs Platform	423
Multiple Y Columns	423
The Matched Pairs Report	424
Difference Plot and Report	425
Across Groups	425
Matched Pairs Platform Options	426
Example Comparing Matched Pairs across Groups	427
Statistical Details for the Matched Pairs Platform	428
Statistical Details for Graphics of Matched Pairs	429
Statistical Details for the Correlation of Responses	430

21	Explore Outliers	431
	Find Outliers in Univariate and Multivariate Data	
	Overview of Explore Outliers	433
	Quantile Range Outliers	433
	Robust Fit Outliers	433
	Robust PCA Outliers	433
	K Nearest Neighbor Outliers	434
	Example of Explore Outliers	435
	Launch the Explore Outliers Platform	438
	The Explore Outliers Report	442
	Quantile Range Outliers Report	442
	Robust Fit Outliers Report	447
	Robust PCA Outliers Report	450
	K Nearest Neighbors Report	452
	Explore Outliers Platform Options	453
	Additional Example of the Explore Outliers Platform	454
22	Explore Missing Values	457
	Explore and Impute Missing Values	
	Overview of the Explore Missing Values Platform	459
	Multivariate Normal Imputation	459
	Multivariate SVD Imputation	459
	Multivariate RPCA Imputation	460
	Automated Data Imputation	460
	Example of the Explore Missing Values Platform	461
	Launch the Explore Missing Values Platform	463
	The Explore Missing Values Report	464
	Commands	465
	Imputation Report	467
	Explore Missing Values Platform Options	468
23	Explore Patterns	469
	Explore Data for Unusual Features or Patterns	
	Overview of the Explore Patterns Platform	471
	Example of the Explore Patterns Platform	471
	Launch the Explore Patterns Platform	474
	The Explore Patterns Report	475
	Control Panel	475
	Univariate Summary	475
	Univariate Patterns	476
	Explore Patterns Options	477

Univariate Patterns Report Options	480
Additional Example of the Explore Patterns Platform	481
Statistical Details for the Explore Patterns Platform	483
24 Response Screening	485
Test Many Responses in Large-Scale Data	
Overview of the Response Screening Platform	487
Example of Response Screening	488
Launch the Response Screening Platform	493
The Response Screening Report	497
Response Screening Plots	497
Response Screening Report Tables	499
Response Screening Platform Options	503
Means Data Table	506
Means Differences Data Table	507
The PValues Data Table	509
The Response Screening Personality in Fit Model	510
Launch Response Screening in Fit Model	511
The Fit Response Screening Report	512
Fit Response Screening Options	514
Additional Examples of Response Screening	516
Example of Tests of Practical Significance and Equivalence	516
Example of the Max Logworth Option	518
Example of Robust Fit	520
Example of the Response Screening Personality	523
Example of a Volcano Plot	524
Statistical Details for the Response Screening Platform	526
25 Predictor Screening	527
Screen Many Predictors for Significant Effects	
Overview of the Predictor Screening Platform	530
Example of Predictor Screening	530
Launch the Predictor Screening Platform	532
The Predictor Screening Report	533
Predictor Screening Platform Options	533
26 Association Analysis	535
Perform Market Basket Analysis	
Overview of the Association Analysis Platform	537
Example of the Association Analysis Platform	538
Launch the Association Analysis Platform	540

Data Format	541
The Association Analysis Report	542
Frequent Item Sets	542
Rules	542
Association Analysis Platform Options	543
Overview of Singular Value Decomposition	544
SVD Report	545
SVD Report Options	546
Overview of Topic Analysis	547
Topic Analysis Report	547
Topic Analysis Report Options	548
Additional Example of the Association Analysis Platform	549
Statistical Details for the Association Analysis Platform	551
Statistical Details for Frequent Item Set Generation	551
Statistical Details for Performance Measures	552
27 Process History Explorer	555
Identify Problem Components in Complex Process Histories	
Example of Process History Explorer	557
Launch the Process History Explorer Platform	559
Data Format	560
The Process History Explorer Report	560
Process History Explorer Platform Options	561
A Statistical Details	565
Predictive and Specialized Modeling	
Entropy RSquare	567
Validation in JMP Modeling	567
Validation Column Role	568
K-Fold and Holdback Validation	570
Predicted Probabilities in Tree Methods	572
B References	573
C Technology License Notices	577

Chapter **1**

Learn about JMP

Documentation and Additional Resources

Learn about JMP documentation, such as book conventions, descriptions of each JMP document, the Help system, and where to find additional support.

Contents

Formatting Conventions in JMP Documentation.....	19
JMP Help	20
JMP Documentation Library	20
Additional Resources for Learning JMP	27
Search JMP	27
JMP Tutorials	27
Sample Data Tables	28
Learn about Statistical and JSL Terms	28
Learn JMP Tips and Tricks.....	28
JMP Tooltips.....	28
JMP User Community	29
Free Online Statistical Thinking Course	29
JMP New User Welcome Kit	29
Statistics Knowledge Portal.....	29
JMP Training	30
JMP Books by Users	30
The JMP Starter Window	30
JMP Technical Support.....	30

Formatting Conventions in JMP Documentation

These 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 (or sans-serif online) font.
- Code appears in *Lucida Sans Typewriter* (or monospace online) font.
- Code output appears in *Lucida Sans Typewriter* italic (or monospace italic online) font and is indented farther than the preceding code.
- **Helvetica bold** formatting (or bold sans-serif online) indicates items that you select to complete a task:
 - buttons
 - check boxes
 - commands
 - list names that are selectable
 - menus
 - options
 - tab names
 - text boxes
- The following items appear in italics:
 - words or phrases that are important or have definitions specific to JMP
 - book titles
 - variables
- Features that are for JMP Pro only are noted with the JMP Pro icon . For an overview of JMP Pro features, visit jmp.com/software/pro.

Note: Special information and limitations appear within a Note.

Tip: Helpful information appears within a Tip.

JMP Help

JMP Help in the Help menu enables you to search for information about JMP features, statistical methods, and the JMP Scripting Language (or *JSL*). You can open JMP Help in several ways:

- Search and view JMP Help on Windows by selecting **Help > JMP Help**.
- On Windows, press the F1 key to open the Help system in the default browser.
- 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.

Note: The JMP Help is available for users with Internet connections. Users without an Internet connection can search all books in a PDF file by selecting **Help > JMP Documentation Library**. See “[JMP Documentation Library](#)” for more information.

JMP Documentation Library

The Help system content is also available in one PDF file called *JMP Documentation Library*. Select **Help > JMP Documentation Library** to open the file. You can also download the Documentation PDF Files add-in if you prefer searching individual PDF files of each document in the JMP library. Download the available add-ins from <https://community.jmp.com>.

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

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

Document Title	Document Purpose	Document Content
<i>Basic Analysis</i>	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.
<i>Essential Graphing</i>	Find the ideal graph for your data.	Describes these Graph menu platforms: • Graph Builder • Scatterplot 3D • Contour Plot • Bubble Plot • Parallel Plot • Cell Plot • Scatterplot Matrix • Ternary Plot • Treemap • Chart • Overlay Plot The book also covers how to create background and custom maps.
<i>Profilers</i>	Learn how to use interactive profiling tools, which enable you to view cross-sections of any response surface.	Covers all profilers listed in the Graph menu. Analyzing noise factors is included along with running simulations using random inputs.

Document Title	Document Purpose	Document Content
<i>Design of Experiments Guide</i>	Learn how to design experiments and determine appropriate sample sizes.	Covers all topics in the DOE menu.
<i>Fitting Linear Models</i>	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 • Generalized Linear Mixed Model • MANOVA • Loglinear Variance • Nominal Logistic • Ordinal Logistic • Generalized Linear Model

Document Title	Document Purpose	Document Content
<i>Predictive and Specialized Modeling</i>	Learn about additional modeling techniques.	Describes these Analyze > Predictive Modeling menu platforms: • Neural • Partition • Bootstrap Forest • Boosted Tree • K Nearest Neighbors • Naive Bayes • Support Vector Machines • Model Comparison • Model Screening • Make Validation Column • Formula Depot Describes these Analyze > Specialized Modeling menu platforms: • Fit Curve • Nonlinear • Functional Data Explorer • Gaussian Process • Time Series • Matched Pairs Describes these Analyze > Screening menu platforms: • Explore Outliers • Explore Missing Values • Explore Patterns • Response Screening • Predictor Screening • Association Analysis • Process History Explorer

Document Title	Document Purpose	Document Content
<i>Multivariate Methods</i>	Read about techniques for analyzing several variables simultaneously.	Describes these Analyze > Multivariate Methods menu platforms: • Multivariate • Principal Components • Discriminant • Partial Least Squares • Multiple Correspondence Analysis • Structural Equation Models • Factor Analysis • Multidimensional Scaling • Multivariate Embedding • Item Analysis Describes these Analyze > Clustering menu platforms: • Hierarchical Cluster • K Means Cluster • Normal Mixtures • Latent Class Analysis • Cluster Variables

Document Title	Document Purpose	Document Content
<i>Quality and Process Methods</i>	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 (EMP and Type 1 Gauge) • Variability / Attribute Gauge Charts • Process Screening • Process Capability • Model Driven Multivariate Control Chart • Legacy Control Charts • Pareto Plot • Diagram • Manage Spec Limits • OC Curves
<i>Reliability and Survival Methods</i>	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 • Repeated Measures Degradation • Destructive Degradation • Reliability Forecast • Reliability Growth • Reliability Block Diagram • Repairable Systems Simulation • Survival • Fit Parametric Survival • Fit Proportional Hazards

Document Title	Document Purpose	Document Content
<i>Consumer Research</i>	Learn about methods for studying consumer preferences and using that insight to create better products and services.	Describes these Analyze > Consumer Research menu platforms: • Categorical • Choice • MaxDiff • Uplift • Multiple Factor Analysis
<i>Genetics</i>	Learn about methods that are available in JMP to help you analyze your genetic data and use that data to simulate a breeding program to predict the optimum genetic crosses to make.	Describes these Analyze > Genetics menu platforms: • Marker Statistics • Marker Simulation
<i>Scripting Guide</i>	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.
<i>JSL Syntax Reference</i>	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.

Additional Resources for Learning JMP

In addition to reading JMP help, you can also learn about JMP using the following resources:

- [“Search JMP”](#)
- [“JMP Tutorials”](#)
- [“Sample Data Tables”](#)
- [“Learn about Statistical and JSL Terms”](#)
- [“Learn JMP Tips and Tricks”](#)
- [“JMP Tooltips”](#)
- [“JMP User Community”](#)
- [“Free Online Statistical Thinking Course”](#)
- [“JMP New User Welcome Kit”](#)
- [“Statistics Knowledge Portal”](#)
- [“JMP Training”](#)
- [“JMP Books by Users”](#)
- [“The JMP Starter Window”](#)

Search JMP

If you are not sure where to find a statistical procedure, do a search across JMP. Results are tailored to the window that you launch the search from, such as a data table or report.

1. Click **Help > Search JMP**. Or, press Ctrl+comma.
2. Enter your search text.
3. Click the result that contains the procedure that you want.

On the right, you can see a description and the location of the procedure.

4. Click the corresponding button to open or go to a result.

JMP 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, 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 Folder** to open the sample data directory.

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

Sample data tables are installed in the following directory:

On Windows: C:\Program Files\SAS\JMP\17\Samples\Data

On macOS: \Library\Application Support\JMP\17\Samples\Data

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

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

Learn about Statistical and JSL Terms

For help with statistical terms, select **Help > Statistics Index**. For help with JSL scripting and examples, select **Help > Scripting Index**.

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 and get help on the commands.

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.

JMP Tooltips

JMP provides descriptive tooltips (or *hover labels*) when you hover 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 macOS.

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 on the Web > JMP User Community** or visit <https://community.jmp.com>.

Free Online Statistical Thinking Course

Learn practical statistical skills in this free online course on topics such as exploratory data analysis, quality methods, and correlation and regression. The course consists of short videos, demonstrations, exercises, and more. Visit jmp.com/statisticalthinking.

JMP New User Welcome Kit

The JMP New User Welcome Kit is designed to help you quickly get comfortable with the basics of JMP. You'll complete its thirty short demo videos and activities, build your confidence in using the software, and connect with the largest online community of JMP users in the world. Visit jmp.com/welcome.

Statistics Knowledge Portal

The Statistics Knowledge Portal combines concise statistical explanations with illuminating examples and graphics to help visitors establish a firm foundation upon which to build statistical skills. Visit jmp.com/skp.

JMP Training

SAS offers training on a variety of topics led by a seasoned team of JMP experts. Public courses, live web courses, and on-site courses are available. You might also choose the online e-learning subscription to learn at your convenience. Visit jmp.com/training.

JMP Books by Users

Additional books about using JMP that are written by JMP users are available on the JMP website. Visit jmp.com/books.

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 macOS) > 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 macOS, select **JMP > Preferences > Initial JMP Starter Window**.

JMP 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 jmp.com/support, including the technical support phone number.

Introduction to Predictive and Specialized Modeling

Overview of Modeling Techniques

Predictive and Specialized Modeling provides details about more technical modeling techniques, such as Response Screening, Partitioning, and Neural Networks.

- The Neural platform implements a fully connected multi-layer perceptron with one or two layers. Use neural networks to predict one or more response variables using a flexible function of the input variables. See [“Neural Networks”](#).
- The Partition platform recursively partitions data according to a relationship between the X and Y values, creating a decision tree of partitions. See [“Partition Models”](#).
- **JMP PRO** The Bootstrap Forest platform enables you to fit an ensemble model by averaging many decision trees each of which is fit to a random subset of the training data. See [“Bootstrap Forest”](#).
- **JMP PRO** The Boosted Tree platform produces an additive decision tree model that consists of many smaller decision trees that are constructed in layers. The tree in each layer consists of a small number of splits, typically five or fewer. Each layer is fit using the recursive fitting methodology. See [“Boosted Tree”](#).
- **JMP PRO** The K Nearest Neighbors platform predicts a response value for a given observation using the responses of the observations in that observation’s local neighborhood. It can be used with a categorical response for classification and with a continuous response for prediction. See [“K Nearest Neighbors”](#).
- **JMP PRO** The Naive Bayes platform classifies observations into groups that are defined by the levels of a categorical response variable. The variables (or factors) that are used for classification are often called features in the data mining literature. See [“NaiveBayes”](#).
- **JMP PRO** The Support Vector Machines platform classifies observations into groups that are defined by levels of a categorical response variable. The model classifies data by optimizing a hyperplane that separates the classes. See [“Support Vector Machines”](#).
- **JMP PRO** The Model Screening platform enables you to quickly run multiple predictive models and compare the results. Measures of fit are provided for each model along with overlaid diagnostic plots. See [“Model Screening”](#).
- **JMP PRO** The Model Comparison platform enables you to compare the predictive ability of different models. Measures of fit are provided for each model along with overlaid diagnostic plots. See [“Model Comparison”](#).

- **JMP PRO** The Make Validation Column platform lets you partition the data into two or three sets, using one of five different methods to create these partitions. See [“Make Validation Column”](#).
- **JMP PRO** The Formula Depot platform enables you to organize, compare, profile, and score models for deployment. For model exploration work, you can use the Formula Depot to store candidate models outside of your JMP data table. See [“Formula Depot”](#).
- The Fit Curve platform provides predefined models, such as polynomial, logistic, Gompertz, exponential, peak, and pharmacokinetic models. Compare different groups or subjects using a variety of analytical and graphical techniques. See [“Fit Curve”](#).
- The Nonlinear platform lets you fit custom nonlinear models, which include a model formula and parameters to be estimated. See [“Nonlinear Regression”](#).
- The Gaussian Process platform models the relationship between a continuous response and one or more continuous predictors. These models are common in areas like computer simulation experiments, such as the output of finite element codes, and they often perfectly interpolate the data. See [“Gaussian Process”](#).
- **JMP PRO** The Functional Data Explorer platform enables you to convert functional data into a form that can be analyzed in another JMP platform. See [“Functional Data Explorer”](#).
- The Time Series platform lets you explore, analyze, and forecast univariate time series. See [“Time Series Analysis”](#).
- The Time Series Forecast platform lets you model and forecast multiple time series. The best fitting model is automatically selected from a set of up to 30 exponential smoothing models. See [“Time Series Forecast”](#).
- The Matched Pairs platform compares the means between two or more correlated variables and assesses the differences. See [“Matched Pairs Analysis”](#).
- The Explore Outliers platform enables you to identify, explore, and manage outliers in both univariate and multivariate data. See [“Explore Outliers”](#).
- The Explore Missing Values platform provides several ways to identify, understand, and impute missing values in your data. See [“Explore Missing Values”](#).
- The Explore Patterns platform is used to detect unusual or unexpected patterns in data. See [“Explore Patterns”](#).
- The Response Screening platform automates the process of conducting tests across a large number of responses. Your test results and summary statistics are presented in data tables, rather than reports, to enable data exploration. See [“Response Screening”](#).
- The Predictor Screening platform enables you to screen a data set for significant predictors. See [“Predictor Screening”](#).
- **JMP PRO** The Association Analysis platform enables you to identify items that have an affinity for each other. It is frequently used to analyze transactional data (also called market

- baskets) to identify items that often appear together in transactions. See [“Association Analysis”](#).
- The Process History Explorer platform enables you to identify problem components in complex process histories. See [“Process History Explorer”](#).

Chapter 3

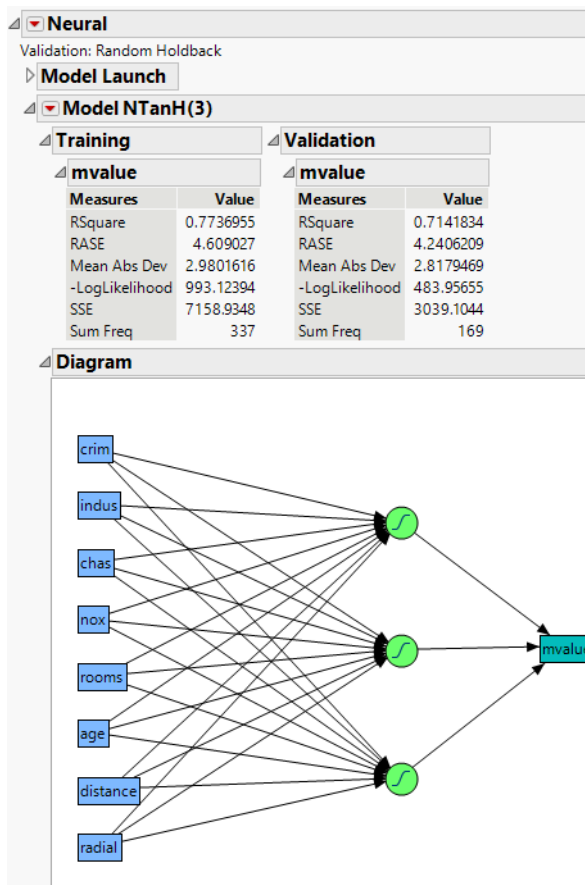
Neural Networks

Fit Nonlinear Models Using Nodes and Layers

Most features in this platform are available only in JMP Pro and noted with this icon.

The Neural platform implements a fully connected multi-layer perceptron with one or two layers. Use neural networks to predict one or more response variables using a flexible function of the input variables. Neural networks can be very good predictors when it is not necessary to describe the functional form of the response surface, or to describe the relationship between the inputs and the response.

Figure 3.1 Example of a Neural Network



Contents

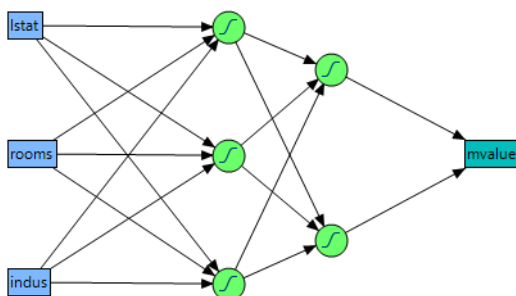
Overview of the Neural Platform	37
Example of a Neural Network	37
Launch the Neural Platform	40
The Neural Launch Window	40
The Model Launch Control Panel	42
Validation Methods for Neural	43
Hidden Layer Structure	44
Boosting	45
Neural Fitting Options	45
Model Reports	47
Training and Validation Measures of Fit	47
Confusion Statistics	48
Neural Platform Options	48
Neural Model Options	49

Overview of the Neural Platform

Neural networks are a function of a set of derived inputs, called hidden nodes. The hidden nodes are nonlinear functions of the original inputs. You can specify up to two layers of hidden nodes, where each layer can contain as many hidden nodes as you want.

Figure 3.2 shows a two-layer neural network with three X variables and one Y variable. In this example, the first layer has two nodes, and each node is a function of all three nodes in the second layer. The second layer has three nodes, and all nodes are a function of the three X variables. The predicted Y variable is a function of both nodes in the first layer.

Figure 3.2 Neural Network Diagram



The functions applied at the nodes of the hidden layers are called activation functions. The activation function is a transformation of a linear combination of the X variables. For more information about the activation functions, see [“Hidden Layer Structure”](#). The function applied at the response is a linear combination (for continuous responses), or a logistic transformation (for nominal or ordinal responses).

The main advantage of a neural network model is that it can efficiently model different response surfaces. Given enough hidden nodes and layers, any surface can be approximated to any accuracy. The main disadvantage of a neural network model is that the results are not easily interpretable. This is because there are intermediate layers rather than a direct path from the X variables to the Y variables, as in the case of regular regression.

Example of a Neural Network

Suppose you want to create a model to predict the median home value as a function of several demographic characteristics. Follow the steps below to build the neural network model:

1. Select **Help > Sample Data Folder** and Boston Housing.jmp.
2. Launch the Neural platform by selecting **Analyze > Predictive Modeling > Neural**.

- 3. Assign mvalue to the **Y, Response** role.
- 4. Assign the other columns (crim through lstat) to the **X, Factor** role.
- 5. Click **OK**.
- 6. Enter 0.2 for the **Holdback Proportion**.
- 7. Enter 1234 for the **Random Seed**.

Note: In general, results vary due to the random nature of choosing a validation set. Entering the seed above enable you to reproduce the results shown in this example.

- 8. Enter 3 for the number of TanH nodes in the first layer.
- 9. Check the **Transform Covariates** option.
- 10. Click **Go**.

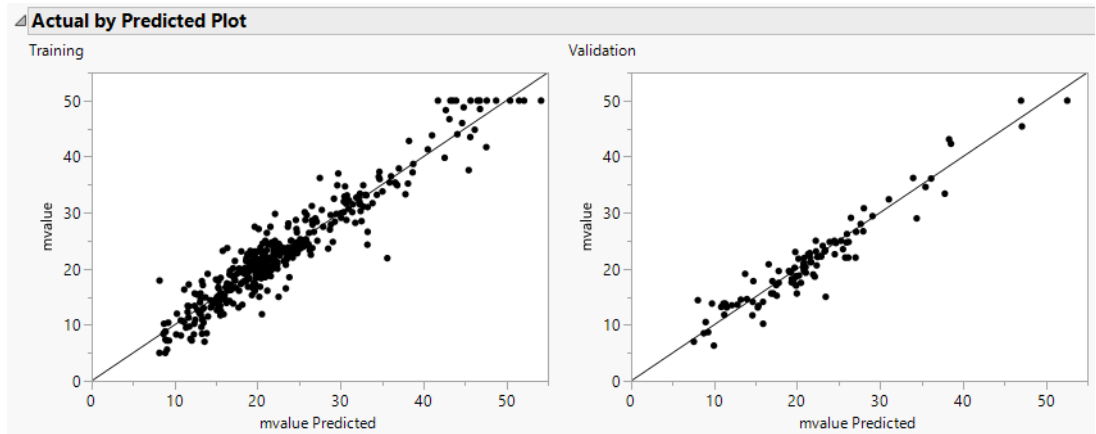
Figure 3.3 Neural Report

Model NTanH(3)			
Training		Validation	
mvalue		mvalue	
Measures	Value	Measures	Value
RSquare	0.907698	RSquare	0.9124801
RASE	2.8528693	RASE	2.4320654
Mean Abs Dev	2.125991	Mean Abs Dev	1.8299258
-LogLikelihood	996.77457	-LogLikelihood	235.3833
SSE	3288.1007	SSE	603.32408
Sum Freq	404	Sum Freq	102

Results are provided for both the training and validation sets. Use the results of the validation set as a representation of the model’s predictive power on future observations.

The R-Square statistic for the Validation set is 0.913, signifying that the model is predicting well on data not used to train the model. As an additional assessment of model fit, click the red triangle next to Model NTanH(3) and select **Plot Actual by Predicted**.

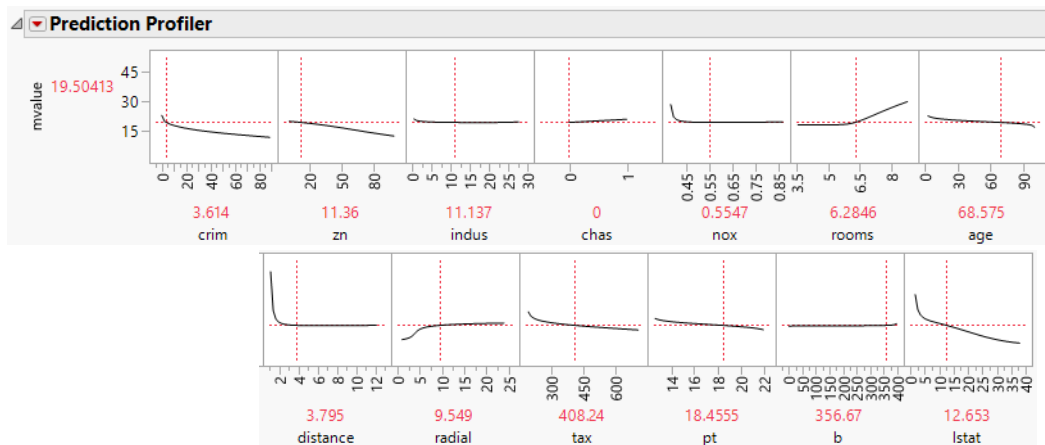
Figure 3.4 Actual by Predicted Plot



The points fall along the line, signifying that the predicted values are similar to the actual values.

To get a general understanding of how the X variables are impacting the predicted values, click the red triangle next to Model NTanH(3) and select **Profiler**.

Figure 3.5 Profiler



Some of the variables have profiles with positive slopes, and some negative. For example, the variable rooms has a positive slope. This indicates that the more rooms a home has, the higher the predicted median value. The variable pt is the pupil teacher ratio by town. This variable has a negative slope, indicating that the higher the pupil to teacher ratio, the lower the median value.

Launch the Neural Platform

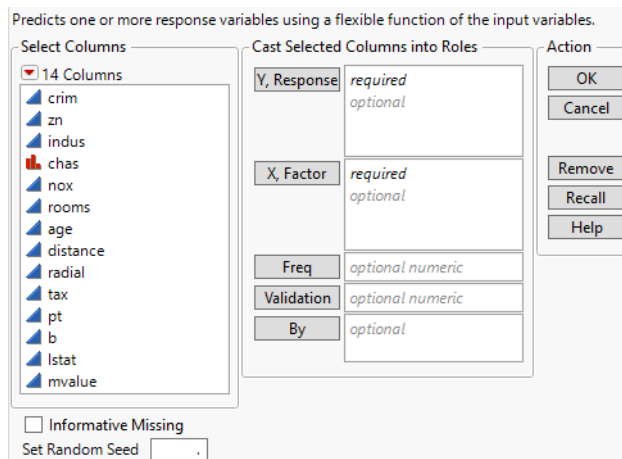
To launch the Neural platform, select **Analyze > Predictive Modeling > Neural**.

Launching the Neural platform is a two-step process. First, enter your variables on the Neural launch window. Second, specify your options in the Model Launch control panel.

The Neural Launch Window

Use the Neural launch window to specify X and Y variables, a validation column, and to enable Informative Missing value coding.

Figure 3.6 The Neural Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Response The response variable or variables that you want to analyze. When multiple responses are specified, the models for the responses share all parameters in the hidden layers (those parameters not connected to the responses). The responses are conditionally independent given the predictor variables, but are marginally correlated through those variables to create one overall neural model.

X, Factor The predictor variables.

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

 **Validation** A numeric column that defines the validation sets. See [“Validation Methods for Neural”](#). 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 [“Make Validation Column”](#).

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.

JMP[®] PRO **Informative Missing** Check this box to enable informative coding of missing values. This coding allows estimation of a predictive model despite the presence of missing values. It is useful in situations where missing data are informative. If this option is not checked, rows with missing values are ignored.

For a continuous variable, missing values are replaced by the mean of the variable. Also, a missing value indicator, named <colname> Is Missing, is created and included in the model. If a variable is transformed using the Transform Covariates fitting option on the Model Launch control panel, missing values are replaced by the mean of the transformed variable.

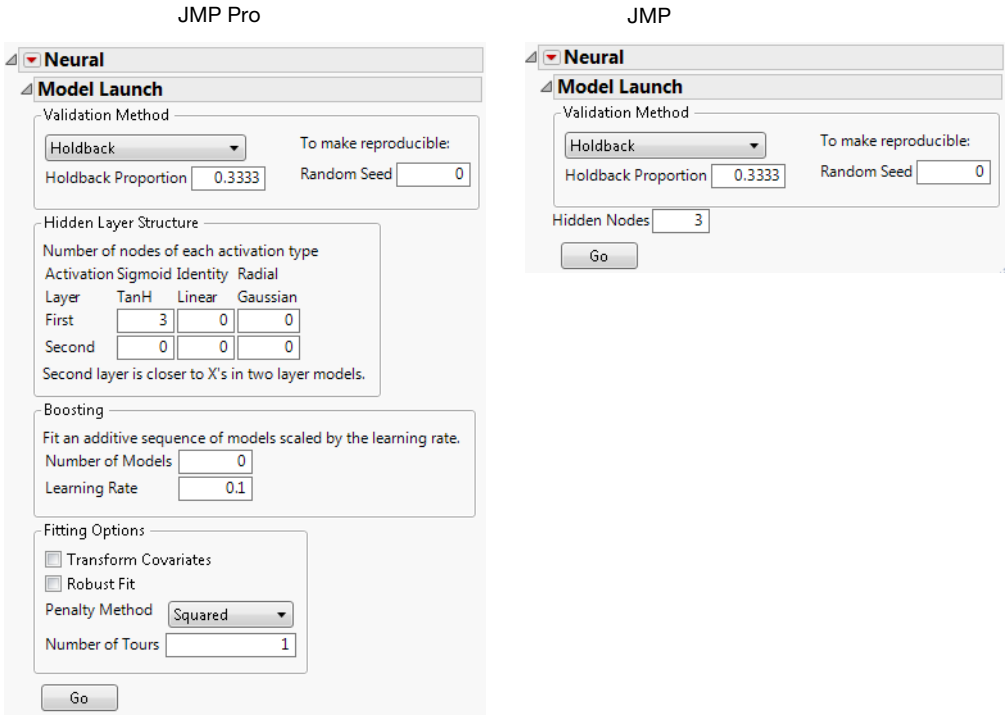
For a categorical variable, missing values are treated as a separate level of that variable.

Set Random Seed Sets the seed for the starting values used in the fitting procedure. If you do not specify a validation column, this value is automatically assigned as the Random Seed in the Model Launch control panel and is also used for validation assignment. Set Random Seed is useful if you want to reproduce an analysis. If you set a random seed and save the script, the seed is automatically saved in the script.

The Model Launch Control Panel

Use the Model Launch control panel in the Neural platform to specify the validation method, the structure of the hidden layer, whether to use gradient boosting, and other fitting options.

Figure 3.7 The Model Launch Control Panel



Validation Method Select the method that you want to use for model validation. See [“Validation Methods for Neural”](#).

Random Seed Specify a nonzero numeric random seed if you want to reproduce the starting values and validation assignment for future launches of the Neural platform. By default, the Random Seed is set to zero, which does not produce reproducible results. When you save the analysis to a script, the random seed that you enter is saved to the script.

Hidden Layer Structure or Hidden Nodes Specify the number of hidden nodes of each type in each layer. See [“Hidden Layer Structure”](#).

Note: The standard edition of JMP uses only the TanH activation function, and can fit only neural networks with one hidden layer.

JMP PRO Boosting Specify options for gradient boosting. See [“Boosting”](#).

JMP PRO Fitting Options Specify options for variable transformation and model fitting. See “Neural Fitting Options”.

Go Fits the neural network model and shows the model reports.

After you click **Go** to fit a model, you can reopen the Model Launch Control Panel and change the settings to fit another model.

Validation Methods for Neural

The Neural platform utilizes a validation method in its fitting routine. Choices of validation method include holdback, K-fold, or (in JMP Pro) the use of a validation column. To fit a model, the Neural platform does the following:

- applies a penalty on the model parameters
- uses the validation set to tune the penalties on the parameters.

Select one of the following validation methods:

Excluded Rows Holdback Uses row states to subset the data. Rows that are unexcluded are used as the training set, and excluded rows are used as the validation set.

For more information about using row states and how to exclude rows, see *Using JMP*.

Holdback Randomly divides the original data into the training and validation sets. You specify the proportion of the original data to use as the validation set (holdback). The random selection is based on stratified sampling across the model factors to attempt to create training and validation sets that are more balanced than ones based on simple random sampling.

KFold Divides the data into K subsets, or folds. In turn, each of the K folds is used to validate the model fit on the rest of the data, fitting a total of K models. The final model is for the fold that resulted in the model with the best validation statistic.

This method is useful for small data sets, because it makes efficient use of limited amounts of data.

JMP PRO Validation Column Uses a numeric column that defines the validation sets. The column's values determine how the data is split:

- If the validation column has two levels, the smaller value defines the training set and the larger value defines the validation set.
- If the validation column has three levels, the values, in order of increasing size, define the training, validation, and test sets.
- If the validation column has more than three unique values, then K-Fold validation is performed. The number of folds is determined by the number of values in the validation column.

The Neural platform uses the validation column to train and tune the model or to train, tune, and evaluate the model. For more information about validation, see [“Validation in JMP Modeling”](#).

Hidden Layer Structure

The Neural platform can fit one or two-layer neural networks. Increasing the number of nodes in the first layer, or adding a second layer, makes the neural network more flexible. You can add an unlimited number of nodes to either layer. The second layer nodes are functions of the X variables. The first layer nodes are functions of the second layer nodes. The Y variables are functions of the first layer nodes.

Note: The standard edition of JMP uses only the TanH activation function, and can fit only neural networks with one hidden layer.

Caution: You cannot use boosting with a two-layer neural network. If you specify any non-zero values in the second layer and also specify boosting, the second layer is ignored.

The functions applied at the nodes of the hidden layers are called activation functions. An activation function is a transformation of a linear combination of the X variables. The following activation functions are available:

TanH The hyperbolic tangent function is a sigmoid function. TanH transforms values to be between -1 and 1, and is the centered and scaled version of the logistic function. The hyperbolic tangent function is:

$$\frac{e^{2x} - 1}{e^{2x} + 1}$$

where x is a linear combination of the X variables.

Linear The identity function. The linear combination of the X variables is not transformed.

The Linear activation function is most often used in conjunction with one of the non-linear activation functions. In this case, the Linear activation function is placed in the second layer, and the non-linear activation functions are placed in the first layer. This is useful if you want to first reduce the dimensionality of the X variables, and then have a nonlinear model for the Y variables.

For a continuous Y variable, if only Linear activation functions are used, the model for the Y variable reduces to a linear combination of the X variables. For a nominal or ordinal Y variable, the model reduces to a logistic regression.

Gaussian The Gaussian function. Use this option for radial basis function behavior, or when the response surface is Gaussian (normal) in shape. The Gaussian function is:

$$e^{-x^2}$$

where x is a linear combination of the X variables.

Boosting

Boosting is the process of building a large additive neural network model by fitting a sequence of smaller models. Each of the smaller models is fit on the scaled residuals of the previous model. The models are combined to form the larger final model. The process uses validation to assess how many component models to fit, not exceeding the specified number of models.

Boosting is often faster than fitting a single large model. However, the base model should be a 1 to 2 node single-layer model. The benefit of faster fitting can be lost if a large number of models is specified.

Use the Boosting panel in the Model Launch control panel to specify the number of component models and the learning rate. Use the Hidden Layer Structure panel in the Model Launch control panel to specify the structure of the base model.

The learning rate must be $0 < r \leq 1$. Learning rates close to 1 result in faster convergence on a final model, but also have a higher tendency to overfit data. Use learning rates close to 1 when a small Number of Models is specified.

As an example of how boosting works, suppose you specify a base model consisting of one layer and two nodes, with the number of models equal to eight. The first step is to fit a one-layer, two-node model. The predicted values from that model are scaled by the learning rate, then subtracted from the actual values to form a scaled residual. The next step is to fit a different one-layer, two-node model, where the response values are the scaled residuals of the previous model. This process continues until eight models are fit, or until the addition of another model fails to improve the validation statistic. The component models are combined to form the final, large model. In this example, if six models are fit before stopping, the final model consists of one layer and $2 \times 6 = 12$ nodes.

Caution: You cannot use boosting with a two-layer neural network. If you specify any non-zero values in the second layer and also specify boosting, the second layer is ignored.

Neural Fitting Options

The following model fitting options are available:

Transform Covariates Transforms all continuous variables to near normality using either the Johnson Su or Johnson Sb distribution. Transforming the continuous variables helps

mitigate the negative effects of outliers or heavily skewed distributions. See the Save Transformed Covariates option in [“Neural Model Options”](#).

Robust Fit Trains the model using least absolute deviations instead of least squares. This option is useful if you want to minimize the impact of response outliers. This option is available only for continuous responses.

Penalty Method Choose the penalty method. To mitigate the tendency neural networks have to overfit data, the fitting process incorporates a penalty on the likelihood. See [“Penalty Method”](#).

Number of Tours Specify the number of times to restart the fitting process, with each iteration using different random starting points for the parameter estimates. The iteration with the best validation statistic is chosen as the final model.

Penalty Method

The penalty is $\lambda p(\beta_i)$, where λ is the penalty parameter, and $p(\)$ is a function of the parameter estimates, called the penalty function. Validation is used to find the optimal value of the penalty parameter.

Table 3.1 Descriptions of Penalty Methods

Method	Penalty Function	Description
Squared	$\sum \beta_i^2$	Use this method if you think that most of your X variables are contributing to the predictive ability of the model.
Absolute	$\sum \beta_i $	Use either of these methods if you have a large number of X variables, and you think that a few of them contribute more than others to the predictive ability of the model.
Weight Decay	$\sum \frac{\beta_i^2}{1 + \beta_i^2}$	
NoPenalty	none	Does not use a penalty. You can use this option if you have a large amount of data and you want the fitting process to go quickly. However, this option can lead to models with lower predictive performance than models that use a penalty.

RSquare Gives the RSquare for the model.

RASE Gives the root average squared error. When the response is nominal or ordinal, the differences are between 1 and p (the fitted probability for the response level that actually occurred).

Mean Abs Dev The average of the absolute values of the differences between the response and the predicted response. When the response is nominal or ordinal, the differences are between 1 and p (the fitted probability for the response level that actually occurred).

Misclassification Rate The rate for which the response category with the highest fitted probability is not the observed category. Appears only when the response is nominal or ordinal.

-LogLikelihood Gives the negative of the log-likelihood. See *Fitting Linear Models*.

SSE Gives the error sums of squares. Available only when the response is continuous.

Sum Freq Gives the number of observations that are used. If you specified a Freq variable in the Neural launch window, Sum Freq gives the sum of the frequency column.

If there are multiple responses, fit statistics are given for each response, and an overall Generalized RSquare and negative Log-Likelihood is given.

Confusion Statistics

When nominal or ordinal responses are fit in the Neural platform, a Confusion Matrix report and Confusion Rates report are given ([Figure 3.8](#)). The Confusion Matrix report shows a two-way classification of the actual response levels and the predicted response levels. For a categorical response, the predicted level is the one with the highest predicted probability. The Confusion Rates report is equal to the Confusion Matrix report, with the numbers divided by the row totals.

Neural Platform Options

The Neural red triangle menu contains the following options:

See *Using JMP* 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.

- Platform Preferences** Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.
- 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.

Neural Model Options

Each model report has a red triangle menu that contains options for producing additional output or saving results. The model report red triangle menu contains the following options:

- Diagram** Shows a diagram representing the hidden layer structure.
- Show Estimates** Shows the parameter estimates in a report.
- Profiler** Launches the Prediction Profiler. For nominal or ordinal responses, each response level is represented by a separate row in the Prediction Profiler. For more information about the options in the red triangle menu, see *Profilers*.
- Categorical Profiler** Launches the Prediction Profiler. Similar to the Profiler option, except that all categorical probabilities are combined into a single profiler row. Available only for nominal or ordinal responses. For more information about the options in the red triangle menu, see *Profilers*.
- Contour Profiler** Launches the Contour Profiler. This is available only when the model contains more than one continuous factor. For more information about the options in the red triangle menu, see *Profilers*.
- Surface Profiler** Launches the Surface Profiler. This is available only when the model contains more than one continuous factor. For more information about the options in the red triangle menu, see *Profilers*.
- ROC Curve** Creates an ROC curve. Available only for nominal or ordinal responses. For more information about ROC Curves, see [“ROC Curve”](#).
- Lift Curve** Creates a lift curve. Available only for nominal or ordinal responses. For more information about Lift Curves, see [“Lift Curve”](#).
- Decision Threshold** (Available only for binary categorical responses.) Shows or hides Decision Thresholds reports for the training, validation, and test sets, if specified. Each report contains a graph of the distribution of fitted probabilities for each model, confusion

matrices for each model, classification graphs to compare the model fits, and a table of classification accuracy metrics. See [“Decision Thresholds Report”](#).

Plot Actual by Predicted Plots the actual versus the predicted response. Available only for continuous responses.

Plot Residual by Predicted Plots the residuals versus the predicted responses. Available only for continuous responses.

Save Formulas Creates new columns in the data table containing formulas for the predicted response and the hidden layer nodes.

Save Profile Formulas Creates new columns in the data table containing formulas for the predicted response. Formulas for the hidden layer nodes are embedded in this formula. This option produces formulas that can be used by the Interactive HTML version of the Profiler.

Save Fast Formulas Creates new columns in the data table containing formulas for the predicted response. Formulas for the hidden layer nodes are embedded in this formula. This option produces formulas that evaluate faster than the other options, but cannot be used in the Interactive version of the Profiler.

JMP[®] PRO Publish Prediction Formula Creates prediction formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See [“Formula Depot”](#).

Make SAS Data Step Creates SAS code that you can use to score a new data set.

Save Validation Creates a new column in the data table that identifies which rows were used in the training and validation sets. This option is not available when a Validation column is specified on the Neural launch window. See [“The Neural Launch Window”](#).

JMP[®] PRO Save Transformed Covariates Creates new columns in the data table showing the transformed covariates. The columns contain formulas that show the transformations. This option is available only when the Transform Covariates option is checked on the Model Launch control panel. See [“Neural Fitting Options”](#).

Remove Fit Removes the entire model report.

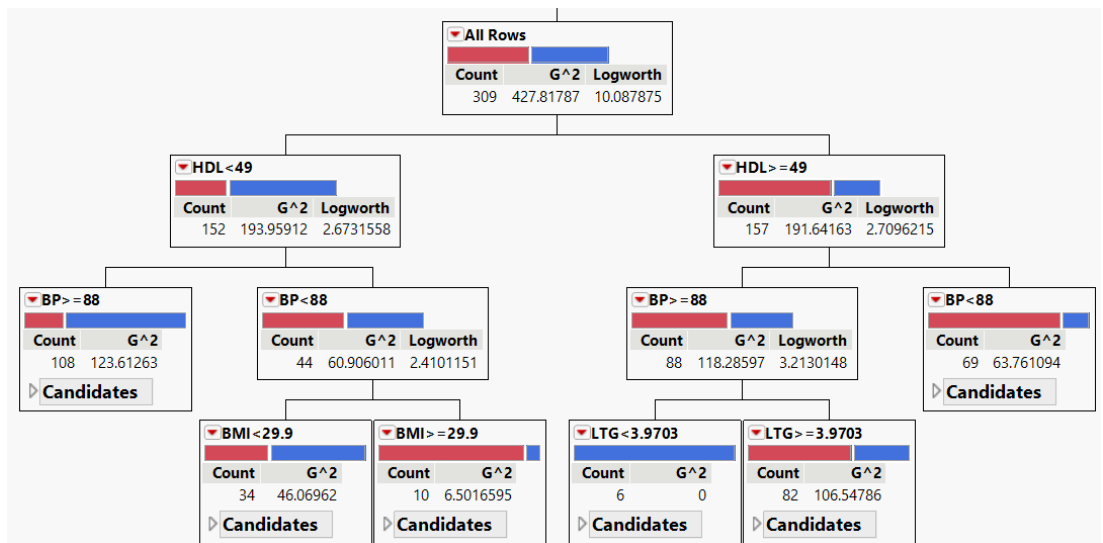
Chapter 4

Partition Models

Use Decision Trees to Explore and Model Your Data

The Partition platform recursively partitions data according to a relationship between the predictors and response values, creating a decision tree. The partition algorithm searches all possible splits of predictors to best predict the response. These splits (or *partitions*) of the data are done recursively to form a tree of decision rules. The splits continue until the desired fit is reached. The partition algorithm chooses optimum splits from a large number of possible splits, making it a powerful modeling, and data discovery tool.

Figure 4.1 Example of a Decision Tree



Contents

Overview of the Partition Platform	53
Example of the Partition Platform	53
Launch the Partition Platform	56
Informative Missing	57
The Partition Report	58
Partition Plot	58
Partition Buttons	60
Summary of Fit Statistics	60
Node Reports	61
Partition Platform Options	63
Show Fit Details	66
Specify Profit Matrix	68
ROC Curve	69
Lift Curve	71
Validation in Partition	72
Additional Examples of the Partition Platform	73
Example of a Continuous Response	73
Example of Informative Missing	76
Example of Profit Matrix and Decision Matrix Report	79
Statistical Details for the Partition Platform	82
Statistical Details for Responses and Factors	82
Statistical Details for the Splitting Criterion	83

Overview of the Partition Platform

The Partition platform recursively partitions data according to a relationship between the predictors and response values, creating a decision tree. Variations of partitioning go by many names and brand names: decision trees, CARTTM, CHAIDTM, C4.5, C5, and others. The technique is often considered as a data mining technique for the following reasons:

- it is useful for exploring relationships without having a good prior model
- it handles large problems easily
- the results are interpretable

A classic application of partitioning is to create a diagnostic heuristic for a disease. Given symptoms and outcomes for a number of subjects, partitioning can be used to generate a hierarchy of questions to help diagnose new patients.

Predictors can be either continuous or categorical (nominal or ordinal). If a predictor is continuous, then the splits are created by a *cutting value*. The sample is divided into values below and above this cutting value. If a predictor is categorical, then the sample is divided into two groups of levels.

The response can also be either continuous or categorical (nominal or ordinal). If the response is continuous, then the platform fits the means of the response values and the split is chosen to minimize the sum of squared errors. If the response is categorical, then the fitted value is a probability for the levels of the response and the split is chosen to minimize the residual log-likelihood chi-square.

For more information about split criteria, see [“Statistical Details for the Partition Platform”](#).

For more information about recursive partitioning, see Hawkins and Kass (1982) and Kass (1980).

Example of the Partition Platform

In this example, you use the Partition platform to construct a decision tree that predicts the one-year disease progression (low or high) of patients with diabetes.

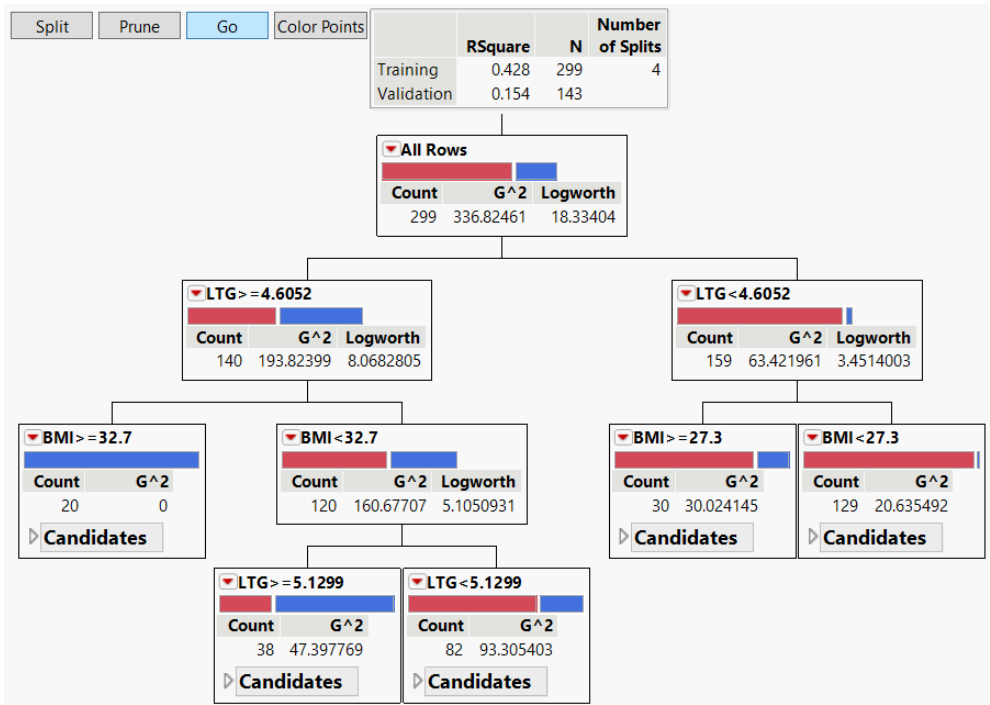
1. Select **Help > Sample Data Folder** and Diabetes.jmp.
2. Select **Analyze > Predictive Modeling > Partition**.
3. Select Y Binary and click **Y, Response**.
4. Select Age through Glucose and click **X, Factor**.
5. Enter 0.33 for the **Validation Portion**.

Note: In JMP Pro, a validation column can be used for validation. Select Validation and click **Validation**. Set the **Validation Portion** to 0.

- 6. Click **OK**.
- 7. On the platform report window, click **Go** to perform automatic splitting.

Note: Because you are using a random Validation Portion, your results differ from those in [Figure 4.2](#).

Figure 4.2 Partition Report for Diabetes



Automatic splitting resulted in four splits. The final RSquare for the Validation set is 0.154. The decision tree shows the four splits and the counts of observations in each split.

- 8. Click the red triangle next to Partition for Y Binary and select **Column Contributions**.

Figure 4.3 Column Contributions Report

Column Contributions				
Term	Number of Splits	G ²		Portion
LTG	2	99.5525528		0.6844
BMI	2	45.9092475		0.3156
Age	0	0		0.0000
Gender	0	0		0.0000
BP	0	0		0.0000
Total Cholesterol	0	0		0.0000
LDL	0	0		0.0000
HDL	0	0		0.0000
TCH	0	0		0.0000
Glucose	0	0		0.0000

The Column Contributions report shows that LTG and BMI are the only predictors in the decision tree model. Each column is used in two splits. Your results can differ. When the Validation Portion is used, the validation set is selected at random from the data table. If you redo your analysis, a new random validation set is selected and your results can differ from your first run.

- Click the red triangle next to Partition for Y Binary and select **Profiler**.

Figure 4.4 Profiler for Partition Model



The profiler enables you to change the values of BMI and LTG to obtain the predicted Y Binary outcome. No other factors had a split in the partition model. Their profiles are flat lines.

- Click the red triangle next to Partition for Y Binary and select **Save Columns > Save Prediction Formula**.

In the Diabetes.jmp data table, columns called Prob(Y Binary==Low), Prob(Y Binary==High), and Most Likely Y Binary are added. To see how these response probabilities are calculated, in the Columns panel, next to each column, double-click the Formula icon .

Launch the Partition Platform

Launch the Partition platform by selecting **Analyze > Predictive Modeling > Partition**.

Figure 4.5 Partition Launch Window

Builds a decision tree to predict a response.

Select Columns	Cast Selected Columns into Roles	Action												
14 Columns Y Y Binary Y Ordinal Age Gender BMI BP Total Cholesterol LDL HDL TCH LTG Glucose Validation	<table border="1"> <tr> <td>Y, Response</td> <td>required optional</td> </tr> <tr> <td>X, Factor</td> <td>required optional</td> </tr> <tr> <td>Weight</td> <td>optional numeric</td> </tr> <tr> <td>Freq</td> <td>optional numeric</td> </tr> <tr> <td>Validation</td> <td>optional numeric</td> </tr> <tr> <td>By</td> <td>optional</td> </tr> </table>	Y, Response	required optional	X, Factor	required optional	Weight	optional numeric	Freq	optional numeric	Validation	optional numeric	By	optional	OK Cancel Remove Recall Help
Y, Response	required optional													
X, Factor	required optional													
Weight	optional numeric													
Freq	optional numeric													
Validation	optional numeric													
By	optional													

Options

Method: Decision Tree

Validation Portion: 0.3

☒ Informative Missing

☒ Ordinal Restricts Order

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Response The response variable or variables that you want to analyze.

X, Factor The predictor variables.

Weight A column whose numeric values assign a weight to each row in the analysis.

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

JMP PRO Validation A numeric column that defines the validation sets. See [“Validation in Partition”](#). 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 [“Make Validation Column”](#).

By A column or columns whose levels define separate analyses. For each level of the column, the corresponding rows are analyzed using the other variables that you specify. The results appear 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.

JMP PRO Method Enables you to select the partition method (Bootstrap Forest, Boosted Tree, K Nearest Neighbors, or Naive Bayes).

For more information about these methods, see [“Bootstrap Forest”](#), [“Boosted Tree”](#), [“K Nearest Neighbors”](#), and [“NaiveBayes”](#).

Validation Portion The portion of the data to be used as the validation set. See [“Validation in Partition”](#).

Informative Missing If selected, enables missing value categorization for categorical predictors and informative treatment of missing values for continuous predictors. See [“Informative Missing”](#).

Ordinal Restricts Order If selected, restricts consideration of splits to those that preserve the ordering.

Informative Missing

The Informative Missing option in the Partition platform enables informative treatment of missing values on the predictors. The model that is fit is deterministic. The Informative Missing option is found on the launch window and is selected by default. When the Informative Missing option is selected, categorical and continuous predictors are handled differently:

- Rows containing missing values for a categorical predictor are entered into the analysis as a separate level of the variable.
- Rows containing missing values for a continuous predictor are assigned to a split as follows: The values of the continuous predictor are sorted. Missing rows are first considered to be on the low end of the sorted values. All splits are constructed. The missing rows are then considered to be on the high end of the sorted values. Again, all splits are constructed. The optimal split is determined using the logworth criterion. For further splits on the given predictor, the algorithm commits the missing rows to high or low values, as determined by the first split induced by that predictor.

If the Informative Missing option is not selected, the missing values are handled as follows:

- When a predictor with missing values is used as a splitting variable, each row with a missing value on that predictor is randomly assigned to one of the two sides of the split.
- The first time a predictor with missing values is used as a splitting variable an *Imputes* column is added to the Summary Report showing the number of imputations. As additional imputations are made, the Imputes column updates, where five imputations were performed.

Note: The number of Imputes can be greater than the number of rows that contain missing values. The imputation occurs at each split. A row with missing values can be randomly assigned multiple times. Each time a row is randomly assigned it increments the imputation count.

The Partition Report

The initial Partition report shows a partition plot, control buttons, a summary panel, and a decision tree. The partition plot and decision tree are initialized without any splits. The reports details are different for categorical and continuous responses.

- [“Partition Plot”](#)
- [“Partition Buttons”](#)
- [“Summary of Fit Statistics”](#)
- [“Node Reports”](#)

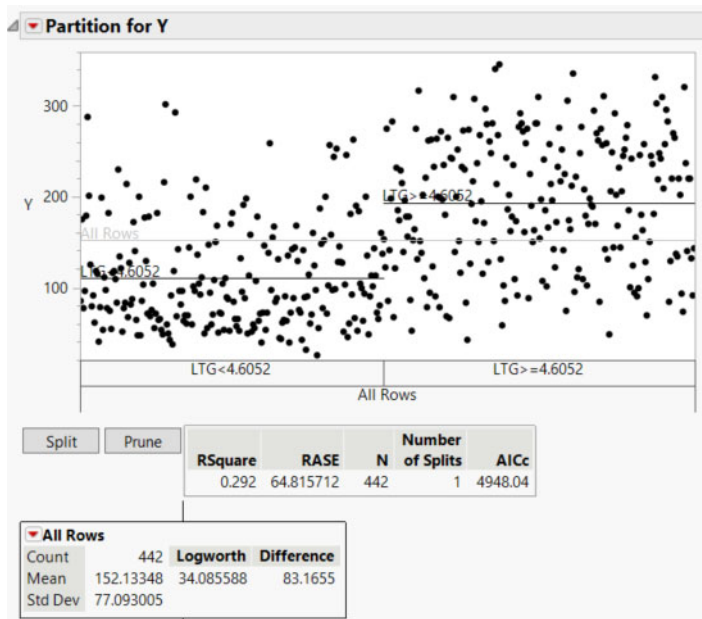
Partition Plot

Each point in the Partition Plot represents an observation in the data table. If validation is used, the plot shows only observations in the training data. The initial partition plot does not show splits. The appearance of the Partition plot is different for continuous and categorical responses.

Partition Plot for a Continuous Response

If the response is continuous, the left vertical axis represents the values of the observations. There is an initial horizontal line on the plot that is the overall mean of the response. Once you begin splitting, additional horizontal lines are added that show the mean response value for each node of the decision tree. Splits are shown below the horizontal axis with a text description and a vertical line. Observations are reorganized into their respective nodes as splits are created or removed. The most recent split appears directly below the horizontal axis and on top of existing splits. The plot is updated with each split or prune of the decision tree.

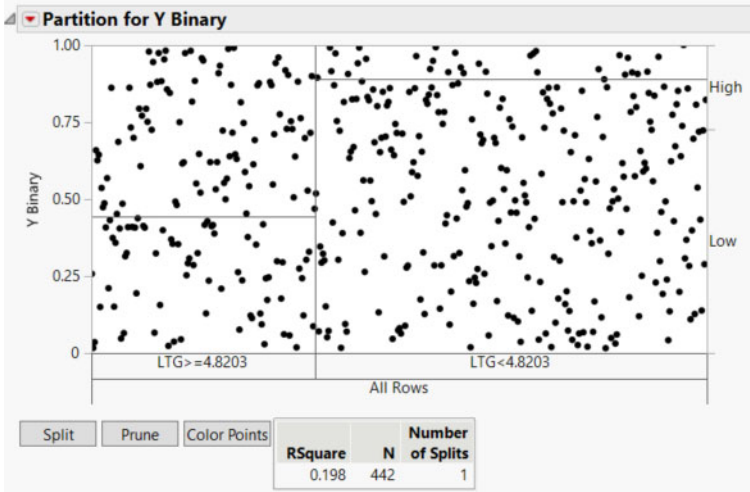
Figure 4.6 Partition Report for a Continuous Response



Partition Plot for a Categorical Response

If the response is categorical, the left vertical axis represents the proportion of each response outcome and the right vertical axis shows the order in which the response levels are plotted. There is an initial horizontal line on the plot that is the overall proportion of the first plotted response level. Once you begin splitting, additional horizontal lines are added that show the proportion of the first plotted response level for each node of the decision tree. Splits are shown below the horizontal axis with a text description and a vertical line that splits the observations in the plot. The vertical lines extend into the plot and indicate the boundaries for each node. The most recent split appears directly below the horizontal axis and on top of existing splits. The plot is updated with each split or prune of the decision tree.

Figure 4.7 Partition Report for a Categorical Response



Partition Buttons

Use the buttons in the Partition report to interact with the decision tree.

Split Creates a partition of the data using the optimal split. To specify multiple splits, hold the Shift key as you click **Split**.

Prune Removes the worst split.

Go (Available only when validation is used.) Automatically adds splits to the decision tree until the validation statistic is optimized. See “[Validation in Partition](#)”. Without validation, you simply decide the number of splits to use in the partition model.

Color Points (Available only for categorical responses.) Colors observations according to response level. These colors are added to the data table.

Summary of Fit Statistics

The Summary section in the Partition report provides fit statistics for the training data and validation and test data (if used). The fit statistics in the Summary Panel update as you add splits or prune the decision tree.

RSquare The current value of R^2 .

N The number of observations.

Number of Splits The current number of splits in the decision tree.

RASE (Available only for continuous responses.) The root average square error. See *Fitting Linear Models*.

AICc (Available only for continuous responses.) The corrected Akaike's Information Criterion. See *Fitting Linear Models*.

Node Reports

Each node in the Partition report tree has a report and a red triangle menu with additional options. See [“Node Options”](#). Terminal nodes also have a Candidates report.

Count Number of training observations that are characterized by the node.

Mean (Available only for continuous responses.) The average response for all observations in that node.

Std Dev (Available only for continuous responses.) The standard deviation of the response for all observations in that node.

G² (Available only for categorical responses.) A fit statistic used for categorical responses (instead of sum of squares that is used for continuous responses). Lower values indicate a better fit. See [“Statistical Details for the Partition Platform”](#).

Logworth The Logworth statistic, defined as $-\log_{10}(p\text{-value})$. The optimal split is the one that maximizes the logworth. See [“Statistical Details for the Partition Platform”](#)

Difference (Available only for continuous responses.) The difference in the mean of the response between the two nodes.

Candidates For each column, the Candidates report provides details about the optimal split for that column. The optimal split over all terms is marked with an asterisk.

Term Shows the candidate columns.

Candidate G² (Available only for categorical responses.) Likelihood ratio chi-square for the best split. Splitting on the predictor with the largest G² maximizes the reduction in the model G².

Candidate SS (Available only for continuous responses.) Sum of squares for the best split.

Logworth The Logworth statistic, defined as $-\log_{10}(p\text{-value})$. The optimal split is the one that maximizes the logworth. See [“Statistical Details for the Partition Platform”](#).

Cut Point The value of the predictor that determines the split. For a categorical term, the levels in the left-most split are listed.

The optimal split is noted by an asterisk. However, there are cases where the Candidate G^2 or the Candidate SS is higher for one variable, but the logworth is higher for a different variable. In this case > and < are used to point in the best direction for each variable. The asterisk corresponds to the condition where they agree. See [“Statistical Details for the Partition Platform”](#).

Node Options

This section describes the options in the red triangle menu for each node.

Split Best Finds and executes the best split at or below this node.

Split Here Splits at the selected node on the best column to split by.

Split Specific Lets you specify where a split takes place. This is useful in showing what the criterion is as a function of the cut point, as well as in determining custom cut points. When specifying a splitting column, you can choose the following options for how the split is performed:

Optimal Value Splits at the optimal value of the selected variable.

Specified Value Enables you to specify the level where the split takes place.

Output Split Table Produces a data table showing all possible splits and their associated split value.

Prune Below Eliminates the splits below the selected node.

Prune Worst Finds and removes the worst split below the selected node.

Select Rows Selects the data table rows corresponding to this leaf. You can extend the selection by holding down the Shift key and choosing this command from another node.

Show Details Produces a data table that shows the split criterion for a selected variable. The data table, composed of split intervals and their associated criterion values, has an attached script that produces a graph for the criterion.

Lock Prevents a node or its subnodes from being chosen for a split. When checked, a lock icon appears in the node title.

Partition Platform Options

The options in the Partition red triangle menu give you the ability to customize reports according to your needs. The available options are determined by the type of data that you use for your analysis.

Display Options Contains options that show or hide report elements.

Show Points Shows the points. For categorical responses, this option shows the points or colored panels.

Show Tree Shows the large tree of partitions.

Show Graph Shows the partition graph.

Show Split Bar (Available only for categorical responses.) Shows the colored bars that indicate the split proportions in each leaf.

Show Split Stats Shows the split statistics. For more information about the categorical split statistic G^2 , see [“Statistical Details for the Partition Platform”](#).

Show Split Prob (Available only for categorical responses.) Shows the Rate and Prob statistics in the node reports.

JMP automatically shows the Rate and Prob statistics when you select **Show Split Count**. For more information about Rate and Prob, see [“Statistical Details for the Partition Platform”](#).

Show Split Count (Available only for categorical responses.) Shows frequency counts in the node reports. When you select this option, JMP automatically selects **Show Split Prob**. And when you deselect **Show Split Prob**, the counts do not appear.

Show Split Candidates Shows the Candidates report.

Sort Split Candidates Sorts the Candidates reports by the statistic or the log(worth), whichever is appropriate.

Split Best Splits the tree at the optimal split point. This is equivalent to clicking the **Split** button.

Prune Worst Removes the terminal split that has the least discrimination ability. This is equivalent to clicking the **Prune** button.

Minimum Size Split Define the minimum size split allowed by entering a number or a fractional portion of the total sample size. 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 is set to the maximum of 5, or the floor of the number of rows divided by 10,000.

Lock Columns Interactively lock columns so that they are not considered for splitting. You can turn the display off or back on without affecting the individual locks.

Plot Actual by Predicted (Available only for continuous responses.) Shows a plot of actual response values by predicted response values. When you fit a Decision Tree, all observations in a leaf have the same predicted value. If there are n leaves, then the Actual by Predicted plot shows at most n distinct predicted values. The actual values form a scatter of points around each leaf mean on n vertical lines.

The plot contains a diagonal line that is the $Y = X$ line. For a perfect fit, all the points would be on this diagonal. When validation is used, plots are shown for both the training and the validation sets.

Small Tree View Shows a small version of the partition tree to the right of the partition plot.

Tree 3D Shows a 3-D plot of the tree structure. To access this option, press Shift and click the red triangle menu.

Leaf Report Shows the mean and count or rates for the bottom-level leaves of the report.

Column Contributions Shows a report indicating each input column's contribution to the fit. The report also shows how many times it defined a split and the total G^2 or Sum of Squares attributed to that column.

Split History Shows a plot of RSquare versus the number of splits. If you use excluded row validation, holdback validation, or a validation column, separate curves are drawn for training and validation RSquare values. The RSquare curve is blue for the training set and red for the validation set. If you select K Fold Crossvalidation, the RSquare curve for all of the data is blue, and the curve for the cross validation RSquare is green.

ROC Curve (Available only for categorical responses.) Receiver Operating Characteristic (ROC) curves display the efficiency of a model's fitted probabilities to sort the response levels. See ["ROC Curve"](#).

Lift Curve (Available only for categorical responses.) Lift curves display the predictive ability of a partition model. See ["Lift Curve"](#).

Decision Threshold (Available only for binary categorical responses.) Shows or hides Decision Thresholds reports for the training, validation, and test sets, if specified. Each report contains a graph of the distribution of fitted probabilities for each model, confusion matrices for each model, classification graphs to compare the model fits, and a table of classification accuracy metrics. See ["Decision Thresholds Report"](#).

Show Fit Details (Appears only for categorical responses.) The Fit Details report shows several measures of fit and provides a Confusion Matrix report. See ["Show Fit Details"](#).

Save Columns Contains options for saving model and tree results, and creating SAS code.

Save Residuals Saves the residual values from the model to the data table.

Save Predicteds Saves the predicted values from the model to the data table.

Save Leaf Numbers Saves the leaf numbers of the tree to a column in the data table.

Save Leaf Labels Saves leaf labels of the tree to the data table. The labels document each branch that the row would trace along the tree. Each branch is separated by "&". An example label might be: "size(Small,Medium)&size(Small)". However, JMP does not include redundant information in the form of category labels that are repeated. A category label for a leaf might refer to an inclusive list of categories in a higher tree node. A caret (^) appears where the tree node with redundant labels occurs. Therefore, "size(Small,Medium)&size(Small)" is presented as ^&size(Small).

Save Prediction Formula Saves prediction formulas to a column or multiple columns in the data table. The formulas consist of nested conditional clauses that describe the tree structure. If the response is continuous, one column that contains a Predicting property is added. If the response is categorical, columns that contain a Response Probability property are added for each level of the response. In addition, a Most Likely column that contains the response level with the highest probability of occurrence for each observation is added.

Save Tolerant Prediction Formula Saves a formula that predicts even when there are missing values and when Informative Missing has not been checked. The prediction formula tolerates missing values by randomly allocating response values for missing predictors to a split. If the response is continuous, the column contains a Predicting property. If the response is categorical, the column contains a Response Probability property. If you have checked Informative Missing, you can save the Tolerant Prediction Formula by holding the Shift key as you click the report's red triangle.

Save Leaf Number Formula Saves a column containing a formula in the data table that computes the leaf number.

Save Leaf Label Formula Saves a column containing a formula in the data table that computes the leaf label.

Make SAS DATA Step Creates SAS code for scoring a new data set.



Publish Prediction Formula Creates a prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See "[Formula Depot](#)".



Publish Tolerant Prediction Formula Creates a tolerant prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See "[Formula Depot](#)". If you have checked Informative Missing, you can use this option by holding the Shift key as you click the report's red triangle.

Specify Profit Matrix (Available only for categorical responses.) Enables you to specify profits or costs associated with correct or incorrect classification decisions. For a nominal response, you can specify the profit matrix entries using a probability threshold. See [“Specify Profit Matrix”](#).

Profiler Shows an interactive profiler report. Changes in the factor values are reflected in the estimated classification probabilities. See *Profilers*.

Color Points (Available only for categorical responses.) Colors points based on their response level. This is equivalent to clicking the Color Points button. See [“The Partition Report”](#).

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Show Fit Details

The Fit Details report in the Partition platform contains the following measures of fit:

Entropy RSquare A measure of fit that compares the log-likelihoods from the fitted model and the constant probability model. Entropy RSquare ranges from 0 to 1, where values closer to 1 indicate a better fit. See [“Entropy RSquare”](#).

Generalized RSquare A measure that can be applied to general regression models. It is based on the likelihood function L and is scaled to have a maximum value of 1. The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the Nagelkerke or Craig and Uhler R^2 , which is a normalized version of Cox and Snell's pseudo R^2 . See Nagelkerke (1991). Values closer to 1 indicate a better fit.

Mean -Log p The average of $-\log(p)$, where p is the fitted probability associated with the event that occurred. Smaller values indicate a better fit.

RASE The root average square error, where the differences are between the response and p (the fitted probability for the event that actually occurred). Smaller values indicate a better fit.

Mean Abs Dev The average of the absolute values of the differences between the response and p (the fitted probability for the event that actually occurred). Smaller values indicate a better fit.

Misclassification Rate The rate for which the response category with the highest fitted probability is not the observed category. Smaller values indicate a better fit.

Confusion Matrix Report

The Confusion Matrix report shows matrices for the training set and for the validation and test sets (if defined). The Confusion Matrix Report contains confusion matrices and confusion rates matrices. A confusion matrix is a two-way classification of actual and predicted responses. A confusion rates matrix is equal to the confusion matrix, with the numbers divided by the row totals.

Decision Matrix Report

When a profit matrix is defined, the partition algorithm uses the values in the matrix to calculate the profit for each decision. You can define a profit matrix with a Profit Matrix column property or by specifying costs using the Specify Profit Matrix option. See [“Specify Profit Matrix”](#).

When you select Show Fit Details, a Decision Matrix report appears. In the Decision Matrix report, the decision counts reflect the most profitable prediction decisions based on the weighting in the profit matrix. The report gives Decision Count and Decision Rate matrices for the training set and for validation and test sets (if defined). For reference, the profit matrix is also shown.

Note: If you change the weights in your Profit Matrix using the Specify Profit Matrix option, the Decision Matrix report automatically updates to reflect your changes.

Decision Count Matrix Shows a two-way classification with actual responses in rows and classification counts in columns.

Specified Profit Matrix Gives the weights that define the Profit Matrix.

Decision Rate Matrix Shows rate values corresponding to the proportion of a given row's observations that are classified into each category. If all observations are correctly classified, the rates on the diagonal are all equal to one.

Tip: You can obtain a decision rate matrices for a response using the default profit matrix with costs of 1 and -1. Select **Specify Profit Matrix** from the red triangle menu, make no changes to the default values, and click **OK**.

The matrices are arranged in two rows:

- The Decision Count matrices are in the first row.
- The Specified Profit Matrix is to the right in the first row.
- The Decision Rate matrices are in the second row.

Specify Profit Matrix

For categorical responses, you can specify a profit matrix in the Partition platform. A profit matrix is used to assign costs to undesirable outcomes and profits to desirable outcomes.

Figure 4.8 Specify Profit Matrix Window

Specify Profit Matrix

Each matrix entry is the profit if you predict the response in the column when the response in the row is the actual response.

Enter values that reflect profits for correct decisions on the diagonal.
Enter values (usually negative) that reflect profits for incorrect decisions elsewhere.
Use the Undecided column to reflect profits for an alternative decision.

When you save prediction formulas, these values will be used to create best decision columns.
The best decision is the one with greatest expected profit.

		Decision or Prediction		
		High	Low	Undecided
Actual	High	1	-1	.
	Low	-1	1	.

To create a profit matrix for a binary response, enter a Target and Probability Threshold.
If the predicted probability exceeds the threshold, the best decision will be the target.

Target: Low Set

Probability Threshold: 0.5

☐ Save to column as property.

OK Cancel

You can assign profit and cost values to each combination of actual and predicted response categories. To specify the costs of classifying into an alternative category, enter values in the Undecided column. To save your assignments to the response column as a property, check **Save to column as property**. Leaving this option unchecked applies the Profit Matrix only to the current Partition report.

Probability Threshold Specification for Profit Matrix

When the response is binary, instead of entering weights into the profit matrix, you can specify a probability threshold in the Profit Matrix window. For more information about how values are calculated for the profit matrix, see *Using JMP*.

Target The level whose probability is modeled.

Probability Threshold A threshold for the probability of the target level. If the probability that an observation falls into the target level exceeds the probability threshold, the observation is classified into that level.

When you define costs using the Specify Profit Matrix option and then select Show Fit Details, a Decision Matrix report appears. See [“Decision Matrix Report”](#).

When you specify a profit matrix and save the model prediction formula, the formula columns saved to the data table include the following:

- Profit for <level>: For each level of the response, a column gives the expected profit for classifying each observation into that level.
- Most Profitable Prediction for <column name>: For each observation, gives the level of the response with the highest expected profit.
- Expected Profit for <column name>: For each observation, gives the expected profit for the classification defined by the Most Profitable Prediction column.
- Actual Profit for <column name>: For each observation, gives the actual profit for classifying that observation into the level specified by the Most Profitable Prediction column.

ROC Curve

Receiver Operating Characteristic (ROC) curves display the efficiency of a model's fitted probabilities in sorting the response levels. ROC Curves are available only for categorical responses. An introduction to ROC curves is found in *Basic Analysis*.

The predicted response for each observation in a partition model is a value between 0 and 1. To use the predicted response to classify observations as positive or negative, a *cut point* is used. For example, if the cut point is 0.5, an observation with a predicted response at or above 0.5 would be classified as positive, and an observation below 0.5 as negative. There are trade offs in classification as the cut point is varied.

To generate a ROC curve, each predicted response level is considered as a possible cut point and the following values are computed for each possible cut point:

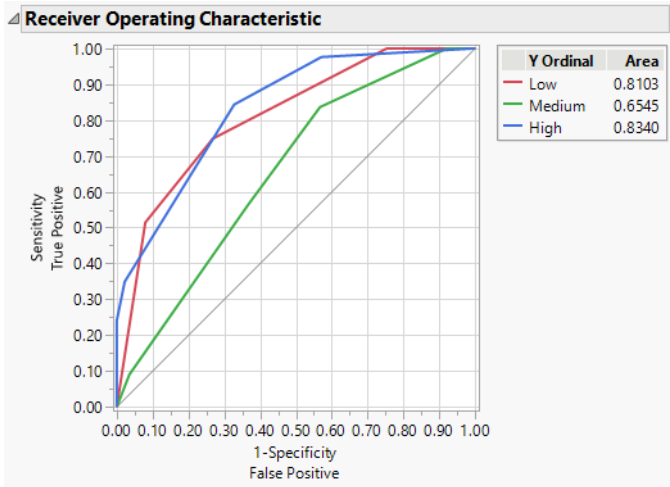
- The *sensitivity* is the proportion of true positives or the percent of positive observations with a predicted response greater than the cut point.

- The *specificity* is the proportion of true negatives or the proportion of negative observations with a predicted response less than the cut point.

The ROC curve plots sensitivity against 1 - specificity. A partition model with n splits has $n+1$ predicted values. The ROC curve for the partition model has $n+1$ line segments.

If your response has more than two levels, the Partition report contains a separate ROC curve for each response level versus the other levels. Each curve is the representation of a level as the positive response level. If there are only two levels, one curve is the reflection of the other.

Figure 4.9 ROC Curves for a Three Level Response



If the model perfectly rank-orders the response values, then the sorted data contains all of the positive values first, followed by all of the other values. In this situation, the curve moves all the way to the top before it moves at all to the right. If the model does not predict well, the curve follows the diagonal line from the bottom left to top right of the plot.

In practice, the ROC curve lies above the diagonal. The area under the curve is the indicator of the goodness of fit for the model. A value of 1 indicates a perfect fit and a value near 0.5 indicates that the model cannot discriminate among groups.

When your response has more than two levels, the ROC curve plot enables you to see which response categories have the largest area under the curve.

Lift Curve

The Lift Curve option provides another plot to display the predictive ability of a partition model. The lift curve plots the lift versus the portion of the observations. There is a point for each unique predicted probability value. Each predicted probability of a response level defines a portion of the observations that have a predicted probability greater than or equal to the unique predicted probability value. For a particular level of the response, the *lift* value is the ratio of the proportion of observed responses in that portion to the overall proportion of observed responses.

Note: For smaller models, it is possible that a large portion of the points have the same predicted probability value. If this probability is the highest predicted probability for the response level, the lift curve does not start at Portion = 0. For example, this is shown in the Low lift curve in [Figure 4.10](#).

Figure 4.10 Lift Curve

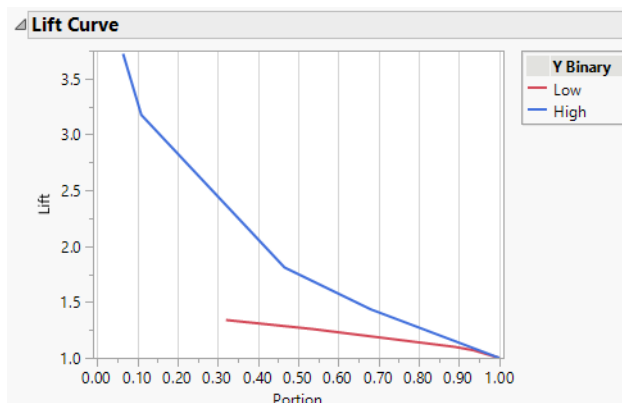


Figure 4.11 Lift Table for Lift Curve

Prob High	N > Prob High	Portion	N High in Portion	Portion High	Lift = portion high/ overall high of .27
0.97	20	0.06	20	1.00	3.72
0.77	44	0.14	39	0.89	3.30
0.33	68	0.22	47	0.69	2.57
0.31	168	0.54	78	0.46	1.73
0.04	309	1.00	83	0.27	1.00

[Figure 4.11](#) provides a table of values to demonstrate the calculation of Lift and Portion used for the High lift curve shown in [Figure 4.10](#). A partition model with five splits was built to predict the response, Y Binary. Y Binary has two levels: Low and High. The lift curve is based on 309 observations. There are 83 observed High responses for an overall rate of 0.27.

- Prob High: The five unique predicted probability values from the partition model for the High response level.

- **N > Prob High:** The number of observations that have a predicted probability value equal to or greater than the value in Prob High.
- **Portion:** N > Prob High divided by 309, the total number of observations.
- **N High in Portion:** The number of observations in each portion that have an observed High response.
- **Portion High:** N High in Portion divided by N > Prob High.
- **Lift:** Portion High divided by 0.27, the overall rate of the observed High response.

Lift measures how many High responses fall in each portion as compared to the expected number of High responses for that portion. For the first 6% of the data set the lift is 3.72. Using the model to select the 6% of the observations with the highest predicted values results in 3.72 more High responses than if that 6% were selected at random.

Validation in Partition

The use of validation with partition models is important given that partition models are easily overfit. When this happens, the model predicts the data used to build the model very well, but predicts future observations poorly. Validation is the process of using part of a data set to estimate model parameters, and using the other part to assess the predictive ability of the model. For more information about validation, see [“Validation in JMP Modeling”](#).

In Partition, when a validation method is used, the **Go** button appears. The **Go** button provides for repeated splitting without having to repeatedly click the **Split** button. When you click the **Go** button, splitting occurs until the validation RSquare is better than what the next 10 splits would obtain. This rule can result in complex trees that are not very interpretable, but have good predictive power.

Using the **Go** button turns on the **Split History** command. If using the **Go** button results in a tree with more than 40 nodes, the **Show Tree** command is turned off.

Select one of the following validation methods:

Excluded Rows Uses row states to subset the data. Rows that are unexcluded are used as the training set, and excluded rows are used as the validation set.

For more information about using row states and how to exclude rows, see *Using JMP*.

Holdback Randomly divides the original data into the training and validation data sets. The Validation Portion on the platform launch window is used to specify the proportion of the original data to use as the validation data set (holdback). See [“Launch the Partition Platform”](#) for more information about the Validation Portion.

JMP PRO Validation Column Uses a numeric column that defines the validation sets. This column should contain at most three distinct values:

- If the validation column has two levels, the smaller value defines the training set and the larger value defines the validation set.
- If the validation column has three levels, the values, in order of increasing size, define the training, validation, and test sets.
- If the validation column has more than three levels, the rows that contain the smallest three values define the validation sets. All other rows are excluded from the analysis.

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 [“Make Validation Column”](#).

Tip: To use K Fold or Nested K Fold crossvalidation, fit a partition model through the Model Screening platform. See [“Model Screening”](#).

Additional Examples of the Partition Platform

This section contains examples using the Partition platform.

- [“Example of a Continuous Response”](#)
- [“Example of Informative Missing”](#)
- [“Example of Profit Matrix and Decision Matrix Report”](#)

Example of a Continuous Response

In this example, you use the Partition platform to construct a decision tree that predicts the one-year disease progression measured on a quantitative scale for patients with diabetes.

1. Select **Help > Sample Data Folder** and Diabetes.jmp.
2. Select **Analyze > Predictive Modeling > Partition**.
3. Select Y and click **Y, Response**.
4. Select Age through Glucose and click **X, Factor**.
5. Select a validation procedure based on your JMP installation:
 - For JMP Pro, select Validation and click **Validation**.
 - For JMP, enter 0.3 as the **Validation Portion**.

Note: Results using the validation proportion can differ from those shown here due to the random selection of validation rows.

Figure 4.12 Completed Launch Window with Validation Portion = 0.3

Builds a decision tree to predict a response.

Select Columns	Cast Selected Columns into Roles	Action												
14 Columns - Y - Y Binary - Y Ordinal - Age - Gender - BMI - BP - Total Cholesterol - LDL - HDL - TCH - LTG - Glucose - Validation	<table><tr><td>Y, Response</td><td>Y Binary <i>optional</i></td></tr><tr><td>X, Factor</td><td>Age Gender BMI BP</td></tr><tr><td>Weight</td><td><i>optional numeric</i></td></tr><tr><td>Freq</td><td><i>optional numeric</i></td></tr><tr><td>Validation</td><td><i>optional numeric</i></td></tr><tr><td>By</td><td><i>optional</i></td></tr></table>	Y, Response	Y Binary <i>optional</i>	X, Factor	Age Gender BMI BP	Weight	<i>optional numeric</i>	Freq	<i>optional numeric</i>	Validation	<i>optional numeric</i>	By	<i>optional</i>	OK Cancel Remove Recall Help
Y, Response	Y Binary <i>optional</i>													
X, Factor	Age Gender BMI BP													
Weight	<i>optional numeric</i>													
Freq	<i>optional numeric</i>													
Validation	<i>optional numeric</i>													
By	<i>optional</i>													

Options

Method Decision Tree

Validation Portion 0.3

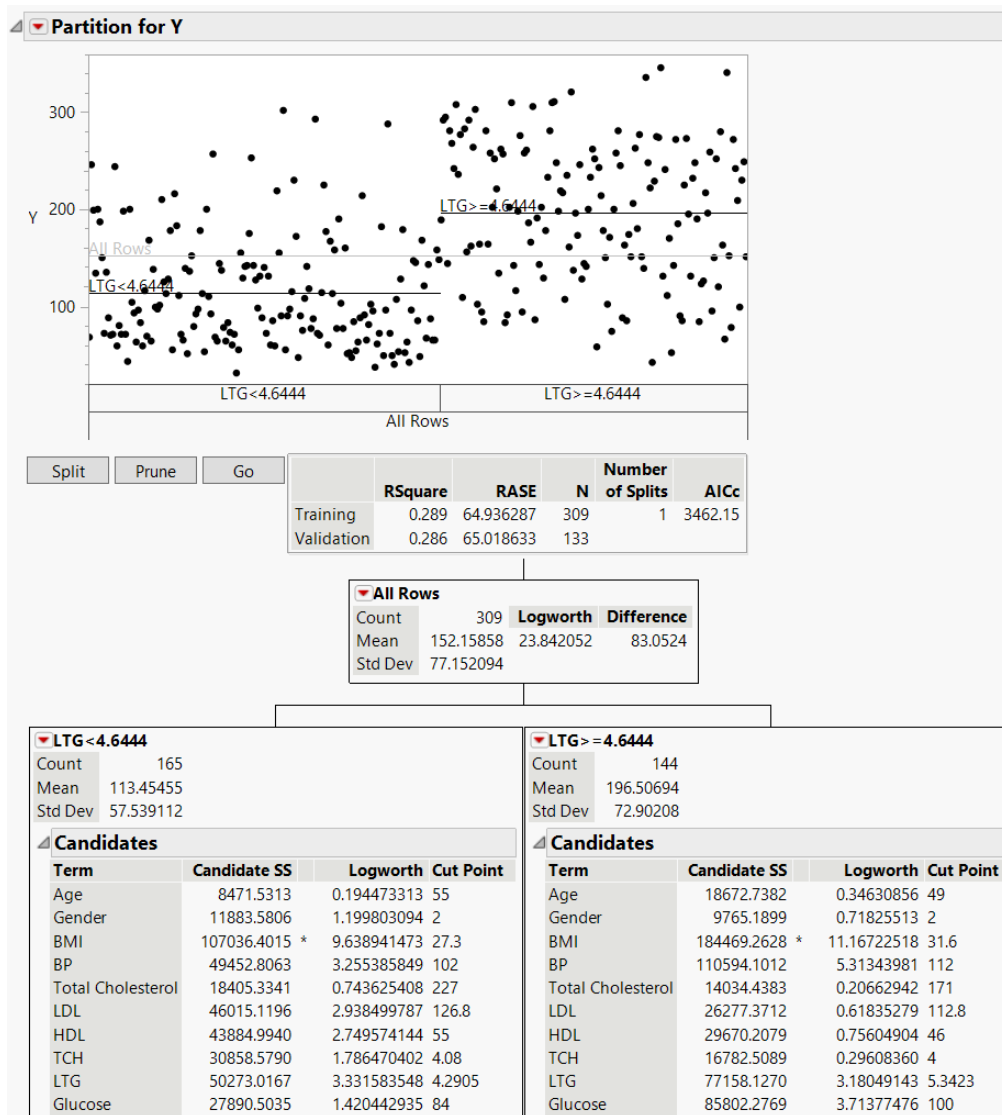
☒ Informative Missing

☒ Ordinal Restricts Order

6. Click **OK**.

7. On the platform report window, click **Split** once to perform a split.

Figure 4.13 Report after First Split with Decision Tree Hidden



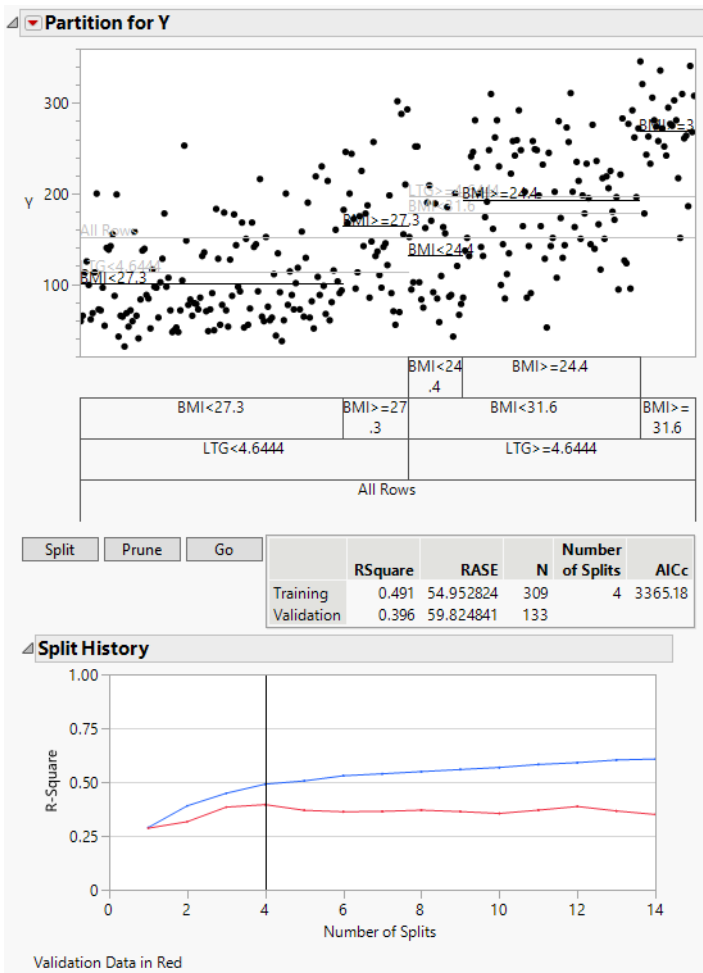
The original 309 values in the training data set are now split into two parts:

- The left leaf, corresponding to $LTG < 4.6444$, has 165 observations.
- The right leaf, corresponding to $LTG \geq 4.6444$ has 144 observations.

For both the right and left leaf the next split would be on BMI. The Candidate SS for BMI on the right leaf is higher than the Candidate SS for BMI on the left leaf. Thus, the next split is on the right leaf.

8. Click **Go** to use automatic splitting.

Figure 4.14 Report after Automatic Splitting with Validation



The solution found has four splits. The Split History plot shows that there is no further improvement in the validation data set after four splits. The RSquare value of 0.39 on the validation data does not support this model as a strong predictor of disease progression. The scatter across partitions in the partition plot further indicate that this model does not separate the Y levels well.

Example of Informative Missing

In this example, you use the Partition platform to construct a decision tree model to predict if a customer is a credit risk. Since your data set contains missing values, you also explore the effectiveness of the Informative Missing option.

Launch the Partition Platform

1. Select **Help > Sample Data Folder** and open Equity.jmp.
2. Select **Analyze > Predictive Modeling > Partition**.
3. Select BAD and click **Y, Response**.
4. Select LOAN through DEBTINC and click **X, Factor**.
5. Click **OK**.

Create the Decision Tree and ROC Curve with Informative Missing

1. Press Shift and click **Split**.
2. Enter 5 for the number of splits and click **OK**.
3. Click the red triangle next to Partition for BAD and select **ROC Curve**.
4. Click the red triangle next to Partition for BAD and select **Save Columns > Save Prediction Formula**.

The columns Prob(BAD==Good Risk) and Prob(BAD==Bad Risk) contain the formulas that Informative Missing utility uses to classify the credit risk of future loan applicants. You are interested in how this model performs in comparison to a model that does not use informative missing.

Create the Decision Tree and ROC Curve without Informative Missing

1. Click the red triangle next to Partition for BAD and select **Redo > Relaunch Analysis**.
2. De-select **Informative Missing**.
3. Click OK and repeat the steps in [“Create the Decision Tree and ROC Curve with Informative Missing”](#).

The columns Prob(BAD==Good Risk) 2 and Prob(BAD==Bad Risk) 2 contain the formulas that do not use the informative missing utility.

Compare the ROC Curves

Visually compare the ROC curves from the two models. The model at left is with Informative Missing, and the model at right is without Informative Missing.

The Measures of Fit report shows that the first model, which was fit with informative missing, performs better than the second model, which was not fit with informative missing. The first model has higher RSquare values as well as a lower RMSE value and a lower Misclassification Rate. These findings align the ROC curves comparison.

Note: Again, your results can differ due to the random differences when Informative Missing is not used.

Example of Profit Matrix and Decision Matrix Report

In this example, you use the Decision Matrix report in the Partition platform to help assess the classification rates relative to the costs in your profit matrix. The profit matrix is defined for a study of patients who have liver cancer. Based on various measurements and markers, you want to classify patients according to their disease severity (high or low). There are two errors that one can make in classification of patients: classifying a subject who has high severity into the low group, or classifying a patient with low severity into the high group. Clinically, the misclassification of a high patient as low is a costly error, as that patient might not receive the aggressive treatment needed. Classifying a patient with low severity into the high severity group is a less costly error. That patient might receive the more aggressive treatment than needed, but this is not a major concern.

1. Select **Help > Sample Data Folder** and open Liver Cancer.jmp.
2. Select **Analyze > Predictive Modeling > Partition**.
3. Select Severity and click **Y, Response**.
4. Select BMI through Jaundice and click **X, Factor**.
5. Select a validation procedure based on your JMP installation:
 - For JMP Pro, select Validation and click **Validation**.
 - For JMP, enter 0.3 as the **Validation Proportion**.

Note: Results using the validation proportion can differ from those shown here, due to the random selection of validation rows.

Figure 4.17 Completed Launch Window with Validation Portion = 0.3

Builds a decision tree to predict a response.

Select Columns

9 Columns

- Node Count
- Severity
- BMI
- Age
- Time
- Markers
- Hepatitis
- Jaundice
- Validation

Options

Method: Decision Tree

Validation Portion: 0.3

☒ Informative Missing

☒ Ordinal Restricts Order

Cast Selected Columns into Roles

Y, Response: Severity (optional)

X, Factor: BMI, Age, Time, Markers

Weight: optional numeric

Freq: optional numeric

Validation: optional numeric

By: optional

Action

OK

Cancel

Remove

Recall

Help

6. Click **OK**.
7. Press Shift and click **Split**.
8. Enter 10 for the number of splits and click **OK**.
Check that the Number of Splits is 10 in the panel beneath the plot.
9. Click the red triangle next to Partition for Severity and select **Specify Profit Matrix**.
10. Change the entries to the following values:
 - Enter 1 in the High, High box.
 - Enter -5 in the High, Low box.
 - Enter -3 in the Low, High box.
 - Enter 1 in the Low, Low box.

Figure 4.18 Completed Profit Matrix

		Decision or Prediction		
		Low	High	Undecided
Actual	Low	1	-5	.
	High	-3	1	.

Tip: You can save this profit matrix as a column property for use in later analyses. Select the check box “Save column as property” at the bottom of the profit matrix window.

Note the following:

- Each value of 1 reflects your profit when you make a correct decision.

- The -3 value indicates that if you classify a Low severity patient as High severity, your loss is 3 times as much as the profit of a correct decision.
- The -5 value indicates that if you classify a High severity patient as Low severity, your loss is 5 times as much as the profit of a correct decision.

11. Click **OK**.

12. Click the red triangle next to Partition for Severity and select **Show Fit Details**.

Figure 4.19 Confusion Matrix and Decision Matrix Reports

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.3788	0.0367	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5382	0.0647	$(1-(L(0)/L(model))^{2/n})/(1-L(0)^{2/n})$
Mean -Log p	0.4134	0.6402	$\sum -\text{Log}(p[j])/n$
RASE	0.3649	0.4694	$\sqrt{\sum (y[j]-p[j])^2/n}$
Mean Abs Dev	0.2854	0.3861	$\sum y[j]-p[j] /n$
Misclassification Rate	0.1915	0.3810	$\sum (p[j] \neq pMax)/n$
N	94	42	n

Confusion Matrix

Training

	Predicted Count	
Actual Severity	Low	High
Low	30	6
High	12	46

Validation

	Predicted Count	
Actual Severity	Low	High
Low	11	5
High	11	15

	Predicted Rate	
Actual Severity	Low	High
Low	0.833	0.167
High	0.207	0.793

	Predicted Rate	
Actual Severity	Low	High
Low	0.688	0.313
High	0.423	0.577

Decision Matrix

Training

	Decision Count	
Actual Severity	Low	High
Low	33	3
High	16	42

Validation

	Decision Count	
Actual Severity	Low	High
Low	12	4
High	12	14

Specified Profit Matrix

	Decision	
Actual	Low	High
Low	1	-5
High	-3	1

	Decision Rate	
Actual Severity	Low	High
Low	0.917	0.083
High	0.276	0.724

	Decision Rate	
Actual Severity	Low	High
Low	0.750	0.250
High	0.462	0.538

Misclassification Rate	
0.2021	

Misclassification Rate	
0.3810	

The Confusion Matrix and Decision Matrix reports follow the list of Measures in the Fit Details report. Notice that the Confusion Matrix report and the confusion matrices in the

Decision Matrix report show different counts. This is because the weighting in the profit matrix results in different decisions than do the predicted probabilities without weighting.

13. Click the red triangle next to Partition for Severity and select **Save Columns > Save Prediction Formula**.

Eight columns are added to the data table.

Tip: To quickly return to the data table, click the View Associated Data icon  in the bottom right corner of the report window (Windows) or the Show Data Table icon on the tool bar menu (macOS).

- The first three columns involve only the predicted probabilities. The confusion matrix counts are based on the Most Likely Severity column, which classifies a patient into the level with the highest predicted probability. These probabilities are given in the Prob(Severity == High) and Prob(Severity == Low) columns.
- The last five columns involve the profit matrix weighting. The column called Most Profitable Prediction for Severity contains the decision based on the profit matrix. The decision for a patient is the level that results in the largest profit. The profits are given in the Profit for High and Profit for Low columns.

Statistical Details for the Partition Platform

This section contains statistical details for the Partition platform.

- [“Statistical Details for Responses and Factors”](#)
- [“Statistical Details for the Splitting Criterion”](#)

Statistical Details for Responses and Factors

In the Partition platform, the response can be either continuous or categorical (nominal or ordinal):

- If the response is categorical, then it is fitting the probabilities estimated for the response levels, minimizing the residual log-likelihood chi-square $[2 \times \text{entropy}]$.
- If the response is continuous, then the platform fits means, minimizing the sum of squared errors.

The factors can be either continuous or categorical (nominal or ordinal):

- If the factor is continuous, then the partition is done according to a splitting “cut” value for the factor.

- If the factor is categorical, then it divides the X categories into two groups of levels and considers all possible groupings into two levels.

Statistical Details for the Splitting Criterion

Node splitting in the Partition platform is based on the logworth statistic, which is reported in Candidate reports for nodes. Logworth is calculated as follows:

$$-\log_{10}(p\text{-value})$$

where the adjusted p -value is calculated in a complex manner that takes into account the number of different ways splits can occur. This calculation is very fair compared to the unadjusted p -value, which favors X s with many levels, and the Bonferroni p -value, which favors X s with small numbers of levels. Details about the method are discussed in Sall (2002).

For continuous responses, the Sum of Squares (SS) is reported in node reports. This is the change in the error sum-of-squares due to the split.

A candidate SS that has been chosen is:

$$SS_{\text{test}} = SS_{\text{parent}} - (SS_{\text{right}} + SS_{\text{left}}) \text{ where SS in a node is just } s^2(n - 1).$$

Also reported for continuous responses is the Difference statistic. This is the difference between the predicted values for the two child nodes of a parent node.

For categorical responses, the G^2 (likelihood ratio chi-square) appears in the report. This is actually twice the [natural log] entropy or twice the change in the entropy. Entropy is $\Sigma -\log(p)$ for each observation, where p is the probability attributed to the response that occurred.

A candidate G^2 that has been chosen is:

$$G^2_{\text{test}} = G^2_{\text{parent}} - (G^2_{\text{left}} + G^2_{\text{right}}).$$

Partition actually has two rates; one used for training that is the usual ratio of count to total, and another that is slightly biased away from zero. By never having attributed probabilities of zero, this allows logs of probabilities to be calculated on validation or excluded sets of data, used in Entropy R-Square.

Chapter 5

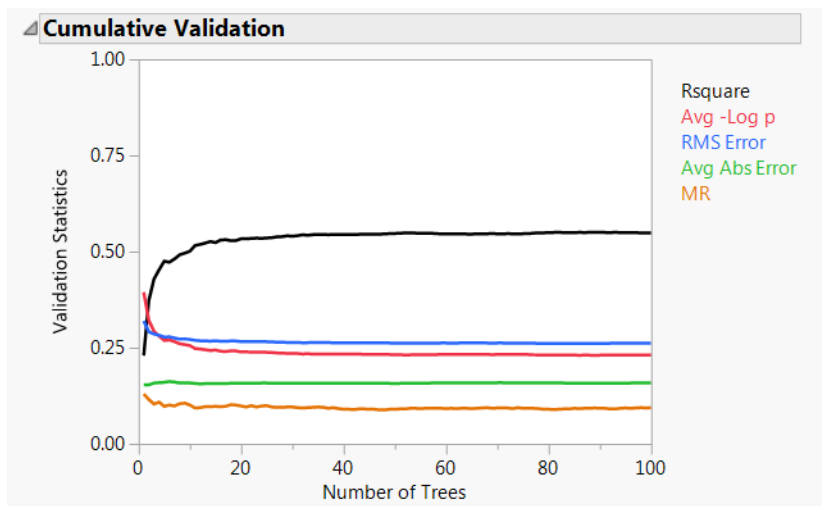
JMP[®] PRO Bootstrap Forest

Fit a Model By Averaging Many Trees

The Bootstrap Forest platform is available only in JMP Pro.

The Bootstrap Forest platform fits an ensemble model by averaging many decision trees each of which is fit to a bootstrap sample of the training data. Each split in each tree considers a random subset of the predictors. In this way, many weak models are combined to produce a more powerful model. The final prediction for an observation is the average of the predicted values for that observation over all the decision trees.

Figure 5.1 Example of a Cumulative Validation Report



Contents

Overview of the Bootstrap Forest Platform..... 87

Example of Bootstrap Forest with a Categorical Response..... 88

Example of Bootstrap Forest with a Continuous Response 91

Launch the Bootstrap Forest Platform 93

 Specification Window..... 95

The Bootstrap Forest Report 97

 Model Validation Set Summaries 99

 Specifications 99

 Overall Statistics 99

 Cumulative Validation 101

 Per Tree Summaries 102

Bootstrap Forest Platform Options..... 102



Overview of the Bootstrap Forest Platform

The Bootstrap Forest platform predicts a response value by averaging the predicted response values across many decision trees. Each tree is grown on a *bootstrap sample* of the training data. A bootstrap sample is a random sample of observations, drawn with replacement. In addition, the predictors are sampled at each split in the decision tree. The decision tree is fit using the recursive partitioning methodology described in “[Partition Models](#)”.

This is the fitting process for the training set:

1. For each tree, select a bootstrap sample of observations.
2. Fit the individual decision tree, using recursive partitioning.
 - Select a random set of predictors for each split and randomize the order of the predictors in the set.
 - Continue splitting until a stopping rule that is specified in the Bootstrap Forest Specification window is met.
3. Repeat [step 1](#) and [step 2](#) until the number of trees specified in the Bootstrap Forest Specification window is reached or until Early Stopping occurs.

For an individual tree, the bootstrap sample of observations that is used to fit the tree is drawn with replacement. You can specify the proportion of observations to be sampled. If you specify that 100% of the observations are to be sampled, because they are drawn with replacement, the expected proportion of unused observations is $1/e$, or approximately 36.8%. For each individual tree, these unused observations are called the *out-of-bag* observations. The observations used in fitting the tree are called *in-bag* observations. For a continuous response, the Bootstrap Forest platform provides measures for the error rate for out-of-bag observations, called *out-of-bag error*.

For a continuous response, the predicted value for an observation is the average of its predicted values over the collection of individual trees. For a categorical response, the predicted probability for an observation is the average of its predicted probabilities over the collection of individual trees. The observation is classified into the level for which its predicted probability is the highest.

For more information about bootstrap forests, see Hastie et al. (2009).



Example of Bootstrap Forest with a Categorical Response

In this example, you construct a bootstrap forest model to predict whether a customer is a bad credit risk. Then, you explore the degree to which predictor values are missing.

Bootstrap Forest Model

1. Select **Help > Sample Data Folder** and open Equity.jmp.
2. Select **Analyze > Predictive Modeling > Bootstrap Forest**.
3. Select BAD and click **Y, Response**.
4. Select LOAN through DEBTINC and click **X, Factor**.
5. Select Validation and click **Validation**.
6. Click **OK**.
7. Enter 4 next to Number of Terms Sampled per Split.
8. Enter 30 next to Maximum Splits per Tree.
9. Select **Multiple Fits over Number of Terms** and enter 10 next to Max Number of Terms.
10. (Optional) Select **Suppress Multithreading** and enter 123 next to Random Seed.

Because the bootstrap forest method involves random sampling, these actions ensure that your results are exactly the same as the results shown below.
11. Click **OK**.

Figure 5.3 Column Contributions Report

Column Contributions				
Term	Number of Splits	G ²		Portion
DEBTINC	82	551.046172		0.5305
DELINQ	83	109.254056		0.1052
VALUE	73	101.104219		0.0973
DEROG	73	85.0941864		0.0819
CLAGE	79	49.6260994		0.0478
CLNO	57	30.3844858		0.0293
JOB	55	27.1814486		0.0262
LOAN	54	26.5515552		0.0256
NINQ	57	25.4950118		0.0245
YOJ	44	17.4271135		0.0168
MORTDUE	27	12.0098272		0.0116
REASON	12	3.58243083		0.0034

The Column Contributions report suggests that the strongest predictor of a customer's credit risk is DEBTINC, which is the debt to income ratio. The next highest contributors to the model are DELINQ, the number of delinquent credit lines, and VALUE, the assessed value of the customer.

Missing Values

1. Select **Analyze > Screening > Explore Missing Values**.
2. Select Bad through DEBTINC and click **Y, Columns**.
3. Click **OK** in the Alert that appears.

The columns REASON and JOB are not added to the Y, Columns list because they have a Character data type. You can see how many values are missing for these two columns using Distribution (not illustrated in this example).

4. Click **OK**.

Figure 5.4 Missing Values Report

Explore Missing Values

Commands

- Missing Value Report** Lists the number of missing values for each column.
- Missing Value Clustering** Provides a hierarchical clustering analysis of the missing data to determine whether groups of columns have similar patterns of missingness.
- Missing Value Snapshot** Finds patterns of missing values using a graphical map.
- Multivariate Normal Imputation** Imputes missing values using least squares predictions from the nonmissing columns. Avoid this method when there are hundreds of columns. Use the shrinkage option to improve estimation of the covariance matrix.
- Multivariate SVD Imputation** Imputes missing values quickly for large problems using an iterated low-rank SVD matrix completion method.
- Multivariate RPCA Imputation** Imputes missing values using robust principal components, which replaces missing values using a low-rank matrix factorization (SVD) that is robust to outliers. This method is useful for wide problems.
- Automated Data Imputation** Imputes missing values using a low-rank matrix approximation method. This method automatically selects the best dimension for the low-rank approximation based on the data.

Automated Data Imputation Controls

Missing Columns

☐ Show only columns with missing
Close

Select columns and choose an action.
Select Rows Color Cells
Exclude Rows Color Rows

Column	Number Missing
BAD	0
LOAN	0
MORTDUE	518
VALUE	112
YOJ	515
DEROG	708
DELINQ	580
CLAGE	308
NINQ	510
CLNO	222
DEBTINC	1267

The DEBTINC column contains 1267 missing values, which amounts to about 21% of the observations. Most other columns involved in the Bootstrap Forest analysis also contain missing values. The Informative Missing option in the launch window ensures that the missing values are treated in a way that acknowledges any information that they carry. See [“ROC Curve”](#).

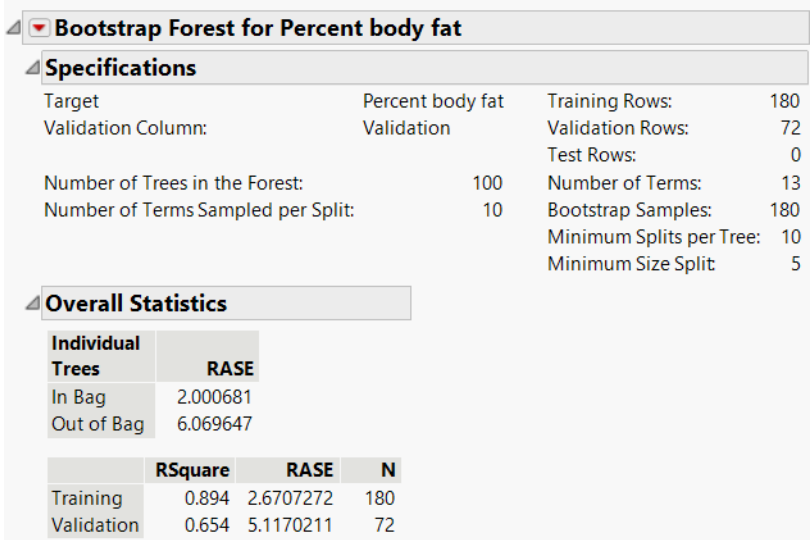
JMP PRO Example of Bootstrap Forest with a Continuous Response

In this example, you construct a bootstrap forest model to predict the percent body fat for male subjects.

1. Select **Help > Sample Data Folder** and open Body Fat.jmp.
2. Select **Analyze > Predictive Modeling > Bootstrap Forest**.

- 3. Select Percent body fat and click **Y, Response**.
- 4. Select Age (years) through Wrist circumference (cm) and click **X, Factor**.
- 5. Select Validation and click **Validation**.
- 6. Click **OK**.
- 7. (Optional) Select **Suppress Multithreading** and enter 123 next to Random Seed.
Because the bootstrap forest method involves random sampling, these actions ensure that your results are exactly the same as the results shown below.
- 8. Click **OK**.

Figure 5.5 Overall Statistics



The Overall Statistics report shows that the Validation RSquare is 0.654. You are interested in obtaining a model-independent indication of the most important predictors.

- 9. Click the red triangle next to Bootstrap Forest for Percent body fat and select **Column Contributions**.

Figure 5.6 Column Contributions

Column Contributions				
Term	Number of Splits	SS		Portion
Abdomen circumference (cm)	101	4185.96516		0.6194
Chest circumference (cm)	39	850.607271		0.1259
Knee circumference (cm)	33	254.906105		0.0377
Height (inches)	50	242.813772		0.0359
Age (years)	45	239.973342		0.0355
Ankle circumference (cm)	45	167.897611		0.0248
Thigh circumference (cm)	25	162.416449		0.0240
Weight (lbs)	25	145.612142		0.0215
Biceps (extended) circumference (cm)	44	142.127633		0.0210
Wrist circumference (cm)	38	117.953731		0.0175
Neck circumference (cm)	32	111.864186		0.0166
Hip circumference (cm)	24	90.0947826		0.0133
Forearm circumference (cm)	25	46.3944108		0.0069

The Column Contributions report suggests that Abdomen circumference (cm), Chest circumference (cm), and Age (years) are the strongest predictors for Percent body fat.

Launch the Bootstrap Forest Platform

Launch the Bootstrap Forest platform by selecting **Analyze > Predictive Modeling > Bootstrap Forest**.

Figure 5.7 Bootstrap Forest Launch Window

Builds a collection of decision trees using random sampling and averages the results to predict a response.

Select Columns

14 Columns

- BAD
- LOAN
- MORTDUE
- VALUE
- REASON
- JOB
- YOJ
- DEROG
- DELINQ
- CLAGE
- NINQ
- CLNO
- DEBTINC
- Validation

Cast Selected Columns into Roles

Y, Response *required*
optional

X, Factor *required*
optional

Weight *optional numeric*

Freq *optional numeric*

Validation *optional numeric*

By *optional*

Action

OK

Cancel

Remove

Recall

Help

Options

Method **Bootstrap Forest**

Validation Portion

☒ Informative Missing

☒ Ordinal Restricts Order

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Bootstrap Forest platform launch provides the following options:

Y, Response The response variable or variables that you want to analyze.

X, Factor The predictor variables.

Weight A column whose numeric values assign a weight to each row in the analysis.

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

Validation A numeric column that defines the validation sets. This column should contain at most three distinct values:

- If the validation column has two levels, the smaller value defines the training set and the larger value defines the validation set.
- If the validation column has three levels, the values, in order of increasing size, define the training, validation, and test sets.
- If the validation column has more than three levels, the rows that contain the smallest three values define the validation sets. All other rows are excluded from the analysis.

The Bootstrap Forest platform uses the validation column to train and tune the model or to train, tune, and evaluate the model. For more information about validation, see [“Validation in JMP Modeling”](#).

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 [“Make Validation Column”](#).

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.

Method Enables you to select the partition method (Decision Tree, Bootstrap Forest, Boosted Tree, K Nearest Neighbors, or Naive Bayes). These alternative methods, except for Decision Tree, are available in JMP Pro.

For more information about these methods, see [“Partition Models”](#), [“Boosted Tree”](#), [“K Nearest Neighbors”](#), and [“NaiveBayes”](#).

Validation Portion The portion of the data to be used as the validation set.

Informative Missing If selected, enables missing value categorization for categorical predictors and informative treatment of missing values for continuous predictors. See [“ROC Curve”](#).

Ordinal Restricts Order If selected, restricts consideration of splits to those that preserve the ordering.

JMP[®] PRO Specification Window

In the Bootstrap Forest Specification window, you can specify the number of rows and terms, details about the forest, multiple fits, and reproducibility. The Bootstrap Forest Specification window appears after you select OK in the Bootstrap Forest launch window.

Figure 5.8 Bootstrap Forest Specification Window

Bootstrap Forest Specification

Number of Rows: 252
Number of Terms: 13

Forest

Number of Trees in the Forest: 100
Number of Terms Sampled per Split: 10
Bootstrap Sample Rate: 1
Minimum Splits per Tree: 10
Maximum Splits per Tree: 2000
Minimum Size Split: 5
☒ Early Stopping

Multiple Fits

☐ Multiple Fits over Number of Terms
Max Number of Terms: 10
☐ Use Tuning Design Table

Reproducibility

☐ Suppress Multithreading
Random Seed: 0

OK Cancel

JMP[®] PRO Specification Panel

Number of Rows The number of rows in the data table.

Number of Terms The number of columns that are specified as predictors.

JMP[®] PRO Forest Panel

Number of Trees in the Forest The number of trees to grow and then average.

Number of Terms Sampled per Split The number of predictors to consider as splitting candidates at each split. For each split, a new random sample of predictors is taken as the candidate set.

Bootstrap Sample Rate The proportion of observations to sample (with replacement) for growing each tree. A new random sample is generated for each tree.

Minimum Splits Per Tree The minimum number of splits for each tree.

Maximum Splits Per Tree The maximum number of splits for each tree.

Minimum Size Split The minimum number of observations needed on a candidate split.

Early Stopping (Available only if validation is used.) If selected, the process stops growing additional trees if the additional trees do not improve the validation statistic. The validation statistic is the validation set's Entropy RSquare value for a categorical response and its RSquare value for a continuous response. If not selected, the process continues until the specified number of trees is reached.

JMP[®] PRO Multiple Fits Panel

Multiple Fits over Number of Terms If selected, creates a bootstrap forest for several values of number of terms sampled per split. The model for which results are displayed is the model whose Validation Set's Entropy RSquare value (for a categorical response) or RSquare (for a continuous response) is the largest.

The lower bound is the Number of Terms Sampled per Split specification. The upper bound is specified by the following option:

Max Number of Terms The maximum number of terms to consider for a split.

Use Tuning Design Table Opens a window where you can select a data table containing values for the Forest panel tuning parameters, called a *tuning design table*. A tuning design table has a column for each option that you want to specify and has one or multiple rows that each represent a single Bootstrap Forest model design. If an option is not specified in the tuning design table, the default value is used.

For each row in the table, JMP creates a Bootstrap Forest model using the tuning parameters specified. If more than one model is specified in the tuning design table, the Model Validation-Set Summaries report lists the RSquare value for each model. The Bootstrap Forest report shows the fit statistics for the model with the largest RSquare value.

You can create a tuning design table using the Design of Experiments facilities. A bootstrap forest tuning design table can contain the following case-insensitive columns in any order:

- Number Trees
- Number Terms
- Portion Bootstrap
- Minimum Splits per Tree
- Maximum Splits per Tree

- Minimum Size Split

JMP[®] PRO Reproducibility Panel

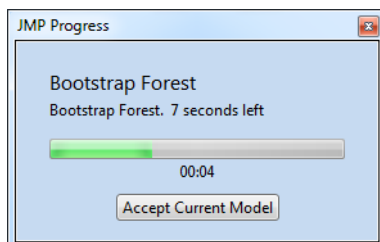
Suppress Multithreading If selected, all calculations are performed on a single thread.

Random Seed Specify a nonzero numeric random seed to reproduce the results for future launches of the platform. By default, the Random Seed is set to zero, which does not produce reproducible results. When you save the analysis to a script, the random seed that you enter is saved to the script.

JMP[®] PRO The Bootstrap Forest Report

After you click OK in the Bootstrap Forest Specification window, a window appears that informs you of approximately how long it will take JMP to fit the model. If the approximated time is too long, you have the option to view the current model by selecting **Accept Current Model**. The Bootstrap Forest report appears after the complete model is fit or after the current model is accepted.

Figure 5.9 Accept Current Model Dialog



The following reports are provided, depending on whether the response is categorical or continuous:

- “Model Validation Set Summaries”
- “Specifications”
- “Overall Statistics”
- “Cumulative Validation”
- “Per Tree Summaries”

Model Validation Set Summaries

In the Bootstrap Forest platform, the Model Validation Set Summaries report shows fit statistics for all of the fitted models. This report is available only when you select the Multiple Fits over Number of Terms option in the Bootstrap Forest Specification window. See [Figure 5.10](#) and “Multiple Fits Panel”.

Specifications

In the Bootstrap Forest platform, the Specifications report shows the settings used in fitting the model.

Overall Statistics

In the Bootstrap Forest platform, the Overall Statistics report provides fit statistics for the training set and for the validation and test sets if they are specified. The specific form of the report depends on the modeling type of the response.

Suppose that multiple models are fit using the Multiple Fits over Multiple Terms option in the Bootstrap Forest Specification window. Then the model for which results are displayed in the Overall Statistics and Cumulative Validation reports is the model for which the validation set’s Entropy RSquare value (for a categorical response) or RSquare (for a continuous response) is the largest.

Categorical Response

Measures Report

Gives the following statistics for the training set, and for the validation and test sets if they are specified.

Note: For Entropy RSquare and Generalized RSquare, values closer to 1 indicate a better fit. For Mean -Log p , RASE, Mean Abs Dev, and Misclassification Rate, smaller values indicate a better fit.

Entropy RSquare A measure of fit that compares the log-likelihoods from the fitted model and the constant probability model. It ranges from 0 to 1. See [“Entropy RSquare”](#).

Generalized RSquare A measure that can be applied to general regression models. It is based on the likelihood function L and is scaled to have a maximum value of 1. The value is 1 for a perfect model, and 0 for a model no better than a constant model. The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the Nagelkerke or Craig and Uhler R^2 , which is a normalized version of Cox and Snell’s pseudo R^2 .

Mean -Log P The average of negative $\log(p)$, where p is the fitted probability associated with the event that occurred.

RASE The root average squared prediction error. The differences are between 1 and p , the fitted probability for the response level that actually occurred.

Mean Abs Dev The average of the absolute values of the differences between the response and the predicted response. The differences are between 1 and p , the fitted probability for the response level that actually occurred.

Misclassification Rate The rate for which the response category with the highest fitted probability is not the observed category.

N The number of observations.

Confusion Matrix Report

(Available only for categorical responses.) Shows classification statistics for the training set, and for the validation and test sets if they are specified. The Confusion Matrix Report contains confusion matrices and confusion rates matrices. A confusion matrix is a two-way classification of actual and predicted responses. A confusion rates matrix is equal to the confusion matrix, with the numbers divided by the row totals.

Decision Matrix

(Available only for categorical responses and if the response has a Profit Matrix column property or if you specify costs using the Specify Profit Matrix option.) Gives Decision Count and Decision Rate matrices for the training set, and for the validation and test sets if they are specified. See [“Additional Examples of the Partition Platform”](#).

JMP[®] PRO Continuous Response

Individual Trees Report

Gives RASE values, which are averaged over all trees, for In Bag and Out of Bag observations. Training set observations that are used to construct a tree are called *in-bag* observations. Training observations that are not used to construct a tree are called *out-of-bag* (OOB) observations.

For each tree, the Out of Bag RASE is computed as the square root of the sum of squared errors divided by the number of OOB observations. The squared Out of Bag RASE for each tree is given in the Per-Tree Summaries report as OOB SSE/N.

RSquare and RASE Report

Gives Rsquare, root average squared prediction error, and the number of observations for the training set, and for the validation and test sets, if they are defined.

JMP[®] PRO Cumulative Validation

In the Bootstrap Forest platform, the Cumulative Validation report shows a plot of the fit statistics for the validation set versus the number of trees. This report is available only if validation is used.

For a continuous response, the single fit statistic is R-Square. For a categorical response, the fit statistics are listed below and are described in [“Measures Report”](#).

- RSquare (Entropy RSquare)
- Avg - Log p (Mean - Log p)
- Root Average Squared Prediction Error (RASE)
- Avg Abs Error (Mean Abs Dev)
- MR (Misclassification Rate)

The Cumulative Details report below the Cumulative Validation plot gives the values used in the plot.

JMP PRO Per Tree Summaries

In the Bootstrap Forest platform, the Per Tree Summaries report involves the concepts of in-bag and out-of-bag observations. For an individual tree, the bootstrap sample of observations used in fitting the tree is drawn with replacement. Even if you specify that 100% of the observations are to be sampled, because they are drawn with replacement, the expected proportion of unused observations is $1/e$. For each individual tree, the unused observations are called the *out-of-bag* observations. The observations used in fitting the tree are called *in-bag* observations.

The Per-Tree Summaries report shows the following summary statistics for each tree:

Splits The number of splits in the decision tree.

Rank The rank of the tree's OOB Loss in ascending order. The tree with the smallest OOB loss has Rank 1.

OOB Loss A measure of the total predictive inaccuracy of the tree when applied to the Out Of Bag rows. Lower values indicate a higher predictive accuracy.

OOB Loss/N The OOB Loss divided by the number of OOB rows, OOB N.

RSquare (Available only for continuous responses.) The RSquare value for the tree.

IB SSE (Available only for continuous responses.) Sum of squared errors for the In Bag rows.

IB SSE/N (Available only for continuous responses.) Sum of squared errors for the In Bag rows divided by the number of In Bag observations. The number of In Bag observations is equal to the number of observations in the training set multiplied by the bootstrap sampling rate that you specify in the Bootstrap Forest Specification window.

OOB N (Available only for continuous responses.) The number of Out Of Bag rows.

OOB SSE (Available only for continuous responses.) Sum of squared errors when the tree is applied to the Out Of Bag rows.

OOB SSE/N (Available only for continuous responses.) The OOB SSE divided by the number of OOB rows, OOB N.

JMP PRO Bootstrap Forest Platform Options

The Bootstrap Forest red triangle menu contains the following options:

Plot Actual by Predicted (Available only for continuous responses.) Provides a plot of actual versus predicted values.

Column Contributions Displays a report that shows each input column's contribution to the fit. The report also shows:

- The total number of instances over all of the trees when the specified column is used to split the data.
- The total G^2 (for a categorical response) or SS, sum of squares (for a continuous response), attributed to the column.
- A bar chart of G^2 or SS.
- The proportion of G^2 or SS attributed to the column.

Show Trees Provides various options for displaying trees in the Tree Views report. The report gives a picture of the tree that is fit at each layer of the boosting process. For a description of the Prob column shown by the Show names categories estimates option, see [“Predicted Probabilities in Tree Methods”](#).

ROC Curve (Available only for categorical responses.) See [“ROC Curve”](#).

Lift Curve (Available only for categorical responses.) See [“Lift Curve”](#).

Decision Threshold (Available only for binary categorical responses.) Shows or hides Decision Thresholds reports for the training, validation, and test sets, if specified. Each report contains a graph of the distribution of fitted probabilities for each model, confusion matrices for each model, classification graphs to compare the model fits, and a table of classification accuracy metrics. See [“Decision Thresholds Report”](#).

Save Columns Contains options for saving model and tree results, and creating SAS code.

Save Predicteds Saves the predicted values from the model to the data table.

Save Prediction Formula Saves the prediction formula to a column in the data table. The formula consists of nested conditional clauses that describe the tree structure. If the response is continuous, the column contains a Predicting property. If the response is categorical, the column contains a Response Probability property.

Save Tolerant Prediction Formula (The Save Prediction Formula option should be used instead of this option. Use this option only when Save Prediction Formula is not available.) Saves a formula that predicts even when there are missing values and when Informative Missing has not been selected. The prediction formula tolerates missing values by randomly allocating response values for missing predictors to a split. If the response is continuous, the column contains a Predicting property. If the response is categorical, the column contains a Response Probability property. If you have selected Informative Missing, you can save the Tolerant Prediction Formula by holding the Shift key as you click the report's red triangle.

Save Residuals (Available only for continuous responses.) Saves the residuals to the data table.

Save Cumulative Details (Available only if validation is used.) Creates a data table containing the fit statistics for each tree.

Publish Prediction Formula Creates a prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See [“Formula Depot”](#).

Publish Tolerant Prediction Formula (The Publish Prediction Formula option should be used instead of this option. Use this option only when Publish Prediction Formula is not available.) Creates a tolerant prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See [“Formula Depot”](#). If you have selected Informative Missing, you can use this option by holding the Shift key as you click the report’s red triangle.

Make SAS DATA Step Creates SAS code for scoring a new data set.

Specify Profit Matrix (Available only for categorical responses.) Enables you to specify profit or costs associated with correct or incorrect classification decisions. See [“Show Fit Details”](#).

Profiler Shows a Prediction Profiler. See *Profilers*.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Chapter 6

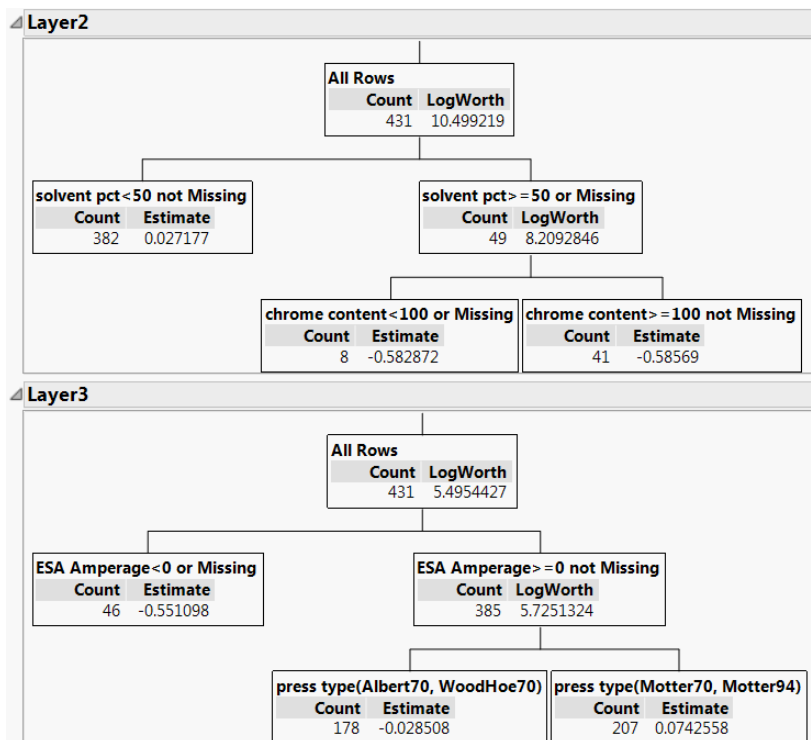
JMP PRO Boosted Tree

Fit Many Layers of Trees, Each Based on the Previous Layer

The Boosted Tree platform is available only in JMP Pro.

Boosting is the process of building a large, additive decision tree by fitting a sequence of smaller decision trees, called *layers*. The tree at each layer consists of a small number of splits. The tree is fit based on the residuals of the previous layers, which allows each layer to correct the fit for bad fitting data from the previous layers. The final prediction for an observation is the sum of the predictions for that observation over all of the layers.

Figure 6.1 Example of Boosted Tree Layers



Contents

- Overview of the Boosted Tree Platform..... 107
- Example of Boosted Tree with a Categorical Response..... 107
- Example of Boosted Tree with a Continuous Response 109
- Launch the Boosted Tree Platform..... 111
 - Specification Window..... 113
- The Boosted Tree Report 115
 - Model Validation Set Summaries 116
 - Specifications 117
 - Overall Statistics 117
 - Cumulative Validation 118
- Boosted Tree Platform Options..... 119
- Statistical Details for the Boosted Tree Platform 121
 - Statistical Details for the Overfit Penalty..... 121

Overview of the Boosted Tree Platform

The Boosted Tree platform produces an additive decision tree model that is based on many smaller decision trees that are constructed in *layers*. The tree in each layer consists of a small number of splits, typically five or fewer. Each layer is fit using the recursive fitting methodology described in “[Partition Models](#)”. The only difference is that fitting stops at a specified number of splits. For a given tree, the predicted value for an observation in a leaf is the mean of all observations in that leaf.

This is the fitting process:

1. Fit an initial layer.
2. Compute residuals. These are obtained by subtracting the predicted mean for observations within a leaf from their actual value.
3. Fit a layer to the residuals.
4. Construct the additive tree. For a given observation, sum its predicted values over the layers.
5. Repeat [step 2](#) to [step 4](#) until the specified number of layers is reached, or, if validation is used, until fitting an additional layer no longer improves the validation statistic.

The final prediction is the sum of the predictions for an observation over all the layers.

By fitting successive layers on residuals from previous layers, each layer can improve the fit.

For categorical responses, only those with two response levels are supported. For a categorical response, the residuals fit at each layer are offsets of linear logits. The final prediction is a logistic transformation of the sum of the linear logits over all the layers.

For more information about boosted trees, see Hastie et al. ([2009](#)).

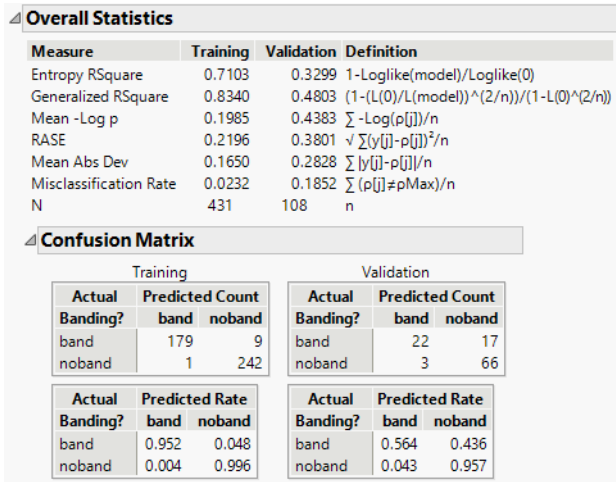
Example of Boosted Tree with a Categorical Response

In this example, you construct a boosted tree model to predict which printing jobs are affected by a defect called banding.

1. Select **Help > Sample Data Folder** and open Bands Data.jmp.
2. Select **Analyze > Predictive Modeling > Boosted Tree**.
3. Select Banding? and click **Y, Response**.
4. Select the Predictors column group and click **X, Factor**.
5. Enter 0.2 for **Validation Portion**.

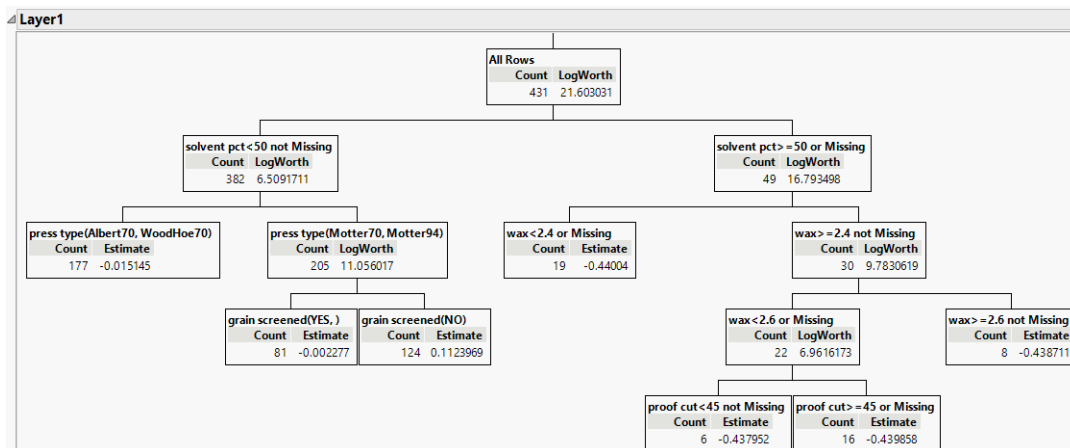
- 6. Click **OK**.
The Boosted Tree Specification window appears.
- 7. (Optional) In the Reproducibility panel, select **Suppress Multithreading** and enter 123 for Random Seed.
Because the boosted tree fit involves a random component, these actions ensure that you obtain the exact results shown below.
- 8. Click **OK**.

Figure 6.2 Overall Statistics for Nominal Response



- Because the response, Banding?, is categorical, the Boosted Tree analysis provides a Misclassification Rate under Measure and a Confusion Matrix report. The Misclassification Rate for the validation set is 0.1852, or about 19%.
- 9. Click the red triangle next to Boosted Tree for Banding? and select **Show Trees > Show names categories estimates**.
A Tree Views report appears, with outlines for the layers. You can examine the layers to see the trees that are fit and the predicted values.

Figure 6.3 Layer 1 of the Boosted Tree



- Click the red triangle next to Boosted Tree for Banding? and select **Save Columns > Save Prediction Formula**.

Columns called Prob(Banding?==noband), Prob(Banding?==band), and Most Likely Banding? are added to the data table. Examine the Prob(Banding?==noband) column to see how model predictions are calculated from the layers.

JMP PRO Example of Boosted Tree with a Continuous Response

In this example, you construct a boosted tree model to predict the percent body fat given a combination of nominal and continuous factors.

- Select **Help > Sample Data Folder** and open the Body Fat.jmp sample data table.
- Select **Analyze > Predictive Modeling > Boosted Tree**.
- Select Percent body fat and click **Y, Response**.
- Select Age (years) through Wrist circumference (cm) and click **X, Factor**.
- Select Validation and click **Validation**.
- Click **OK**.
- Click **OK**.

Figure 6.4 Overall Statistics for Continuous Response

Overall Statistics			
	RSquare	RASE	N
Training	0.847	3.2116281	180
Validation	0.611	5.4301412	72

The Overall Statistics report provides the R-square and RASE for the boosted tree model. The R-square for the validation set is 0.611. The RASE for the validation set is about 5.43.

You are interested in obtaining a model-independent indication of the important predictors for Percent body fat.

- 8. Click the red triangle next to Boosted Tree for Percent body fat and select **Profiler**.
- 9. Click the red triangle next to Prediction Profiler and select **Assess Variable Importance > Independent Uniform Inputs**.

Note: Because Assess Variable Importance uses randomization, your results might not exactly match those in [Figure 6.5](#).

Figure 6.5 Summary Report for Variable Importance

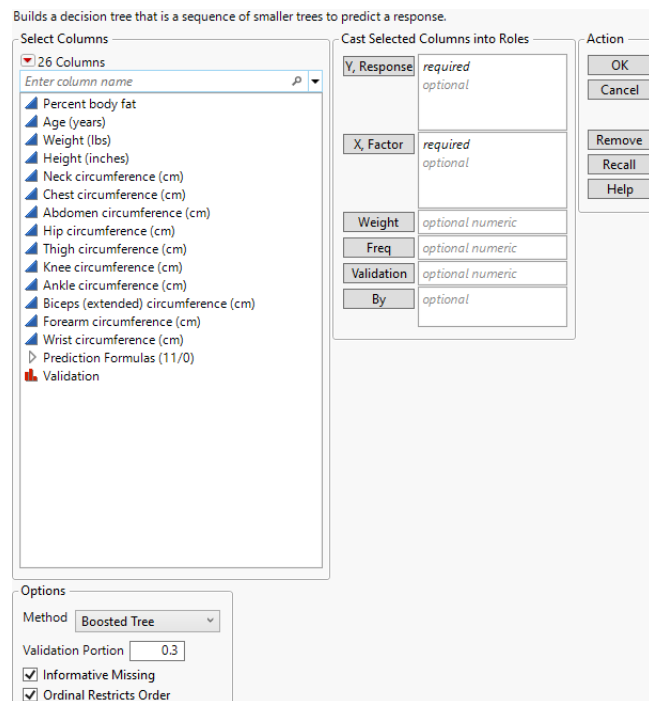
Summary Report						
Column	Main Effect	Total Effect	.2	.4	.6	.8
Abdomen circumference (cm)	0.865	0.887				
Age (years)	0.063	0.085				
Wrist circumference (cm)	0.007	0.013				
Biceps (extended) circumference (cm)	0.003	0.006				
Hip circumference (cm)	0.002	0.005				
Chest circumference (cm)	0.002	0.004				
Height (inches)	0.002	0.003				
Ankle circumference (cm)	0.001	0.003				
Thigh circumference (cm)	0.001	0.002				
Weight (lbs)	0.001	0.002				
Knee circumference (cm)	3e-4	0.001				
Neck circumference (cm)	8e-18	2e-17				
Forearm circumference (cm)	8e-18	2e-17				

The Summary Report shows that Abdomen circumference (cm) is the most important predictor of Percent body fat.

Launch the Boosted Tree Platform

Launch the Boosted Tree platform by selecting **Analyze > Predictive Modeling > Boosted Tree**.

Figure 6.6 Boosted Tree Launch Window Using Body Fat.jmp



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Boosted Tree platform launch window has the following options:

Y, Response The response variable or variables that you want to analyze.

X, Factor The predictor variables.

Weight A column whose numeric values assign a weight to each row in the analysis.

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

Validation A numeric column that defines the validation sets. This column should contain at most three distinct values:

- If the validation column has two levels, the smaller value defines the training set and the larger value defines the validation set.

- If the validation column has three levels, the values, in order of increasing size, define the training, validation, and test sets.
- If the validation column has more than three levels, the rows that contain the smallest three values define the validation sets. All other rows are excluded from the analysis.

The Boosted Tree platform uses the validation column to train and tune the model or to train, tune, and evaluate the model. For more information about validation, see [“Validation in JMP Modeling”](#).

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 [“Make Validation Column”](#).

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.

Method Enables you to select the partition method (Decision Tree, Bootstrap Forest, Boosted Tree, K Nearest Neighbors, or Naive Bayes). These alternative methods, except for Decision Tree, are available in JMP Pro.

For more information about these methods, see [“Partition Models”](#), [“Bootstrap Forest”](#), [“K Nearest Neighbors”](#), and [“NaiveBayes”](#).

Validation Portion The portion of the data to be used as the validation set.

Informative Missing If selected, enables missing value categorization for categorical predictors and informative treatment of missing values for continuous predictors. See [“ROC Curve”](#).

Ordinal Restricts Order If selected, restricts consideration of splits to those that preserve the ordering.

JMP^{PRO} Specification Window

After you select OK in the Boosted Tree launch window, the Gradient-Boosted Trees Specification window appears.

Figure 6.7 Boosted Tree Specification Window

Gradient-Boosted Trees Specification

Boosting

Number of Layers: 152

Splits per Tree: 2

Learning Rate: 0.066

Minimum Size Split: 5

Multiple Fits

☐ Multiple Fits over Splits and Learning Rate

Max Splits Per Tree: 3

Max Learning Rate: 0.1

☐ Use Tuning Design Table

Stochastic Boosting

Row Sampling Rate: 1.0000

Column Sampling Rate: 1.0000

Reproducibility

☒ Suppress Multithreading

Random Seed: 0

☒ Early Stopping

OK Cancel

JMP^{PRO} Boosting Panel

Number of Layers The maximum number of layers to include in the final tree.

Splits per Tree The number of splits for each layer.

Learning Rate A number such that $0 < r \leq 1$. Learning rates close to 1 result in faster convergence on a final tree, but also have a higher tendency to overfit data. Use learning rates closer to 1 when a small Number of Layers is specified. The learning rate is a small fraction typically between 0.01 and 0.1 that slows the convergence of the model. This preserves opportunities for later layers to use different splits than the earlier layers.

Overfit Penalty (Available only for categorical responses.) A biasing parameter that helps protect against fitting probabilities equal to zero. See [“Statistical Details for the Overfit Penalty”](#).

Minimum Size Split The minimum number of observations needed on a candidate split.

JMP^{PRO} Multiple Fits Panel

Multiple Fits over Splits and Learning Rate If selected, creates a boosted tree for every combination of Splits per Tree (in integer increments) and Learning Rate (in 0.1 increments).

The lower bounds for the combinations are specified by the Splits per Tree and Learning Rate options. The upper bounds for the combinations are specified by the following options:

Max Splits per Tree Upper bound for Splits per Tree.

Max Learning Rate Lower bound for Learning Rate.

Use Tuning Design Table Opens a window where you can select a data table containing values for some tuning parameters, called a *tuning design table*. A tuning design table has a column for each option that you want to specify and has one or multiple rows that each represent a single Boosted Tree model design. If an option is not specified in the tuning design table, the default value is used.

For each row in the table, JMP creates a Boosted Tree model using the tuning parameters specified. If more than one model is specified in the tuning design table, the Model Validation-Set Summaries report lists the R-Square value for each model. The Boosted Tree report shows the fit statistics for the model with the largest R-Square value.

You can create a tuning design table using the Design of Experiments facilities. A boosted tree tuning design table can contain the following case-insensitive columns in any order:

- Number of Layers
- Splits per Tree
- Learning Rate
- Minimum Size Split
- Row Sampling Rate
- Column Sampling Rate

JMP PRO Stochastic Boosting Panel

Row Sampling Rate Proportion of training rows to sample for each layer.

Note: When the response is categorical, the training rows are sampled using stratified random sampling.

Column Sampling Rate Proportion of predictor columns to sample for each layer.

JMP PRO Reproducibility Panel

Suppress Multithreading If selected, all calculations are performed on a single thread.

Random Seed Specify a nonzero numeric random seed to reproduce the results for future launches of the platform. By default, the Random Seed is set to zero, which does not produce reproducible results. When you save the analysis to a script, the random seed that you enter is saved to the script.

JMP PRO Early Stopping

Early Stopping If selected, the boosting process stops fitting additional layers when the additional layers do not improve the validation statistic. If not selected, the boosting process continues until the specified number of layers is reached. This option appears only if validation is used.

JMP PRO The Boosted Tree Report

After you click OK in the Gradient-Boosted Trees Specification window, the Boosted Tree report opens.

Figure 6.8 Boosted Tree Report for a Continuous Response

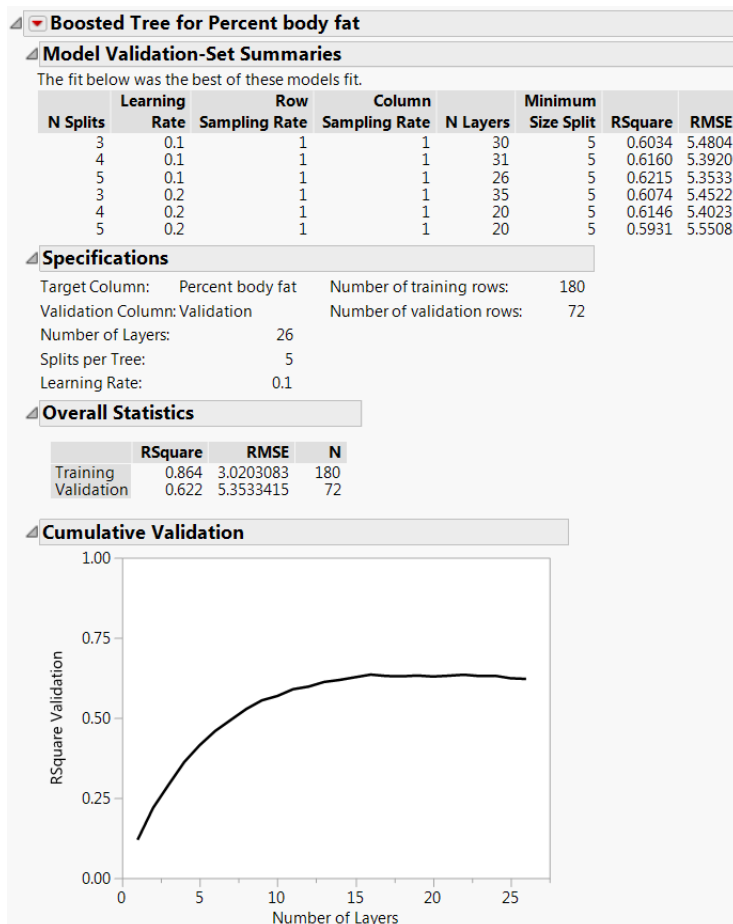
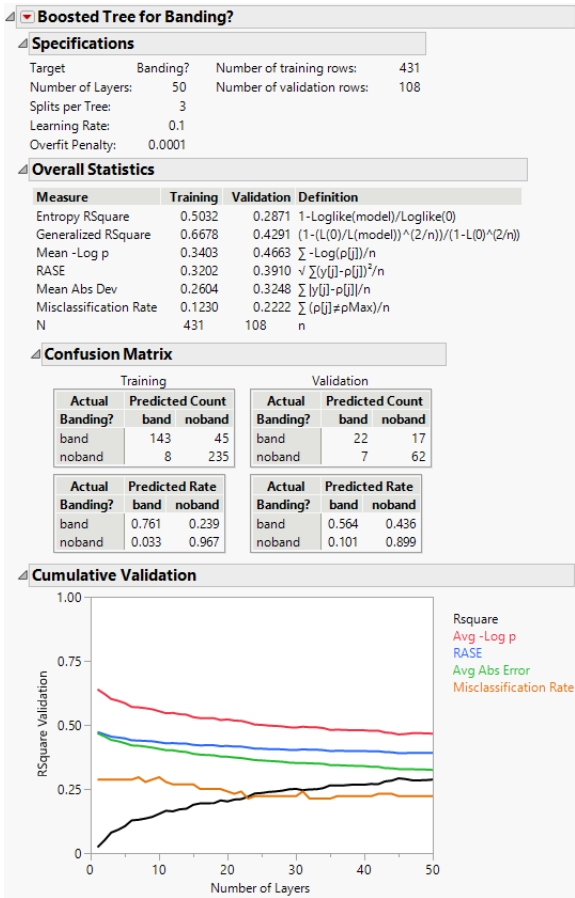


Figure 6.9 Boosted Tree Report for a Categorical Response



The following reports are provided, depending on whether the response is categorical or continuous:

- “Model Validation Set Summaries”
- “Specifications”
- “Overall Statistics”
- “Cumulative Validation”

JMP^{PRO} Model Validation Set Summaries

In the Boosted Tree platform, the Model Validation Set Summaries report shows fit statistics for all the models fit if you selected the Multiple Fits over Splits and Learning Rate option in the Specification window. See [Figure 6.8](#) and “Multiple Fits Panel”.

JMP PRO Specifications

In the Boosted Tree platform, the Specifications report shows the settings used in fitting the model.

JMP PRO Overall Statistics

In the Boosted Tree platform, the Overall Statistics report shows fit statistics for the training set, and for the validation and test sets if they are specified.

Suppose that you fit multiple models using the Multiple Fits over Splits and Learning Rate option in the Boosted Tree Specification window. Then the model for which results are displayed in the Overall Statistics and Cumulative Validation reports is the model for which the validation set's Entropy R-square value (for a categorical response) or R-square (for a continuous response) is the largest.

JMP PRO Measures Report

(Available only for categorical responses.) Gives the following statistics for the training set, and for the validation and test sets if they are specified.

Note: For Entropy R-Square and Generalized R-Square, values closer to 1 indicate a better fit. For Mean -Log p, RASE, Mean Abs Dev, and Misclassification Rate, smaller values indicate a better fit.

Entropy RSquare A measure of fit that compares the log-likelihoods from the fitted model and the constant probability model. It ranges from 0 to 1. See [“Entropy RSquare”](#).

Generalized RSquare A measure that can be applied to general regression models. It is based on the likelihood function L and is scaled to have a maximum value of 1. The value is 1 for a perfect model, and 0 for a model no better than a constant model. The Generalized R-Square measure simplifies to the traditional R-Square for continuous normal responses in the standard least squares setting. Generalized R-Square is also known as the Nagelkerke or Craig and Uhler R^2 , which is a normalized version of Cox and Snell's pseudo R^2 .

Mean -Log P The average of negative $\log(p)$, where p is the fitted probability associated with the event that occurred.

RASE The root average squared prediction error. The differences are between 1 and p , the fitted probability for the response level that actually occurred.

Mean Abs Dev The average of the absolute values of the differences between the response and the predicted response. The differences are between 1 and p , the fitted probability for the response level that actually occurred.

Misclassification Rate The rate for which the response category with the highest fitted probability is not the observed category.

N The number of observations.

JMP PRO Confusion Matrix Report

(Available only for categorical responses.) Shows classification statistics for the training set, and for the validation and test sets if they are specified. The Confusion Matrix Report contains confusion matrices and confusion rates matrices. A confusion matrix is a two-way classification of actual and predicted responses. A confusion rates matrix is equal to the confusion matrix, with the numbers divided by the row totals.

JMP PRO Decision Matrix

(Available only for categorical responses and if the response has a Profit Matrix column property or if you specify costs using the Specify Profit Matrix option.) Gives Decision Count and Decision Rate matrices for the training set, and for the validation and test sets if they are specified. See [“Additional Examples of the Partition Platform”](#).

JMP PRO Cumulative Validation

In the Boosted Tree platform, the Cumulative Validation report shows a plot of the fit statistics for the Validation set versus the number of layers. This report is available only if validation is used.

For a continuous response, the single fit statistic is R-Square. For a categorical response, the fit statistics are listed below and are described in [“Measures Report”](#).

- R-Square (Entropy R-Square)
- Avg - Log p (Mean - Log p)
- Root Average Squared Prediction Error (RASE)
- Avg Abs Error (Mean Abs Dev)
- MR (Misclassification Rate)

The Cumulative Details report below the Cumulative Validation plot gives the values used in the plot.

Boosted Tree Platform Options

The Boosted Tree red triangle menu contains the following options:

Show Trees Provides options for displaying trees in the Tree Views report. The report gives a picture of the tree that is fit at each layer of the boosting process.

Plot Actual by Predicted (Available only for continuous responses.) Provides a plot of actual versus predicted values.

Column Contributions Displays a report showing each input column's contribution to the fit. The report also shows:

- The total number of instances over all of the trees when the specified column is used to split the data.
- The total G^2 (for a categorical response) or SS, sum of squares (for a continuous response) attributed to the column.
- A bar chart of G^2 or SS.
- The proportion of G^2 or SS attributed to the column.

ROC Curve (Available only for categorical responses.) See [“ROC Curve”](#).

Lift Curve (Available only for categorical responses.) See [“Lift Curve”](#).

Decision Threshold (Available only for binary categorical responses.) Shows or hides Decision Thresholds reports for the training, validation, and test sets, if specified. Each report contains a graph of the distribution of fitted probabilities for each model, confusion matrices for each model, classification graphs to compare the model fits, and a table of classification accuracy metrics. See [“Decision Thresholds Report”](#).

Save Columns Contains options for saving model and tree results, and creating SAS code.

Save Predicteds Saves the predicted values from the model to the data table.

Save Prediction Formula Saves the prediction formula to a column in the data table. The formula consists of nested conditional clauses that describe the tree structure. If the response is continuous, the column contains a Predicting column property. If the response is categorical, the column contains a Response Probability column property.

Save Tolerant Prediction Formula (The Save Prediction Formula option should be used instead of this option. Use this option only when Save Prediction Formula is not available.) Saves a formula that predicts even when there are missing values and when Informative Missing has not been selected. The prediction formula tolerates missing values by randomly allocating response values for missing predictors to a split. If the response is continuous, the column contains a Predicting column property. If the response is categorical, the column contains a Response Probability column property. If you have selected Informative Missing, you can save the Tolerant Prediction Formula by holding the Shift key as you click the report's red triangle.

Save Residuals (Available only for continuous responses.) Saves the residuals to the data table.

Save Offset Estimates (Available only for categorical responses.) Saves the sums of the linear components. These are the logits of the fitted probabilities.

Save Tree Details Creates a data table containing split details and estimates for each layer.

Save Cumulative Details (Available only if validation is used.) Creates a data table containing the fit statistics for each layer.

Publish Prediction Formula Creates a prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See [“Formula Depot”](#).

Publish Tolerant Prediction Formula (The Publish Prediction Formula option should be used instead of this option. Use this option only when Publish Prediction Formula is not available.) Creates a tolerant prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See [“Formula Depot”](#). If you have selected Informative Missing, you can use this option by holding the Shift key as you click the report's red triangle.

Make SAS DATA Step Creates SAS code for scoring a new data set.

Specify Profit Matrix (Available only for categorical responses.) Enables you to specify profit or costs associated with correct or incorrect classification decisions. See [“Show Fit Details”](#).

Profiler Shows a Prediction Profiler. See *Profilers*.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Statistical Details for the Boosted Tree Platform

This section describes details specific to the Boosted Tree Platform. For statistical details about recursive decision trees, see [“Statistical Details for the Partition Platform”](#).

Statistical Details for the Overfit Penalty

When the response is categorical, a parametric penalty is imposed. For each layer, the estimates minimize the negative log-likelihood plus the penalty value multiplied by the sum of squares of the estimates for each observation. This penalty encourages each new layer not to overfit the training data.

Chapter 7

JMP PRO K Nearest Neighbors

Predict Response Values Using Nearby Observations

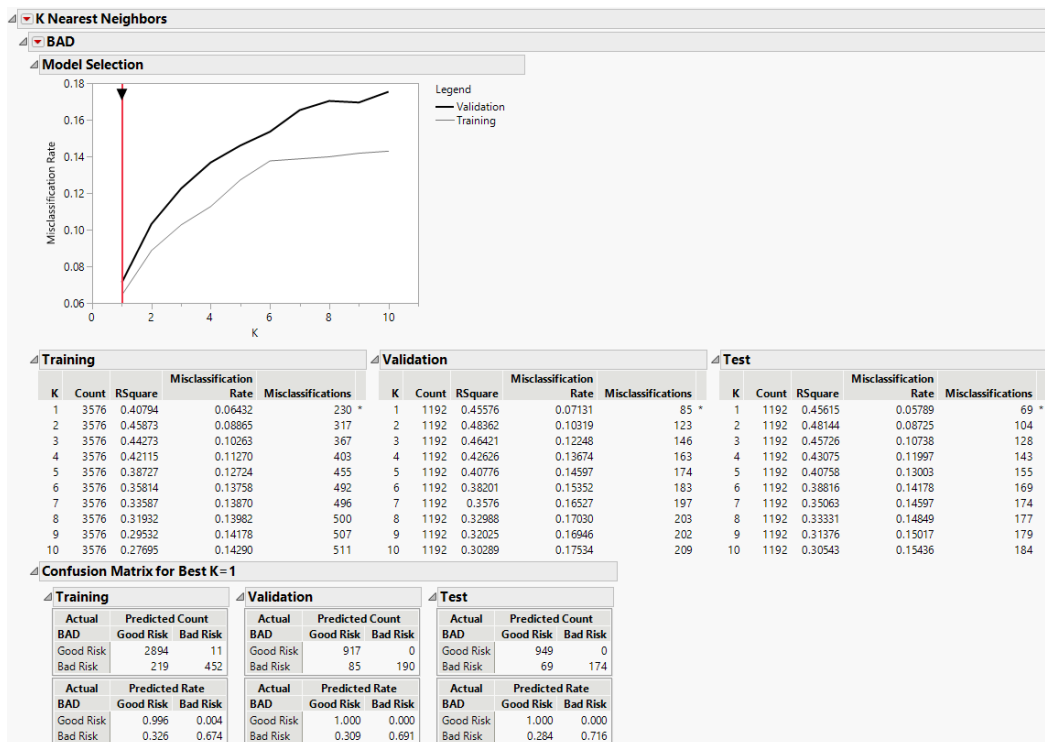
The K Nearest Neighbors platform is available only in JMP Pro.

The K Nearest Neighbors platform predicts a response value for a given observation using the responses of the observations in that observation's local neighborhood. It can be used for classification of a categorical response as well as for prediction of a continuous response.

K Nearest Neighbors is a nonparametric method that is based on the distance to neighboring observations. Because of this fact, K Nearest Neighbors is able to classify observations using irregular predictor value boundaries. However, because the algorithm is sensitive to irrelevant predictors, the selection of predictors can impact your results.

K Nearest Neighbors has been used successfully in many applications, such as classifying satellite imagery and electrocardiogram (EKG) patterns.

Figure 7.1 Example of the K Nearest Neighbors Platform



Contents

Overview of the K Nearest Neighbors Platform 125

Example of K Nearest Neighbors with a Categorical Response 126

Example of K Nearest Neighbors with a Continuous Response 128

Launch the K Nearest Neighbors Platform..... 130

The K Nearest Neighbors Report 132

 Continuous Responses 132

 Categorical Responses 133

K Nearest Neighbors Platform Options..... 134

 Model Fit Options 135



Overview of the K Nearest Neighbors Platform

The K Nearest Neighbors platform predicts a response value based on the responses of the k nearest neighbors. The k nearest neighbors to a given observation are determined by identifying the k smallest Euclidean distances between the predictor values for that observation and the predictor values for each of the other observations. The K Nearest Neighbors platform models both continuous and categorical responses.

A potential drawback of the k nearest neighbors method is that for large scale problems, the prediction formula is often complex and hard to interpret, limiting its usefulness. In addition, K Nearest Neighbors does not calculate probabilities for categorical responses. For more information about the k nearest neighbors method, see Hastie et al. (2009), Hand et al. (2001), and Shmueli et al. (2017).

Continuous Responses

For a continuous response, the predicted value is the average of the responses for the k nearest neighbors. Each continuous predictor is scaled by its standard deviation. With this scaling, a single predictor with a large range does not excessively influence the distance calculation. Missing values for a continuous predictor are replaced by the mean of that predictor. See [“Example of K Nearest Neighbors with a Continuous Response”](#).

Categorical Responses

For a categorical response, the predicted value is the most frequent response level for the k nearest neighbors. If two or more levels are tied as the most frequent levels, the predicted response is assigned by selecting one of these levels at random.

Note: Because ties for most frequent levels in the case of a categorical response are broken at random, results from independent runs of the platform might differ. To obtain reproducible results, use the Set Random Seed option in the launch window or include the `Set Random Seed()` function in a JSL script.

In the categorical prediction models, each categorical predictor is expressed in terms of indicator variables, with one indicator variable representing each level. A row with a missing value for a categorical predictor is represented by values of zero on all indicator variables for that predictor.



Example of K Nearest Neighbors with a Categorical Response

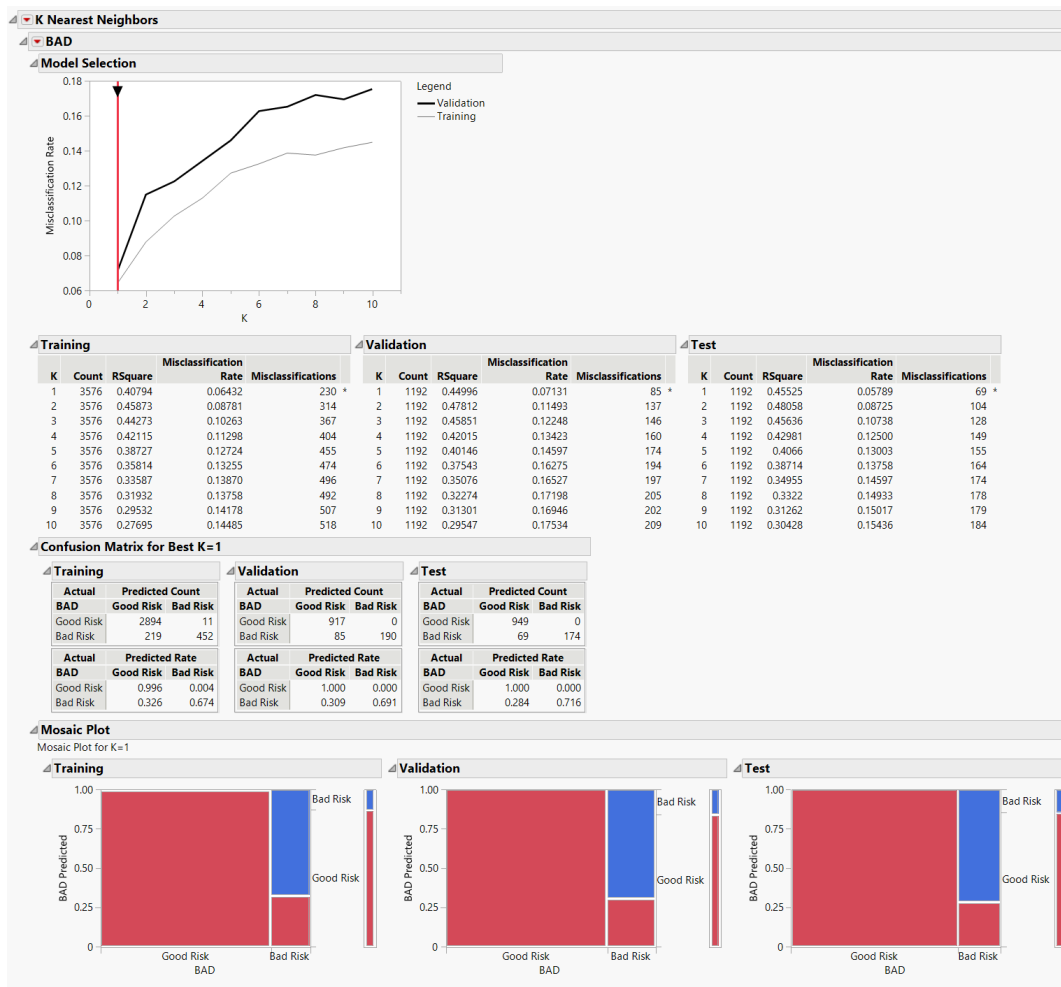
In this example, you want to construct a model to classify the credit risk of customers who apply for home equity loans. You have historical financial data for 5,960 customers where each customer was classified as being a Good Risk or Bad Risk. There are missing values for many of the predictors.

1. Select **Help > Sample Data Folder** and open Equity.jmp.
2. Select **Analyze > Predictive Modeling > K Nearest Neighbors**.
3. Select BAD and click **Y, Response**.
4. Select LOAN through CLNO and click **X, Factor**.

Because one of the potential predictors, DEBTINC, has many missing values, you do not include it in your model. Missing values for continuous predictors are replaced by the average of the predictor. This procedure sometimes works well for values that are missing at random. Although the high missing rate of the DEBTINC indicates that the missing might be informative, we do not investigate that in this example.

5. Select Validation and click **Validation**.
6. Click **OK**.

Figure 7.2 K Nearest Neighbors Report



For each value of K , JMP constructs a model using only the training set observations. Each of these models is used to classify the validation set observations. The validation set results are used to select a best model. In this example, the model based on the single nearest neighbor ($K = 1$) has the smallest misclassification rate. The test set verifies that the single nearest neighbor model is the best performer for independent data.

- Click the BAD red triangle and select **Publish Prediction Formula**.
- Next to **Number of Neighbors, K**, leave the default value of 1.
- Click **OK**.

The prediction equation is saved in the Formula Depot. You can compare the performance of alternative models published to the Formula Depot with that of the $K = 1$ nearest

neighbor model using the Model Comparison option in the Formula Depot. See [“Formula Depot”](#).

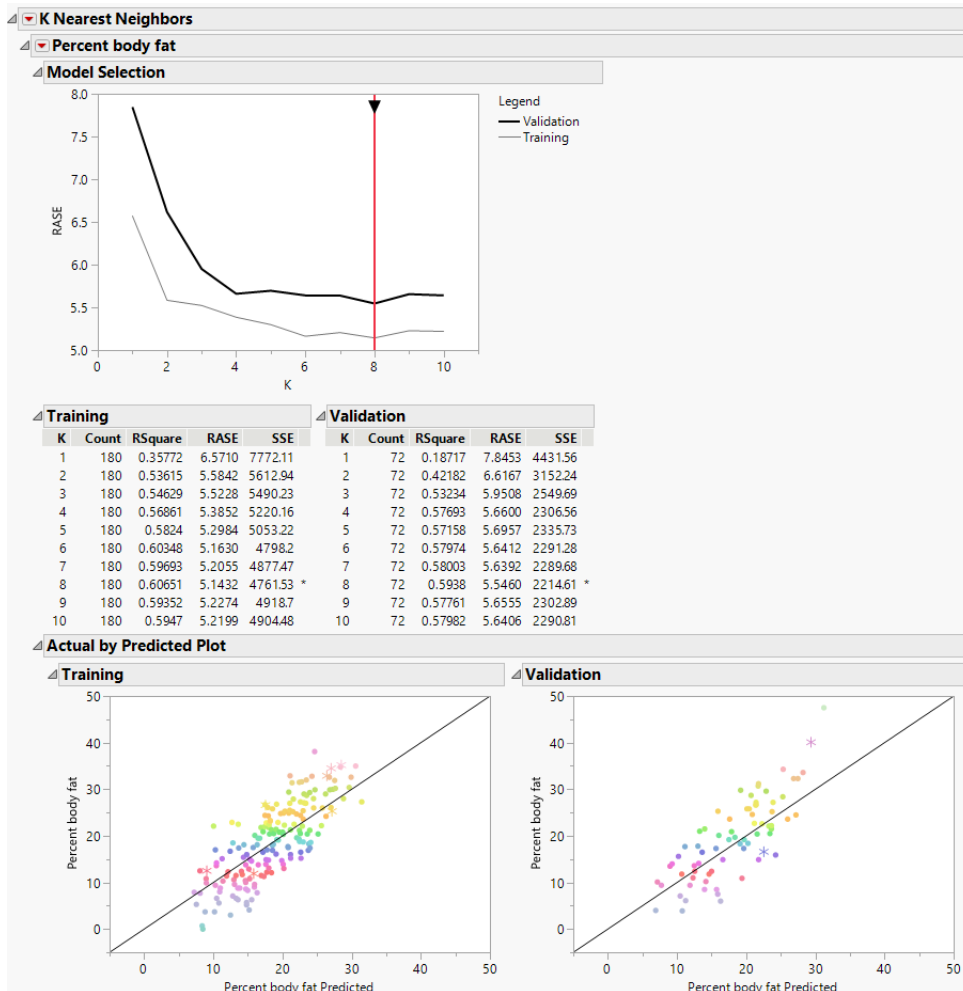


Example of K Nearest Neighbors with a Continuous Response

In this example, you want to predict the percent body fat for males using 13 predictors. The sample data table contains percent body fat estimates that are based on underwater weighing and on various body circumference measurements.

1. Select **Help > Sample Data Folder** and open Body Fat.jmp.
2. Select **Analyze > Predictive Modeling > K Nearest Neighbors**.
3. Select Percent body fat and click **Y, Response**.
4. Select Age (years) through Wrist circumference (cm) and click **X, Factor**.
5. Select Validation and click **Validation**.
6. Click **OK**.
7. Click the Percent body fat red triangle and select **Plot Actual by Predicted**.

Figure 7.3 K Nearest Neighbors Report



The $K = 8$ model had the lowest RASE for the validation set. Among k nearest neighbor models, the model based on 8 nearest neighbors seems to perform the best. The Actual by Predicted plot for the training set shows that the points fall along the line, signifying that the predicted values are similar to the actual values. Most of the points on the plot for the validation set fall along the line, with a few observations on the upper end that are a bit farther away.



Launch the K Nearest Neighbors Platform

Launch the K Nearest Neighbors platform by selecting **Analyze > Predictive Modeling > K Nearest Neighbors**.

Figure 7.4 K Nearest Neighbors Launch Window

Predicts a response based on the responses of the k nearest neighbors.

Select Columns	Cast Selected Columns into Roles	Action
▼ 14 Columns ■ BAD ▲ LOAN ▲ MORTDUE ▲ VALUE ■ REASON ■ JOB ▲ YOJ ▲ DEROG ▲ DELINQ ▲ CLAGE ▲ NINQ ▲ CLNO ▲ DEBTINC ■ Validation	Y, Response optional X, Factor required optional Validation optional numeric By optional	OK Cancel Remove Recall Help

Validation Portion
 K
 Category Bias
 Set Random Seed

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The K Nearest Neighbors launch window provides the following options:

Y, Response The response variable or variables that you want to analyze.

Note: The K Nearest Neighbors platform can be used as a utility to determine the distances between neighboring observations, even without the presence of a response variable. If you do not specify a response variable, a blank report appears. However, the red triangle menu options Save Near Neighbor Rows and Save Near Neighbor Distances are available.

X, Factor The predictor variables.

Validation A numeric column that defines the validation sets. This column should contain at most three distinct values:

- If the validation column has two levels, the smaller value defines the training set and the larger value defines the validation set.

- If the validation column has three levels, the values, in order of increasing size, define the training, validation, and test sets.
- If the validation column has more than three levels, the rows that contain the smallest three values define the validation sets. All other rows are excluded from the analysis.

The K Nearest Neighbors platform uses the validation column to train and tune the model or to train, tune, and evaluate the model. For more information about validation, see [“Validation in JMP Modeling”](#).

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 [“Make Validation Column”](#).

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.

Validation Portion The portion of the data to be used as the validation set.

Number of Neighbors, K Maximum number of nearest neighbors to analyze. Models are fit for one nearest neighbor up to the value that you specify for K.

Note: The maximum number of neighbors, K , must be no larger than one less than the number of rows in the training data table. If you specify a K that is larger than the maximum allowable K , a warning appears.

Category Bias Specifies a tuning parameter that ensures that the fitted probabilities for categorical responses are always positive. A proportion of the bias value is added to the counts of observations to ensure that there is never a count of zero. This procedure avoids computing the logarithm of zero, which is what ensures that the fitted probabilities are always positive. The Category Bias value must be between 0 and 1. The default value is 0.5.

Set Random Seed Sets the seed for the randomization process used in tie-breaking for nominal and ordinal responses. If you specify a Validation Portion, this option also sets the seed for the rows used for validation. Set Random Seed is useful if you want to reproduce an analysis. If you set a random seed and save the script, the seed is automatically saved in the script.



The K Nearest Neighbors Report

The K Nearest Neighbors report contains a separate report for each response variable. Each response variable report contains information about the fitted model for that response. This information includes a Model Selection report and summary information for each of the k models that were fit. The report shows tables for the training set and for the validation and test sets if you defined these using validation.

The Model Selection report displays a solution path plot across K based on the Misclassification Rate for categorical responses or the RASE for continuous responses. By default, the slider is placed on the value of K that corresponds to the best performing model. You can drag the slider to change the value of K in the report.

The statistics reported depend on the modeling type of the response. Each row in the summary tables corresponds to a model defined by k nearest neighbors, where K ranges from one to the value that you specified as Number of Neighbors, K in the launch window.



Continuous Responses

By default, in addition to the Model Selection graph, the K Nearest Neighbors report for a continuous response contains a summary table.

Summary Table

An asterisk marks the model for the value of K that has the smallest RASE. The report for a continuous response contains the following columns:

K Number of nearest neighbors used in the model. K ranges from 1 to the Number of Neighbors, K that you specified in the launch window.

Count Number of observations.

RSquare The RSquare value for the model.

RASE Root mean average squared prediction error for the model. The model with the smallest RASE is marked with an asterisk. If there are tied RASE values, the model with the smallest K is marked with the asterisk.

SSE Sum of squared errors for the model.

JMP PRO Categorical Responses

By default, in addition to the Model Selection graph, the K Nearest Neighbors report for a categorical response contains a summary table, a confusion matrix, and a mosaic plot.

Summary Table

An asterisk marks the model for the value of K that has the smallest misclassification rate. The report for a categorical response contains the following columns:

K Number of nearest neighbors used in the model. K ranges from 1 to the Number of Neighbors, K that you specified in the launch window.

Count Number of observations.

RSquare The Entropy RSquare for the model.

Misclassification Rate Proportion of observations misclassified by the model. This is calculated as Misclassifications divided by Count. The model with the smallest misclassification rate is marked with an asterisk. If there are tied misclassification rates, the model with the smallest K is marked with the asterisk.

Misclassifications Number of observations that are incorrectly predicted by the model.

Confusion Matrix

By default, a Confusion Matrix Report is shown for the model with the smallest Misclassification Rate. If there are ties for the smallest misclassification rate, a report is shown for the model with the smallest K . The Confusion Matrix Report contains confusion matrices and confusion rates matrices. A confusion matrix is a two-way classification of actual and predicted responses. A confusion rates matrix is equal to the confusion matrix, with the numbers divided by the row totals. If you use validation, confusion matrices and confusion rates matrices for the validation and test sets appear. Use the Confusion Matrix Report and the misclassification rates to evaluate your model.

Tip: If you change the position of the slider in the solution path plot, an additional Confusion Matrix Report is displayed for the chosen value of K . Use the additional report to compare an alternative model to the default best model.

Mosaic Plot

By default, a mosaic plot is shown for the model with the smallest Misclassification Rate. If there are ties for the smallest misclassification rate, a mosaic plot is shown for the model with the smallest K . A mosaic plot is a stacked bar chart where each segment is proportional to its group's frequency count. For more information about mosaic plots, see *Basic Analysis*. If you use validation, mosaic plots for the validation and test sets are shown.

Tip: If you change the position of the slider in the solution path plot, the mosaic plot updates to display the results for the chosen value of K .



K Nearest Neighbors Platform Options

The K Nearest Neighbors red triangle menu contains the following options:

Save Near Neighbor Rows Saves K columns to the data table. The columns are named RowNear < k >. For a given row, the k^{th} column contains the row number of its k^{th} nearest neighbor.

Caution: The row numbers in the columns RowNear < k > do not update when you reorder the rows in your data table. If you reorder the rows, the values in those columns are incorrect.

Save Near Neighbor Distances Saves K columns to the data table. The columns are named Distance < k >. For a given row, the k^{th} column contains the distance to that row's k^{th} nearest neighbor.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

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

Model Fit Options

The Model Fit section of the K Nearest Neighbors report contains the following red triangle menu options:

Mosaic Plot (Available only for nominal or ordinal responses.) Shows or hides the mosaic plot. See [“Mosaic Plot”](#).

Plot Actual by Predicted (Available only for continuous responses.) Plots the actual versus the predicted responses for the model with the smallest RASE. If there are ties for the smallest RASE, the plot is based on the model with the smallest K .

Tip: If you change the position of the slider on the solution path plot to a different K , the Actual by Predicted plot is updated to reflect the model for the chosen value of K .

Plot Residual by Predicted (Available only for continuous responses.) Plots the residuals versus the predicted responses for the model with the smallest RASE. If there are ties for the smallest RASE, the plot is based on the model with the smallest K .

Tip: If you change the position of the slider on the solution path plot to a different K , the Residual by Predicted plot is updated to reflect the model for the chosen value of K .

Save Predicteds Saves K predicted value columns to the data table. The columns are named Predicted <Response> < k >. The k^{th} column contains predictions for the model based on the k nearest neighbors, where Response is the name of the response column. The report statistics that appear in the platform are generated using these raw predictions.

Save Prediction Formula Saves a column that contains a prediction formula for a specific k nearest neighbor model. Enter a value for K when prompted. The prediction formula contains all the training data, so this option might not be practical for large data tables. This option is useful for scoring new observations or predicting missing response values.

Caution: The predicted values for the training data rows that are obtained from Save Prediction Formula and Save Predicteds do not match. The predicted values that are obtained from the Save Prediction Formula option use all of the rows in the training set, including the row for the predicted value. The predicted values that are obtained from the Save Predicteds option do not use the row for the predicted value in the training set, only all other rows. Since the Save Prediction Formula option uses all of the training data, each training data row uses the row itself for the first neighbor and has perfect prediction for $k = 1$. This means that any predictions for the training data might be too accurate and create inflated estimates of model accuracy if used to calculate model statistics. Therefore, the report statistics that appear in the platform are not generated using the predictions from the prediction formula.

Publish Prediction Formula Creates a prediction formula for the specified k nearest neighbor model 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 [“Formula Depot”](#).

Chapter 8

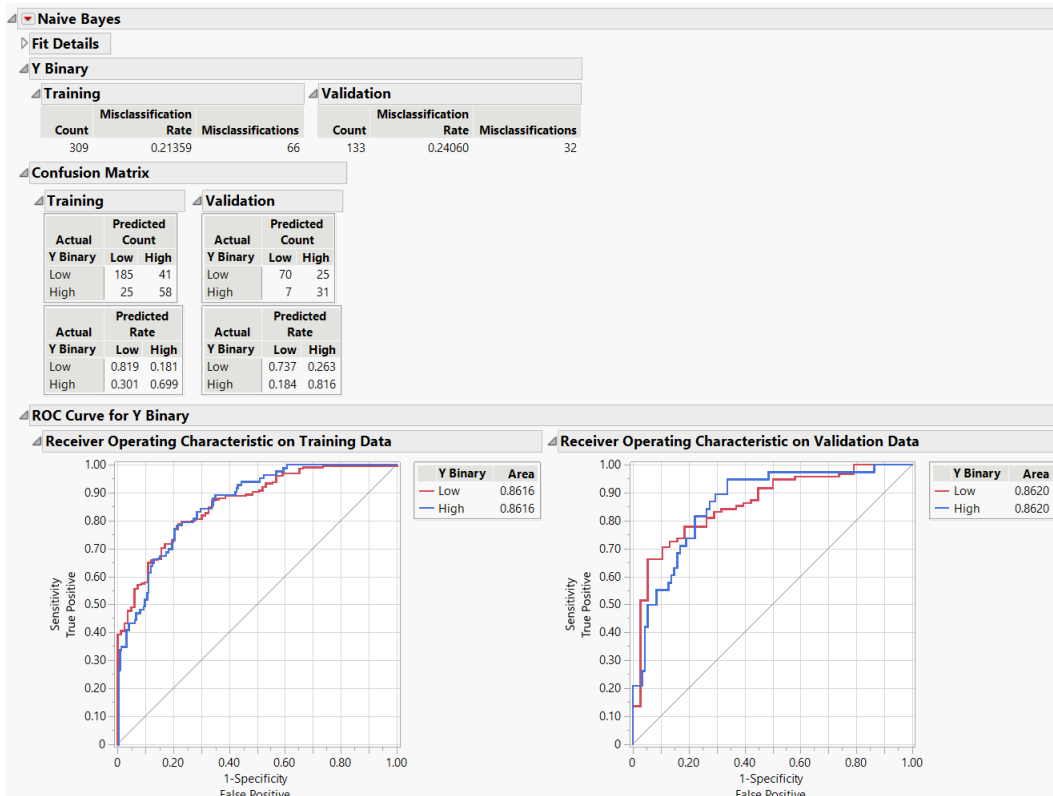
JMP PRO NaiveBayes

Classify Observations Using Probabilistic Assumptions

The Naive Bayes platform is available only in JMP Pro.

The Naive Bayes platform fits a model to predict the value of a categorical variable. Naive Bayes is a fast and computationally simple method. It is especially suitable for situations where there are a large number of predictors.

Figure 8.1 Example of Naive Bayes Analysis



Contents

Overview of the Naive Bayes Platform 139

Example of Naive Bayes..... 140

Launch the Naive Bayes Platform..... 142

The Naive Bayes Report..... 143

 Fit Details Report..... 144

 Response Column Report 144

 Confusion Matrix Report..... 144

 ROC Curves..... 144

Naive Bayes Platform Options 145

Additional Example of Naive Bayes..... 146

Statistical Details for the Naive Bayes Platform 147

 Statistical Details for the Naive Bayes Algorithm 147

 Statistical Details for Saved Probability Formulas..... 149

Overview of the Naive Bayes Platform

The Naive Bayes platform classifies observations into *classes* that are defined by the levels of a categorical response variable. The variables (or factors) that are used for classification are often called *features* in the data mining literature.

For each class, the naive Bayes algorithm computes the conditional probability of each feature value occurring. If a feature is continuous, its conditional marginal density is estimated. The naive Bayes technique assumes that, within a class, the features are independent. (This is the reason that the technique is referred to as “naive”.) Classification is based on the idea that an observation whose feature values have high conditional probabilities within a certain class has a high probability of belonging to that class. See Hastie et al. (2009).

Because the algorithm estimates only one-dimensional densities or distributions, the algorithm is extremely fast. This makes it suitable for large data sets, and in particular, data sets with large numbers of features. All nonmissing feature values for an observation are used in calculating the conditional probabilities.

Each observation is assigned a *naive score* for each class. An observation’s naive score for a given class is the proportion of training observations that belong to that class multiplied by the product of the observation’s conditional probabilities. The *naive probability* that an observation belongs to a class is its naive score for that class divided by the sum of its naive scores across all classes. The observation is assigned to the class for which it has the highest naive probability.

Caution: Because the conditional probabilities of class membership are assumed independent, the naive Bayes estimated probabilities are inefficient.

Naive Bayes requires a large number of training observations to ensure representation for all predictor values and classes. If an observation in the Validation set is being classified and it has a categorical predictor value that was missing in the Training set, then the platform uses the non-missing features to predict. If an observation is missing all predictor values, the predicted response is the most frequent response. The prediction formulas handle missing values by having them contribute nothing to the observation’s score.

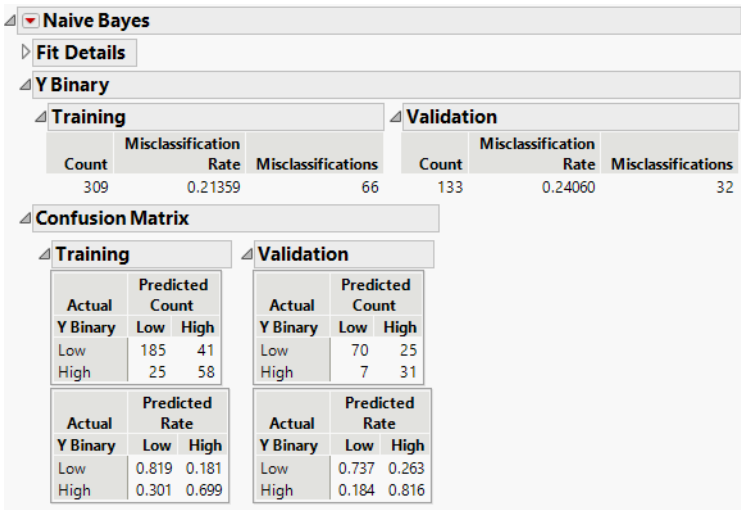
For more information about the naive Bayes technique, see Hand et al. (2001), and Shmueli et al. (2010).

JMP[®] PRO **Example of Naive Bayes**

You want to construct a classification model to be used in predicting the disease progression for future patients as High or Low. You have baseline medical data for 442 diabetic patients. You also have a binary measure of diabetes disease progression obtained one year after each patient’s initial visit. This measure quantifies disease progression as being either Low or High.

- 1. Select **Help > Sample Data Folder** and open Diabetes.jmp.
- 2. Select **Analyze > Predictive Modeling > Naive Bayes**.
- 3. Select Y Binary and click **Y, Response**.
- 4. Select Age through Glucose and click **X, Factor**.
- 5. Select Validation and click **Validation**.
- 6. Click **OK**.

Figure 8.2 Naive Bayes Report

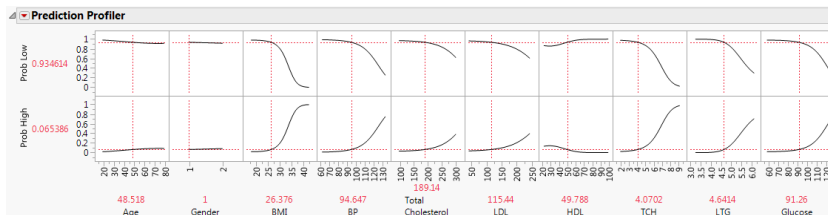


The Training Set has about a 21% misclassification rate and the Validation Set has about a 24% misclassification rate. The Confusion matrix suggests that, for both the Training and Validation sets, the larger source of misclassification comes from classifying patients with Low disease progression as having High disease progression. The Validation set results indicate how your model extends to independent observations.

You are interested in which individual predictors have the greatest impact on the naive Bayes classification.

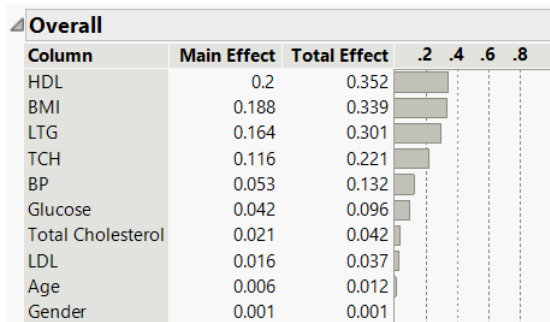
- 7. Click the Naive Bayes red triangle and select **Profiler**.

Figure 8.3 Prediction Profiler for Disease Progression



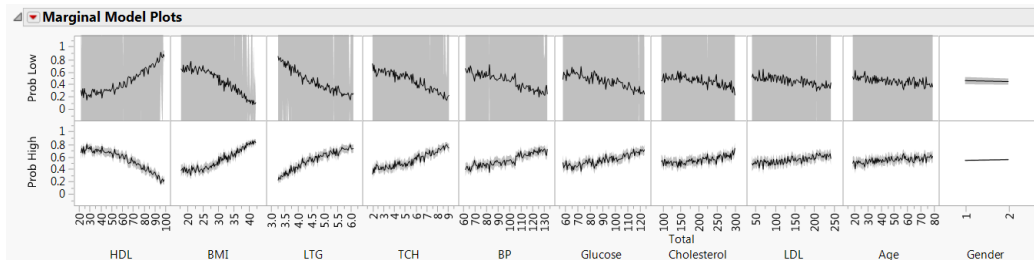
8. Click the Prediction Profiler red triangle and select **Assess Variable Importance > Independent Uniform Inputs**.

Figure 8.4 Variable Importance



The Summary Report indicates that HDL, BMI, and LTG have the greatest impact on the estimated probabilities.

Figure 8.5 Marginal Model Plots Report

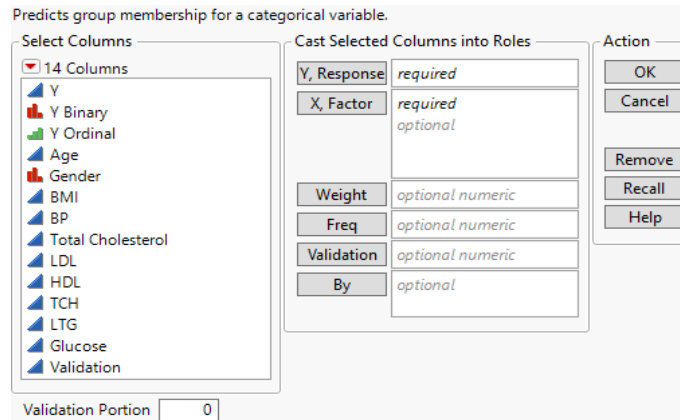


The second row of plots in the Marginal Model Plots report shows that higher values of HDL are associated with a lower probability of classifying a patient as High. Also, higher BMI and LTG values are associated with a higher probability of classifying a patient as High.

Launch the Naive Bayes Platform

Launch the Naive Bayes platform by selecting **Analyze > Predictive Modeling > Naive Bayes**.

Figure 8.6 Naive Bayes Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Naive Bayes launch window provides the following options:

Y, Response The categorical response column whose values are the classes of interest.

X, Factor Categorical or continuous predictor columns.

Weight A column whose numeric values assign a weight to each row in the analysis.

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

Validation A numeric column that defines the validation sets. This column should contain at most three distinct values:

- If the column has three unique values, then:
 - rows with the smallest value are used for the Training set.
 - rows with the middle value are used for the Validation set.
 - rows with the largest value are used for the Test set.
- If the column has two unique values, then only Training and Validation sets are used.
- If the validation column has more than three levels, the rows that contain the smallest three values define the validation sets. All other rows are excluded from the analysis.

JMP PRO Fit Details Report

The Fit Details report in the Naive Bayes platform shows various measures of fit for the model for the Training set, and for the Validation and Test sets if they are specified. The Measure column lists the different fit statistics and the Definition column shows the formulas for the corresponding fit statistics. See [“Measures of Fit for Categorical Responses”](#). By default, the Fit Details report in the Naive Bayes report window is closed.

JMP PRO Response Column Report

In the Naive Bayes platform, the report for the response column is named after the column name of the response. The report shows performance statistics for the naive Bayes classification in a summary table for the Training set, and the Validation and Test sets if they are specified. The summary tables contain the following columns:

Count The number of observations in the set corresponding to the table (Training, Validation, or Test set).

Misclassification Rate Proportion of observations in the corresponding set that are misclassified by the model. This is calculated as Misclassifications divided by Count.

Misclassifications The number of observations in the corresponding set that are classified incorrectly.

JMP PRO Confusion Matrix Report

The Confusion Matrix report in the Naive Bayes platform shows a confusion matrix and a confusion rates matrix for the Training set, and for the Validation and Test sets if they are specified. A confusion matrix is a two-way classification of actual and predicted responses. A confusion rates matrix is equal to the confusion matrix, with the numbers divided by the row totals.

JMP PRO ROC Curves

The report displays the Receiver Operating Characteristic (ROC) curve for the Training set, and for the Validation and Test sets if they are specified. The ROC curve measures the ability of the fitted probabilities to classify response levels correctly. The further the curve from the diagonal, the better the fit. An introduction to ROC curves is found in *Basic Analysis*.

See [“ROC Curve”](#) for more information about ROC curves.

Naive Bayes Platform Options

The Naive Bayes red triangle menu contains the following options:

ROC Curve Shows or hides the ROC plot. The ROC plot is shown by default. See [“ROC Curves”](#).

Lift Curve Shows or hides the lift curve for the model. If you used validation, Lift curve is shown for each of the Training, Validation, and Test sets, if specified.

A lift curve shows how effectively response levels are classified as their fitted probabilities decrease. The fitted probabilities are plotted along the horizontal axis in descending order. The vertical coordinate for a fitted probability is the proportion of correct classifications for that probability or higher, divided by the overall correct classification rate. Use the lift curve to see whether you can correctly classify a large proportion of observations if you select only those with a fitted probability that exceeds a threshold value.

If the response has two levels, the Lift curve plot displays a lift curve for the first level of the response only. If the response has more than two levels, the Lift curve plot displays a sub-outline of the curves for each response level. See [“Lift Curve”](#) for more information about lift curves.

Save Predicteds Saves the predicted classifications to the data table in a column called Naive Predicted <Y, Response>.

Save Prediction Formula Saves a column called Naive Predicted Formula <Y, Response> to the data table. This column contains the prediction formula for the classifications.

Save Probability Formula Saves columns to the data table that contain formulas used for classifying each observation. Three groups of columns are saved:

Naive Score <Class>, Sum For each column that represents a class, this column gives a score formula that measures strength of membership in the given class. In the Naive Score Sum column., these scores are summed across classes. See [“Statistical Details for Saved Probability Formulas”](#).

Naive Prob <Class> For each class, this column gives a formula for the conditional probability that an observation is in that class. See [“Statistical Details for Saved Probability Formulas”](#).

Naive Predicted Formula <Y, Response> Gives the formula for the predicted class.

Publish Probability Formula Creates probability formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option opens the Formula Depot window. See [“Formula Depot”](#).

Profiler Shows or hides an interactive profiler report. Changes in the factor values are reflected in the estimated classification probabilities. See *Profilers*.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

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



Additional Example of Naive Bayes

Use Naive Bayes to construct a classification model using a data set that has missing data on most of the predictors. You have historical financial data for 5,960 customers who applied for home equity loans. Each customer was classified as being a Good Risk or Bad Risk.

1. Select **Help > Sample Data Folder** and open Equity.jmp.
2. Select **Analyze > Predictive Modeling > Naive Bayes**.
3. Select BAD and click **Y, Response**.

One of the potential predictors, DEBTINC, has many missing values that might be informative. However, naive Bayes is not prepared to handle large number of missing values well, so you do not include DEBTINC in your model.

4. Select LOAN through CLNO and click **X, Factor**.
5. Select Validation and click **Validation**.
6. Click **OK**.

Denote the possible classifications by $C_1, \dots, C_K$. Denote the features, or predictors, by $X_1, X_2, \dots, X_p$.

The conditional probability that an observation with predictor values $x_1, x_2, \dots, x_p$ belongs in the class C_k is computed as follows:

$$P(C_k | (x_1, \dots, x_p)) = \frac{\left(P(C_k) \prod_{j=1}^p [P(x_j | C_k)] \right) / (\mathbf{R})}{\sum_{k=1}^K \left(P(C_k) \left(\prod_{j=1}^p [P(x_j | C_k)] \right) / \mathbf{R} \right)},$$

where $\mathbf{R}$ is a regularization constant. In the formula above, the conditional probability that an observation with $X_j = x_j$ belongs to the class C_k , $P(x_j | C_k)$, is determined as follows:

- If X_j is categorical:

$$P(x_j | C_k) = \frac{\# \text{ of observations in } x_j | C_k + (0.5/K)}{\# \text{ of observations in } C_k + 0.5}$$

- If X_j is continuous:

$$P(x_j | C_k) = \frac{1}{s_{jk}} \phi((x_j - m_{jk})/s_{jk})$$

Here, ϕ is the standard normal density function, and m and s are the mean and standard deviation, respectively, of the predictor values within the class C_k .

The unconditional probability that an observation belongs in class C_k , $P(C_k)$, is computed as follows:

$$P(C_k) = \frac{\# \text{ of observations in } C_k + (0.5/K)}{\text{Total \# of observations} + 0.5}$$

An observation is classified into the class for which its conditional probability is the largest.

Note: In the formula for $P(C_k)$, 0.5 is the prior bias factor. This value is the default value. To change the prior default factor, go to File > Preferences > Platforms > Naive Bayes, select the Prior Bias check box and change the value.

Statistical Details for Saved Probability Formulas

This section describes the formulas saved using the Save Probability Formula option in the Naive Bayes red triangle menu. The conditional probability that an observation with predictor values $x_1, x_2, \dots, x_p$ belongs in the class C_k differs slightly from that given by $P(C_k | (x_1, \dots, x_p))$, shown in the section [“Statistical Details for the Naive Bayes Algorithm”](#). This is done for computational efficiency.

Naive Score Formulas

The Naive Score formula for a given class C_k , $S(C_k)$, is a variation of the numerator in the expression for $P(C_k | (x_1, \dots, x_p))$ and is computed as follows:

$$S(C_k) = \exp[\ln\{P(C_k)\} + \text{Continuous} + \text{Categorical} + \ln(\mathbf{R})]$$

The Naive Score formula is a combination of scores from continuous and categorical predictors. Recall that $\mathbf{R}$ is a regularization constant. The continuous portion of the formula is computed as follows:

$$\text{Continuous} = \sum_{j=1}^{p_1} \left\{ \text{Normal Log Density} \left(\frac{x_j - m_{jk}}{s_{jk}} \right) + \ln(1/s_{jk}) \right\}$$

where

$j = 1, \dots, p_1$ continuous predictors.

The categorical portion of the formula is computed as follows:

$$\text{Categorical} = \sum_{r=1}^{p_2} \left\{ \sum_{l=1}^{L_r} 1_{rl} \ln(P(x_{rl} | C_k)) \right\}$$

where

$r = 1, \dots, p_2$ categorical variables

$l = 1, \dots, L_r$ levels of the r^{th} categorical variable

1_{rl} is an indicator variable that equals 1 when x_{rl} is the l^{th} level of the r^{th} categorical predictor and 0 otherwise

Naive Score Sum Formulas

The Naive Score Sum formula, S , sums the Naive Score formulas over all classes. This is a variation of the denominator in the expression for $P(C_k | (x_1, \dots, x_p))$.

$$S = \sum_{k=1}^K S(C_k)$$

Naive Prob Formulas

The Naive Prob formula for a given class C_k equals $P(C_k | (x_1, \dots, x_p))$. In the JMP formulas,

$$P(C_k | (x_1, \dots, x_p)) = \frac{S(C_k)}{S}$$

Naive Predicted Formula

The Naive Predicted Formula for an observation classifies that observation into the class for which $P(C_k | (x_1, \dots, x_p))$ is the largest. This is equivalent to classifying an observation into the class for which its Naive Score formula is the largest.

Chapter 9

JMP PRO Support Vector Machines

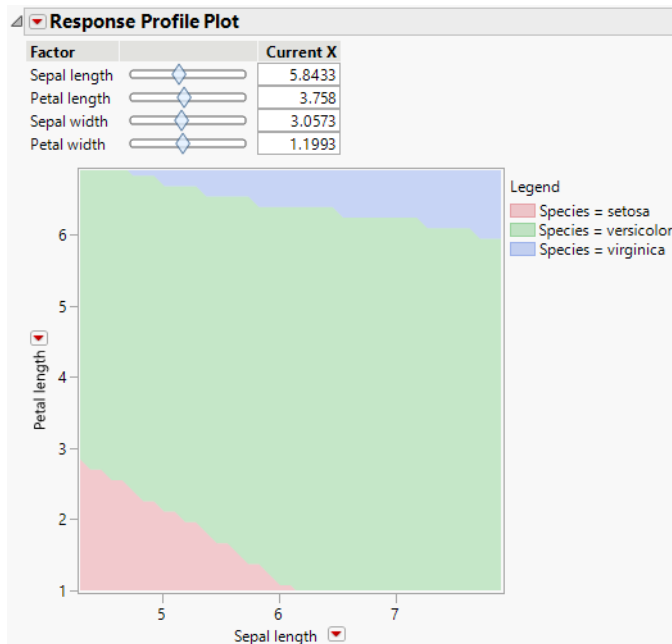
Predict Response Categories Using Support Vectors

The Support Vector Machines platform is available only in JMP Pro.

The Support Vector Machines (SVM) platform provides a machine learning tool for both nonlinear regression and classification. The SVM algorithm is a supervised learning algorithm that uses the training data to build a model to predict or classify new observations.

If you have a categorical response, you can visualize the model classification in the Response Profile Plot. The SVM platform also provides misclassification rates, confusion matrices, and ROC and Lift curves to help you determine how well your model fits the data. If you have a continuous response, fit statistics and actual by predicted plots are used to determine how well the model fits the data.

Figure 9.1 Response Profile Plot



Contents

Overview of Support Vector Machines 153

Example of Support Vector Machines 153

Launch the Support Vector Machines Platform 156

 Model Launch Control Panel..... 157

The Support Vector Machine Report 160

 Model Comparison Report 160

 Support Vector Machine Model Report..... 161

Support Vector Machines Platform Options 163

 Support Vector Machine Model Report Options 164

Additional Example of Support Vector Machines 166



Overview of Support Vector Machines

A support vector machine (SVM) model is a supervised learning algorithm that is used to predict or classify new observations. A model is fit on a set of training data where the responses are known. Then, the model is used to predict the responses of new observations.

When the response is categorical, SVM models classify data by optimizing a hyperplane that separates the classes. This can also be viewed as finding the hyperplane that maximizes the margin between the classes. In simple problems, this hyperplane is linear. However, more complicated data often cannot be separated linearly. For these scenarios, the SVM platform provides the option to use a radial basis function kernel to map the points to a nonlinear dimension that can make the classes easier to separate.

When the response is continuous, the models that are fit are known as support vector regression (SVR) models. In a typical regression problem, the goal is to fit a model that minimizes the error between a predicted response and the actual response. In an SVR problem, the goal is to fit a model such that the error between a predicted response and the actual response falls within a range of $-\epsilon$ to ϵ . This provides a more flexible fit. In JMP Pro, ϵ is equal to 0.1. The SVR algorithm doubles the data by creating two classes, $Y + \epsilon$ and $Y - \epsilon$. Then the same algorithm that is used for the classification problem is also used for the prediction (SVR) problem.

The maximization in SVM algorithms is performed by solving a quadratic programming problem. In JMP Pro, the algorithm used by the SVM platform is based on the Sequential Minimal Optimization (SMO) algorithm introduced by John Platt in 1998. Typically, the SVM quadratic programming problem is very large. The SMO algorithm splits the overall quadratic programming problem into a series of smaller quadratic programming problems. Smaller quadratic programming problems are solved analytically instead of numerically, meaning they produce closed-form solutions. Therefore, the SMO algorithm is more efficient than solving the overall quadratic programming problem (Platt 1998).



Example of Support Vector Machines

In this example, you use the Tuning Design option to fit several Support Vector Machine models with different kernel functions and parameter values. You want to find the best classification model to predict disease progression in diabetic patients. You have baseline medical data and a binary measure of diabetes disease progression obtained one year after each patient's initial visit. This measure quantifies disease progression as either Low or High

1. Select **Help > Sample Data Folder** and open Diabetes.jmp.
2. Select **Analyze > Predictive Modeling > Support Vector Machines**.

3. Select Y Binary and click **Y, Response**.
4. Select Age through Glucose and click **X, Factor**.
5. Select Validation and click **Validation**.
6. Click **OK**.
7. In the Model Launch control panel, check that the kernel function is a Radial Basis Function and select Tuning Design.
8. Enter 10 next to Number of Runs.
9. Click **Go**.
10. Click the gray triangle next to Model Launch to open the Model Launch control panel.
11. Change the kernel function to a Linear function and select Tuning Design.
12. Enter 10 next to Number of Runs.
13. Click **Go**.

Figure 9.2 Model Comparison Report

Model Comparison												
Show	Method	Kernel Function	Cost	Gamma	# SV	Training Misclassification Rate	Validation Misclassification Rate	Validation Generalized RSquare	Probability Threshold	Threshold Validation Misclassification Rate	Best	
☐	Model 1	Radial Basis Function	2.57061	0.43877	251	0.00324	0.25564	0.00838	0.39079	0.25564		
☐	Model 2	Radial Basis Function	3.04271	0.31899	218	0.01294	0.28571	-0.4206	0.23223	0.28571		
☐	Model 3	Radial Basis Function	3.93906	0.46899	255	0.00324	0.28571	-0.0081	0.53372	0.28571		
☐	Model 4	Radial Basis Function	1.073	0.27968	208	0.0712	0.21805	0.22158	0.17372	0.21805		
☐	Model 5	Radial Basis Function	0.41677	0.0281	156	0.15534	0.18797	0.50483	0.36015	0.18797		
☐	Model 6	Radial Basis Function	0.23369	0.39898	246	0.26861	0.28571	0.22323	6e-8	0.28571		
☐	Model 7	Radial Basis Function	4.8013	0.15144	167	0.05178	0.2782	-0.4245	0.26219	0.2782		
☐	Model 8	Radial Basis Function	3.51202	0.20295	181	0.04207	0.28571	-0.454	0.22753	0.28571		
☐	Model 9	Radial Basis Function	4.2693	0.06093	140	0.11974	0.19549	0.42841	0.32812	0.19549		
☐	Model 10	Radial Basis Function	1.87641	0.0919	150	0.13269	0.21053	0.43499	0.30598	0.21053		
☐	Model 11	Linear	4.42735	.	122	0.17476	0.17293	0.51943	0.47722	0.17293		
☐	Model 12	Linear	4.99046	.	123	0.16828	0.17293	0.52037	0.47551	0.17293		
☐	Model 13	Linear	3.8272	.	123	0.17476	0.17293	0.51923	0.47763	0.17293		
☐	Model 14	Linear	2.74788	.	124	0.17152	0.17293	0.51724	0.48019	0.17293		
☐	Model 15	Linear	3.22207	.	123	0.17476	0.17293	0.51851	0.47842	0.17293		
☐	Model 16	Linear	1.29304	.	124	0.17152	0.17293	0.51518	0.4819	0.17293		
☐	Model 17	Linear	1.74671	.	124	0.16828	0.17293	0.51557	0.48143	0.17293		
☐	Model 18	Linear	0.83575	.	123	0.16505	0.15789	0.51498	0.48207	0.15789		
☐	Model 19	Linear	2.22	.	125	0.17152	0.17293	0.51579	0.48255	0.17293		
☒	Model 20	Linear	0.04975	.	138	0.15534	0.14286	0.52302	0.41946	0.14286	Smallest ...	

The Model Comparison report shows that the best model in terms of misclassification rate and RSquare is Model 20. This model has a linear kernel function with cost parameter 0.04975. This is the model to further analyze.

Figure 9.3 Model Report for Best Fitting Model

Support Vector Machine Model 20

Model Summary

Estimation Details

Response	Y Binary		Cost	0.05
Validation Method	Validation Column			
Kernel Function	Linear			
Measure	Training	Validation		
Number of rows	309	133		
Sum of Frequencies	309	133		
Misclassification Rate	0.1553398	0.1428571		
Number of Support Vectors	138	138		

Fit Details

Measure	Training	Validation	Definition
Entropy RSquare	0.3855	0.3795	1-Loglike(model)/Loglike(0)
Generalized RSquare	0.5257	0.5230	$(1-(L(0)/L(model))^{(2/n)})/(1-L(0)^{(2/n)})$
Mean -Log p	0.3576	0.3712	$\sum -\text{Log}(p[j])/n$
RASE	0.3405	0.3404	$\sqrt{\sum (y[j]-p[j])^2/n}$
Mean Abs Dev	0.2311	0.2453	$\sum y[j]-p[j] /n$
Misclassification Rate	0.1715	0.1579	$\sum (p[j]\neq p[\text{Max}])/n$
N	309	133	n

Confusion Matrix

Set Probability Threshold

Training

Validation

			Misclassification Rate				Misclassification Rate
Actual	Predicted Rate		0.1553	Actual	Predicted Rate		0.1429
Y Binary	Low	High		Y Binary	Low	High	
Low	0.960	0.040		Low	0.926	0.074	
High	0.470	0.530		High	0.316	0.684	
Actual	Predicted Count			Actual	Predicted Count		
Y Binary	Low	High		Y Binary	Low	High	
Low	217	9		Low	88	7	
High	39	44		High	12	26	

The Model Summary report shows that the misclassification rates for the training and validation sets are very similar. This is a good indication that the model did not overfit the data. The confusion matrices provide more information about the types of observations that are misclassified by the model. In the confusion matrices, the upper left corner shows that the model correctly categorizes the Low responses most of the time (96% in training and 92.6% in validation). However, fewer of the High responses are correctly categorized (53% in training and 68.4% in validation). Therefore, most of the misclassifications are the High responses being misclassified as Low.

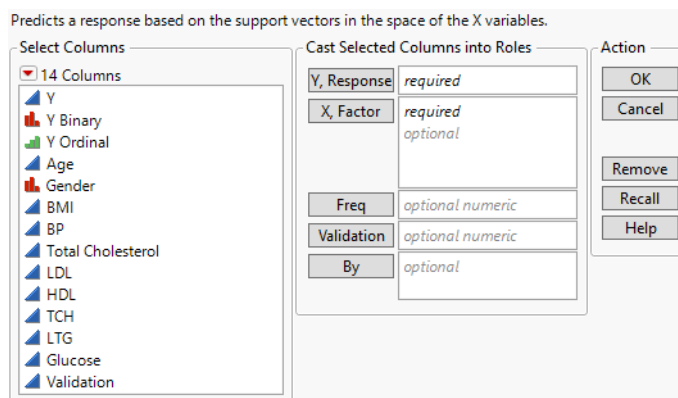
Note: Results vary due to the random nature of the design points in the tuning design.



Launch the Support Vector Machines Platform

Launch the Support Vector Machines (SVM) platform by selecting **Analyze > Predictive Modeling > Support Vector Machines**.

Figure 9.4 Support Vector Machines Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Support Vector Machines launch window provides the following options:

Y, Response The response variable that you want to analyze.

X, Factor The predictor variables.

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

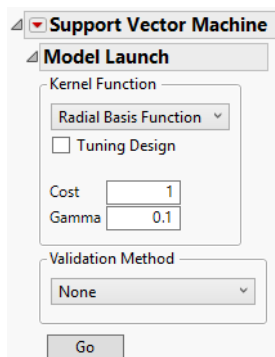
Validation A numeric column that defines the validation sets. See [“Validation Method”](#). If you click the Validation button with no columns selected in the Select Columns list, you can add a validation column to your data table. For more information about the Make Validation Column utility, see [“Make Validation Column”](#).

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.

JMP PRO Model Launch Control Panel

After you launch the Support Vector Machines platform, a Model Launch control panel for fitting models appears. Use the Model Launch control panel to specify the kernel function and associated parameter values, as well as the validation method.

Figure 9.5 The Model Launch Control Panel



The Model Launch control panel contains the following options:

Kernel Function Specifies the kernel function used in the model. Choose from the following kernel functions:

Radial Basis Function Selects the radial basis function kernel to create a nonlinear hyperplane to separate the classes.

- The Cost parameter is the penalty associated with misclassifying an observation in the training set. A higher cost parameter implements an algorithm that is less likely to misclassify a point in the training set, whereas a lower cost parameter produces a wider margin. The Cost parameter must be greater than 0, and the default value is 1.
- The Gamma parameter is the parameter in the kernel function. This parameter determines the amount of curvature there is to the decision line; a higher Gamma value indicates more curvature. A nonlinear decision line provides a more flexible fit, but too much curvature can lead to overfitting. The Gamma parameter must be greater than 0, and the default value is $1/(\text{\# of predictors})$.

Linear Selects the linear kernel function to create a linear hyperplane to separate the classes.

- The Cost parameter is the penalty associated with misclassifying an observation in the training set. A higher cost parameter implements an algorithm that is less likely to misclassify a point in the training set, whereas a lower cost parameter produces a wider margin. The Cost parameter must be greater than 0, and the default value is 1.

Note: If you specify parameter values that are out of range, the default values are used.

Tip: To find the best fitting model, fit a range of kernel functions and parameter values and use the Model Comparison report.

Tuning Design Enables you to fit a range of parameter values for the specified kernel. The models with the largest RSquare and the smallest Misclassification Rate or RASE are identified in the Model Comparison report. After you select Tuning Design, you must specify minimum and maximum values for the parameters. Default values are provided based on the data and the minimum must be greater than zero. You must also specify a value for the Number of Runs. The SVM platform fits that many models over a grid of parameter values determined by the minimum and maximum values.

Validation Method Specifies the model validation method. When you click the Go button for the first time, the first SVM model is fit using the specified validation method. This Validation Method is then used for all SVM models fit from within the SVM window. This ensures that all models in the report window are fit using the same validation method and validation set.

Holdback Randomly divides the original data into training and validation sets. You can specify the proportion of the original data to use as the validation set (holdback).

KFold (Available only when Y is continuous or nominal.) Randomly divides the original data into K subsets. In turn, each of the K sets is used to validate the model fit on the rest of the data, fitting a total of K models. If Y is continuous, the model that has the best validation RASE statistic is chosen as the final model. If Y is nominal, the model that has the best validation misclassification rate is chosen as the final model.

Validation Column (Available only if you specified a Validation column in the launch window.) Uses the values in the specified Validation column to divide the data into parts. The column's values determine how the data are split, and what method is used for validation:

- If there are two values, the smaller value defines the training set and the larger value defines the validation set.
- If there are three values, these values define the training, validation, and test sets in order of increasing size.
- If the validation column has more than three levels, then Validation Column K Fold is used.

The SVM platforms uses the validation column to train and evaluate the model, unless a Tuning Design is used. If the Tuning Design option is selected, the SVM platform uses the validation column to train and tune the model or to train, tune, and evaluate the model. For more information about validation, see [“Validation in JMP Modeling”](#).

Note: If the validation column does not lead to a valid partition of the data, the Holdback validation method is used instead.

Validation Column K Fold (Available only when the Y, Response column has exactly two levels and a Validation column is specified in the launch window.) Uses the values in the specified Validation column to divide the data into K sets, where K is the number of unique values in the column. Then, K-Fold validation is performed.

None No validation used.

Go Fits the specified SVM model and shows the model report.

Note: If you have a large data table, a progress bar is shown for each model that is fit to the data. The total number of models fit is $k!/2(k-2)!$, where k is the number of levels of the response variable. Each progress bar has an **Accept Current Estimates** button. Click this button if you want to stop the fitting algorithm early and accept the current estimates. Because prediction calculations are performed after you click this button, it may take some time for the report to appear.

Missing Values

Rows that contain any missing predictor values are not included in the SVM modeling procedure. Therefore, any columns saved to the data table will contain missing values in those rows. If you want to include data with missing values in an SVM model, some form of preprocessing is required. See [“Explore Missing Values”](#).



The Support Vector Machine Report

When you click **Go** in the Model Launch control panel, a Model Comparison report appears above the control panel and a model report appears below the control panel. The model reports are numbered sequentially. Each time a new model is fit, it is added to the Model Comparison report and a new model report is added to the window. See [“Model Comparison Report”](#) and [“Support Vector Machine Model Report”](#). Click on the column headings to sort the models by any of the columns in the Model Comparison report. The first click sorts in ascending order; click the column heading a second time to sort in descending order.



Model Comparison Report

In the Support Vector Machines platform, the Model Comparison report contains the following information:

Show A check box that indicates whether the Support Vector Machine Model report for the corresponding fit should appear in the report window.

Method The model number.

Kernel Function The kernel function for the corresponding model.

Cost The value of the cost parameter for the corresponding model.

Gamma The value of the gamma parameter for the corresponding model. If the model uses a linear kernel, this value is missing.

SV The number of support vectors used in the corresponding model.

Training Misclassification Rate (Appears only if the response is categorical.) The misclassification rate for the observations in the training set. This rate is based on the classification decision rule that is calculated by the platform.

Validation Misclassification Rate (Appears only if the response is categorical and validation is used.) The misclassification rate for the observations in the validation set. This rate is based on the classification decision rule that is calculated by the platform.

Test Misclassification Rate (Appears only if the response is categorical and a test set is specified using a validation column.) The misclassification rate for the observations in the test set. This rate is based on the classification decision rule that is calculated by the platform.

Training RASE (Appears only if the response is continuous.) The square root of the mean squared prediction error (Root Average Square Error) for the observations in the training set. See [“RASE”](#).

Validation RASE (Appears only if the response is continuous and validation is used.) The square root of the mean squared prediction error (Root Average Square Error) for the observations in the validation set. See [“RASE”](#).

Test RASE (Appears only if the response is continuous and a test set is specified using a validation column.) The square root of the mean squared prediction error (Root Average Square Error) for the observations in the test set. See [“RASE”](#).

Validation Generalized RSquare (Appears only if validation is used.) The Generalized RSquare value for observations in the validation set. See [“Fit Details”](#).

Probability Threshold (Appears only if the response is binary and validation is used.) The cutoff probability determined by the classification decision rule that is calculated by the platform. An observation is classified into the target level when its predicted probability exceeds this value.

Tip: You can change this value by using the Set Probability Threshold option in the Confusion Matrix report. See [“Confusion Matrix”](#).

Conditional Validation Misclassification Rate (Appears only if the response is binary and validation is used.) The misclassification rate for the observations in the validation set, conditioned on the probability threshold value.

Best Indicates which model fit has the smallest misclassification rate. If validation is used without a test set, this is the model with the smallest validation misclassification rate. If validation is used with a test set, this is the model with the smallest test misclassification rate.

Model Comparison Plot

Shows a contour plot of the model performance of the validation set over the grid of parameter values. This report is available only if Tuning Design is specified in the Model Launch Control Panel.

Support Vector Machine Model Report

Each model that is fit produces a model report that contains the following sections:

Response Profile Plot (Available only when the response is categorical.) Gives a visual representation of the classification model. The points on the plot are the actual data observations and are on by default only when there are exactly two continuous variables. For the two variables plotted, the shaded contours represent a plane of the prediction space determined by fixed values of the remaining model factors. The predictions are based on the classification decision rule that is calculated by the platform. Controls for the

fixed values are located above the plot. When you change the fixed values of the factors, using either the slider or number box, the prediction space for the plotted variables is automatically updated. You can also change the axes of the plot to display any continuous factor using the red triangles on each axis.

The Response Profile Plot red triangle menu contains the following options:

Grid Density Determines the fineness of the prediction grid underlying the shaded contours. A higher grid density provides a smoother decision line.

Show Points Shows or hides the points on the plot. On by default when there are only two variables.

Model Summary Gives the name of the response column, the validation method, and the type of kernel function used in the model fit. The Model Summary table also contains information about the model fit for the training, validation, and test sets. The number of observations and the number of support vectors are reported for each set. If the response is categorical, the misclassification rate is reported for each set. The misclassification rate is the proportion of observations misclassified by the model. This is calculated as the number of misclassifications divided by the total number of observations. If the response is continuous, the RASE and R Square values are reported for each set.

Estimation Details Gives the values of the parameters used in the model.

Fit Details (Available only if the response is categorical.) Gives the following statistics for the training set, and for the validation and test sets if they are specified:

Entropy RSquare A measure of fit that compares the log-likelihoods from the fitted model and the constant probability model. Entropy RSquare ranges from 0 to 1, where values closer to 1 indicate a better fit. See [“Entropy RSquare”](#).

Generalized RSquare A measure that can be applied to general regression models. It is based on the likelihood function L and is scaled to have a maximum value of 1. The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the Nagelkerke or Craig and Uhler R^2 , which is a normalized version of Cox and Snell's pseudo R^2 . See Nagelkerke (1991). Values closer to 1 indicate a better fit.

Mean -Log p The average of $-\log(p)$, where p is the fitted probability associated with the event that occurred. Smaller values indicate a better fit.

RASE The square root of the mean squared prediction error (Root Average Square Error). RASE is computed as follows, where *Source* indicates the Training, Validation, or Test set.

$$RASE_{Source} = \sqrt{\frac{SSE_{Source}}{n}}$$

Mean Abs Dev The average of the absolute values of the differences between the response and the predicted response. The differences are between 1 and p , the fitted probability for the response level that actually occurred. Smaller values indicate a better fit.

Misclassification Rate The rate for which the response category with the highest fitted probability is not the observed category.

Note: The misclassification rates in the Fit Details report might not match the misclassification rates in the Confusion Matrix report. When the response is binary, the rates in the Fit Details report use a probability cutoff of 0.5, but the rates in the Confusion Matrix report use the probability threshold value as the cutoff.

N The number of observations.

Confusion Matrix (Available only if the response is categorical.) A confusion matrix is shown for each set (training, validation, and test) specified in the model. A confusion matrix is a two-way classification of actual and predicted responses. Use the confusion matrices and the misclassification rates to evaluate your model.

The confusion matrices and the misclassification rates use the value in the probability threshold box as the cutoff value. By default, this value is based on the classification decision rule that is calculated by the platform. You can change the cutoff value by dragging the slider or entering a new value in the box next to Probability Threshold. If you change the probability threshold, the confusion matrices and misclassification rates automatically update. The probability threshold and conditional validation misclassification rate columns in the Model Comparison report are also updated.

Actual by Predicted Plot (Available only if the response is continuous.) Plots the actual response versus the predicted response. The diagonal line is the $Y = X$ line. The closer the points are to this line, the better the model fits the data. When validation is used, plots are shown for the training, validation, and, if specified, test sets.

Support Vector Machines Platform Options

The Support Vector Machine red triangle menu contains the following options:

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Support Vector Machine Model Report Options

Each Support Vector Machine Model report contains the following options:

Response Profile Plot (Available only if the response is categorical.) Shows or hides the Response Profile Plot.

Support Vector Coefficients Shows or hides the table of support vector coefficients. Select rows in the table to highlight the support vectors in the original data table and the Response Profile Plot.

Confusion Matrix (Available only if the response is categorical.) Shows or hides the confusion matrix. See [“Confusion Matrix”](#).

ROC Curve (Available only if the response is categorical.) Shows or hides the ROC Curve. If you use validation, ROC Curves for the validation and, if specified, test sets appear in addition to the curve for the training set. For more information, see [“ROC Curve”](#).

Lift Curve (Available only if the response is categorical.) Shows or hides the Lift Curve. If you use validation, Lift Curves for the validation and, if specified, test sets appear in addition to the curve for the training set. For more information, see [“Lift Curve”](#).

Plot Actual by Predicted (Available only if the response is continuous.) Shows or hides a plot of the actual values versus the predicted values.

Plot Residual by Predicted (Available only if the response is continuous.) Show or hides a plot of the residuals versus the predicted values.

- Profiler** Shows or hides the Prediction Profiler. Each response level is represented by a separate row in the Prediction Profiler. For more information about the options in the red triangle menu, see *Profilers*.
- Contour Profiler** (Available only when the model contains more than one continuous factor.) Shows or hides the Contour Profiler. For more information about the options in the red triangle menu, see *Profilers*.
- Surface Profiler** (Available only when the model contains at least one continuous factor.) Shows or hides a surface profiler. For more information about the options in the red triangle menu, see *Profilers*.
- Save Predicteds** Saves the predicted classifications based on the threshold probability cutoff value to the data table in a column called Predicted <Y, Response>.
- Save Prediction Formula** Saves the prediction formulas based on the threshold probability cutoff value to the data table. If the response is binary, a single prediction formula column is added to the data table. If the response has more than two levels, a prediction formula column is added to the data table for each pairwise model, in addition to a formula column for the final prediction.
- Publish Prediction Formula** Creates probability formulas based on the threshold probability cutoff value and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option opens the Formula Depot window. See [“Formula Depot”](#).
- Save Probabilities** Saves the probability of each response level as a separate column in the data table, as well as a column called Most Likely <Y, Response>.
- Save Probability Formula** (Available only if the response is binary.) Saves the following formula columns to the data table:
- Decision Value** A hidden column that contains the formula for the decision value.
 - Probability** Two probability formula columns, one for each level of the binary response.
 - Most Likely** The predicted classification based on which probability value is higher.
 - Threshold Predicted** The predicted classification based on the threshold probability cutoff value obtained from the confusion matrix. By default, this value is based on the classification decision rule that is calculated by the platform. If the threshold value is changed to 0.5, the Threshold Predicted column is not saved to the data table because it would be equivalent to the Most Likely column.
- Publish Probability Formula** (Available only if the response is binary.) Creates probability formulas and saves them as formula column scripts in the Formula Depot platform. If a Formula Depot report is not open, this option opens the Formula Depot window. See [“Formula Depot”](#).

Save Validation (Available only when the Holdback or KFold validation methods are used.)
Creates a new column in the data table that identifies which rows were used in the training and validation sets.

Remove Fit Removes the entire model report.

Additional Example of Support Vector Machines

In this example, you use Support Vector Regression (SVR) to model a continuous response. SVR models use the SVM algorithm to predict continuous response data. Use the same medical data on diabetic patients that was used in [“Example of Support Vector Machines”](#). However, instead of a binary measure of diabetes disease progression, you now have a continuous measure. Higher values correspond to more disease progression. Vary the cost parameter of the radial basis function to find the best fitting model.

1. Select **Help > Sample Data Folder** and open Diabetes.jmp.
2. Select **Analyze > Predictive Modeling > Support Vector Machines**.
3. Select Y and click **Y, Response**.
4. Select Age through Glucose and click **X, Factor**.
5. Select Validation and click **Validation**.
6. Click **OK**.
7. In the Model Launch control panel, check that the kernel function is a Radial Basis Function with Cost parameter 1 and Gamma parameter 0.1.
8. Click **Go**.
9. Click the gray triangle next to Model Launch to open the Model Launch control panel.
10. Change the cost parameter to 0.1.
11. Click **Go**.
12. Click the gray triangle next to Model Launch to open the Model Launch control panel.
13. Change the cost parameter to 2.
14. Click **Go**.

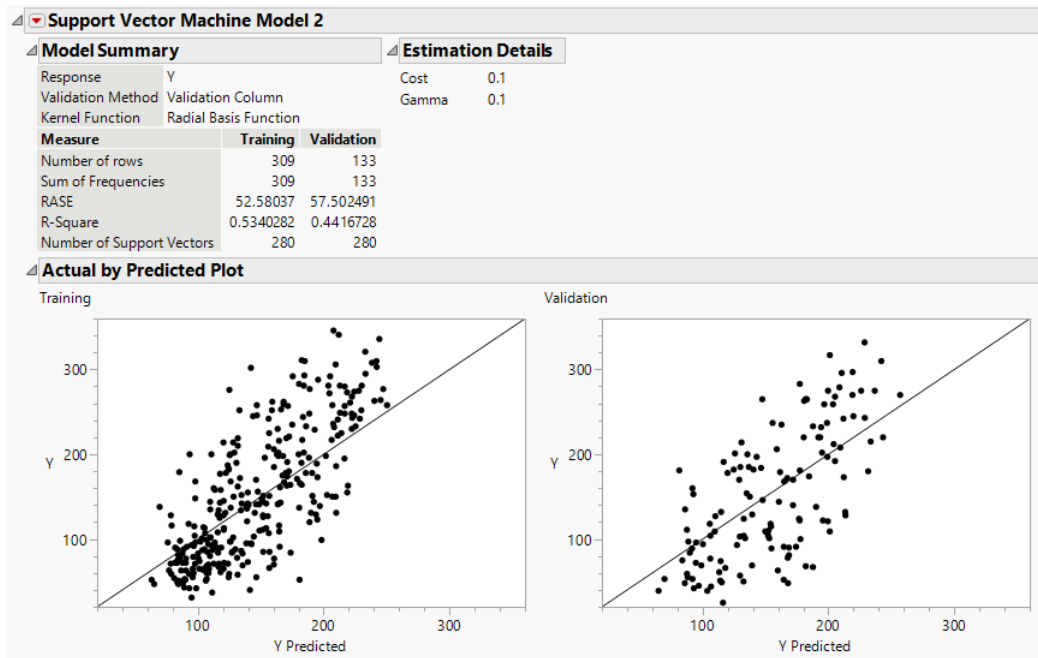
Figure 9.6 Model Comparison Report

Model Comparison									
Show	Method	Kernel Function	Cost	Gamma	# SV	Training RASE	Validation RASE	Validation Generalized RSquare	Best
☐	Model 1	Radial Basis Function	1	0.1	266	43.8147	58.3285	0.42552	
☒	Model 2	Radial Basis Function	0.1	0.1	280	52.5804	57.5025	0.44167	Smallest RASE Larg
☐	Model 3	Radial Basis Function	2	0.1	265	40.553	60.7801	0.37621	

The Model Comparison report contains the results of the models using different cost parameters. Recall that a higher cost parameter implements an algorithm that is less likely to misclassify a point, whereas a lower cost parameter implements a more flexible algorithm. In this scenario, the more flexible model (cost parameter equal to 0.1) produces the best model fit.

15. In the Show column of the Model Comparison report, deselect Model 1 and Model 3.

Figure 9.7 Model Report for Continuous Response



The model report for a continuous response contains the Model Summary, Estimation Details, and Actual by Predicted Plot. In the Model Summary report, the RASE is 52.58 for the training set and 57.50 for the validation set.

Chapter 10

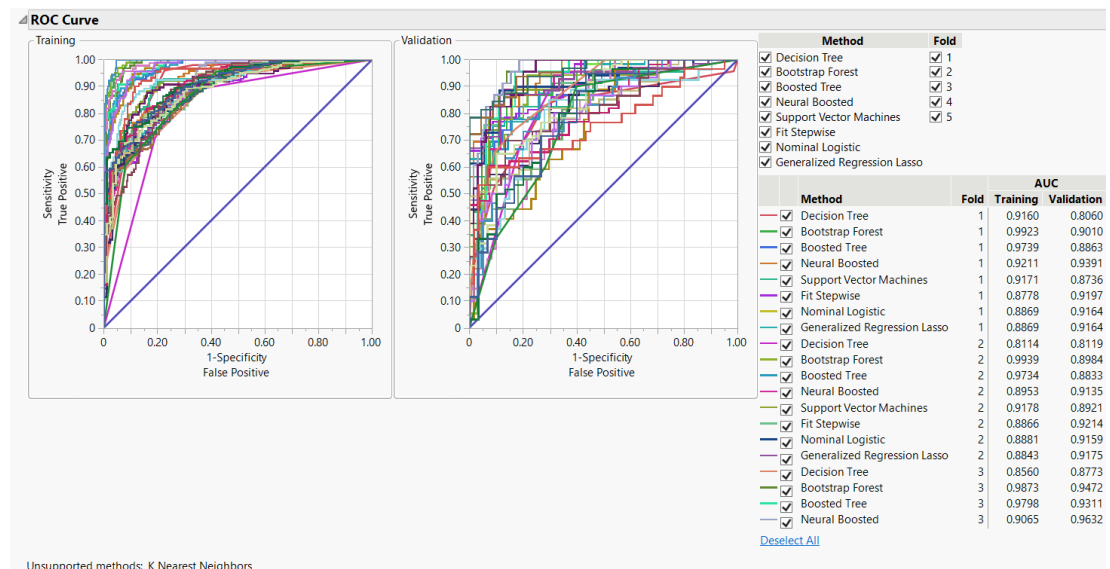
JMP[®] PRO Model Screening

Fit and Compare Multiple Prediction Models

The Model Screening platform is available only in JMP Pro.

The Model Screening platform in JMP Pro enables you to run multiple predictive modeling platforms from one launch window and assemble summaries from the different methods. This helps you pick the best performing model without launching each individual model platform. You can also easily launch an individual platform report for any of the models from the Model Screening report window for further refinement.

Figure 10.1 Model Screening Report



Contents

Example of Model Screening..... 171

Launch the Model Screening Platform..... 173

The Model Screening Report..... 176

 Summary Across the Folds Report..... 177

 Training, Validation, and Test Measures of Fit..... 178

Model Screening Platform Options 179

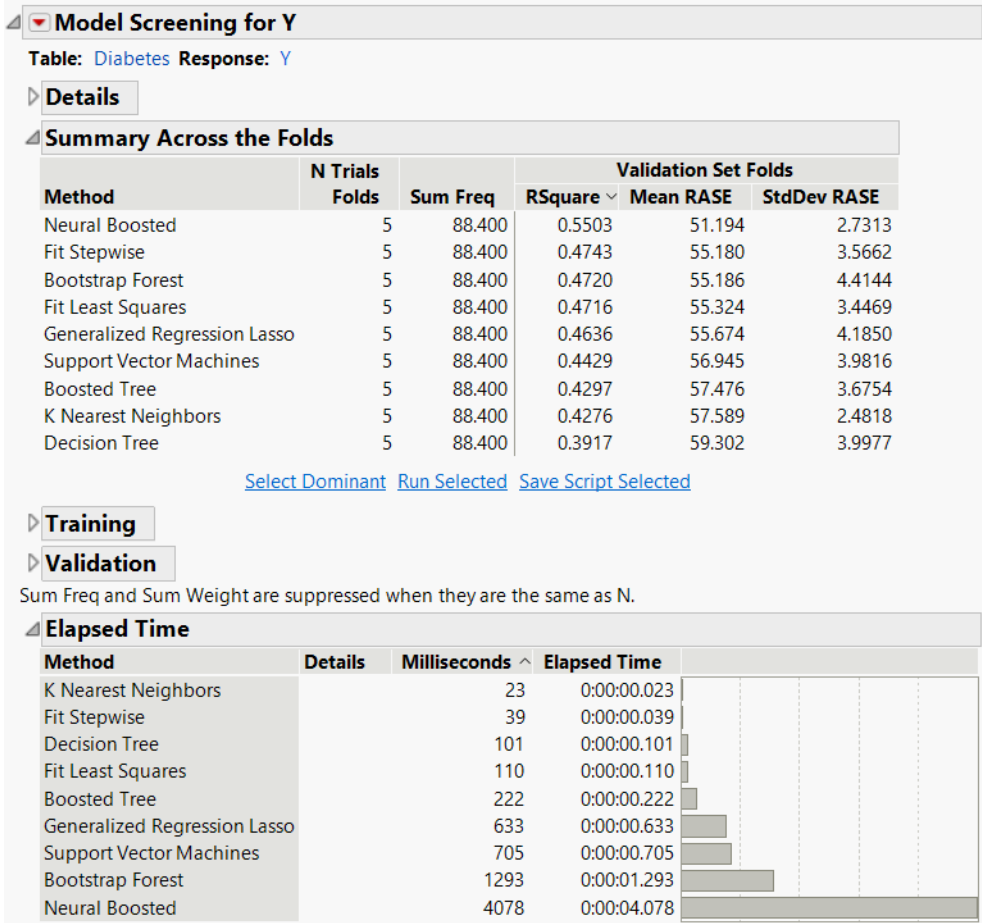
Decision Thresholds Report..... 182

Example of Model Screening

You want to assess various models for their ability to predict diabetes disease progression. The data set includes baseline medical data for 442 diabetic patients and a continuous measure of diabetes disease progression obtained one year after each patient's initial visit. Higher values correspond to more disease progression.

1. Select **Help > Sample Data Folder** and open Diabetes.jmp.
2. Select **Analyze > Predictive Modeling > Model Screening**.
3. Select Y and click **Y, Response**.
4. Select Age through Glucose and click **X, Factor**.
5. In the Folded Crossvalidation section, select the box next to **K Fold Crossvalidation**.
6. (Optional) In the Operational Options section, type 42920 next to **Set Random Seed**.
Use the random seed in order to match the example output.
7. Click **OK**.
8. Click the red triangle next to Model Screening for Y and select **Optional Reports > Elapsed Time**.

Figure 10.2 Model Report for Continuous Response



The best performing model based on the Validation RSquare averaged across the folds, is Neural Boosted. The average Validation RSquare for Neural Boosted is 0.5503. The Elapsed Time report shows that the Neural Boosted models also took the longest to fit. In this case, the elapsed time was only 5 seconds, but this information could be important for larger data sets or more complex models.

Tip: To see performance metrics for the individual models, click the gray disclosure icon next to Validation.

JMP^{PRO} Launch the Model Screening Platform

To launch the Model Screening platform, select **Analyze > Predictive Modeling > Model Screening**.

Figure 10.3 The Model Screening Launch Window

Fits many different predictive models and provides summaries of measures of fit.

Select Columns

15 Columns

- Y
- Y Binary
- Y Ordinal
- Age
- Gender
- BMI
- BP
- Total Cholesterol
- LDL
- HDL
- TCH
- LTG
- Glucose
- Validation
- Hidden Validation

Cast Selected Columns into Roles

Y, Response: Y (optional)

X, Factor: Age, Gender, BMI, BP

Weight: optional numeric

Freq: optional numeric

Validation: optional numeric

By: optional

Action

OK

Cancel

Remove

Recall

Help

Method

- ☒ Decision Tree
- ☒ Bootstrap Forest
- ☒ Boosted Tree
- ☒ K Nearest Neighbors
- ☐ Naive Bayes
- ☒ Neural
- ☒ Support Vector Machines
- ☒ Discriminant
- ☒ Fit Least Squares
- ☒ Fit Stepwise
- ☒ Logistic Regression
- ☒ Generalized Regression
- ☒ Additional Methods
- ☐ Partial Least Squares
- ☐ XGBoost

Modeling Options

- ☐ Add Two Way Interactions
- ☐ Add Quadratics
- ☐ Informative Missing

Operational Options

Set Random Seed: .

Time Limit Each: .

- ☐ Remove Live Reports
- ☐ Show method in Log when run

Folded Crossvalidation

Fit repeatedly with sequenced folds.

- ☒ K Fold Crossvalidation K: 5
- ☐ Nested Crossvalidation K: 4 L: 5

Repeated K Fold: 0

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Response The response variable or variables that you want to analyze.

X, Factor The predictor variables.

Weight (Not applicable to the K Nearest Neighbors, Support Vector Machines, or Neural modeling platforms.) A column whose numeric values assign a weight to each row in the analysis.

Freq (Not applicable to the K Nearest Neighbors modeling platform.) A column whose numeric values assign a frequency to each row in the analysis.

Validation (Not applicable if any of the Crossvalidation options are selected in the launch window.) A numeric column that defines the validation sets. 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 [“Make Validation Column”](#).

Note: If you specify a validation column with more than three levels, this column is used to perform K Fold crossvalidation.

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.

Methods Enables you to select the desired modeling platforms. By default, the modeling platforms that are fit are Decision Tree (Partition), Bootstrap Forest, Boosted Tree, K Nearest Neighbors, Neural, Support Vector Machines, Discriminant, Fit Least Squares, Fit Stepwise, Logistic Regression, and Generalized Regression. Naive Bayes, Partial Least Squares, and XGBoost are also available.

Notes:

- XGBoost is not supported by JMP and is available only if the XGBoost add-in is installed. For more information about XGBoost, see community.jmp.com.
- Decision Tree (Partition), Discriminant, and Partial Least Squares all require some type of validation set in order to fit a model.
- If there are fewer than 20 observations in a validation set, a Decision Tree (Partition) model cannot be fit.
- The modeling platforms use default options and tuning parameters in model fitting. You can try to improve the fit past what the default yields by calling platforms directly and choosing different options.
- The Additional Methods option under Generalized Regression calls several additional methods, such as Ridge, Elastic Net and Lasso, in the Generalized Regression platform. For the Lasso method, Early Stopping is disabled when there are less than 1000 observations and less than 100 variables. See *Fitting Linear Models*.

Caution: This results in additional model fits.

Modeling Options

Provides additional options for the modeling platforms.

Add Two Way Interactions Adds all two way interaction effects to linear models.

Add Quadratics Adds effects for the squares of continuous variables to linear models.

Informative Missing Enables informative missing for all platforms.

Operational Options

Provides additional options.

Set Random Seed Sets a random seed that is used for any random components of the model fit routines. This enables you to rerun the platform and obtain the same model fits.

Time Limit Each Specifies a time limit, in seconds, for each fit. For platforms that support early stopping, the best estimates up to that point are provided. For platforms that do not support early stopping, no result is provided.

Remove Live Reports Does not include the individual model platform reports in the Model Screening report window.

Tip: Select this option to free up memory when you have a large problem with many methods and fits.

Show method in Log when run Writes out a progress message to the log each time a fitting platform is called.

Folded Crossvalidation

Provides options for various types of crossvalidation.

K Fold Crossvalidation Divides the data randomly into K parts or folds. A model is fit to the data using $K-1$ folds to build the model and the remaining fold used for crossvalidation. This is repeated K times for a total of K models. The default value of K is 5.

- K specifies the number of folds for K Fold Crossvalidation. The default is 5 and K must be greater than 1.
- The results for the best model is provided.

Nested Crossvalidation Divides the data into nested folds for crossvalidation. First, the data are divided into $k = 1, \dots, K$ equal parts, or *folds*. For each fold, the k^{th} fold is used as a test set and the remaining data are divided further into L equal parts. These L subdivisions are called *inner folds*. Then, a model is fit to the data using $L-1$ inner folds with the remaining inner fold held out each time as a crossvalidation set. The L models then use the k^{th} fold as

a common test data set. In all, a total of $K \times L$ models are fit. The default value of K is 4 and the default value of L is 5.

For example, set $K = 2$ and $L = 3$. The data are initially divided into two folds. The first fold is held out as a test set and the second fold is divided into 3 inner folds. Three models are fit to the data, each time with a different inner fold held out as a crossvalidation set. Then, all three models are tested on the first fold.

The second fold is then held out as a test set and the first fold is divided into 3 inner folds. Three models are fit to the data, each time with a different inner fold held out as a crossvalidation set. Then, all three models are tested on the second fold.

- K specifies the number of folds for Nested Crossvalidation. The default is 4 and K must be greater than 1.
- L specifies the number of inner folds for Nested Crossvalidation. The default is 5 and L must be greater than 1.

Note: If both K Fold Crossvalidation and Nested Crossvalidation are selected, Nested Crossvalidation is performed.

Repeated K Fold Specifies the number of times the K Fold Crossvalidation or Nested Crossvalidation process is repeated.

When you click OK, the specified models are fit and a set of progress bars are shown. The upper progress bar reports the progress across all fits. The lower progress bar reports the progress for the current individual model fit. You can stop the lower progress bar to employ early stopping and the upper progress bar will continue to run.

The Model Screening Report

The content of the Model Screening report depends on both the type of crossvalidation implemented and the modeling type of the response variable. At the top of the Model Screening report, the name of the data table and column name of the response variable are shown. If a validation column is specified, the name of that column is also shown.

By default, the Details report appears next. This report contains the individual model reports for all of the models for each fold and trial. These reports are the actual output from each platform and contain all reports and command features. For example, if you decide on a model that is best, you can go to the corresponding platform report in the Details report and save the desired prediction formula column. You can hide the Details report by selecting the Remove Live Reports option in the launch window.

If either the K Fold Crossvalidation option or the Nested Crossvalidation option is specified in the launch window, the Summary Across the Folds report appears. This report is a summary of the measures of fit averaged across the test folds.

Measures of fit are shown for each individual model fit, for the training, validation, and test sets if applicable. If crossvalidation is used by specifying a validation column in the launch window, only the individual model fit measures of fit are shown in the report.

Summary Across the Folds Report

The Summary Across the Folds report in the Model Screening platform contains a summary of the measures of fit across the folds, as well as across the trials if Repeated K Fold is specified. If the K Fold Crossvalidation option is specified, the measures of fit are summarized across the validation sets. If the Nested Crossvalidation option is specified, the measures of fit are summarized across the test sets. The report contains a table with the following columns:

Method The name of the method used to fit the model.

N Trials Folds The total number of models that were fit across all folds and trials, if applicable.

Sum Freq The average number of observations in the validation or test sets. The test sets are used to estimate model performance.

RSquare The mean RSquare across all validation or test set folds. This column contains the Entropy RSquare when the response is categorical.

Mean RASE The mean RASE (Root Average Square Error) across all validation or test set folds.

StdDev RASE The standard deviation of the RASE across all validation or test sets.

Mean AUC (Available only for categorical responses.) The mean area under the ROC curve (AUC) across all validation or test sets.

Mean MR (Available only for categorical responses.) The mean misclassification rate (MR) across all validation or test sets.

The following options are available below the table:

Select Dominant Selects each model that is better than or equal to all of the other models in terms of a combination of model fitting criteria. This is also referred to as selecting the Pareto Frontier. For continuous responses, RSquare and Sum Freq are considered when determining the dominant model. For categorical responses, Entropy RSquare, Misclassification Rate, AUC, and Sum Freq are considered when determining the dominant model.

Run Selected

Runs the individual model for each selected row. If any type of folded crossvalidation is specified in the launch window, the model run uses the fold, inner fold, and trial combination that corresponds to the final model. The final model is the model that produces the highest weighted average RSquare. The weight average RSquare is the average, weighted by number of observations, of the training RSquare, validation RSquare, and when a test set is used, the test RSquare. A validation column is created in the data table for any fold, inner fold, and trial combination that is required by clicking Run Selected.

Save Script Selected

Saves a model script to the script window for each selected row. If any type of folded crossvalidation is specified in the launch window, the validation set specified in the script uses the fold, inner fold, and trial combination that corresponds to the final model. The final model is the model that produces the highest weighted average RSquare. The weight average RSquare is the average, weighted by number of observations, of the training RSquare, validation RSquare, and when a test set is used, the test RSquare. A validation column is created in the data table for any fold, inner fold, and trial combination that is requested by clicking Save Script Selected.

Training, Validation, and Test Measures of Fit

In the Model Screening platform, there is a measures of fit report for each model data set that is specified. These could be Training, Training and Validation, or Training, Validation, and Test sets. Each report contains a table with the following columns:

- Method** The name of the method used to fit the model.
- N** The number of observations in the set.
- Sum Wgt** The sum of the weights.
- RSquare** (Available only for continuous responses.) The RSquare value of the fitted model.
- Entropy RSquare** (Available only for categorical responses.) A measure of fit that compares the log-likelihoods from the fitted model and the constant probability model. Entropy RSquare ranges from 0 to 1, where values closer to 1 indicate a better fit. See [“Entropy RSquare”](#).
- Misclassification Rate** (Available only for categorical responses.) The proportion of observations misclassified by the model. Smaller values indicate a better fit.

Note: In these tables, the misclassification rate is always calculated using a probability threshold of 0.5.

AUC (Available only for categorical responses.) The area under the ROC curve. Values closer to 1 indicate a better fit.

RASE The square root of the mean squared prediction error (Root Average Square Error). RASE is computed as follows, where *Source* indicates the Training, Validation, or Test set.

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

Generalized RSquare (Available only for categorical responses.) A measure that can be applied to general regression models. It is based on the likelihood function *L* and is scaled to have a maximum value of 1. The value is 1 for a perfect model, and 0 for a model no better than a constant model. The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the Nagelkerke or Craig and Uhler R^2 , which is a normalized version of Cox and Snell's pseudo R^2 . See Nagelkerke (1991).

Fold (Available only if the K Fold Crossvalidation option or the Nested Crossvalidation option is specified in the launch window.) Identifies the fold that is held out for the model fit in that row.

Inner Fold (Available only if the Nested Crossvalidation option is specified in the launch window.) Identifies the inner fold that is held out for the model fit in that row.

Trial (Available only if the Repeated K Fold option is specified in the launch window.) Identifies the trial number for the model fit in that row.

The following options are available below each table:

Select Dominant Selects each model that is better than or equal to all of the other models in terms of a combination of model fitting criteria. For continuous responses, RSquare and Sum Freq are considered when determining the dominant model. For categorical responses, Entropy RSquare, Misclassification Rate, AUC, and Sum Freq are considered when determining the dominant model.

Run Selected Runs the individual models specified in each selected row.

Save Script Selected Saves a model script to the script window for each selected row.

Model Screening Platform Options

The Model Screening red triangle menu contains the following options:

Plot Actual by Predicted (Available only for continuous responses.) Shows or hides a report of plots that overlay the actual by predicted points from all of the model fits. There are separate plots for the Training, Validation, and Test sets. The report contains a Methods legend with check boxes that enable you to specify which fits appear on the plots based on

the method, fold, inner fold, and trial combination. Click the Deselect All button at the bottom of the legend to remove all model fits from the plots.

Notes:

- If you select a subset of models from the Training, Validation, or Test reports and then select Plot Actual by Predicted, only the plots for the selected models are shown.
- The actual by predicted plots are not supported by all of the modeling platforms. A list of unsupported platforms is shown below the plots.
- This option is not available if Remove Live Reports is selected in the launch window.

Profiler (Available only for continuous responses.) Shows or hides prediction profilers for each type of model fit. See *Profilers*. If any type of folded crossvalidation is specified in the launch window, the selected model in the prediction profiler uses the fold, inner fold, and trial combination that corresponds to the final model. The final model is the model that produces the highest weighted average RSquare. The weight average RSquare is the average, weighted by number of observations, of the training RSquare, validation RSquare, and when a test set is used, the test RSquare.

Notes:

- Confidence intervals are shown in prediction profilers for Standard Least Squares models and Generalized Regression models.
- Prediction profilers are not available for K Nearest Neighbors models.
- This option is not available if Remove Live Reports is selected in the launch window.

ROC Curve (Available only for categorical responses.) Shows or hides a report of plots that overlay the ROC curves for all of the model fits. There are separate plots for the Training, Validation, and Test sets. The report contains two legends with check boxes that enable you to specify which fits appear on the plots. Click the Deselect All button at the bottom of the legend to remove all model fits from the plots. For more information about ROC Curves, see [“ROC Curve”](#).

Notes:

- If you select a subset of models from the Training, Validation, or Test reports and then select ROC Curves, only the plots for the selected models are shown.
- The ROC plots are not supported by all of the modeling platforms. A list of unsupported platforms is shown below the plots.
- This option is not available if Remove Live Reports is selected in the launch window.

Show Profit (Available only for binary categorical responses that have a Profit Matrix column property.) Shows or hides the expected profit for each model.

Decision Threshold (Available only for binary categorical responses.) Shows or hides Decision Thresholds reports for the training, validation, and test sets, if specified. Each

report contains a graph of the distribution of fitted probabilities for each model, confusion matrices for each model, and classification graphs to adjust thresholds and compare the model fits. See [“Decision Thresholds Report”](#).

Notes:

- If you select a subset of models from the Training, Validation, or Test reports and then select Decision Threshold, only the information for the selected models is shown.
- If you specified any type of folded crossvalidation in the launch window, the probabilities shown in the Decision Threshold report are the averaged probabilities over all of the folds.

Save Results Table Saves the information in the Validation report to a new data table titled Model Screening Statistics Validation Set. If there is a test set, such as when the Nested Crossvalidation option is specified, the information in the Test report is also saved to a new data table titled Model Screening Statistics Test Set.

Save KFold Results Table (Available only if a Crossvalidation option is specified in the launch window.) Saves the information in the Summary Across the Folds report to a new data table titled KFold Results.

Save Prediction Formulas (Available only if the live reports are included in the model report.) Saves the prediction formulas for the selected models to the data table. If you select a model in the Summary Across the Folds report and any type of fold crossvalidation is specified in the launch window, the prediction formula is for the fold, inner fold, and trial combination that corresponds to the final model. The final model is the model that produces the highest weighted average RSquare. The weight average RSquare is the average, weighted by number of observations, of the training RSquare, validation RSquare, and when a test set is used, the test RSquare.

Notes:

- A prediction formula is not saved for Stepwise models.
- A prediction formula is not saved for SVM models when the response is ordinal.
- This option is not available if Remove Live Reports is selected in the launch window.

Optional Reports Shows a submenu of additional report options.

Elapsed Time Shows or hides a report that contains the total elapsed time that was spent fitting each type of model.

Cardinality of Predictors Shows or hides a report of the number of levels and how many parameters are used in the linear model fit for each categorical predictor.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Decision Thresholds Report

In the Model Screening platform, the Decision Thresholds report enables you to explore thresholds for binary classification models. There is a Decision Thresholds report for each model data set that is specified by the validation method. These could be Training, Training and Validation, or Training, Validation, and Test sets. Each Decision Thresholds report contains a graph of the distribution of fitted probabilities, a bar chart of classifications, and confusion matrices. All are organized by model fit, fold, trial, and class level. The report also contains a tabbed section on classification accuracy measures and an option to set the profit matrix. The report updates as you adjust the probability threshold.

Distribution of Fitted Probabilities

The distribution of fitted probabilities, or model scores, enables you to see how each individual model fit differentiates between the two classes. A vertical line on the graph represents the probability threshold, which determines the classification of each observation. By default, the probability threshold is 0.5. You can change the probability threshold by dragging the vertical line or by clicking the Probability Threshold value and entering a new value. This changes the probability threshold value across the whole Decision Thresholds report. The value of the probability threshold must be between 0 and 1.

Classification Counts

The bar chart of classifications shows the classification counts for each level of the response variable at the current threshold. Green bars represent correctly classified observations; red bars represent incorrectly classified observations.

Confusion Matrices

The confusion matrices, also known as contingency tables, show the two-way classification of actual and predicted responses for each individual model fit. Confusion rates matrices are shown as well. The rates are equal to the values in the confusion matrices divided by the row totals.

Classification Accuracy Measures

False Classification by Threshold Shows a plot of the misclassification count by probability threshold and a plot of the misclassification rate by probability threshold. Each plot contains two curves for each individual model fit. The curves for the low response category are solid and the curves for the high response category are dashed. The curves intersect at the threshold that yields equal misclassification, counts or rates, of each response level. There is also a vertical line on each graph that represents the current probability threshold value. You can change the probability threshold value by dragging the vertical line. This changes the probability threshold value across the whole report.

False Classification by Portion Shows a plot of the misclassification count or rate by the portion of the rank ordered scores. Each plot contains two curves for each individual model fit. The curves for the low response category are solid and the curves for the high response category are dashed.

True Classification by Threshold Shows a plot of the true count by the probability threshold and a plot of the true rate by the probability threshold. Each plot contains two curves for each individual model fit. The curves for the low response category are dashed and those for the high response level are solid. The curves intersect at the threshold that yields equal correct classifications, counts or rates, for each response. There is also a vertical line on each graph that represents the current probability threshold value. You can change the probability threshold value by dragging the vertical line. This changes the probability threshold value across the whole report.

True Classification by Portion Shows a plot of the true count or rate by the portion of the rank ordered scores. Each plot contains two curves for each individual model fit. The curves for the low response category are dashed and the curves for the high response category are solid.

Profit by Threshold (Available only if a profit matrix is specified.) Shows a plot of the average profit by the probability threshold. There is a curve for each individual model fit and a vertical line that represents the current probability threshold value. The specified profit matrix is also shown next to the plot.

Metrics Shows a table of classification accuracy metrics for each model. A legend is provided to describe how the metric in each column is calculated.

Note: Two less common classification accuracy metrics are F1 and MCC. The F1 score is a combination of precision and recall, or sensitivity. Another way to calculate F1 is $2(\text{Precision} \times \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity})$. The Mathews correlation coefficient (MCC) is equivalent to the Pearson correlation coefficient estimated for two binary variables. See [“Statistical Details for the Pearson Product-Moment Correlation”](#).

Set Profit Matrix

Enables you to assign costs to undesirable outcomes and profits to desirable outcomes. See [“Specify Profit Matrix”](#). If you change the probability threshold in the profit matrix window and click OK, the Decision Thresholds report is updated using that value as the probability threshold.

Chapter 11

JMP[®] PRO Model Comparison

Compare the Predictive Ability of Fitted Models

The Model Comparison platform is available only in JMP Pro.

The Model Comparison platform enables you to compare the predictive ability of different models. Measures of fit are provided for each model along with overlaid diagnostic plots.

Figure 11.1 Example of Comparing Models



Contents

Example of Model Comparison 187

Launch the Model Comparison Platform 191

The Model Comparison Report..... 192

Model Comparison Platform Options 193

 Continuous Responses 193

 Categorical Responses 194

Additional Example of Model Comparison 195



Example of Model Comparison

This example compares a regression model and a bootstrap forest model. The data are demographic data and the goal is to build a model for median home price.

Begin by selecting **Help > Sample Data Folder** and opening Boston Housing.jmp.

Create a Validation Column

1. Select **Analyze > Predictive Modeling > Make Validation Column**.

2. Do not select any columns in the launch window.

This indicates that the platform will create a simple random validation column

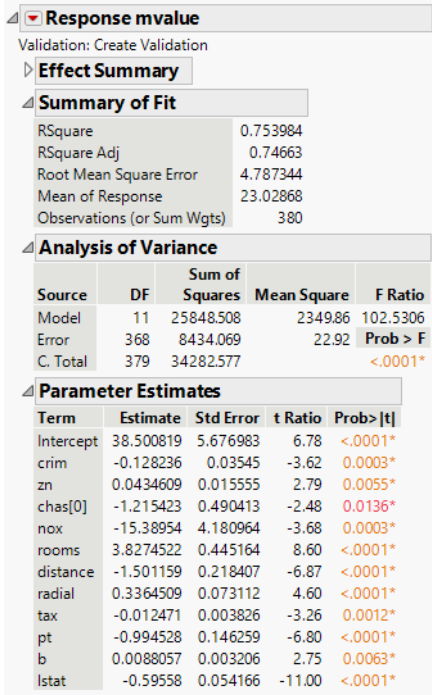
3. Click **OK**.
4. In the box next to New Column Name, type Create Validation.
5. In the box next to Random Seed, enter 1234.
6. Click **Go**.

A new Validation column is created. The rows assigned a 0 are the training set. The rows assigned a 1 are the validation set.

Create the Regression Model and Save the Prediction Formula to a Column

1. Select **Analyze > Fit Model**.
2. Select mvalue and click **Y**.
3. Select crim through lstat and click **Add**.
4. Select Create Validation and click **Validation**.
5. Select **Stepwise** in the Personality list.
6. Click the **Run** button.
7. Select **P-value Threshold** from the Stopping Rule list.
8. Click the **Go** button.
9. Click the **Run Model** button.

Figure 11.2 Fit Model Report



10. To save the prediction formula to a column, click the Response red triangle and select **Save Columns > Prediction Formula**.

Create the Bootstrap Forest Model and Save the Prediction Formula to a Column

1. Select **Analyze > Predictive Modeling > Bootstrap Forest**.
2. Select mvalue and click **Y, Response**.
3. Select crim through lstat and click **X, Factor**.
4. Select Create Validation and click **Validation**.
5. Click **OK**.
6. Select the **Early Stopping** check box.
7. Enter 617 in the box next to **Random Seed**.
8. Click **OK**.

Figure 11.3 Bootstrap Forest Model

Bootstrap Forest for mvalue			
Specifications			
Target	mvalue	Training Rows:	380
Validation Column:	Create Validation	Validation Rows:	126
		Test Rows:	0
Number of Trees in the Forest:	100	Number of Terms:	13
Number of Terms Sampled per Split:	10	Bootstrap Samples:	380
		Minimum Splits per Tree:	10
		Minimum Size Split:	5
Overall Statistics			
Individual Trees	RASE		
In Bag	2.040028		
Out of Bag	5.079566		
	RSquare	RASE	N
Training	0.937	2.37545	380
Validation	0.838	3.2154605	126

- To save the prediction formula to a column, click the Bootstrap Forest red triangle and select **Save Columns > Save Prediction Formula**.

Compare the Models

- Select **Analyze > Predictive Modeling > Model Comparison**.
- Select the two prediction formula columns and click **Y, Predictors**.
- Select Create Validation and click **Group**.

Tip: If a **Group** column is not specified, JMP automatically recognizes when the same validation column has been used for all predictors and prompts you to add it as a grouping variable.

- Click **OK**.

Figure 11.4 Model Comparison Report

Model Comparison								
Predictors								
Measures of Fit for mvalue								
Create Validation	Predictor	Creator	.2 .4 .6 .8	RSquare	RASE	AAE	Freq	
Training	Pred Formula mvalue 2	Fit Least Squares		0.7540	4.7111	3.3770	380	
Training	mvalue Predictor 2	Bootstrap Forest		0.9375	2.3755	1.4373	380	
Validation	Pred Formula mvalue 2	Fit Least Squares		0.6550	4.6972	3.3155	126	
Validation	mvalue Predictor 2	Bootstrap Forest		0.8383	3.2155	1.8809	126	

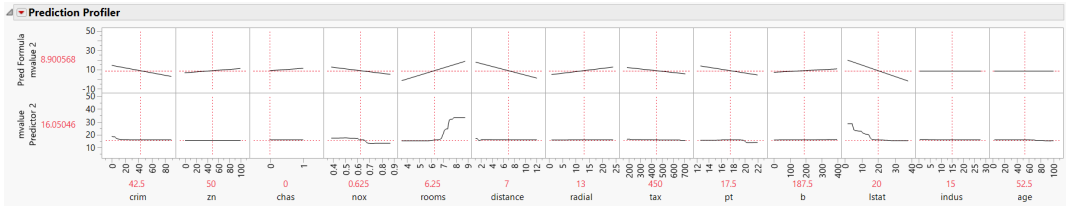
The rows in the training set were used to build the models, so the RSquare statistics for Create Validation = Training might be artificially inflated. In this case, the statistics are not

representative of the models’ future predictive ability. This is especially true for the bootstrap forest model.

Compare the models using the statistics for Create Validation = Validation. In this case, the bootstrap forest model predicts better than the regression model.

- 5. Click the Model Comparison red triangle and select Profiler.

Figure 11.5 Prediction Profiler for All Models



The prediction profiler enables you to compare the impact of each factor in the different models. The profiler is especially interesting when comparing different types of models such as here where you have a regression model and a partition model.

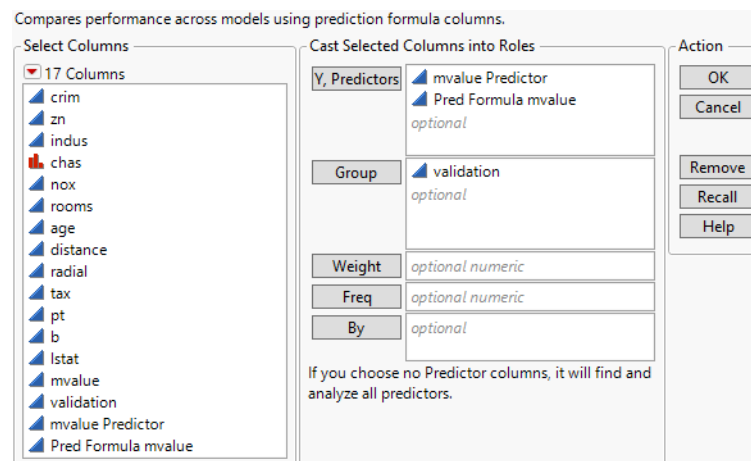
Related Information

- *Fitting Linear Models*
- [“Partition Models”](#)

JMP^{PRO} Launch the Model Comparison Platform

To launch the Model Comparison platform, select **Analyze > Predictive Modeling > Model Comparison**.

Figure 11.6 The Model Comparison Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Predictors The columns that contain the predictions for the models that you want to compare. They can be either formula columns or just data columns. Prediction formula columns created by JMP platforms have either the Predicting or Response Probability column property. If you specify a column that does not contain one of these properties, the platform prompts you to specify which column is being predicted by the specified Y column.

For a categorical response with k levels, most model fitting platforms save k columns to the data table, each predicting the probability for a level. All k columns need to be specified as **Y, Predictors**. For platforms that do not save k columns of probabilities, the column containing the predicted response level can be specified as a **Y, Predictors** column.

If you do not specify any **Y, Predictors** columns, JMP uses the prediction formula columns in the data table that have either the Predicting or Response Probability column property.

Group The column that separates the data into groups, which are evaluated separately. If a **Group** column is not specified, JMP automatically recognizes when the same validation column has been used for all predictors and prompts you to add it as a grouping variable.

The other role buttons are common among JMP platforms. See *Using JMP*.

JMP PRO The Model Comparison Report

The initial Model Comparison report contains a Predictors report and a Measures of Fit report. The Predictors report shows all responses and all models being compared for each response. The fitting platform that created the predictor column is also listed. The Measures of Fit report shows measures of fit for each model. The columns are different for continuous and categorical responses.

Measures of Fit for Continuous Responses

RSquare The r -squared statistic. In data tables that contain no missing values, the r -squared statistics in the Model Comparison report and original models match. However, if there are any missing values, the r -squared statistics differ.

RASE The square root of the mean squared prediction error. This is computed as follows:

- Square and sum the prediction errors (differences between the actual responses and the predicted responses) to obtain the SSE .
- Denote the number of observations by n .
- RASE is:

$$\text{RASE} = \sqrt{\frac{SSE}{n}}$$

AAE The average absolute error.

Freq The column that contains frequency counts for each row.

Measures of Fit for Categorical Responses

Entropy RSquare One minus the ratio of the negative log-likelihoods from the fitted model and the constant probability model. It ranges from 0 to 1.

Generalized RSquare A measure that can be applied to general regression models. It is based on the likelihood function L and is scaled to have a maximum value of 1. The value is 1 for a perfect model, and 0 for a model no better than a constant model. The Generalized RSquare measure simplifies to the traditional RSquare for continuous normal responses in the standard least squares setting. Generalized RSquare is also known as the Nagelkerke or Craig and Uhler R^2 , which is a normalized version of Cox and Snell's pseudo R^2 . See Nagelkerke (1991).

Mean -Log p The average of $-\log(p)$, where p is the fitted probability associated with the event that occurred.

RASE The root average squared prediction error. For categorical responses, the differences are between 1 and p , the fitted probability for the response level that actually occurred.

Mean Abs Dev The average of the absolute values of the differences between the response and the predicted response. For categorical responses, the differences are between 1 and p (the fitted probability for the response level that actually occurred).

Misclassification Rate The rate for which the response category with the highest fitted probability is not the observed category.

N The number of observations.

Related Information

[“Training and Validation Measures of Fit”](#) provides more information about measures of fit for categorical responses.



Model Comparison Platform Options

The options in the Model Comparison red triangle menu depend on whether the response is continuous or categorical.

- [“Continuous Responses”](#)
- [“Categorical Responses”](#)



Continuous Responses

When the response is continuous, the Model Comparison red triangle menu contains the following options:

Model Averaging Creates a new column of the arithmetic mean of the predicted values across models.

Plot Actual by Predicted Shows or hides a scatterplot of the actual versus the predicted values. The plots for the different models are overlaid.

Plot Residual by Row Shows or hides a plot of the residuals by row number. The plots for the different models are overlaid.

Profiler Shows or hides a profiler for each response based on prediction formula columns in your data. The profilers have a row for each model being compared.

Categorical Responses

When the response is categorical, the Model Comparison red triangle menu contains the following options:

Model Averaging Creates a new column of the arithmetic mean of the predicted probabilities across models.

ROC Curve Shows or hides ROC curves for each level of the response variable. The curves for the different models are overlaid.

AUC Comparison Shows or hides a comparison of the area under the ROC curve (AUC) from each model. The area under the curve is the indicator of the goodness of fit, where 1 is a perfect fit.

The report includes the following information:

- standard errors and confidence intervals for each AUC
- standard errors, confidence intervals, and hypothesis tests for the difference between each pair of AUCs
- an overall hypothesis test for testing whether all AUCs are equal

Lift Curve Shows or hides lift curves for each level of the response variable. The curves for the different models are overlaid.

Cum Gains Curve Shows or hides cumulative gains curves for each level of the response variable. A cumulative gains curve is a plot of the proportion of a response level that is identified by the model against the proportion of all responses. A cumulative gains curve for a perfect model would reach 1.0 at the overall proportion of the response level. The curves for the different models are overlaid.

Confusion Matrix Shows or hides confusion matrices for each model. A confusion matrix is a two-way classification of actual and predicted responses. Count and rate confusion matrices are shown. Separate confusion matrices are produced for each level of the Group variable.

If the response has a Profit Matrix column property, then Actual by Decision Count and Actual by Decision Rate matrices are shown to the right of the confusion matrices. For more information about these matrices, see [“Additional Examples of the Partition Platform”](#).

Profiler Shows or hides a profiler for each response based on prediction formula columns in your data. The profilers have a row for each model being compared.

Decision Threshold (Available only for binary categorical responses.) Shows or hides Decision Thresholds reports for the training, validation, and test sets, if specified. Each report contains a graph of the distribution of fitted probabilities for each model, confusion

matrices for each model, classification graphs to compare the model fits, and a table of classification accuracy metrics. See [“Decision Thresholds Report”](#).

Related Information

- [“ROC Curve”](#)
- [“Lift Curve”](#)



Additional Example of Model Comparison

This examples compares a logistic regression model and a decision tree model. The data are automobile data and the goal is to predict the size of a purchased car.

Begin by selecting **Help > Sample Data Folder** and opening Car Physical Data.jmp.

Create the Logistic Regression Model

1. Select **Analyze > Fit Model**.
2. Select Type and click **Y**.
3. Select the following columns and click **Add**: Country, Weight, Turning Cycle, Displacement, and Horsepower.
4. Click **Run**.

The Nominal Logistic Fit report appears.

5. To save the prediction formulas to columns, click the Nominal Logistic red triangle and select **Save Probability Formula**.

Create the Decision Tree Model and Save the Prediction Formula to a Column

1. Select **Analyze > Predictive Modeling > Partition**.
2. Select Type and click **Y, Response**.
3. Select the Country, Weight, Turning Cycle, Displacement, and Horsepower columns and click **X, Factor**.
4. Make sure that **Decision Tree** is selected in the Method list.
5. Click **OK**.

The Partition report appears.

6. Press Shift and click **Split**.
7. Enter 10 next to Enter Number of Splits.
8. Click **OK**.

- 9. To save the prediction formulas to columns, click the Partition red triangle and select **Save Columns > Save Prediction Formula**.

Compare the Models

- 1. Select **Analyze > Predictive Modeling > Model Comparison**.
- 2. Select all columns that begin with Prob and click **Y, Predictors**.
- 3. Click **OK**.

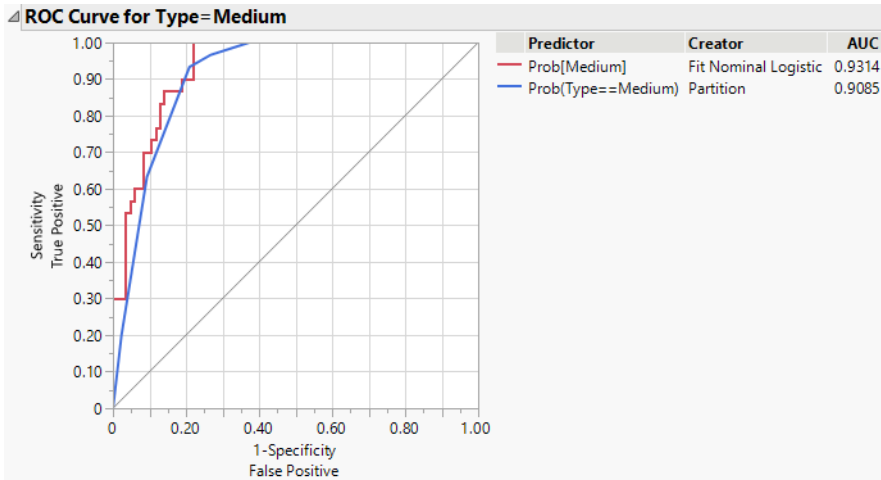
Figure 11.7 Initial Model Comparison Report

Model Comparison									
Predictors									
Measures of Fit for Type									
Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N	
Fit Nominal Logistic		0.5821	0.8798	0.6656	0.4780	0.3900	0.3103	116	
Partition		0.6248	0.9006	0.5976	0.4575	0.3986	0.2759	116	

The report shows that the Partition model has slightly higher values for Entropy RSquare and Generalized RSquare and a slightly lower value for Misclassification Rate.

- 4. Click the Model Comparison red triangle and select **ROC Curve**.
ROC curves appear for each Type, one of which is shown in [Figure 11.8](#).

Figure 11.8 ROC Curve for Medium



Examining all the ROC curves, you see that the two models are similar in their predictive ability.

- 5. Click the Model Comparison red triangle and select **AUC Comparison**.

AUC Comparison reports appear for each Type, one of which is shown in [Figure 11.9](#).

Figure 11.9 AUC Comparison for Medium

AUC Comparison for Type=Medium							
Predictor	AUC	Std Error	Lower 95%	Upper 95%			
Prob[Medium]	0.9314	0.0218	0.8742	0.9637			
Prob(Type==Medium)	0.9085	0.0255	0.8448	0.9477			
Predictor	vs. Predictor	AUC					
		Difference	Std Error	Lower 95%	Upper 95%	ChiSquare	Prob> ChiSq
Prob[Medium]	Prob(Type==Medium)	0.0229	0.0234	-0.023	0.0687	0.9553	0.3284
Test	ChiSquare	DF	Prob> ChiSq				
All AUCs equal	0.95534	1	0.3284				

The report shows results for a hypothesis test for the difference between the AUC values (area under the ROC curve). Examining the results, you see there is no statistical difference between the values for any level of Type.

You conclude that there is no large difference between the predictive abilities of the two models for the following reasons:

- The R Square values and the ROC curves are similar.
- There is no statistically significant difference between AUC values.

Chapter 12

Make Validation Column

Divide Data into Training, Validation, and Test Sets

The Make Validation Column platform is available only in JMP Pro.

Validation is the process of using part of a data set to estimate model parameters and using another part to assess the predictive ability of a model. You can partition the data into two or three sets or create four or more categories for K-Fold crossvalidation. The Make Validation Column platform provides several different methods to create these partitions.

Contents

Overview of the Make Validation Column Platform	201
Example of the Make Validation Column Platform	201
Launch the Make Validation Column Platform	203
Make Validation Column Report	206
Specify Rates or Relative Rates	206
Set Cutpoints	206
Set Number of Folds	207
Options	207
Go	208
Make Validation Column Platform Options	208
Additional Example of the Make Validation Column Platform	208

Overview of the Make Validation Column Platform

Validation is the process of using part of a data set to estimate model parameters and using another part to assess the predictive ability of a model. With complex data, this can reduce the risk of model overfitting.

One use for a validation column is to partition the data into two or three parts.

- The training set is used to estimate the model parameters.
- The validation set is used to help choose a model with good predictive ability.
- The testing set checks the model's predictive ability after a model has been chosen.

Another use for a validation column is to partition the data into four or more folds to use in K-Fold crossvalidation.

A validation column can be used as a validation method in many JMP platforms, but K-Fold crossvalidation through a validation column is only supported by a few platforms. See [“Validation in JMP Modeling”](#).

The Make Validation Column platform enables you to create training, validation, and test sets using a variety of methods. You can specify stratification, grouping, or cutpoint columns to determine the method used to create the validation column.

Example of the Make Validation Column Platform

Use the Make Validation Column platform to create a stratified validation column to balance the levels of a particular variable across the training and validation sets.

1. Select **Help > Sample Data Folder** and open Lipid Data.jmp.
2. Select **Analyze > Distribution**.
3. Select Gender and click **Y, Columns**.
4. Click **OK**.

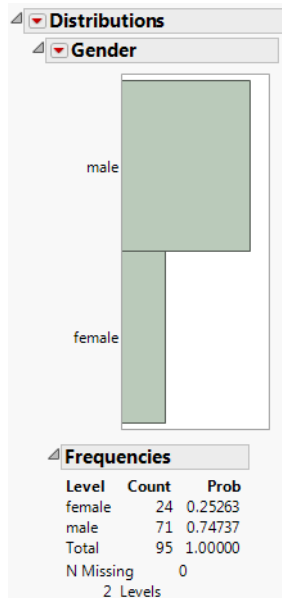
Figure 12.1 Distribution of Gender in Lipid Data.jmp

Figure 12.1 illustrates the distribution of Gender in the data set. Notice that there is not an equal proportion of males and females represented. Because there are fewer females within the data, you want to be sure to balance the genders across the validation and training sets.

5. Select **Analyze > Predictive Modeling > Make Validation Column**.
6. Select Gender and click **Stratification Columns**.
7. Click **OK**.

The Make Validation Column report appears with a description of the validation method you selected. There are also options to change the rates, column types, or set a seed.

8. (Optional) Type 1234 next to **Random Seed** in the Options section of the report.
9. Click **Go**.

A Validation column is added to the data table. You can explore the distribution of the validation and training sets by creating a Mosaic Plot.

10. Select **Analyze > Fit Y by X**.
11. Assign Validation to **Y, Response**, and Gender to **X, Factor**.
12. Click **OK**.

Figure 12.2 Distribution of Gender across Validation and Training Sets

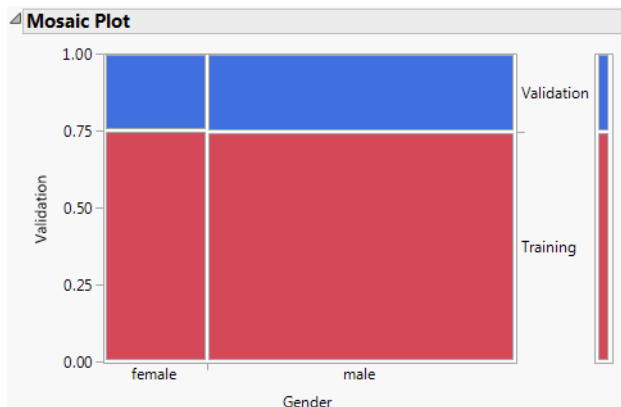


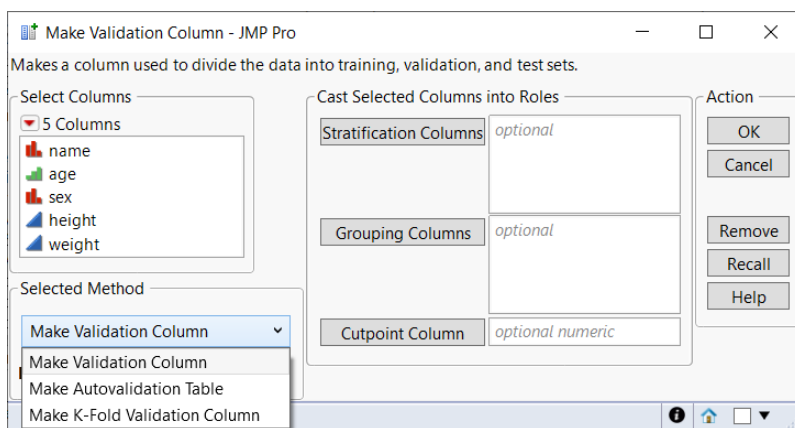
Figure 12.2 illustrates the distribution of Gender across each of the validation and training sets. Note that about 75% of both females and males are in the training set and about 25% of both females and males are in the validation set.



Launch the Make Validation Column Platform

Launch the Make Validation Column platform by selecting **Analyze > Predictive Modeling > Make Validation Column**.

Figure 12.3 Make Validation Column Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Make Validation Column launch window provides the following options:

Stratification Columns Assigns one or more stratification columns.

Grouping Columns Assigns one or more categorical grouping columns.

Cutpoint Column Assigns a numeric cutpoint column.

Cutpoint Batch ID When a cutpoint column is assigned, you can also assign a column for cutpoint batch IDs. This enables you to determine cutpoint values within each level of the Cutpoint Batch ID column.

Selected Method

Provides three methods for validation.

Make Validation Column creates a validation column based on the specified stratification, grouping, and cutpoint columns. The validation column method determined by the specified stratification, grouping, and cutpoint columns is described below the box. After a method is selected and you click OK, you specify the allocations for each set in the Make Validation Column report. See [“Specify Rates or Relative Rates”](#) and [“Set Cutpoints”](#). There are five methods for constructing the holdback sets. All of these methods, except for Cutpoint Validation, are also used to create the folds for K-Fold Validation. See [“Make K-Fold Validation Column”](#).

Random Validation Column The default method if there are no column assignments in the launch window. This method partitions the data into sets or folds based on the allocations entered in the Make Validation Column report.

Stratified Validation Column The selected method if one or more stratification columns are assigned. This method partitions the data into balanced sets based on the levels of the specified stratification columns. As in the Random Validation Column method, rows are randomly assigned to the holdback sets or folds based on the allocations entered in the Make Validation Column report. However, this is done at each level or combination of levels of the stratifying columns. Use this method when you want a balanced representation of the levels of a column in each of the training, validation, and test sets or in each of the folds in K-Fold crossvalidation.

Grouped Validation Column The selected method if one or more grouping columns are specified. This method partitions the data into sets in such a way that entire levels of a specified column or combinations of levels of two or more columns are placed in the same set or fold. Because of this, the sizes of the resulting sets vary slightly from the sizes that you specified. Use this option when splitting levels across holdback sets or folds is not desirable.

Stratify by Group Validation Column The selected method if both stratification and grouping columns are specified. This method partitions the data to balance the levels across the stratification column while requiring that the specified groups stay together in the same holdback sets or folds. As in Grouped Validation Column, groups can be created as levels of a specified column or combination of levels of two or more columns. The sizes of the resulting sets vary slightly from the sizes you specified.

Cutpoint Validation Column The selected method if a cutpoint column is specified. This method partitions the data into sets based on the time series cutpoints. Use this option when you want to assign your data to holdback sets based on time periods. The training set consists of rows between the first cutpoint and the second cutpoint. The validation set consists of rows between the second and third cutpoints. The test set consists of the remaining rows. These sets are chosen based on options in the Set Cutpoints report.

Make Autovalidation Table Creates a new data table that contains a duplication of the rows in the original data table concatenated to the rows in the original data table. The new data table, which can be used for crossvalidation, has four additional columns:

Valid Set Assigns a value of 0 to the original data and a value of 1 to the duplicated data. The values in this column designate the training and validation sets. Use this column in the Validation role in the launch window of an analysis.

Valid ID Assigns the row number of the original observation. This allows matching of training and validation set rows for each original observation.

Valid Weight Assigns the autovalidation weight, to be used in the Freq role in the launch window of an analysis. For each value of Valid ID, the same uniform random number is generated for the training observation and the validation observation. For the training set, Valid Weight is calculated as follows:

$$\text{Valid Weight} = -\log(1 - \text{Valid Uniform})$$

For the validation set, Valid Weight is calculated as follows:

$$\text{Valid Weight} = -\log(\text{Valid Uniform})$$

The Valid Weight column is constructed so that the training data weights are negatively correlated with the validation data weights. This ensures that the difference in the fit of the validation data yields an effective crossvalidation of the fitting method.

Null Factor Assigns the same normal random number for each value of Valid ID.

Tip: Use Make Autovalidation Table for small data tables, where using a subset as the training data could cause estimation problems.

Make K-Fold Validation Column Creates a validation column with four or more categories, based on the specified stratification and grouping columns. Each category represents a fold to use in K-Fold crossvalidation. The Y column is used to order the rows and each row is then sequentially assigned to a fold. The validation column method determined by the specified stratification and grouping columns is described below the box. These are the same methods described in [“Make Validation Column”](#). After a method is selected and you click OK, you specify the number of folds, K , in the Make Validation Column report. See [“Set Number of Folds”](#).

Missing Values

A missing value in a stratification, grouping, or cutpoint column results in a missing value in the validation column for that row.



Make Validation Column Report

The Make Validation Column report contains a description of the validation method you selected in the launch window and relevant options for rates, cutpoints, column type, and setting a random seed.



Specify Rates or Relative Rates

This section of the Make Validation Column report enables you to specify the allocations for the training, validation, and test sets. In the boxes next to Training Set, Validation Set, and Test Set, enter values that represent the proportions or numbers of rows that you would like to include in each of these sets. The default values construct a training set that contains about 75% of the rows and a validation set that contains about 25% of the rows.

Depending on the number of rows in the data table and the selected validation method, the actual sizes of the resulting sets might vary slightly from the sizes that you specified. The Adjusted Rates, Row Counts, and Group Count columns provide the actual sizes of the sets.



Set Cutpoints

This section of the Make Validation Column report enables you to specify how cutpoints are determined. It appears only when a cutpoint column is specified in the launch window. The cutpoints are determined by one of the following four methods:

Proportions Determines the cutpoints based on the proportions of rows specified for each set. In the boxes next to Training Set, Validation Set, and Test Set, enter values that represent the proportions that you would like to include in each of these sets. Depending

on the number of rows and the proportions, the actual sizes of the resulting sets might vary slightly from the sizes you specified.

Number of Rows Determines the cutpoints based on the number of rows specified for each set. In the boxes next to Training Set, Validation Set, and Test Set, enter values that represent the number of rows that you would like to include in each of these sets. This option enables you to specify the sets exactly.

Fixed Time or Date Determines the cutpoints based on fixed data points in the specified cutpoint column. If you select this option, the minimum and maximum values for the cutpoint column are shown for reference. In the boxes next to Training Set, Validation Set, and Test Set, enter the value that represents the minimum value that you would like to include in each of these sets.

Elapsed Time Determines the cutpoints based on the amount of time that has elapsed since the first timestamp in the cutpoint column. If you select this option, the total elapsed time from the first value to the last value is shown for reference. In the boxes next to Training Set, Validation Set, and Test Set, enter the values that represent the amounts of elapsed time that you would like to include in each of these sets.

Set Number of Folds

This section of the Make Validation Column report enables you to specify K , the number of folds for K -Fold crossvalidation. The default and minimum value for K is 4 and the maximum value is the number of levels in the Grouping Columns, if specified. Specifying a grouping column that has less than four levels results in an error message.

Options

This section of the Make Validation Column report enables you to specify the following additional options:

New Column Name Enables you to specify the name of the validation column.

Validation Column Type Enables you to specify the validation column type as fixed or formula.

- The fixed column type produces a column with fixed values.
- The formula column type produces a formula column as the validation column.

Random Seed (Not available if a formula column type is specified.) Enables you to set a random seed to reproduce the same validation column in the future.

**JMP[®]
PRO** Go

Click **Go** in the Make Validation Column report to add the validation column to the data table. The validation column has a Notes property that gives the stratifying, grouping, or cutpoint variables.

**JMP[®]
PRO** Make Validation Column Platform Options

The Make Validation Column red triangle menu contains the following options:

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

**JMP[®]
PRO** Additional Example of the Make Validation Column Platform

In this example, you compare two cutpoint validation columns. The first is created using only a cutpoint column and the second is created using both a cutpoint column and a batch ID column. The data are weekly weather data collected over one year from 16 weather stations across the United States. Not every weather station has a weekly temperature measurement for every week of the year. For each weather station, you want to use the first 60% of observations for training, the next 25% of observations for validation, and the final 15% of observations for testing.

Create Validation Column with Cutpoints

1. Select **Help > Sample Data Folder** and open Functional Data/Weekly Weather Data.jmp.
2. Select **Analyze > Predictive Modeling > Make Validation Column**.
3. Select Week of Year and click **Cutpoint Column**.
4. Click **OK**.

5. In the list next to Determine cutpoints using, select **Proportions**.
6. In the boxes next to Training Set, Validation Set, and Test Set, enter 0.60, 0.25, and 0.15, respectively.
7. In the box next to New Column Name, type Cutpoint Validation.
8. Click **Go**.

A validation column called Cutpoint Validation is added to the data table.

9. Select **Analyze > Tabulate**.
10. Click ID and drag it to the Drop zone for rows.
11. Click Cutpoint Validation and drag it on top of N.
12. Click Row% and drag it on top of the cells.

Figure 12.4 Cutpoint Validation Column Proportions

ID	Cutpoint Validation		
	Training	Validation	Test
	Row %	Row %	Row %
Bristol, TN	62.79%	20.93%	16.28%
Corpus Christi, TX	70.27%	18.92%	10.81%
Denver, CO	58.14%	23.26%	18.60%
Goodwell, OK	64.71%	17.65%	17.65%
Grand Canyon National Park, AZ	53.66%	26.83%	19.51%
Greenville, ME	60.47%	20.93%	18.60%
Harrisburg, PA	60.53%	26.32%	13.16%
Helena, MT	60.00%	25.00%	15.00%
Lynchburg, VA	71.79%	17.95%	10.26%
Madera, CA	67.50%	22.50%	10.00%
Miami Beach, FL	63.16%	26.32%	10.53%
Minneapolis/St. Paul, MN	57.50%	27.50%	15.00%
N. Myrtle Beach, SC	69.05%	16.67%	14.29%
Pearson Ridge Range, LA	57.89%	28.95%	13.16%
Philip, SD	59.52%	23.81%	16.67%
Terra Haute, IN	47.22%	33.33%	19.44%

Figure 12.4 shows that not all of the weather stations have the correct proportions for the training, validation, and test sets. Use a Batch ID column to obtain the correct percentages.

Create Validation Column with Cutpoints and Batch ID

1. Select **Analyze > Predictive Modeling > Make Validation Column**.
2. Select Week of Year and click **Cutpoint Column**.
3. Select ID and click **Cutpoint Batch ID**.
4. Click **OK**.
5. In the list next to Determine cutpoints using, select **Proportions**.
6. In the boxes next to Training Set, Validation Set, and Test Set, enter 0.60, 0.25, and 0.15, respectively.
7. In the box next to New Column Name, type Cutpoint Batch Validation.

8. Click **Go**.
- A validation column called Cutpoint Batch Validation is added to the data table.
9. Select **Analyze > Tabulate**.
10. Click ID and drag it to the Drop zone for rows.
11. Click Cutpoint Batch Validation and drag it on top of N.
12. Click Row% and drag it on top of the cells.

Figure 12.5 Cutpoint Validation Column with Batch ID Proportions

	Cutpoint Batch Validation		
	Training	Validation	Test
ID	Row %	Row %	Row %
Bristol, TN	60.47%	25.58%	13.95%
Corpus Christi, TX	59.46%	24.32%	16.22%
Denver, CO	60.47%	25.58%	13.95%
Goodwell, OK	58.82%	26.47%	14.71%
Grand Canyon National Park, AZ	60.98%	24.39%	14.63%
Greenville, ME	60.47%	25.58%	13.95%
Harrisburg, PA	60.53%	26.32%	13.16%
Helena, MT	60.00%	25.00%	15.00%
Lynchburg, VA	58.97%	25.64%	15.38%
Madera, CA	60.00%	25.00%	15.00%
Miami Beach, FL	60.53%	26.32%	13.16%
Minneapolis/St. Paul, MN	60.00%	25.00%	15.00%
N. Myrtle Beach, SC	59.52%	26.19%	14.29%
Peason Ridge Range, LA	60.53%	26.32%	13.16%
Philip, SD	59.52%	26.19%	14.29%
Terra Haute, IN	61.11%	25.00%	13.89%

Figure 12.5 shows that using a Cutpoint Batch ID column ensures that each weather station has proportions for the training, validation, and test sets that are much closer to the specified values.

Chapter 13

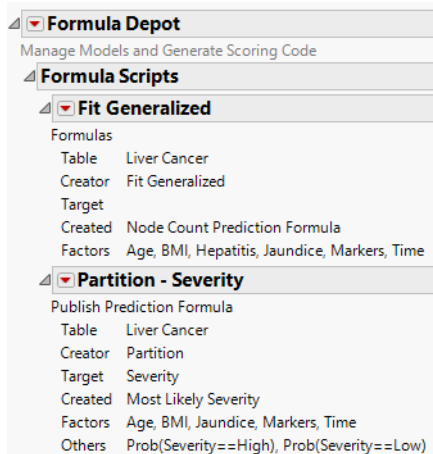
JMP[®] PRO Formula Depot

Manage Models and Generate Scoring Code

The Formula Depot Platform is available only in JMP Pro.

The Formula Depot is a repository to organize, compare, and profile models. Scoring code for deployment within or outside of JMP can be generated for models published to the Formula Depot. For model exploration work, you can use the Formula Depot to store candidate models outside of your JMP data table. The model profiler and model compare platforms are accessible from the Formula Depot. A model that is selected for further use can be saved to your JMP table or saved to a JMP table with new data for scoring. For use in an environment outside of JMP, you can generate scoring code in C, Python, JavaScript, SAS, or SQL.

Figure 13.1 Example of the Formula Depot



Contents

Overview of the Formula Depot Platform..... 213

Example of the Formula Depot..... 213

Launch the Formula Depot Platform..... 214

 Platforms That Publish Prediction Formulas to the Formula Depot..... 215

Formula Depot Platform Options..... 215

 Formula Depot Model Options..... 217

Generate Scoring Code from the Formula Depot Platform 218



Overview of the Formula Depot Platform

The Formula Depot is a repository to organize, compare, profile, and score models for deployment. Models are prediction formulas. Prediction formulas are saved to the Formula Depot as column scripts. You can add prediction formulas to JMP tables to score data. You can also use the Formula Depot to generate scoring code for prediction formulas to facilitate the deployment of models in environments outside of JMP.

The Formula Depot enables you to perform the following tasks:

- save prediction formulas outside of data tables
- save prediction formulas from multiple data tables in a common location
- save intermediate cleaning formulas within a prediction formula
- compare models
- profile models
- add prediction formulas to a data table for scoring of new data within JMP
- generate scoring code (C, Python, JavaScript, SAS, or SQL) for deploying models outside of JMP
- The Formula Depot includes intermediate formulas in the generated scoring code. A model constructed from inputs with column formulas includes the corresponding formulas in the scoring code. For example, if your model includes a recoded variable the recode formula is included in the scoring code. This enables you to apply your prediction formula to raw data that has not been processed.

Note: Informative missing is not a supported intermediate formula.



Example of the Formula Depot

This example uses an existing script in a sample data table to generate a model and then publishes the prediction formula from the model to the Formula Depot. The data table contains data on the severity of liver cancer in patients when they entered a study.

1. Select **Help > Sample Data Folder** and open Liver Cancer.jmp.
2. Click the green triangle next to the Lasso Poisson, Validation Column script.
3. Click the red triangle next to Poisson Adaptive Lasso with Validation Column and select **Save Columns > Publish Prediction Formula**.

This option opens a Formula Depot that contains the prediction formula for the Fit Generalized model.

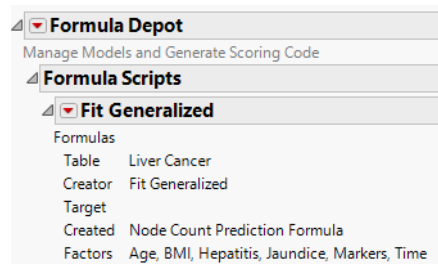
- To add the prediction formula to the data table, select **Run Script** from the Fit Generalized - Node Count red triangle in the Formula Depot.

Note: The result of [step 3](#) and [step 4](#) can be obtained using the **Save Columns > Save Prediction Formula** at [step 3](#). However, then the prediction formula is not part of the Formula Depot

- To generate scoring code for use outside of JMP select a code type from the Fit Generalized - Node Count red triangle. A script window appears containing the code. See [“Generate Scoring Code from the Formula Depot Platform”](#).
- To save the Formula Depot, select **File > Save**.

Note: The Formula Depot is saved as a JMP Report File (*.jrp).

Figure 13.2 Formula Depot with a Generalized Model

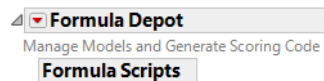


**JMP
PRO**

Launch the Formula Depot Platform

Launch the Formula Depot by selecting **Analyze > Predictive Modeling > Formula Depot**.

Figure 13.3 Empty Formula Depot from Launch



Alternatively, if there is not an open Formula Depot then a Formula Depot opens when you select a Publish command.

Platforms That Publish Prediction Formulas to the Formula Depot

The platforms that publish prediction formulas and generate scoring code include:

- Discriminant
- Least Squares Regression
- Logistic Regression
- Partition
- Uplift
- K Nearest Neighbors
- Naive Bayes
- Neural
- Latent Class Analysis
- Principal Components
- Generalized Regression
- PLS
- Gaussian Process

In platforms that do not publish prediction formulas to the Formula Depot, you can save the prediction formula to the data table. From the data table, add it to the Formula Depot by selecting **Add Formula from Column**. However, the scoring code might not be fully functional for such models. For more information about scoring code see [“Generate Scoring Code from the Formula Depot Platform”](#).

In addition, you can publish a formula directly from the formula editor. Select the formula, right-click, and select **Publish to Formula Depot**.

Formula Depot Platform Options

The Formula Depot red triangle menu contains the following options:

Add Formula From Column Enables you to add existing prediction formulas from data table columns to the current formula depot. If your table has only one formula, that formula is added to the depot. If you have multiple column formulas, a dialog box enables you to select the formulas of interest. If your formula columns are selected in your data table, they will be pre-selected in the dialog box.

Show Scripts Opens a new Formula Window (or appends to an open Formula Window) that contains scripts for all of the formulas that are in the current formula depot.

Copy Scripts Copies all scripts from the current formula depot to the clipboard.

Copy Formulas as Functions Enables you to select models from the current formula depot to be copied as functions for use in JSL. Selected models are copied onto the clipboard. The functions act on scalar variables rather than columns. One Namespace is generated for each selected model. Each Namespace has one function, called `predict`, that takes the model inputs as parameters and returns the model predictor.

Copy Formulas as Transforms Enables you to select models from the current formula depot to be copied. Selected models are copied onto the clipboard within a `Transform Columns()` statement.

Run Scripts Enables you to save models from the current formula depot to new columns in your JMP data table. Formulas with intermediate column formulas will add the intermediate columns to your JMP data table.

Note: If you have more than one data table open, a Choose Table window appears that enables you to specify which data table to use for Run Scripts.

Generate C Code, Generate Python Code, Generate JavaScript Code, Generate SAS Code, and Generate SQL Code Enables you to select models from the current formula depot for code generation. A new script window appears that contains scoring code for the selected models in C, Python, JavaScript, SAS DS2, or SQL, respectively. You can use this code to facilitate the deployment of the model in the environment or framework of your choice. See [“Generate Scoring Code from the Formula Depot Platform”](#).

Model Comparison Enables you to select models from the current formula depot to be compared using the model comparison utility. If models from multiple tables are stored in the depot, first select the table of interest and then the models of interest. Hold the shift key when selecting the Model Comparison option to launch the Model Comparison in a new window. See [“Model Comparison”](#).

Note: If you have more than one data table open, a Choose Table window appears that enables you to specify which data table to use for the Model Comparison.

Remove Model Comparison Removes all Model Comparison reports from the current formula depot.

Profiler Enables you to select models from the current formula depot to be profiled using the profiler. If models from multiple tables are stored in the depot you first select the table of interest and then the models of interest. Hold the shift key when selecting the profiler option to launch the profiler in a new window. See *Profilers*.

Note: If you have more than one data table open, a Choose Table window appears that enables you to specify which data table to use for the Profiler.

Remove Profiler Removes all Profilers from the current formula depot.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

Formula Depot Model Options

Each prediction formula that is saved in the Formula Depot has an individual options menu. In addition to Show, Copy, Run, and Code Generation options that correspond to the main menu options, the individual menus include the following options:

Rename New Column Enables you to rename the model. This new name is applied as the column name if you run the script to add the prediction formula to a JMP data table. In addition, this name is included in generated code.

Remove Removes the model from the Formula Depot. This command cannot be undone.



Generate Scoring Code from the Formula Depot Platform

Scoring code generation is intended to facilitate using models built in JMP in a production environment or other framework of your choice. Many platforms publish prediction formulas to the Formula Depot, however; not all prediction formulas generate complete code. Code can be complete, it can be a code fragment, or the code can include unsupported functions that require additional programming for implementation. Error messages often indicate an unsupported function call that requires additional programming in the targeted language. Target language specific libraries might be required for the implementation of more complex functions. These libraries are discussed in the language specific sections below. For example, you can perform the following tasks:

- Deploy your model to SAS Model Manager using the generated SAS code.
- Augment your ETL process with in-database scoring using the generated SQL code.
- Create a node for a data transformation pipeline with an application built with the generated C code.
- Create a Jupyter notebook to show live scoring results using the generated Python code.
- Enable customers to score their own data with a web application that includes the generated JavaScript.

For the C, Python, and JavaScript languages, you must include supporting code such as .h files and utility libraries when deploying or compiling the generated code. These files are available in your JMP installation folder inside the Scoring folder.

Tip: To find the location of installation folder on your machine, use the JSL command `Get Path Variable("$ALL_HOME/Scoring");`.

C Code The C scoring code that is generated must be compiled into a library and then linked into an application. You might use either a static or a dynamic link approach. The files `jmp_lib.h`, `jmp_parms.h`, and `jmp_score.h` needed for compiling and linking can be found in the Scoring/C folder in your JMP installation folder.

Python Code The file `jmp_score.py` that is needed to run your Python scoring application can be found in the Scoring/Python folder in your JMP installation folder. When deploying your models, copy `jmp_score.py` to the same location as your executable scoring code.

When calling the generated code to score your data, it is important that the input data is passed in a data structure compatible with the code generated by JMP. The Python scoring code expects the input and output arguments to be of a dictionary-like type. Values can be

accessed and written by key, such as DataFrame row objects from the Spark, Dask and Pandas frameworks, or the standard Python “dict” class.

Use of linear algebra operators or advanced operators such as Vec Quadratic and Design Norm require the NumPy Python library.

JavaScript Code The file `jmp_score.js` that is needed to run your JavaScript scoring application can be found in `Scoring/JavaScript` folder.

SAS Code The generated code fragment that once wrapped by PROC DS2 statements can be used in SAS applications including the SAS In-Database Code Accelerator. The code is added to a SAS Window and includes a variable name mapping section in the comment block that precedes the code. Any temporary variable names associated with intermediate columns are deleted at the end of the DATA Step by the drop command.

Tip: For models with an ifmax call, such as logistic or neural models for a categorical response, move temporary variable declarations before the **method run()** statement.

SQL Code The SQL code fragment that once wrapped in a select statement can be used in SQL queries against most major database servers.

Note: Code that contains “placeholder” or “ERROR” indicates an unsupported function call.

Chapter 14

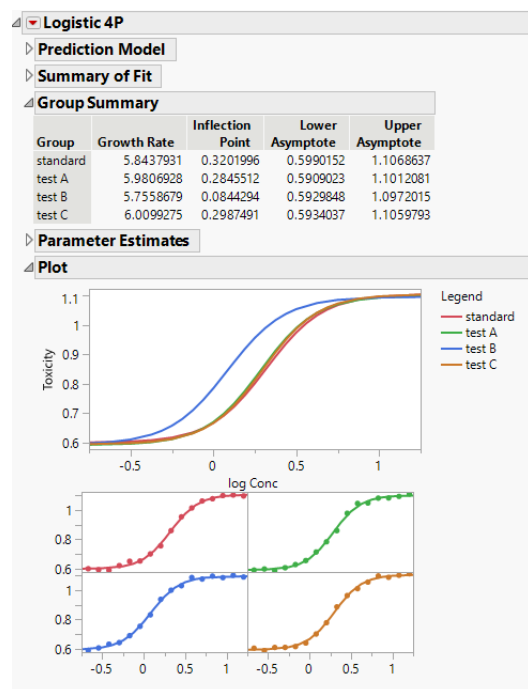
Fit Curve

Fit Built-In Nonlinear Models to Your Data

The Fit Curve platform enables you to fit models that are *nonlinear* in the parameters. In many situations, especially in the physical and biological sciences, well-known nonlinear equations describe the relationship between variables. For example, pharmacological bioassay experiments can demonstrate how the strength of the response to a drug changes as a function of drug concentration. Sigmoid curves often accurately model response strength as a function of drug concentration. Another example is exponential growth curves, which can model the size of a population over time.

The Fit Curve platform provides predefined models, such as polynomial, logistic, probit, Gompertz, exponential, peak, pharmacokinetic, rate, and dissolution models. The use of predefined models means that you are not required to create model formulas or specify starting values for parameter estimates. To specify your own starting values and create model formulas, use the Nonlinear platform, which can fit specified nonlinear models. See [“Nonlinear Regression”](#).

Figure 14.1 Example of Nonlinear Fit in the Fit Curve Platform



Contents

Overview of the Fit Curve Platform.....	223
Example of the Fit Curve Platform.....	224
Launch the Fit Curve Platform	227
The Fit Curve Report	229
Model Comparison Report	229
Model Fit Report	231
Fit Curve Options	231
Model-Free Comparisons Reports	234
Model Fit Options	236
Test Parallelism	238
Compare Parameter Estimates	241
Equivalence Test	242
Additional Example of the Fit Curve Platform.....	243
Statistical Details for the Fit Curve Platform.....	245
Statistical Details for Fit Curve Models	246
Statistical Details for Dissolution Curve Analysis	256

Overview of the Fit Curve Platform

Some models are *linear* in the parameters (for example, a quadratic or other polynomial); others can be transformed to be such (for example, when you use a log transformation of x). The Fit Model or Fit Y by X platforms are more appropriate in these situations. An example in *Fitting Linear Models* shows a significant linear relationship between oxygen uptake and time spent running. For more information about Fit Model, see *Fitting Linear Models*. For more information about Fit Y by X, see *Basic Analysis*.

The Fit Curve platform enables you to fit models that are *nonlinear* in the parameters. The initial example in this chapter shows the analysis of a nonlinear relationship: drug toxicity as a function of concentration. The effect of concentration on toxicity changes from low to high doses, so this relationship is nonlinear.

The following are examples of equations for linear and nonlinear functions.

Linear function: $Y = \beta_0 + \beta_1 e^x$

Nonlinear function: $Y = \beta_0 + \beta_1 e^{\beta_2 x}$

The Fit Curve platform provides predefined models, such as polynomial, logistic, probit, Gompertz, exponential, peak, pharmacokinetic, rate, and dissolution models. Specifying a grouping variable lets you estimate separate model parameters for each level of the grouping variable. The fitted models and estimated parameters can be compared across the levels of the grouping variable. There is also a set of options to compare dissolution profiles, with options for both parametric and nonparametric techniques.

Fit Curve also enables you to build a model to create the prediction formula. Then you set upper and lower parameter limits in Nonlinear. See [“Example of Setting Parameter Limits”](#). In JMP Pro, you can also specify a set of supplementary variables and fit a generalized regression model within the Fit Curve platform to determine how these variables affect the response.

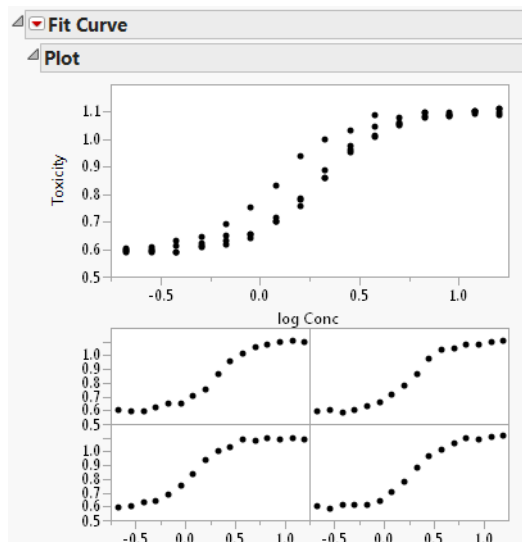
Example of the Fit Curve Platform

This example shows how to build a model for toxicity as a function of the concentration of a drug. You have a standard formulation of the drug and want to compare it to three new formulations.

You are interested in a toxicity ratio of non-surviving to surviving cells at a specific concentration of each drug. From prior research, you know the toxicity ratios for 16 different concentrations of each drug formulation. A higher ratio indicates more toxicity, which could be detrimental to development of the drug. Log concentration was calculated to decrease the range of concentration values and make it easier to detect differences in the curves.

1. Select **Help > Sample Data Folder** and open Nonlinear Examples/Bioassay.jmp.
2. Select **Analyze > Specialized Modeling > Fit Curve**.
3. Assign Toxicity to the **Y, Response** role.
4. Assign log Conc to the **X, Regressor** role.
5. Assign Formulation to the **Group** role.
6. Click **OK**.

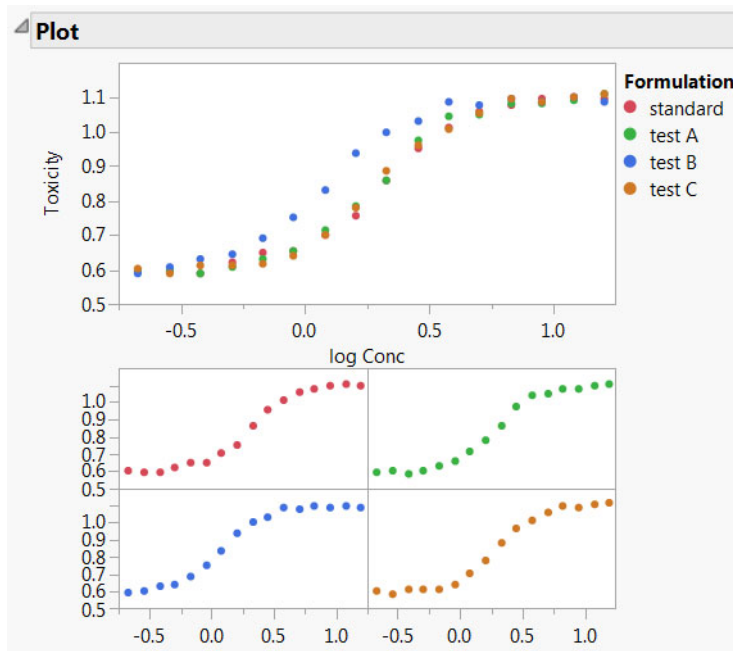
Figure 14.2 Initial Fit Curve Report



The Plot report contains an overlaid plot of the fitted model of each formulation.

7. To see a legend identifying each drug formulation, right-click one of the graphs and select **Row Legend**. Select **Formulation** for the column and click **OK**.

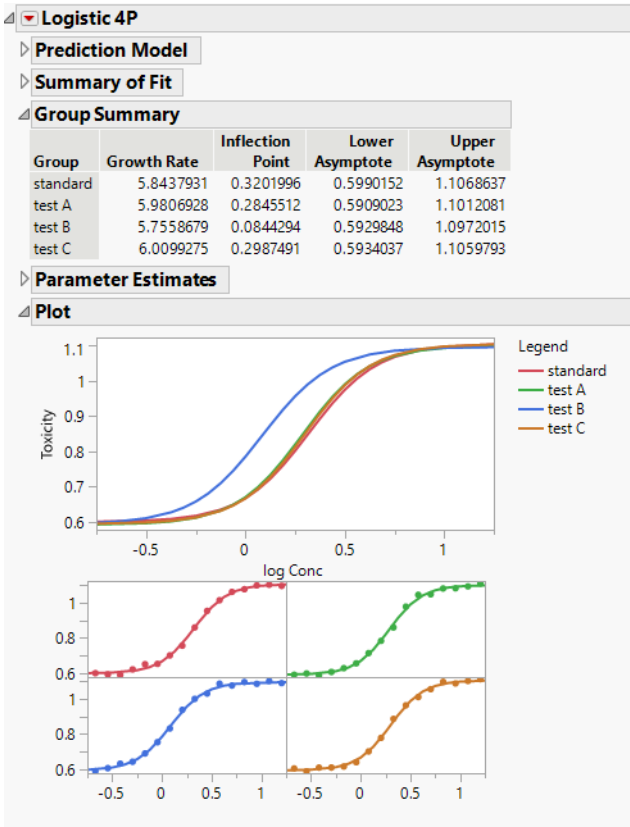
Figure 14.3 Fit Curve Report with Plot Legend



The curves appear S-shaped, so a sigmoid curve would be an appropriate fit. [Table 14.1](#) shows formulas and graphical depictions of the different types of models that the Fit Curve platform offers.

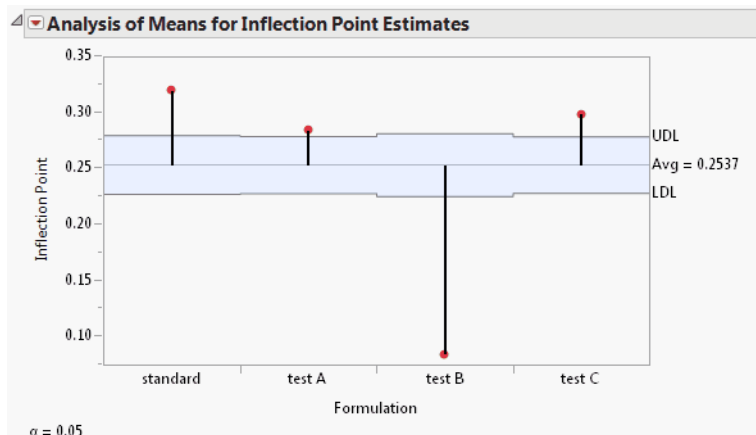
8. Click the Fit Curve red triangle and select **Sigmoid Curves > Logistic Curves > Fit Logistic 4P**.

Figure 14.4 Logistic 4P Report



- The Logistic 4P report appears. There is also a separate plot for each drug formulation. The plot of the fitted curves suggests that formulation B might be different, because the test B curve starts to rise sooner than the others. Inflection point parameters cause this rise.
- Click the Logistic 4P red triangle and select **Compare Parameter Estimates**.

Figure 14.5 Parameter Comparison Report

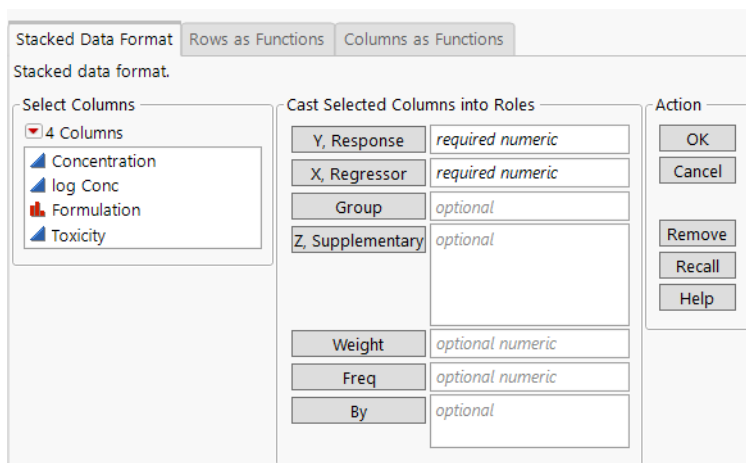


Notice that the Inflection Point parameter for the test B formulation is significantly lower than the average inflection point. This agrees with the plots shown in [Figure 14.4](#). Drug formulation B has a higher toxicity ratio than the other formulations.

Launch the Fit Curve Platform

To launch the Fit Curve platform, select **Analyze > Specialized Modeling > Fit Curve**.

Figure 14.6 Fit Curve Platform Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Select Data Format

The Fit Curve launch window includes tabs for different types of data formats. Select the tab based on your data format.

Stacked Data Format Select for data tables where each row corresponds to a single observation. There are separate columns for the response, regressors, and group variables.

Rows as Functions Select for data tables where each row corresponds to the full response function for one level of the Group variable. Each column corresponds to a level of the regressor variable.

Tip: If the columns contain the FDE X column property, the column property is used to specify the x values. If the columns do not contain the column property, the Rows as Functions data format uses the levels of the x values specified in the column names. If the column name is too complicated, an error message is shown to indicate which columns were problematic and the x values are specified as $x = 1, 2, \dots, n$, where n is the number of observations.

Columns as Functions Select for data tables where each column corresponds to the entire response function for one level of the Group variable. Each row corresponds to a level of a regressor variable.

Specify Column Roles

The launch window contains the following options for specifying the column roles.

Y, Response Specifies the Y variable.

X, Regressor (Available only for the Stacked Data Format and Columns as Functions tabs.) Specifies the X variable.

Group (Available only for the Stacked Data Format and Rows as Functions tabs.) Specifies a grouping variable. The fitted model has separate parameters for each level of the grouping variable. This enables you to compare fitted models and estimated parameters across the levels of the grouping variable.

Z, Supplementary (Available only for the Stacked Data Format and Rows as Functions tabs.) Specifies one or more supplementary variables. Supplementary variables are not used in any of the calculations in the Fit Curve platform and including them does not affect the results. Supplementary variables are variables you might want to use in future analyses of the results from Fit Curve. When you specify supplementary variables, they are included in the table that is created by the Make Parameter Table option. In JMP Pro, if you specify one or more supplementary variables and a grouping variable, a Curve DOE Analysis option is made available for the fitted models. See [“Curve DOE Analysis”](#).

Weight Specify a variable that contains the weights of the observations.

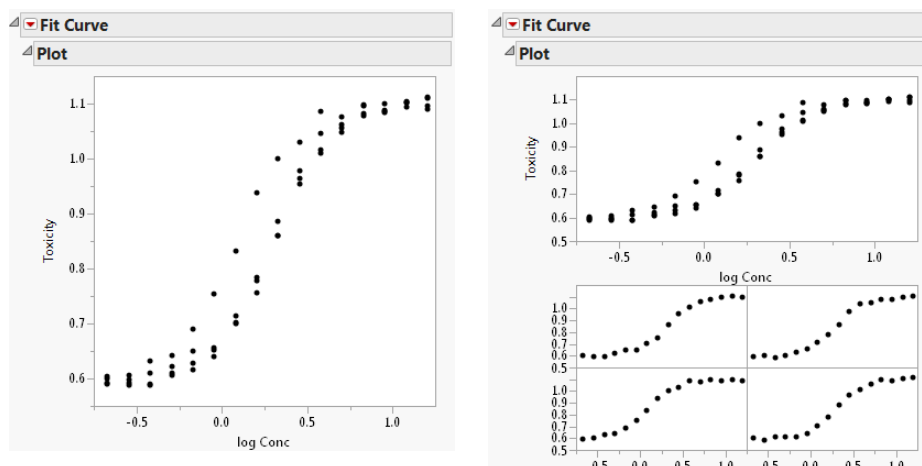
Freq (Available only for the Stacked Data Format tab.) Specifies a variable that contains the frequencies of the observations.

By Specifies a variable to perform a separate analysis for every level of the variable.

The Fit Curve Report

The Fit Curve report initially contains only a plot of Y versus X. If you specify a Group variable or use the Columns as Functions tab, the report includes overlaid and individual plots for each group of the fitted model.

Figure 14.7 Fit Curve Reports: No Grouping Variable (left) and with Group Variable (right)



After fitting a model, the fitted model appears on the plot (when no grouping variable is specified on the platform launch window). A Model Comparison report appears above the plots and a Model Report appears below the plot. Each time a new model is fit, it is added to the Model Comparison report and a new model report is added to the report window. See [“Model Comparison Report”](#) and [“Model Fit Report”](#).

Note: Any rows that are excluded in the data table are also hidden in the Fit Curve plot.

Model Comparison Report

The Model Comparison report in the Fit Curve platform shows fit statistics used for comparing multiple models.

Model Fit Report

A report is created for each individual model that is fit in the Fit Curve platform. Each model report contains the following sections.

Prediction Model Shows the algebraic form of the prediction formula and the parameters.

Summary of Fit Shows the same fit statistics as the Model Comparison report.

Group Summary (Appears only if there is a grouping variable.) Shows the group estimates of the parameters.

Parameter Estimates Shows the estimates of the parameters, standard errors, Chi-squared statistics, *p*-values, and confidence intervals. There are also separate tables for the correlations and covariances of the estimates.

Tip: You can sort the Parameter Estimates report table by clicking on the column that you want to sort by.

Plot Shows plots of the data with the fitted model ([Figure 14.7](#)). The plots are shown only when you select a Grouping variable on the platform launch window.

Fit Curve Options

The Fit Curve red triangle menu contains several categories of built-in models. Models that are not available for your particular data are grey menu options. For individual model details and data requirements, see [Table 14.1](#).

Polynomials Fits first degree to fifth degree polynomials, as well as a Power model.

Sigmoid Curves Fits Logistic, Probit, Gompertz, and Weibull models. These models are S-shaped and have both upper and lower asymptotes. The Logistic 2P, 3P, and 4P and Probit 2P and 4P models are symmetric. The Logistic 5P, Probit 3P, and both Gompertz models are not symmetric. The Logistic 2P and Probit 2P are available only when the response is between 0 and 1. The Weibull Growth is available only when both the response values and regressor values are nonnegative. Examples of Sigmoid curves include learning curves and modeling tumor growth, both of which increase initially and then taper off.

Exponential Growth and Decay Fits Exponential, Biexponential, Mechanistic Growth, Hybrid Exponential, and Cell Growth models. The Exponential 2P and 3P are similar, but the 3P model has an asymptote. The Biexponential models assume there are two separate growth or decay processes. The Mechanistic Growth and Exponential 3P models always increase (or decrease), but the rate of growth (or decay) slows so that the model has an

asymptote. Examples of exponential growth and decay functions are virus spread and drug half-life, respectively.

Peak Models Fits Gaussian Peak, ExGaussian Peak (Exponentially Modified Gaussian Peak), and Lorentzian Peak models. These models increase up to a peak and then decrease. The Gaussian Peak model is a scaled version of the Gaussian probability density function (PDF). The ExGaussian Peak model is similar to the Gaussian Peak model, but it can take on a skewed shape. The Lorentzian Peak model is a scaled version of the Cauchy distribution, a continuous probability distribution. These models can be used for some chemical concentration assays and artificial neural networks.

Pharmacokinetic Models Fits the One Compartment Oral Dose model, the Two Compartment IV Bolus Dose model, and the Biexponential 4P model. This option is used to model the concentration of drugs in the body.

Rate Equations Fits the Michaelis-Menten and Inverse Michaelis-Menten models, as well as several first order and second order rate models. The Michaelis-Menten model is a biochemical kinetics model, which relates the rate of enzymatic reactions to substrate concentration. The first order and second order rate models are useful in modeling chemical reactions.

Fit Antoine Equation Fits the Antoine model, which is often used to model vapor pressure as a function of temperature. The Antoine model has both horizontal and vertical asymptotes.

Dissolution Curve Analysis (Available only if both the response values and the regressor values are nonnegative.) Contains a submenu of options for dissolution curve analysis. A dissolution curve measures the rate at which a tablet dissolves over time. Dissolution curve analysis compares the dissolution curves of a new tablet to the dissolution curves of a standard or reference tablet. A separate analysis is done for each new tablet. The Dissolution Curve Analysis menu provides both parametric and nonparametric techniques for comparing dissolution curves. If you use one of the built-in dissolution curve models, a standard model fit report is created for each individual model that is fit.

Model-Free Comparisons Provides nonparametric techniques to compare dissolution curves. If you use one of the nonparametric techniques from the Model-Free Comparisons submenu, a method specific dissolution curve report is created. See [“Model Comparison Report”](#).

Note: All of the model-free comparison techniques require that the curves contain no missing values.

F1 Analysis Performs an analysis using the F1 difference factor, which measures the percent difference between the curves of the reference tablet and the curves of the test tablet at each time point. See [“Statistical Details for Fit Curve Models”](#). Confidence intervals for F1 are computed using a bootstrapping procedure. The bootstrapping procedure uses the bias-corrected and accelerated (BCa) percentile method. See Efron (1981).

When you select the F1 Analysis option, a Dissolution Curve Specification window appears. Use this window to specify the reference level, the alpha level used to calculate the confidence intervals, the number of bootstrap samples, and a random seed. If the data are in the Stacked Data Format, you must also specify the curve ID column, which identifies the individual curves within the reference group and the test groups.

F2 Analysis Performs an analysis using the F2 similarity factor, which measures the similarity of percent dissolution between the curves of the reference tablet and the curves of the test tablet. See [“Statistical Details for Fit Curve Models”](#). Confidence intervals for F2 are computed using a bootstrapping procedure. The bootstrapping procedure uses the bias-corrected and accelerated (BCa) percentile method. See Efron (1981). If the lower limit for F2 is greater than 50, then the conclusion is that the dissolution of the test tablet is similar to the dissolution of the reference tablet (Paixão et al. 2017).

When you select the F2 Analysis option, a Dissolution Curve Specification window appears. Use this window to specify the reference level, the alpha level used to calculate the confidence intervals, the number of bootstrap samples, and a random seed. If the data are in the Stacked Data Format, you must also specify the curve ID column, which identifies the individual curves within the reference group and the test groups.

Multivariate Distance Performs an analysis using the Mahalanobis distance M , which measures the multivariate distance between the curves of the reference tablet and the curves of the test tablet. See [“Statistical Details for Fit Curve Models”](#). Confidence intervals are calculated for M using a multivariate normal assumption. If the upper limit for M is less than the maximum difference, then the conclusion is that the dissolution of the test tablet is similar to the dissolution of the reference tablet (Paixão et al. 2017).

When you select the Multivariate Distance option, a Dissolution Curve Specification window appears. Use this window to specify the reference level and the alpha level used to calculate the confidence intervals. If the data are in the Stacked Data Format, you must also specify the curve ID column, which identifies the individual curves within the reference group and the test groups.

Higuchi Curves Fits Higuchi models, including models with a time lag component and models with a burst component.

Hixson-Crowell Curves Fits Hixson-Crowell models, including models with a time lag component.

Korsmeyer-Peppas Curves Fits Korsmeyer-Peppas models, including models with a time lag component and models with a burst component.

Sigmoid Curves Fits Logistic, Probit, and Weibull models. See [“Sigmoid Curves”](#).

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

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

Model-Free Comparisons Reports

In the Fit Curve platform, a model-free comparison report is created when one of the three model-free comparison techniques is used: F1 Analysis, F2 Analysis, or Multivariate Distance. The report title corresponds to the name of the method. Each model-free comparison report contains a tabbed report of curve comparisons. If there are multiple test groups, each model-free comparison report contains a summary table. There is also a note at the top of each report that indicates which level of the grouping variable was used as the reference group.

Summary Table

The summary table has a row for each test group that is being compared to the reference group. There are columns for the observed factor value and the lower and upper confidence limits for the corresponding method (F1 Analysis, F2 Analysis, or Multivariate Distance). For the F2 Analysis and Multivariate Distance options, the summary table also contains a conclusion column that indicates whether the test curve for that row is similar or not similar to the reference curve.

Comparisons

The Comparisons section contains a tab for each test group that is being compared to the reference group. Each tab contains a plot, table of values, and a table of results for the specified technique. If the F1 Analysis or F2 Analysis technique is used, a plot of bootstrap results is also shown.

The plot shows the two mean dissolution curves, labeled as Reference and Test. There is also a table of summary statistics that provide the mean, standard deviation, and coefficient of variation (CV) for each time point for both the reference group and the test group. If a value is highlighted yellow in the table of summary statistics, one of the following conditions is true:

- The test group has more than one time point with a mean dissolution greater than 85%.
- The first time point in the test group has a CV value that is greater than 20%.
- There is a time point after the first time point in the test group that has a CV value that is greater than 10%.

If the F1 Analysis or F2 Analysis technique is used, the table of results are from the bootstrap procedure. The table includes the observed F1 or F2 value, the number of bootstrap samples, the bootstrap mean, the bootstrap median, the lower and upper BCa confidence limits, and the skewness and kurtosis. There is also a conclusion section for the F2 Analysis. The Bootstrap Results section contains the value of F1 or F2 for each bootstrap sample and a normal quantile plot of the bootstrap samples.

If the Multivariate Distance technique is used, the table includes the observed Mahalanobis distance value, the calculated lower and upper confidence limits, and maximum difference. The maximum difference is what the Mahalanobis distance would be if the difference between the reference tablet and test tablet was 10% at every time point. See [“Statistical Details for Dissolution Curve Analysis”](#). If the upper limit for *M* is less than the maximum difference, then the conclusion is that the dissolution of the test tablet is similar to the dissolution of the reference tablet.

Model-Free Comparison Analysis Options

The red triangle menu for the F1 Analysis, F2 Analysis, and Multivariate Distance reports contain the following options:

Curve DOE Analysis (Available only if a grouping variable and at least one supplementary variable are specified in the launch window. Not available if the Columns as Functions format is used.) Launches a Generalized Regression report within the model-free comparison report to model the corresponding factor (F1, F2, or Multivariate Distance) as a function of the supplementary variables. See [“Curve DOE Analysis”](#).

Remove Fit Removes the model-free comparison report from the Fit Curve report.

Model Fit Options

Each model report in the Fit Curve platform contains a red triangle menu with some or all of the following options:

Test Parallelism Helps determine whether the curves are similar in shape when they are shifted along the X axis. In certain situations, it is important to establish parallelism before making further comparisons between groups. This option is available only when a Group variable is specified on the platform launch window. This option is available for the Sigmoid models, as well as the Linear Regression model, with the exception of higher-order polynomials. See [“Test Parallelism”](#).

Area Under Curve Gives the area under the fitted curve. This option is available for the following models: One Compartment, Two Compartment, Gaussian Peak, and Lorentzian Peak. This option is also available for Bi-Exponential 4P Models, but only when all parameters are positive. The range of integration depends on the type of model and is specified in the report.

If a Grouping variable is specified on the platform launch window, an Analysis of Means is performed for comparing the estimates across groups. If the result for a group exceeds a decision limit, the result is considered different from the overall mean of AUC.

Time to Peak Response Displays the estimate of the regressor, X, at the peak of the fitted curve. The standard error of the estimate is also shown. This option is available for Cell Growth 4P and One Compartment models.

Peak Response Displays the estimate of the response, Y, at the peak of the fitted curve. The standard error of the estimate is also shown. This option is available for Cell Growth 4P and One Compartment models.

Compare Parameter Estimates Gives an analysis for testing the equality of parameters across levels of the grouping variable. This option is available only when a Group variable is specified on the platform launch window. See [“Compare Parameter Estimates”](#).

Equivalence Test Gives an analysis for testing the equivalence of models across levels of the grouping variable. This option is available only when a Group variable is specified on the platform launch window. See [“Equivalence Test”](#).



Curve DOE Analysis (Available only if a grouping variable and at least one supplementary variable are specified in the launch window. Not available if the Columns as Functions tab is used.) Launches a Generalized Regression report within the Fit Curve platform. A generalized regression model is fit to each parameter of the nonlinear model using the supplementary variables as model effects. By default, a two degree factorial model is fit and the Estimation Method is Best Subset. If the number of terms in the model is greater than 20 or the number of functions is greater than 1000, the Estimation Method

automatically switches to Pruned Forward. If the data table is from a Definitive Screening Design, the Estimation Method automatically switches to Two Stage Forward Selection. Alternatively, you can specify a model script in the original data table that defines the desired model fit. The models are then combined to create a profiler of the response as a function of the regressor variable and the supplementary variables. You can then use the CDOE Profiler to explore how the supplementary variables affect the response.

The Curve DOE Analysis report contains the following red triangle menu options:

Generalized Regression for Model Parameters Shows or hides the Generalized Regression reports for each model parameter. For more information on Generalized Regression model reports, see *Fitting Linear Models*.

Diagnostic Plots Shows or hides actual by predicted and residual plots for the response variable.

CDOE Profiler Shows or hides the CDOE Profiler, which enables you to explore how the response changes based on the supplementary variables. For more information about the CDOE Profiler red triangle menu options, see *Profilers*.

Save Prediction Formula Saves the prediction formula for the response to a new column in the data table.

Make Parameter Table Saves the parameter estimates, standard errors, and t-ratios in a data table. This option is available only when a Group variable is specified on the platform launch window.

Plot Actual by Predicted Plots actual Y values on the vertical axis and predicted Y values on the horizontal axis.

Plot Residual by Predicted Plots the residuals on the vertical axis and the predicted Y values on the horizontal axis.

Profiler Shows or hides a profiler of the fitted prediction function. The derivatives are derivatives of the prediction function with respect to the X variable. For more information about profilers, see *Profilers*.

Save Formulas Contains options for saving a variety of formula columns in the data table. If you use either of the Rows as Functions or Columns as Functions tabs in the launch window, the Save Stacked Data option is the only save option available.

Save Prediction Formula Saves the prediction equation.

Save Std Error of Predicted Saves the standard error of the predicted values.

Save Parametric Prediction Formula Saves the prediction equation in parametric form. This is helpful if you want to use the fitted model in the custom Nonlinear platform.

Save Residual Formula Saves the residuals.

Save Studentized Residual Formula Saves the studentized residual formula, a standard residual that is divided by its estimated standard deviation.

Save First Derivative Saves the derivative of the prediction function with respect to the X variable.

Save Std Error of First Derivative Saves the equation of the standard error of the first derivative.

Save Inverse Prediction Formula Saves the equation for predicting X from Y .

Save Stacked Data Saves columns to a new data table. The data table contains the original data in the stacked format as well as a column for the predicted values of the response and a column for the residuals.

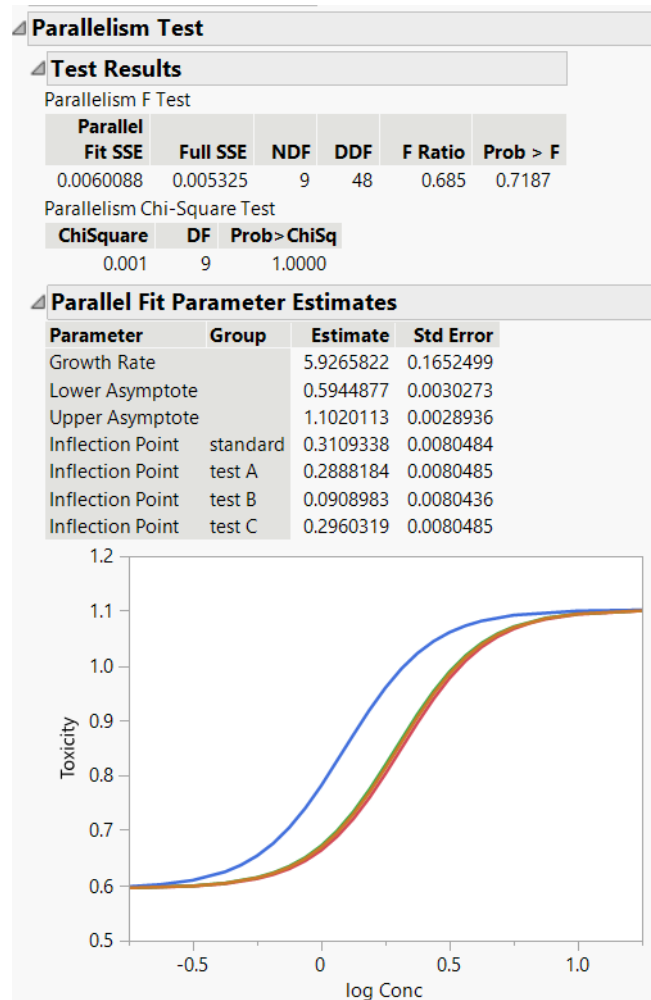
Custom Inverse Prediction Predicts an X value for a specific Y value. For more information about inverse prediction, see *Fitting Linear Models*.

Remove Fit Removes the model report, the entry from the Model Comparison report, and the fitted line from the plot.

Test Parallelism

The Test Parallelism option in the Fit Curve platform OR in the Model Fit red triangle menu of the Fit Curve platform provides an analysis for testing if the fitted models between groups have the same shape, but are shifted along the X axis ([Figure 14.9](#)). In the Bioassay example, the curve for drug formulation B is shifted to the left of the other three curves. However, you do not know whether the curves still have the same shape (are parallel), or if formulation B is different. The Parallelism Test tells us if the shapes for the different drug formulations have similar shapes and are shifted along the horizontal axis. Select **Test Parallelism** from the fitted model's red triangle menu to add the report.

Figure 14.9 Parallelism Test



The report gives the following results:

Test Results Gives the results of an F Test and a Chi-Square Test for parallelism. The tests compare the error sums-of-squares for a full model (Full SSE) and a reduced model (Fit SSE). The full model gives each group different parameters. The reduced model forces the groups to share every parameter except for the inflection point. A small p -value indicates that the group models are significantly different from one another. Therefore, the reduced model with shared parameters is insufficient and the full model should be used. In [Figure 14.9](#), the p -value is greater than 0.05, indicating that there is not enough evidence to conclude that differences exist between the curves. In this case, the reduced model with shared parameters is appropriate.

The fit statistics are calculated as follows:

F Ratio $[(\text{Fit SSE} - \text{Full SSE}) / \text{NDF}] / [\text{Full SSE} / \text{DDF}]$

ChiSquare $\text{Fit SSE} - \text{Full SSE}$

Parallel Fit Parameter Estimates Gives the parameter estimates under the reduced model (same parameters, except for inflection point). A plot of the fitted curves under the reduced model is provided. The inflection point for drug formulation B is much lower than that of the other three drug formulations.

Relative Potencies Gives the potency and relative potency for each level of the grouping variable. The potency is $10^{(\text{EC}_{50})}$, where EC_{50} is the concentration at which the response half way between baseline and maximum is obtained. For the Logistic 2P, 3P, and 4P, the potency is $10^{(\text{inflection point parameter})}$. The relative potency is the potency of one level of the grouping variable divided by the potency of another level. For example, in [Figure 14.10](#), the relative potency of test A to standard is $\text{Potency}_{\text{test A}} / \text{Potency}_{\text{standard}}$.

Figure 14.10 Relative Potencies by Group

Relative Potencies			
Relative Potency versus standard			
Group	Potency	Relative Potency	Std Error
standard	2.0461329	1	0
test A	1.9445466	1.0522416	0.0246535
test B	1.2328161	1.6597227	0.0388878
test C	1.977115	1.0349084	0.0242474
Relative Potency versus test A			
Group	Potency	Relative Potency	Std Error
standard	2.0461329	0.950352	0.0222663
test A	1.9445466	1	0
test B	1.2328161	1.5773209	0.036957
test C	1.977115	0.9835273	0.0230435
Relative Potency versus test B			
Group	Potency	Relative Potency	Std Error
standard	2.0461329	0.6025103	0.014117
test A	1.9445466	0.6339864	0.0148545
test B	1.2328161	1	0
test C	1.977115	0.6235429	0.0146098
Relative Potency versus test C			
Group	Potency	Relative Potency	Std Error
standard	2.0461329	0.9662691	0.0226392
test A	1.9445466	1.0167486	0.0238219
test B	1.2328161	1.6037388	0.037576
test C	1.977115	1	0

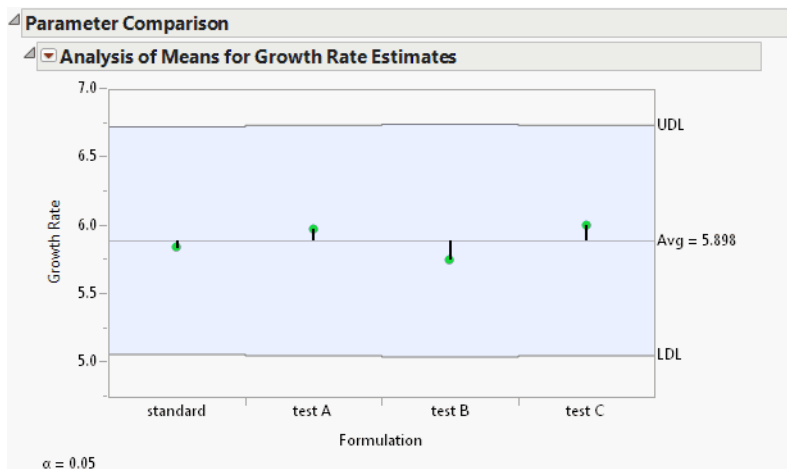
In the **Relative Potency versus standard** panel from [Figure 14.10](#), note that the relative potencies for drug formulations A and C are nearly one. This indicates that their potencies are similar to that of the standard formulation. The potency for drug formulation B is lower than that of the standard. This means that drug formulation B increases in toxicity as a function of concentration faster than the standard.

In the parallelism test, the curves are parallel, which enables you to calculate relative potencies. Based on the relative potencies, you conclude that formulation B is more potent than the other drug formulations. Taken with the prior findings, drug formulation B appears to be more toxic.

Compare Parameter Estimates

In the Fit Curve platform, the Compare Parameter Estimates option gives results for testing the equality of parameters across the levels of the grouping variable. There is an Analysis of Means (ANOM) report for each parameter, which tests whether the parameters are equal to an overall mean. If the result for a parameter exceeds the decision limits, then the parameter is different from the overall mean. [Figure 14.11](#) shows the ANOM report for growth rate estimates. Select **Compare Parameter Estimates** from the fitted model's red triangle menu to add the report.

Figure 14.11 Parameter Comparison for Growth Rate Estimates



The Analysis of Means red triangle menu contains the following options:

Set Alpha Level Sets the alpha level for the test.

Show Summary Report Shows or hides a report containing the parameter estimates, the decision limits, and whether the parameter exceeded the limits.

Display Options Contains options for showing or hiding decision limits, shading, and the center line. Also contains options for changing the appearance of the points.

For more information about the Analysis of Means report, see *Basic Analysis*.

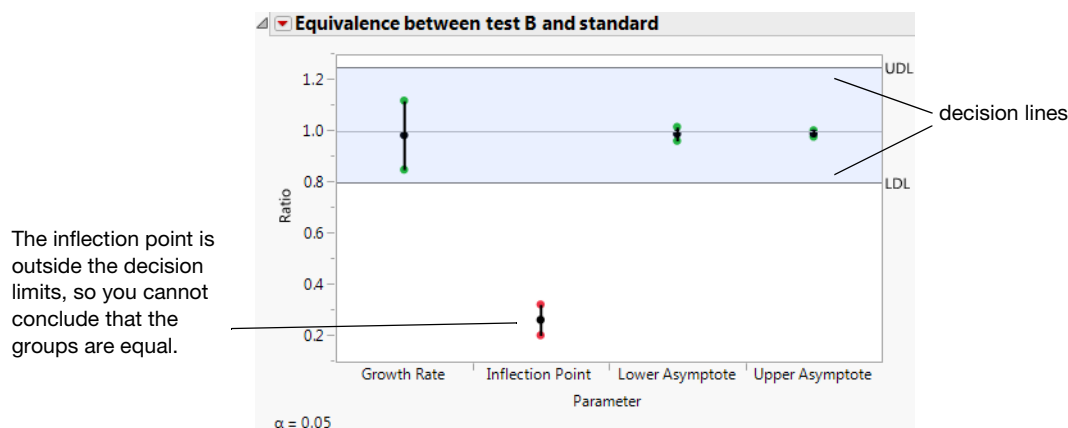
Equivalence Test

In the Fit Curve platform, the Equivalence Test option gives an analysis for testing the equivalence of models across levels of the grouping variable (Figure 14.12). After selecting the option, you specify the level of the grouping variable that you want to test against every other level. There is a report for every level versus the chosen level. Select **Equivalence Test** from the fitted model's red triangle menu to add the report.

The equality of the parameters is tested by analyzing the ratio of the parameters. The default decision lines are placed at ratio values of 0.8 and 1.25, representing a 25% difference.

If all of the confidence intervals are inside the decision lines, then the two groups are practically equal. If a single interval falls outside the lines (Figure 14.12), then you cannot conclude that the groups are equal. The inflection point for drug formulation B is lower than the standard, which agrees with the previous findings.

Figure 14.12 Equivalence Test



The Equivalence red triangle menu contains the following options:

Set Alpha Level Sets the alpha level for the test. The default value is 0.05.

Set Decision Lines Changes the decision limits for the ratio. The default values are set at 0.8 and 1.25, representing a 25% difference.

Show Summary Report Shows or hides a report containing the parameter estimates, the decision limits, and whether the parameter exceeded the limits.

Display Options Contains options for showing or hiding decision limits, shading, and the center line. Also contains options for changing the appearance of the points. For additional formatting options, right-click the graph and select **Customize**.

Additional Example of the Fit Curve Platform

Use the Dissolution Curve Analysis option to determine which factor settings create a tablet that is most similar to a reference tablet in terms of dissolution profiles. There are 16 groups of test tablets, with 6 individual tablets in each group. Each group of test tablets is made up of a different combination of four factors. The factor settings for the reference tablet are unknown. First, you compare the dissolution curves of the test tablets to the dissolution curve of the reference tablet. Then, you use the Curve DOE Analysis option to determine appropriate factor settings.

Dissolution Curve Analysis

1. Select **Help > Sample Data Folder** and open Dissolution DoE.jmp.
2. Select **Analyze > Specialized Modeling > Fit Curve**.
3. Click the Rows as Functions tab.
4. Select Dissolution60 through Dissolution360 and click **Y, Response**.
5. Select Batch and click **Group**.
6. Select Polymer A, Polymer B, Total Polymer, and Compression Force and click **Z, Supplementary**.
7. Click **OK**.
8. Click the Fit Curve red triangle and select **Dissolution Curve Analysis > Model-Free Comparisons > F2 Analysis**.
9. In the Dissolution Curve Specification window, specify R01 as the reference level of Batch.
10. (Optional) In the Dissolution Curve Specification window, type 1234 next to Random Seed.
11. Click **OK**.

Figure 14.13 F2 Analysis Report

F2 Analysis				
Reference: R01				
Group	F2	Lower 90% BCa Limit	Upper 90% BCa Limit	Conclusion
A01	38.98	37.067	41.073	Not Similar
A02	76.189	70.283	83.053	Similar
A03	40.263	38.118	42.75	Not Similar
A04	60.558	56.102	65.004	Similar
A05	37.199	35.593	38.613	Not Similar
A06	36.14	34.365	38.095	Not Similar
A07	70.32	64.452	77.655	Similar
A08	42.429	40.117	44.955	Not Similar
A09	42.943	39.548	46.749	Not Similar
A10	50.25	47.025	54.547	Not Similar
A11	33.246	31.524	35.227	Not Similar
A12	29.53	28.316	30.677	Not Similar
A13	84.864	79.669	93.739	Similar
A14	33.312	32.29	34.607	Not Similar
A15	37.005	35.454	38.685	Not Similar
A16	28.922	27.894	29.979	Not Similar

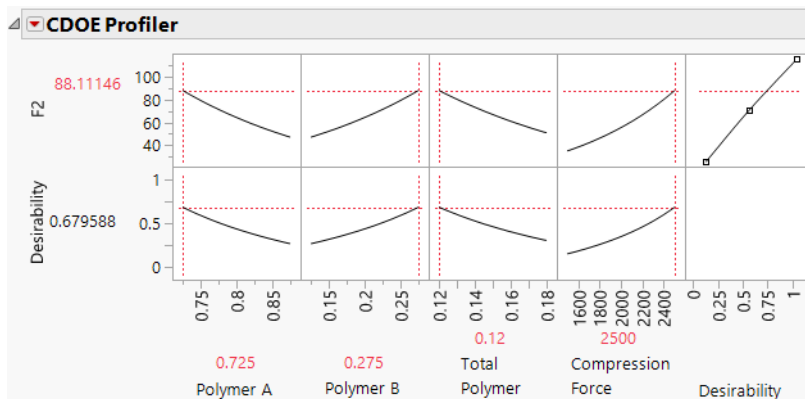
Based on the F2 similarity factor, four of the tablets (A02, A04, A07, and A13) have similar dissolution profiles to the reference tablet. For all of these tablets, the lower 90% confidence limit for F2 is greater than 50. You can select different tabs in the Comparisons section to view the individual comparisons between each test group and the reference group.

Curve DOE Analysis for F2

The F2 Analysis determines which tablets are similar to the reference tablet, but it does not determine why the tablets are similar. The Curve DOE Analysis option performs a Generalized Regression analysis that models F2 as a function of the four supplementary variables that were specified in the launch window.

1. Click the F2 Analysis red triangle and select **Curve DOE Analysis**.
Since a higher F2 value indicates a more similar tablet to the reference tablet, you can maximize F2 to determine the best factor settings.
2. Click the CDOE Profiler red triangle and select **Optimization and Desirability > Desirability Functions**.
3. Click the CDOE Profiler red triangle and select **Optimization and Desirability > Maximize Desirability**.

Figure 14.14 Curve DOE Analysis for F2



The following factor settings create a tablet that is most similar to the reference tablet in terms of dissolution profiles:

Polymer A at 0.725

Polymer B at 0.275

Total Polymer at 0.12

Compression Force at 2500

Statistical Details for the Fit Curve Platform

This section contains statistics details for the Fit Curve platform.

- [“Statistical Details for Fit Curve Models”](#)
- [“Statistical Details for Dissolution Curve Analysis”](#)

Statistical Details for Fit Curve Models

The formulas for the models that are available in the Fit Curve red triangle menu are summarized in Table 14.1.

Table 14.1 Fit Curve Model Formulas

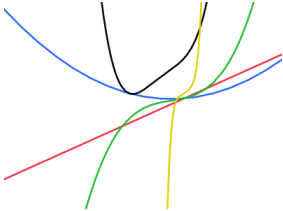
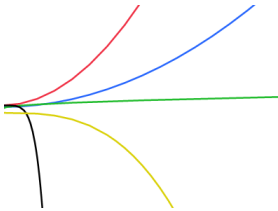
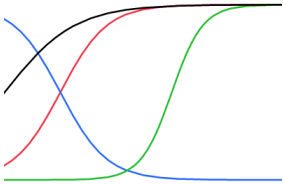
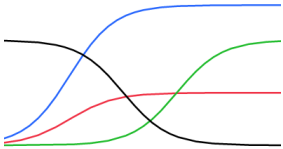
Model	Formula
Polynomials 	$\beta_0 + \sum_{i=1}^k \beta_i x^i$ where k is the order of the polynomial. These models can also be fit using the Fit Model and Fit Y by X platforms.
Power Model 	$a + bx^c$ a = Intercept b = Slope c = Power Available only when both the response values and the x values are nonnegative.
Logistic 2P 	$\frac{1}{1 + \exp(-a(x - b))}$ a = Growth Rate b = Inflection Point Available only when all response values are between zero and one.
Logistic 3P 	$\frac{c}{1 + \exp(-a(x - b))}$ a = Growth Rate b = Inflection Point c = Asymptote

Table 14.1 Fit Curve Model Formulas (*Continued*)

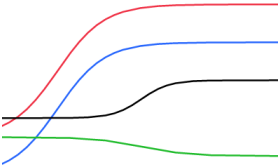
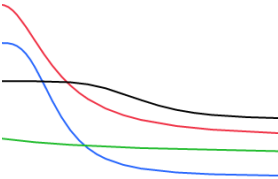
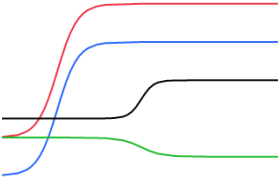
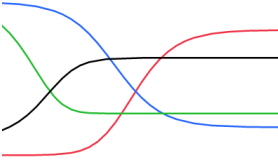
Model	Formula
Logistic 4P 	$c + \frac{d - c}{1 + \exp(-a(x - b))}$ a = Growth Rate b = Inflection Point c = Lower Asymptote d = Upper Asymptote
Logistic 4P Rodbard 	$c + \frac{d - c}{1 + (x/b)^a}$ a = Growth Rate b = Inflection Point c = Lower Asymptote d = Upper Asymptote Available only when the x values are positive.
Logistic 4P Hill 	$c + \frac{d - c}{1 + 10^{(-a(x - b))}}$ a = Growth Rate b = Inflection Point c = Lower Asymptote d = Upper Asymptote
Logistic 5P 	$c + \frac{d - c}{(1 + \exp(-a(x - b)))^f}$ a = Growth Rate b = Inflection Point c = Asymptote 1 d = Asymptote 2 f = Power

Table 14.1 Fit Curve Model Formulas (*Continued*)

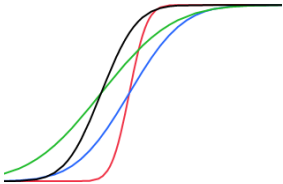
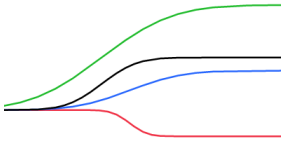
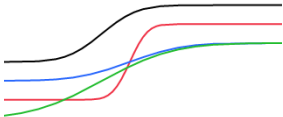
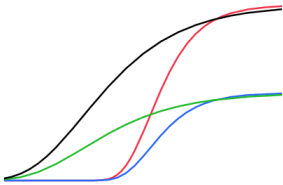
Model	Formula
Probit 2P 	$\Phi(a \cdot (x - b))$ a = Growth Rate b = Inflection Point Φ = Normal Distribution CDF Available only when all response values are between zero and one.
Probit 3P 	$c \cdot \Phi(a \cdot (x - b))$ a = Growth Rate b = Inflection Point c = Asymptote Φ = Normal Distribution CDF There is also an asymptote at zero.
Probit 4P 	$c + (d - c) \cdot \Phi(a \cdot (x - b))$ a = Growth Rate b = Inflection Point c = Lower Asymptote d = Upper Asymptote Φ = Normal Distribution CDF
Gompertz 3P 	$a \exp(-\exp(-b(x - c)))$ a = Asymptote b = Growth Rate c = Inflection Point

Table 14.1 Fit Curve Model Formulas (*Continued*)

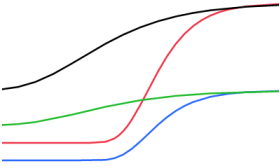
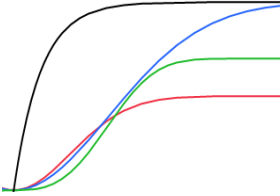
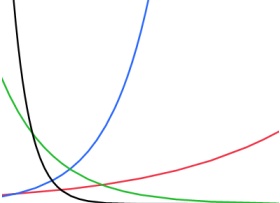
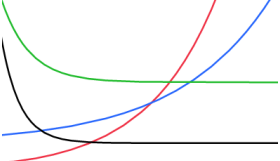
Model	Formula
Gompertz 4P 	$a + (b - a)\exp(-\exp(-c(x - d)))$ a = Lower Asymptote b = Upper Asymptote c = Growth Rate d = Inflection Point
Weibull Growth 	$a(1 - \exp(-(x/c)^b))$ a = Upper Asymptote b = Growth Rate c = Inflection Point Available only when both the response values and the x values are nonnegative.
Exponential 2P 	$a \exp(bx)$ a = Scale b = Growth Rate
Exponential 3P 	$a + b \exp(cx)$ a = Asymptote b = Scale c = Growth Rate

Table 14.1 Fit Curve Model Formulas (Continued)

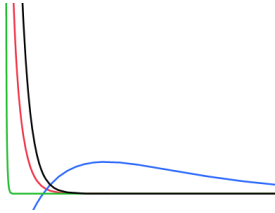
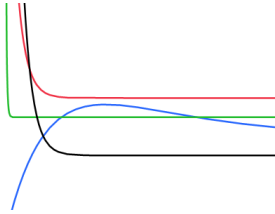
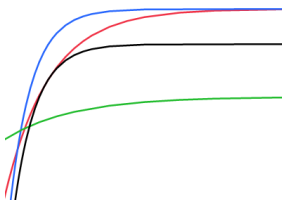
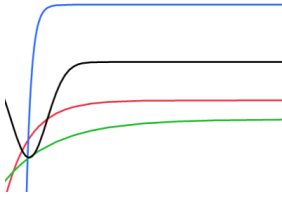
Model	Formula
Biexponential 4P 	$a \exp(-bx) + c \exp(-dx)$ a = Scale 1 b = Decay Rate 1 c = Scale 2 d = Decay Rate 2 Available only when the response values are positive.
Biexponential 5P 	$a + b \exp(-cx) + d \exp(-fx)$ a = Asymptote b = Scale 1 c = Decay Rate 1 d = Scale 2 f = Decay Rate 2
Mechanistic Growth 	$a(1 - b \exp(-cx))$ a = Asymptote b = Scale c = Growth Rate
Hybrid Exponential 	$ab - a(b - 1) \exp(-(cx)^d)$ a = Maximum b = Scale c = Rate d = Power Available only when both the response values and the x values are nonnegative.

Table 14.1 Fit Curve Model Formulas (*Continued*)

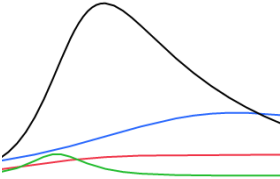
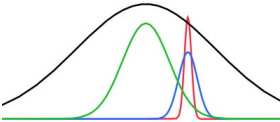
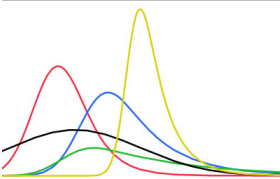
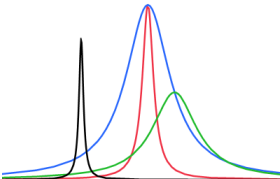
Model	Formula
Cell Growth 4P 	$\frac{ab}{(a-b)\exp(-cx) + b\exp(dx)}$ a = Peak value if mortality rate, d, is zero b = Response at time zero c = Cell Division Rate d = Cell Mortality Rate Available only when the response values are positive and the x values are nonnegative.
Gaussian Peak 	$a \exp\left(-\frac{1}{2}\left(\frac{x-b}{c}\right)^2\right)$ a = Peak Value b = Critical Point c = Growth Rate
ExGaussian Peak (Exponentially Modified Gaussian Peak) 	$ad \cdot \exp\left(0.5d(2b + dc^2 - 2x) + \log\left(\phi\left(\frac{x-b-dc^2}{c}\right)\right)\right)$ a = AUC (area under the curve) b = Location c = Scale d = Lambda ϕ = the standard normal pdf
Lorentzian Peak 	$\frac{ab^2}{(x-c)^2 + b^2}$ a = Peak Value b = Growth Rate c = Critical Point

Table 14.1 Fit Curve Model Formulas (*Continued*)

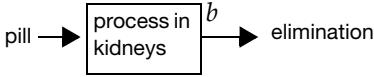
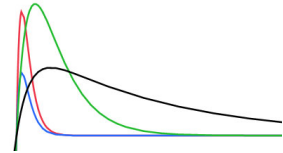
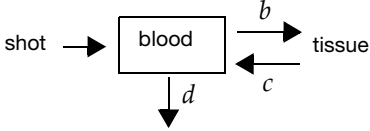
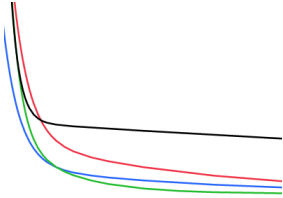
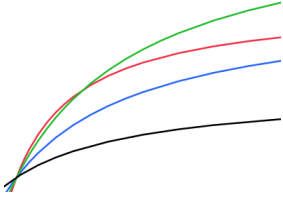
Model	Formula
One Compartment Oral Dose  	$\frac{abc}{c-b}(\exp(-bx) - \exp(-cx))$ $a = \text{Area Under Curve}$ $b = \text{Elimination Rate}$ $c = \text{Absorption Rate}$ Available only when the response values and the x values are all positive.
Two Compartment IV Bolus Dose  	$\frac{a}{\alpha - \beta}((\alpha - b)\exp(-\alpha x) - (\beta - b)\exp(-\beta x))$ $\alpha = \frac{1}{2}(b + c + d + \sqrt{(b + c + d)^2 - 4bd})$ $\beta = \frac{1}{2}(b + c + d - \sqrt{(b + c + d)^2 - 4bd})$ $a = \text{Initial Concentration}$ $b = \text{Transfer Rate In}$ $c = \text{Transfer Rate Out}$ $d = \text{Elimination Rate}$ Available only when the response values and the x values are all positive.
Michaelis-Menten 	$\frac{ax}{b+x}$ $a = \text{Maximum Velocity}$ $b = \text{Michaelis Constant}$ Available only when the response values and the x values are all positive.

Table 14.1 Fit Curve Model Formulas (*Continued*)

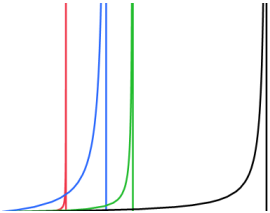
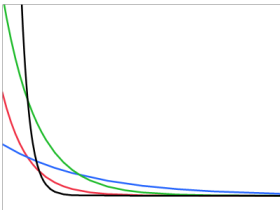
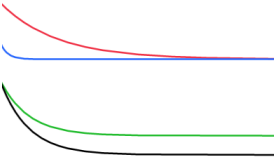
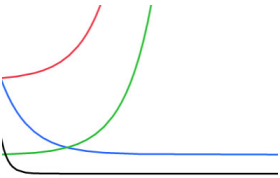
Model	Formula
Inverse Michaelis-Menten 	$\frac{bx}{a - x}$ a = Maximum Velocity b = Michaelis Constant Available only when the response values and the x values are all positive.
First Order Rate 	$a \exp(-bx)$ a = Initial Value b = Rate Constant Available only if the x values are nonnegative.
First Order with Limits 	$b + (a - b) \exp(-cx)$ a = Initial Value b = Limiting Value c = Rate Constant Available only if the x values are nonnegative.
First Order with Equilibrium 	$\frac{a(c + b \exp(-(b - c)x))}{(b + c)}$ a = Initial Value b = Forward Rate c = Backward Rate Available only if the x values are nonnegative.

Table 14.1 Fit Curve Model Formulas (*Continued*)

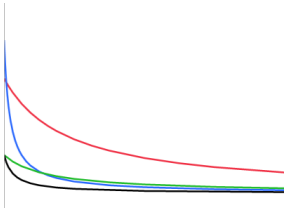
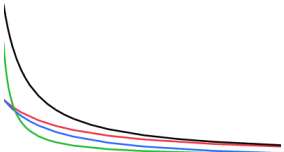
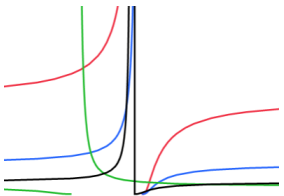
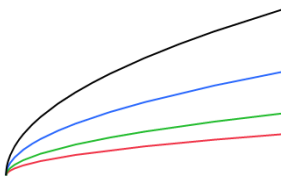
Model	Formula
Second Order 	$\frac{1}{(bx + 1/a)}$ a = Initial Value b = Rate Constant Available only if the x values are nonnegative.
Second Order with Two Components 	$\frac{ba}{(b + a)\exp(bcx) - a}$ a = Initial value b = Offset c = Rate Constant Available only if all x values are nonnegative.
Antoine Equation 	$10^{a - \frac{b}{c + x}}$ a = Component Specific Constant b = Component Specific Constant c = Component Specific Constant Available only when the response values are greater than 0. The vertical asymptote is at 10^a. The horizontal asymptote is at $-c$.
Higuchi 	$a\sqrt{x}$ a = Release Constant Available only if all x and y values are nonnegative.

Table 14.1 Fit Curve Model Formulas (*Continued*)

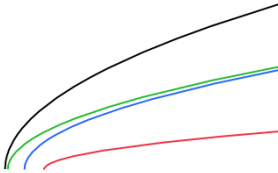
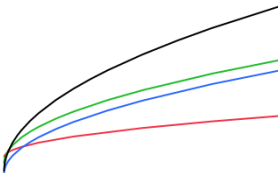
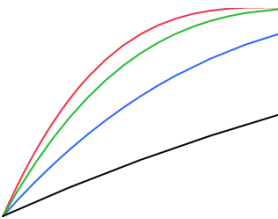
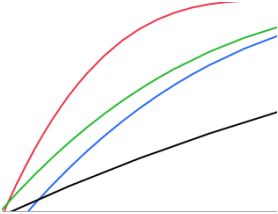
Model	Formula
 Higuchi with Lag	$a\sqrt{x-b}$ a = Release Constant b = Time Lag Available only if all x and y values are nonnegative.
 Higuchi with Burst	$b + a\sqrt{x}$ a = Release Constant b = Burst Available only if all x and y values are nonnegative.
 Hixson-Crowell	$100(1 - (1 - ax)^3)$ a = Release Constant Available only if all x and y values are nonnegative.
 Hixson-Crowell with Lag	$100(1 - (1 - a(x - b))^3)$ a = Release Constant Available only if all x and y values are nonnegative.

Table 14.1 Fit Curve Model Formulas (*Continued*)

Model	Formula
Korsmeyer-Peppas	ax^b a = Release Constant b = Diffusional Exponent Available only if all x and y values are nonnegative.
Korsmeyer-Peppas with Lag	$a(x - c)^b$ a = Release Constant b = Diffusional Exponent c = Time Lag Available only if all x and y values are nonnegative.
Korsmeyer-Peppas with Burst	$c + ax^b$ a = Release Constant b = Diffusional Exponent c = Burst Available only if all x and y values are nonnegative.

Statistical Details for Dissolution Curve Analysis

This section contains the formulas to calculate the factors that are used in model-free dissolution analyses in the Fit Curve platform. The formulas use the following notation:

- μ_{Ri} is the mean of the dissolution values for the batch of reference curves at time point t_i
- μ_{Ti} is the mean of the dissolution values for the batch of test curves at time point t_i
- $i = 1, \dots, N$ is the number of time points

F1 Difference Factor

The F1 difference factor measures the percent difference between the curves of the reference tablet and the curves of the test tablet at each time point. F1 is calculated using the following formula:

$$F1 = 100 \times \left(\frac{\sum_{i=1}^N |(\mu_{Ri} - \mu_{Ti})|}{\sum_{i=1}^N \mu_{Ri}} \right)$$

F2 Similarity Factor

The F2 similarity factor measures the similarity of percent dissolution between the curves of the reference tablet and the curves of the test tablet. F2 is calculated using the following formula (Paixão et al. 2017):

$$F2 = 50 \log \left(\frac{100}{\sqrt{1 + \frac{D^2}{p}}} \right)_{10}$$

where:

$$D^2 = \sum_{i=1}^N (\mu_{Ri} - \mu_{Ti})^2$$

p is the number of time points

Mahalanobis Distance

The Mahalanobis distance M , measures the multivariate distance between the curves of the reference tablet and the curves of the test tablet. M is calculated using the following formula (Paixão et al. 2017):

$$M = \sqrt{(\mathbf{R} - \mathbf{T})^T (S_{pooled})^{-1} (\mathbf{R} - \mathbf{T})}$$

where:

$$\mathbf{R} = [\mu_{R1}, \dots, \mu_{RN}]^T$$

$$\mathbf{T} = [\mu_{T1}, \dots, \mu_{TN}]^T$$

$$S_{pooled} = (\mathbf{SR} + \mathbf{ST})/2$$

SR = covariance matrix of the reference curve values

ST = covariance matrix of the test curve values

The 10% maximum difference is calculated using the following formula:

$$MaxDiff\ 10\% = \sqrt{m^T (S_{pooled})^{-1} m}$$

where m is an $N \times 1$ vector of 10's.

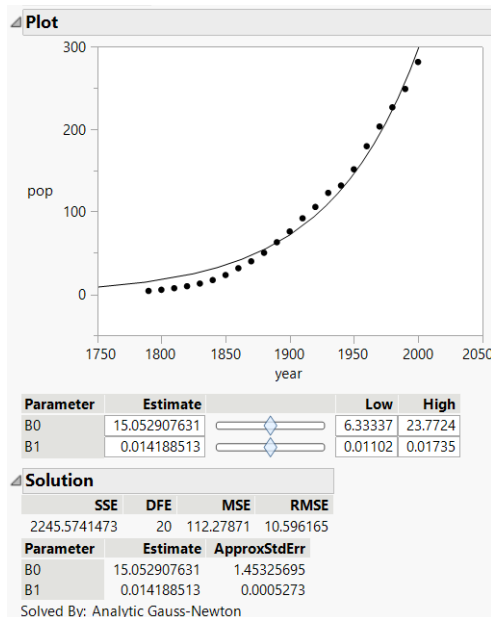
Chapter 15

Nonlinear Regression

Fit Custom Nonlinear Models to Your Data

The Nonlinear platform is a good choice for models that are *nonlinear* in the parameters. This chapter focuses on custom nonlinear models, which include a model formula and parameters to be estimated. Use the default least squares loss function or a custom loss function to fit models. The platform minimizes the sum of the loss function across the observations.

Figure 15.1 Example of a Custom Nonlinear Fit



The Nonlinear platform also provides predefined models, such as polynomial, logistic, Gompertz, exponential, peak, and pharmacokinetic models. See [“Fit Curve”](#).

Note: Some models are *linear* in the parameters (for example, a quadratic or other polynomial) or can be transformed to be such (for example, when you use a log transformation of x). The Fit Model or Fit Y by X platforms are more appropriate in these situations. For more information about these platforms, see *Fitting Linear Models* and *Basic Analysis*.

Contents

Example of the Nonlinear Platform	261
Launch the Nonlinear Platform	264
The Nonlinear Fit Report	265
Nonlinear Platform Options	269
Create a Formula Using the Model Library	273
Customize the Nonlinear Model Library	275
Additional Examples of the Nonlinear Platform	277
Example of Maximum Likelihood	277
Example of a Probit Model with Binomial Errors	280
Example of a Poisson Loss Function	281
Example of Setting Parameter Limits	283
Example of Analyzing Left-Censored Data	286
Example of Fitting a Weibull Loss Function	287
Example of Fitting Simple Survival Distributions	290
Statistical Details for the Nonlinear Platform	293
Statistical Details for Profile Likelihood Confidence Limits	293
Statistical Details for Custom Loss Functions	294
Statistical Details on Derivatives	295
Statistical Details on Effective Nonlinear Modeling	297

Example of the Nonlinear Platform

This example fits a custom model by creating a formula column for the model and then fitting the model using the Nonlinear platform. This method requires a few more steps than fitting a built-in model, but it allows any nonlinear model to be fit. You can also provide a custom loss function, and specify several other options for the fitting process.

Tip: If the custom model is one that you use often for several data tables, you can add the model to the Model Library instead. See [“Customize the Nonlinear Model Library”](#).

Create a Formula Column

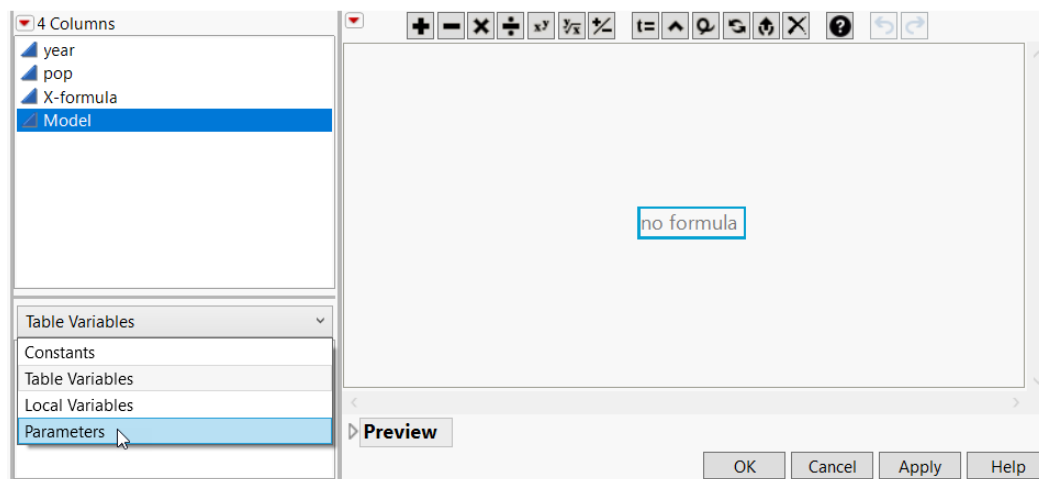
To fit a custom model, you must first create a model column with initial parameter estimates.

1. Select **Help > Sample Data Folder** and open Nonlinear Examples/US Population.jmp.
2. Create a new column called Model.
3. Right-click the Model column and select **Column Properties > Formula**.

The Formula Editor appears.

4. Select **Parameters** from the list below the list of columns.

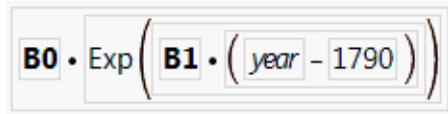
Figure 15.2 Select Parameters



5. Select **New Parameter**.
6. Type B0 for Name.
7. Type 3.9 for Value. This is the initial estimate of the parameter.
8. Click **OK**.

9. Select **New Parameter**.
10. Type B1 for Name and enter 0.022 for Value.
11. Click **OK**.
12. Enter the model formula using the Formula Editor functions, the column *year*, and the parameters.

Figure 15.3 Completed Model Formula



The screenshot shows the completed model formula in a software interface. The formula is displayed as $B0 \cdot \exp(B1 \cdot (year - 1790))$. The formula is enclosed in a light gray border, and the components are nested within boxes, indicating the use of the Formula Editor functions.

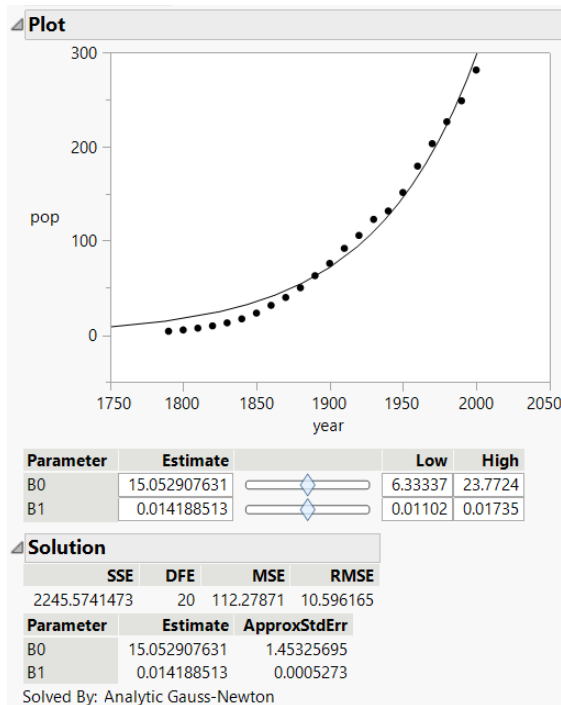
Tip: Click the gray triangle next to Transcendental to find the Exp command.

13. Click **OK**.

Fit a Nonlinear Model

1. Select **Analyze > Specialized Modeling > Nonlinear**.
2. Assign Model to the **X, Predictor Formula** role.
3. Assign pop to the **Y, Response** role.
4. Click **OK**.
5. Click **Go** on the Control Panel to fit the model.

Figure 15.4 Plot and Solution Report



The final parameter estimates are shown in the **Solution** report, along with other fit statistics. The fitted model is shown on the plot.

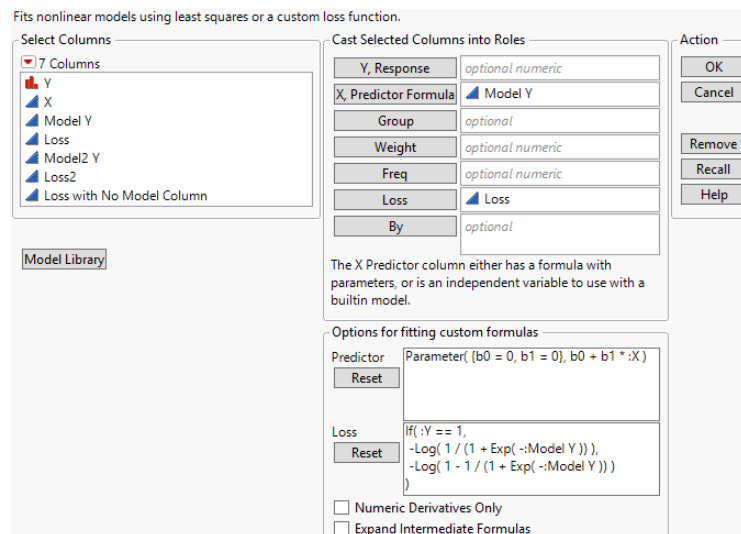
Parameters for Models with a Grouping Variable

In the formula editor, when you add a parameter, note the check box for **Expand Into Categories, selecting column**. This option is used to add several parameters (one for each level of a categorical variable for example) at once. When you select this option, a dialog appears that enables you to select a column. After selection, a new parameter appears in the Parameters list with the name *D_column*, where D is the name that you gave the parameter. When you use this parameter in the formula, a Match expression is inserted, containing a separate parameter for each level of the grouping variable.

Launch the Nonlinear Platform

To launch the Nonlinear platform, select **Analyze > Specialized Modeling > Nonlinear**.

Figure 15.5 Nonlinear Platform Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Nonlinear platform launch window has the following features:

Y, Response Select the Y variable.

X, Predictor Formula Select either the X variable or a column containing the model formula with parameters.

Note: If you select a column that does not contain a model formula with parameters and you also do not specify a custom formula, the Fit Curve platform is launched instead.

Note: If you select a formula column that contains a parameter list item that does not have an assignment, JMP automatically sets that parameter to zero and displays a message in the log.

Group Specify a grouping variable. The fitted model has separate parameters for each level of the grouping variable. This enables you to compare fitted models and estimated parameters across the levels of the grouping variable.

Note: If the formula does not have separate parameters for each group level, the Group variable is ignored by JMP.

Weight Specify a variable containing the weights of observations.

Freq Specify a variable representing the frequency of an observation.

Loss Specify a formula column giving a loss function.

By Specify a variable to perform a separate analysis for every level of the variable.

Model Library Launches the Model Library tool, which helps you choose initial values to create a formula column. See [“Create a Formula Using the Model Library”](#).

Options for fitting custom formulas If you specify a column in the X, Predictor Formula role that does not contain a model formula with parameters, you can use this option to create a custom formula. In the Predictor field, specify parameters, parameter values, and a custom formula that uses the selected X variable. Then, click the **Reset** button and launch the platform.

Numeric Derivatives Only Uses numeric derivatives only. This option is useful when you have a model for which it is too messy to take analytic derivatives. It can also be valuable in obtaining convergence in tough cases. This option is used only when a formula column is provided in the X, Predictor Formula role.

Expand Intermediate Formulas Tells JMP that if an ingredient column to the model is a column that itself has a formula, to substitute the inner formula, as long as it refers to other columns. To prevent an ingredient column from expanding, use the **Other** column property with a name of “Expand Formula” and a value of 0. This option is used only when a formula column is provided in the X, Predictor Formula role.

The Nonlinear Fit Report

The initial Nonlinear Fit report includes the Control Panel and Plot, shown in [Figure 15.6](#).

Control Panel Provides options for controlling the fitting process.

Go Starts the fitting process.

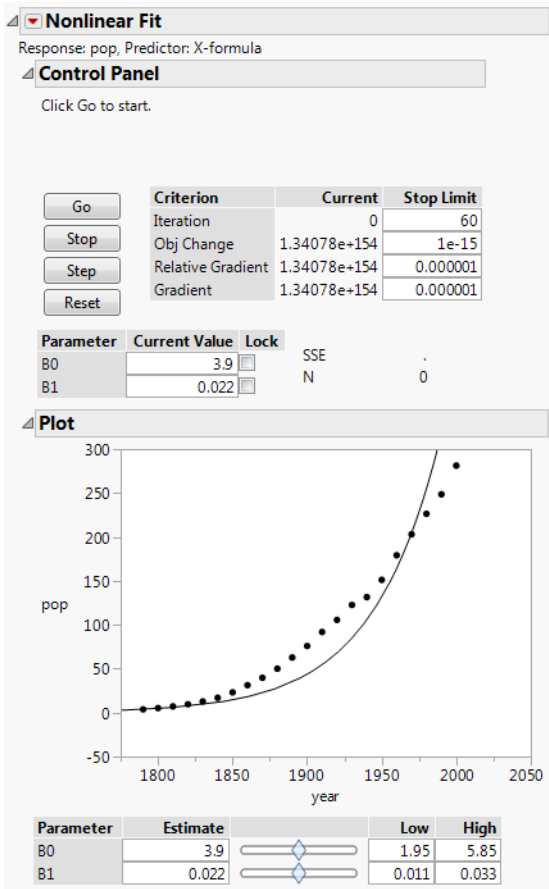
Stop Stops the fitting process.

Step Proceeds through the fitting process one iteration at a time.

Reset Resets the editable values into the formula, resets the iteration values, and calculates the SSE at these new values.

- Criterion** Shows iteration measures from the fitting process.
- Current** Shows the current value of each Criterion.
- Stop Limit** Sets limits on the measures listed under Criterion.
- Plot** Shows a plot of the X and Y variables for models with only one X variable. The model based on the current values is shown on the plot. To change the current values of the parameters, use the sliders or edit boxes beneath the plot.

Figure 15.6 Initial Nonlinear Fit Report



After you click **Go** to fit a model, the report includes the following additional items, shown in [Figure 15.7](#).

- Save Estimates** Saves the current parameter values to the parameters in the formula column.

Confidence Limits Computes confidence intervals for all parameters. The intervals are profile likelihood confidence intervals, and are shown in the Solution report. The confidence limit computations involve a new set of iterations for each limit of each parameter, and the iterations often do not find the limits successfully. The **Edit Alpha** and **Convergence Criterion** options are for the confidence interval computations. For more information about the **Goal SSE for CL**, see [“Statistical Details for Profile Likelihood Confidence Limits”](#).

Solution Shows the parameters estimates and other statistics.

SSE Shows the residual sum of squared errors. SSE is the objective that is to be minimized. If a custom loss function is specified, this is the sum of the loss function.

DFE Shows the degrees of freedom for error, which is the number of observations used minus the number of parameters fitted.

MSE Shows the mean squared error. It is the estimate of the variance of the residual error, which is the SSE divided by the DFE.

RMSE Estimates the standard deviation of the residual error, which is square root of the MSE.

RSquare (Available only if the model has an intercept and the sum of the residuals is close to zero.) Shows the RSquare value.

Parameter Lists the names that you gave the parameters in the fitting formula.

Estimate Lists the parameter estimates produced. Keep in mind that with nonlinear regression, there might be problems with this estimate even if everything seems to work.

ApproxStdErr Lists the approximate standard error, which is computed analogously to linear regression. It is formed by the product of the RMSE and the square root of the diagonals of the derivative cross-products matrix inverse.

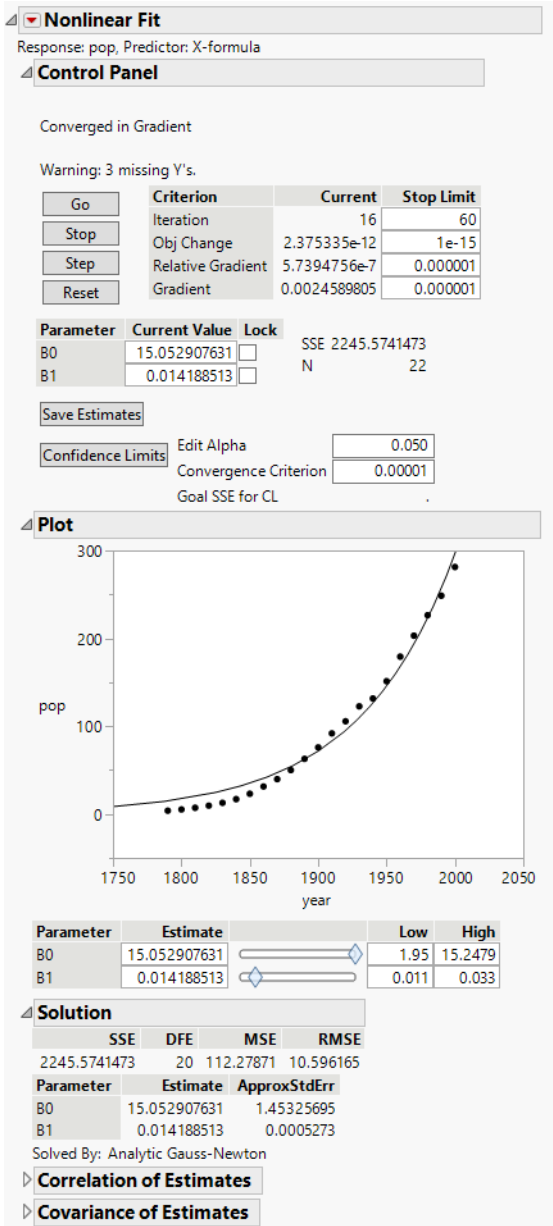
Note: If the number of observations is less than or equal to the number of parameters in the fitted model, ApproxStdErr is not reported.

Lower CL and Upper CL Shows the confidence limits for the parameters. They are missing until you click the **Confidence Limits** on the Control Panel. For more information about the confidence intervals, see [“Statistical Details for Profile Likelihood Confidence Limits”](#).

Excluded Data Shows a report showing fit statistics for excluded rows. This is useful for validating the model on observations not used to fit the model. You can use this feature in conjunction with the Remember Solution option to change the exclusions, and get a new report reflecting the different exclusions

- Correlation of Estimates** Displays the correlations between the parameter estimates.
- Covariance of Estimates** Displays the covariances between the parameter estimates.

Figure 15.7 Fitted Model Report



Nonlinear Platform Options

The Nonlinear Fit red triangle menu contains the following options:

Parameter Bounds Sets bounds on the parameters. When the option is selected, editable boxes appear in the Control Panel. Unbounded parameters are signified by leaving the field blank.

Plot Shows or hides a plot of the X and Y variables for models with only one X variable. The model shown on the plot is based on the current values of the parameters. To change the current values of the parameters, use the sliders or edit boxes beneath the plot. If you specify a Group variable at launch, then a curve shows for each group.

Iteration Options Specifies options for the fitting algorithm.

Iteration Log Shows or hides the Iterations table. Once this option is selected, the platform records subsequent iterations in the table.

Numeric Derivatives Only Useful when you have a model that is too messy to take analytic derivatives for. It can also be valuable in obtaining convergence in tough cases.

Expand Intermediate Formulas Tells JMP that if an ingredient column to the formula is a column that itself has a formula, to substitute the inner formula, as long as it refers to other columns. To prevent an ingredient column from expanding, use the **Other** column property with a name of “Expand Formula” and a value of 0.

Newton Specifies whether Gauss-Newton (for regular least squares) or Newton-Raphson (for models with loss functions) is the optimization method.

QuasiNewton SR1 Specifies QuasiNewton SR1 as the optimization method.

QuasiNewton BFGS Specifies QuasiNewton BFGS as the optimization method.

Accept Current Estimates Tells JMP to produce the solution report with the current estimates, even if the estimates did not converge.

Show Derivatives Shows the derivatives of the nonlinear formula in the JMP log. See [“Statistical Details on Derivatives”](#), for technical information about derivatives.

Unthreaded Runs the iterations in the main computational thread. In most cases, JMP does the computations in a separate computational thread. This improves the responsiveness of JMP while doing other things during the nonlinear calculations. However, there are some isolated cases (models that have side effects that call display routines, for example) that should be run in the main thread, so this option should be turned on.

Profilers Provides various profilers for viewing response surfaces.

Profiler Shows the Prediction Profiler. The Profiler lets you view vertical slices of the surface across each x -variable in turn, as well as find optimal values of the factors.

Contour Profiler Shows the Contour Profiler. The Contour profiler lets you see two-dimensional contours as well as three dimensional mesh plots.

Surface Profiler Creates a three-dimensional surface plot. This option is available only for models with two or more X variables.

Parameter Profiler Shows the Prediction Profiler and profiles the SSE or loss as a function of the parameters.

Parameter Contour Profiler Shows the Contour Profiler and contours the SSE or loss as a function of the parameters.

Parameter Surface Profiler Creates a three-dimensional surface plot and profiles the SSE or loss as a function of the parameters. This option is available only for models with two or more parameters.

SSE Grid Create a grid of values around the solution estimates and compute the error sum of squares for each value. The solution estimates should have the minimum SSE. When the option is selected, the **Specify Grid for Output** report is shown with these features:

Parameter Lists the parameters in the model.

Min Displays the minimum parameter values used in the grid calculations. By default, Min is the solution estimate minus 2.5 times the `ApproxStdErr`.

Max Displays the maximum parameter value used in the grid calculations. By default, Max is the solution estimate plus 2.5 times the `ApproxStdErr`.

Number of Points Gives the number of points to create for each parameter. To calculate the total number of points in the new grid table, multiply all the Number of Points values. Initially Number of Points is 11 for the first two parameters and 3 for the rest. If you specify new values, use odd values to ensure that the grid table includes the solution estimates. Setting Number of Points to 0 for any parameter records only the solution estimate in the grid table.

When you click **Go**, JMP creates the grid of points in a new table. A highlighted row marks the solution estimate row if the solution is in the table.

Revert to Original Parameters Resets the platform to the original parameter values (the values given in the formula column parameters).

Remember Solution Creates a report called Remembered Models, which contains the current parameter estimates and summary statistics. Results of multiple models can be remembered and compared. This is useful if you want to compare models based on different parameter restrictions, or models fit using different options. Click the radio

button for a particular model to display that model in the Plot and the parameter estimates in the Control Panel.

Custom Estimate Gives an estimate of a function of the parameters. You provide an expression involving only parameters. JMP calculates the expression using the current parameter estimates, and also calculates a standard error of the expression using a first-order Taylor series approximation.

Custom Estimation Profiler Enables you to construct a profiler for a custom expression. This type of profiler is useful for nonlinear models you have specified. You provide an expression involving parameters and at least one factor. There is also an option to apply a transformation to the expression. After you click OK, you specify the initial values for the factors. The profiler for the expression is based on the current parameter estimates across a range of the factors. The profiler is set at the specified initial values. You can add several custom profilers using the Custom Estimation Profiler option. Each custom profiler red triangle menu contains options for the profiler graph and factor settings. See *Profilers*. To remove a profiler, click the corresponding red triangle and select Remove Profiler. See [“Create Custom Profiler”](#).

Note: By default, the profiler axis ranges for the expression and the factors might be narrow. You can adjust the range by double clicking an axis and specifying new minimum and maximum values.

Custom Inverse Prediction Estimates the X value for a given Y value. It also calculates a standard error for the estimated X . JMP must be able to invert the model. The standard error is based on the first-order Taylor series approximation using the inverted expression. The confidence interval uses a t -quantile with the standard error, and is a Wald interval.

Show Prediction Expression Shows the prediction model or the loss function at the top of the report.

Save Pred Confid Limits Saves asymptotic confidence limits for the model prediction. This is the confidence interval for the average Y at a given X value.

Save Indiv Confid Limits Saves asymptotic confidence limits for an individual prediction. This is the confidence interval for an individual Y value at a given X value.

Save Formulas Gives options for saving model results to data table columns:

Save Prediction Formula Saves the prediction formula with the current parameter estimates.

Save Std Error of Predicted Saves the standard error for a model prediction. This is the standard error for predicting the average Y for a given X . The formula is of the form $\text{Sqrt}(\text{VecQuadratic}(\text{matrix1}, \text{vector1}))$. **matrix1** is the covariance matrix associated with the parameter estimates, and **vector1** is a composition of the partial derivatives of the model with respect to each parameter.

Save Std Error of Individual Saves the standard error for an individual prediction. This is the standard error for predicting an individual Y value for a given X value. The formula is of the form $\text{Sqrt}(\text{VecQuadratic}(\text{matrix1}, \text{vector1}) + \text{mse})$. **matrix1** is the covariance matrix associated with the parameter estimates, **vector1** is a composition of the partial derivatives of the model with respect to each parameter, and **mse** is the estimate of error variance.

Save Residual Formula Saves the formula for computing the residuals.

Save Pred Confid Limit Formula Saves the formula to calculate the confidence interval for a model prediction. This is a confidence interval for the average Y for a given X .

Save Indiv Confid Limit Formula Saves the formula to calculate the confidence interval for an individual prediction. This is a confidence interval for an individual Y for a given X .

Save Inverse Prediction Formula Saves formulas for the inverse of the model, the standard error of an inverse prediction, and the standard error of an individual inverse prediction.

Save Specific Solving Formula Equivalent to Save Inverse Prediction Formula in simple cases. However, this command allows the formula to be a function of several variables and allows expressions to be substituted. This feature works only for solving easily invertible operators and functions that occur just once in the formula.

After selecting this command, a dialog appears that enables you to select the variable to solve for. You can also edit the names of the columns in the resulting table. You can also substitute values for the names in the dialog. In these cases, the formula is solved for those values.

Note: The standard errors, confidence intervals, and hypothesis tests are correct only if least squares estimation is done, or if maximum likelihood estimation is used with a proper negative log-likelihood.

Save Estimates to Table Saves the parameter estimates to a new data table.

Create a Formula Using the Model Library

Use the Model Library in the Nonlinear platform to create a formula column with parameters and initial values. Click the **Model Library** button on the Nonlinear launch window to open the library. Select a model in the list to see its formula in the **Formula** box.

Figure 15.8 Nonlinear Model Library Dialog

Select a Model

- Logistic 2p
- Logistic 3p**
- Logistic 4p
- Logistic 5p (Richard Function)
- Bioassay 5PL Grouped
- Sigmoid Shape, polarograph, steady state voltammetry (3P)
- Weibull model (4P)
- CES Production Function (4P and 2X)
- Rodbard model (4P)
- Michaelis Menten Model (2P)
- First-order Decay Kinetics (3P)
- Gaussian Peak Shape (3P)
- Lorentzian Peak Shape (3P)
- Thermodynamic Linkage (3P)
- Thermodynamics: Pressure and temperature relationship (3P)
- Atmosphere: Wind speed profile (3P)
- Chemical reaction(3P and 2X)
- Enzyme Kinetics: competitive inhibition model(3P and 2X)
- Mechanistic Growth Model (3P)
- Model C (2P)
- Model D (2P and 2X)
- Model E (2P)
- Model F (Asymptotic Reg. 3P)
- Model H (Gompertz growth model, 3P)
- Model I (3P)
- Model J (LogLogistic growth model, 3P)
- Model K (3P)
- Model L (3P)
- Model M (4P and 3X)
- Model O (Morgan Mercer, 4P)
- Model P (Richards growth model, 4P)
- Model S (7P and 2X, Polynomial)

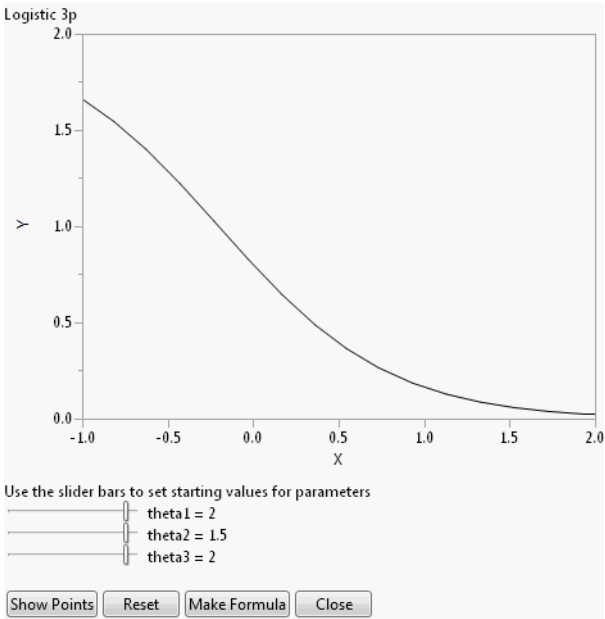
Formula

$$\frac{\text{theta1}}{1 + \text{theta2} * \text{Exp}(\text{theta3} * X)}$$

Show Graph Make Formula Close

Click **Show Graph** to show a 2-D theoretical curve for one-parameter models and a 3-D surface plot for two-parameter models. No graph is available for models with more than two explanatory (X) variables. On the graph window, change the default initial values of parameters using the slider, or clicking and entering values in directly.

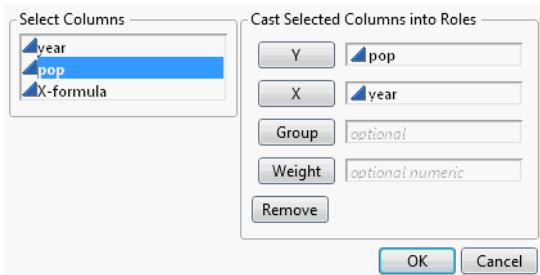
Figure 15.9 Example Graph in Model Library



The **Reset** button sets the initial values of parameters back to their default values.

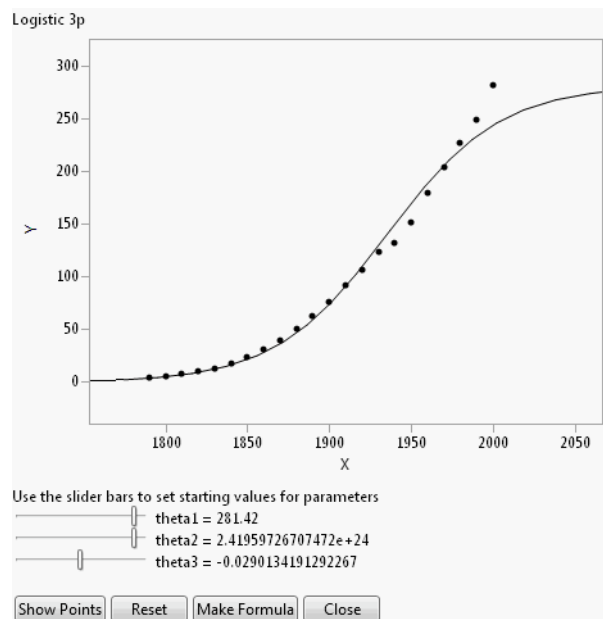
Click **Show Points** to overlay the actual data points to the plot. The dialog in [Figure 15.10](#) opens, asking you to assign columns into X and Y roles, and an optional Group role. The Group role allows for fitting the model to every level of a categorical variable. If you specify a Group role here, also specify the Group column on the platform launch window.

Figure 15.10 Select Roles



For most models, the starting values are constants. Showing points enables you to adjust the parameter values to see how well the model fits for different values of the parameters. For the US population example, the points are shown in [Figure 15.11](#).

Figure 15.11 Show Points



Clicking **Make Formula** at this point (after using **Show Points**) creates a new data table column named after the model that you chose from the Model Library. This column has the formula as a function of the latest parameter starting values.

Note: If you click **Make Formula** before using the **Show Graph** or **Show Points** buttons, you are asked to provide the X and Y roles, and an optional Group role (Figure 15.10). After that, you are brought back to the plot so that you have the opportunity to adjust the parameters starting values if desired. At that point click **Make Formula** again to create the new column.

Once the formula is created in the data table, continue the analysis by assigning the new column as the **X, Predictor Formula** in the Nonlinear launch dialog.

Customize the Nonlinear Model Library

The Model Library in the Nonlinear platform is created by a built-in script named NonlinLib.jsl, located in the Resources/Builtins folder in the folder that contains JMP (Windows) or in the Application Package (macOS). You can customize the nonlinear library script by modifying this script.

Note: You might not be able to directly edit the file. If this is the case, open the script, make the desired edits, and save to a different location on your computer. Use the same filename. Then, delete the old `NonlinLib.jsl` script in the `Resources/Builtins` folder and move the new `NonlinLib.jsl` file into the `Resources/Builtins` folder.

To add a model, you must add three lines to the list named `Listofmodellist#`. These three lines are actually a list themselves, which consists of the following three parts.

- Model name, a quoted string
- Model formula, an expression
- Model scale

For example, suppose you want to add a model called “Simple Exponential Growth” that has the form

$$y = b_1 e^{kx}$$

Add the following lines to the `NonlinLib.jsl` script

```
{//Simple Exponential Growth
  "Simple Exponential Growth",
  Expr(Parameter({b1=2, k=0.5}, b1*exp(k * :X))),
  lowx = -1; highx = 2; lowy = 0;  highy = 2},
```

Some things to note:

- The first line is simply an open bracket (starting the list) and an optional comment. The second line is the string that is displayed in the model library window.
- The values of `lowx`, `highx`, `lowy`, and `highy` specify the initial window for the theoretical graph.
- There is a comma as the last character in the example above. If this is the final entry in the `Listofmodellist#` list, the comma can be omitted.
- If the model uses more than two parameters, replace the last line (containing the graph limits) with the quoted string “String Not Available”.

To delete a model, delete the corresponding three-lined list from the `Listofmodellist#` list.

Additional Examples of the Nonlinear Platform

This section contains examples using the Nonlinear platform.

- [“Example of Maximum Likelihood”](#)
- [“Example of a Probit Model with Binomial Errors”](#)
- [“Example of a Poisson Loss Function”](#)
- [“Example of Setting Parameter Limits”](#)
- [“Example of Analyzing Left-Censored Data”](#)
- [“Example of Fitting a Weibull Loss Function”](#)
- [“Example of Fitting Simple Survival Distributions”](#)

Example of Maximum Likelihood

This example uses the Nonlinear platform to minimize a loss function. The loss function is the negative of a log-likelihood function, thus producing maximum likelihood estimates.

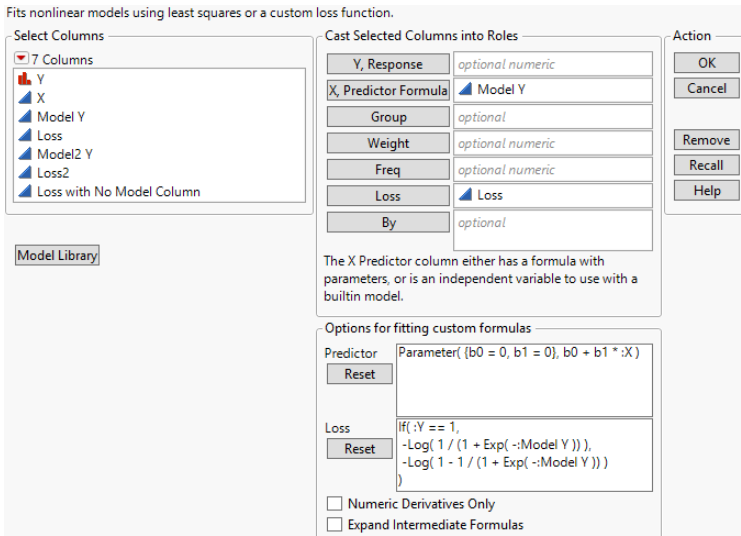
Estimate the Model

1. Select **Help > Sample Data Folder** and open Nonlinear Examples/Logistic w Loss.jmp.
2. Select **Analyze > Specialized Modeling > Nonlinear**.
3. Assign Model Y to the **X, Predictor Formula** role.

This column contains the linear model. The model is shown in the Predictor section of the launch window.

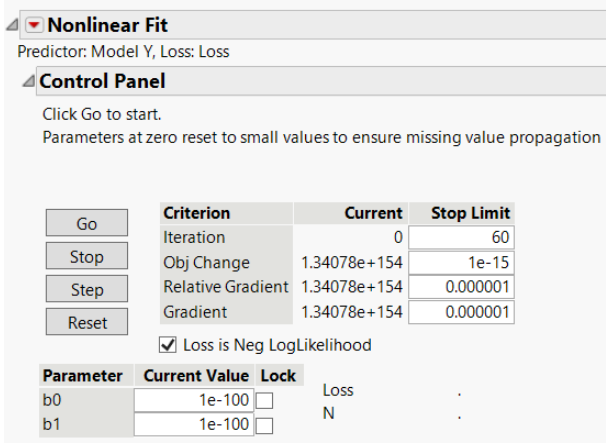
4. Assign Loss to the **Loss** role to define a custom loss function.

Figure 15.12 Nonlinear Launch Window



5. Click **OK**.

Figure 15.13 Nonlinear Fit Control Panel



6. Click **Go** in the Control Panel.

Figure 15.14 Solution Report

Solution			
Loss	DFE	Avg Loss	Sqrt Avg Loss
7.8070527905	18	0.4337252	0.6585781
Parameter	Estimate	ApproxStdErr	
b0	-5.511451052	2.43728165	
b1	0.0347553478	0.01595111	
Solved By: Analytic NR			

The Loss value in the Solution report is the negative log-likelihood evaluated at the parameter estimates.

Create Custom Profiler

Because our model was fit using a custom loss function, a standard profiler is not available. Use the custom profiler to explore model predictions.

1. Click the Nonlinear Fit red triangle and select **Custom Estimation Profiler**.
2. Enter the model as the custom estimation expression. To do so enter the following expression in the Enter Expression window:

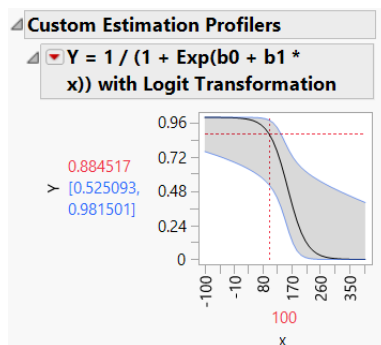
$$1/(1 + \text{Exp}(b_0 + b_1 * x))$$
3. Select Logit from the Transformation list.

Figure 15.15 Completed Enter Expression Window

4. Click **OK**.
5. Enter 100 as the initial value.
6. Click **OK**.

By default, the profiler is difficult to interpret because the ranges for response and the factor are narrow. You can adjust the axis settings to use a meaningful range.

7. Double click the vertical axis. In the Scale outline, specify 0 as the Minimum and 1 as the Maximum.
8. Click **OK**.
9. Double click the horizontal axis. In the Scale outline, specify -100 as the Minimum and 400 as the Maximum.
10. Click **OK**.

Figure 15.16 Custom Estimation Profiler


The custom profiler shown is titled by the custom expression and shows the value of the response across the values of the x variable. The profiler also displays the 95% confidence intervals.

Example of a Probit Model with Binomial Errors

Use numerical derivatives in the Nonlinear platform to obtain maximum likelihood estimates for a probit model with binomial errors. The data table in this example includes the numbers of ingots tested for readiness after different treatments of heating and soaking times. The average number of ingots ready is the product of the number tested and the probability that an ingot is ready for use given the amount of time it was heated and soaked.

1. Select **Help > Sample Data Folder** and open Ingots2.jmp.
2. Select **Analyze > Specialized Modeling > Nonlinear**.
3. Assign P to the **X, Predictor Formula** role.

Using a probit model, the P column contains the following model formula:

Normal Distribution(**b0+b1*Heat+b2*Soak**)

The argument to the Normal Distribution function is a linear model of the treatments.

4. Assign Loss to the **Loss** role.

To specify binomial errors, the Loss column contains a loss function with the following formula:

$$-(N_{\text{ready}} * \text{Log}(p) + (N_{\text{total}} - N_{\text{ready}}) * \text{Log}(1 - p))$$

5. Select the **Numeric Derivatives Only** option.
6. Click **OK**.
7. Click **Go**.

The platform used the Numerical SR1 method to obtain the parameter estimates shown in [Figure 15.17](#).

Figure 15.17 Solution for the Ingots2 Data

Solution				
Loss		DFE	Avg Loss	Sqrt Avg Loss
47.479945327		16	2.9674966	1.7226423
Parameter	Estimate	ApproxStdErr		
b0	-2.893415255	0.51357485		
b1	0.0399554549	0.01202873		
b2	0.0362537778	0.15060404		
Solved By: Numerical SR1				

Example of a Poisson Loss Function

Use the Nonlinear platform to fit a Poisson model to count data. The Poisson distribution has the following parameterization:

$$P(Y = n) = \frac{e^{-\mu} \mu^n}{n!}, n = 0, 1, 2, \dots$$

where μ can be a single parameter, or a linear model with many parameters. Many texts and papers show how the model can be transformed and fit with iteratively reweighted least squares (Nelder and Wedderburn [1972](#)). However, in JMP it is more straightforward to fit the model directly.

1. Select **Help > Sample Data Folder** and open Ship Damage.jmp.
2. Select **Analyze > Specialized Modeling > Nonlinear**.
3. Assign model to the **X, Predictor Formula** role.
4. Assign Poisson to the **Loss** role.
5. Click **OK**.
6. Set the **Current Value** (initial value) for b0 to 1, and the other parameters to 0.

Figure 15.18 Enter New Parameters

Control Panel

Click Go to start.

Go

Stop

Step

Reset

Criterion	Current	Stop Limit
Iteration	0	60
Obj Change	1.34078e+154	1e-15
Relative Gradient	1.34078e+154	0.000001
Gradient	1.34078e+154	0.000001

☒ Loss is Neg LogLikelihood

Parameter	Current Value	Lock
b0	1	☐
B	0	☐
C	0	☐
D	0	☐
E	0	☐
Yr 65	0	☐
Yr 70	0	☐
Yr 75	0	☐
Used 75	0	☐

Loss

N

.

.

- 7. Click **Go**.
- 8. Click the **Confidence Limits** button.

The Solution report appears. The results include the parameter estimates and confidence intervals, and other summary statistics.

Figure 15.19 Solution Table for the Poisson Loss Example

Solution				
	Loss	DFE	Avg Loss	Sqrt Avg Loss
	68.281234087	25	2.7312494	1.6526492
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	-6.405914771	0.21744445	-6.8430512	-5.9896833
B	-0.543353982	0.17758996	-0.881379	-0.183552
C	-0.687418388	0.32904722	-1.3764541	-0.0745299
D	-0.075979835	0.2905789	-0.6715296	0.47523936
E	0.325581683	0.23587934	-0.143451	0.78520431
Yr 65	0.6971496668	0.1496412	0.40752821	0.99512758
Yr 70	0.8184263427	0.16977314	0.48728046	1.15369087
Yr 75	0.4534435316	0.23317069	-0.0123098	0.90388895
Used 75	0.3844880718	0.11827198	0.15341509	0.61742312
Solved By: Analytic NR				

Note: The standard errors, confidence intervals, and hypothesis tests are correct only if least squares estimation is done, or if maximum likelihood estimation is used with a proper negative log-likelihood.

Example of Setting Parameter Limits

This example shows how to save the prediction formula from Fit Curve and then set parameter limits in Nonlinear. The Fit Curve personality enables you to fit a model and then use the prediction equation in the full personality of the Nonlinear platform. This method requires more steps and user input but allows any nonlinear model to be fit.

1. To fit the model, complete [step 1](#) through [step 8](#) of “[Example of the Fit Curve Platform](#)”.
2. Click the Logistic 4P red triangle and select **Save Formulas > Save Parametric Prediction Formula**.

A new column named Toxicity Predictor appears in the data table.

3. Select **Analyze > Specialized Modeling > Nonlinear**.
4. Assign Toxicity to the **Y, Response** role.
5. Assign Toxicity Predictor to the **X, Predictor Formula** role.
6. Assign Formulation to the **Group** role.
7. Click **OK**.

The Nonlinear Fit window appears. In the Control Panel, parameter values and locking options are shown. The letters listed before each parameter correspond to variables from the Prediction Model in the Fit Curve function.

Figure 15.20 Nonlinear Fit Control Panel

Nonlinear Fit

Response: Toxicity, Predictor: Toxicity Predictor, Group: Formulation

Control Panel

Click Go to start.

Go

Stop

Step

Reset

Criterion	Current	Stop Limit
Iteration	0	60
Obj Change	1.34078e+154	1e-15
Relative Gradient	1.34078e+154	0.000001
Gradient	1.34078e+154	0.000001

Parameter	Current Value	Lock
a_standard	5.8437931223	☐
b_standard	0.320199595	☐
c_standard	0.5990151869	☐
d_standard	1.1068636768	☐
a_test A	5.9806928009	☐
b_test A	0.284551222	☐
c_test A	0.5909023298	☐
d_test A	1.101208093	☐
a_test B	5.7558678529	☐
b_test B	0.0844294427	☐
c_test B	0.592984828	☐
d_test B	1.0972015134	☐
a_test C	6.0099274621	☐
b_test C	0.298749146	☐
c_test C	0.5934036815	☐
d_test C	1.1059793	☐

SSE

N

.

.

Tip: You can lock parameters if you know the values from prior information.

8. Click the Nonlinear Fit red triangle and select **Parameter Bounds**.
Options for setting the lower and upper parameters appear next to the parameters.
9. Set the lower bounds for the parameters as shown in [Figure 15.21](#). You know from prior experience that the maximum toxicity of the drug is at least 1.1.

Figure 15.21 Setting Parameter Bounds

Control Panel

Click Go to start.

Go

Stop

Step

Reset

Criterion	Current	Stop Limit
Iteration	0	60
Obj Change	1.34078e+154	1e-15
Relative Gradient	1.34078e+154	0.000001
Gradient	1.34078e+154	0.000001

Parameter	Current Value	Lock	Lower Bound	Upper Bound
a_standard	5.8437931223	☐	.	.
b_standard	0.320199595	☐	.	.
c_standard	0.5990151869	☐	.	.
d_standard	1.1068636768	☐	1.1	.
a_test A	5.9806928009	☐	.	.
b_test A	0.284551222	☐	.	.
c_test A	0.5909023298	☐	.	.
d_test A	1.101208093	☐	1.1	.
a_test B	5.7558678529	☐	.	.
b_test B	0.0844294427	☐	.	.
c_test B	0.592984828	☐	.	.
d_test B	1.0972015134	☐	1.1	.
a_test C	6.0099274621	☐	.	.
b_test C	0.298749146	☐	.	.
c_test C	0.5934036815	☐	.	.
d_test C	1.1059793	☐	1.1	.

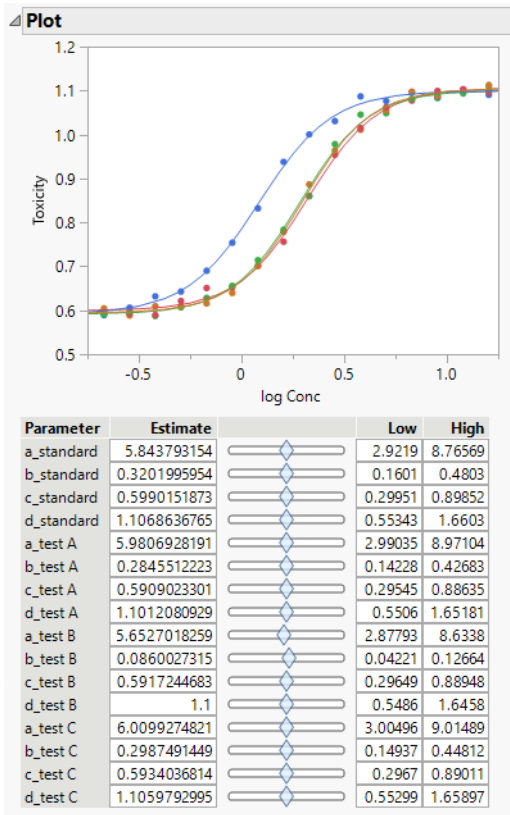
SSE

N

10. Click **Go**.

The final parameter estimates are shown in the **Solution** report, along with other fit statistics. The fitted model is shown on the plot.

Figure 15.22 Nonlinear Fit Plot and Parameter Estimates



Options below the plot allow for adjusting parameter limits and estimates.

Example of Analyzing Left-Censored Data

You can analyze left-censored data using the Nonlinear platform. For the left-censored data in this example, zero is the censored value because it also represents the smallest known time for an observation.

Note: The Tobit model is popular in economics for responses that must be positive or zero, with zero representing a censored point. It assumes a normal distribution that is censored at zero. An observation of zero is considered to be left-censored.

Note the following about the Tobit2.jmp data table:

- The response variable is a measure of the durability of a product and cannot be less than zero (Durable, is left-censored at zero).
- Age and Liquidity are independent variables.

- The table also includes the model and Tobit loss function. The model in residual form is $\text{durable} - (b_0 + b_1 * \text{age} + b_2 * \text{liquidity})$.

Tip: To see the formula associated with Tobit Loss, right-click the column and select **Formula**.

1. Select **Help > Sample Data Folder** and open Reliability/Tobit2.jmp.
2. Select **Analyze > Specialized Modeling > Nonlinear**.
3. Select Model and click **X, Predictor Formula**.
4. Select Tobit Loss and click **Loss**.
5. Click **OK**.
6. Click **Go**.
7. Click **Confidence Limits**.

The parameter estimates with confidence limits are shown in [Figure 15.23](#).

Figure 15.23 Solution Report

Solution				
Loss	DFE	Avg Loss	Sqrt	
28.92596097	16	1.8078726	1.3445715	
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	15.277120063	16.0327167	-22.059913	54.7012259
b1	-0.134007501	0.21893135	-0.7365041	0.32841861
b2	-0.045135584	0.05826851	-0.1858906	0.09330941
Sigma	5.5693492202	1.72814365	3.31613117	11.5291865
Solved By: Analytic NR				

Example of Fitting a Weibull Loss Function

In this example, models are fit to the survival time using the Weibull, lognormal, and exponential distributions in the Nonlinear platform. Model fits include a simple survival model containing only two effects, a more complex model with all the effects, and the creation of dummy variables for the discrete effect Cell Type to be included in the full model.

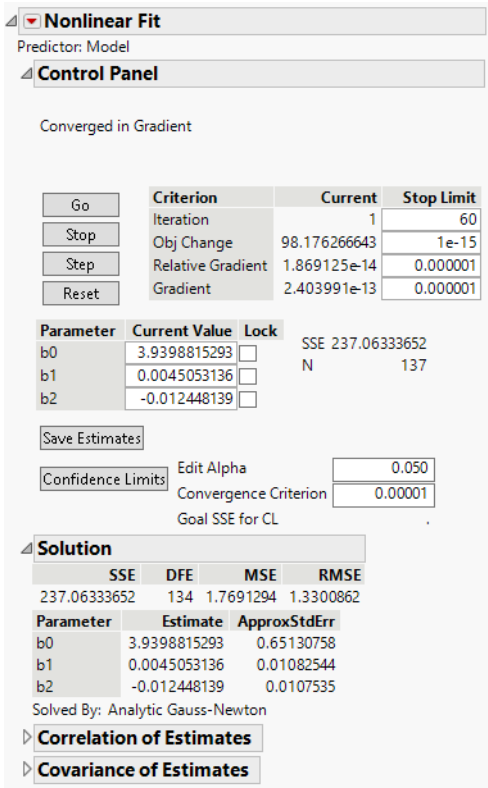
Nonlinear model fitting is often sensitive to the initial values that you give to the model parameters. In this example, one way to find reasonable initial values is to first use the Nonlinear platform to fit only the linear model. When the model converges, the solution values for the parameters become the initial parameter values for the nonlinear model.

1. Select **Help > Sample Data Folder** and open VA Lung Cancer.jmp.

The first model and all the loss functions have already been created as formulas in the data table. The Model column has the following formula:

- $\text{Log}(\text{:Time}) - (b_0 + b_1 * \text{Age} + b_2 * \text{Diag Time})$
2. Select **Analyze > Specialized Modeling > Nonlinear**.
 3. Select Model and click **X, Predictor Formula**.
 4. Click **OK**.
 5. Click **Go**.

Figure 15.24 Initial Parameter Values in the Nonlinear Fit Control Panel



- The report computes the least squares parameter estimates for this model.
6. Click **Save Estimates**.
- The parameter estimates in the column formulas are set to those estimated by this initial nonlinear fitting process.
- The Weibull column contains the Weibull formula. See *Reliability and Survival Methods*.
- To continue with the fitting process:
7. Select **Analyze > Specialized Modeling > Nonlinear** again.
 8. Select Model and click **X, Predictor Formula**.

9. Select Weibull loss and click **Loss**.
10. Click **OK**.

Figure 15.25 Initial Nonlinear Fit Control Panel

Nonlinear Fit
Predictor: Model, Loss: Weibull loss

Control Panel
Click Go to start.

Criterion	Current	Stop Limit
Iteration	0	60
Obj Change	1.34078e+154	1e-15
Relative Gradient	1.34078e+154	0.000001
Gradient	1.34078e+154	0.000001

☒ Loss is Neg LogLikelihood

Parameter	Current Value	Lock
b0	3.9398815293	☐
b1	0.0045053136	☐
b2	-0.012448139	☐
sigma	1	☐

Loss: .
N: .

There is now the additional parameter called sigma in the loss function. Because it is in the denominator of a fraction, a starting value of 1 is reasonable for sigma. When using any loss function other than the default, the **Loss is Neg LogLikelihood** box on the Control Panel is checked by default.

11. Click **Go**.

Figure 15.26 Nonlinear Fit

Nonlinear Fit
Predictor: Model, Loss: Weibull loss

Control Panel
Converged in Gradient

Criterion	Current	Stop Limit
Iteration	13	60
Obj Change	4.145281e-10	1e-15
Relative Gradient	1.6644587e-9	0.000001
Gradient	3.3666278e-8	0.000001

☒ Loss is Neg LogLikelihood

Parameter	Current Value	Lock
b0	5.6130465522	☐
b1	-0.012429709	☐
b2	-0.01091692	☐
sigma	1.1599937998	☐

Loss: 227.61384948
N: 137

Save Estimates

Confidence Limits: Edit Alpha: 0.050
Convergence Criterion: 0.00001
Goal SSE for CL: .

The fitting process estimates the parameters by maximizing the negative log of the Weibull likelihood function.

12. (Optional) Click **Confidence Limits** to show lower and upper 95% confidence limits for the parameters in the Solution table.

Figure 15.27 Solution Report

Solution				
Loss	DFE	Avg Loss	Sqrt	
227.61384948	133	1.7113823	1.3081981	
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
b0	5.6130465522	0.67171997	4.31169667	6.96484024
b1	-0.012429709	0.01114031	-0.0346969	0.00930037
b2	-0.01091692	0.010399	-0.0289695	0.01220355
sigma	1.1599937998	0.07823465	1.0209347	1.33040517
Solved By: Analytic NR				
Correlation of Estimates				
Covariance of Estimates				

Note: Because the confidence limits are profile likelihood confidence intervals instead of the standard asymptotic confidence intervals, they can take time to compute.

You can also run the model with the predefined exponential and lognormal loss functions. Before you fit another model, reset the parameter estimates to the least squares estimates, as they might not converge otherwise. To reset the parameter estimates:

13. (Optional) Click the Nonlinear Fit red triangle and select **Revert to Original Parameters**.

Example of Fitting Simple Survival Distributions

This section contains examples of fitting simple survival distributions in the Nonlinear platform.

Exponential, Weibull, and Extreme-Value Loss Function

You can use the Nonlinear platform to fit exponential, Weibull, and extreme value loss functions to failure time data. The data in this example are from a study of 70 diesel fans that accumulated a total of over 300 thousand hours in service. The fans were placed in service at different times. The response is failure time of the fans or run time at censoring.

1. Select **Help > Sample Data Folder** and open Reliability/Fan.jmp.

Tip: To view the formulas for the loss functions, in the Fan.jmp data table, right-click the Exponential, Weibull, and Extreme value columns and select **Formula**.

2. Select **Analyze > Specialized Modeling > Nonlinear**.

Lognormal Loss Function

The lognormal distribution is useful when the range of the data is several powers of e . The lognormal loss function can be very sensitive to starting values for its parameters. Because the lognormal distribution is similar to the normal distribution, you can create a new variable that is the natural log of Time and use **Distribution** to find the mean and standard deviation of this column. Then, use those values as starting values for the Nonlinear platform. In this example, the mean of the natural log of Time is 4.72 and the standard deviation is 0.35.

1. Select **Help > Sample Data Folder** and open Reliability/Locomotive.jmp.

Tip: To view the formula for the loss function, in the Locomotive.jmp data table, right-click the logNormal column and select **Formula**.

2. Select **Analyze > Specialized Modeling > Nonlinear**.
3. Select logNormal and click **Loss**.
4. Click **OK**.
5. Type 4.72 in the box next to Mu.
6. Type 0.35 in the box next to sigma.
7. Click **Go**.
8. Click **Confidence Limits**.

Figure 15.29 Solution Report

Solution				
Loss	DFE	Avg Loss	Sqrt	
237.09354534	94	2.5222718	1.5881662	
Parameter	Estimate	ApproxStdErr	Lower CL	Upper CL
Mu	5.1169247193	0.10416697	4.93735853	5.35966436
sigma	0.7054940302	0.0932136	0.55446171	0.93294904

Solved By: Numerical NR

The maximum likelihood estimates of the lognormal parameters are 5.11692 for Mu and 0.7055 for Sigma. The corresponding estimate of the median of the lognormal distribution is the antilog of 5.11692, ($e^{5.11692}$), which is approximately 167. This represents the typical life for a locomotive engine.

Statistical Details for the Nonlinear Platform

This section contains statistical details for the Nonlinear platform.

- [“Statistical Details for Profile Likelihood Confidence Limits”](#)
- [“Statistical Details for Custom Loss Functions”](#)
- [“Statistical Details on Derivatives”](#)
- [“Statistical Details on Effective Nonlinear Modeling”](#)

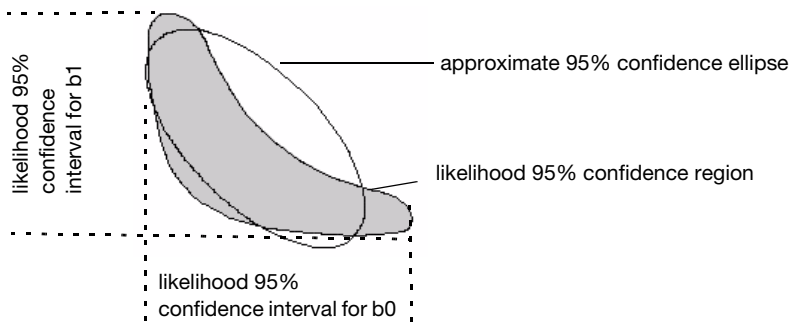
Statistical Details for Profile Likelihood Confidence Limits

The upper and lower confidence limits for the parameters in the Nonlinear platform are based on a search for the value of each parameter after minimizing with respect to the other parameters. The search looks for values that produce an SSE greater by a certain amount than the solution’s minimum SSE. The goal of this difference is based on the F -distribution. The intervals are sometimes called *likelihood confidence intervals* or *profile likelihood confidence intervals* (Bates and Watts 1988; Ratkowsky 1990).

Profile confidence limits all start with a *goal SSE*. This is a sum of squared errors (or sum of loss function) that an F test considers significantly different from the solution SSE at the given alpha level. If the loss function is specified to be a negative log-likelihood, then a Chi-square quantile is used instead of an F quantile. For each parameter’s upper confidence limit, the parameter value is increased until the SSE reaches the goal SSE. As the parameter value is moved up, all the other parameters are adjusted to be least squares estimates subject to the change in the profiled parameter. Conceptually, this is a compounded set of nested iterations. Internally there is a way to do this with one set of iterations developed by Johnston and DeLong. See the *SAS/ETS User’s Guide* (SAS Institute Inc. 2020).

[Figure 15.30](#) shows the contour of the goal SSE or negative likelihood, with the least squares (or least loss) solution inside the shaded region:

- The asymptotic standard errors produce confidence intervals that approximate the region with an ellipsoid and take the parameter values at the extremes (at the horizontal and vertical tangents).
- Profile confidence limits find the parameter values at the extremes of the true region, rather than the approximating ellipsoid.

Figure 15.30 Diagram of Confidence Limits for Parameters


Likelihood confidence intervals are more trustworthy than confidence intervals calculated from approximate standard errors. If a particular limit cannot be found, computations begin for the next limit. When you have difficulty obtaining convergence, try the following:

- use a larger alpha, resulting in a shorter interval, more likely to be better behaved
- relax the confidence limit criteria.

Statistical Details for Custom Loss Functions

The Nonlinear platform can minimize or maximize functions other than the default sum of squares residual. This section shows the mathematics of how it is done.

Suppose that $f(\beta)$ is the model. Then the Nonlinear platform attempts to minimize the sum of the loss functions defined as follows:

$$L = \sum_{i=1}^n \rho(f(\beta))$$

The loss function $\rho(\bullet)$ for each row can be a function of other variables in the data table. It must have nonzero first- and second-order derivatives. The default $\rho(\bullet)$ function, squared-residuals, is

$$\rho(f(\beta)) = (y - f(\beta))^2$$

To specify a model with a custom loss function, construct a variable in the data table and build the loss function. After launching the Nonlinear platform, select the column containing the loss function as the loss variable.

The nonlinear minimization formula works by taking the first two derivatives of $\rho(\bullet)$ with respect to the model, and forming the gradient and an approximate Hessian as follows:

$$L = \sum_{i=1}^n \rho(f(\beta))$$

$$\frac{\partial L}{\partial \beta_j} = \sum_{i=1}^n \frac{\partial \rho(f(\beta))}{\partial f} \frac{\partial f}{\partial \beta_j}$$

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} = \sum_{i=1}^n \left[\frac{\partial^2 \rho(f(\beta))}{(\partial f)^2} \frac{\partial f}{\partial \beta_j} \frac{\partial f}{\partial \beta_k} + \frac{\partial \rho(f(\beta))}{\partial f} \frac{\partial^2 f}{\partial \beta_k \partial \beta_j} \right]$$

If $f(\bullet)$ is linear in the parameters, the second term in the last equation is zero. If not, you can still hope that its sum is small relative to the first term, and use

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_k} \cong \sum_{i=1}^n \frac{\partial^2 \rho(f(\beta))}{(\partial f)^2} \frac{\partial f}{\partial \beta_j} \frac{\partial f}{\partial \beta_k}$$

The second term is probably small if ρ is the squared residual because the sum of residuals is small. The term is zero if there is an intercept term. For least squares, this is the term that distinguishes Gauss-Newton from Newton-Raphson.

Note: The standard errors, confidence intervals, and hypothesis tests are correct only if least squares estimation is done, or if maximum likelihood estimation is used with a proper negative log-likelihood.

Statistical Details on Derivatives

The Nonlinear platform takes symbolic derivatives for formulas with most common operations. This section shows what type of derivative expressions result.

If you open the `Negative Exponential.jmp` nonlinear sample data example, the actual formula for the Nonlinear column looks something like this:

```
Parameter({b0=0.5, b1=0.5}, b0*(1-Exp(-b1*X)))
```

The `Parameter` block in the formula is hidden if you use the formula editor. That is how it is stored in the column and how it appears in the Nonlinear Launch dialog. Two parameters named **b0** and **b1** are given initial values and used in the formula to be fit.

The Nonlinear platform makes a separate copy of the formula, and edits it to extract the parameters from the expression. Then it maps the references to them to the place where they are estimated. Nonlinear takes the analytic derivatives of the prediction formula with respect to the parameters. If you use the **Show Derivatives** command, you get the resulting formulas listed in the log, like this:

Prediction Model:

```
b0 * First(T#1=1-(T#2=Exp(-b1*X)), T#3=-(-1*T#2*X))
```

The Derivative of Model with respect to the parameters is:

```
{T#1, T#3*b0}
```

The derivative facility works like this:

- In order to avoid calculating subexpressions repeatedly, the prediction model is threaded with assignments to store the values of subexpressions that it needs for derivative calculations. The assignments are made to names like **T#1**, **T#2**, and so on.
- When the prediction model needs additional subexpressions evaluated, it uses the **First** function, which returns the value of the first argument expression, and also evaluates the other arguments. In this case additional assignments are needed for derivatives.
- The derivative table itself is a list of expressions, one expression for each parameter to be fit. For example, the derivative of the model with respect to **b0** is **T#1**; its thread in the prediction model is $1 - (\text{Exp}(-b1 * X))$. The derivative with respect to **b1** is **T#3*b0**, which is $-(-1 * \text{Exp}(-b1 * X) * X) * b0$ if you substitute in the assignments above. Although many optimizations are made, it does not always combine the operations optimally. You can see this by the expression for **T#3**, which does not remove a double negation.

If you specify a loss function, then the formula editor takes derivatives with respect to parameters, if it has any. And it takes first and second derivatives with respect to the model, if there is one.

If the derivative mechanism does not know how to take the analytic derivative of a function, then it takes numerical derivatives, using the **NumDeriv** function. If this occurs, the platform shows the delta that it used to evaluate the change in the function with respect to a delta change in the arguments. You might need to experiment with different delta settings to obtain good numerical derivatives.

Tips

There are always many ways to represent a given model, and some ways behave much better than other forms. Ratkowsky (1990) covers alternative forms in his text.

If you have repeated subexpressions that occur several places in a formula, then it is better to make an assignment to a temporary variable. Then refer to it later in the formula. For example, one of the model formulas above was this:

```
If(Y==0, Log(1/(1+Exp(model))), Log(1 - 1/(1 + Exp(model))));
```

This could be simplified by factoring out an expression and assigning it to a local variable:

```
temp=1/(1+Exp(model));  
If(Y==0, Log(temp), Log(1-temp));
```

The derivative facility can track derivatives across assignments and conditionals.

Statistical Details on Effective Nonlinear Modeling

When using the Nonlinear platform, it is strongly encouraged to *center polynomials*.

If you have a complete polynomial term that is linear in the parameters, it is always good to center the polynomials. This improves the condition of the numerical surface for optimization. For example, if you have an expression like the following:

$$a_1 + b_1x + c_1x^2$$

you should transform it to

$$a_2 + b_2(x - \bar{x}) + c_2(x - \bar{x})^2$$

The two models are equivalent, apart from a transformation of the parameters, but the second model is far easier to fit if the model is nonlinear.

The transformation of the parameters is easy to solve.

$$\begin{aligned} a_1 &= a_2 - b_2\bar{x} + c_2\bar{x}^2 \\ b_1 &= b_2 - 2c_2\bar{x} \\ c_1 &= c_2 \end{aligned}$$

If the number of iterations still goes to the maximum, increase the maximum number of iterations or relax one of the convergence criteria.

There is really no one omnibus optimization method that works well on all problems. JMP has options like **Newton**, **QuasiNewton BFGS**, **QuasiNewton SR1**, and **Numeric Derivatives Only** to expand the range of problems that are solvable by the Nonlinear Platform.

If the default settings are unable to converge to the solution for a particular problem, using various combinations of these settings to increase the odds of obtaining convergence.

Some models are very sensitive to starting values of the parameters. Working on new starting values is often effective. Edit the starting values and click **Reset** to see the effect. The plot often helps. Use the sliders to visually modify the curve to fit better. The parameter profilers can help, but might be too slow for anything but small data sets.

Chapter 16

JMP[®] PRO Functional Data Explorer

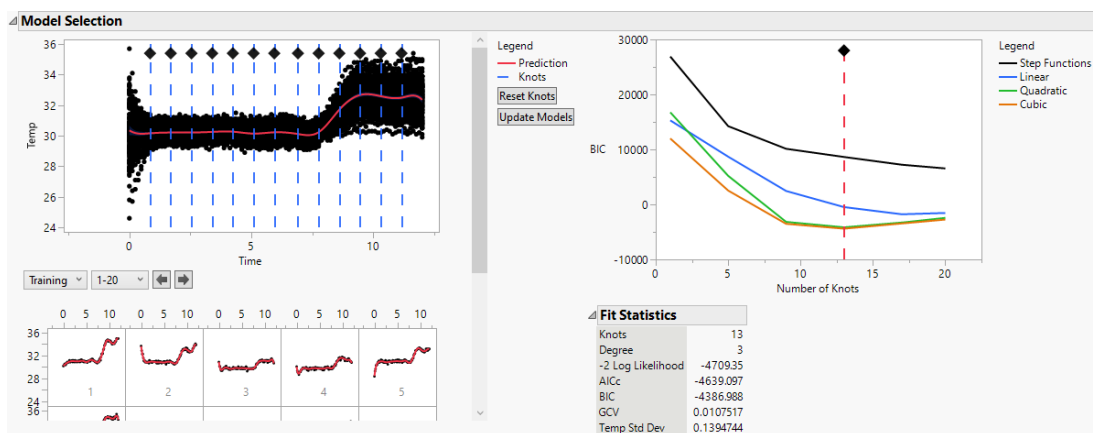
Perform Functional Data Analysis

The Functional Data Explorer platform is available only in JMP Pro.

The Functional Data Explorer (FDE) platform is designed for data that are functions, signals, spectra, or series. It can be used as an exploratory data analysis tool or as a dimension-reduction technique. In the case of dimension reduction, the FDE platform converts functional data into a form that can be analyzed in another JMP platform.

Data preprocessing tools are available in the FDE platform, including several types of transformations for output data and alignments for input data. A functional model is created by fitting a B-spline, P-spline, Fourier, or Wavelets basis model to the data. When a model is fit, functional principal components analysis (functional PCA) is automatically performed on the functional model. Results from the functional PCA, such as the functional principal component scores, are saved to a separate data table for feature extraction or dimension reduction. There is also an option to perform functional principal components analysis directly on the data, without fitting a basis function model first.

The FDE platform also contains options for design of experiments (DOE). Using one of the DOE options, you can model features of the functional data using the supplementary variables that were specified in the launch window. This enables you to use the model fit to determine how the functional response changes based on the supplementary variables. B-Spline Model Selection in FDE



Contents

Overview of the Functional Data Explorer Platform	301
Example of Functional Data Explorer	301
Launch the Functional Data Explorer Platform	305
The Functional Data Explorer Report	308
Model Reports	309
Functional Data Explorer Platform Options	316
Functional Data Explorer Group Options	318
Data Processing Options	319
Model Report Options	323
Additional Examples of the Functional Data Explorer Platform	326
Example for Multiple Functional Processes	327
Example of Functional DOE Analysis	331
Statistical Details for the Functional Data Explorer Platform	333
Statistical Details for Functional Model Fits	334
Statistical Details for the Function Summaries Report	335

Overview of the Functional Data Explorer Platform

Functional data can be defined as data that are recorded over a continuous domain, where a set of measurements form a curve or image. Often, the domain is time and the sets of measurements are defined by an ID variable. A functional data observation for ID level i at a specific point t on the domain is written as $f_i(t)$. The Functional Data Explorer platform enables you to explore and analyze functional data.

The form of functional data can be dense or sparse. Dense functional data occur when observations are on the same equally spaced grid of points for all levels of the ID variable. Sparse functional data occur when ID levels have different numbers of observations that are unequally spaced across the domain. The Functional Data Explorer platform can handle both forms of functional data.

Although functional data can be expressed in many ways, it can generally be classified into the following two cases:

- The response of interest, $f(t)$, has a functional form.
- There are one or more covariates, $f(t)$'s, that have functional forms. These are sometimes referred to as functional or signal processes.

The Functional Data Explorer platform is useful as an exploratory tool for any type of functional data. However, the strength of the platform is taking many functional processes (that might be associated with a scalar response) and extracting key features to use in further modeling. This can be done by first fitting a functional model to the data. Then, a functional principal components analysis (functional PCA) is performed on the functional model. Results from the functional PCA, such as the functional principal component (FPC) scores, are saved and used for feature extraction and analysis in another modeling platform, such as the Generalized Regression personality of the Fit Model platform. Alternatively, you can specify a set of supplementary variables and fit a generalized regression model within the FDE platform to determine how these variables affect the response.

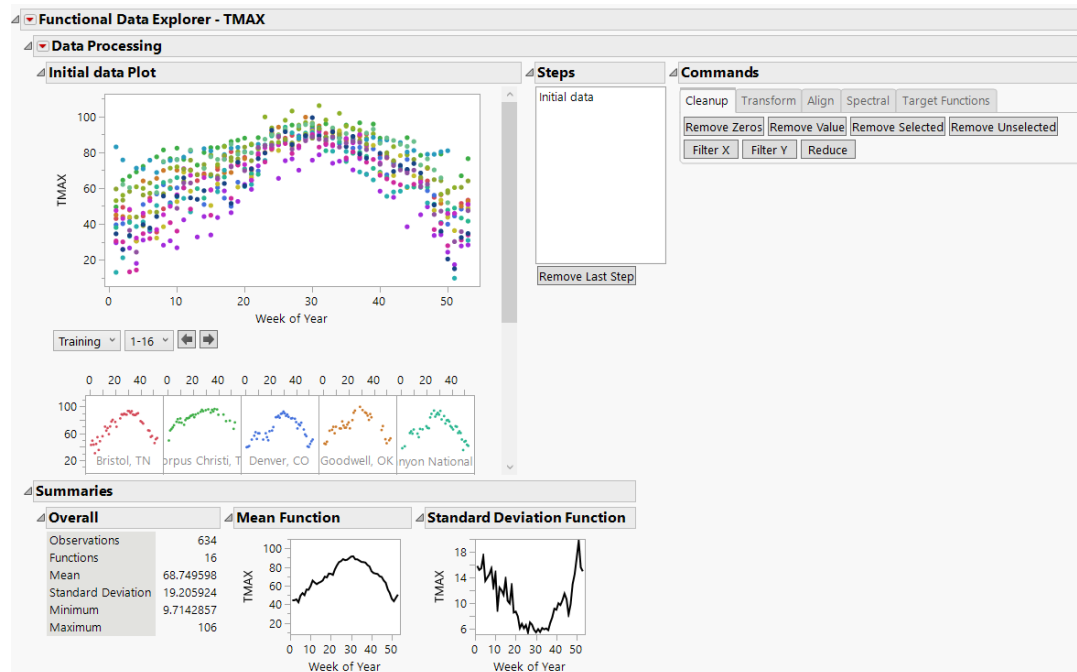
For more information about functional data analysis, see Ramsay and Silverman (2005).

Example of Functional Data Explorer

This example analyzes weekly weather data collected from 16 weather stations across the United States. This is an example of sparse functional data because not every weather station has weekly temperature measurement for every week of the year. Daily temperatures are summarized as weekly averages. Run the Weather Station Locations script in the data table to view a map of the locations.

1. Select **Help > Sample Data Folder** and open Functional Data/Weekly Weather Data.jmp.

2. Select **Analyze > Specialized Modeling > Functional Data Explorer**.
3. Select TMAX and click **Y, Output**.
4. Select Week of Year and click **X, Input**.
5. Select ID and click **ID, Function**.
6. Click **OK**.

Figure 16.1 Initial Functional Data Explorer Report


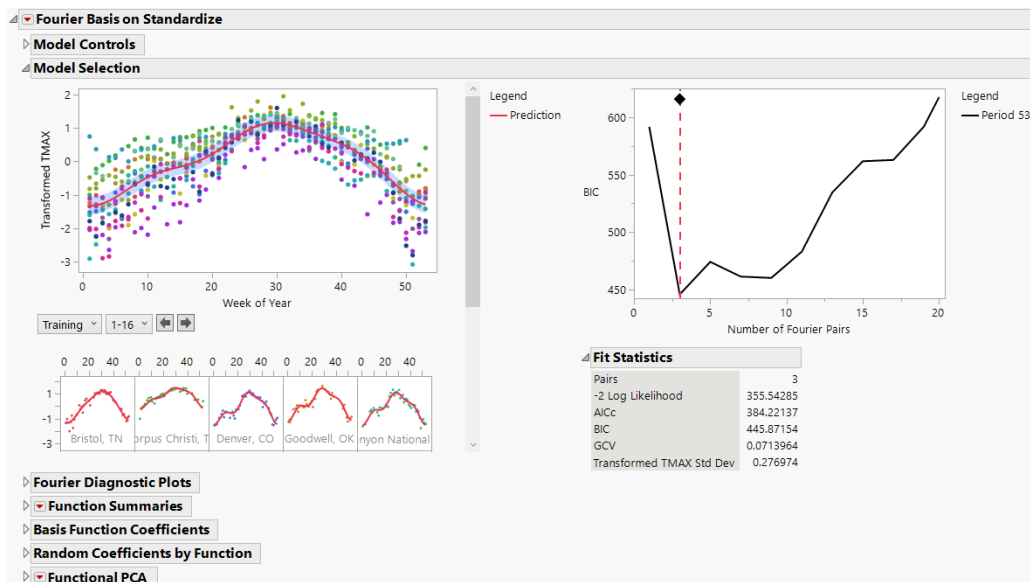
The initial Functional Data Explorer report contains plots of the raw data, summary statistics, and summary plots for the functional mean and functional standard deviation of the data. There is also a Commands section that contains options for data processing. Data processing options are also accessible from the Data Processing red triangle menu. Prior to modeling, it is often a good idea to standardize your output data.

7. Click the Transform tab in the Commands section.
8. Click the **Standardize** button.

The data plots and summary statistics are updated based on the specified transformation. Standardized is added to the Steps list.

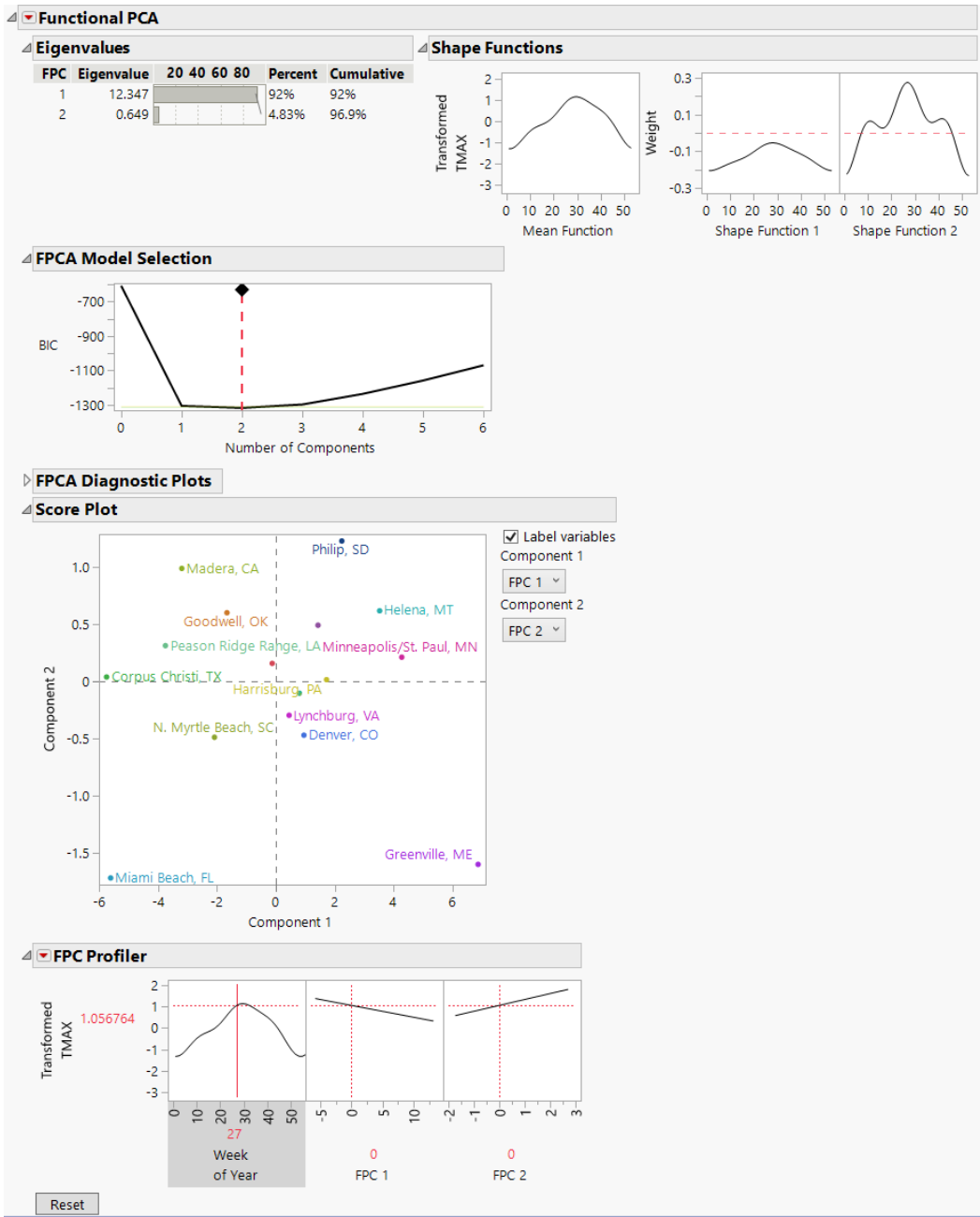
9. Click the Functional Data Explorer red triangle and select **Models > Fourier Basis**.

Figure 16.2 Fourier Basis Model Report



The Fourier Basis report includes several reports that contain information about the selected model. In the Model Selection report, the displayed model is the best fitting model according to the BIC fit criterion. For the weather data, the Fourier Basis model that is chosen has a period of 53 and three basis function pairs. Fit statistics and coefficients are also available for the model. Scroll down to view the Functional PCA report.

Figure 16.3 Functional PCA Report



The Functional PCA report shows that the first two eigenvalues explain nearly 97% of the variation in the data. In the Model Selection graph, click and drag the red dashed line to 3 FPCs to see that the first three eigenvalues explain 99% of the variation in the data.

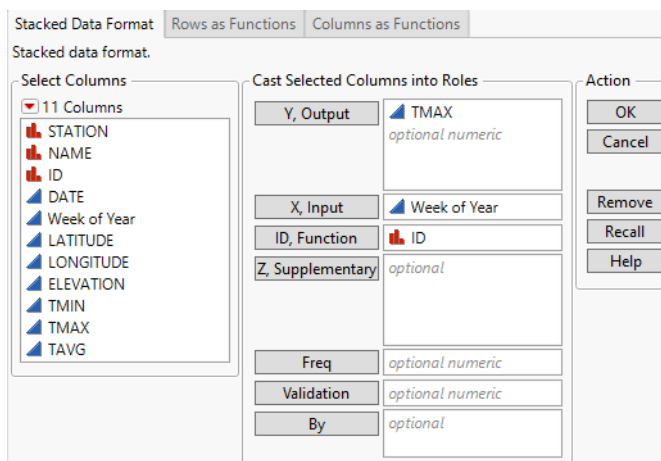
However, the first eigenvalue alone explains 92%. You can use the Score Plot to detect individual functions that are outliers from the other functions. In the Score Plot, most of the locations are clustered together except for the Miami Beach, FL and Greenville, ME locations. Scroll up to the individual function plots. The function for the Miami Beach location is flatter, indicating less temperature variability than the rest of the locations. The function for the Greenville location has a lower maximum, indicating consistently colder temperatures than the rest of the locations.

Tip: Deselect the Label variables option in the Score Plot report to better identify outliers.

JMP PRO Launch the Functional Data Explorer Platform

Launch the Functional Data Explorer platform by selecting **Analyze > Specialized Modeling > Functional Data Explorer**.

Figure 16.4 Functional Data Explorer Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Select Data Format

The Functional Data Explorer launch window includes tabs for different types of data formats. Select the tab based on your data format.

Stacked Data Format Select for data tables where each row corresponds to a single observation. There are separate columns for the output, input, and ID variables.

Note: The Stacked Data Format is the only data format that enables you to specify multiple functional processes. If you assign more than one column to **Y, Output** in the Stacked Data Format tab, each Y variable is analyzed separately. A Fit Group report contains the individual reports for the Y variables.

Rows as Functions Select for data tables where each row corresponds to the full output function for one level of the ID variable. Each column corresponds to a level of the input variable.

Tip: If the columns contain the FDE X column property, the column property is used to specify the values of the input variable. If the columns do not contain the column property, the Rows as Functions data format uses the levels of the input variable specified in the column names. If the column names are too complicated, an error message is shown to indicate which columns were problematic and the values are specified as $x = 1, 2, \dots, n$, where n is the number of observations in the function. For example, t20 2 and x30v20 are problematic column names.

Columns as Functions Select for data tables where each column corresponds to the entire output function for one level of the ID variable. Each row corresponds to a level of the input variable.

Specify Column Roles

The launch window contains the following options for specifying the column roles.

Y, Output Assigns the functional process, $f(t)$. There must be at least two observed output values for each level of the ID variable.

Note: Functions with fewer than two observed output values are removed from the analysis.

X, Input (Available only for the Stacked Data Format and Columns as Functions tabs.) Assigns the input variable t . If no variable is specified for X, Input, the row number is used. Using the row number assumes that the observations are equally spaced in the input domain.

ID, Function (Available only for the Stacked Data Format and Rows as Functions tabs.) Assigns the ID variable to each function. For Stacked Data, if no ID variable is assigned all observations are assumed to come from one function.

Z, Supplementary (Available only for the Stacked Data Format and Rows as Functions tabs.) Assigns one or more supplementary variables. Supplementary variables are not used in any of the calculations in the Functional Data Explorer platform and including them does not affect the results. Supplementary variables are variables you might want to use in

future analyses of the results from Functional Data Explorer. When you specify supplementary variables, they are included in a Supplementary column group in the tables that are created by the Save Data and Save Summaries options. These columns retain any column properties that were specified in the original data table. A Functional DOE Analysis option is also made available for the fitted models. See [“Functional DOE Analysis”](#).

Freq (Available only for the Stacked Data Format tab.) Assigns a column whose numeric values represent a frequency for each row in the analysis. The effect of a frequency column is to expand the data table, so that any row with integer frequency k is expanded to k identical rows.

Validation Assigns an optional numeric column that defines the validation sets. This column should contain only two distinct values. The smaller value defines the training set and the larger value defines the validation set. If there are more than two values, the smallest value defines the training set and all other values define the validation set. The FDE platform uses the validation column to train and evaluate the model. For more information about validation, see [“Validation in JMP Modeling”](#).

Note: The Validation option enables you to hold out complete functions, not a sample of observations from each function. Therefore, all observations that have the same ID value must be classified as either test or validation. You cannot have observations with the same ID value in both sets. For more information about this type of validation column, see [“Grouped Validation Column”](#).

Caution: If you click the Validation button with no columns selected in the Select Columns list, and add a validation column to your data table it will not be a grouped validation column. Instead, use the Make Validation Column Platform, see [“Grouped Validation Column”](#).

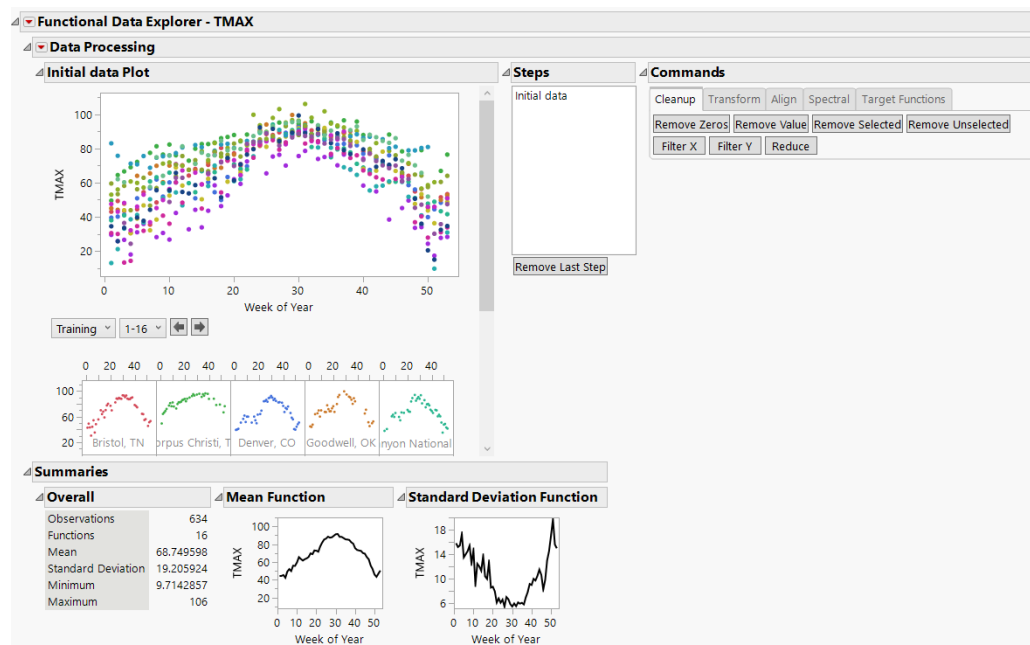
By Assigns a column that creates a report consisting of separate analyses for each level of the variable. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.



The Functional Data Explorer Report

The initial Functional Data Explorer report contains a Data Processing report and a Summaries report.

Figure 16.5 Initial Functional Data Explorer Report



The Data Processing report includes an initial data plot of all observations, as well as a grid of individual data plots that correspond to levels of the ID variable. The maximum number of individual plots shown in the grid is twenty. There are drop-down menus and arrows above the grid to interactively select which individual plots are shown. If you specified a validation set or target functions, the first drop-down menu enables you to select if the individual plots shown are from the training set, the validation set, or the target functions. If there are more than twenty functions, the second drop-down menu and the arrows enable you to view different groups of plots. All graphs plot the functional response data over the values of the input variable. The Data Processing report also contains data processing buttons in the Commands section.

The Summaries report contains a table of overall summary statistics, including the number of observations, number of functions, and the overall mean, standard deviation, minimum, and maximum values. There are also plots of the functional mean and functional standard deviation. The functional summary statistics displayed in the plots are computed at each unique value of the input variable.

The Data Processing and Summaries reports are useful for preprocessing your data prior to fitting a model or performing direct functional principal components analysis. There are data cleaning, transformation, alignment, and spectral buttons available in the Data Processing report that correspond to the options in the Data Processing red triangle menu. See [“Data Processing Options”](#). Each time you perform a preprocessing step, the graphs and summary results are automatically updated. You can perform multiple preprocessing steps in sequence. Each step is added to the Steps list of the Data Processing report. To remove the most recent step, click the **Remove Last Step** button below the Steps outline.

Model Reports

Use the options in the Models submenu in the Functional Data Explorer red triangle menu to fit models to your data. See [“Models”](#) for the available models. Each time you fit a different type of model to the data, a model report appears. Each model report contains the following reports:

- [“Model Controls”](#)
- [“Model Selection”](#)
- [“Diagnostic Plots”](#)
- [“Function Summaries”](#)
- [“Coefficients”](#)
- [“Random Coefficients by Function”](#)
- [“Functional PCA”](#)

Model Controls

The Model Controls report enables you to define parameters of models to compare in the Model Selection report. The appearance of the Model Controls report depends on the type of model that is fit. Model Controls are not available for Wavelets models.

B-Spline and P-Spline Model Controls

When a B-Spline or P-Spline model is fit, you can specify the following parameters:

Number of Knots Add, remove, or specify a range for the number of knots in each spline.
The knots must be non-zero integers.

Note: The maximum number of knots allowed for B-Spline models is one less than the maximum number of observations per function or the number of unique inputs. The maximum number of knots allowed for P-Spline models is two less than the number of unique inputs. If you specify a number larger than the maximum, a warning message appears.

Spline Degree Add or remove spline degree fits from the Model Selection report.

Fourier Basis Model Controls

When a Fourier Basis model is fit, you can specify the following parameters:

Number of Fourier Pairs Add, remove, or specify a range for the number of Fourier pairs to compare.

Period Change the period of the function.

After you specify the model controls, click **Go** to view the updated models in the Model Selection report.

Tip: To specify the Model Controls prior to fitting a model, press Shift, click the Functional Data Explorer red triangle, and select the desired model. See [“Models”](#).

Model Selection

The Model Selection report contains an overall prediction plot, a grid of individual prediction plots, a solution path plot, and a table. The grid of individual prediction plots has the same layout and controls as the grid of individual plots in the Data Processing report. At most, there are twenty plots shown at a time. There are drop-down menus and arrows that enable you to view different groups of individual prediction plots.

The prediction plots show the raw data and prediction curves for the current model. If there is a validation set, the predicted curves are not shown for functions that are in the validation set. The curve in the overall prediction plot is a prediction of the mean curve. The curves in the individual prediction plots are prediction curves for each specific function. For B-Spline models, the overall prediction plot also displays the location of the knots. You can change the location of the knots by dragging the blue slider bars to different locations. To update the model reports according to the new knot locations, click the **Update Models** button. To reset the knots to their default locations, click the **Reset Knots** button.

For Wavelets models, there is also a Coefficients tab. This tab contains an overall coefficients plot and a grid of individual coefficients plots. The coefficients plots are made up of dashed lines that represent the relative importance of the coefficients within each resolution. The placement of the lines on the plot is determined by the input variable (horizontal axis) and the resolution number (vertical axis). These numbers are located in the column names of the

Wavelets Coefficients table. For the individual plots, the color of each line on the plot is determined by the sign of the corresponding coefficient. Blue indicates positive coefficients and red indicates negative coefficients. The length of each line in the plot is determined by the coefficients. The length is the corresponding coefficient scaled by the largest absolute value coefficient within each resolution. Therefore, the largest coefficient has the longest line length and the lines get shorter as the coefficients get smaller. For the mean plot, the coefficients are averaged over each function. The color and length of each line is determined by the averaged and scaled coefficient.

Note: The Coefficients plots are available only for the best fitting wavelets model.

The appearance of the solution path plot and associated table depend on the model type.

B-Spline and P-Spline Models

The solution path plot shows a model selection criterion plotted over the defined number of knots. There is a separate solution path for each spline degree. The Bayesian Information Criterion (BIC) is the default fitting criterion. Use the model red triangle option **Model Selection** to change the selection criterion. The current solution is designated by the dotted vertical line on the solution path plot. By default, the spline degree and number of knots selected corresponds to the model that has the smallest model selection criterion value. To change the current model selection, drag the slider at the top of a vertical line or click a specific spline in the solution path plot or legend. These actions automatically update the prediction plots in the Model Selection report, as well as the information in all other reports.

The table below the solution path plot is the Fit Statistics table, which contains information about the current solution model. It shows the number of knots, the spline degree, the -2 Log Likelihood, the values for the AICc, BIC, and GCV model fitting criteria, and a value for the response standard deviation. The response standard deviation is defined as the residual sigma from the fitted model. When a P-Spline model is selected, the penalty parameter λ (Lambda) is also displayed.

Fourier Models

The solution path plot shows a model selection criterion plotted across the number of Fourier pairs for a defined period. The Bayesian Information Criterion (BIC) is the default fitting criterion. Use the model red triangle option **Model Selection** to change the selection criterion. The current solution is designated by the dotted vertical line in the solution path plot. By default, the slider is placed at the number of Fourier pairs that correspond to the model that has the smallest model selection criterion value. Drag the slider at the top of the dotted vertical line to change the number of Fourier pairs in the current model. Dragging the slider automatically updates the prediction plots in the Model Selection report, as well as the information in all other reports.

The table below the solution path plot is the Fit Statistics table, which contains information about the current solution model. It shows the number of Fourier pairs, the -2 Log Likelihood, the values for the AICc, BIC, and GCV model fitting criteria, and a value for the response standard deviation. The response standard deviation is defined as the residual sigma from the fitted model.

Wavelets Models

The solution path plot shows a model selection criterion plotted across the model number that defines the wavelets model. The Bayesian Information Criterion (BIC) is the default fitting criterion. Use the model red triangle option **Model Selection** to change the selection criterion. The current model selection is designated by the dotted vertical line on the solution path plot. By default, the selected model is the model that has the smallest model selection criterion value.

The table below the solution path plot shows the model numbers, corresponding model names, and current model selection. This table is sorted by the model selection criterion, with the best fitting model at the top. Select different wavelets models by dragging the slider at the top of the dotted vertical line or by selecting a model directly in the table. Selecting a different model automatically updates the prediction plots in the Model Selection report, as well as the information in all other report.

Diagnostic Plots

The Diagnostic Plots report contains the Actual by Predicted plot and the Residual by Predicted plot. These plots help assess how well the current model fits the data. The Diagnostic Plots report is closed by default.

Function Summaries

Displays summaries from the Functional PCA for each level of the ID variable. The functional principal components associated with eigenvalues that explain more than 1% variation in the data are displayed by default. The mean, standard deviation, median, minimum, maximum, integrated difference, root integrated square error (RISE), and root integrated function square (RIFS) are also shown. The integrated difference and RISE summary values are used to determine how much the ID specific function differs from the overall mean function. The RIFS summary value is used for optimal curve fitting. See [“Statistical Details for the Function Summaries Report”](#).

Function Summaries Options

The Function Summaries red triangle menu contains the following options:

Customize Function Summaries Displays a window that enables you to select the number of FPCs and the summary statistics that are shown in the Function Summaries report. If the number of FPCs to be shown is specified, the Functional PCA report is also updated. There is also a checkbox, *Save Graphs*, that determines whether a graph for each function is included in the data table produced by the *Save Summaries* option.

Note: The maximum number of FPCs that can be specified for B-Spline, P-Spline, or Fourier models is 10. For Wavelets models or Direct Functional PCA, the maximum number of FPCs that can be specified is 20.

Tip: If you have multiple functional processes, you can customize all Function Summaries reports to show the same summary values by clicking *Ctrl* and selecting *Customize Function Summaries*.

Save Summaries Saves the summary statistics specified in the Function Summaries report to a new data table. The name of the new data table describes the model fit. This data table contains formula columns for the eigenfunctions, mean function, prediction function, and conditional prediction function. If the *Save Graphs* option is selected in the *Customize Function Summaries* window, there is also a column that contains the image of a graph of the raw data and the specified model fit for each function. In the data table, there is a profiler script that launches the prediction profilers for the prediction and conditional prediction formulas. These formulas are functions of the input variable, the ID variable, and the eigenfunctions.

Tip: For Wavelets models, the wavelet coefficients are also saved to the data table that is created using the *Save Summaries* option.

Control Chart Builder Analyzes the functional principal components using the Control Chart Builder. An individual chart and Limit Summaries table is created for each functional principal component in the Function Summaries table. See *Quality and Process Methods*.

Coefficients

Displays the estimated coefficients for the specified model.

For basis function models, the estimated basis function coefficients and their standard deviations are shown. These coefficients are common across all levels of the ID variable and are fixed estimates in the mixed model framework. To view standard errors and confidence intervals for the coefficients, right-click in the table and select **Columns**.

For Wavelets models, the estimated wavelet coefficients are shown for each level of the ID variable. There is a column for the father wavelet and then a series of columns for the remainder of the coefficients. Each coefficient column is identified by the corresponding resolution and location on the input domain. The coefficients table contains a sparse representation of the wavelet coefficients that is found using thresholding (Donoho, 1995).

JMP PRO Random Coefficients by Function

Displays the estimated random coefficients for each basis function and functional process combination. These are unique to each level of the ID variable and are random effects estimates in the mixed model framework. The Random Coefficients by Function report is not show for wavelets models.

JMP PRO Functional PCA

Functional principal components analysis (functional PCA) is performed on the fitted functional model. The Functional PCA report lists the eigenvalues that correspond to each functional principal component (FPC) in order from largest to smallest. The percent of variation accounted for by each FPC and the cumulative percent is listed and shown in a bar chart. There is a graph of the mean function as well as a graph for each shape function. These are the values of the eigenfunctions.

You can perform model selection in the Functional PCA report to refine the selected number of functional principal components. There is a solution path plot that shows the Bayesian Information Criterion (BIC) plotted versus the number of FPCs. The current number of FPCs is designated by the dotted vertical line in the solution path plot. It is possible that models with different numbers of FPCs might have similar fits. Therefore, the solution path plot provides zones, which are intervals of values of the BIC statistic. There is a green zone and a yellow zone. The green zone contains values in the interval of the minimum BIC to the minimum BIC plus four and the yellow zone contains values in the interval of the minimum BIC plus four to the minimum BIC plus 10. By default, the model with the smallest number of FPCs within the green zone is selected. You can drag the slider at the top of the vertical line to change the number of FPCs. Dragging the slider automatically updates the other information in the Functional PCA report.

Note: Narrow zones relative to the full y-axis scale may be difficult to view on your plot. Zoom in on the y-axis to better visualize the zones.

When Direct Functional PCA is performed, there is also an overall prediction plot and a grid of individual prediction plots. The grid of individual prediction plots has the same layout and controls as the grid of individual plots in the Data Processing report. At most, there are twenty plots shown at a time and there are drop-down menus and arrows that enable you to view different groups of individual prediction plots. Updating the number of FPCs automatically updates the prediction plots as well.

The prediction plots show the raw data and prediction curves that correspond to the current model. If there is a validation set, the predicted curves are not shown for functions that are in the validation set. The curve in the overall prediction plot is a prediction of the mean curve, given the specified number of FPCs. The curves in the individual prediction plots are prediction curves for each specific function, given the specified number of FPCs.

Note: The Functional PCA report is not shown if only a single function is modeled. Otherwise, if JMP is unable to perform Functional PCA, an error message appears in the Functional PCA report.

The following options are available in the Functional PCA red triangle menu:

Diagnostic Plots Shows or hides the Actual by Predicted and the Residual by Predicted plots. Use these plots help assess how well the model fits the data, given the selected number of functional principal components.

Score Plot Shows or hides a score plot of the FPC scores. Use the lists under Select Component to specify which FPCs are plotted on each axis of the Score Plot. If there is only one FPC, the FPC scores are plotted on the line $y = x$ and the lists to change the components are not shown. Score plots are useful for detecting outliers. In the case of FPC scores, the Score Plot is useful for detecting levels of the ID variable that have outlier functions. If you select a point in the score plot, the FPC Profiler is set to the scores for that function.

Tip: Hover over a point in the score plot to view a prediction plot of the fitted curve for that level of the ID variable.

FPC Profiler Shows or hides a profiler of the FPC scores. The FPC Profiler includes a column for the input variable and a column for each FPC score. If a target function is specified, there are additional button options above the profiler graphs. You can optimize the target function and show or hide the target profilers. If you select Show Target Profilers, two additional profilers are added to the report. One measures the difference from the target function, and the other measures the integrated error from the target function. For more information about FPC Profiler red triangle menu options, see *Profilers*.

Tip: Use the Reset button to reset all of the FPC scores to 0 in the profiler.

Customize Number of FPC's Specifies the number of FPC scores to show in the Functional PCA. Specifying the number of FPC scores in this option also updates the Function Summaries report. To view the mean model, set the number of FPC's to 0.



Functional Data Explorer Platform Options

The Functional Data Explorer red triangle menu contains the following options:

Summaries A submenu of the following options for functional summary statistics:

Plot Mean Shows or hides a plot of the functional mean in the Summaries report. On by default.

Plot Standard Deviation Shows or hides a plot of the functional standard deviation in the Summaries report. On by default.

Plot Median Shows or hides a plot of the functional median in the Summaries report.

Models A submenu of the following model options:

B-Splines Fits a basis spline (B-Spline) model to the data. Use the B-Spline model for non-periodic data.

P-Splines Fits a penalized basis spline (P-Spline) model to the data.

Fourier Basis Fits a Fourier Basis model to the data. Use the Fourier Basis model for periodic data. A periodic model assumes that the function finishes where it starts. See [“Fourier Basis Model”](#).

Wavelets Fits several wavelets models to the data. A Wavelets model is a type of basis function model that is useful for data that contain a lot of peaks. This option requires data to be on an evenly spaced grid. If data is not evenly spaced, a grid is automatically created before the wavelet routine.

When the Wavelets option is selected, several wavelets models are fit from the following families of models: Haar, Daubechies, Symlet, Coiflet, and Biorthogonal. These are all flexible functions with different shapes and types of peaks based on the parameters. The model fit results from these models are shown in a table in the Model Selection report, ordered with the best fitting model at the top. See [“Wavelets Models”](#).

For more information about wavelets, see Nason ([2008](#)).

Note: If there are fewer than three unique input values, neither a Fourier basis model or a P-Spline model can be fit to the data, and a warning message appears.

Model Controls Shows a submenu that enables you to open the Model Controls panel prior to fitting a model. See [“Model Controls”](#). This option also enables the Model Selection red triangle menu option to change to the model selection criteria prior to fitting a model.

Direct Functional PCA (Not available if there is only a single function.) Performs functional principal components analysis directly on the data, without fitting a basis function model first. This reduces computation time, particularly for large data sets. The implementation of Direct Functional PCA is as follows:

1. Align the input data to be between 0 and 1 and interpolate the observations to a common grid of input values.
2. Perform functional principal components analysis on the data.
3. Smooth the first eigenfunction using a P-Spline model with a knot at each grid point.
4. Remove the first smoothed eigenfunction from the data and repeat [step 2](#) to [step 4](#) until a large amount of the variation in the data is explained.

Once you perform a Direct Functional PCA, a Functional PCA report is shown. See [“Functional PCA”](#).

Save Data Saves the processed data to a new data table. The processed data are saved in the stacked data format.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Functional Data Explorer Group Options

Functional Data Explorer Group options are available only when there are multiple functional processes specified in the Stacked Data Format tab in the launch window.

Summaries A submenu of options to show or hide plots of functional summary statistics for all functional processes. See [“Summaries”](#).

Data Processing A submenu of data processing options that can be applied to all functional processes. See [“Data Processing Options”](#).

Models A submenu of model options that can be fit to all functional processes. See [“Models”](#).

Save Data Saves the processed data for all functional processes to a new data table. The processed data are saved in the stacked data format.

Save Summaries Saves the summary statistics specified in each Function Summaries report to a new data table. The name of the new data table describes the model fit. This data table contains formula columns for the eigenfunctions, mean function, prediction function, and conditional prediction function for each of the Y variables. In the data table, there is a profiler script for each Y variable that launches the prediction profilers for the prediction and conditional prediction formulas. These formulas are functions of the input variable, the ID variable, and the eigenfunctions.

Arrange in Rows Specifies how many reports are displayed across the window.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Data Processing Options

The Data Processing red triangle menu in the Functional Data Explorer platform contains the following options:

Cleanup A submenu of the following data cleanup options:

Remove Zeros Removes observations with zero values. If there are no zeros in the data, an alert appears, indicating that no zero values were found.

Remove Value Displays a specifications window that enables you to specify a value to remove from the data.

Remove Selected Removes observations that correspond to rows that are selected in the data table.

Remove Unselected Removes observations that correspond to rows that are not selected in the data table.

Caution: Remove Selected and Remove Unselected remove the row numbers. When Auto Recalc is enabled, you must add or delete rows before using these options.

Filter X Removes X values that fall outside of a specified interval. When you select the Filter X option, you must specify Below and Above values. The X values that fall outside of the specified interval are not used for the analysis.

Filter Y Removes Y values that fall outside of a specified interval. When you select the Filter Y option, you must specify Below and Above values. The Y values that fall outside of the specified interval are not used for the analysis.

Reduce Reduces the data over the X values using one of the following techniques:

- Use the Grid tab to interpolate observations to a common grid of values. You can specify the grid size. By default, the grid size is the number of values in the longest function. This is also the maximum allowable grid size.

- Use the Bin tab to create a specified number of bins that are evenly spaced over the unique X values. For each function (or level of the ID, Function variable), the observations within a bin are averaged to produce a Y value for the corresponding bin level.

- Use the Thin tab to remove every N observation over the X values, where N is determined by the specified thinning rate. This is done for each function (or level of the ID, Function variable). By default, the thinning rate is 2, which removes half of the observations in each function.

Note: The Remove options exclude the specified observations from the analysis and modeling reports, but the observations remain unchanged in the data table.

Transform A submenu of the following options to transform the data:

Center Centers the output.

Standardize Standardizes the output by centering and scaling the data to have mean 0 and variance 1.

Range 0 to 1 Scales the output to lie within the range of 0 and 1.

Square Root Transforms the data by computing the square root of the output. The output values must be nonnegative.

Square Transforms the data by computing the square of the output.

Log Transforms the data by computing the natural logarithm of the output.

Exp Transforms the data by computing the exponential function of the output.

Negation Transforms the data by negating the output.

Logit Transforms the data by computing the logit function of the output. The output values must be between 0 and 1.

Log X Transforms the data by computing the natural logarithm of the input.

Align A submenu of the following options to align the input data:

Row Alignment Replaces the input values with the row number.

Align Maximum Aligns the functions using the observed maximum output value for each ID level. The input value associated with the observed maximum output value is set to zero for each ID level and the other input values are shifted up or down based on the difference between the observed maximum and zero.

Align Minimum Aligns the functions using the observed minimum output value for each ID level. The input value associated with the observed minimum output value is set to zero for each ID level and the other input values are shifted up or down based on the difference between the observed minimum and zero.

Align 0 to 1 Aligns the output functions such that the range of the input values is 0 to 1.

Tip: Align 0 to 1 is particularly useful when you fit a P-Spline model.

Align by Function Aligns the output functions such that each function starts at the overall minimum of the input values and ends at the overall maximum of the input values.

Dynamic Time Warping (Available only when there is more than one function.) Aligns the output functions using dynamic time warping (DTW). DTW is a function alignment technique that finds an optimal warping to align two or more functions together. When you select the DTW option, a Select Reference Function window appears. Use this to select the reference function. The reference function is the function that the remaining functions are aligned to.

Once you select a reference function and click OK, a warping function plot is shown along with a list for the remaining query functions. On the warping function plot, the reference function is on the y-axis and the selected query function is on the x-axis. Deviations from the red diagonal line ($y = x$) indicate that the inputs of the query function have been warped for better alignment.

Spectral A submenu of the following options that are useful for spectral data:

SNV Applies the Standard Normal Variate method to the data. This method standardizes the output by centering and scaling each individual function (level of the ID variable) to have a mean of 0 and a standard deviation of 1.

MSC Applies the Multiplicative Scatter Correction to the data. A simple linear regression is fit for each individual function (level of the ID variable) where the response is output values for the function and the regressor is the output values for the mean function. The original output values, y_{it} , are then replaced by new values, y_{it}^* , using the following equation:

$$y_{it}^* = \frac{y_{it} - \bar{y}_t}{b_i}$$

where b_i is the slope obtained from the simple linear regression for function i . For more information, see Geladi et al (1985).

Savitzky-Golay Provides options to use the Savitzky-Golay method. See Savitzky and Golay (1964).

Note: All options involving the Savitzky-Golay method require that the input data be on an evenly spaced grid and that at least one function contains 7 or more data points. If the data is not on an evenly spaced grid, it is automatically placed on an evenly spaced grid when you select a Savitzky-Golay option.

Filter Applies a Savitzky-Golay filter to the data. This method fits local polynomials to several collections of points across the domain. The polynomials are fit using least squares and the number of points in each fit is determined by the bandwidth. When you select this option, several fits are made for polynomials of order 0, 1, and 2 and bandwidths up to 10. The best fitting models for each function are selected based on the AIC. The order of the polynomial and the bandwidth can be different for each function.

First Derivative Applies a Savitzky-Golay filter to the data using only polynomials of order 2 or 3 and then takes the first derivative. Since the filter fits polynomials, the derivatives are computed analytically.

Second Derivative Applies a Savitzky-Golay filter to the data using only polynomials of order 3 and then takes the second derivative. Since the filter fits polynomials, the derivatives are computed analytically.

Baseline Correction Subtracts a baseline function from each individual function. A baseline correction is used when there is a known trend, or baseline, that you want to remove. For example, this could be due to an artifact of how the data is measured. Usually, the information is in the peaks of the data, so these regions are not included in the baseline model.

When you select this option, a baseline correction window is shown. This window contains a selection plot that displays the data and a set of options to specify the baseline model. The baseline correction window contains the following options:

Baseline Model Specifies the type of model for the baseline function. You can specify a linear, quadratic, cubic, two parameter exponential, or three parameter exponential model.

Correction Region Specifies the region that the baseline function is subtracted from. You can subtract the baseline from the entire function region or from only the regions that were used to construct the baseline model.

Baseline Regions Adds or removes a pair of blue vertical lines to the selection plot. The lines are initially on top of one another. Move the lines to specify regions of the data that you do not want to include in the baseline model. The region of the data that falls in between a pair of blue lines is not included in the baseline model.

Anchor Points Adds or removes a red vertical line to the selection plot. This line specifies data points that are forced into the baseline model.

Target Functions (Available only when there is more than one function.) Enables you to load a target function.

Load Targets Shows a window that enables you to specify a target function. A target function is used for curve matching, where it is desirable for all of the functions to look like the target function, also known as a reference function or a golden curve.

If you specify a target function, the data from the function is not used in model fitting. When you specify a target function, there are additional options added to the FPC Profiler. See [“FPC Profiler”](#).

Note: A target function must be loaded before any other preprocessing steps are performed.

Dynamic Time Warping Options

Plot Warping Functions Shows or hides the warping function plot. On by default.

Save Distance Matrix Saves the distance matrix to a separate data table. The distance matrix can be useful for clustering the functions. The distance matrix data table contains a hierarchical clustering script.

Save Warping Functions Saves the warping functions to a separate data table. Each row of the data table contains the DTW adjusted input variable, the original input variable, and the ID variable.

Model Report Options

When you fit a model to the data in the Functional Data Explorer platform, a model report appears. The red triangle menu for a model report contains the following options:

Model Selection Displays a submenu of choices to use as the model selection criteria. The choices are AICc, BIC, and GCV. See *Fitting Linear Models*.

Plot Basis (Not available for Wavelets or Direct Functional PCA.) Shows or hides a plot of all the basis functions on one graph.

Diagnostic Plots (Not available for Wavelets or Direct Functional PCA.) Shows or hides the Diagnostics Plots report. See [“Diagnostic Plots”](#)

Function Summaries Shows or hides the Function Summaries report. See [“Function Summaries”](#).

Basis Function Coefficients (Not available for Direct Functional PCA.) Shows or hides the coefficients report for the corresponding model fit. See [“Coefficients”](#).

Random Coefficients by Function (Not available for Wavelets or Direct Functional PCA.)

Shows or hides the Random Coefficients by Function report. See [“Random Coefficients by Function”](#).

Functional PCA Shows or hides the Functional PCA report. See [“Functional PCA”](#).

Wavelets DOE Analysis (Available only if at least one supplementary variable is specified in the launch window and a wavelet model is fit to the data.) Launches a Generalized Regression report within the FDE platform. Generalized regression models are fit to the wavelet coefficients using the supplementary variables as model effects. Since there can be many wavelet coefficients, only a subset of the coefficients that have the most impact are fit. This subset of coefficients is determined by the energy of each wavelet coefficient. The energy for a wavelet coefficient is calculated as follows:

$$E_{j,k} = \sum_{i=1}^N d_{j,k,i}^2$$

where j = the resolution, k = the location in the input domain, i = the function, and d = the wavelet coefficient for (j,k,i) . The total energy is the sum of all the energy values for the coefficients:

$$E = \sum_j \sum_k E_{j,k}$$

The energy values are used in the Energy table that is shown in the Wavelets DOE Analysis report. The Energy table contains the following columns

Coefficient The wavelet coefficient, identified by the resolution and location in the input domain.

Energy The relative energy for the corresponding wavelet coefficient. The relative energy is the energy value standardized by the maximum energy across all wavelet coefficients. This is calculated as $E_{j,k}/\max_{j,k}(E_{j,k})$.

Percent The percent of energy for the corresponding wavelet coefficient compared to the total energy. This describes how large the coefficient is relative to the sum of all of the coefficients.

Cumulative The cumulative percent of the total energy.

The coefficients that cumulatively explain at least 90% of the total energy are shown in the Energy table.

A generalized regression model is fit for the father wavelet and for each wavelet coefficient in the Energy table. By default, a two degree factorial model is fit and the Estimation Method is Best Subset. Note that categorical quadratic effects are not included. If the number of terms in the model is greater than 20 or the number of functions is greater than

1000, the Estimation Method automatically switches to Pruned Forward. Alternatively, you can specify a model script in the original data table that defines the desired model fit.

Modeling the individual wavelet coefficients using the supplementary variables enables you see how the supplementary values effect the location of peaks around a particular input value. This is because each wavelet coefficient is directly related to a specific location on the input domain. You can then use the FDOE Profiler to explore how the supplementary variables affect the response.

Tip: In the unlikely event that the Wavelets DOE Analysis option fails to load, use the Standardize pre-processing step on the data, refit the wavelets models on the centered and scaled data, and then try launching the Wavelets DOE Analysis.

The Wavelets DOE Analysis report contains the following red triangle menu options:

Generalized Regression for Wavelets Coefficients Shows or hides the Generalized Regression reports for each wavelet coefficient. For more information on Generalized Regression model reports, see *Fitting Linear Models*.

FDOE Profiler Shows or hides the FDOE Profiler, which enables you to explore how the response changes based on the supplementary variables. For more information about the FDOE Profiler red triangle menu options, see *Profiler*.

Note: The FDOE Profiler in the Wavelets DOE Analysis report does not support target functions. If you wish to examine target functions, use the Functional DOE Analysis option instead.

Save Columns Shows a list with the options Save Prediction Formula and Save Residual Formula. These options save the corresponding formula to a new column in the data table.

Functional DOE Analysis (Available only if at least one supplementary variable is specified in the launch window.) Launches a Generalized Regression report within the FDE platform. A generalized regression model is fit to each of the FPC score functions using the supplementary variables as model effects. By default, a two degree factorial model is fit and the Estimation Method is Best Subset. Note that categorical quadratic effects are not included. If the number of terms in the model is greater than 20 or the number of functions is greater than 1000, the Estimation Method automatically switches to Pruned Forward. Alternatively, you can specify a model script in the original data table that defines the desired model fit. Modeling the FPC scores using the supplementary variables enables you to use the model fit to determine how the response changes based on the supplementary variables. Use the FDOE Profiler to explore how the supplementary variables affect the response.

The Functional DOE Analysis report contains the following red triangle menu options:

Generalized Regression for FPC Scores Shows or hides the Generalized Regression reports for each FPC score. For more information on Generalized Regression model reports, see *Fitting Linear Models*.

Diagnostic Plots Shows or hides actual by predicted and residual plots for the response variable.

FDOE Profiler Shows or hides the FDOE Profiler, which enables you to explore how the response changes based on the supplementary variables. If a target function is specified, there are additional button options above the profiler graphs. You can optimize the target function and show or hide the target profilers. If you select Show Target Profilers, two additional profilers are added to the report. One measures the difference from the target function, and the other measures the integrated error from the target function. For more information about the FDOE Profiler red triangle menu options, see *Profilers*.

Save Columns Shows a list with the options Save Prediction Formula and Save Residual Formula. These options save the corresponding formula to a new column in the data table.

Save Data Saves the modeled data to a new data table. The modeled data are saved in the stacked data format.

Remove Fit Removes the model report for the specified fit.

Additional Examples of the Functional Data Explorer Platform

This section contains examples using the Functional Data Explorer platform.

- [“Example for Multiple Functional Processes”](#)
- [“Example of Functional DOE Analysis”](#)

Example for Multiple Functional Processes

This example uses the Functional Data Explorer platform to fit models to the data and save functional principal components to a new data table. The functional principal components are then analyzed in the Generalized Regression personality of the Fit Model platform.

You want to analyze enzyme production. Yield is the amount of an enzyme produced by genetically modified yeast. There are 100 process measurements per batch that were taken at equally spaced times over a 12-hour period.

Fit Functional Models

1. Select **Help > Sample Data Folder** and open Functional Data/Fermentation Process.jmp.
2. Select **Analyze > Specialized Modeling > Functional Data Explorer**.
3. In the Stacked Data Format tab, select Ethanol through pH and click **Y, Output**.
4. Select Time and click **X, Input**.
5. Select BatchID and click **ID, Function**.
6. Click **OK**.
7. Click the Functional Data Explorer Group red triangle and select **Data Processing > Align > Align 0 to 1**. This aligns the input variable to be between 0 and 1 in each Functional Data Explorer report.
8. Click the Functional Data Explorer Group red triangle and select **Models > B-Splines**. This fits a B-spline model to each of the functional processes.

Figure 16.6 and Figure 16.7 show the model reports for one of the functional process variables, Ethanol. Scroll through the full report to view the models fit for each of the process variables. Next, use the FPCs in the Function Summaries report in an analysis.

Save FPCs and Link Yield Results

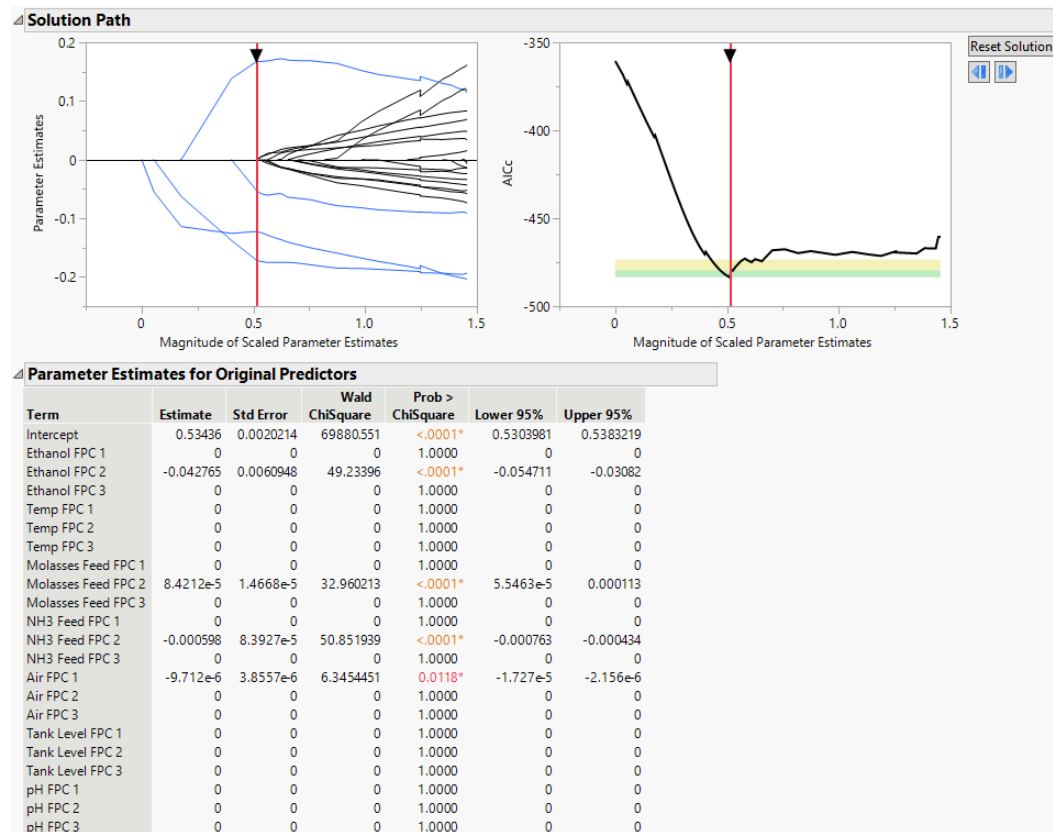
1. Press Ctrl, click any Function Summaries red triangle, and select **Customize Function Summaries**.
2. In the box next to **Enter number of FPCs to show**, type 3.
3. Click the **Deselect All Summaries** box.
4. Click **OK**.
5. Click the Functional Data Explorer Group red triangle and select **Save Summaries**.
6. Select **Help > Sample Data Folder** and open Functional Data/Fermentation Process Batch Yield Results.jmp.
7. In the Functional Data Explorer Model Summaries.jmp data table, right-click BatchID and deselect **Link ID**.
8. In the Functional Data Explorer Model Summaries.jmp data table, right-click BatchID and select **Link Reference > Fermentation Process Batch Yield Results.jmp**.

This virtually joins the yield data table and the summaries data table.

Fit a Generalized Regression Model

Use the Generalized Regression personality of the Fit Model platform to determine how Yield is affected by the functional process variables.

1. In the Functional Data Explorer Model Summaries.jmp data table, select **Analyze > Fit Model**.
2. Click the triangle next to referenced by BatchID to Fermentation Process Batch Yield Results.
3. Select Yield[BatchID] and click **Y**.
4. Select the remaining columns, except Time and BatchID, and click **Add**.
5. Change the Personality to Generalized Regression.
6. Click **Run**.
7. Select the **Adaptive** box.
8. Click **Go**.

Figure 16.8 Generalized Regression Report for Batch Yield

The Generalized Regression report shows that Yield is significantly affected by certain components of Ethanol, Molasses Feed, NH3 Feed, and Air. The RSquare for the model is 0.73225. By using FDE to perform dimension reduction on the functional processes first, you greatly reduce the number of variables, while still retaining the ability to build a reasonable prediction models.

Example of Functional DOE Analysis

This example uses the Functional Data Explorer to see how combinations of ingredients effect the homogeneity grade of a compound across different temperatures. The three ingredients that form the compound are the primary ingredient labeled **Active**, **Water**, and **Solvent**. There are 32 combinations of ingredients; these combinations are denoted by the values in the **Formulation** column. For each formulation, the **Homogeneity Grade** is measured from 20 to 70 degrees Celsius at 5 degree intervals. By plotting these measurements across the different temperatures, you can form a curve for each formulation. Use the Functional Data Explorer platform to determine how the curves change depending on the values of **Active**, **Water**, and **Solvent**.

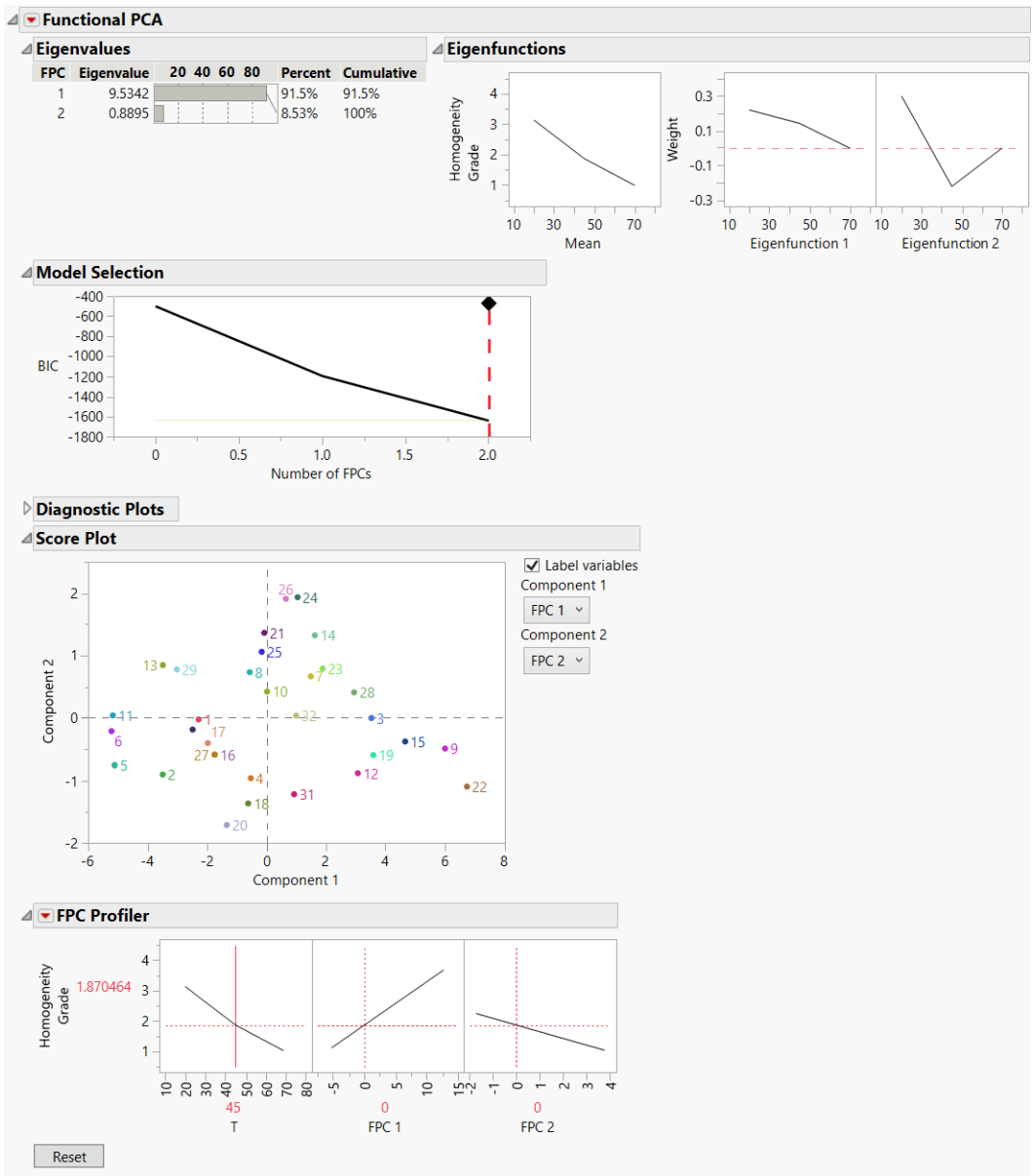
1. Select **Help > Sample Data Folder** and open **Functional Data/Formulation For Homogeneity DOE.jmp**.
2. Select **Analyze > Specialized Modeling > Functional Data Explorer**.
3. In the **Stacked Data Format** tab, select **Homogeneity Grade** and click **Y, Output**.
4. Select **T** and click **X, Input**.
5. Select **Formulation** and click **ID, Function**.
6. Select **Solvent**, **Active**, and **Water** and click **Z, Supplementary**.

These variables are not used in the initial functional data analysis. However, by specifying them here as supplementary variables, you can use them later when fitting models.

7. Click **OK**.
8. Click the Functional Data Explorer red triangle and select **Models > B-Splines**.

The best fitting model is a linear spline model with one knot.

Figure 16.9 Functional PCA for Homogeneity Grade

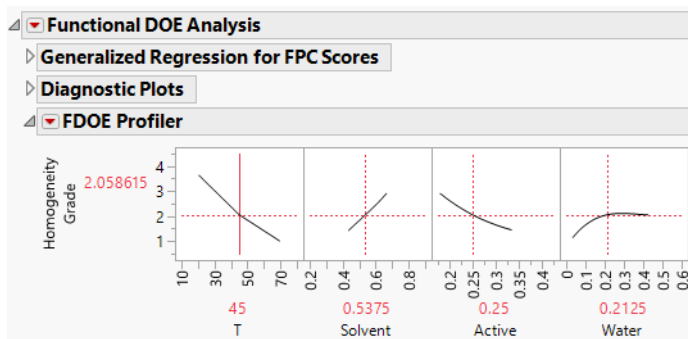


- The Functional PCA report shows that two functional principal components are needed to explain the variability in the data. The FPC Profiler shows how Homogeneity Grade is affected by the FPCs across the different temperatures. However, there is no way to tell how the actual ingredients affect the Homogeneity Grade.
9. Click the red triangle next to B-Spline on Initial data and select **Functional DOE Analysis**.

The Functional DOE Analysis option fits a generalized regression model to each principal component score using the three supplementary variables as predictors. Because the data table contains a Model script, the models that are fit are based on the model defined in the script.

Note: If your data table does not contain a Model script, a two degree factorial model is fit.

Figure 16.10 FDOE Profiler for Homogeneity Grade



The FDOE Profiler enables you to explore how Homogeneity Grade changes based on the predictors you are actually interested in. In general, Homogeneity Grade seems to decrease as the amount of Active increases.

Note: In this example, the values of Active, Solvent, and Water are constrained to sum to 1. This constraint is reflected in the profiler.

JMP PRO Statistical Details for the Functional Data Explorer Platform

This section contains statistical details for the Functional Data Explorer platform.

- [“Statistical Details for Functional Model Fits”](#)
- [“Statistical Details for the Function Summaries Report”](#)

Statistical Details for Functional Model Fits

All of the model fits in the Functional Data Explorer platform rely on basis function expansion. Basis functions are a set of independent functions. Any function, $f(t)$, can be approximated by taking a linear combination of K basis functions, denoted as ϕ_k . In general, a function is approximated as follows:

$$\hat{f}(t) = \sum_{k=1}^K c_k \phi_k(t)$$

where the c_k parameters are the basis coefficients. The amount of smoothing is determined by the number of basis functions, K . For more information about basis function expansion, see Ramsay and Silverman (2005).

JMP[®] PRO Fourier Basis Model

The $k = 1, \dots, K$ functions that define the Fourier basis are defined such that $\phi_0 = 1$, $\phi_{2k-1} = \sin(r\omega t)$, and $\phi_{2k} = \cos(r\omega t)$. Then, the approximated function is defined as follows:

$$\hat{f}_i(t) = c_{i0} + c_{i1} \sin \omega t + c_{i2} \cos \omega t + c_{i3} \sin 2\omega t + c_{i4} \cos 2\omega t + \dots$$

This approximates a periodic function with period A defined as $A = 2\pi/\omega$. The coefficients are a combination of fixed basis function coefficients and random coefficients by function. Each c is defined as follows:

$$c_{ik} = \beta_k + \alpha_{ik}$$

where β_k is the fixed coefficient for basis function k and α_{ik} is the random coefficient for basis function k for a specific functional process i . For the Fourier Basis model, the estimates for the β_k and α_{ik} parameters are found in the Basis Function Coefficients table and Random Coefficients by Function table, respectively.

Fourier basis models have an intercept term and an equal number of sine and cosine terms, which are referred to as Fourier pairs. Therefore, K is always odd. For example, if $K = 7$, there is an intercept term and three Fourier pairs.

Statistical Details for the Function Summaries Report

The Function Summaries report in the Functional Data Explorer platform includes the following summary statistics: integrated difference, root integrated square error (RISE), and root integrated function squared (RIFS). These summary statistics are defined in this section. For all equations in this section, it is assumed that the input values have been aligned between 0 and 1. The following notation is used for all functions:

$\hat{f}_i(x)$ is the estimated function for curve i at input value x

$\hat{m}(x)$ is the estimated mean function at input value x

Integrated Difference

The integrated difference is the average difference between points on the ID specific curve and the overall mean curve. It can be used to determine if the ID specific curve falls above or below the mean curve, on average. The integrated difference for curve i is defined as follows:

$$ID_i = \int_0^1 (\hat{f}_i(x) - \hat{m}(x)) dx$$

RISE

Root integrated square error (RISE) is the average distance between the ID specific curve and the overall mean curve. It can be used to determine which ID specific curves are most similar or most different from the overall mean curve. The RISE summary value for curve i is defined as follows:

$$RISE_i = \sqrt{\int_0^1 (\hat{f}_i(x) - \hat{m}(x))^2 dx}$$

RIFS

Root integrated function squared (RIFS) is a summary value for analyzing optimal curve scenarios. If you pre-subtract off an optimal curve, you can use the smallest values of RIFS to determine which functions are closest to the optimal curve. The RIFS summary value for curve i is defined as follows:

$$RIFS_i = \sqrt{\int_0^1 \hat{f}_i(x)^2 dx}$$

Chapter 17

Gaussian Process

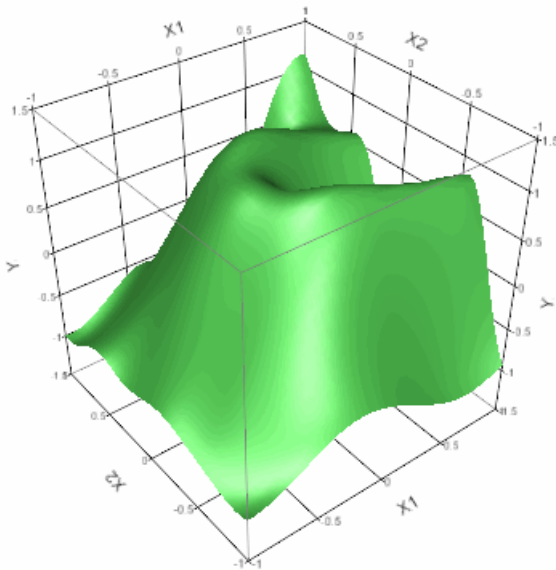
Fit Data Using Smoothing Models

Use the Gaussian Process platform to model the relationship between a continuous response and one or more predictors. These types of models are common in computer simulation experiments, such as the output of finite element codes, and they often perfectly interpolate the data. Gaussian processes can deal with these no-error-term models, in which the same input values always results in the same output value.

The Gaussian Process platform fits a spatial correlation model to the data. The correlation of the response between two observations decreases as the values of the independent variables become more distant.

One purpose for using this platform is to obtain a prediction formula that can be used for further analysis and optimization.

Figure 17.1 Gaussian Process Prediction Surface Example



Contents

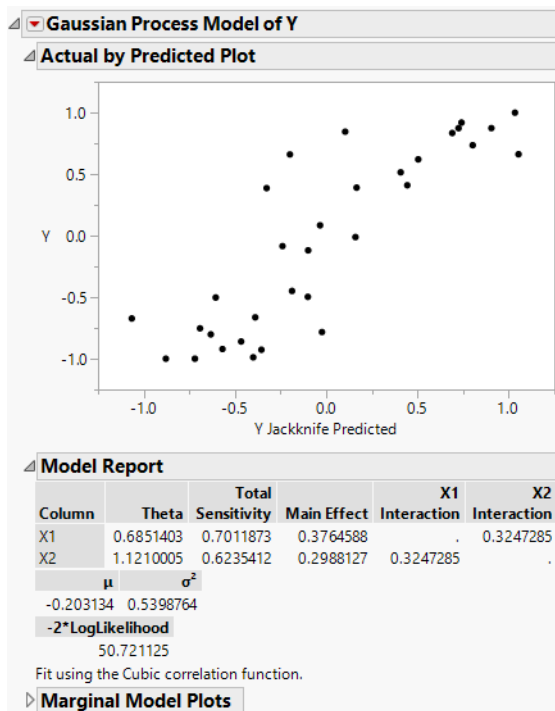
Example of the Gaussian Process Platform	339
Launch the Gaussian Process Platform	341
The Gaussian Process Report	342
Actual by Predicted Plot	342
Model Report	342
Marginal Model Plots	343
Gaussian Process Platform Options	343
Additional Examples of the Gaussian Process Platform	344
Example of a Gaussian Process Model	345
Example of a Gaussian Process Model with Categorical Predictors	346
Statistical Details for the Gaussian Process Platform	348
Statistical Details for Models with Continuous Predictors	348
Statistical Details for Models with Categorical Predictors	349
Statistical Details for Variance Formula Parameterization	350
Statistical Details for the Model Fit	350

Example of the Gaussian Process Platform

This example uses the Gaussian Process platform to find the exploratory power of two variables on a response. The data are from a space filling design in two variables with a deterministic equation for the response. You can view the equation for the response in the column formula.

1. Select **Help > Sample Data Folder** and open 2D Gaussian Process Example.jmp.
2. Select **Analyze > Specialized Modeling > Gaussian Process**.
3. Select X1 and X2 and click **X**.
4. Select Y and click **Y**
5. Select **Correlation Type > Cubic**.
6.  Deselect **Fast GASP**.
7. Click **OK**.

Figure 17.2 Gaussian Process Report

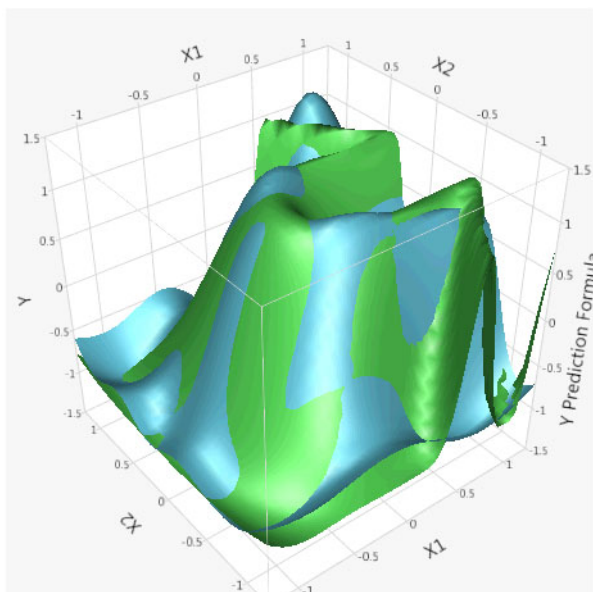


Note: The estimated parameters can be different due to different starting points in the minimization routine, the choice of correlation type, and the inclusion of a nugget parameter.

Now, visualize the fitted surface compared to the original surface.

8. Click the red triangle next to Gaussian Process Model of Y and select **Save Prediction Formula**.
9. Select **Graph > Surface Plot**.
10. Select X1 through Y Prediction Formula and click **Columns**.
11. Click **OK**.
12. In the Surface column, select **Both sides** for the Y Prediction Formula.

Figure 17.3 3D Surface Plot of the Actual and Predicted Ys

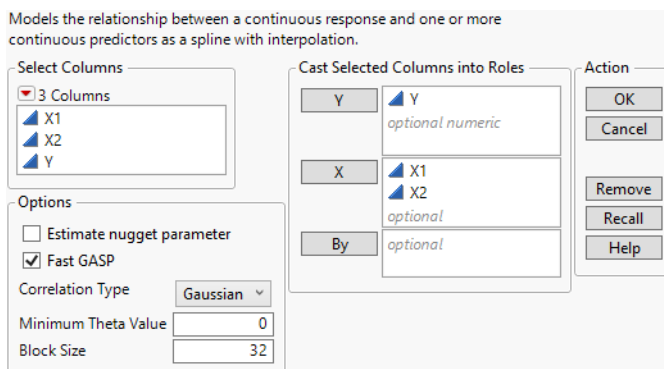


The two surfaces are similar. The impact of X1 and X2 on the response Y can be visualized. You can rotate the plot to view it from different angles. Marginal plots are another tool to use to understand the impact of the factors on the response.

Launch the Gaussian Process Platform

Launch the Gaussian Process platform by selecting **Analyze > Specialized Modeling > Gaussian Process**.

Figure 17.4 Gaussian Process Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y Assigns the continuous columns to analyze.

X Assigns the columns to use as explanatory variables. Categorical variables are allowed in JMP Pro when the Fast GASP option is specified.

Estimate Nugget Parameter Introduces a ridge parameter into the estimation procedure. A ridge parameter is useful if there is noise or randomness in the response, and you want the prediction model to smooth over the noise instead of perfectly interpolating.

JMP PRO Fast GASP Option to use the Fast GASP algorithm. Fast GASP breaks the Gaussian process model into small pieces (called blocks) to speed computation time. Blocks allow for the use of multiple CPUs and parallel processing.

Note: When there are more than 2,500 observations, the Fast GASP algorithm is required.

For more information about Fast GASP, see Parker (2015).

Correlation Type Choose the correlation structure for the model. The platform fits a spatial correlation model to the data, where the correlation of the response between two observations decreases as the values of the independent variables become more distant.

Gaussian Restricts the correlation between two points to always be nonzero, no matter the distance between the points.

Cubic Allows the correlation between two points to be zero for points that are far enough apart. This method is a generalization of a cubic spline.

JMP PRO The Fast GASP algorithm does not support the cubic correlation function.

Minimum Theta Value Sets the minimum theta value to use in the fitted model. The default is 0. The theta values are analogous to a slope parameter in regular regression models. Small theta values indicate that a variable has little influence on the predicted values.

JMP PRO Block Size The number of observations in each computational block used by the Fast GASP algorithm. There must be at least 25 observations per block and a maximum of the number of rows in the data set up to a maximum of 2,500.

The Gaussian Process Report

The Gaussian Process report shows an actual by predicted plot, a model report, and marginal model plots.

- [“Actual by Predicted Plot”](#)
- [“Model Report”](#)
- [“Marginal Model Plots”](#)

Actual by Predicted Plot

The Actual by Predicted plot in the Gaussian Process report shows the actual Y values on the Y axis and the jackknife predicted values on the X axis. One measure of goodness-of-fit is how well the points lie along the diagonal ($Y = X$) of the plot.

The jackknife values are not true jackknife values in that the model is not re-fit with the associated row for each Y excluded. Rather, the row is excluded from the prediction model for each associated Y but the correlation parameters retain the contribution of the row in them. For Gaussian processes that perfectly interpolate the data this jackknife procedure provides predictions that are not equal to the input.

Model Report

The Model Report in the Gaussian Process platform shows a functional ANOVA table for the model parameter estimates. Specifically, it is an analysis of variance table where the variation is computed using a function-driven method.

Theta Gaussian Process model parameter estimates. See [“Statistical Details for the Gaussian Process Platform”](#).

Total Sensitivity Sum of the main effect and all interaction terms for each factor. It is a measure of the amount of influence a factor and all its two-way interactions have on the response variable.

Total variation is the integrated variability over the entire experimental space.

Main Effect The functional main effect of each factor is the integrated total variation due to that factor alone. The main effect is the ratio of the functional effect and the total variation for each factor in the model.

Interactions Functional interaction effects are computed in a similar way to main effects.

JMP PRO Categorical Input When the model includes categorical factors, a correlation matrix for each categorical factor is provided. The off-diagonal entries correspond to Gaussian Process model parameter estimates. See [“Statistical Details for Models with Categorical Predictors”](#).

Mu and Sigma² The mean and variance model parameters.

Nugget The estimated nugget value. A nugget value is reported if you selected **estimate nugget parameter** in the Gaussian Process launch window. A nugget value is also reported if JMP has added a nugget parameter in order to avoid a singular covariance matrix.

-2LogLikelihood The estimated value of twice the negative log-likelihood function. See *Fitting Linear Models*.

Marginal Model Plots

In the Gaussian Process report, a marginal plot appears for each factor in the model. For each plot, all factors except one are integrated out using a distribution that is uniform over the ranges of the other factors in the data. The marginal prediction of the response for the remaining factor of interest is drawn in blue. The data points for the observed response values are included to show how well the marginal model fits the data. For models with a small number of factors, these plots can be used as diagnostic plots.

Gaussian Process Platform Options

The Gaussian Process red triangle menu contains the following options:

Profiler Opens the standard Profiler. See *Profilers*.

Contour Profiler Opens the Contour Profiler. See *Profilers*.

Surface Profiler Opens the Surface Profiler. See *Profilers*.

Save Prediction Formula Creates a new prediction formula column in the active data table.

Save Variance Formula Creates a new variance formula column in the active data table.



Publish Prediction Formula Creates a prediction formula and saves it as a formula column script in the Formula Depot platform. If a Formula Depot report is not open, this option creates a Formula Depot report. See [“Formula Depot”](#).



Publish Variance Formula Creates a variance 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 [“Formula Depot”](#).

Save Jackknife Predicted Values Saves the jackknife predicted values to the active data table. These are the X-axis values for the Actual by Predicted Plot.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

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

Additional Examples of the Gaussian Process Platform

This section contains examples using the Gaussian Process platform.

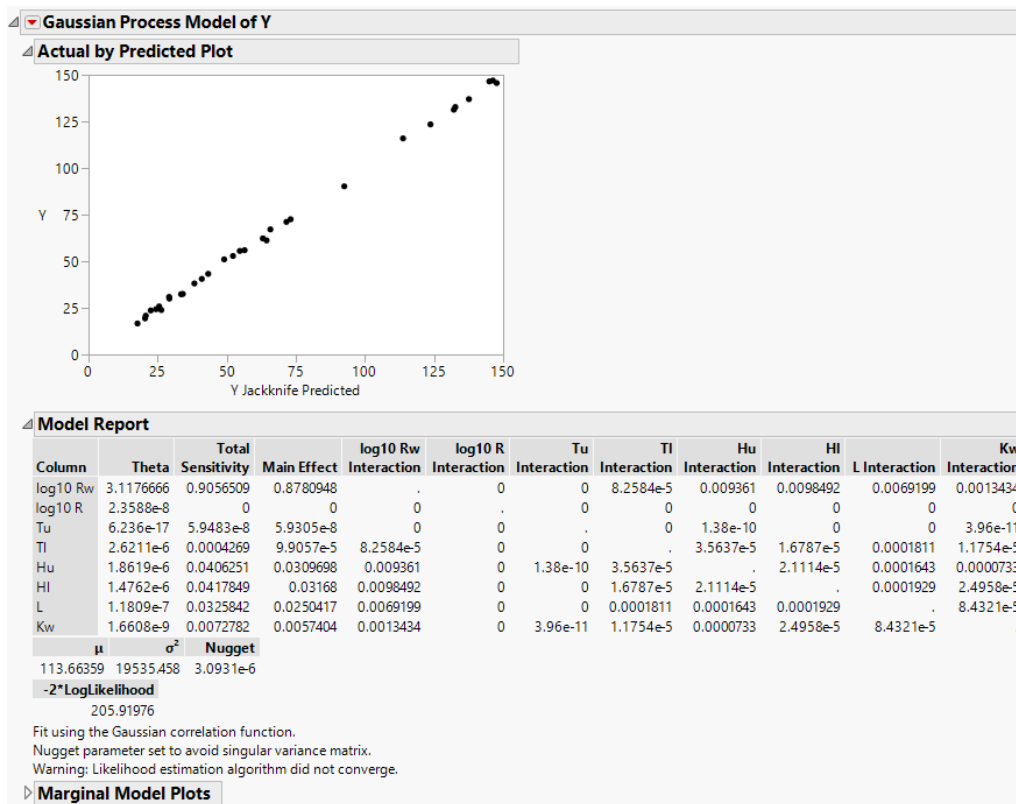
- [“Example of a Gaussian Process Model”](#)
- [“Example of a Gaussian Process Model with Categorical Predictors”](#)

Example of a Gaussian Process Model

Given a specified engineering model, use a Gaussian process model to understand the impact of factors included in the model on the response. The data demonstrate the flow of water through a Borehole that is drilled from the ground surface through two aquifers.

1. Select **Help > Sample Data Folder** and open Design Experiment/Borehole Latin Hypercube.jmp.
2. Select **Analyze > Specialized Modeling > Gaussian Process**.
3. Select log10 Rw through Kw and click **X**.
4. Select Y and click **Y**.
5.  In JMP Pro, to run the analysis faster, leave the Fast GASP checked.
6. Click **OK**.

Figure 17.5 Borehole Latin Hypercube Report



The data on the actual by predicted plot fall along the $Y = X$ line, indicating that the Gaussian process prediction model is a good approximation of the true function. In the

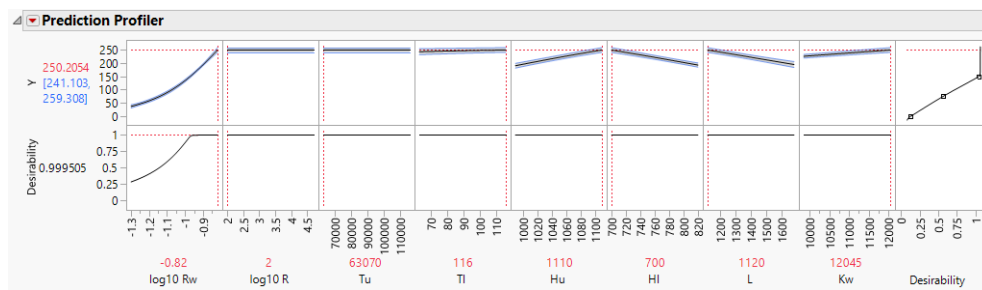
Model Report, you see that the first factor, log10 Rw, has the highest total sensitivity. The estimated total sensitivity for log10 Rw explains more than 90% of the variation in the response. Factors with small theta values have little (or no) impact on the prediction formula. Use the profiler to visualize the sensitivities.

- Click the Gaussian Process Model of Y red triangle and select **Profiler**.

The desirability functions automatically appear because the response, Y, has a Response Limits column property.

- Click the Prediction Profiler red triangle and select **Optimize and Desirability > Maximize Desirability**.

Figure 17.6 Gaussian Process Model Profiler



The desirability function is set to maximize the response Y because Y has goal set to Maximize in the Response Limits column property. Maximizing the desirability functions identifies the values of the factors that maximize the response. The shaded bands represent the 95% confidence intervals.

Note: Your estimates can differ from those shown in [Figure 17.5](#), which were found using the Fast GASP algorithm.

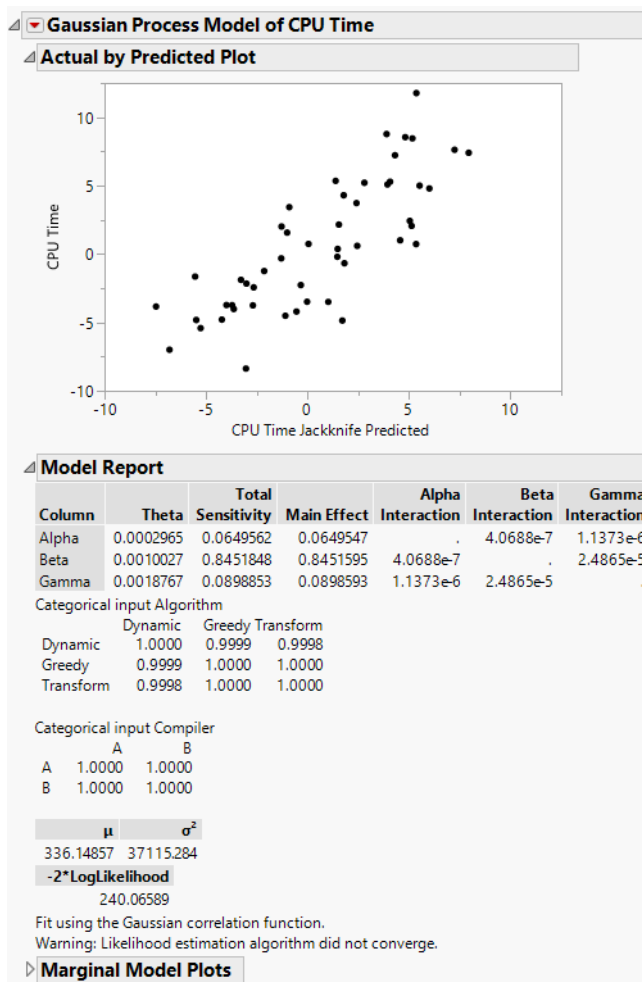
JMP PRO Example of a Gaussian Process Model with Categorical Predictors

Predict the response using a Gaussian Process model that contains both continuous and categorical factors. These data are simulated CPU times from a 50 run space filling designed experiment that has three continuous and two categorical factors.

- Select **Help > Sample Data Folder** and open Design Experiment/Algorithm Data.jmp.
- Select **Analyze > Specialized Modeling > Gaussian Process**.
- Select Alpha through Compiler and click **X**.
- Select CPU Time and click **Y**.
- JMP PRO** To run the analysis, leave the Fast GASP checked. Click **OK**.

Note: The Fast GASP option must be used for models that contain categorical factors. See [“Statistical Details for Models with Categorical Predictors”](#).

Figure 17.7 Algorithm Data Report



The actual by predicted plot shows a strong correlation between the actual and predicted CPU times. This is an indication that the Gaussian process prediction model is a good approximation of the true function. In the Model Report, the Beta predictor has the highest total sensitivity. This indicates that of the continuous predictors, Beta explains the most variation in the response. There is a separate Categorical Input matrix for each of the categorical predictors, Algorithm and Compiler. These matrices are correlation matrices and show the correlation between levels for each categorical predictor. The off-diagonals of the matrices are the τ parameters.

Statistical Details for the Gaussian Process Platform

This section contains statistical details for the Gaussian Process platform.

- [“Statistical Details for Models with Continuous Predictors”](#)
- [“Statistical Details for Models with Categorical Predictors”](#)
- [“Statistical Details for Variance Formula Parameterization”](#)
- [“Statistical Details for the Model Fit”](#)

Statistical Details for Models with Continuous Predictors

If the Gaussian Process model contains only continuous predictors, the Gaussian Process platform implements two possible correlation structures, the Gaussian and the Cubic.

The Gaussian correlation structure uses the product exponential correlation function with a power of 2 as the estimated model. This model assumes that Y is normally distributed with mean μ and covariance matrix $\sigma^2 \mathbf{R}$. The elements of the $\mathbf{R}$ matrix are defined as follows:

$$r_{ij} = \exp \left(- \sum_{k=1}^K \theta_k (x_{ik} - x_{jk})^2 \right)$$

where

K = # of continuous predictors

θ_k = theta parameter for the k^{th} predictor

x_{ik} = the value of the k^{th} predictor for subject i

x_{jk} = the value of the k^{th} predictor for subject j

The Cubic correlation structure also assumes that Y is normally distributed with mean μ and covariance matrix $\sigma^2 \mathbf{R}$. The $\mathbf{R}$ matrix consists of the following elements:

$$r_{ij} = \prod_k \rho(d; \theta_k)$$

where

$$d = x_{ik} - x_{jk}$$

$$\rho(d;\theta) = \begin{cases} 1 - 6(d\theta)^2 + 6(|d|\theta)^3, & |d| \leq \frac{1}{2\theta} \\ 2(1 - |d|\theta)^3, & \frac{1}{2\theta} < |d| \leq \frac{1}{\theta} \\ 0, & \frac{1}{\theta} < |d| \end{cases}$$

See Santer (2003). The theta parameter used in the Cubic correlation structure is the reciprocal of the parameter often used in the literature. The reciprocal is used so that when theta has no effect on the model, then rho has a value of zero, rather than infinity.

Statistical Details for Models with Categorical Predictors

If the Gaussian Process model includes categorical predictors, the Gaussian correlation structure is used for the correlation structure. The elements of the **R** matrix are defined as follows:

$$r_{ij} = \left(\prod_{p=1}^P \tau_{p_{ij}} \right) \exp \left(- \sum_{k=1}^K \theta_k (x_{ik} - x_{jk})^2 \right)$$

where

K = # of continuous predictors

P = # of categorical predictors

θ_k = theta parameter for the k^{th} continuous predictor

x_{ik} = the value of the k^{th} continuous predictor for subject i

x_{jk} = the value of the k^{th} continuous predictor for subject j

$\tau_{p_{ij}}$ = the correlation between the observed level of predictor p for subject i and the observed level of predictor p for subject j

There is a τ parameter for each combination of levels of a categorical variable, where τ_{ij} corresponds to the unique combination formed by the observed levels of subject i and subject j . Thus, the covariance element, r_{ij} , depends on the combination of levels of the categorical predictors obtained from the i^{th} and j^{th} observations. See Qian et al. (2012).

Statistical Details for Variance Formula Parameterization

In the Gaussian Process platform, the saved variance formula uses the previously defined parameterization of $\mathbf{R}$, except when the model includes categorical predictors. When the Gaussian Process model includes categorical predictors, the saved variance formula uses the following parameterization of $\mathbf{R}$:

$$r_{ij} = \exp \left(- \sum_{k=1}^K \theta_k (x_{ik} - x_{jk})^2 - \sum_{p=1}^P \phi_{p_{ij}} \right)$$

where $\phi_{p_{ij}} = -\ln(\tau_{p_{ij}})$ and all other variables are as previously defined.

Statistical Details for the Model Fit

The model parameters of a Gaussian Process model are fit via maximum likelihood. The fitted parameters are provided in the platform report. These are the parameters:

- μ is the Gaussian Process mean,
- σ^2 is the Gaussian Process variance,
- Theta corresponds to the values of θ_k in the definition of $\mathbf{R}$.
- The off-diagonals of the categorical input correlation matrices correspond to the values of $\tau_{p_{ij}}$ in the definition of $\mathbf{R}$.

Note: If your report contains the note **Nugget parameters set to avoid singular variance matrix**, JMP has added a ridge parameter to the variance matrix so that it is invertible.

Chapter 18

Time Series Analysis

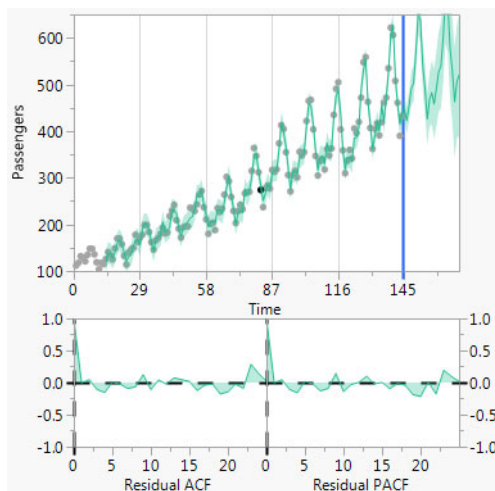
Fit Time Series Models and Transfer Functions

The Time Series platform enables you to explore, analyze, and forecast univariate time series. A time series is a set of observations taken over a series of equally spaced time periods. Observations that are close together in time are typically correlated. Time series methodology takes advantage of this dependence between observations to better predict what the series will look like in the future.

Characteristics that are common in time series data include seasonality, trend, and autocorrelation. The Time Series platform provides options to handle these characteristics. Graphs such as variograms, autocorrelation plots, partial autocorrelation plots, and spectral density plots can be used to identify the type of model appropriate for describing and predicting (forecasting) the time series. There are also several decomposition methods in the platform that enable you to remove seasonal or general trends in the data to simplify the analysis. Alternatively, the platform can fit more sophisticated ARIMA models and State Space Smoothing models that have the ability to incorporate seasonality and long term trends all in one model. You can also perform a Box-Cox transformation and analyze and model the transformed series.

There are several methods to assess the forecasting performance of models. The Forecast on Holdback feature partitions the time series into a training portion to build models and a holdback portion to assess forecasting performance.

Figure 18.1 Forecast Plot



Contents

Overview of the Time Series Platform	353
Example of the Time Series Platform	353
Launch the Time Series Platform	356
The Time Series Report	358
Time Series Graph	358
Time Series Basic Diagnostics Chart	359
Time Series Platform Options	361
Time Series Diagnostics	361
Differencing and Decomposition	362
Time Series Models	365
Additional Options	372
Additional Time Series Reports	373
Difference Report	374
Decomposition Reports	374
Model Comparison Report	376
Model Report	379
State Space Smoothing Report	383
Transfer Function Model Report	386
Spectral Density Report	387
Additional Examples of the Time Series Platform	388
Example of Creating a Time ID Column	388
Example of a Box-Cox Transformation	393
Example Using a Holdback Set	395
Statistical Details for the Time Series Platform	397
Statistical Details for Spectral Density	398
Statistical Details for X-11 Decomposition	398
Statistical Details for Exponential Smoothing Models	399
Statistical Details for ARIMA Models	402
Statistical Details for Transfer Functions	404

Overview of the Time Series Platform

A time series is a set $y_1, y_2, \dots, y_N$ of observations that are observed over a series of equally spaced time periods. Some examples of time series data include quarterly sales reports, monthly average temperatures, and counts of sunspots. The Time Series platform enables you to explore patterns and trends found in these types of data. You can then use these patterns and trends to forecast, or predict, into the future.

Characteristics that are common in time series data include seasonality, trend, and autocorrelation. *Seasonality* refers to patterns that occur over a known period of time. For example, data that are collected monthly might look similar in summer months across all years of data collection. *Trend* refers to long term movements of a series, such as gradual increases or decreases of values across time. *Autocorrelation* is the degree to which each point in a series is correlated with earlier values in the series.

There are many different models and forecasting methods available in the Time Series platform. However, not all methods can handle trend or seasonality. In order to choose an appropriate model, it is essential to determine which characteristics are present in the series. The Time Series platform provides graphs such as variograms, autocorrelation plots, partial autocorrelation plots, and spectral density plots that can be used to identify the type of model appropriate for describing and forecasting the evolution of the time series. There are also several differencing and decomposition methods in the platform that enable you to remove seasonal or general trends in the data to explore and simplify the analysis. You can also view and apply a Box-Cox transformation to your data.

Alternatively, the platform can fit more sophisticated models that can incorporate seasonality and long term trends. One such model in the platform that has this ability is Winter's Additive Method, which is an advanced exponential smoothing model. In addition, the platform can fit AutoRegressive Integrated Moving Average (ARIMA) models and State Space Smoothing models. Both of these types of models are the most statistically complex, but also provide the most flexibility. Advanced exponential smoothing, ARIMA, and State Space Smoothing models are harder to interpret, but they are excellent tools for forecasting.

The Time Series platform can also fit transfer function models when supplied with an input series.

Example of the Time Series Platform

Use the Time Series platform to examine a series of maximum monthly temperatures and predict those temperatures for the next two years.

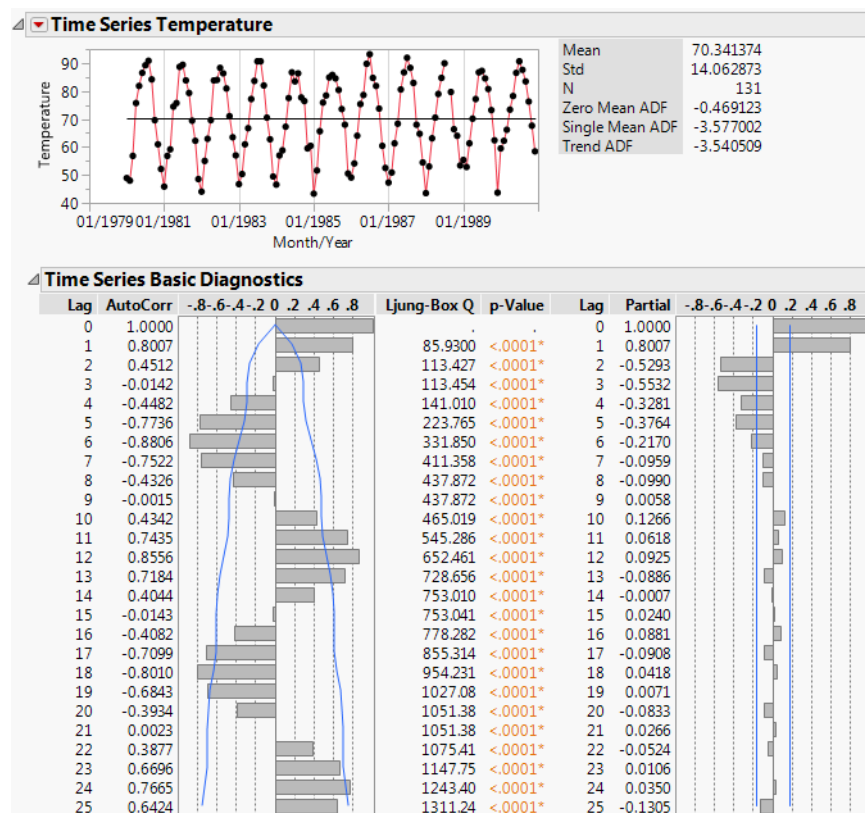
1. Select **Help > Sample Data Folder** and open Time Series/Raleigh Temps.jmp.

2. Select **Analyze > Specialized Modeling > Time Series**.
3. Select Temperature and click **Y, Time Series**.
4. Select Month/Year and click **X, Time ID**.
5. In the box next to **Forecast Periods**, type 24.

This is the number of future periods that are forecast by the models fit to the data. You want to predict the monthly temperature for the next two years, which is 24 months.

6. Click **OK**.

Figure 18.2 Time Series Analysis Report for Raleigh Temps.jmp

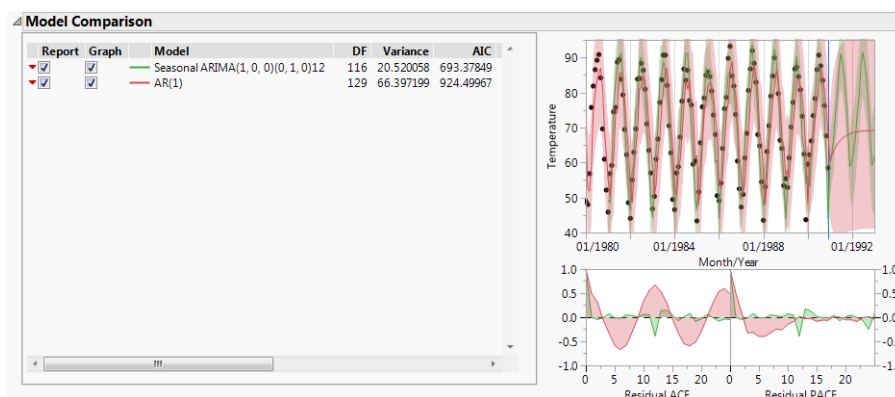


The Time Series graph shows that the series is cyclic. This cyclic component is also apparent in the autocorrelation chart. Points that are 1 lag apart are positively correlated, with an AutoCorr value of 0.8007. As points become farther apart, they become negatively correlated, then positively correlated again, and then the pattern repeats. The Time Series graph and the autocorrelation chart provide evidence of seasonality in the time series.

7. Click the Time Series red triangle and select **ARIMA**.
8. Set p , the autoregressive order, to 1 because the series showed evidence of autocorrelation.

9. Click **Estimate**.
10. Click the Time Series red triangle and select **Seasonal ARIMA**.
11. In the ARIMA box, set p , the autoregressive order, to 1 because the series showed evidence of autocorrelation.
12. In the Seasonal ARIMA box, set D , the seasonal differencing order, to 1 because the series showed evidence of seasonality.
13. Click **Estimate**.
14. In the Model Comparison table, check the box under Graph for both models.

Figure 18.3 Model Comparison Table for Raleigh Temps.jmp

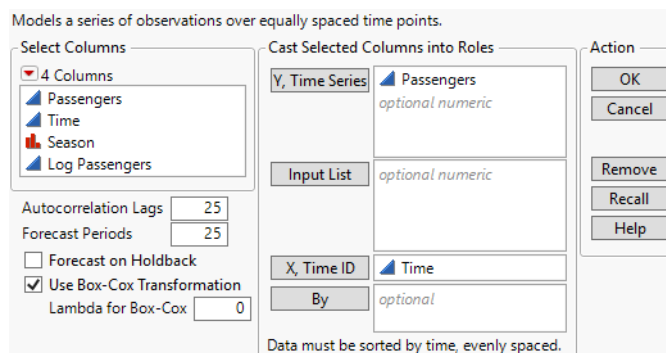


The Model Comparison table is sorted by the AIC statistic, in decreasing order. This means that the best fitting model appears at the top of the report. The AIC value for the seasonal ARIMA model (693.4) is much smaller than the value for the regular ARIMA model (924.5). The graph shows that while the ARIMA model predicts the observed points relatively well, the residuals are larger than those from the Seasonal ARIMA model. Also, the Seasonal ARIMA model has more realistic predictions for future observations with narrower prediction intervals. These results make sense since the series showed evidence of a seasonal component.

Launch the Time Series Platform

Launch the Time Series platform by selecting **Analyze > Specialized Modeling > Time Series**.

Figure 18.4 The Time Series Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Time Series platform launch window contains the following options:

Y, Time Series Assigns one or more columns as time series variables. Displayed on the Y axis.

Input List Assigns one or more columns as input series variables. Displayed in the Input Time Series Panel and used in transfer function models. The input series variable must be numeric, either as a time series or an indicator.

X, Time ID Assigns one variable for labeling the time axis (X axis). If no variable is specified for Time ID, the row number is used instead.

Note: If you use an X, Time ID variable, you can specify the time frequency by using the Time Frequency column property. You can choose Annual, Quarterly, Monthly, Weekly, Daily, Hourly, By Minute, and By Second. This helps JMP determine the spacing of the data when plotting the forecast values. If no frequency is specified, the data is treated as equally spaced numeric data.

Caution: It is assumed that the observations of the variable assigned to **X, Time ID** are equally spaced. However, the Time Series platform checks only whether the time stamps are increasing. The platform does not check if the observations are equally spaced.

By Assigns a column that creates a report consisting of separate analyses for each level of the variable. If more than one By variable is assigned, a separate analysis is produced for each possible combination of the levels of the By variables.

Note: If you use a By variable, you might have to change the number of autocorrelation lags depending on how many observations there are for each level of the By variable. The number of lags must be greater than one but less than the number of observations per level.

Autocorrelation Lags Specifies the number of lags to use in computing the autocorrelations and partial correlations. This is the maximum number of periods between points used in the computation of these correlations. It must be greater than one but less than the number of rows. The default number of lags is 25.

Tip: A commonly used rule for the maximum number of lags is $n/4$, where n is the number of observations.

Forecast Periods Specifies the number of observations that are forecast using each model fitted to the data. The default number of forecasts is 25.

- If you select the Forecast on Holdback option, the Forecast Periods value specifies the number of observations at the end of the series that are not included in the model fitting. These values are then forecast using the fitted model.
- If you do not select the Forecast on Holdback option, the Forecast Periods value specifies the number of future observations that are forecast after the end of the series.

Forecast on Holdback Determines whether forecasts are made on future observations or on the holdback observations. If the Forecast on Holdback option is selected, the forecasts are made on the holdback set that is determined by the number specified in the Forecast Periods option.

Use Box-Cox Transformation Transforms the original data using a Box-Cox transformation with the lambda that is specified in the Lambda for Box-Cox option. When this option is selected, all analyses in the Time Series report are performed on the transformed data. However, all forecasts are transformed back and reported on the original scale.

Lambda for Box-Cox (Available only when Use Box-Cox Transformation is selected.) Specifies the lambda parameter used for the Box-Cox transformation of the original data.

Note: In order to preserve the order and spacing of a time series, the Time Series platform treats excluded rows in the data table as missing values.

The Time Series Report

The initial Time Series report displays the time series graph, summary statistics and tests for the time series variable, and a basic diagnostics report. If a column is specified for Input List in the launch window, the Transfer Function Analysis and Input Time Series Panel reports are shown. Both the Transfer Function Analysis report and the Input Time Series Panel report contain the same initial information as the Time Series report.

- [“Time Series Graph”](#)
- [“Time Series Basic Diagnostics Chart”](#)

Time Series Graph

The Time Series graph plots each times series by the time ID. If no time ID is specified, the row number is used instead. If a holdback set is specified, there is a vertical reference line on the graph that separates the training data from the holdback data.

The platform also performs several tests for stationarity using *Augmented Dickey-Fuller* (ADF) tests. The following tests and summary statistics are displayed next to the time series graph:

Lambda for Box-Cox (Appears only if a Box-Cox Transformation is specified.) The value of lambda used in the Box-Cox transformation.

Mean The sample mean.

SD The sample standard deviation.

N The series length.

Zero Mean ADF A test against a random walk with zero mean, which is defined as follows:

$$x_t = \phi x_{t-1} + e_t$$

Single Mean ADF A test against a random walk with a non-zero mean, which is defined as follows:

$$x_t - \mu = \phi(x_{t-1} - \mu) + e_t$$

Trend ADF A test against a random walk with a non-zero mean and a linear trend, which is defined as follows:

$$x_t - \mu - \beta t = \phi[x_{t-1} - \mu - \beta(t-1)] + e_t$$

Time Series Basic Diagnostics Chart

The information that is shown in the Time Series Basic Diagnostics chart depends on the Time Series Report red triangle menu options. The red triangle menu options that show or hide information from the diagnostics chart are Autocorrelation, Partial Autocorrelation, Variogram, and AR Coefficients. By default, Autocorrelation and Partial Autocorrelation are shown.

Autocorrelation Chart

The Autocorrelation option shows or hides the following columns in the Time Series Basic Diagnostics chart:

Lag The number of periods between points.

Note: The number of lags begins with 0 as the default. To compute correlations beginning with lag 1, modify the JMP preferences before generating the graph. Select **File > Preferences > Platforms > Time Series**, and then select **Suppress Lag 0 in ACF and PACF**.

AutoCorr The autocorrelation for the k th lag, which is computed as follows:

$$r_k = \frac{c_k}{c_0} \text{ where } c_k = \frac{1}{N} \sum_{t=k+1}^N (y_t - \bar{y})(y_{t-k} - \bar{y})$$

and $\bar{y}$ is the mean of the N non-missing points in the time series. By definition, the first autocorrelation (lag 0) always has length 1.

The bars graphically depict the autocorrelations. The blue curves represent twice the large-lag standard error (± 2 standard errors), which is computed as follows:

$$SE_k = \sqrt{\frac{1}{N} \left(1 + 2 \sum_{i=1}^{k-1} r_i^2 \right)}$$

Ljung-Box Q Used to test whether a group of autocorrelations is significantly different from zero, or to test that the residuals from a model can be distinguished from white noise. Q is the test statistic.

p-Value The p -value from the Ljung-Box test.

Partial Autocorrelation Chart

The Partial Autocorrelation option shows or hides the following columns in the Time Series Basic Diagnostics chart:

Lag The number of periods between points.

Partial The partial autocorrelation for the k th lag.

The bars graphically depict the partial autocorrelations. The blue lines represent ± 2 standard errors for approximate 95% prediction limits, where the standard error is computed as follows:

$$SE_k = \frac{1}{\sqrt{n}} \text{ for all } k$$

Variogram Chart

The Variogram option shows or hides the following columns in the Time Series Basic Diagnostics chart:

Lag The number of periods between points.

Variogram The variogram measures the variance of the differences of points k lags apart and compares it to that for points one lag apart. The variogram is computed from the autocorrelations as follows:

$$V_k = \frac{1 - r_{k+1}}{1 - r_1}$$

where r_k is the autocorrelation at lag k .

AR Coefficients Chart

The AR Coefficients option shows or hides the following columns in the Time Series Basic Diagnostics chart:

Lag The number of periods between points.

AR Coef The coefficients approximate those that you would obtain from fitting a high-order, purely autoregressive model.

Time Series Platform Options

The Time Series red triangle menu contains the following options:

- [“Time Series Diagnostics”](#)
- [“Differencing and Decomposition”](#)
- [“Time Series Models”](#)
- [“Additional Options”](#)

Time Series Diagnostics

The Time Series red triangle menu contains the following diagnostic options to explore the time series data.

Graph Shows a submenu of options to control the time series plot appearance.

Time Series Graph Shows or hides the time series graph.

Show Points Shows or hides the points in the time series graph.

Connecting Lines Shows or hides the lines connecting the points in the time series graph.

Mean Line Shows or hides a horizontal line in the time series graph that depicts the mean of the time series.

Autocorrelation Shows or hides the Autocorrelation plot in the Time Series Basic Diagnostics Chart. The autocorrelation graph describes the correlation between all pairs of points in the time series for a given separation in time (lag). See [“Autocorrelation Chart”](#).

Tip: The autocorrelation graph of the sample is often called the *sample autocorrelation function*.

Partial Autocorrelation Shows or hides the Partial Autocorrelation plot in the Time Series Basic Diagnostics Chart. The partial autocorrelation graph describes the partial correlation between all the pairs of points in the time series for a given separation in time (lag). See [“Partial Autocorrelation Chart”](#).

Tip: The Autocorrelation and Partial Autocorrelation graphs can help you determine whether the time series is stationary (meaning it has a fixed mean and standard deviation over time) and what model might be appropriate to fit the time series.

Variogram Shows or hides the graph of the variogram in the Time Series Basic Diagnostics Chart. See [“Variogram Chart”](#).

AR Coefficients Shows or hides the graph of the least squares estimates of the autoregressive (AR) coefficients in the Time Series Basic Diagnostics Chart. See [“AR Coefficients Chart”](#).

Spectral Density Shows or hides graphs of the spectral density as a function of period and frequency. The spectral density option also displays the White Noise test report, which gives results from two tests on the data. See [“Spectral Density Report”](#) and [“Statistical Details for Spectral Density”](#).

Differencing and Decomposition

The Time Series red triangle menu contains the following options to prepare the data for modeling.

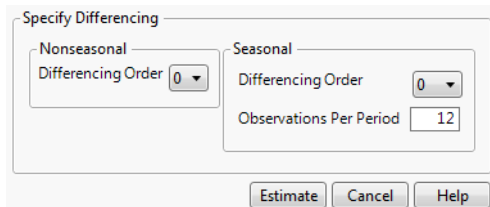
Difference Shows the Differencing Specification window ([Figure 18.5](#)). The window enables you to specify the differencing operation that you want to apply to the time series. Differencing the values in a time series can transform a nonstationary series into a stationary series. The differenced series is given by the following equation:

$$w_t = (1 - \mathbf{B})^d (1 - \mathbf{B}^s)^D y_t$$

where t is the time index and $\mathbf{B}$ is the backshift operator defined by $\mathbf{B}y_t = y_{t-1}$.

Note: Many time series do not exhibit a fixed mean, such as time series with trend or seasonality. Such nonstationary series are not suitable for description by time series models that assume a stationary time series such as ARMA models. Removing the trend and/or seasonality creates a differenced series that is stationary and enables you to describe the series using the models that assume stationarity.

Figure 18.5 Differencing Specification Window



The image shows a software dialog box titled "Specify Differencing". It contains two main sections: "Nonseasonal" and "Seasonal". In the "Nonseasonal" section, there is a label "Differencing Order" followed by a dropdown menu showing the value "0". In the "Seasonal" section, there is a label "Differencing Order" followed by a dropdown menu showing the value "0", and below it, a label "Observations Per Period" followed by a text input field containing the value "12". At the bottom of the dialog box, there are three buttons: "Estimate", "Cancel", and "Help".

The Differencing Specification window enables you to specify the Nonseasonal Differencing Order, d , the Seasonal Differencing Order, D , and the number of Observations per Period, s . Selecting zero for the value of the differencing order is

equivalent to no differencing of that kind. Each time you specify a differencing operation and click **Estimate**, a new Difference Report is displayed in the report window. See [“Additional Examples of the Time Series Platform”](#).

Decomposition (Not available when either Forecast on Holdback or Use Box-Cox Transformation are selected in the launch window.) Shows a submenu of decomposition methods. Decomposition of time series data isolates and removes linear trends and seasonal cycles from a time series. This can help with better model estimation. Three Decomposition options are provided.

Remove Linear Trend Estimates the linear trend of the time series using a linear regression model and removes the linear trend from the data. A Time Series report for the detrended series is added to the report window, along with the linear trend information. See [“The Time Series Report”](#) and [“Linear Trend Report”](#).

Remove Cycle Estimates the cyclic component of a time series using a single cosine wave and then removes the cyclic component from the data. When you select the Remove Cycle option, the Define Cycle dialog appears. This dialog window enables you to specify the number of units per cycle and indicate whether a constant should be subtracted from the data. A Time Series report for the decycled series is added to the report window, along with the cycle information. See [“The Time Series Report”](#) and [“Cycle Report”](#).

X11 Removes trend and seasonal effects using the X-11 method developed by the US Bureau of the Census (Shiskin et. al. 1967). For more information about the X-11 method, see [“Statistical Details for X-11 Decomposition”](#). When selected, the Select Decomposition Type dialog appears. This dialog window enables you to specify a multiplicative or additive X-11 adjustment. Once you click **OK**, an X11 report is added to the report window. See [“X11 Report”](#).

The X11 option is available only for monthly or quarterly data. The X, Time ID column must contain numeric values equally spaced by month or quarter without any gaps or missing values. JMP returns an error if you request X11 for a time column that does not satisfy these requirements. For an example of an appropriate time column for an X11 analysis, see [“Create Appropriate Time ID Column”](#).

Note: When you select the **Remove Linear Trend** or the **Remove Cycle** options, JMP adds a column to the data table that contains the detrended or decycled data. If this column is already present in the data table when you select the option, JMP overwrites the existing column.

Tip: Typically, you would begin decomposition by removing any linear trend, and then removing long cycles, such as a 12-month cycle. Then you could start removing short cycles, such as 6-month cycles.

Show Lag Plot Shows or hides a plot with observations at time t on the Y axis and observations at time $t \pm p$ on the X axis. The $\pm p$ is known as the lag. This plot is useful in determining how an observation at time t is related to another observation at time $t \pm p$. If there is not an identifiable structure to the plot, the observations are not related. However, if there is a structure to the plot, this indicates that there is some relationship between observations across time. Identifying the structure helps when building a time series model.

Show Box-Cox Transformation Plot (Not available if you specified Use Box-Cox Transformation in the launch window.) Shows or hides the Box-Cox Transformation plot. The Box-Cox transformation is a power transformation based on the Lambda parameter (λ) and uses the following formula:

$$Y_{\lambda} = \begin{cases} \frac{y^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases}$$

The Box-Cox Transformation Plot shows a graph of the transformed time series and options for setting the Lambda parameter (λ). The Lambda box shows the current value of the parameter. You can also use the Lambda slider to change the value of the parameter, which then automatically updates the plot. The Low and High boxes set the range of the slider.

Tip: Use the Box-Cox Transformation plot to see the transformed time series for different values of Lambda. Once you decide on a value, relaunch the analysis and select **Use Box-Cox Transformation** and enter the desired Lambda next to Lambda for Box-Cox. This runs the Time Series platform on the transformed series.

Cross Correlation (Available only in the Transfer Function Analysis red triangle menu.) Shows or hides a cross-correlation plot to the report. The length of the plot is twice that of an autocorrelation plot, or $2 \times \text{ACF length} + 1$. The plot includes plots of the output series versus all input series, in both numerical and graphical forms. The blue lines indicate two standard errors.

Note: For the cross correlation plot, the standard errors are calculated under the null hypothesis as $1/\sqrt{n - k}$, where n is the number of nonmissing data values and k is the number of autocorrelation lags.

Prewhitening (Available only in the Input Series red triangle menu.) Shows the Prewhitening Specification window that enables you to set the prewhitening order. Prewhitening is a technique used to help identify the transfer function model. This technique involves fitting an ARIMA model to the input series such that the residuals are equivalent to white noise. The same model is then used to estimate the output series. You

can use the cross correlation between the filtered input and filtered output series to determine an appropriate lag for the transfer function model. For information about prewhitening, see Box et al. (1994).

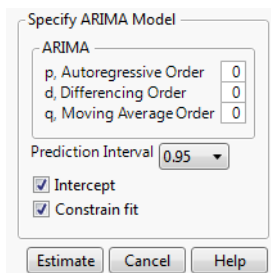
Time Series Models

The Time Series red triangle menu contains options to fit the following types of models.

ARIMA Shows the ARIMA Specification window, which enables you to specify the ARIMA model that you want to fit. An ARIMA model predicts future values of a time series by a linear combination of its past values and a series of errors (also known as *random shocks* or *innovations*). The ARIMA model performs a maximum likelihood fit of the specified ARIMA model to the time series. See “[ARIMA Model](#)”.

Note: An ARIMA model is commonly denoted $ARIMA(p,d,q)$. If any of p , d , or q are zero, the corresponding letters are often dropped. For example, if p and d are zero, then the model would simply be a moving average model, denoted as $MA(q)$.

Figure 18.6 ARIMA Specification Window



p , Autoregressive Order The order p of the polynomial $\phi(\mathbf{B})$ operator.

d , Differencing Order The order d of the differencing operator.

q , Moving Average Order The order q of the differencing operator $\theta(\mathbf{B})$.

Prediction Interval Enables you to set the prediction level between 0 and 1 for the forecast prediction intervals.

Intercept Determines whether the intercept term μ is a part of the model.

Constrain fit If checked, the fitting procedure constrains the autoregressive parameters to always remain within the stable region and the moving average parameters within the invertible region.

Tip: Deselect the Constrain fit option if the fitter is having difficulty finding the true optimum or if you want to speed up the fit. You can use the Model Summary table to see whether the resulting fitted model is stable and invertible.

Once you specify the model and click **Estimate**, a Model Report is added to the report window. See [“Additional Time Series Reports”](#).

Seasonal ARIMA Shows the Seasonal ARIMA Specification window, which enables you to specify the Seasonal ARIMA model that you want to fit. This window has the same elements as the ARIMA specification window, but it also contains the seasonal element specifications. The additional Observations per Period option enables you to specify the number of observations per period, denoted as s . For more information about the Seasonal ARIMA model, see [“Seasonal ARIMA Model”](#).

Note: Seasonal ARIMA models are denoted as Seasonal ARIMA(p,d,q)(P,D,Q) s .

Once you specify the model and click **Estimate**, a Model Report is added to the report window. See [“Additional Time Series Reports”](#).

Smoothing Models Shows a submenu of smoothing models. Once you select a smoothing model, a specification window appears. See [“Simple Smoothing Average Specification Window”](#) and [“Smoothing Model Windows”](#). For each model that is specified, a Smoothing Model Report appears in the report window. See [“Model Report”](#). Smoothing models represent the evolution of a time series by the model:

$$y_t = \mu_t + \beta_t t + s(t) + a_t$$

where

μ_t is the time-varying mean term

β_t is the time-varying slope term

$s(t)$ is one of the s time-varying seasonal terms

a_t are the random shocks

For more information about the general smoothing model equation, see [“Statistical Details for Exponential Smoothing Models”](#). The following smoothing models are available:

Simple Moving Average (Not available when either Forecast on Holdback or Use Box-Cox Transformation are selected in the launch window.) A model that estimates values by using an average of several adjacent points, defined by the smoothing window. The Simple Smoothing Average Specification window enables you to specify aspects of the smoothing window. Once the smoothing window options are specified, a Simple Moving Average report is shown. By default, this report produces plotted values that are equal to the average of consecutive observations in a time window. Multiple Simple Moving Average models can be added and shown on the same plot. See [“Simple Smoothing Average Specification Window”](#).

Simple Exponential Smoothing A model with a level component. See [“Simple Exponential Smoothing”](#).

Double Exponential Smoothing A model with a level component and a trend component. This is a special case of Linear Exponential Smoothing. See [“Double \(Brown\) Exponential Smoothing”](#).

Linear Exponential Smoothing A model with a level component and a trend component. See [“Linear \(Holt\) Exponential Smoothing”](#).

Damped-Trend Linear Exponential Smoothing A model with a level component and a damped trend component. This model is appropriate for a series that exhibits a trend more complicated than a linear trend. See [“Damped-Trend Linear Exponential Smoothing”](#).

Seasonal Exponential Smoothing A model with a level component and a seasonal component. See [“Seasonal Exponential Smoothing”](#).

Winters Method A model with a level component, a trend component, and a seasonal component. See [“Winters Method \(Additive\)”](#).

Note: Each smoothing model, except for the Simple Moving Average model, is a traditional exponential smoothing model and has an ARIMA model equivalent. You might not be able to specify the equivalent ARIMA model using the ARIMA option because some smoothing models intrinsically constrain the ARIMA model parameters in ways that the ARIMA option does not allow.

State Space Smoothing Models Shows the Specify State Space Smoothing Models window, which enables you to fit a variety of state space smoothing models as defined by Hyndman et al. (2008). A state space smoothing model is defined based on its error, trend component, and seasonal component.

- The errors can be additive (A) or multiplicative (M).
- The trend component can be none (N), additive (A), additive damped (A_d), multiplicative (M), or multiplicative damped (M_d).

- The seasonal component can be none (N), additive (A), or multiplicative (M).

A specific model can be represented by its ETS (Error, Trend, Seasonal). Use the check boxes in the Specify State Space Smoothing Models window to select the error, trend, and seasonality for the desired models. Click **Select Recommended** to select the check boxes that correspond to the models recommended by the platform. The window opens with the recommended models selected. Click **Select All** to select all check boxes or **Deselect All** to deselect all check boxes.

The window also contains the following options:

Period Specifies seasonality values to be considered in the model fitting process.

Constrain Parameters Constrains the parameters in such a way that the further an observation is from the present, the less effect it has on the present state of the model. In State Space Smoothing models, a forecast at time t , given all previous observations, is the same as the weighted sum of all observations up to time t . The weights are a function of the parameters. Therefore, constraining the parameters ensures that the weights for past observations go to zero and that the further an observation is from the present, the faster the weight goes to zero.

When you click **OK**, the specified set of models is fit. Each State Space Smoothing model that is fit is added to the Model Comparison table. If the Report check box for a state space smoothing model in the Model Comparison table is selected, a State Space Smoothing report appears. See [“State Space Smoothing Report”](#).

Transfer Function (Available only in the Transfer Function Analysis red triangle menu.) Shows the Transfer Function Model Specification window. Building a transfer function model is similar to building an ARIMA model; it is an iterative process of exploring, fitting, and comparing models. Before building a model and during the data exploration process, it is sometimes useful to prewhiten the data. See [“Prewhitening”](#). For more information about transfer functions, see [“Statistical Details for Transfer Functions”](#).

Note: Currently, the Transfer Function option has limited support of missing values. Because the Time Series platform also treats excluded rows as missing values, transfer function models cannot be fit on a data table with excluded rows.

Figure 18.7 Transfer Function Model Specification Window

Specify Transfer Function Model																																			
Noise Series Orders <table border="1"> <thead> <tr> <th></th> <th>Output CO2</th> </tr> </thead> <tbody> <tr><td>p, Autoregressive Order</td><td>2</td></tr> <tr><td>d, Differencing Order</td><td>0</td></tr> <tr><td>q, Moving Average Order</td><td>0</td></tr> <tr><td>P, Autoregressive Order</td><td>0</td></tr> <tr><td>D, Differencing Order</td><td>0</td></tr> <tr><td>Q, Moving Average Order</td><td>0</td></tr> <tr><td>S, Observations Per Period</td><td>12</td></tr> </tbody> </table> Choose Inputs ☒ Input Gas Rate Inputs Series Orders <table border="1"> <thead> <tr> <th></th> <th>Input Gas Rate</th> </tr> </thead> <tbody> <tr><td>s1, Order of Numerator Operator</td><td>2</td></tr> <tr><td>d1, Order of Differencing Operator</td><td>0</td></tr> <tr><td>r1, Order of Denominator Operator</td><td>2</td></tr> <tr><td>s2, Order of Seasonal Numerator Operator</td><td>0</td></tr> <tr><td>d2, Order of Seasonal Differencing Operator</td><td>0</td></tr> <tr><td>r2, Order of Seasonal Denominator Operator</td><td>0</td></tr> <tr><td>S, Observations Per Period</td><td>12</td></tr> <tr><td>L, Input Lag</td><td>3</td></tr> </tbody> </table>			Output CO2	p, Autoregressive Order	2	d, Differencing Order	0	q, Moving Average Order	0	P, Autoregressive Order	0	D, Differencing Order	0	Q, Moving Average Order	0	S, Observations Per Period	12		Input Gas Rate	s1, Order of Numerator Operator	2	d1, Order of Differencing Operator	0	r1, Order of Denominator Operator	2	s2, Order of Seasonal Numerator Operator	0	d2, Order of Seasonal Differencing Operator	0	r2, Order of Seasonal Denominator Operator	0	S, Observations Per Period	12	L, Input Lag	3
	Output CO2																																		
p, Autoregressive Order	2																																		
d, Differencing Order	0																																		
q, Moving Average Order	0																																		
P, Autoregressive Order	0																																		
D, Differencing Order	0																																		
Q, Moving Average Order	0																																		
S, Observations Per Period	12																																		
	Input Gas Rate																																		
s1, Order of Numerator Operator	2																																		
d1, Order of Differencing Operator	0																																		
r1, Order of Denominator Operator	2																																		
s2, Order of Seasonal Numerator Operator	0																																		
d2, Order of Seasonal Differencing Operator	0																																		
r2, Order of Seasonal Denominator Operator	0																																		
S, Observations Per Period	12																																		
L, Input Lag	3																																		
☒ Intercept ☐ Alternative Parameterization ☒ Constrain fit Forecast Periods: 8 Prediction Interval: 0.95																																			
Estimate Cancel Help																																			

The Transfer Function Model Specification window contains the following sections:

Noise Series Orders Contains specifications for the noise series. Lowercase letters are coefficients for non-seasonal polynomials, and uppercase letters are coefficients for seasonal polynomials.

Choose Inputs Enables you select the input series for the model.

Input Series Orders Contains specifications for the input series. The first three orders relate to non-seasonal polynomials. The next four orders relate to seasonal polynomials. The final option is for an input lag.

There are three additional options that control model fitting:

Intercept Specifies whether the model has an intercept. If unchecked, it is assumed that μ is zero.

Alternative Parameterization Specifies whether the general regression coefficient is factored out of the numerator polynomials.

Constrain Fit Toggles the constraining of the AR and MA coefficients.

Forecast Periods Specifies the number of forecasting periods that are used for forecasting. If there are rows at the end of the data table that contain missing values for the Y variable and nonmissing values for the input variables, these rows are used in the initial forecasting settings. The values for the input variables are treated as future values of the input variables.

Prediction Interval Specifies the confidence level for the prediction interval.

ARIMA Model Group Shows the ARIMA Model Group window, which enables you to fit a range of ARIMA or Seasonal ARIMA models by specifying the range of orders. As you enter ranges into the window, the Total Number of Models updates accordingly.

Figure 18.8 ARIMA Model Group Specification Window

Specify ARIMA Model

ARIMA

p, Autoregressive Order	0	0
d, Differencing Order	0	0
q, Moving Average Order	0	0

Seasonal ARIMA

P, Autoregressive Order	0	0
D, Differencing Order	0	0
Q, Moving Average Order	0	0
Observations Per Period	12	12

Prediction Interval: 0.95

☒ Intercept

☒ Constrain fit

Total Number of Models: 0

Estimate Cancel Help

Once you specify the models and click **Estimate**, a Model Report for each specified model is added to the report window. See [“Additional Time Series Reports”](#).

Simple Smoothing Average Specification Window

The Simple Smoothing Average Specification window appears when you select Simple Moving Average as the smoothing model. Let w be the smoothing window width in a simple moving average (SMA) model. Let $f_t = (y_t + y_{t-1} + y_{t-2} + \dots + y_{t-(w-2)} + y_{t-(w-1)}) / w$ be the average of w consecutive observations for some time point t .

Figure 18.9 Simple Smoothing Average Specification Window

Enter smoothing window width: 5

☒ No Centering

☐ Centered

☐ Centered and Double Smoothed for Even Number of Terms

OK Cancel

Enter smoothing window width The smoothing window width, w , that defines the number of consecutive points to average. The larger the window width, the more the series is smoothed.

No Centering The smoothing window is constructed from the points leading up to and including the time point, t , the point at which the series is being estimated. In other words, f_t is the plotted value for time t .

Centered The smoothing window is centered around the time point at which the series is being estimated.

- For odd w , f_t is the plotted value for time $t - (w - 1) / 2$.

- For even w , f_t is the plotted value for time $t-(w-1)/2$. When saved to a data table, f_t is at $t-(w-2)/2$.

Centered and Double Smoothed for Even Number of Terms For even w , the smoothing window cannot be centered around the time point at which the series is being estimated. This option creates two smoothing windows that are almost centered, and averages them together. The smoothing estimates are calculated as follows:

$$f_{t-\frac{w}{2}} = \frac{y_t + 2 \sum_{i=1}^{w-1} y_{t-i} + y_{t-w}}{2w}$$

Smoothing Model Windows

The Smoothing Model specification windows appear when you select one of the smoothing model options other than Simple Moving Average. The title of the window and the available options depend on the smoothing model option that you select.

Figure 18.10 Smoothing Model Specification Window

Prediction Interval: 0.95

Observations per Period: 12

Constraints: Zero to One

Estimate Cancel Help

Prediction Interval Enables you to set the prediction level for the forecast prediction intervals.

Observations per Period (Available only for seasonal smoothing models.) Enables you to set the number of observations per period in a seasonal smoothing model.

Constraints Enables you to specify what type of constraint you want to enforce on the smoothing weights during the fit. The following constraint options are available:

Zero To One Constrains the values of the smoothing weights to the range zero to one.

Unconstrained Allows the parameters to range freely.

Stable Invertible Constrains the parameters such that the equivalent ARIMA model is stable and invertible.

Custom Expands the dialog to enable you to set constraints on individual smoothing weights. Each smoothing weight can be **Bounded**, **Fixed**, or **Unconstrained** as determined by the setting of the popup menu next to the weight's name. When entering values for fixed or bounded weights, the values can be positive or negative real numbers.

Figure 18.11 Custom Smoothing Weights

Prediction Interval: 0.95

Constraints: Custom

Custom Smoothing Weights

Weight	Constrain	Fixed Weight	Lower Bound	Upper Bound
Level	Fixed	0.3		
Trend	Bounded		0.1	0.8

The example shown in [Figure 18.11](#) has the Level weight (α) fixed at a value of 0.3 and the Trend weight (γ) bounded by 0.1 and 0.8. In this case, the value of the Trend weight is allowed to move within the range 0.1 to 0.8 while the Level weight is held constant at 0.3. Note that you can specify all the smoothing weights in advance by using these custom constraints. In that case, none of the weights would be estimated from the data although forecasts and residuals would still be computed.

Additional Options

The Time Series red triangle menu also contains the following options:

Combine and Save Forecasts from Models Creates a new data table with the combined results from all model fits in the report.

Save Spectral Density Creates a new data table containing the spectral density and periodogram where the $(i+1)$ th row corresponds to the frequency $f_i = i / N$ (that is, the i th harmonic of $1 / N$). The new data table has the following columns:

Period The period of the i th harmonic, $1 / f_i$.

Frequency The frequency of the harmonic, f_i .

Angular Frequency The angular frequency of the harmonic, $2\pi f_i$.

Sine The Fourier sine coefficients, a_i .

Cosine The Fourier cosine coefficients, b_i .

Periodogram The periodogram, $I(f_i)$.

Spectral Density The spectral density, a smoothed version of the periodogram.

Number of Forecast Periods Shows a window that enables you to set the number of future periods that are forecast for the fitted models. The initial value is set in the Time Series launch window. All existing and future forecast results will use the new number of periods once it is changed.

Maximum Iterations Shows a window that enables you to reset the maximum number of iterations for future optimizations used in fitting ARIMA models.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

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

Additional Time Series Reports

Many options create additional reports in the Time Series platform.

- [“Difference Report”](#)
- [“Decomposition Reports”](#)
- [“Model Comparison Report”](#)
- [“Model Report”](#)
- [“State Space Smoothing Report”](#)
- [“Transfer Function Model Report”](#)
- [“Spectral Density Report”](#)

Difference Report

The Difference Report in the Time Series platform contains graphs of the autocorrelations and partial autocorrelations of the differenced series. These graphs can be used to determine whether the differenced series is stationary.

The ARIMA and Seasonal ARIMA models that are available in the Time Series platform accommodate a differencing operation. In a two step process, these models first difference the time series according to the differencing operation, and then fit the differenced series. The Difference option is a useful preprocessing tool for determining the order of differencing to specify for the ARIMA model.

The Difference red triangle menu contains the following options:

Graph Shows a submenu of options to control the appearance of the differenced series plot. See [“Time Series Platform Options”](#).

Autocorrelation Shows or hides the autocorrelation of the differenced series.

Partial Autocorrelation Shows or hides the partial autocorrelations of the differenced series.

Variogram Shows or hides the variogram of the differenced series.

Save Saves a new column that contains the values in the differenced series to the original data table. Some of the leading elements are lost in the differencing process. They are represented as missing values in the saved Difference column.

Remove Fit Removes the Difference report from the report window.

Decomposition Reports

In the Time Series platform, there are three decomposition options. Each option creates a different report.

- [“Linear Trend Report”](#)
- [“Cycle Report”](#)
- [“X11 Report”](#)

Linear Trend Report

Contains the values of β_0 and β_1 from the linear regression model that is fit to the data:

$$\text{Trend}_t = \beta_0 + \beta_1 \text{time}$$

The detrended series is equal to $D_t = O_t - \text{Trend}_t$, where O_t is the original time series.

Cycle Report

Contains the values of the cyclical component that is fit to the data:

$$\text{Cycle}_t = C + A \cdot \cos\left(2 \cdot \pi \cdot \left(\left(\frac{1}{U}\right) \cdot t + P\right)\right)$$

The parameter values are defined as follows:

- C is the (optional) Constant
- A is the Amplitude of the cosine wave
- U is the number of Units per Cycle
- P is the Phase of the cosine wave
- t is one less than the row number of a given observation

The decycled series is equal to $D_t = O_t - \text{Cycle}_t$, where O_t is the original time series.

X11 Report

Depending on your selection of Decomposition Type, the X11 option adds an X11 - Multiplicative report or an X11 - Additive report. The reports contain the same four plots:

Original and Adjusted Overlays the X11-adjusted time series on the original time series, O_t . The X11-adjusted values are O_t/S_t for the multiplicative adjustment and $O_t - S_t$ for the additive adjustment.

D10 - Final Seasonal Factors Plots the seasonal factor components, S_t , over time.

D12 - Final Trend Cycle Plots the trend cycle components, C_t , over time.

D13 - Final Irregular Series Plots the irregular components, I_t , over time.

X11 Report Options

The X11 reports have the following red triangle options:

Show Tables Shows or hides the X11 summary tables, as described in Shiskin et. al. (1967). The tables are grouped into five categories (labeled B through F), described in [Table 18.1](#).

Save Columns Saves four columns to the data table: the seasonally adjusted time series, the seasonal components (S_t), the trend cycle components (C_t), and the irregular series components (I_t).

Save All Columns Saves columns to the data table for all of the tables produced in the report by the **Show Tables** option.

Remove Fit Removes the X11 report from the report window.

Table 18.1 Descriptions of the Categories of X11 Output Tables

Letter Prefix	Category Description
B	preliminary estimates of seasonal, trend cycle, and irregular components
C	intermediate estimates of seasonal, trend cycle, and irregular components
D	final estimates of seasonal, trend cycle, and irregular components
E	analytical tables
F	summary measures

For more information about the contents of the X11 output tables, see Shisken et. al. (1967) or *SAS/ETS 15.2 User’s Guide* (search for “The output from PROC X11”).

Model Comparison Report

The Model Comparison report in the Time Series platform contains a table of fit statistics and plots for the models. Each time a new model is fit, a new row is added to the Model Comparison table, with a unique color-coding. The Model Comparison table summarizes the fit statistics for each model and is used to compare several models fitted to the same time series. By default, the models are sorted by the AIC statistic, in increasing order. To sort the model by a different statistic, right-click any column in the Model Comparison table and select Sort by Column. Alternatively, you can also click on the column header to sort. For definitions of the fit statistics, see “Model Summary Table”. The only fit statistic that is unique to the Model Comparison Table is Weights. This fit statistic is the normalized AIC Weight. The AIC Weight for a model is calculated as follows:

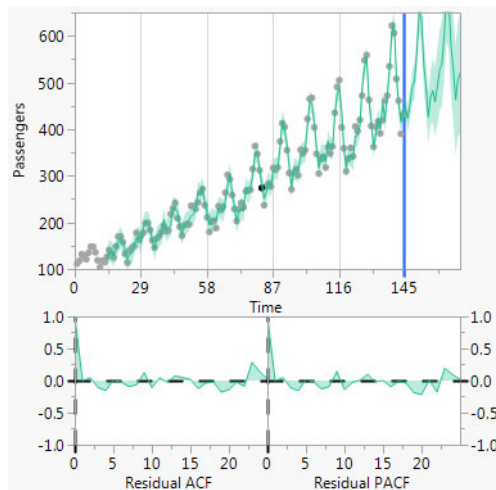
$$\text{AICWeight} = \exp[-0.5(\text{AIC} - \text{BestAIC})] / \sum_{k=1}^K (\exp[-0.5(\text{AIC}_k - \text{BestAIC})])$$

K is the total number of models, AIC_k is the AIC value for model k , and BestAIC is the AIC value for the model with the minimum AIC value.

Caution: If you do not specify a holdback set and fit a combination of state space smoothing models and ARIMA models, a caution message is shown at the top of the Model Comparison table. In this case, the state space smoothing models are not comparable to the ARIMA models because likelihood values and likelihood-based information criteria, such as the AIC statistic, are not comparable between the two different model classes. However, if you specify a holdback set, the models can be compared because the RMSE statistic is used.

Model Comparison															
Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2	.4	.6	.8	MAPE	MAE
☒	☐	SeasonalARIMA(0, 1, 1)(0, 1, 1) ₁₂	128	138.49122	1020.9047	1029.5302	0.990	1014.9047	1.000000					3.197010	8.990978
☒	☐	ARIMA(1, 1, 1)	140	978.9437	1394.1215	1403.0101	0.932	1388.1215	0.000000					8.687058	24.364166
☒	☐	ARMA(1, 1)	141	989.13835	1407.7483	1416.6577	0.919	1401.7483	0.000000					9.704486	25.348507
☒	☐	AR(1)	142	1134.3871	1426.1794	1432.1190	0.909	1422.1794	0.000000					9.873726	26.556219
☒	☐	MA(1)	142	4264.6121	1616.8626	1622.8022	0.693	1612.8626	0.000000					23.865796	53.716974

You can select which full model reports are shown in the report window using the Report check box. Two model plots appear to the right of the Model Comparison table. The top plot is a time series plot of the data, forecasts, and prediction limits. Below that are plots of the autocorrelation and partial autocorrelation functions. You can select which models are displayed on the model plots using the Graph check box.

Figure 18.13 Model Plots


Model Comparison Report Options

Each model in the Model Comparison report has the following red triangle menu options:

Fit New Opens a specification window for the model. You can change the settings to fit a different model.

Simulate Once (Not available if a holdback set is specified.) Provides one simulation of the model to the upper end of the time axis. The simulation is shown on the Model Comparison time series plot.

Simulate More (Not available if a holdback set is specified.) Provides the specified number of simulations of the model to the upper end of the time axis. The simulations are shown on the Model Comparison time series plot.

Remove Model Simulation (Not available if a holdback set is specified.) Removes the simulations for the model.

Remove All Simulation (Not available if a holdback set is specified.) Removes the simulations for all models.

Generate Simulation (Not available if a holdback set is specified.) Generates simulations for the model, and stores the results in a data table. You can specify the random seed, number of simulations, and the number of forecast periods.

Set Seed (Not available if a holdback set is specified.) Specifies the seed for generating the simulated trajectories.

Remove Fit Removes the corresponding model from the report window.

Remove All Models Removes all models that are listed in the Model Comparison table from the report window.

Remove Unselected Removes all models that do not have the Report check box selected from the report window.

Remove Selected Removes all models that have the Report check box selected from the report window.

Hide All Reports Hides all of the models that are listed in the Model Comparison table from the report window. This is equivalent to deselecting all of the Report check boxes.

Hide All Graphs Hides all of the models that are listed in the Model Comparison table from the Model Plots. This is equivalent to deselecting all of the Graph check boxes.

Note: The Time Series platform treats excluded rows as missing values. If the last several observations in the data table are excluded or missing, the simulation options in the Model Comparison Report will not work properly. The simulation procedure uses the last several observations of the time series. If the last several observations are missing, the simulation produces all missing values into the future.

Model Report

In the Time Series platform, the ARIMA, Seasonal ARIMA, and Smoothing Model options produce similar model reports. These modeling options are used to fit theoretical models to the series and use the fitted model to predict (forecast) future values of the series. These options also produce statistics and residuals that enable you to determine the adequacy of the model that you have chosen to use. You can select the modeling options multiple times. Each time you select a model, that model is added to the Model Comparison table. When the Report check box next to a model in the Model Comparison table is selected, a report is produced for that model. The report specifies the model in its title.

The following reports are shown by default:

- Model Summary Table
- Parameter Estimates Table
- Forecast Plot
- Residuals
- Iteration History

Model Summary Table Contains fit statistics for the model. In the formulas below, n is the length of the series and k is the number of fitted parameters in the model.

DF The number of degrees of freedom in the fit, $n - k$.

Sum of Squared Innovations The sum of the squared innovations.

Sum of Squared Residuals The sum of the squared residuals.

Variance Estimate The sum of squared innovations divided by the number of degrees of freedom $(n - k)$. This is the sample estimate of the variance of the random shocks a_t , described in the section [“ARIMA Model”](#).

Standard Deviation The square root of the variance estimate. This is a sample estimate of the standard deviation of the random shocks a_t .

Akaike’s ‘A’ Information Criterion The AIC value, computed as $-2\log\text{likelihood} + 2k$. Smaller AIC values indicate better fit.

Schwarz’s Bayesian Criterion The SBC value, computed as $-2\log\text{likelihood} + k*\ln(n)$. Smaller SBC values indicate better fit. Schwarz’s Bayesian Criterion is equivalent to the Bayesian Information Criterion (BIC).

RSquare The R-Square value, computed as follows:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$\hat{y}_i$ are the one-step-ahead forecasts

$\bar{y}$ is the mean

If the model does not fit the series well, the model error sum of squares, SSE, might be larger than the total sum of squares, SST. As a result, R^2 can be negative.

RSquare Adj The adjusted R^2 value, computed as follows:

$$1 - \left[\frac{(n-1)}{(n-k)} (1 - R^2) \right]$$

MAPE The Mean Absolute Percentage Error value, computed as follows:

$$\frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

MAE The Mean Absolute Error value, computed as follows:

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

-2LogLikelihood Twice the negative log-likelihood function evaluated at the best-fit parameter estimates. Smaller values are better fits. See *Fitting Linear Models*.

Stable Indicates whether the autoregressive operator is stable. That is, whether all the roots of $\phi(z) = 0$ lie outside the unit circle.

Invertible Indicates whether the moving average operator is invertible. That is, whether all the roots of $\theta(z) = 0$ lie outside the unit circle.

Note: The ϕ and θ operators are defined in the section [“ARIMA Model”](#).

Parameter Estimates Table Shows the estimates for the time series model parameters. Each type of model has its own set of parameters, which are described in the sections on specific time series models. Each Parameter Estimates table contains the following columns:

Term The name of the parameter, which are described in the sections for each model type. Some models contain an *intercept* or mean term. In those models, the related *constant estimate* is also shown. The definition of the constant estimate is given under the description of ARIMA models.

Factor (Shown only for multiplicative Seasonal ARIMA models.) The factor of the model that contains the parameter. In the multiplicative seasonal models, Factor 1 is nonseasonal and Factor 2 is seasonal.

Lag (Shown only for ARIMA and Seasonal ARIMA models.) The degree of the lag or backshift operator that is applied to the term to which the parameter is multiplied.

Estimate The parameter estimates of the time series model.

Std Error The estimates of the standard errors of the parameter estimates. These estimates are used to calculate tests and prediction intervals.

t Ratio The test statistics for the hypotheses that each parameter is zero. The test statistic for a parameter is the ratio of the parameter estimate to its standard error. If the hypothesis is true, then this statistic has an approximate Student’s *t* distribution. Looking for a *t*-ratio greater than 2 in absolute value is a common rule for judging significance because it approximates the 0.05 significance level.

Prob>|t| The observed p -value calculated for each parameter. The p -value is the probability of getting a t -ratio greater (in absolute value) than the computed value, given a true hypothesis.

Constant Estimate (Shown only for models that contain an intercept or mean term.) The definition of the constant estimate is given under ARIMA model.

Mu (Shown only for ARIMA and Seasonal ARIMA models.) The estimate for the intercept value of an ARIMA or seasonal ARIMA model.

Forecast A plot that shows both the observed and predicted values for the time series. The plot is divided by a vertical line into two regions.

If you specified a holdback set, a vertical line separates the training data from the holdback data. To the left of the vertical line, the one-step-ahead forecasts are overlaid with the training data points. To the right of the vertical line, the values forecast by the model are overlaid with the holdback data points. The prediction intervals for the forecast are also shown. You can control the number of forecast values using the Forecast Periods option in the platform launch window.

If you did not specify a holdback set, the vertical line separates the observed data from the future forecasts. To the left of the vertical line, the one-step-ahead forecasts are overlaid with the observed data points. To the right of the vertical line, the future values forecast by the model and the prediction intervals for the forecast are shown. You can control the number of future forecast values using the Forecast Periods option in the platform launch window or by selecting Number of Forecast Periods from the Time Series red triangle menu.

Residuals A graph that shows the values of the residuals based on the fitted model. These values are the observed values of the time series minus the one-step-ahead predicted values. The autocorrelation and partial autocorrelation reports for these residuals are also shown. These reports can be used to determine whether the fitted model is adequate to describe the data. If the fitted model is adequate, the points in the residual plot should be normally distributed about zero and the autocorrelation and partial autocorrelation of the residuals should not have any significant components for lags greater than zero.

Forecasting Errors on Holdback Series (Available only if a holdback set is specified.) A plot of the forecasting errors on the observations in the holdback set. The autocorrelation and partial autocorrelation reports for these errors are also shown.

Iteration History Contains the value of the objective function at each iteration. This can be useful for diagnosing problems with the fitting procedure. Attempting to fit a model that is poorly suited to the data can result in a failure to converge on an optimum value for the likelihood. The Iteration History table contains the following quantities:

Iter The iteration number.

Iteration History The objective function value for each step.

Step The type of iteration step.

Obj-Criterion The norm of the gradient of the objective function.

ARIMA and Seasonal ARIMA Model Report Options

Each ARIMA and Seasonal ARIMA model report has a red triangle menu that contains the following options:

Show Points Shows or hides the data points in the forecast graph.

Show Prediction Interval Shows or hides the prediction intervals in the forecast graph.

Save Columns Creates a new data table that contains columns that represent the results of the model.

Save Prediction Formula (Not available if a holdback set is specified.) Saves the data and prediction formula to a new data table.

Create SAS Job Creates SAS code that duplicates the model analysis in SAS.

Submit to SAS Submits SAS code to SAS that duplicates the model analysis. If you are not connected to a SAS server, this option guides you through the connection process.

Residual Statistics Controls the displays of residual statistics are shown for the model. These displays are described in the section [“Time Series Platform Options”](#). However, they are applied to the series of residuals.

Remove Fit Removes the individual model fit from the report.

State Space Smoothing Report

Each State Space Smoothing report in the Time Series platform has a unique outline name that identifies the model. Each model is added to the Model Comparison table. If you select the Report check box for a model in the Model Comparison table, a State Space Smoothing report appears. The following reports are included:

Model Type Describes the selected model for the series in terms of error, trend, and seasonality.

Model Summary Contains fit statistics for the model. In the formulas below, n is the length of the series and k is the number of fitted parameters in the model.

-2LogLikelihood Twice the negative log-likelihood function evaluated at the best-fit parameter estimates. Smaller values are better fits. See *Fitting Linear Models*.

AIC The Akaike's Information Criterion (AIC) value, computed as $-2\log\text{likelihood} + 2k$. Smaller AIC values indicate better fit.

BIC The Schwarz's Bayesian Criterion (SBC) value, computed as $-2\log\text{likelihood} + k*\ln(n)$. Smaller SBC values indicate better fit. Schwarz's Bayesian Criterion is equivalent to the Bayesian Information Criterion (BIC).

Nparm The number of parameters in the calculation of the information criteria. This number is the number of parameters in the likelihood function, which is the same as the number of parameters in the forecast formula plus one.

Note: Nparm is not the number of parameters in the parameter estimates table.

Sigma The estimate of the one-step-ahead forecast standard deviation.

RSquare The R-Square value, computed as follows:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

where

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$$

$$\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$\hat{y}_i$ are the one-step-ahead forecasts

$\bar{y}$ is the mean

If the model does not fit the series well, the model error sum of squares, SSE, might be larger than the total sum of squares, SST. As a result, R^2 can be negative.

RSquare Adj The adjusted R^2 value, computed as follows:

$$1 - \left[\frac{(n-1)}{(n-k)} (1 - R^2) \right]$$

MAPE The Mean Absolute Percentage Error value, computed as follows:

$$\frac{100}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

where

$\hat{y}_i$ are the one-step-ahead forecasts

MAE The Mean Absolute Error value, computed as follows:

$$\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

where

$\hat{y}_i$ are the one-step-ahead forecasts

Parameter Estimates A table that shows the estimate, standard error, and 95% confidence interval for each parameter in the model.

Forecast A plot that shows both the observed and predicted values for the time series. See [“Forecast”](#).

One-Step-Ahead Forecasting Errors A plot of the one-step-ahead forecasting errors over time. For each time point t , the error is calculated as the observed value of the time series at t minus the one-step-ahead predicted value at t . If a holdback set is specified, the errors are computed only for the training data. The autocorrelation and partial autocorrelation reports for these errors are also shown.

Forecasting Errors on Holdback Series (Available only if a holdback set is specified.) A plot of the forecasting errors on the observations in the holdback set. The autocorrelation and partial autocorrelation reports for these errors are also shown.

Component States A report that contains a plot for each state of the model and the corresponding formula for the model. Each plot includes the state estimates of the series and the state forecasts. If a holdback set is specified, the state estimates are for the training portion and the state forecasts are for the holdback portion. If a holdback set is not specified, the state estimates are for the portion up to the last observation of the time series and the state forecasts are for after the last observation of the time series. A model can have up to three states: Level, Trend, and Seasonal.

The Component States red triangle menu has a Save States option. Select this option to save the component states at each time point to a separate data table.

One-Step-Ahead Relative Forecasting Errors for Multiplicative Error Model (Available only for multiplicative error models.) A plot of the one-step-ahead relative forecasting errors over time. For each time point t , the error is calculated as the one-step-ahead-forecasting error at t divided by the one-step-ahead predicted value at t .

State Space Smoothing Report Options

Each state space smoothing report has a red triangle menu that contains the following options:

Show Points Shows or hides the data points in the forecast graph.

Show Prediction Interval Shows or hides the prediction intervals in the forecast graph.

Save Columns Creates a new data table that contains columns that represent the results of the model.

Remove Fit Removes the individual model fit from the report.

Transfer Function Model Report

Each Transfer Function Model report in the Time Series platform contains the following reports:

- Model Summary
- Parameter Estimates
- Residuals
- Forecasting Errors on Holdback Series
- Interactive Forecasting
- Iteration History

Each transfer function model is added to the Model Comparison table. If you select the Report check box for a transfer function model in the Model Comparison table, a Transfer Function Model report appears. If a holdback set is specified, a model plot is also available in the Model Comparison report.

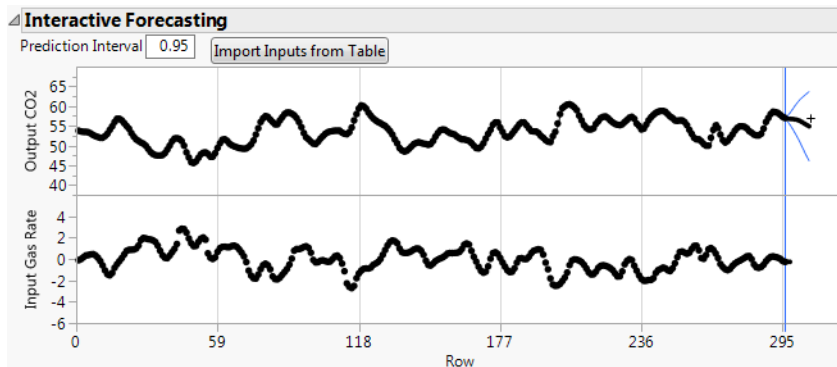
The information in the Model Summary, Parameter Estimates, Residuals, Forecasting Errors on Holdback Series, and Iteration History reports is the same as in the ARIMA and Seasonal ARIMA model report. For more information about these reports, see [“Model Report”](#). The Parameter Estimates table is followed by the formula of the model, where **B** is the backshift operator.

Interactive Forecasting

The Interactive Forecasting report provides a forecasting graph based on a prediction interval specified in the Transfer Function Model Specification window. The prediction interval around the prediction is shown in blue. Change the confidence level for this prediction interval by entering a number in the Prediction Interval box above the graph.

You can drag the plus sign in the graph to change the number of forecast periods in the graph. In the forecast periods, you can change the input values using the Import Inputs from Table button or by dragging points in the input graph to different values. The results of your changes are reflected in the forecast periods of the output graph.

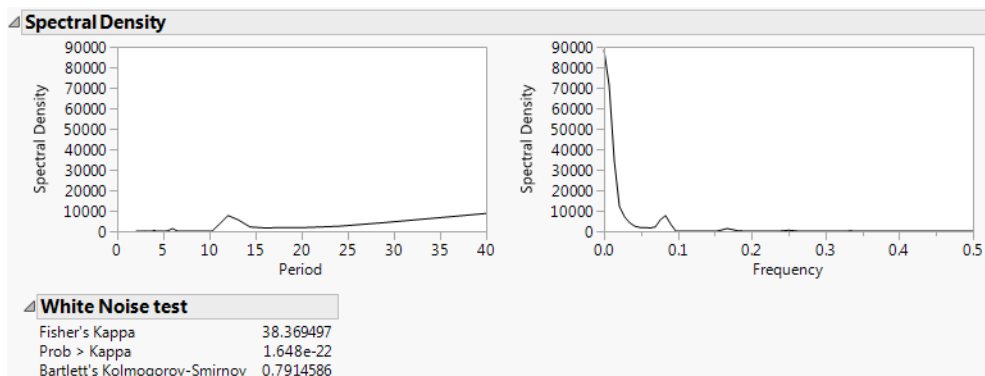
Figure 18.14 Interactive Forecasting Graph



Spectral Density Report

The Spectral Density report in the Time Series platform contains two spectral density plots and a White Noise Test table.

Figure 18.15 Spectral Density Plots and White Noise Test Report



The White Noise Test table contains the following statistics:

Fisher's Kappa Tests the null hypothesis that the values in the series are drawn from a normal distribution with variance 1 against the alternative hypothesis that the series has a periodic component. Kappa is the ratio of the maximum value of the periodogram, $I(f_i)$, and its average value.

Prob > Kappa The probability of observing a value larger than Kappa if the null hypothesis is true, given by the following equation:

$$Pr(k > \kappa) = 1 - \sum_{j=0}^q (-1)^j \binom{q}{j} \left[\max\left(1 - \frac{j\kappa}{q}, 0\right) \right]^{q-1}$$

where

$q = N / 2$ if N is even, $q = (N - 1) / 2$ if N is odd

κ is the observed value of Kappa

The null hypothesis is rejected if this probability is less than the significance level α .

Bartlett's Kolmogorov-Smirnov Compares the normalized cumulative periodogram to the cumulative distribution function of the uniform distribution on the interval (0, 1). The test statistic equals the maximum absolute difference of the cumulative periodogram and the uniform CDF. If it exceeds $a/(\sqrt{q})$, then one typically rejects the hypothesis that the series comes from a normal distribution. The values $a = 1.36$ and $a = 1.63$ correspond to significance levels 5% and 1% respectively.

Additional Examples of the Time Series Platform

This section contains examples using the Time Series platform.

- [“Example of Creating a Time ID Column”](#)
- [“Example of a Box-Cox Transformation”](#)
- [“Example Using a Holdback Set”](#)

Example of Creating a Time ID Column

In this example, you create a new column that is appropriate for the Time ID and then use the Time Series platform to perform a time series analysis.

Create Appropriate Time ID Column

1. Select **Help > Sample Data Folder** and open Time Series/SeriesP.jmp.

The SeriesP.jmp data table contains a Year column and a Quarter column to identify the time period during which the responses were observed. However, the Time Series platform requires one column with unique, equally spaced time points to label the X axis. If no Time ID is specified, then the row number is used to identify the time periods. To avoid this and make the report easier to interpret, you construct a Time ID column from Year and Quarter.

2. Select **Cols > New Columns**. In the Column Name box, type Year.Quarter.
3. Select **Best > Date > yyyyQq** next to Format.
4. Select **Column Properties > Formula**.
5. Click the gray triangle next to Date Time and click **Informat**.
6. Click the gray triangle next to Character.
7. Select Year, click **Char**, and then click **Concat**.
8. Type "Q" in the box (including quotation marks) and press Enter.
9. Click **Concat**.
10. Select Quarter and click **Char**.
11. Click the <formatName> box, type yyyyQq, and press Enter.
12. Click **OK**.

The completed New Column window should appear as in [Figure 18.16](#).

Figure 18.16 New Column

'Year.Quarter' in table 'SeriesP'

Column Name: Year.Quarter

☒ Lock

Data Type: Numeric

Modeling Type: Continuous

Format: yyyyQq Width: 10

Input Format: yyyyQq

Column Properties: Formula

Remove

Formula

Edit Formula ☐ Suppress Eval ☐ Ignore Errors

Informat

Char (Year, ^) || "Q" || Char (Quarter, ^) yyyyQq ~

13. Click **OK**.

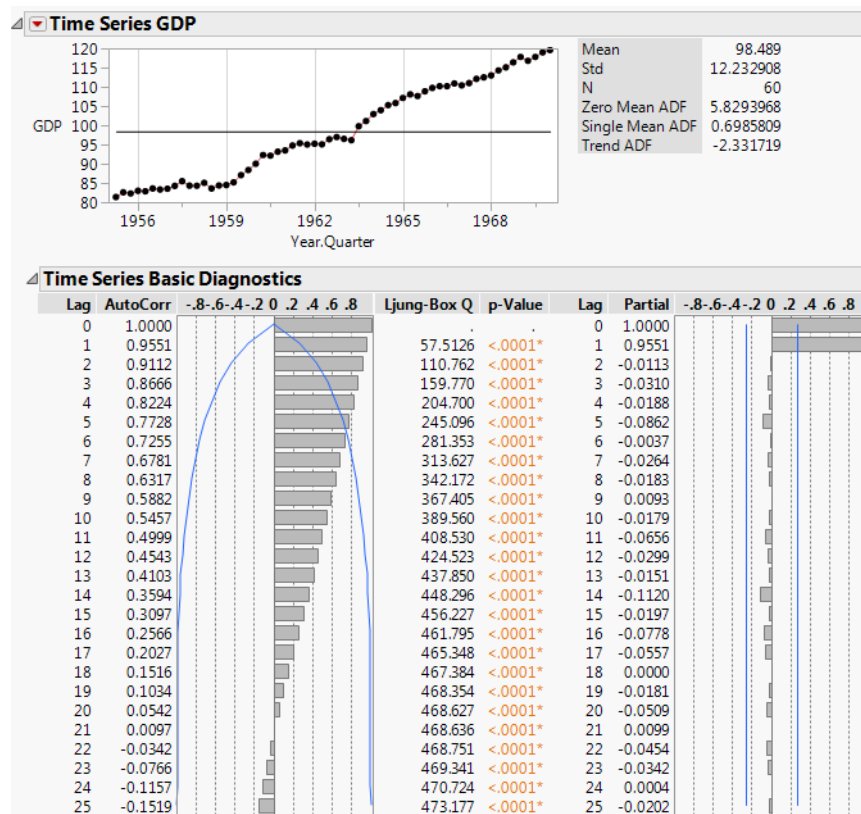
Note: This time column also works for X11 analyses.

Time Series Analysis

Now that the data table contains an appropriate Time ID column, proceed with the analysis.

1. Select **Analyze > Specialized Modeling > Time Series**.
2. Select GDP and click **Y, Time Series**.
3. Select Year.Quarter and click **X, Time ID**.
4. Click **OK**.

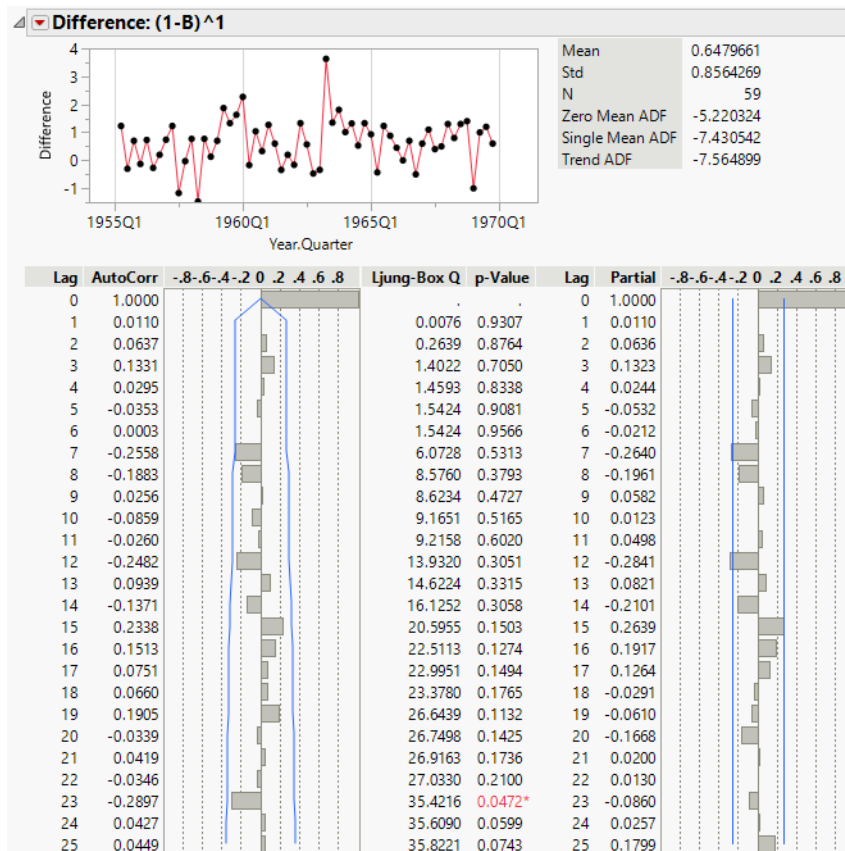
Figure 18.17 Time Series Report for SeriesP.jmp



The series shows an increasing trend over time that is fairly linear. In addition, the autocorrelation chart shows that there is strong correlation between points that are close together. The AutoCorr values for points with lags of 1, 2, and 3 are 0.9551, 0.9112, 0.8666, respectively.

5. Click the Time Series GDP red triangle and select **Difference**.
6. Select 1 for the Nonseasonal Differencing Order and click **Estimate**.

Figure 18.18 Difference Report for SeriesP.jmp



The Difference report helps determine an appropriate model to be fit to the original time series. The plot of differences shows that the differenced series no longer has the trend that was observed in the original data. This indicates that lag-1 differencing is an appropriate choice. Also, even after removing the trend, the series shows no sign of seasonality. For these reasons, models to fit the original series should be able to handle linear trends, but do not necessarily need to handle seasonality. Linear exponential smoothing and ARIMA models would be appropriate.

- Click the Time Series GDP red triangle and select **Smoothing Model > Linear Exponential Smoothing**.
- Click **Estimate**.
- Click the Time Series GDP red triangle and select **ARIMA Model Group**. This enables you to fit multiple ARIMA models for a range of values of $(p,d,q)(P,D,Q)$.
- In the ARIMA box, set the following ranges:

- Fix d , the differencing order, at 1 by setting the range from 1 to 1 because the differencing report showed lag-1 differencing was appropriate.
- Set p , the autoregressive order, to range from 0 to 1 because the original series showed evidence of autocorrelation.
- Set q , the moving average order, to range from 0 to 1.

Note: In most cases, it is sufficient to keep p and q small.

- Leave P , D , and Q set at 0, since the series showed no evidence of seasonality.
- These settings lead to the fitting of four total models.

Figure 18.19 ARIMA Model Group Specification

Specify ARIMA Model

ARIMA

p, Autoregressive Order: 0 1
d, Differencing Order: 1 1
q, Moving Average Order: 0 1

Seasonal ARIMA

P, Autoregressive Order: 0 0
D, Differencing Order: 0 0
Q, Moving Average Order: 0 0
Observations per Period: 12 12

Prediction Interval: 0.95

☒ Intercept
☒ Constrain fit

Total Number of Models: 4

Estimate Cancel Help

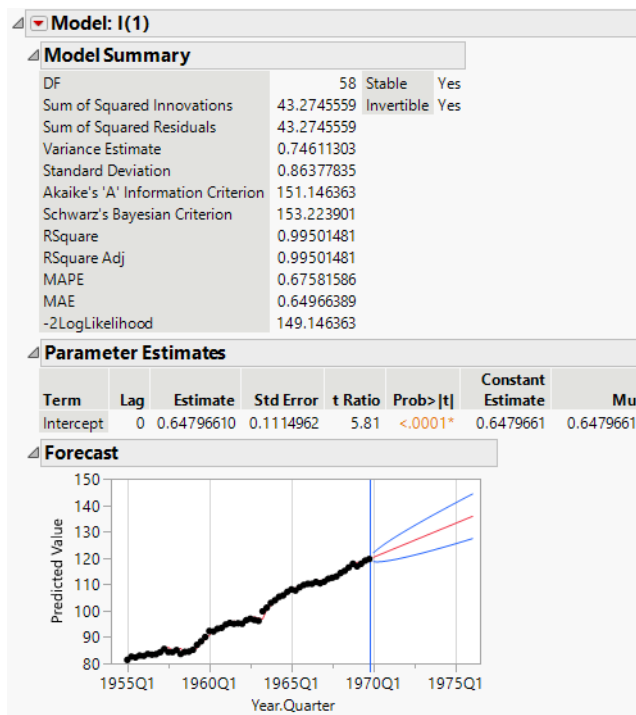
11. Click **Estimate**.

Figure 18.20 Model Comparison Table

Model Comparison													
Report	Graph	Model	DF	Variance	AIC	SBC	RSquare	-2LogLH	Weights	.2	.4	.6	.8
☒	☐	I(1)	58	0.746113	151.14636	153.22390	0.995	149.14636	0.477865				
☒	☐	ARI(1, 1)	57	0.7591096	153.13924	157.29432	0.995	149.13924	0.176424				
☒	☐	IMA(1, 1)	57	0.7591199	153.14002	157.29510	0.995	149.14002	0.176355				
☒	☐	ARIMA(1, 1, 1)	56	0.748633	153.91882	160.15143	0.995	147.91882	0.119474				
☒	☐	Linear (Holt) Exponential Smoothing	56	0.7878954	155.66571	159.78660	0.994	151.66571	0.049882				
												MAPE	MAE
												0.675816	0.649664
												0.676216	0.649960
												0.676179	0.649933
												0.668189	0.641981
												0.716452	0.686737

The Model Comparison Table is sorted such that the best fitting model, according to the AIC criterion, is at the top of the list. In this case, the ARIMA(0,1,0) model (denoted I(1) in the report) best fits the original time series. It should also be noted that although the I(1) model is “best,” all of the models have extremely similar values for the fit statistics. They could all be considered appropriate.

Figure 18.21 Model Report for ARIMA(0,1,0)



The model report for I(1) shows the forecast graph. The blue lines indicate the prediction intervals. GDP is predicted to continue increasing at a linear rate.

Example of a Box-Cox Transformation

In this example, you compare two time series analyses. In one analysis, you use a Box-Cox transformation and in the other, you analyze a manually transformed time series. When the lambda value is set to zero, the Box-Cox transformation is an approximate log transformation. However, using the Box-Cox transformation for the analysis makes the results easier to interpret since all forecasts are transformed back and reported on the original scale.

1. Select **Help > Sample Data Folder** and open Time Series/Seriesg.jmp.
2. Select **Analyze > Specialized Modeling > Time Series**.
3. Select Passengers and click **Y, Time Series**.
4. Select Time and click **X, Time ID**.
5. Select **Use Box-Cox Transformation**.
Use the default value of 0 for Lambda.
6. Click **OK**.

7. Click the red triangle next to Time Series Box-Cox Transformed Passengers and select **Seasonal ARIMA**.
8. In the ARIMA section of the Seasonal ARIMA Specification window, set the following parameters.
 - Set p , Autoregressive Order to 0.
 - Set d , Differencing Order to 1
 - Set q , Moving Average Order to 1
9. In the Seasonal ARIMA section of the Seasonal ARIMA Specification window, set the following parameters.
 - Set P , Autoregressive Order to 0
 - Set D , Differencing Order to 1
 - Set Q , Moving Average Order to 1
10. Deselect Intercept to run a model without an intercept.

Your window should now match the window shown in [Figure 18.22](#).

Figure 18.22 Seasonal ARIMA Specifications

Specify ARIMA Model

ARIMA		Seasonal ARIMA	
p, Autoregressive Order	0	P, Autoregressive Order	0
d, Differencing Order	1	D, Differencing Order	1
q, Moving Average Order	1	Q, Moving Average Order	1
		Observations per Period	12

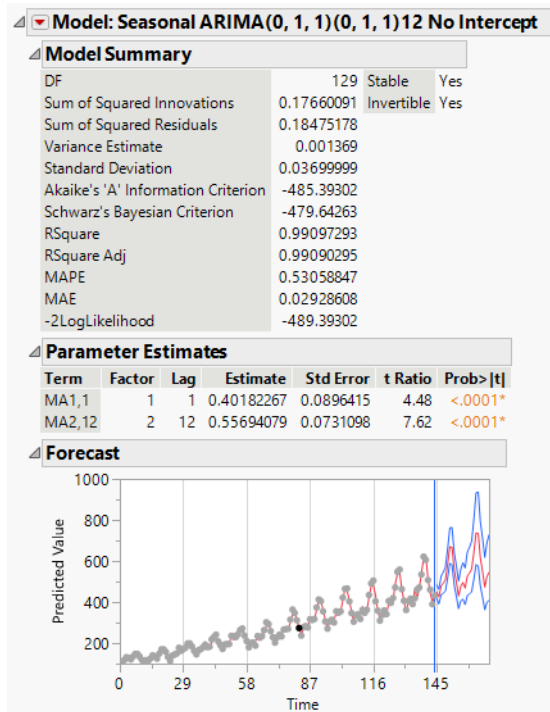
Prediction Interval: 0.95

☐ Intercept
☒ Constrain fit

Estimate Cancel Help

11. Click **Estimate**.

Figure 18.23 Seasonal ARIMA Model for Box-Cox Transformed Passengers Data



The parameter estimates are equivalent to those obtained from running the same model on the Log Passengers column. However, the forecasts and prediction intervals in [Figure 18.23](#) are on the original scale.

Example Using a Holdback Set

In this Time Series example, use a holdback set to evaluate predictions made on monthly data on lead production. Since a holdback set is used, both State Space Smoothing and ARIMA models can be fit and compared to find the best fitting model for the data.

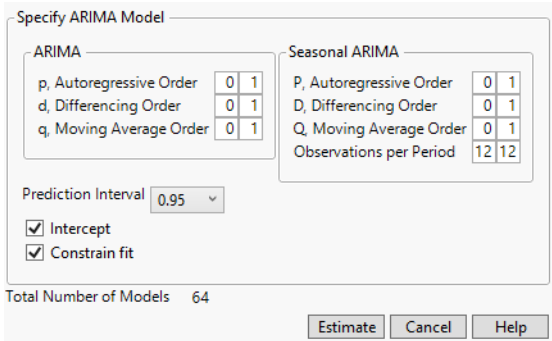
1. Select **Help > Sample Data Folder** and open Time Series/Lead Production.jmp.
2. Select **Analyze > Specialized Modeling > Time Series**.
3. Select LEADPROD and click **Y, Time Series**.
4. Select DATE and click **X, Time ID**.
5. Select **Forecast on Holdback**.

Selecting this options means that forecasts are made on the holdback set instead of on future observations.

6. Enter 12 next to Forecast Periods.

- This assigns the last 12 observations in the data to the holdback set. Since this is monthly data, this is equivalent to the last year.
- Click **OK**.
 - Click the Time Series LEADPROD red triangle and select **State Space Smoothing**.
 - Type 12 next to Period.
 - Select Constrain Parameters.
 - Click **OK**.
 - Click the Time Series LEADPROD red triangle and select **ARIMA Model Group**.
 - Set the range for all parameters from 0 to 1. The Specify ARIMA Model window looks like the one in [Figure 18.24](#).

Figure 18.24 ARIMA Model Group Specifications



The 'Specify ARIMA Model' dialog box is shown. It has two tabs: 'ARIMA' and 'Seasonal ARIMA'. The 'ARIMA' tab is active, showing parameters: p, Autoregressive Order (0 to 1); d, Differencing Order (0 to 1); q, Moving Average Order (0 to 1). The 'Seasonal ARIMA' tab shows: P, Autoregressive Order (0 to 1); D, Differencing Order (0 to 1); Q, Moving Average Order (0 to 1); Observations per Period (12 to 12). Below the tabs, there is a 'Prediction Interval' dropdown set to '0.95', and two checked checkboxes: 'Intercept' and 'Constrain fit'. At the bottom, it says 'Total Number of Models 64' and has 'Estimate', 'Cancel', and 'Help' buttons.

- Click **Estimate**.

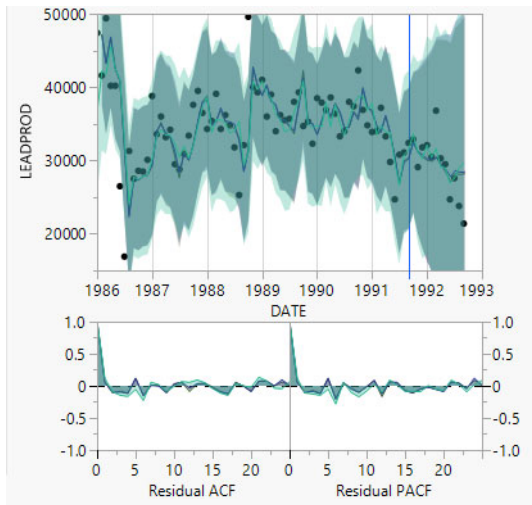
Figure 18.25 Model Comparison

Model Comparison							
Report	Graph	Model Class	Model	Holdback Evaluation			
				RMSE ^	MSE	MAPE	MAE
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 1)(1, 0, 0)12	3219.4131	10364621	8.900164	2416.3319
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 1)(0, 0, 1)12	3316.8746	11001657	9.148390	2456.0873
☐	☐	State Space Smoothing	AAA12 Constrained	3350.9303	11228734	8.623519	2245.5657
☐	☐	ARIMA	Seasonal ARIMA(1, 1, 0)(1, 0, 0)12	3397.0855	11540190	9.082427	2392.5823
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 1)(1, 1, 0)12	3448.6372	11893099	9.829459	2649.8142
☐	☐	ARIMA	Seasonal ARIMA(1, 1, 0)(1, 1, 0)12	3484.3764	12140879	9.708056	2604.6290
☐	☐	ARIMA	Seasonal ARIMA(1, 1, 0)(0, 0, 1)12	3491.2648	12188930	9.302681	2435.7441
☐	☐	State Space Smoothing	MAM12 Constrained	3512.1513	12335206	10.552196	2813.4655
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 0)(1, 0, 0)12	3525.0858	12426230	9.374006	2440.6892
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 0)(1, 1, 0)12	3551.3973	12612423	9.367947	2480.3210
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 1)(1, 0, 1)12	3579.0242	12809414	10.520274	2739.9652
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 0)(0, 1, 0)12	3597.3098	12940638	11.622868	3219.0476
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 0)(0, 0, 1)12	3631.8176	13190099	9.792052	2544.4269
☐	☐	ARIMA	Seasonal ARIMA(1, 1, 0)(0, 1, 0)12	3661.8395	13409068	12.030672	3366.0935
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 1)(0, 1, 0)12	3673.6369	13495608	12.063523	3379.4644
☐	☐	ARIMA	Seasonal ARIMA(1, 1, 0)(1, 0, 1)12	3700.6621	13694900	11.077791	2890.6426
☐	☐	ARIMA	Seasonal ARIMA(0, 1, 0)(1, 0, 1)12	3723.7664	13866437	11.222827	2935.8075

The fit statistics are comparable between the State Space Smoothing models and the ARIMA models because they are evaluated on the holdback set.

15. Select the Graph check box next to the first three models in the Model Comparison report.

Figure 18.26 Model Comparison Graph



The evaluation statistics and the forecasting plots show that there is no substantial differences in forecasting performance among the top models. Based on the report, either an ARIMA model or a State Space Smoothing model is appropriate for this data.

Statistical Details for the Time Series Platform

This section contains statistical details for the Time Series platform.

- [“Statistical Details for Spectral Density”](#)
- [“Statistical Details for X-11 Decomposition”](#)
- [“Statistical Details for Exponential Smoothing Models”](#)
- [“Statistical Details for ARIMA Models”](#)
- [“Statistical Details for Transfer Functions”](#)

Statistical Details for Spectral Density

In the Time Series platform, the least squares estimates of the coefficients of the Fourier series are computed as follows:

$$a_t = \frac{2}{N} \sum_{i=1}^N y_i \cos(2\pi f_i t)$$

$$b_t = \frac{2}{N} \sum_{i=1}^N y_i \sin(2\pi f_i t)$$

Then the $f_i = i/N$ are combined to form the periodogram $I(f_i) = (N/2)(a_i^2 + b_i^2)$, which represents the intensity at frequency f_i .

The periodogram is then smoothed and scaled by $1/(4\pi)$ to form the spectral density.

Statistical Details for X-11 Decomposition

In the Time Series platform, the X-11 Decomposition method adjusts the original time series using either a multiplicative or an additive decomposition. The model is fit using an iterative process to estimate the three X-11 components: trend cycle, seasonal, and irregular. The trend cycle component contains both the long-term trend and the long-term cyclical effects. The irregular component contains the effects of variation unexplained by the trend and seasonal components. For a historical overview of the development of the X-11 method, see *SAS/ETS 15.2 User's Guide* (search for "Historical Development of X-11").

The multiplicative adjustment fits the following model:

$$O_t = C_t \cdot S_t \cdot I_t$$

where

O_t is the original time series

C_t is the trend cycle component

S_t is the seasonal component

I_t is the irregular component

The adjusted multiplicative trend is O_t/S_t .

The additive adjustment fits the following model:

$$O_t = C_t + S_t + I_t$$

The adjusted additive trend is $O_t - S_t$.

Statistical Details for Exponential Smoothing Models

Smoothing models are defined as follows:

$$y_t = \mu_t + \beta_t t + s(t) + a_t$$

where

μ_t is the time-varying mean term

β_t is the time-varying slope term

$s(t)$ is one of the s time-varying seasonal terms

a_t are the random shocks

Models without a trend have $\beta_t = 0$ and nonseasonal models have $s(t) = 0$. The estimators for these time-varying terms are defined as follows:

L_t is a smoothed level that estimates μ_t

T_t is a smoothed trend that estimates β_t

S_{t-j} for $j = 0, 1, \dots, s - 1$ are the estimates of the $s(t)$

Each smoothing model defines a set of recursive smoothing equations that describe the evolution of these estimators. The smoothing equations are written in terms of model parameters called *smoothing weights*:

α is the level smoothing weight

γ is the trend smoothing weight

ϕ is the trend damping weight

δ is the seasonal smoothing weight

While these parameters enter each model in a different way (or not at all), they have the common property that larger weights give more influence to recent data while smaller weights give less influence to recent data. For more information on smoothing weights, see *SAS/ETS 15.2 User's Guide* (the Forecasting Process Details chapter).

Simple Exponential Smoothing

The model for simple exponential smoothing is $y_t = \mu_t + \alpha_t$.

The smoothing equation, $L_t = \alpha y_t + (1 - \alpha)L_{t-1}$, is defined in terms of a single smoothing weight α . This model is equivalent to an ARIMA(0, 1, 1) model where the following is true:

$$(1 - B)y_t = (1 - \theta B)\alpha_t \text{ where } \theta = 1 - \alpha$$

The moving average form of the model is defined as follows:

$$y_t = a_t + \sum_{j=1}^{\infty} \alpha a_{t-j}$$

Double (Brown) Exponential Smoothing

The model for double exponential smoothing is $y_t = \mu_t + \beta_1 t + a_t$.

The smoothing equations, defined in terms of a single smoothing weight α , are defined as follows:

$$L_t = \alpha y_t + (1 - \alpha)L_{t-1}$$

$$T_t = \alpha(L_t - L_{t-1}) + (1 - \alpha)T_{t-1}$$

This model is equivalent to an ARIMA(0, 1, 1)(0, 1, 1)₁ model where the following is true:

$$(1 - B)^2 y_t = (1 - \theta B)^2 a_t \text{ where } \theta_{1,1} = \theta_{2,1} \text{ with } \theta = 1 - \alpha$$

The moving average form of the model is defined as follows:

$$y_t = a_t + \sum_{j=1}^{\infty} (2\alpha + (j-1)\alpha^2) a_{t-j}$$

Linear (Holt) Exponential Smoothing

The model for linear exponential smoothing is $y_t = \mu_t + \beta_1 t + a_t$.

The smoothing equations, in terms of smoothing weights α and γ , are defined as follows:

$$L_t = \alpha y_t + (1 - \alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1}$$

This model is equivalent to an ARIMA(0, 2, 2) model where the following is true:

$$(1 - B)^2 y_t = (1 - \theta B - \theta_2 B^2) a_t \text{ with } \theta = 2 - \alpha - \alpha\gamma \text{ and } \theta_2 = \alpha - 1$$

The moving average form of the model is defined as follows:

$$y_t = a_t + \sum_{j=1}^{\infty} (\alpha + j\alpha\gamma) a_{t-j}$$

Damped-Trend Linear Exponential Smoothing

The model for damped-trend linear exponential smoothing is $y_t = \mu_t + \beta_t t + a_t$.

The smoothing equations, in terms of smoothing weights α , γ , and ϕ , are defined as follows:

$$L_t = \alpha y_t + (1 - \alpha)(L_{t-1} + \phi T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)\phi T_{t-1}$$

This model is equivalent to an ARIMA(1, 1, 2) model where the following is true:

$$(1 - \phi B)(1 - B)y_t = (1 - \theta_1 B - \theta_2 B^2)a_t$$

where

$$\theta_1 = 1 + \phi - \alpha - \alpha\gamma\phi$$

$$\theta_2 = (\alpha - 1)\phi$$

The moving average form of the model is defined as follows:

$$y_t = \alpha_t + \sum_{j=1}^{\infty} \left(\frac{\alpha + \alpha\gamma\phi(\phi^j - 1)}{\phi - 1} \right) \alpha_{t-j}$$

Seasonal Exponential Smoothing

The model for seasonal exponential smoothing is $y_t = \mu_t + s(t) + a_t$.

The smoothing equations in terms of smoothing weights α and δ are defined as follows:

$$L_t = \alpha(y_t - S_{t-s}) + (1 - \alpha)L_{t-1}$$

$$S_t = \delta(y_t - L_{t-s}) + (1 - \delta)S_{t-s}$$

This model is equivalent to a seasonal ARIMA(0, 1, s+1)(0, 1, 0)_s model:

$$(1 - B)(1 - B^s)y_t = (1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^{s+1})a_t$$

where

$$\theta_1 = 1 - \alpha$$

$$\theta_2 = (1 - \delta)(1 - \alpha)$$

$$\theta_3 = (1 - \alpha)(\delta - 1)$$

The moving average form of the model is defined as follows:

$$y_t = a_t + \sum_{j=1}^{\infty} \psi_j a_{t-j} \text{ where } \psi = \begin{cases} \alpha \text{ for } j \bmod s \neq 0 \\ \alpha + \delta(1 - \alpha) \text{ for } j \bmod s = 0 \end{cases}$$

Winters Method (Additive)

The model for the additive version of the Winters method is $y_t = \mu_t + \beta_t t + s(t) + a_t$.

The smoothing equations in terms of weights α , γ , and δ are defined as follows:

$$L_t = \alpha(y_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + T_{t-1})$$

$$T_t = \gamma(L_t - L_{t-1}) + (1 - \gamma)T_{t-1}$$

$$S_t = \delta(y_t - L_t) + (1 - \delta)S_{t-s}$$

This model is equivalent to a seasonal ARIMA(0, 1, s+1)(0, 1, 0)_s model defined as follows:

$$(1 - B)(1 - B^2)y_t = \left(1 - \sum_{i=1}^{s+1} \theta_i B^i\right) a_t$$

The moving average form of the model is defined as follows:

$$y_t = a_t + \sum_{j=1}^{\infty} \Psi_j a_{t-j}$$

where

$$\Psi = \begin{cases} \alpha + j\alpha\gamma, & j \bmod s \neq 0 \\ \alpha + j\alpha\gamma + \delta(1 - \alpha), & j \bmod s = 0 \end{cases}$$

Statistical Details for ARIMA Models

The Time Series platform fits both ARIMA and Seasonal ARIMA models.

ARIMA Model

For a response series $\{y_t\}$, the general form for the ARIMA model is defined as follows:

$$\phi(B)(w_t - \mu) = \theta(B)a_t$$

where

t is the time index

B is the backshift operator defined as $By_t = y_{t-1}$

$w_t = (1 - B)^d y_t$ is the response series after differencing

μ is the intercept or mean term

$\phi(B)$ and $\theta(B)$ are the autoregressive operator and the moving average operator, respectively, and are defined as follows:

$$\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p$$

$$\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$

where

a_t are the sequence of random shocks

The a_t are assumed to be independent and normally distributed with mean zero and constant variance.

The model can be rewritten as follows:

$$\phi(B)w_t = \delta + \theta(B)a_t$$

The constant estimate δ is given by the relation:

$$\delta = \phi(B)\mu = \mu - \phi_1 \mu - \phi_2 \mu - \dots - \phi_p \mu$$

Seasonal ARIMA Model

In the case of Seasonal ARIMA modeling, the differencing, autoregressive, and moving average operators are the product of seasonal and nonseasonal polynomials:

$$w_t = (1 - B)^d (1 - B^s)^D y_t$$

$$\phi(B) = (1 - \phi_{1,1} B - \phi_{1,2} B^2 - \dots - \phi_{1,p} B^p)(1 - \phi_{2,s} B^s - \phi_{2,2s} B^{2s} - \dots - \phi_{2,p_s} B^{p_s})$$

$$\theta(B) = (1 - \theta_{1,1}B - \theta_{1,2}B^2 - \dots - \theta_{1,q}B^q)(1 - \theta_{2,s}B^s - \theta_{2,2s}B^{2s} - \dots - \theta_{2,Qs}B^{Qs})$$

where s is the number of observations per period. The first index on the coefficients is the factor number (1 indicates nonseasonal, 2 indicates seasonal) and the second is the lag of the term.

Statistical Details for Transfer Functions

A typical transfer function model with m inputs can be represented as follows:

$$Y_t - \mu = \frac{\omega_1(B)}{\delta_1(B)}X_{1,t-d1} + \dots + \frac{\omega_m(B)}{\delta_m(B)}X_{m,m-dm} + \frac{\theta(B)}{\varphi(B)}e_t$$

where

Y_t denotes the output series

X_1 to X_m denote m input series

e_t represents the noise series

$X_{1,t-d1}$ indicates the series X_1 is indexed by t with a $d1$ -step lag

μ represents the mean level of the model

$\varphi(B)$ and $\theta(B)$ represent autoregressive and moving average polynomials from an ARIMA model

$\omega_k(B)$ and $\delta_k(B)$ represent numerator and denominator factors (or polynomials) for individual transfer functions, with k representing an index for the 1 to m individual inputs.

Each polynomial in the above model can contain two parts, either nonseasonal, seasonal, or a product of the two as in seasonal ARIMA. When specifying a model, leave the default 0 for any part that you do not want to include in the model.

Chapter 19

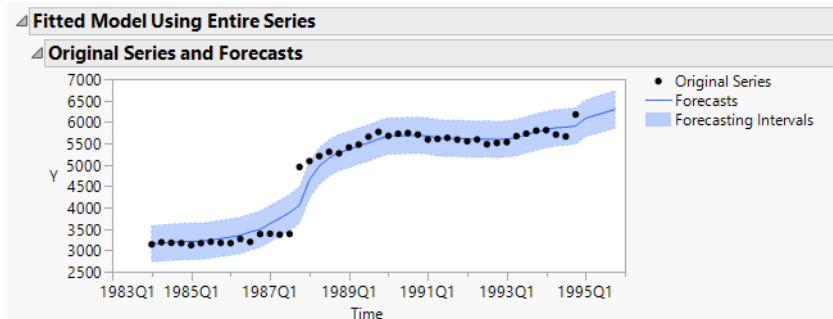
Time Series Forecast

Forecast Multiple Time Series

The Time Series Forecast platform enables you to forecast multiple time series, where the number of series could be very large. The platform provides compact summaries of the multiple time series and provides options for efficiently fitting up to 30 different forecasting models in the exponential smoothing family. The best fitting model is automatically selected.

The Time Series Forecast platform is designed to select a best fitting model automatically from a set of exponential smoothing models. This is in contrast to the Time Series platform which requires manual selection of a best fitting model and uses ARIMA models.

Figure 19.1 Original Series and Forecasts



Contents

Example of Time Series Forecast..... 407

Launch the Time Series Forecast Platform 409

 Data Format 409

The Time Series Forecast Report..... 411

 Analysis of Time Pattern 411

 Group Summary of Time 411

 Modeling Specifications..... 413

Time Series Forecast Platform Options 416

Time Series Forecast Model Reports 417

Example of Time Series Forecast

Use the Time Series Forecast platform to fit a variety of state space smoothing models to data with multiple time series. The final model is selected through the Forecasting Performance option, which utilizes a holdback set.

1. Select **Help > Sample Data Folder** and open Time Series/M3C Quarterly.jmp.
2. Select **Analyze > Specialized Modeling > Time Series Forecast**.
3. Select Y and click **Y**.
4. Select Series and click **Grouping**.
5. Select Time and click **Time**.
6. Click **OK**.

The initial Time Series Forecast report appears. It contains summaries of the Time and Y variables. There is also a Modeling Specifications report that provides options for fitting forecast models.

Figure 19.2 Initial Report for Time Series Forecast

Time Series Forecast

Analysis of Time Pattern

Date Resolution

Quarterly data.

Suggested Seasonality

4

Group Summary of Time

Series	N Rows	Missing Value	Unique Timestamps	Start	End	Missing Record	
N 646	44	No	Yes	1984Q1	1994Q4	No	^
N 647	44	No	Yes	1984Q1	1994Q4	No	
N 648	46	No	Yes	1983Q3	1994Q4	No	
N 649	44	No	Yes	1984Q1	1994Q4	No	
N 650	44	No	Yes	1984Q1	1994Q4	No	
N 651	44	No	Yes	1984Q1	1994Q4	No	
N 652	43	No	Yes	1984Q1	1994Q3	No	
N 653	43	No	Yes	1984Q1	1994Q3	No	
N 654	45	No	Yes	1983Q4	1994Q4	No	
N 655	44	No	Yes	1984Q1	1994Q4	No	^

Group Summary of Y

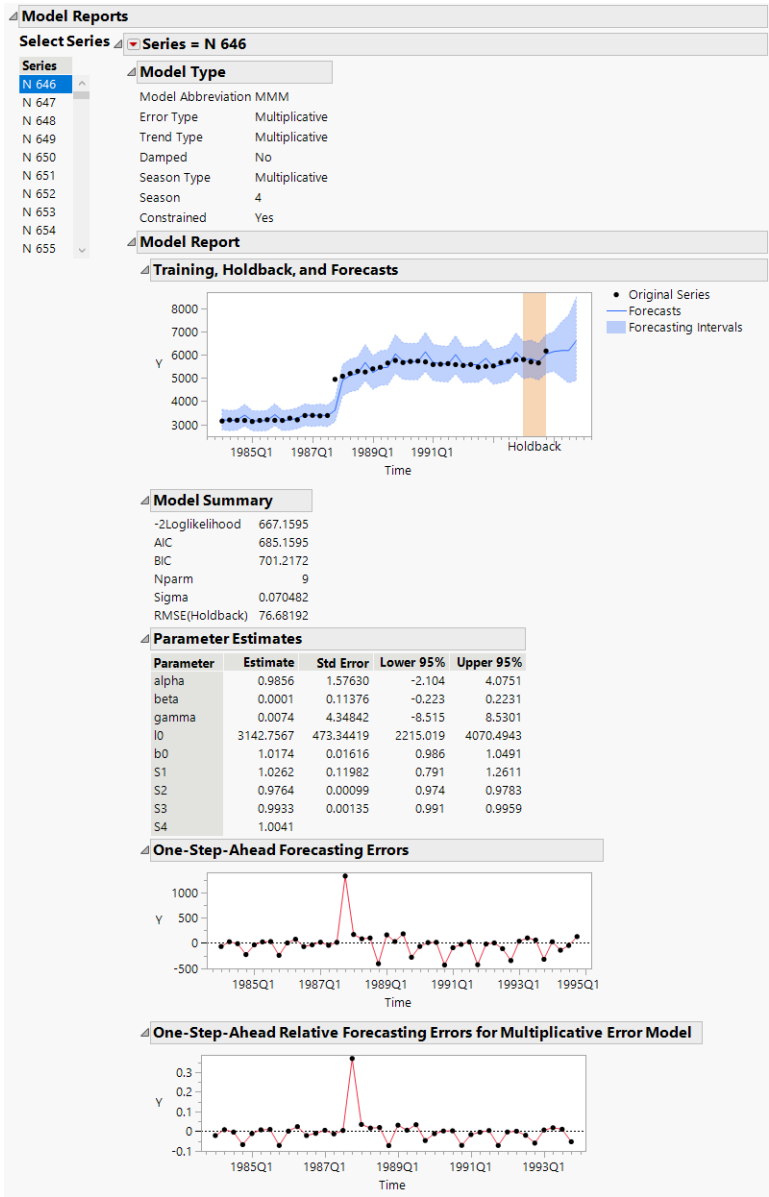
Series	Min	Max	Length	N Obs	N Missing	All Positive	Comment	
N 646	3124.38	6176.6	44	44	0	Yes		^
N 647	1522	4320.8	44	44	0	Yes		
N 648	1549.72	5970	46	46	0	Yes		
N 649	2136.7	5023.2	44	44	0	Yes		
N 650	875.51	6908	44	44	0	Yes		
N 651	7512.4	8526.2	44	44	0	Yes		
N 652	4861.63	9913.57	43	43	0	Yes		
N 653	2797	7202	43	43	0	Yes		
N 654	2372.74	6735.1	45	45	0	Yes		
N 655	2449.06	8904.28	44	44	0	Yes		^

7. In the Model Specifications report, click the Complete Specifications tab.
8. Click the menu below Model Selection Strategy and select Forecasting Performance.

9. Click **Run**.

This fits a set of recommended models to the data. The best model for each level of the grouping variable is chosen and reported in the Model Reports. See “[Model Reports](#)”.

Figure 19.3 Model Reports



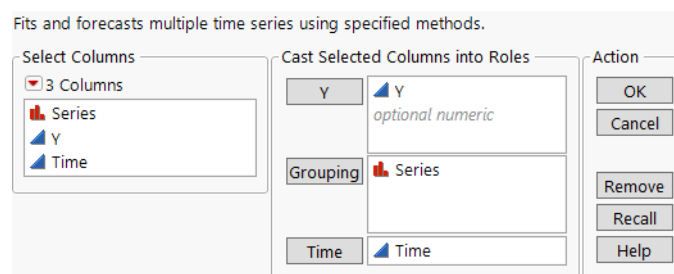
The model report for one series, Series N646, is shown. This corresponds to the series name that is selected in the Select Series list. You can use the down arrow key to scan

through the model reports for each series. Alternatively, you can click a different name in the list to display the model report for the selected series.

Launch the Time Series Forecast Platform

Launch the Time Series Forecast platform by selection **Analyze > Specialized Modeling > Time Series Forecast**.

Figure 19.4 The Time Series Forecast Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The Time Series Forecast platform launch window contains the following options:

Y Assigns one or more columns as time series variables. Multiple time series can be organized as individual columns in a table or they can be stacked in a single column with one or more grouping variables.

Grouping Assigns one or more columns as grouping variables. Grouping variables are used to identify individual time series when they are in stacked format.

Time Assigns a column for the time variable. If the individual time series are organized as individual columns, the timestamps in the Time column must not contain any duplicates.

Data Format

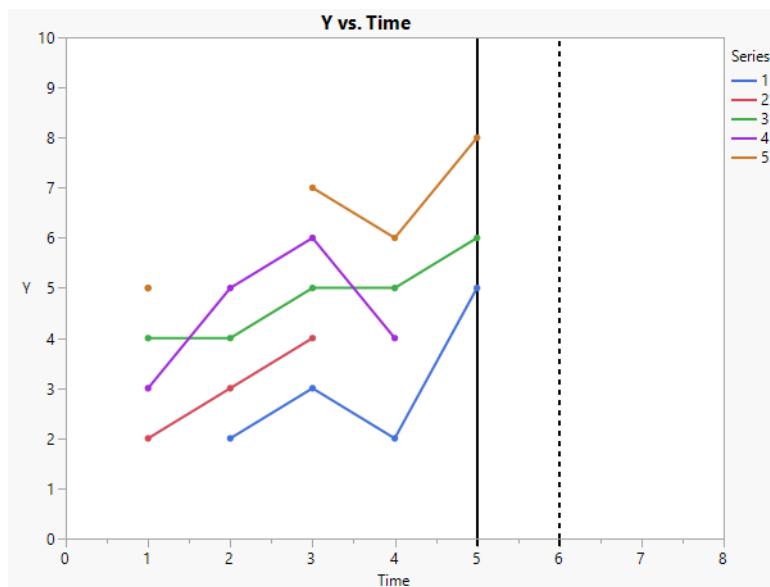
When using the Time Series Forecast platform, multiple time series can be organized as individual columns in a table or they can be stacked in a single column with one or more grouping columns to identify the individual series. You can also have a combination of multiple Y columns and multiple grouping columns.

The platform makes the following assumptions about the data:

- All of the time series are recorded at the same time resolution.

- All of the time series are collected up to the same final observed time point. This final observed time point is defined as the latest time point that occurs in any of the individual time series. A series without data up to the final observed time point is assumed to have missing values. For example, in [Figure 19.5](#), the final observed time point is 5, represented by the solid black line. It is assumed that Series 2 has missing observations at time points 4 and 5 and that Series 4 has a missing observation at time point 5.
- All of the forecasts start at the same time point. This starting time point is the next time point following the final observed time point. For example, in [Figure 19.5](#), the final observed time point is 5, represented by the solid black line. Therefore, all of the forecasts start at time point 6, represented by the dashed black line.

Figure 19.5 Demonstration of Assumptions



The Time Series Forecast Report

The Time Series Forecast report contains summary tables based on the variables specified in the launch window. A summary table of the time series by group is always provided. If a Time column is specified, the report also contains a summary of the time pattern and a summary table of the time by group. The report also contains a Modeling Specifications section that enables you to fit models to each time series.

- [“Analysis of Time Pattern”](#)
- [“Group Summary of Time”](#)
- [“Modeling Specifications”](#)

Analysis of Time Pattern

If a time column is specified, the Time Series Forecast platform analyzes the entries of the column to determine a pattern. The platform is designed to recognize a variety of time patterns, including daily, business daily, monthly, bi-weekly, quarterly, and so on. The pattern is reported as the Date Resolution. A Suggested Seasonality is also provided based on the time pattern.

Group Summary of Time

If a time column is specified, the Time Series Forecast report contains a section of summaries of the individual time series. The summaries contain the following information for each time series:

<Grouping Column Name> (Available only if a Grouping variable is specified.) The level of the grouping variable that defines the individual series when data is provided in the stacked format.

N Rows The number of observations in the series.

Missing Value Indicates whether any of the time values are missing values.

Unique Timestamps Indicates whether the time stamps are unique.

Start The starting time stamp.

End The ending time stamp.

Missing Record Indicates whether there are observations missing from the data table, based on the starting and ending time stamps and the data resolution/pattern.

Tip: The Group Summary of Time table is linked to the Select Series list. When you select a series in the Group Summary of Time table, the corresponding entries are also selected in the Select Series list.

The Time Series Forecast report includes a summary table for each Y variable. The summaries contain the following information for each time series:

Variable (Available only if there are multiple Y columns and at least one grouping column.)
The column name of the time series variables.

Time Series (Available only if there are multiple series.) The name of the individual time series. When the data are provided in the stacked format, this is the level of the grouping variable. When the data are provided in the wide format, with multiple columns, this is the column name of the individual series.

Note: When the data are provided in the stacked format, the column heading in the table is the name of the Grouping column.

Min The minimum value of Y for the individual series.

Max The maximum value of Y for the individual series.

Length The duration of the series. This is determined based on the starting and ending time stamps and the result of the Analysis of Time Pattern.

N Obs The number of observations. If there are no missing records, this number is the same as the Length.

N Missing The number of missing Y values.

All Positive Indicates whether all of the Y values are positive.

Comment Indicates if there is a problem with the series, such as non-unique time stamps.

Tip: The Group Summary of Y table is linked to the Selected Series list. When you select a series in the Group Summary of Y table, the corresponding entries are also selected in the Select Series list.

Modeling Specifications

The Modeling Specifications section in the Time Series Forecast report enables you to fit models to the individual time series. A variety of models are considered. Hyndman et al. (2008) defines state space smoothing models based on their error, trend component, and seasonal component:

- The errors can be additive (A) or multiplicative (M).
- The trend component can be none (N), additive (A), additive damped (A_d), multiplicative (M), or multiplicative damped (M_d).
- The seasonal component can be none (N), additive (A), or multiplicative (M).

A specific model can be represented by its ETS (Error, Trend, Seasonal). These are the models used in Time Series Forecasting.

There are two tabs in the Modeling Specifications report. Use the Recommended Specifications tab to fit a set of state space smoothing models that is chosen by the platform. Use the Complete Specifications tab to select the specific state space smoothing models that you want to fit. For each individual time series, the best fitting model from the given set is then used for forecasting.

Select Models

Available only in the Complete Specifications tab, shown in [“Complete Specifications Tab”](#). The Select Models report enables you to specify the state space smoothing models to fit to the individual time series. Use the check boxes to select the error, trend, and seasonality for the models. Click **Select Recommended** to select the check boxes that correspond to the models recommended by the platform. Click **Select All** to select all check boxes or **Deselect All** to deselect all check boxes. The Constrain Parameters option is selected by default, which constrains the parameters in such a way that the further an observation is from the present, the less effect it has on the present state of the model. In State Space Smoothing models, a forecast at time t , given all previous observations, is the same as the weighted sum of all observations up to time t . The weights are a function of the parameters. Therefore, constraining the parameters ensures that the weights for past observations go to zero and that the further an observation is from the present, the faster the weight goes to zero.

Figure 19.6 Complete Specifications Tab

Recommended Specifications

Complete Specifications

Select Models

State Space Smoothing Models

Additive Error Models

Trend	Seasonal		
	None	Additive	Multiplicative
None	☒	☒	☐
Additive	☒	☒	☐
Additive with Damping	☒	☒	☐
Multiplicative	☐	☐	☐
Multiplicative with Damping	☐	☐	☐

Select Recommended

Select All

Deselect All

Multiplicative Error Models

Trend	Seasonal		
	None	Additive	Multiplicative
None	☒	☒	☒
Additive	☒	☒	☒
Additive with Damping	☒	☒	☒
Multiplicative	☒	☐	☒
Multiplicative with Damping	☒	☐	☒

Select Recommended

Select All

Deselect All

☒ Constrain Parameters

Forecasting Settings

Name	Value	Description
NAhead	4	NAhead >= 0, Optional. Forecast n-ahead future observations.
Period	4	Period > 0, Optional. Integer or comma delimited integers.

Model Selection Strategy

Forecasting Performance

Metric

RMSE

NHoldback

4

Other Options

Preserve Model Selection Criterion

☐

Forecast Interval Level

0.95

Imputation for Applicable Models

None

Run

Modeling Options

Enables you to select or deselect the Constrain Parameters option.

Constrain Parameters If selected, constrains the parameters in such a way that the further an observation is from the present, the less effect it has on the present state of the model. Constrain Parameters is selected by default.

Forecast Settings

Enables you to specify the following optional settings:

- NAhead** Specifies the number of steps ahead to forecast. The number of steps must be nonnegative.
- Period** Specifies seasonality values to be considered in the model fitting process. By default, the period is set to the suggested seasonality from the Analysis of Time Pattern report. If you want to consider models with different seasonality, specify the additional periods here and separate them with a comma. The period must be greater than zero.

Model Selection Strategy

Enables you to specify how the best model is chosen.

Information Criteria Determines the best model based on the specified information criteria. The available information criteria are AIC and BIC.

Tip: Use BIC if the time series is long.

Forecasting Performance Determines the best model based on a performance selection algorithm that involves a holdback set. First, the time series is partitioned into a training set and a holdback set. The value of NHoldback specifies the number of observations in the holdback set. Then, the algorithm fits all recommended or user-specified models on the data in the training set. Forecasts are made on the holdback set using the individual fitted models. These forecasts are compared to the actual holdback observations and the models are evaluated using the specified metric. The model with the best metric is selected for the final model. Last, the selected final model is fit to the entire time series (training set and holdback set) and the refitted model is used to forecast beyond the last observation of the time series.

Metric Specifies the metric to evaluate the forecasts made by the individual model fits. The available metrics are Root Mean Square Error (RMSE), Mean Square Error (MSE), and Mean Absolute Error (MAE).

NHoldback Specifies the number of observations used in the holdback set.

Other Options

Enables you to specify the following options:

Preserve Model Selection Criterion Saves the model selection criterion for the set of models that is fit to the data. After model fitting, this information is shown in the Information Criterion of All Models for All Series report.

Forecast Interval Level The prediction interval level for the forecasts. This changes the width of the shaded area in the forecasting plots.

Imputation for Applicable Models Specifies the imputation method used during the model fitting process.

Note: These options change only the data that are used for model fitting. The raw data are not changed.

None Does not impute missing values.

Last Value Imputes missing values by using the last value that is available before a sequence of missing observations.

Middle Value Imputes missing values by averaging the last value that is available before a sequence of missing observations and the first value that is available after a sequence of missing observations.

Linear Interpolation Imputes missing values by creating a linear interpolation between the last value that is available before a sequence of missing observations and the first value that is available after a sequence of missing observations.

When you click **Run**, the specified set of models is fit. A progress bar appears that reports the number of active threads, the number of finished and total tasks, and the percentage of finished tasks. Note that a task is defined as the fitting of all specified models for one series and response variable combination. When the fitting process is complete, the Model Reports report is shown. See [“Time Series Forecast Model Reports”](#). If you make any changes to the options in the Model Specifications report and click **Run** again, the Model Reports report is replaced with a new report.

Time Series Forecast Platform Options

The Time Series Forecast red triangle menu contains the following options:

Report Options (Effective only after fitting a model.) Contains the following report options:

Set Forecast Interval Level Enables you to change the forecast interval level. The individual model reports update automatically.

Set NAhead Enables you to change the number of steps ahead to forecast. The individual model reports update automatically.

Set NHoldout Enables you to change the duration of the holdout set. The individual model reports update automatically.

Save Results (Effective only after fitting a model.) Saves the forecasts to a separate data table named Forecasts. The Forecasts data table has the same layout and format as the original data table. If a holdout set is specified by NHoldout, an additional data table that saves the actual values from the holdout data and the forecasts is produced. You can use this table to calculate different types of error estimates.

The following additional save options are shown in the Request Outputs window:

Save Forecast Intervals Saves the lower and upper prediction intervals as columns in the Forecasts data table.

Save One-Step-Ahead Predictions Stacks the one-step-ahead predictions and the forecasts and saves them to a separate data table.

Save Original Series Stacks the original series and the forecasts and saves them to a separate data table.

Save Forecast Results to Original Data Table Saves the results to new columns in the original data table.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Time Series Forecast Model Reports

The Model Reports section contains a Select Series list and one or more individual series modeling reports. The individual reports that are displayed are determined by the Select Series list. Click on the name of a series in this list to show the report for the selected series. If you select more than one series name by using either Ctrl or Shift, an individual model fit report for each of the selected series is shown. If you deselect one or more series from the Select Series list, the reports for the corresponding series are removed.

The Select Series list is linked to the data table. When you select a series in the Select Series list, the observations for the series and response variable combination are also selected in the data table. The Select Series list is also linked to the Group Summary of Y table. If a Time variable is specified, the Select Series list is also linked to the Group Summary of Time table. When you select a series in the Select Series list, the corresponding entries are also selected in the tables.

Each individual model report contains the following information:

Model Type Describes the best fitting model for the series in terms of error, trend, and seasonality.

Model Report A group of reports based on the model that is fit to the series.

Series and Forecasts (Available only when Information Criteria is specified as the model selection strategy.) A plot of the original series and forecasts. During the time period of the original series, the one-step-ahead predictions are shown. Beyond the time period of the original series, the forecasts are shown. The shaded region represents the forecasting interval.

Note: All of the forecasts start at the time point that immediately follows the last time point across all individual series. See [“Data Format”](#).

Training, Holdback, and Forecasts (Available only when Forecast Performance is specified as the model selection strategy.) A plot of the training series, holdback series, and the forecasted series. A shaded vertical region represents the holdback series. Since this is the model refit on the entire time series, the one-step-ahead predictions are shown during the time period of the entire original series. During the time period past the last observation, the forecasts are shown. The shaded region represents the forecasting interval.

Model Summary Provides the -2Loglikelihood, the AIC, the number of parameters, and the sigma value. The sigma value is the estimate of the one-step-ahead forecast standard deviation. If Forecast Performance is specified as the model selection strategy, the root mean square error (RMSE) for the holdback set is also shown. See *Fitting Linear Models*.

Parameter Estimates Provides the parameter estimates for the model. The parameters correspond to the parameters in the best fitting model, which is described in the Model Type report.

One-Step-Ahead Forecasting Errors A plot of the one-step-ahead forecasting errors over time. For each time point t , the error is calculated as $e(t) = y(t) - \mu(t)$.

One-Step-Ahead Relative Forecasting Errors for Multiplicative Error Model (Available only for multiplicative error models.) A plot of the one-step-ahead relative forecasting errors over time. For each time point t , the error is calculated as $e(t) = [y(t) - \mu(t)]/\mu(t)$.

Chapter 20

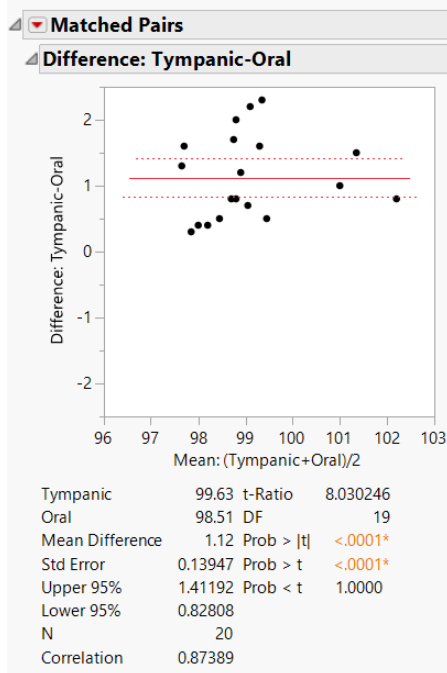
Matched Pairs Analysis

Compare Measurements on the Same Subject

The Matched Pairs platform compares the means between two or more correlated variables and assesses the differences. For example, you might compare a blood pressure measurement taken on the same subject before a treatment and again after the treatment. A statistical method called the *paired t test* takes the correlated responses into account.

The platform produces a graph of the paired differences by the paired means, and the paired *t* test results for all three alternative hypotheses. Additional features provide for more than two matched responses and for a grouping column to test across samples, in a simple version of repeated measures analysis.

Figure 20.1 Example of Matched Pairs Analysis



Contents

Overview of the Matched Pairs Platform	421
Example of Comparing Matched Pairs	421
Launch the Matched Pairs Platform	423
Multiple Y Columns	423
The Matched Pairs Report	424
Difference Plot and Report	425
Across Groups	425
Matched Pairs Platform Options	426
Example Comparing Matched Pairs across Groups	427
Statistical Details for the Matched Pairs Platform	428
Statistical Details for Graphics of Matched Pairs	429
Statistical Details for the Correlation of Responses	430

Overview of the Matched Pairs Platform

The Matched Pairs platform compares row-by-row differences between two response columns using a paired t test. Often, the two columns represent measurements on the same subject before and after some treatment. Alternatively, the measurements could represent data taken on the same subject with two different instruments.

If you have paired data arranged in two data table columns, then you are ready to use the Matched Pairs platform. However, if all of your measurements are in a single column, then perform one of the following tasks:

- Use the **Split** option in the **Tables** menu to split the column of measurements into two columns. Then you can use the Matched Pairs platform.
- For two response columns, create a third column that calculates the difference between the two responses. Then test that the mean of the difference column is zero with the Distribution platform.
- For the two responses stored in a single column, you can do a two-way analysis of variance. One factor (the ID variable) identifies the two responses and the other factor identifies the subject. Use the Fit Y by X Oneway platform with a blocking variable (the subject column), or use the Fit Model platform to do a two-way ANOVA. The test on the ID factor is equivalent to the paired t test.

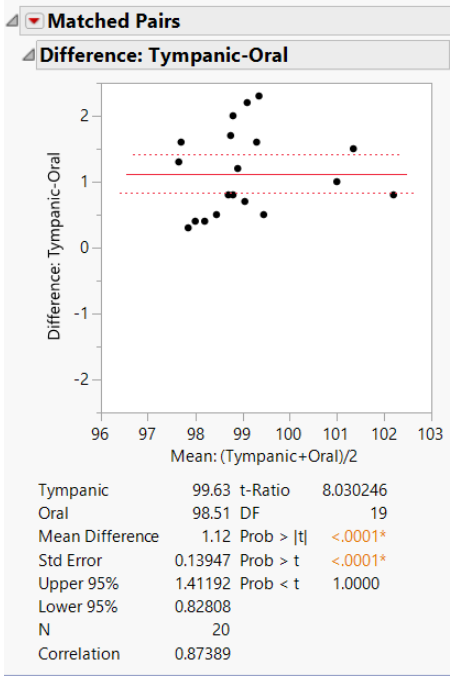
Note: If the data are paired, do not do a regular independent t test. Do not stack the data into one column and use the Fit Y by X One-way ANOVA on the ID without specifying a block variable. To do this has the effect of ignoring the correlation between the responses. This causes the test to overestimate the effect if responses are negatively correlated, or to underestimate the effect if responses are positively correlated.

Example of Comparing Matched Pairs

This example performs a matched pairs analysis to test the difference between two types of thermometers. The data contains temperature measurements on 20 people. Temperature is measured using two types of thermometers: oral and tympanic (ear). You want to determine whether the two types of thermometers produce equal temperature readings. Note that the differences in temperature between the different people are not important.

1. Select **Help > Sample Data Folder** and open Therm.jmp.
2. Select **Analyze > Specialized Modeling > Matched Pairs**.
3. Select Oral and Tympanic and click **Y, Paired Response**.
4. Click **OK**.

Figure 20.2 The Matched Pairs Report Window



The results show that, on average, the tympanic thermometer measures 1.12 degrees higher than the oral thermometer. The small p -value (Prob > |t|) indicates that this difference is statistically significant, and not due to chance.

Note that this matched pairs analysis does not indicate which thermometer is correct (if either), but indicates only that there is a difference between the thermometers.

Launch the Matched Pairs Platform

Launch the Matched Pairs platform by selecting **Analyze > Specialized Modeling > Matched Pairs**.

Figure 20.3 The Matched Pairs Launch Window

Provides a paired t test for matched variables.

Select Columns	Cast Selected Columns into Roles	Action										
4 Columns - Name - Oral - Tympanic - difference	<table border="1"> <tr> <td>Y, Paired Response</td> <td>required numeric required numeric optional numeric</td> </tr> <tr> <td>X, Grouping</td> <td>optional</td> </tr> <tr> <td>Weight</td> <td>optional numeric</td> </tr> <tr> <td>Freq</td> <td>optional numeric</td> </tr> <tr> <td>By</td> <td>optional</td> </tr> </table>	Y, Paired Response	required numeric required numeric optional numeric	X, Grouping	optional	Weight	optional numeric	Freq	optional numeric	By	optional	OK Cancel Remove Recall Help
Y, Paired Response	required numeric required numeric optional numeric											
X, Grouping	optional											
Weight	optional numeric											
Freq	optional numeric											
By	optional											

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Paired Response Provide the two response columns. For information about analyzing more than two responses, see [“Multiple Y Columns”](#).

X, Grouping Provide a grouping variable to compare the differences across groups. See [“Across Groups”](#).

Weight Identifies one column whose numeric values assign a weight to each row in the analysis.

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

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

After you click **OK**, the Matched Pairs report window appears. See [“The Matched Pairs Report”](#).

Multiple Y Columns

In the Matched Pairs platform, you can specify more than two responses. If the number of responses is odd, all possible pairs are analyzed. The following table shows an example for three responses.

Y1 by Y2	Y1 by Y3

Y2 by Y3

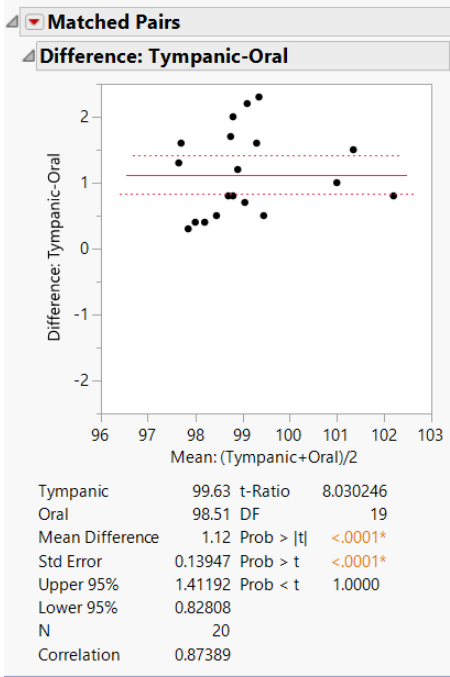
If the number of responses is even, the Matched Pairs platform asks whether you want to do all possible pairs. If you do not do all possible pairs, adjacent responses are analyzed as a pair. The following table shows the arrangement of analyses for four responses.

Y1 by Y2	Y3 by Y4

The Matched Pairs Report

The Matched Pairs report shows a Tukey mean-difference plot, summary statistics, and the results of the paired *t* test. See [“Difference Plot and Report”](#). If you specified an X, Grouping variable, the report also includes the Across Groups report. See [“Across Groups”](#).

Figure 20.4 Example of Matched Pairs Report



Note: The red triangle menu provides additional options that can add reports to the initial report window. See [“Matched Pairs Platform Options”](#).

Difference Plot and Report

The Difference plot in the Matched Pairs report shows differences by means. In the Difference plot the mean difference is shown as the horizontal line. The 95% confidence interval is shown as dotted lines above and below the mean difference. If the confidence region includes zero, then the means are not significantly different at the 0.05 level. If you add a reference frame, the mean of pairs is shown by the vertical line. For more information about a reference frame, see [“Matched Pairs Platform Options”](#).

The Difference report shows the mean of each response, the difference of the means, and a confidence interval for the difference. The Difference report also shows the results of the paired t test.

Across Groups

The Across Groups analysis in the Matched Pairs platform corresponds to a simple repeated measures analysis. You can get the same test results using the Manova personality of the Fit Model platform.

Note: The Across Groups report appears only if you have specified an X, Grouping variable.

Mean Difference Shows the mean of the difference across rows in each group between the two paired columns. In other words, this is the within-subject by across-subject interaction, or split-plot by whole-plot interaction.

Mean Mean Shows the mean of the mean across rows in each group across the two paired columns. In other words, this is the across-subject or whole-plot effect.

Test Across Groups Two F tests determine whether the across-groups values are different:

Mean Difference Tests that the change across the pair of responses is different in different groups.

Mean Mean Tests that the average response for a subject is different in different groups

Related Information

- [“Example Comparing Matched Pairs across Groups”](#)

Matched Pairs Platform Options

The Matched Pairs red triangle menu contains the following options:

Plot Dif by Mean Shows or hides the plot of the paired differences by paired means. For a detailed description of this plot, see [“Difference Plot and Report”](#).

Plot Dif by Row Shows or hides the plot of paired differences by row number.

Reference Frame Shows or hides the reference frame on the Plot Dif by Mean plot. A rectangle showing where a plot of Y2 by Y1 would be located inside the plot, tilted and possibly squished. A vertical red line is shown representing the mean of means. The reference frame is shown initially when the range of the differences is greater than half the range of the data.

Wilcoxon Signed Rank Shows or hides the Wilcoxon signed rank test. The Wilcoxon signed rank test is applied to the paired differences. It is a nonparametric test that compares the sizes of the positive differences to the sizes of the negative differences. The test uses the Pratt method to address zero differences. The test also assumes that the distribution of differences is symmetric. See *Basic Analysis*. See also Lehman (2006), Conover (1999, p. 350), and Cureton (1967).

Sign Test Shows or hides the sign test. This is a nonparametric version of the paired t test that uses only the sign (positive or negative) of the difference for the test.

Bland-Altman Analysis Shows or hides the Bland-Altman report. The report contains the bias, or the mean difference in the X and Y variables. The upper and lower limit of agreement are set at $\text{bias} \pm z_{1-\alpha/2}^*(\text{standard deviation of the bias})$.

Set α Level Changes the alpha level used in the analyses. Affects the confidence intervals in the report and on the plot.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Example Comparing Matched Pairs across Groups

In this example, you produce a Matched Pairs Across Groups report and the corresponding MANOVA report using Fit Model.

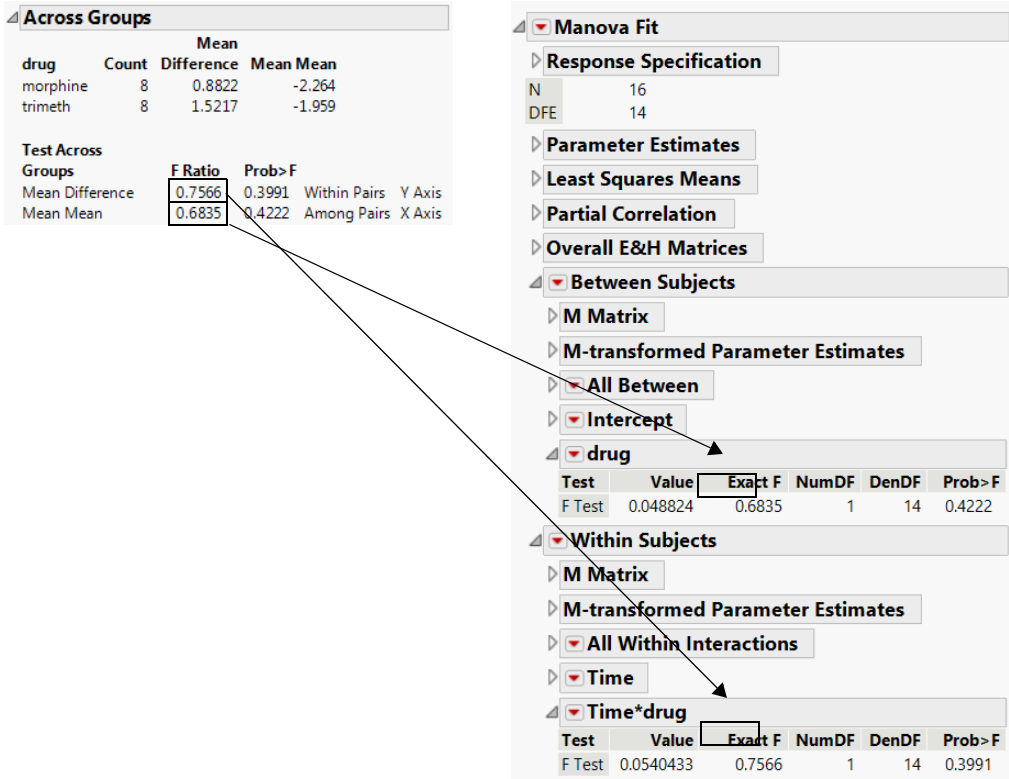
Produce a Matched Pairs Across Groups Report

1. Select **Help > Sample Data Folder** and open Dogs.jmp.
2. Select **Analyze > Specialized Modeling > Matched Pairs**.
3. Select LogHist0 and LogHist1 and click **Y, Paired Response**.
4. Select drug and click **X, Grouping**.
5. Click **OK**.

Produce a Fit Model Report

1. Select **Analyze > Fit Model**.
2. Select LogHist0 and LogHist1 and click **Y**.
3. Select drug and click **Add**.
4. Select the **Manova** personality.
5. Click **Run**.
6. In the Response Specification report, select **Repeated Measures** from the **Choose Response** menu.
7. Click **OK**.

Figure 20.5 Examples of Matched Pairs across Groups and Fit Model MANOVA with Repeated Measures



The F Ratio for the Mean Difference in the Across Groups report corresponds to the F Ratio for Time*drug under the Within Subjects report. The F Ratio for the Mean Mean in the Across Groups report corresponds to the F Ratio for drug under Between Subjects in the Manova Fit report.

Statistical Details for the Matched Pairs Platform

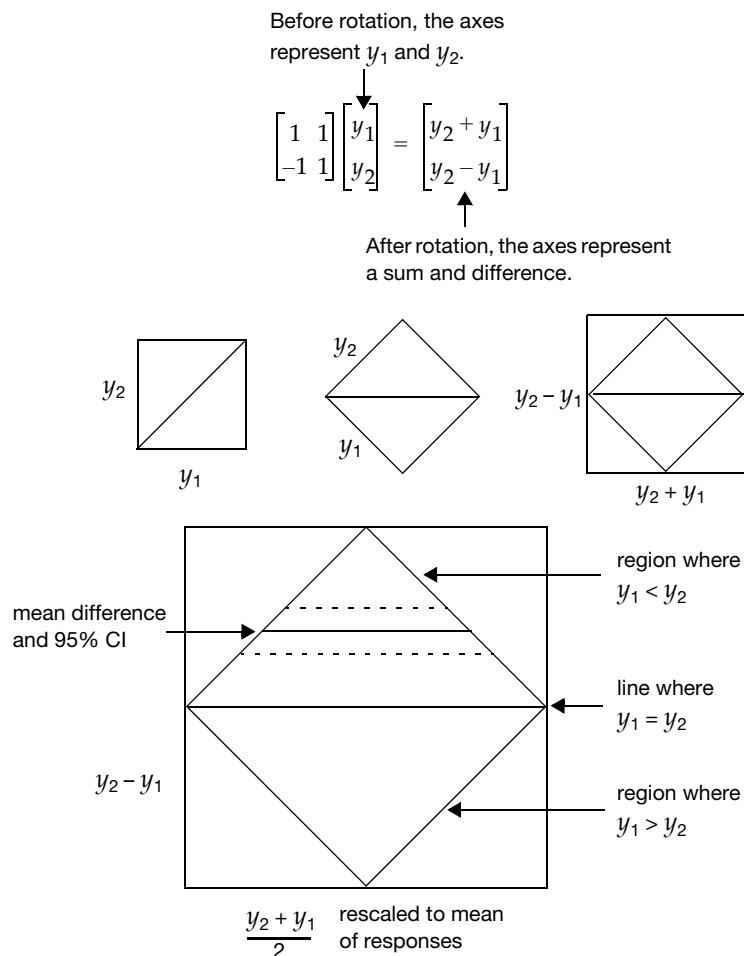
This section contains statistical details for the Matched Pairs platform.

- [“Statistical Details for Graphics of Matched Pairs”](#)
- [“Statistical Details for the Correlation of Responses”](#)

Statistical Details for Graphics of Matched Pairs

The primary graph in the Matched Pairs platform is a Tukey mean-difference (Cleveland 1994, p. 130). This graph plots the difference of the two responses on the vertical axis against the mean of the two responses on the horizontal axis. This graph is the same as a scatterplot of the two original variables, but turned 45 degrees. A 45 degree rotation and rescaling turns the original coordinates into a difference and a mean.

Figure 20.6 Example of Transforming to Difference by Mean, Rotated by 45 Degrees

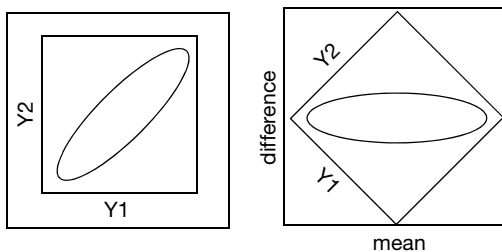


Statistical Details for the Correlation of Responses

It is important to know the correlation between pairs of measurements when performing an analysis in the Matched Pairs platform. In most cases, if the pair of measurements is taken from the same individual at different times, the measurements are positively correlated. However, if the measurements represent competing responses, the correlation can be negative.

Figure 20.7 shows how the positive correlation of the two responses becomes the small variance on the difference (the vertical axis). If the correlation is negative, the ellipse is oriented in the other direction and the variance of the rotated graph is large on the vertical axis.

Figure 20.7 Examples of Positive Correlation Before and After Rotation



Chapter 21

Explore Outliers

Find Outliers in Univariate and Multivariate Data

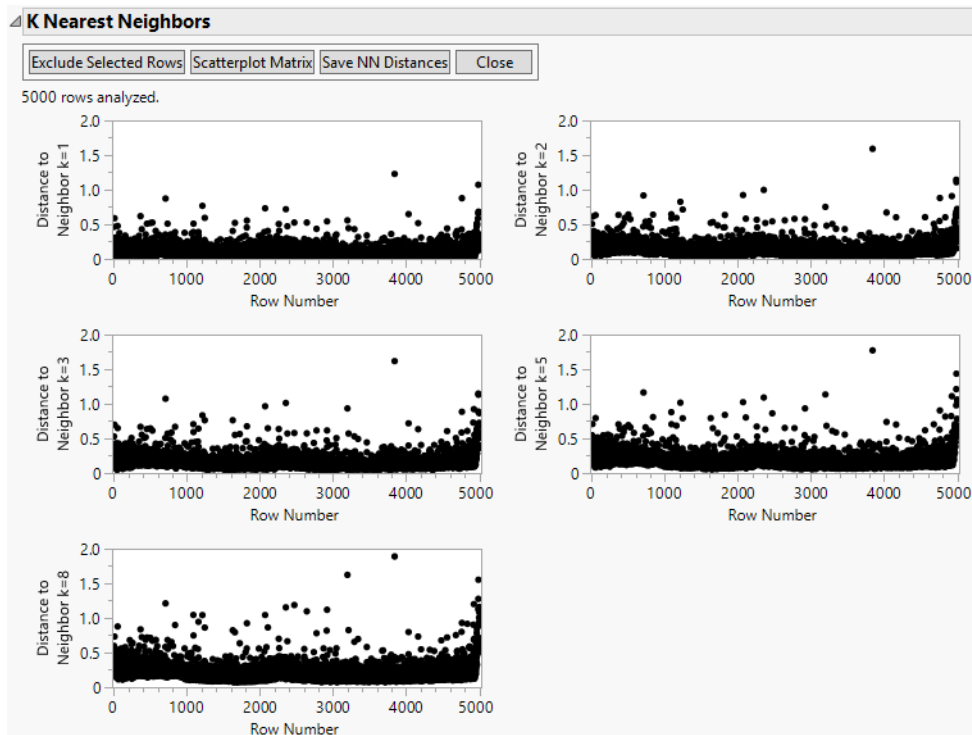
The Explore Outliers platform enables you to identify, explore, and manage outliers.

Exploring and understanding outliers in your data is an important part of analysis. Outliers in data can be due to mistakes in data collection or reporting, measurement systems failure, the inclusion of error or missing value codes in the data set, or simply an unusual value. The presence of outliers can distort estimates and bias results toward those outliers.

Outliers inflate the sample variance. Sometimes retaining outliers in data is necessary, however, and removing them could underestimate the sample variance and bias the data in the opposite direction.

Whether you remove or retain outliers, it is a good practice to locate them. There are many ways to visually inspect for outliers. For example, box plots, histograms, and scatter plots can easily display these extreme values. See *Discovering JMP*.

Figure 21.1 Multivariate k-Nearest Neighbor Outlier Example



Contents

Overview of Explore Outliers	433
Quantile Range Outliers.....	433
Robust Fit Outliers	433
Robust PCA Outliers	433
K Nearest Neighbor Outliers.....	434
Example of Explore Outliers	435
Launch the Explore Outliers Platform	438
The Explore Outliers Report	442
Quantile Range Outliers Report	442
Robust Fit Outliers Report.....	447
Robust PCA Outliers Report	450
K Nearest Neighbors Report	452
Explore Outliers Platform Options.....	453
Additional Example of the Explore Outliers Platform	454

Overview of Explore Outliers

The Explore Outliers platform has four methods for identifying potential outliers in your data. There are two univariate methods and two multivariate methods.

- “Quantile Range Outliers”
- “Robust Fit Outliers”
- “Robust PCA Outliers”
- “K Nearest Neighbor Outliers”

Quantile Range Outliers

The Quantile Range Outliers method in the Explorer Outliers platform uses the quantile distribution of the values in a column to locate extreme values. Quantiles are useful for detecting outliers because there is no distributional assumption associated with them. Data are simply sorted from smallest to largest. For example, the 20th quantile is the value at which 20% of values are smaller. Extreme values are found using a multiplier of the interquantile range, the distance between two specified quantiles. For more information about how quantiles are computed, see *Basic Analysis*.

The Quantile Range Outliers method is also useful for identifying missing value codes stored within the data. In some industries, missing values are entered as nines (such as 999 or 9999). This technique finds any nines greater than the upper quantile as suspected missing value codes. You can then add those missing value codes as a column property in the data table or change these values to missing values in the data table.

Robust Fit Outliers

The Robust Fit Outliers method in the Explore Outliers platform provides several types of robust estimates of the center and spread of your data to determine thresholds for identifying outliers. Robust estimates of parameters are less sensitive to outliers than non-robust estimates.

Robust PCA Outliers

Use the Robust PCA Outliers method in the Explore Outliers platform to identify outlier cells in correlated multivariate data. This method is useful because many other multivariate approaches identify only the outlier rows. Before the method is applied to the data, you have the option to first center and scale the columns. The scaling factor is defined as follows:

$$\max [Q(0.75) - Q(0.50), Q(0.50) - Q(0.25)] / [\text{normalQuantile}(0.75)]$$

where

$Q(p)$ is the p^{th} quantile

Note: If $Q(0.75)$ or $Q(0.25)$ are equal to the median, then more extreme quantiles are used until there is a non-zero range.

After the data are centered and scaled, the Robust PCA Outliers method performs a sequence of singular value decompositions and thresholding steps to decompose the data matrix. The data are decomposed into a low-rank matrix and a sparse matrix of residuals. The thresholding is done so that the residuals are either very large for outliers or very close to zero for non-outliers. The algorithm determines a matrix rank appropriate to capture the systematic variation without the outliers or small noise. Outliers that are not in the low-rank space are detected based on their residuals. See Candes et al (2009) and Lin et al (2013). If there are missing values, they are initially replaced with zeros after the centering and scaling steps. Then, after each singular value decomposition (SVD) iteration, the missing values are updated by their predicted values from the SVD.

K Nearest Neighbor Outliers

Use the K Nearest Neighbor Outliers method in the Explore Outliers platform to identify an outlier based on distance to its nearest neighbor. For each value of k , the K Nearest Neighbor Outliers method displays a plot of the Euclidean distance from each point to its k th nearest neighbor. You specify the largest value of k , denoted as K . Plots are provided for $k = 1, 2, 3, 5, \dots, K$, using the Fibonacci sequence to avoid displaying too many plots.

Before the nearest neighbors are calculated, the columns are centered and scaled. The scaling factor is as follows:

$$\max [Q(0.75) - Q(0.50), Q(0.50) - Q(0.25)] / [\text{normalQuantile}(0.75)]$$

where

$Q(p)$ is the p^{th} quantile

Note: If $Q(0.75)$ or $Q(0.25)$ are equal to the median, then more extreme quantiles are used until there is a non-zero range.

This approach is sensitive to the specified value of k . A small value of k can miss identifying points as outliers and a large value of k can falsely classify points as outliers:

- Suppose that the specified K is small, so that you are studying only a few neighbors. If there is a cluster of more than K points that is far from the rest of the points, then the points within the cluster have small distances to their nearest neighbors. You might be unable to detect the cluster of outliers.

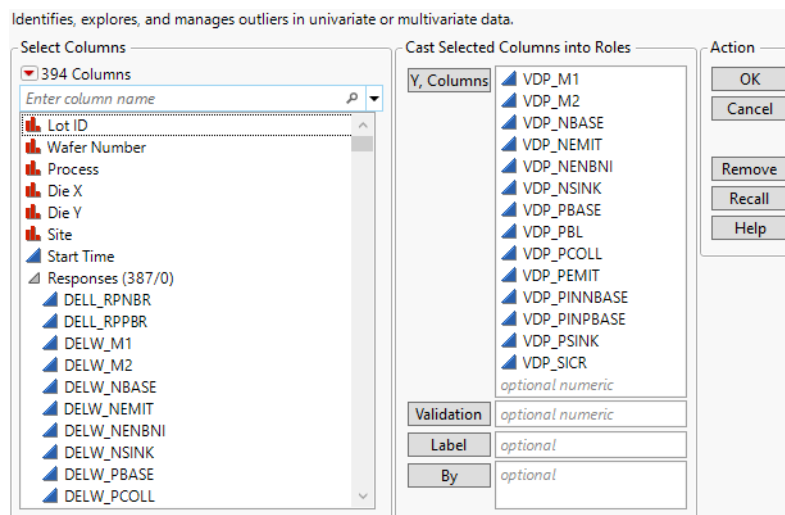
- Suppose that the specified K is large, so that you are studying a large number of neighbors. If there are clusters with fewer than K data points, then the points within these clusters can appear to be outliers. You might overlook the fact that the points form a cluster, interpreting the individual cluster members as outliers instead.

Example of Explore Outliers

Use Explore Outliers to identify outliers within a select group of columns of a data set. The identified outliers can then be examined using the Distribution platform.

1. Select **Help > Sample Data Folder** and open the Probe.jmp sample data table.
2. Select **Analyze > Screening > Explore Outliers**.
3. Click the triangle next to Responses(387/0) to show all of the columns in the group.
4. Select columns VDP_M1 through VDP_SICR and click **Y, Columns**. There should be 14 columns selected.

Figure 21.2 Explore Outliers Launch Window



5. Click **OK**.
6. Click **Quantile Range Outliers**.

The Quantile Range Outliers report uses tabs to organize the results. The Outliers by Column tab of the Quantile Range Outliers report shows each column and lists the number and identity of the outliers found.

7. In the Outliers by Column tab of the Quantile Range Outliers report, select **Show only columns with outliers**. This limits the list of columns to only those that contain outliers.

Note that several columns contain outlier values of 9999. Many industries use nines as a missing value code.

8. Click the Nines tab.
9. In the Nines report, select each column.
10. Click **Add Highest Nines to Missing Value Codes**.

A JMP Alert indicates that you should use the **Save As** command to preserve your original data.

11. Click **OK**.
12. In the Quantile Range Outliers report, click **Rescan**.
13. Select **Restrict search to integers**.

In continuous data, integer values are often error codes or other coded data values. Notice that no additional error codes are included in this set of columns.

14. Deselect **Restrict search to integers**.

Examine the Data

Return to the Outliers by Column tab of the report.

1. Select all of the remaining columns.
2. Click **Select Rows**.
3. Select **Analyze > Distribution**.
4. Assign the selected columns to the **Y, Columns** role. Because you selected these column names in the Quantile Range Outliers report, they are already selected in the Distribution launch window. Simply click **Y, Columns**.
5. Click **OK**.

Figure 21.3 Distribution of Columns with Outliers Selected



In columns VDP_M1 and VDP_PEMIT, notice that some of the selected outliers are somewhat close to the majority of data. For the rest of the columns, the selected outliers appear distant from the majority of data. Now that you have investigated the data points, you can decide which columns to exclude from your analyses.

Refine Excluded Outliers

Return to the Outliers by Column tab of the report.

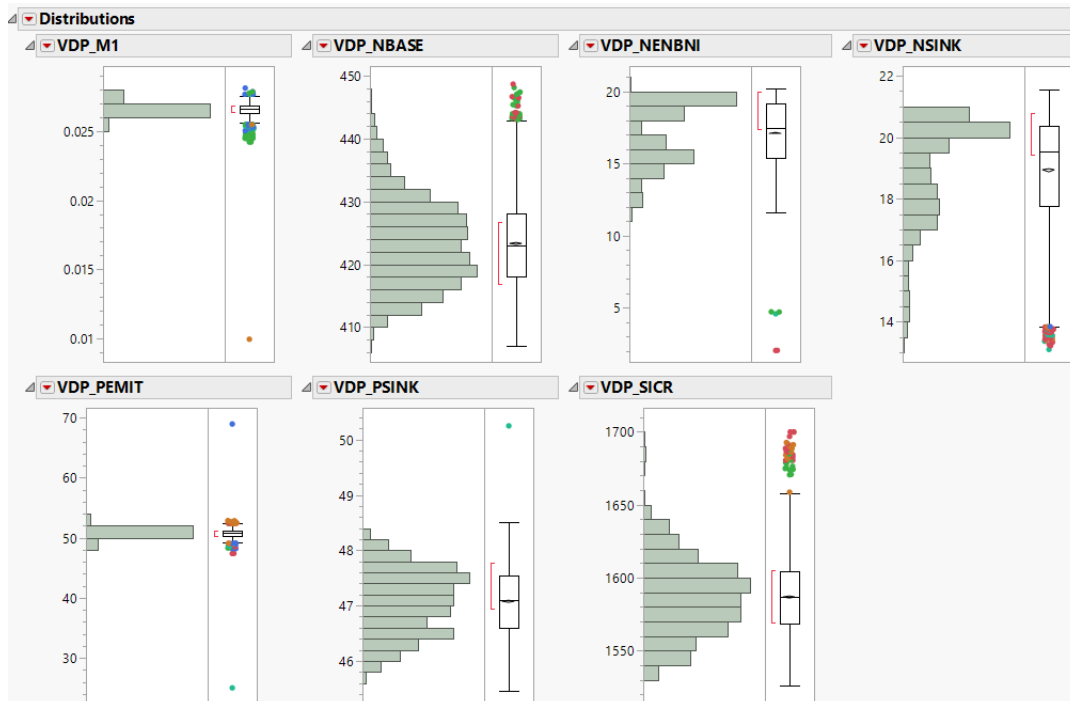
1. Hold Ctrl and deselect columns VDP_M1 and VDP_PEMIT.
2. With the remaining columns selected in the report, click **Exclude Rows**.
3. Change Q to 20.
4. Click **Rescan**.
5. Select columns VDP_M1 and VDP_PEMIT in the report.
6. Click **Select Rows**.

Reexamine the Data

1. Examine the Distributions report again. Notice the selected outliers are now separate enough from the majority of the data to select and exclude them from your analyses.
2. In the Quantile Range Outliers report, click **Exclude Rows**.

3. In the Distributions report, click the Distributions red triangle and select **Redo > Redo Analysis**.

Figure 21.4 Distributions of Columns with Outliers Excluded



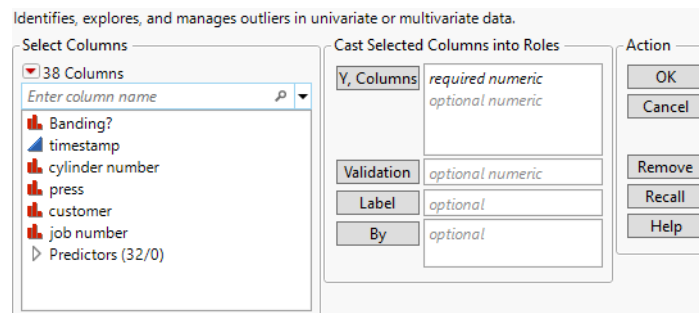
The displays of the distributions of the data are now more informative without the outliers.

Launch the Explore Outliers Platform

To launch Explore Outliers, select **Analyze > Screening > Explore Outliers**.

Note: The Explore Outliers commands analyze only columns with a Continuous modeling type. Other columns can be entered in the launch window but are ignored.

Figure 21.5 Explore Outliers Platform Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns Specifies the columns that you want to analyze.

Validation Specifies a validation column that is used for Robust PCA Outliers.

Label Specifies a column that replaces row numbers in multivariate analysis reports with labels.

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.

Tip: To run an outlier analysis across all levels of a **By** variable, press Ctrl and click the desired outlier analysis command button.

After you click **OK**, the Explore Outliers report appears. The report contains several methods for finding outliers in univariate and multivariate data. There are options for each method that you can specify before making a selection.

Univariate

There are two options for exploring outliers in your univariate data.

Quantile Range Outliers Uses the quantile distribution of each column to identify outliers as extreme values. This tool is useful for discovering missing value or error codes within the data. This is the recommended method to begin exploring outliers in your data. See [“Quantile Range Outliers”](#). You can specify the following options:

Tail Quantile The probability for the lower quantile that is used to calculate the interquantile range. The probability of the upper quantile is considered $1 - \text{Tail Quantile}$. For example, a Tail Quantile value of 0.1 means that the interquantile range is between the 0.1 and 0.9 quantiles of the data. The default value is 0.1.

Q The multiplier that determines the outlier threshold. Values that fall beyond Q times the interquantile range past the Tail Quantile or $1 - \text{Tail Quantile}$ values are identified as outliers. Large values of Q provide a more conservative set of outliers than small values. The default value is 3.

Robust Fit Outliers Finds robust estimates of the center and spread of each column and identifies outliers as those data points that are far from those values. See [“Robust Fit Outliers”](#). You can specify the following options:

K Sigma The multiplier that determines outliers as K times the spread away from the center. Large values of K provide a more conservative set of outliers than small values. The default value is 4.

Huber Uses Huber M-Estimation to estimate center and spread. This option is the default. See Huber and Ronchetti (2009).

Cauchy Assumes a Cauchy distribution to calculate estimates for the center and spread. Cauchy estimates have a high breakdown point and are typically more robust than Huber estimates. However, if your data are separated into clusters, the Cauchy distribution tends to consider only the half of the data that is clustered more closely, ignoring the rest.

Quartile Uses the median as the measure of center and the interquartile range (IQR) divided by 1.34898 as the measure of spread. Dividing the IQR by the 1.34898 factor results in the spread corresponding to one standard deviation if the data are normally distributed.

Multivariate

There are two options for exploring outliers in your multivariate data.

Robust PCA Outliers Decomposes data into a low-rank matrix and residuals and uses the residuals to detect outliers. See [“Robust PCA Outliers”](#). You can specify a value for Lambda and select if the data should be centered. For advanced options, access the Robust PCA Outliers options window by pressing Shift and clicking the Robust PCA Outliers button.

Lambda Specifies a value that determines the sparsity of the matrix of residuals. For larger values of Lambda, the matrix of residuals is more sparse. For a data table with n training rows and p columns, the default value of Lambda is defined as follows:

$$\lambda = \frac{2}{\sqrt{\max(n, p)}}$$

Max Iterations Specifies the maximum number of SVD iterations. The default number of iterations is 100. If there are more than 20,000 columns specified in the launch, the default number of iterations is 50.

Note: If the algorithm does not converge after the maximum number of iterations, a JMP alert is shown. You can continue with more iterations or cancel. If you click Cancel and a less stringent convergence criterion is met, the results are shown. If you click Cancel and the less stringent convergence criterion is not met, another JMP alert is shown to ask whether or not to accept the results.

Converge Criterion Determines when to stop the algorithm. The default convergence criterion values are set based on the number of columns specified in the launch.

- If the number of columns is less than 2,000, the default is 1e-7.
- If the number of columns is greater than or equal to 2,000, the default is 1e-6.
- If the number of columns is greater than or equal to 20,000, the default is 1e-5.
- The less stringent convergence criterion is set to 1000 times the original convergence criterion.

Outlier Threshold Specifies the outlier threshold that determines which outliers are shown in the Cell Large Residuals table. An observation is shown if the scaled residuals is larger than the following:

$$\min[0.99 \times \max\{\text{abs(scaled residuals)}\}, \text{Outlier Threshold}]$$

By default, the Outlier Threshold is 2. If using 2 as the Outlier Threshold results in more than a million outliers, the Outlier Threshold is changed to 3.

Center Determines if the data are centered before the Robust PCA Outlier algorithm is performed.

Note: If the number of rows is less than or equal to 10, the data are not centered.

Scale Determines if the data are scaled before the Robust PCA Outlier algorithm is performed.

Note: If the number of rows is less than or equal to 10, the data are not scaled.

Randomized SVD for Very Wide Problems (Available only if the number of columns specified in the launch is greater than or equal to 1000.) Uses the Randomized SVD method instead of the Lanczos method to decompose the data. This option speeds up the Robust PCA Outlier calculations for wide data. See [“Randomized SVD”](#).

Randomized Dimensions (Available only if the number of columns specified in the launch is greater than or equal to 1000.) Specifies the number of dimensions used in Randomized SVD.

K Nearest Neighbor Outliers Identifies outliers as values that are far from their k-nearest neighbors. See [“K Nearest Neighbor Outliers”](#). You can specify the following options:

K Specifies the upper bound for the farthest neighbor to be considered. The default value is 8.

Impute Missing Specifies whether missing values are imputed.

The Explore Outliers Report

The Explore Outliers report contains a report for each outlier detection method that you select.

- [“Quantile Range Outliers Report”](#)
- [“Robust Fit Outliers Report”](#)
- [“Robust PCA Outliers Report”](#)
- [“K Nearest Neighbors Report”](#)

Quantile Range Outliers Report

The Quantile Range Outliers report in the Explore Outliers platform includes a set of controls and results organized on multiple tabs.

Quantile Range Outliers Controls

The Quantile Range Outliers controls enable you to specify how outliers are to be identified. An outlier is considered any value more than Q times the interquantile range from the lower and upper quantiles. You can adjust the value of Q and the size of the interquantile range.

Figure 21.6 Quantile Range Outliers Initial Options

Tail Quantile The probability for the lower quantile that is used to calculate the interquantile range. The probability of the upper quantile is considered $1 - \text{Tail Quantile}$. For example, a Tail Quantile value of 0.1 means that the interquantile range is between the 0.1 and 0.9 quantiles of the data. The default value is 0.1.

Q The multiplier that determines the outlier threshold. Values that fall beyond Q times the interquantile range past the Tail Quantile or $1 - \text{Tail Quantile}$ values are identified as outliers. Large values of Q provide a more conservative set of outliers than small values. The default is 3.

Restrict search to integers Restricts outlier values to only integer values. This setting limits the search for outliers in order to find industry-specific missing value codes and error codes.

Rescan Rescans the data after outlier actions have been taken.

Tip: Press Ctrl and click **Rescan** to rescan across all open outlier methods.

Close Closes the Quantile Range Outliers report.

Tip: Press Ctrl and click **Close** to close all outlier reports.

Outliers by Column

The Outliers by Column tab in the Quantile Range Outliers report contains a table of columns that have outliers found by the settings specified in the controls. The table shows values for the upper and lower quantiles along with their low and high thresholds. Values outside of these threshold limits are considered outliers. The number of outliers in each column is indicated. The values of each outlier are listed in the last column of the table. Outliers that occur more than once in a column are listed with their count in parentheses.

The Outliers by Column tab contains the following options that can be applied when one or more rows are selected in the outliers table:

Show only columns with outliers Removes columns without outliers from the table in the Outliers by Column tab.

Identify Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Select Rows Selects the rows containing outliers.

Exclude Rows Applies the exclude row state. Click **Rescan** to update the Quantile Range Outliers report.

Color Cells Colors the cells containing outliers. Low valued outliers are colored blue and high valued outliers are colored red.

Color Rows Colors the rows containing outliers.

Clear Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Add to Missing Value Codes Adds outliers to the missing value codes column property. Use this option to identify known missing value or error codes within the data. Missing value and error codes are often integers and are sometimes a series of nines. Click **Rescan** to update the Quantile Range Outliers report.

Note: Add to Missing Value Codes is not available with Quantile Range Outliers if a By variable is specified in the launch window.

Change to Missing Changes the outlier value to a missing value. Use caution when changing values to missing. Change values to missing only if the data are known to be invalid or inaccurate. Click **Rescan** to update the Quantile Range Outliers report.

Note: If the selected outlier has been added to the missing value codes, the outlier is not changed to a missing value.

Formula Columns Creates a new formula column for each column to set outliers to missing. The new columns are prefixed or suffixed by a user specified name to distinguish them from the original columns. By default, the suffix is set to “Culled”.

Formula Script Creates a script that is added to the data table. When the script is run, it creates a new formula column for each column to set outliers to missing. The new columns are prefixed or suffixed by a user specified name to distinguish them from the original columns. By default, the suffix is set to “Culled”.

Outliers by Cell

The Outliers by Cell tab in the Quantile Range Outliers report contains a table of individual outliers found by the settings specified by the controls. The table shows the column name, row number, outlier distance and the actual value of the individual outliers. The outlier distance is a measure of how extreme an outlier is and is calculated using the following equation:

$$\text{Outlier Distance} = \frac{|x - m|}{IQR}$$

where

x = the actual value of the outlier

m = median of the column the outlier is in

IQR = the interquartile range using the specified Tail Quantile

A larger outlier distance indicates a more extreme outlier.

The Outliers by Cell tab contains the following options that can be applied when one or more rows are selected in the outliers table:

Identify Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Select Row and Column Selects the rows and columns that correspond to the selected outliers.

Color Cells Colors the cells of the outliers in the data table. Low valued outliers are colored blue and high valued outliers are colored red.

Clear Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Add to Missing Value Codes Adds the selected outliers to the missing value codes column property. Use this option to identify known missing value or error codes within the data. Missing value and error codes are often integers and are sometimes a series of nines. Click **Rescan** to update the Quantile Range Outliers report.

Note: Add to Missing Value Codes is not available with Quantile Range Outliers if a By variable is specified in the launch window.

Change to Missing Changes the outlier value to a missing value. Use caution when changing values to missing. Change values to missing only if the data are known to be invalid or inaccurate. Click **Rescan** to update the Quantile Range Outliers report.

Note: If the selected outlier has been added to the missing value codes, the outlier is not changed to a missing value.

Outliers by Row

The Outliers by Row tab contains a table of rows that contain outliers found by the settings specified in the controls. The table shows the row number and the number of outliers that were found in that row.

The Outliers by Row tab contains the following options that can be applied when one or more rows are selected in the outliers table:

Identify Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Select Rows Selects the rows with outliers.

Exclude Rows Applies the exclude row state. Click **Rescan** to update the Quantile Range Outliers report.

Color Rows Colors the rows containing outliers.

Nines

The Nines tab within the Quantile Range Outliers report shows a table of columns that contain probable missing value codes. These missing value codes are a series of nines (usually 9999) and are the highest number that is all nines and also higher than the upper quantile. If the count is high, it is likely that these outliers are actually missing value codes. If the count is very low, you should explore further to determine whether the value is an outlier or a missing value code. The table also includes the upper quantile value.

The Nines tab is shown only when probable missing value codes are identified.

The Nines tab contains the following options that can be selected when one or more rows are selected in the table:

Add Highest Nines to Missing Value Codes Adds the selected outlier values to the missing value codes column property. You must click **Rescan** to update the Quantile Range Outliers report.

Note: Add Highest Nines to Missing Value Codes is not available with Quantile Range Outliers if a By variable is specified in the launch window.

Change Highest Nines to Missing Replaces the selected outlier values with missing values in the data table.

Note: The first time you use choose an action (such as Change to Missing or Exclude Rows) to change your data, the alert window warns you to use the Save As command to save your data table as a new file to preserve a copy of your original data. When this window appears, click OK. If you decide to save your new data file, you will automatically be prompted to save the file with a new name.

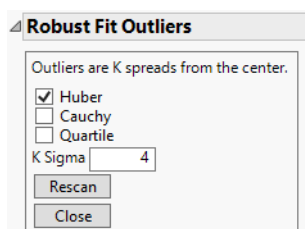
Robust Fit Outliers Report

The Robust Fit Outliers report in the Explore Outliers platform includes a set of controls and results organized on multiple tabs.

Robust Fit Outliers Initial Options

The Robust Fit Outliers controls specify the method used for calculating the robust estimates and the multiplier K . Given a robust estimate of the center and spread, outliers are defined as those values that are K times the robust spread from the robust center.

Figure 21.7 Robust Fit Outliers Controls

A screenshot of the 'Robust Fit Outliers' dialog box. The title bar says 'Robust Fit Outliers'. Inside, there is a text label 'Outliers are K spreads from the center.' Below this are three radio button options: 'Huber' (which is selected), 'Cauchy', and 'Quartile'. Below the radio buttons is a text input field labeled 'K Sigma' with the value '4' entered. At the bottom are two buttons: 'Rescan' and 'Close'.

Huber Uses Huber M-Estimation to estimate center and spread. This option is the default. See Huber and Ronchetti (2009).

Cauchy Assumes a Cauchy distribution to calculate estimates for the center and spread. Cauchy estimates have a high breakdown point and are typically more robust than Huber estimates. However, if your data are separated into clusters, the Cauchy distribution tends to consider only the half of the data that is clustered more closely, ignoring the rest.

Quartile Uses the median as the measure of center and the interquartile range (IQR) divided by 1.34898 as the measure of spread. Dividing the IQR by the 1.34898 factor results in the spread corresponding to one standard deviation if the data are normally distributed.

K Sigma The multiplier that determines outliers as K times the spread away from the center. Large values of K provide a more conservative set of outliers than small values. The default is 4.

Rescan Rescans the data after outlier actions have been taken.

Tip: Press Ctrl and click **Rescan** to rescan across all open outlier methods.

Close Closes the Robust Fit Outliers panel.

Tip: Press Ctrl and click **Close** to close all outlier reports.

Outliers by Column

The Outliers by Column tab in the Robust Fit Outliers report contains a table with a row for each column selected in the launch window. The columns of the table depend on the technique that is used to estimate the center and spread of the data: Huber, Cauchy, or Quartile. For each technique, there is a column of the estimated center, the estimated spread, and the number of outliers based on the center and spread.

The Outliers by Column tab contains the following options that can be applied when on or more rows are selected in the outliers table:

Show only columns with outliers Removes columns without outliers from the table in the Outliers by Column tab.

Identify Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Select Rows Selects the rows containing outliers.

Exclude Rows Applies the exclude row state. Click **Rescan** to update the Robust Fit Outliers report.

Color Cells Colors the cells containing outliers. Low valued outliers are colored blue and high valued outliers are colored red.

Color Rows Colors the rows containing outliers.

Clear Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Add to Missing Value Codes Adds the selected outliers to the missing value codes column property. Use this option to identify known missing value or error codes within the data. Click **Rescan** to update the Robust Fit Outliers report.

Note: Add to Missing Value Codes is not available with Robust Fit Outliers if a By variable is specified in the launch window.

Change to Missing Changes the outlier value to a missing value. Click **Rescan** to update the Robust Fit Outliers report.

Formula Columns Creates a new formula column for each column to set outliers to missing. The new columns are prefixed or suffixed by a user specified name to distinguish them from the original columns. By default, the suffix is set to “Culled”.

Formula Script Creates a script that is added to the data table. When the script is run, it creates a new formula column for each column to set outliers to missing. The new columns are prefixed or suffixed by a user specified name to distinguish them from the original columns. By default, the suffix is set to “Culled”.

Outliers by Cell

The Outliers by Cell tab in the Robust Fit Outliers report contains a table of individual outliers found by the settings specified by the controls. The table shows the column name, row number, outlier distance and the actual value of the individual outliers. The outlier distance is a measure of how extreme an outlier is and is calculated using the following equation:

$$\text{Outlier Distance} = \frac{|x - c|}{s}$$

where

x = the actual value of the outlier

c = the center of column that contains the outlier, measured by the specified outlier method (Huber, Cauchy, or Quartile)

s = the spread of the column that contains the outlier, measured by the specified outlier method (Huber, Cauchy, or Quartile)

A larger outlier distance indicates a more extreme outlier.

The Outliers by Cell tab contains the following options that can be applied when one or more rows are selected in the outliers table:

Identify Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Select Row and Column Selects the rows and columns that correspond to the selected outliers.

Color Cells Colors the cells containing outliers. Low valued outliers are colored blue and high valued outliers are colored red.

Clear Outliers in Table Applies actions to the original data table for selected rows in the outlier summary table.

Add to Missing Value Codes Adds the selected outliers to the missing value codes column property. Use this option to identify known missing value or error codes within the data. Missing value and error codes are often integers and are sometimes a series of nines. Click **Rescan** to update the Robust Fit Outliers report.

Note: Add to Missing Value Codes is not available with Robust Fit Outliers if a By variable is specified in the launch window.

Change to Missing Changes the outlier value to a missing value in the data table. Use caution when changing values to missing. Change values to missing only if the data are known to be invalid or inaccurate. Click **Rescan** to update the Robust Fit Outliers report.

Note: If the selected outlier has been added to the missing value codes, the outlier is not changed to a missing value.

Robust PCA Outliers Report

The Robust PCA Outliers report in the Explore Outliers platform includes a table of information about the algorithm and several tables of results. The initial table includes the rank of the low-rank matrix, the number of SVD iterations, the convergence criterion, the value of Lambda, and the number of imputed missing values. If the Randomized SVD option is enabled, the number of dimensions used in the Randomized SVD option is also included in the summary table. These are the other tables that are included in the report.

Cell Large Residuals A table that shows the largest outlier observations, identified by column name and row number. The number of observations shown is determined by the Outlier Threshold. The table contains the column name and row number of the observation, the residual value, and the scaled residual value.

Tip: To color specific outlier cells in the data table, select rows in the Cell Large Residuals table and click **Colorize**.

Row Root Mean Square A table that shows the root mean square value for each row in the data table. The root mean square is calculated using the scaled residuals.

Tip: If you select a row in the Row Root Mean Square table, the corresponding row is selected in the data table.

Column Root Mean Square A table that shows the root mean square value for each column specified in the launch window. The root mean square is calculated using the scaled residuals.

Tip: If you select a row in the Column Root Mean Square table and click **Select Columns**, the corresponding column is selected in the data table.

Snapshot A graphical representation of the outlier cells in the data table. The outlier cells are colored in red.

Residuals The matrix of residuals from the matrix decomposition. A cell is colored if the absolute value of the scaled residual is greater than the following:

$$\min[0.99 \times \max\{\text{abs}(\text{residuals})\}, \text{Outlier Threshold}]$$

Low Rank Approximation The matrix of scaled residuals from the matrix decomposition.

Singular Values The vector of singular values from the SVD.

Robust PCA Outliers Options

There are buttons at the bottom of the Robust PCA Outliers report that provide options to save different parts of the report.

Close Closes the Robust PCA Outliers report.

Save Large Outliers Saves the information in the Cell Large Residuals table to a new data table.

Save Cleaned Opens a window that provides several techniques to clean the outliers based on thresholds and save new columns to the data table.

Trim Trims outlier cells if the corresponding absolute scaled residual is greater than the specified threshold. By default, the threshold is 10. Select Color  to color the outlier cells red. The trimmed cells are set to the value of the unscaled threshold.

Impute Sets outlier cells to the value of the low-rank approximation if the corresponding absolute scaled residual is greater than the specified threshold. By default, the threshold is 100. Select Color  to color these cells green.

Make Missing Sets outlier cells to missing if the corresponding absolute scaled residual is greater than the specified threshold. By default, the threshold is 1000. Select Color  to color these cells blue.

Color imputed from missing  If selected, colors cells that originally had missing values and were imputed.

Save Residuals Saves the residuals to new columns in the original data table.

Save Scaled Residuals Saves the scaled residuals to new columns in the original data table.

Save Low Rank Approx Saves the low-rank approximation to new columns in the original data table.

K Nearest Neighbors Report

The K Nearest Neighbor Outliers report in the Explore Outliers platform contains plots for select values of k up to the value K . The value of k for each plot is displayed in its vertical axis label. It is of the form Distance to Neighbor $k = \langle a \rangle$, where a is an integer denoting the a^{th} closest neighbor. Each plot shows the distance from the point in the i^{th} row to its a^{th} nearest neighbor. The points that have large distances from their neighbors, across multiple values of k , are likely to be outliers.

The buttons above the plots do the following:

Exclude Selected Rows Excludes rows corresponding to selected points from further analysis. The rows are assigned the Excluded row state in the data table. You are asked if you want to rerun or close the K Nearest Neighbors report. Rerunning the analysis identifies new nearest neighbors. The plots are updated and the excluded points are not shown.

Scatterplot Matrix Opens a separate window containing a scatterplot matrix for all columns in the analysis. You can explore potential outliers by selecting them in the K Nearest Neighbors plots and viewing them in the scatterplot matrix.

Save NN Distances Saves the distances from each row to its n^{th} nearest neighbor as new columns in the data table.

Close Closes the K Nearest Neighbors report.

Largest Outliers

The K Nearest Neighbors report also includes a Largest Outliers table. This table contains the 20 observations with the largest distances from their K^{th} nearest neighbor. The table has the following columns:

Row The row number of the observation.

Distance The distance from the observation in the specified row and its K^{th} nearest neighbor. The table is sorted by this column in descending order.

Nearest Neighbors Lists the row numbers for the k nearest neighbors. The first row number is the closest nearest neighbor. The last row number is the K^{th} nearest neighbor and the distance between this observation and the specified row is found in the Distance column.

Col<n> Specifies the column name for the corresponding RMS value.

RMS<n> The root mean squared differences across the k nearest neighbors for each column. The largest RMS values are displayed in order, where RMS1 is the maximum RMS value. The p^{th} RMS value is calculated as follows:

$$\text{RMS}_p = \sqrt{\frac{\sum_{k=1}^K (D_{p,i} - D_{p,i_k})^2}{K}}$$

where

D_p is the p^{th} column

$D_{p,i}$ is the value of the p^{th} column for row i

D_{p,i_k} is the value of the p^{th} column for the k^{th} nearest neighbor of row i

Note: The number of Col and RMS columns shown in the Largest Outliers table is the minimum of the number of columns specified in the launch and the number five.

Explore Outliers Platform Options

The Explorer Outliers red triangle menu contains the following options:

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

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

Additional Example of the Explore Outliers Platform

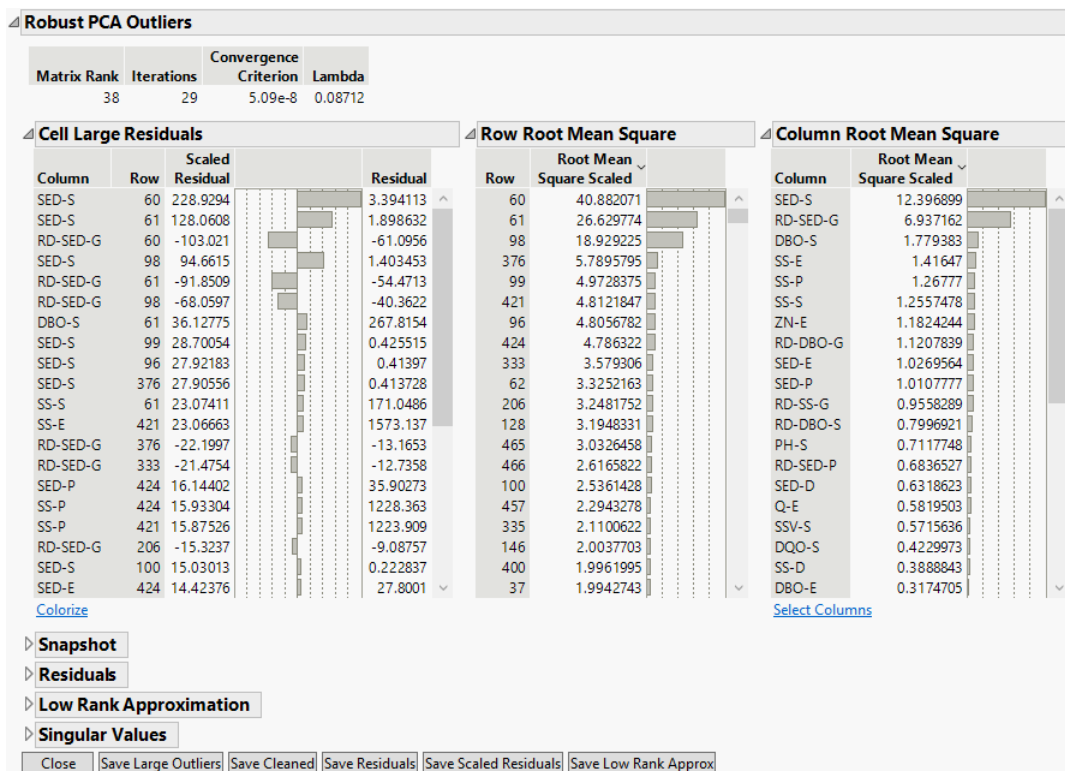
Use the Robust PCA Outliers method to explore data from an urban waste water plant for potential outliers. Potential outliers could include sensor failures, storms, and other situations.

1. Select **Help > Sample Data Folder** and open Water Treatment.jmp.
2. Select **Analyze > Screening > Explore Outliers**.
3. Select the Sensor Measurements column group and click **Y, Columns**.
4. Click **OK**.
5. Shift+Click **Robust PCA Outliers**.
6. Enter 10 next to Outlier Threshold.

Use the default values for the other options.

7. Click **OK**.

Figure 21.8 Robust PCA Outliers Report



The Cell Large Residuals table shows the cells that have a scaled residual greater than 10 or less than -10. There are several cells with large scaled residuals from the SED-S column.

If you look at the Column Root Mean Square table, it shows that the SED-S column does in fact have the largest scaled Root Mean Square value. The row with the largest scaled Root Mean Square value is row 60.

8. Click **Save Cleaned** at the bottom of the report.

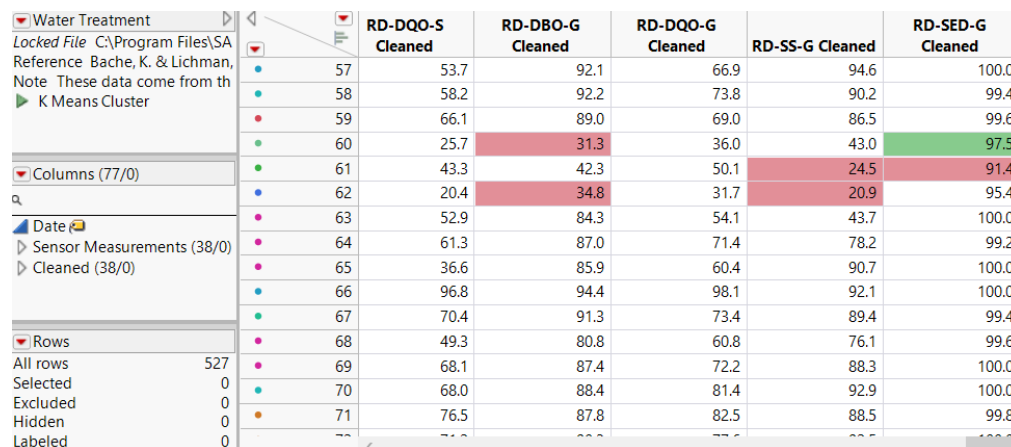
This option enables you to adjust cells with large outliers by trimming the values, imputing new values, or setting the cells to missing. The Save Cleaned Columns report provides these options.

9. Select Impute. Trim is already selected by default.

10. Click **OK**.

There are 38 new Cleaned columns added to the Water Treatment.jmp data table. Since you know row 60 had the largest scaled Root Mean Square value, scroll to this row to view some of the cells that were trimmed and imputed. You can now use the cleaned columns in any type of analysis.

Figure 21.9 Cleaned Columns from Robust PCA Outliers



		RD-DQO-S Cleaned	RD-DBO-G Cleaned	RD-DQO-G Cleaned	RD-SS-G Cleaned	RD-SED-G Cleaned
57		53.7	92.1	66.9	94.6	100.0
58		58.2	92.2	73.8	90.2	99.4
59		66.1	89.0	69.0	86.5	99.6
60		25.7	31.3	36.0	43.0	97.5
61		43.3	42.3	50.1	24.5	91.4
62		20.4	34.8	31.7	20.9	95.4
63		52.9	84.3	54.1	43.7	100.0
64		61.3	87.0	71.4	78.2	99.2
65		36.6	85.9	60.4	90.7	100.0
66		96.8	94.4	98.1	92.1	100.0
67		70.4	91.3	73.4	89.4	99.4
68		49.3	80.8	60.8	76.1	99.6
69		68.1	87.4	72.2	88.3	100.0
70		68.0	88.4	81.4	92.9	100.0
71		76.5	87.8	82.5	88.5	99.8

Chapter 22

Explore Missing Values

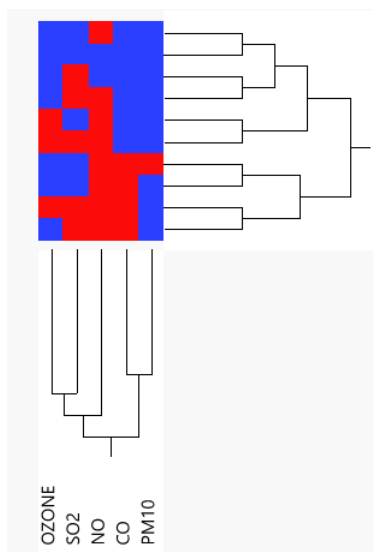
Explore and Impute Missing Values

The Explore Missing Values platform provides several ways to identify and understand the missing values in your data. There are also methods for conducting multivariate imputation for missing values. These imputation methods assume that data are *missing at random*, which means that any differences between missing and non-missing data cannot be explained by the values of the other variables in the study. If you suspect that missing values are *not* missing at random, then consider using the Informative Missing procedure, which is available in a number of platforms. See *Fitting Linear Models*.

The presence of missing values in a data set can affect the conclusions made using the data. If, for example, several healthy participants dropped out of a longitudinal study and their data continued on as missing, the results of the study can be biased toward those unhealthy individuals who remained. Missing data values must not only be identified, they must also be understood before further analysis can be conducted.

Caution: Be careful when analyzing data after imputing missing values, as the results have the potential to be biased. It is best to make a copy of the original data table or save the data with imputed values to a new file.

Figure 22.1 Example of Explore Missing Values



Contents

Overview of the Explore Missing Values Platform	459
Multivariate Normal Imputation	459
Multivariate SVD Imputation	459
Multivariate RPCA Imputation.....	460
Automated Data Imputation	460
Example of the Explore Missing Values Platform	461
Launch the Explore Missing Values Platform.....	463
The Explore Missing Values Report.....	464
Commands	465
Imputation Report.....	467
Explore Missing Values Platform Options.....	468

Overview of the Explore Missing Values Platform

The Explore Missing Values platform provides several methods to identify and understand the missing values in your data.

- “Multivariate Normal Imputation”
- “Multivariate SVD Imputation”
- “Multivariate RPCA Imputation”
- “Automated Data Imputation”

Multivariate Normal Imputation

The Multivariate Normal Imputation method in the Explore Missing Values platform imputes missing values based on the multivariate normal distribution. The procedure requires that all variables have a Continuous modeling type. The algorithm uses least squares imputation. The covariance matrix is constructed using pairwise covariances. The diagonal entries (variances) are computed using all nonmissing values for each variable. The off-diagonal entries for any two variables are computed using all observations that are nonmissing for both variables. In cases where the covariance matrix is singular, the algorithm uses minimum norm least squares imputation based on the Moore-Penrose pseudo-inverse.

Multivariate Normal Imputation allows the option to use a shrinkage estimator for the covariances. The use of shrinkage estimators is a way of improving the estimation of the covariance matrix. For more information about shrinkage estimators, see Schäfer and Strimmer (2005).

Note: If a validation column is specified, the covariance matrices are computed using observations from the Training set.

Multivariate SVD Imputation

The Multivariate SVD Imputation method in the Explore Missing Values platform imputes missing values using the singular value decomposition (SVD). This method is useful for data with hundreds or thousands of variables. Because SVD calculations do not require calculation of a covariance matrix, the SVD method is recommended for wide problems that contain large numbers of variables. The procedure requires that all variables have a Continuous modeling type.

The singular value decomposition represents a matrix of observations $\mathbf{X}$ as $\mathbf{X} = \mathbf{UDV}'$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal matrices and $\mathbf{D}$ is a diagonal matrix.

The SVD algorithm used in the Multivariate SVD Imputation method is the sparse Lanczos method, also known as the *implicitly restarted Lanczos bidiagonalization method* (IRLBA). See Baglama and Reichel (2005). The Multivariate SVD Imputation algorithm does the following:

1. Each missing value is replaced with its column's mean.
2. An SVD decomposition is performed on the matrix of observations, $\mathbf{X}$.
3. Each cell that had a missing value is replaced by the corresponding element of the $\mathbf{UDV}'$ matrix obtained from the SVD decomposition.
4. Steps 2 and 3 are repeated until the SVD converges to the matrix $\mathbf{X}$ or the maximum number of iterations is reached.

Multivariate RPCA Imputation

The Multivariate Robust PCA method in the Explore Missing Values platform imputes missing values using robust principal components, which replaces missing values using a low-rank matrix factorization (SVD) that is robust to outliers. This is the same method that is used in the Robust PCA Outliers method in the Explore Outliers platform. See [“Robust PCA Outliers”](#). This method is useful for wide problems, but can sometimes be computationally expensive for very large dimensions.

Automated Data Imputation

The Automated Data Imputation (ADI) method in the Explore Missing Values platform imputes missing values using a low-rank matrix approximation method, also known as matrix completion. Once trained, the ADI model is capable of performing missing data imputations for streaming data through scoring formulas. Streaming data are added rows of observations that become available over time and were not used for tuning or validating the imputed model. This method is flexible, robust, and automated to select the best dimension for the low-rank approximation. These features enable ADI to work well for many different types of data sets.

A low-rank approximation of a matrix is of the form $\mathbf{X} = \mathbf{UDV}'$ and can be viewed as an extension of singular value decomposition (SVD). ADI uses the Soft-Impute method as the imputation model and is designed such that the data determines the rank of the low-rank approximation.

The ADI algorithm performs the following steps:

1. The data are partitioned into training and validation sets.
2. Each set is centered and scaled using the observed values from the training set.
3. For each partitioned data set, additional missing values are added within each column and are referred to as induced missing (IM) values.

4. The imputation model is fit on the training data set along a solution path of tuning parameters. The IM values are used to determine the best value for the tuning parameter.
5. Additional rank reduction is performed using the training data set by de-biasing the results from the chosen imputation model in [step 4](#).
6. Final rank reduction is performed to calibrate the model for streaming data and to prevent overfitting. This is done by fitting the imputation model on the validation set, using the rank determined in [step 5](#) as an upper bound.

If the algorithm is unable to find a low-rank approximation, the mean is imputed for continuous variables and the mode is imputed for categorical variables. These values are calculated from the training set.

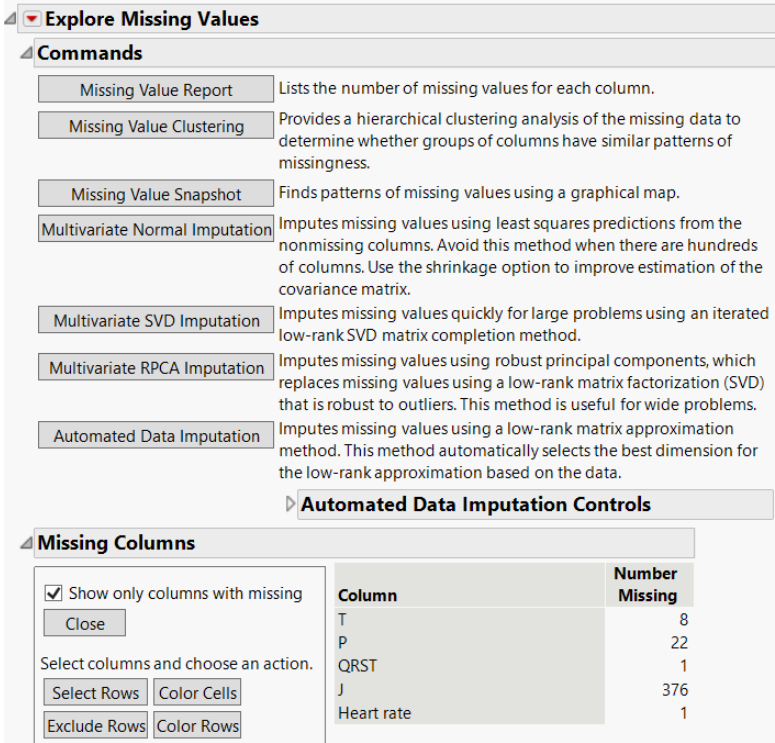
Example of the Explore Missing Values Platform

You are interested in exploring data for missing values and imputing them when appropriate. First explore your data for missing values, then impute values for columns that have a continuous modeling type.

Examine Missing Values

1. Select **Help > Sample Data Folder** and open Arrhythmia.jmp.
2. Select **Analyze > Screening > Explore Missing Values**.
3. Select all columns (280 in all) and click **Y, Columns**.
4. Click **OK**. Select the **Show only columns with missing** check box.

Figure 22.2 Missing Value Report



The Missing Columns report shown in [Figure 22.2](#) indicates that only five columns have missing data. Out of a total of 452 rows, Column J has 376 missing values. Because it is largely missing, imputed values might not result in a meaningful analysis. For such data, you could explore a model using the Informative Missing option for Column J in a platform that supports the Informative Missing option.

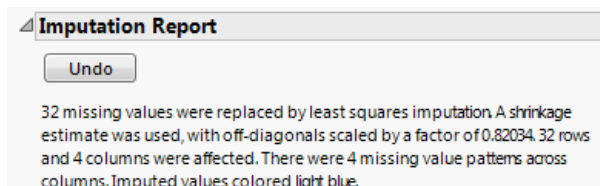
Impute Missing Values

The five columns that have missing values are continuous. You proceed to impute values for the four columns other than Column J using multivariate imputation for the continuous columns in your data table. By doing so, you tacitly assume that the probabilities that values are missing depend only on the values of the continuous variables and not on the values of excluded nominal variables. To conduct this new analysis, you need to launch the Explore Missing Values platform again.

1. Select **Analyze > Screening > Explore Missing Values**.
2. In the launch window, click the red triangle next to **280 Columns**.
Use the columns filter menu to view only the columns with a Continuous modeling type in the Select Columns list.

3. Select **Modeling Type > Uncheck All**.
This removes all columns from the Select Columns list.
4. Select **Modeling Type > Continuous**.
The Select Columns list now contains only the 207 columns that are Continuous.
5. Select all 207 columns. Then Ctrl-click the J column (to deselect it) and click **Y, Columns**.
6. Click **OK**.
7. Click **Multivariate Normal Imputation**.
A window appears and asks whether you want to use a Shrinkage estimator for covariances.
8. Click **Yes Shrinkage**.
A JMP Alert appears, informing you that you should use the **Save As** command to preserve your original data.
9. Click **OK**.

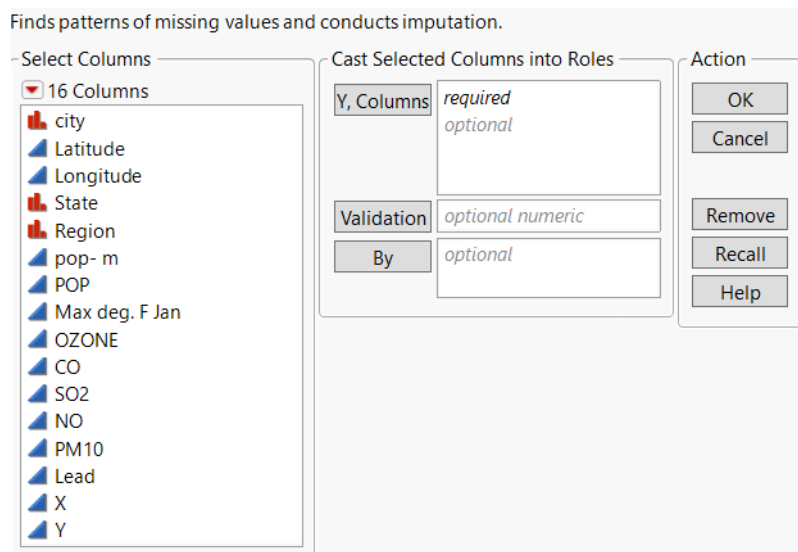
Figure 22.3 Imputation Report



The Imputation Report indicates how many missing values were imputed and the specific imputation details. No missing data remain in the four columns that had missing values.

Launch the Explore Missing Values Platform

Launch the Explore Missing Values platform by selecting **Analyze > Screening > Explore Missing Values**.

Figure 22.4 Explore Missing Values Launch Window


For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns Specifies the columns that you want to analyze. Each column must have a Numeric modeling type.

Validation Specifies a validation column that is used for some imputation methods.

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.

The Explore Missing Values Report

The Explore Missing Values report contains a Commands section, a Missing Columns report, and an Imputation report once an imputation method is selected. The Commands section includes several options for additional reports and imputation methods.

- “Commands”
- “Imputation Report”

Commands

Missing Value Report Shows the Missing Columns report, which lists the name of each column and the number of missing values in that column. The Missing Columns report also contains the following options:

Show only columns with missing Removes columns from the list that do not have missing values.

Close Closes the Missing Columns report.

Select Rows Selects the rows in the data table that contain missing values for the column(s) that you select in the Missing Columns report.

Exclude Rows Applies the excluded row state for rows in the data table that contain missing values for the column(s) that you select in the Missing Columns report.

Color Cells Colors the cells in the data table that contain missing values for the column(s) that you select in the Missing Columns report.

Color Rows Colors the rows in the data table that contain missing values for the column(s) that you select in the Missing Columns report.

To remove the Missing Columns report, click the Close button.

Missing Value Clustering Provides a hierarchical clustering analysis of the missing data. The report include a plot and two dendrograms. The rows of the plot are defined by the missing data patterns; there is a row for each pattern. The columns correspond to the variables. Each red cell indicates a group of missing values for the column listed beneath the plot. Hover over a cell to see the list of values represented. Click in the plot to select missing data pattern rows. Vertical bars are displayed to indicate the selected patterns.

- The dendrogram to the right of the plot shows clusters of missing data pattern rows. These are the rows that you would obtain by using Tables > Missing Data Pattern.
- The dendrogram beneath the plot shows clusters of variables.

Use this report to determine whether certain groups of columns tend to have similar patterns of missing values. To remove the Missing Value Clustering report, click the Close button.

Missing Value Snapshot Shows a cell plot for the missing values. The columns represent the variables. Black cells indicate a missing value. This plot is especially useful in understanding missingness for longitudinal data, where subjects can withdraw from a study before the end of the data collection period. To remove the Missing Value Snapshot report, click the Close button.

Multivariate Normal Imputation Imputes missing values using least squares predictions from the nonmissing columns. Use the shrinkage option to improve estimation of the covariance matrix.

Caution: Avoid this method when there are hundreds of columns.

Multivariate SVD Imputation Imputes missing values quickly for large problems using an iterated low-rank SVD matrix completion method. When you click Multivariate SVD Imputation, the Imputation Method window shows the recommended settings, which can be adjusted.

Number of Singular Vectors Number of singular vectors that are computed and used in the imputation.

Note: It is important not to specify too many singular vectors, otherwise the SVD and the imputations do not change from iteration to iteration.

Maximum Iterations The number of iterations used in imputing the missing values.

Show Iteration Log Opens a Details report that shows the number of iterations and gives details about the criteria.

For large problems, a progress bar shows how many dimensions the SVD has completed. You can stop the imputation and use that number of dimensions at any time.

Multivariate RPCA Imputation Imputes missing values using robust principal components, which replaces missing values using a low-rank matrix factorization that is robust to outliers.

Tip: This method is useful for wide problems.

Automated Data Imputation Imputes missing values using a low-rank matrix approximation method. This method automatically selects the best dimension for the low-rank approximation based on the data. Before selecting this method, you can specify options for saving the imputed values and other advanced controls.

Create New Data Table Creates a new data table that has the same dimensions as the original data table. In the new data table, the columns selected in the launch window contain the imputed values.

Save Scoring Formula to Current Data Table Saves a column group, named `Imputed_`, to the current data table that contains the imputed columns specified in the launch window. A hidden column, `ADI Impute Column`, is also added to the current data table that contains the imputed vectors and the scoring formula used in the data imputation. The column formulas automatically update if any additional rows are added to the data table, enabling missing data imputation for streaming data. This is the default option.

Impute Values in Place Imputes the missing values in the current data table. The imputed values are displayed in light blue.

Include informative missing columns (Available only when **Save Scoring Formula to Current Data Table** is selected.) Adds an additional indicator column to the `Imputed_` column group for each imputed column specified in the launch. This column specifies if each row is missing or not missing for each of the imputed *Y* columns.

Dimension Upper Bound Determines the maximum rank allowed in the low-rank approximation. The default value is determined by the dimension of the matrix formed by the chosen columns.

Maximum Iterations Determines the number of values that are iterated over to determine the tuning parameter for the imputation model. The default is 10.

Proportion of Observations to Induce as Missing Determines the proportion of IM values that are added to the training and validation sets. The default proportion for each set is 0.2.

Proportion of Rows to Use for Validation Determines the proportion of rows to use in the training and validation sets. The default proportion for the validation set is 0.3.

Set Random Seed Determines the random seed for ADI. Use this option to obtain reproducible results.

Tip: To run a missing value command across all levels of a **By** variable, press Ctrl and click the desired command button.

Imputation Report

If you select one of the imputation methods, an Imputation Report is also included in the Explore Missing Values report window. The imputation report explains the results of the selected imputation process. The following results are included, depending on the selected method:

- The number of missing values that were replaced.
- The selected imputation method and any details specific to the selected method.

- The number of rows and columns that were affected.
- (Only for Multivariate Normal Imputation.) The number of different missing value patterns that were found.
- The color of the imputed values in the data table.
- (Only for Multivariate RPCA Imputation.) A Details report shows the number of iterations, the rank of the matrix, the value of the convergence criterion, and the value of the largest absolute scaled residual.

Once the imputation is complete, the cells corresponding to imputed values in the data table are colored. If the Missing Columns report is open, it is updated to show no missing values.

Click **Undo** to undo the imputation and replace the imputed data with missing values.

Explore Missing Values Platform Options

The Explore Missing Values red triangle menu contains the following options:

ADI Loading Matrix (Available only for the ADI method.) Shows or hides a report that contains the columns that correspond to the factor loading for each component.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

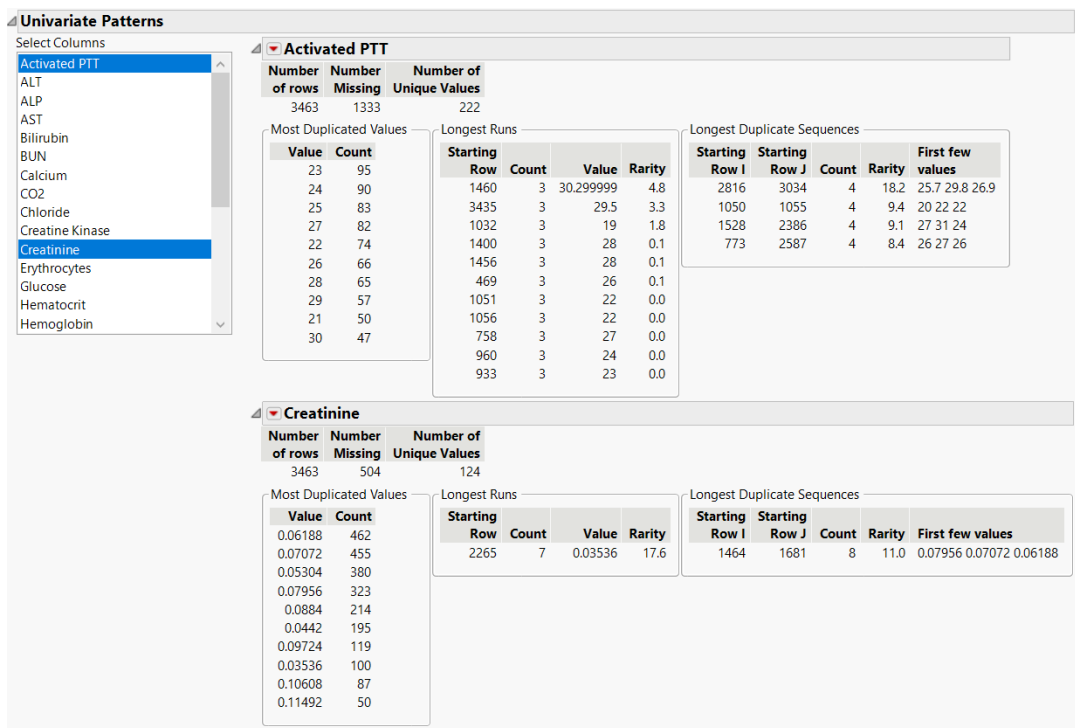
Chapter 23

Explore Patterns

Explore Data for Unusual Features or Patterns

The Explore Patterns platform enables you to detect unusual or unexpected patterns in your data. Although some patterns occur by chance or some innocent mechanism, they might indicate data that has been falsified or tampered with. The Explore Patterns platform is most useful when the values are precise so that matches and patterns are less likely to happen by coincidence. The Explore Patterns platform is also designed to handle large data tables efficiently.

Figure 23.1 Example of Explore Patterns Report



Contents

Overview of the Explore Patterns Platform.....	471
Example of the Explore Patterns Platform.....	471
Launch the Explore Patterns Platform.....	474
The Explore Patterns Report.....	475
Control Panel.....	475
Univariate Summary.....	475
Univariate Patterns.....	476
Explore Patterns Options.....	477
Univariate Patterns Report Options.....	480
Additional Example of the Explore Patterns Platform.....	481
Statistical Details for the Explore Patterns Platform.....	483

Overview of the Explore Patterns Platform

The Explore Patterns platform enables you to detect unusual or unexpected patterns in your data. Although some patterns occur by chance or some innocent mechanism, they might indicate data that has been falsified or tampered with. This platform is most useful when the values are precise so that matches and patterns are less likely to happen by coincidence. The Explore Patterns platform is also designed to handle large data tables efficiently.

The Explore Patterns platform looks for duplicate values and duplicate sequences of values. These types of patterns can indicate data tampering strategies such as pasting the same value many times, pasting the same value across a range of cells, or copying a range of data and pasting it somewhere else. The platform also analyzes properties of the formatted values to determine whether the data are actual measurements or might have been generated by a random number generator. There are also options to examine distribution properties and detect linear relationships between columns across groups of rows.

Note: The Explore Patterns platform ignores excluded rows, but the results are not necessarily equivalent to if you had deleted the row. For runs of rows, such as the Longest Runs and Longest Duplicated Sequences options, an excluded row stops the run.

Example of the Explore Patterns Platform

Use the Explore Patterns platform to determine whether there are any unexpected patterns in the laboratory results from a study of a drug.

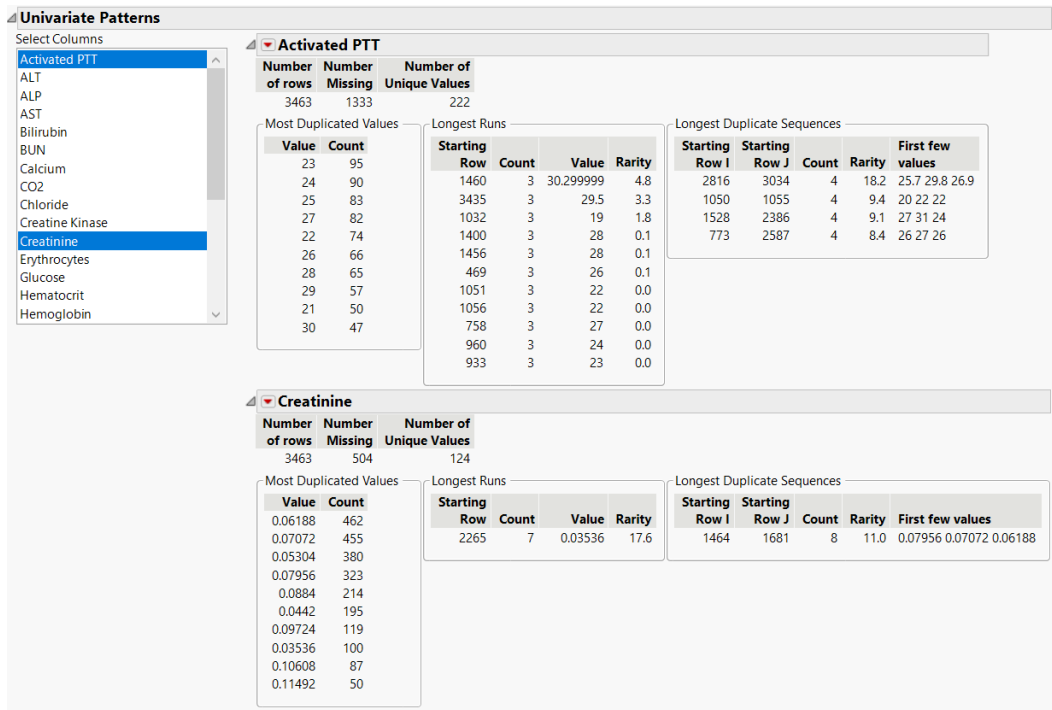
1. Select **Help > Sample Data Folder** and open the Nicardipine Lab Patterns.jmp sample data table.
2. Select **Analyze > Screening > Explore Patterns**.
3. Select the Laboratory Results column group and click **Y, Columns**.
4. Click **OK**.
5. Click the gray disclosure icon next to Univariate Summary to open the Univariate Summary report.

Figure 23.2 Explore Patterns Univariate Summary Report

Univariate Summary									
Longest Runs					Longest Duplicate Sequences				
Column	Starting Row	Count	Value	Rarity	Column	Starting Row I	Starting Row J	Count	Rarity
Creatinine	2265	7	0.03536	17.6	Activated PTT	2816	3034	4	18.2
Hematocrit	3048	4	28.999999	12.3	Prothrombin Time	2942	2966	6	14.9
BUN	2267	5	2.142	7.8	Hematocrit	3051	3057	3	14.6
Prothrombin Time	2931	6	12	6.7	Leukocytes	547	1275	3	11.1
LDH	3456	3	602	5.8	Creatinine	1464	1681	8	11.0
Activated PTT	1460	3	30.299999	4.8	Leukocytes	975	1485	3	10.8
Platelet	521	3	150	4.8	ALT	840	2425	3	10.5
Platelet	1092	3	336	4.1	Sodium	2339	2976	6	9.8
Activated PTT	3435	3	29.5	3.3	Leukocytes	456	2805	3	9.8
CO2	1390	4	82.612	3.1	Activated PTT	1050	1055	4	9.4
Creatine Kinase	3456	3	63	3.1	Potassium	1046	1067	6	9.4
Erythrocytes	815	3	4.3000002	3.0	Hemoglobin	1247	3273	3	9.2

- The Univariate Summary report summarizes the longest runs and longest duplicate sequences found across all of the columns. Each table is sorted by Rarity, a calculation of how likely the run or sequence is to occur by chance. The column with the highest rarity score for longest runs is Creatinine, whereas the column with the highest rarity score for longest duplicate sequences is Activated PTT. You will further examine these columns in the Univariate Patterns Report.
- In the Univariate Patterns report, the report for the first column, Activated PTT, is shown initially. Press Ctrl and select Creatinine from the Select Columns list.

Figure 23.3 Explore Patterns Univariate Patterns Report



As you select column names from the Select Columns list, the corresponding column reports are shown. The Activated PTT report shows that this column contains several three count runs and four count sequences. The Creatinine report shows that the longest run is 7 counts and the longest duplicated sequence is 8 counts.

Launch the Explore Patterns Platform

Launch the Explore Patterns platform by selecting **Analyze > Screening > Explore Patterns**.

Figure 23.4 Explore Patterns Platform Launch Window

Detects unusual patterns in the data.

Select Columns

▼ 37 Columns

Enter column name ⌵

- STUDYID
- SITEID
- USUBJID
- AGE
- SEX
- RACE
- Description of Planned Arm
- Vasoconstriction Flag
- Patient Died
- Measurement
- Laboratory Results (27/0)

Cast Selected Columns into Roles

Y, Columns required numeric
optional numeric

By optional

Action

OK

Cancel

Remove

Recall

Help

Minimum Run Size

Minimum Longest Duplicate Size

Minimum Cross Column Duplicate Run Size

Minimum Rows for Linear Relationship

☐ Include Missing

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Columns Assigns the columns to analyze. The variables must be numeric and continuous.

By Assigns one or more 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.

Minimum Run Size Specifies the minimum number of values in a row for a sequence to be reported as a run.

Minimum Longest Duplicate Size Specifies the minimum length of a sequence for it to be reported as a duplicate sequence within a column.

Minimum Cross Column Duplicate Run Size Specifies the minimum length of a sequence for it to be reported as a duplicate sequence across columns.

Minimum Rows for Linear Relationship Specifies the minimum number of rows used for detecting linear relationships between columns across groups of rows.

Note: The minimum number of rows must be greater than or equal to three.

Include Missing Specifies that missing values be included in pattern detection.

The Explore Patterns Report

The initial Explore Patterns report contains a Control Panel, Univariate Summary report, and Univariate Patterns report. Many of the tables in these reports contain rows that can be selected. If a row is selected and you right-click the first column, a menu appears that enables you to select or colorize the related rows and columns in the data table.

- [“Control Panel”](#)
- [“Univariate Summary”](#)
- [“Univariate Patterns”](#)

Control Panel

The control panel contains options that enable you to select which reports appear in the report window. These are the same options that are in the Explore Patterns red triangle menu. See [“Explore Patterns Options”](#). There is also a section of the control panel to specify report settings. These are the same settings that are found in the launch window. See [“Launch the Explore Patterns Platform”](#). If you change the settings in the control panel, the reports are automatically updated.

Univariate Summary

The Univariate Summary report combines results from several patterns reports across columns. By default, there are summary tables for Longest Runs and Longest Duplicate Sequences. Longest Runs and Longest Duplicate Sequences with rarity ≥ 2 are shown in the tables.

If the Fraction Length option is selected a Fraction Length summary table is also shown. This table shows the proportion of values in each column that have a fraction length greater than or equal to 15. Columns that do not contain any values that have fraction lengths greater than or equal to 15 are not shown in the Fraction Length summary table.

If at least one of the Y columns has a Spec Limit column property and the Distribution wrt Spec Limits option is selected, a Spec Limits Distrib summary table is also shown. This table shows the columns that have a statistically significant number of observations outside of the specification limits.

For more information about the Univariate Summary report contents, see [“Explore Patterns Options”](#).

Univariate Patterns

The Univariate Patterns report contains a Select Columns list and one or more column-specific patterns reports. The column-specific reports that are displayed are determined by the Select Columns list. To examine each column’s report, select the first column and then successively use the down arrow key to advance to the next column. Alternatively, you can click a column name in this list to display the patterns report for the selected column. If you select more than one column name by using either Ctrl or Shift, a patterns report for each of the selected columns is displayed. If you deselect one or more columns from the Select Columns list, the reports for the corresponding columns are removed.

Each individual patterns report displays a summary table that contains the number of rows, number of missing observations, and number of unique values for the given column. By default, there are also tables for Most Duplicated Values, Longest Runs, and Longest Duplicate Sequences. If the Formatted Widths, Fraction Lengths, or Leading and Trailing Digits options are selected from either the control panel or the Explore Patterns red triangle menu, a corresponding table is added to each individual patterns report.

If you right-click on a row in any of the tables in the Univariate Patterns report, a menu of right-click options appears.

Select Rows and Column Selects the column and rows in the original data table that correspond to the selected row in a specific table in the Univariate Patterns report. For example, if you right-click the second row of the Leading Digits table for a specific column, the rows in that column that have a leading digit of 2 are selected in the original data table.

Colorize Cells Colors the cells in the original data table that correspond to the selected row in a specific table in the Univariate Patterns report. The rows of the colored cells are also selected.

There are red triangle options for the column-specific patterns report. See [“Univariate Patterns Report Options”](#).

Explore Patterns Options

The control panel and Explore Patterns red triangle menu contain options for summary and pattern tables.

Most Duplicated Values Shows or hides a table of the values that appear most frequently in the data. Shows up to ten numbers for each column.

Longest Runs Shows or hides a table of the longest run of the same value. The table contains the following information:

Starting Row The row number that corresponds to the start of the run.

Count The length of the run.

Value The value of the entries in the run.

Rarity A measurement of how rare it would be for the run to occur by chance. A high rarity indicates that it is not likely that the run occurred by chance. See [“Rarity in Longest Runs”](#).

Longest Duplicated Sequences Shows or hides a table of the longest sequence of values that appears more than once in the same column. Sequences of repeated values are also considered. If a sequence is duplicated more than once, each instance appears in the table. The table contains the following information:

Starting Row I The row number that corresponds to the start of the first sequence.

Starting Row J The row number that corresponds to the start of the duplicate sequence.

Count The length of the sequence.

Rarity A measurement of how rare it would be for the duplicate sequence to occur by chance. A high rarity indicates that it is not likely that the duplicate sequence occurred by chance. See [“Rarity in Longest Sequences”](#).

First few values The first few values of the duplicate sequence. Shows up to three values for each sequence.

Formatted Widths Shows or hides a table of overall widths and decimal widths of the formatted values. The table contains the following information:

Width The width size.

Overall Count The number of observations in the column whose overall width equals the specified width.

Decimal Count The number of observations in the column whose decimal width equals the specified width.

Fraction Length Shows or hides a table of the continued fraction lengths for the values. A continued fraction is a representation of a number as a sequence of continually divided terms. The fraction length can indicate how rational a number is. Lengths of 15 or more indicate that the number is an irrational number, such as from a root function, a function involving an irrational number, or from a random number generator (Benford,1938).

Note: In Explore Patterns, the maximum fraction length is 15. Numbers with fraction lengths of 15 or more are all reported as 15.

Length The length of the continued fraction.

Count The number of observations in the column whose continued fraction length equals the specified continued fraction length.

Leading and Trailing Digits Shows or hides a table of counts for each leading and trailing digit, 1 through 9. For the leading digits, there are also columns for the expected rate and the minimum and maximum values that contain the corresponding leading digit. The expected rate is based on Benford's Law, which states that the frequency that a number has a first digit that is equal to a is $F_a = \log_{10}(a+1/a)$. This law applies to data that have a large range of numbers that all have four or more digits.

Spec Limit Matches (Available when at least one of the columns has a Spec Limit column property.) Shows or hides a table of values that are exactly the lower specification limit, upper specification limit, or target value. If there are no exact matches, the report is not shown.

Distribution wrt Spec Limits (Available only when at least one of the columns has a Spec Limit column property.) Shows or hides a table of observed versus expected observations that are outside of the specification limits. The Chi-Square and PValue columns appear only if the number of observed observations is significantly higher than the number of expected observations.

Duplicates Across Columns Shows or hides the Duplicates Across Columns report. The report contains a table of sequences of values that appear in the same rows across more than one column. There is a note at the top of the report that states the minimum length of a sequence for it to be considered a duplicate. This number is specified in the launch window or control panel as Minimum Cross Column Duplicate Run Size. There is also a Colorize option at the top of the report. See "[Colorize Duplicates Across Columns](#)".

Colorize Duplicates Across Columns (Available only after Duplicates Across Columns has been selected.) Colors cells in the data table to correspond to the duplicate matches found in the Duplicates Across Columns report. The rows that contain the duplicates are also selected.

Linear Relationships Shows or hides the Linear relationships between variables report. The report contains a table of exact linear relationships across columns over a specified number of sequential rows. An exact linear relationship is defined as a linear relationship that has an R^2 value greater than or equal to 0.999999. There is a note at the top of the report that states the minimum number of rows considered. This number is specified in the launch window or control panel as Minimum Rows for Linear Relationships. There is also a Colorize option at the top of the report. See [“Colorize Linear Relationships”](#).

Colorize Linear Relationships (Available only after Linear Relationships has been selected.) Colors cells in the data table to correspond to the linear relationships found in the Linear relationships between variables table. The rows that contain the linear relationships are also selected.

Save Columns Contains the following save options:

Save Table of Runs Saves the longest runs for all columns to a new data table. The data table contains the same columns as the Longest Runs table in the Univariate Patterns report and is sorted by Starting Row.

Save Table of Duplicate Sequences Saves the longest duplicate sequences for all columns to a new data table. The data table contains the same columns as the Longest Duplicate Sequences table in the Univariate Patterns report and is sorted by Starting Row I.

Save Duplicates Across Columns Saves the Duplicates Across Columns table to a new data table.

Save Linear Relationships Saves the Linear relationships between variables table to a new data table.

Ordering Columns Contains options for ordering. When you change the ordering, the column list in the Univariate Patterns report is reordered. The top column is automatically selected and the report for only that column is displayed.

Order by Runs Rarity Orders the column list by the rarity values of the longest runs. Higher rarity values are at the top.

Order by Sequence Rarity Orders the column list by the rarity values of the longest duplicate sequences. Higher rarity values are at the top.

Order by Column Name Orders the column list in alphabetical order by column name.

Original Order Orders the column list in the original order in which the columns were specified in the launch window.

Clear Cell Colors Clears the cell colors for columns selected in the data table or for all columns if no columns are selected in the data table.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Univariate Patterns Report Options

Each Univariate Patterns red triangle menu contains the following options:

Select Most Duplicated Selects the observations in the data table that correspond to the values in the Most Duplicated Values table.

Select Longest Runs Selects the observations in the data table that correspond to the values in the Longest Runs table.

Select Longest Duplicate Sequences Selects the observations in the data table that correspond to the values in the Longest Duplicate Sequences table.

Select by Formatted Width Selects the observations in the data table that correspond to the values in the Formatted Widths table that have the user-specified overall width.

Select by Decimal Width Selects the observations in the data table that correspond to the values in the Formatted Widths that have the user-specified decimal width.

Select by Fraction Length Selects the observations in the data table that correspond to the values in the Fraction Length table that have the user-specified fraction length.

Select by Leading Digit Selects the observations in the data table that correspond to the values in the Leading and Trailing Digits table that have the user-specified leading digit.

Additional Example of the Explore Patterns Platform

Use the Explore Patterns platform to determine whether there are any unexpected patterns in sensor measurements from sensors in an urban waste water treatment plant.

1. Select **Help > Sample Data Folder** and open the Water Treatment.jmp sample data table.
2. Select **Analyze > Screening > Explore Patterns**.
3. Select the Sensor Measurements column group and click **Y, Columns**.
4. Click **OK**.
5. In the Control Panel, under Across Columns, click the box next to Duplicates Across Columns.

Figure 23.5 Duplicates Across Columns Report

Duplicates Across Columns				
With runs at least 2 Colorize				
from Row	to Row	Number Rows	Number Equal	Duplicate Sets
219	230	12	12	PH-P, PH-D
207	217	11	11	PH-E, PH-P
214	217	4	8	PH-E, PH-P, PH-D
330	333	4	8	PH-E, PH-P, PH-D
198	204	7	7	PH-P, PH-D
91	96	6	6	PH-P, PH-D
94	96	3	6	PH-E, PH-P, PH-D
208	210	3	6	PH-E, PH-P, PH-D
216	217	2	6	PH-E, PH-P, PH-D, PH-S
228	229	2	6	PH-E, PH-P, PH-D, PH-S
247	252	6	6	PH-E, PH-P
290	295	6	6	PH-E, PH-P
293	295	3	6	PH-E, PH-P, PH-D
427	429	3	6	PH-E, PH-P, PH-D
472	477	6	6	PH-E, PH-P
26	30	5	5	PH-E, PH-P
63	67	5	5	PH-P, PH-D
141	145	5	5	PH-E, PH-D
299	303	5	5	PH-E, PH-P
352	356	5	5	PH-P, PH-D

The Duplicates Across Columns report shows a long list of duplicate columns. However, the minimum run size is 2 and there are many small runs in the list. You want to focus on the larger runs of at least 10.

6. In the Control Panel, under Settings, enter 10 in the box next to Minimum Cross Column Duplicate Run Size.

Tip: Click anywhere outside of the box to automatically update the report.

Figure 23.6 Duplicates Across Columns Report With Runs at Least 10

Duplicates Across Columns				
With runs at least 10 Colorize				
from Row	to Row	Number Rows	Number Equal	Duplicate Sets
219	230	12	12	PH-P, PH-D
207	217	11	11	PH-E, PH-P

The updated Duplicates Across Columns report shows that there are two duplicate sets with runs of at least 10. Both sets include the PH-P column.

7. Click **Colorize** in the Duplicates Across Columns Report.

Figure 23.7 Colorized Water Treatments.jmp Data Table

		PH-E	DBO-E	DQO-E	SS-E	SSV-E	SED-E	COND-E	PH-P	DBO-P	SS-P	SSV-P	SED-P	COND-P	PH-D
	206	7.7	124.0	234.0	120.0	61.7	1.8	1550.0	7.9	123.0	220.0	59.0	4.0	1651.0	7.8
	207	7.6	172.0	394.0	226.0	51.3	5.5	1324.0	7.6	197.0	302.0	47.0	4.0	1306.0	7.7
	208	7.6	268.0	324.0	226.0	53.1	4.0	1390.0	7.6	265.0	236.0	53.4	4.5	1389.0	7.6
	209	7.6	176.0	200.0	176.0	50.0	3.0	859.0	7.6	149.0	172.0	51.2	2.0	860.0	7.6
	210	7.6	146.0	243.0	164.0	57.3	3.2	1095.0	7.6	193.0	208.0	55.8	3.0	1089.0	7.6
	211	7.8	139.0	380.0	256.0	53.1	5.5	2040.0	7.8	201.0	340.0	47.6	8.5	2000.0	7.9
	212	7.9	205.0	529.0	264.0	61.4	6.5	2750.0	7.9	220.0	324.0	60.5	6.0	2850.0	7.8
	213	7.8	117.0	366.0	500.0	27.2	4.2	3230.0	7.8	135.0	620.0	32.3	7.0	3170.0	7.7
	214	7.6	148.0	400.0	564.0	31.6	7.0	1554.0	7.6	235.0	798.0	28.6	11.0	1515.0	7.6
	215	7.8	158.0	256.0	194.0	47.4	3.0	1831.0	7.8	218.0	240.0	46.7	4.0	1860.0	7.8
	216	7.7	123.0	246.0	186.0	39.8	2.5	1288.0	7.7	241.0	300.0	25.3	4.5	1328.0	7.7
	217	7.7	133.0	410.0	204.0	62.7	4.5	1490.0	7.7	161.0	256.0	57.8	4.5	1421.0	7.7
	218	7.5	237.0	388.0	178.0	57.3	4.0	1620.0	7.6	300.0	240.0	53.3	4.5	1571.0	7.7
	219	7.7	203.0	357.0	222.0	58.6	3.0	1828.0	7.6	215.0	208.0	73.1	5.0	1817.0	7.6
	220	7.8	199.0	380.0	244.0	55.7	5.0	1453.0	7.8	110.0	196.0	60.2	3.5	1430.0	7.8
	221	7.8	214.0	941.0	304.0	53.3	6.5	1454.0	7.8	396.0	320.0	57.5	7.0	1528.0	7.8
	222	7.7	208.0	298.0	260.0	37.7	4.0	1236.0	7.8	245.0	284.0	46.5	5.0	1285.0	7.8
	223	7.9	206.0	305.0	188.0	56.4	4.0	1200.0	7.9	220.9	272.0	50.0	6.0	1228.0	7.9
	224	7.7	247.0	361.0	286.0	51.0	4.5	1180.0	7.7	146.0	212.0	56.4	2.5	1251.0	7.7
	225	7.5	438.0	681.0	370.0	56.8	6.0	1396.0	7.6	212.0	348.0	64.4	6.0	1446.0	7.6
	226	7.8	255.0	282.0	166.0	67.5	2.0	950.0	7.8	284.0	172.0	65.1	2.0	965.0	7.8
	227	7.7	283.0	310.0	170.0	61.2	2.5	1065.0	7.8	235.0	270.0	51.1	5.5	1100.0	7.8
	228	7.9	313.0	341.0	512.0	33.6	6.0	915.0	7.9	205.0	520.0	33.8	7.5	976.0	7.9
	229	7.7	300.0	610.0	452.0	45.1	8.0	1313.0	7.7	375.0	556.0	43.9	10.0	1312.0	7.7
	230	7.8	177.0	412.0	196.0	69.4	4.5	2190.0	7.8	330.0	476.0	52.1	7.5	2330.0	7.8

The Colorize feature colors the cells in the original data table that correspond to the duplicate sets in the report.

Statistical Details for the Explore Patterns Platform

This section contains statistical details for calculating the rarity in longest runs and longest sequences in the Explorer Patterns platform.

Rarity in Longest Runs

To calculate the rarity for longest runs, first define the following variables:

n = the number of rows in the column

k = the number of times a specific value occurs in the column

$p = k/n$ = the probability of observing the specific value in the column

m = the length of the run

N = the number of unique runs

Then, the rarity for longest runs is calculated as follows:

$$Rarity = -\log_2(1 - (1 - p^m)^N)$$

Rarity in Longest Sequences

To calculate the rarity for longest sequences, first define the following variables:

p = the probability of observing the specific sequence one time in the column

k = the number of times the starting value of the sequence occurs in the column

Then, the rarity for longest sequences is calculated as follows:

$$Rarity = -\log_2(1 - (1 - p)^k)$$

Response Screening

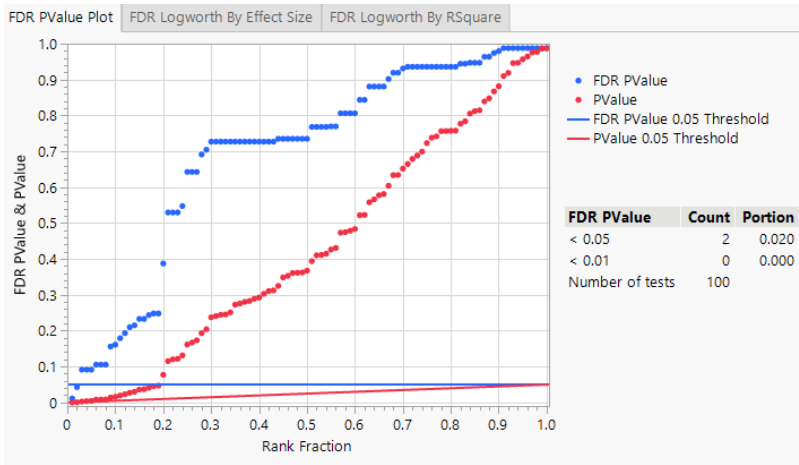
Test Many Responses in Large-Scale Data

Response Screening automates the process of conducting tests across a large number of responses. The test results and summary statistics are presented directly in the report, or in data tables to enable further data exploration. Testing many responses for the effects of factors can be challenging, if not misleading, without appropriate methodology. A False-Discovery Rate approach guards against incorrect declarations of significance. Plots of p -values are scaled using the logworth, making them easily interpretable.

Because large scale data sets are often messy, Response Screening presents methods that address irregularly distributed and missing data. A robust estimate method allows outliers to remain in the data, but reduces the sensitivity of tests to these outliers. Missing data options allow missing values to be included in the analysis. These features enable you to analyze your data without first conducting an extensive analysis of data quality.

When you have many observations, even differences that are of no practical interest can be statistically significant. Response Screening presents tests of practical difference, where you specify the difference that you are interested in detecting. On the other hand, you might want to know whether differences do not exceed a given magnitude, that is, if the means are equivalent. For this purpose, Response Screening presents equivalence tests.

Figure 24.1 Example of a Response Screening Plot



Contents

Overview of the Response Screening Platform	487
Example of Response Screening	488
Launch the Response Screening Platform	493
The Response Screening Report	497
Response Screening Plots	497
Response Screening Report Tables	499
Response Screening Platform Options	503
Means Data Table	506
Means Differences Data Table	507
The PValues Data Table	509
The Response Screening Personality in Fit Model	510
Launch Response Screening in Fit Model	511
The Fit Response Screening Report	512
Fit Response Screening Options	514
Additional Examples of Response Screening	516
Example of Tests of Practical Significance and Equivalence	516
Example of the Max Logworth Option	518
Example of Robust Fit	520
Example of the Response Screening Personality	523
Example of a Volcano Plot	524
Statistical Details for the Response Screening Platform	526

Overview of the Response Screening Platform

Response Screening automates the process of conducting tests across a large number of responses. It tests each response that you specify against each factor that you specify. Response Screening addresses two main issues connected to large-scale data. These are the need to conduct many tests and the requirement to deal effectively with outliers and missing values.

Response Screening is available as a platform and as a Fit Model personality. In both cases, it performs tests analogous to those found in the Fit Y by X platform, as shown in [Table 24.1](#). As a personality, it performs tests of the responses against the individual model effects.

To facilitate and support the multiple inferences that are required, Response Screening provides these features:

Data Tables Results are available in data tables, as well as in a report, to enable you to explore, sort, search, and plot your results. Statistics that facilitate plot interpretation are provided, such as the logworth of p -values ($-\log_{10}(p\text{-value})$).

False Discovery Rates Because you are conducting a large number of tests, you need to control the overall rate of declaring tests significant. Response screening controls the *false discovery rate*. The False Discovery Rate (FDR) is the expected proportion of significant tests that are incorrectly declared significant (Benjamini and Hochberg [1995](#); Westfall et al. [2011](#)).

Tests of Practical Significance When your data table consists of a large number of rows (large n), the standard error used in testing can be very small. As a result, tests might be statistically significant, when in fact, the observed difference is too small to be of practical consequence. To address this issue, you can define an effect size that you consider to be of *practical significance*. You then conduct tests of practical significance, thereby only detecting effects large enough to be of pragmatic interest.

Equivalence Tests When you are studying many factors, you are often interested in those that have essentially equivalent effects on the response. In this case, you can specify an effect size that defines practical equivalence and then conduct equivalence tests.

To address issues that arise when dealing with messy data, Response Screening provides features to deal with outliers and missing data. These features enable you to analyze your data directly, without expending effort to address data quality issues:

Robust Estimation Outliers in your data increase estimates of standard error, causing tests to be insensitive to real effects. Select the Robust option to conduct Huber M-estimation. Outliers remain in the data, but the sensitivity of tests to these outliers is reduced.

Missing Value Options The platform contains an option to treat missing values in categorical predictors in an informative fashion.

Table 24.1 Analyses Performed by Response Screening

Response	Factor	Fit Y by X Analysis	Description
Continuous	Categorical	Oneway	Analysis of Variance
Continuous	Continuous	Bivariate	Simple Linear Regression
Categorical	Categorical	Contingency	Chi-Square
Categorical	Continuous	Logistic	Simple Logistic Regression

The Response Screening platform generates a section of plots and a section of reports. The FDR PValue plot and the Result Table are shown by default. The Response Screening personality generates a report that contains an Effect Tests table, the FDR PValue Plot for Effects plot, and the FDR Logworth by Effect Size plot.

The JSL command `Summarize Y by X` performs the same function as the Response Screening platform but without creating a platform window. See the *JSL Syntax Reference*.

Example of Response Screening

Use the Response Screening platform to determine which of the response characteristics show different values across a process change.

1. Select **Help > Sample Data Folder** and open `Probe.jmp`.
2. Select **Analyze > Screening > Response Screening**.
3. Select the Responses column group and click **Y, Response**.
4. Select Process and click **X**.
5. Click the gray triangle next to Advanced Options.
6. Enter 100 in the box next to **Max Logworth**.

A logworth ($-\log_{10}(p\text{-value})$) of 100 or larger corresponds to an extremely small p -value. Setting a value for the Max Logworth helps control the scale of plots as it limits the reported logworth value.

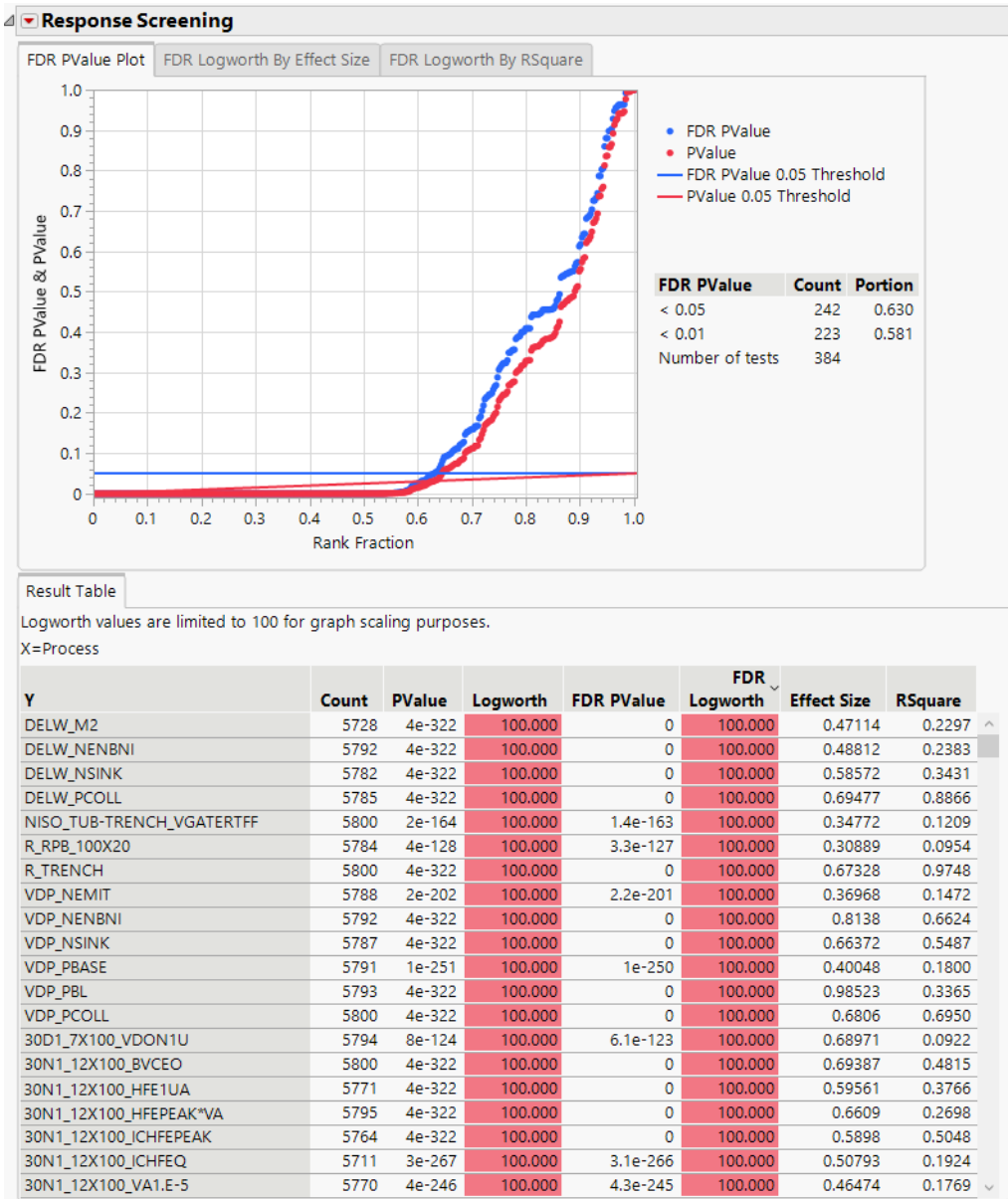
7. Click **OK**.

The report shows a table of p -value results and three plots in a tabbed view. The FDR PValue Plot is visible by default.

The FDR PValue Plot shows two types of p -values, FDR PValue and PValue, for each of the 384 tests. These are plotted against Rank Fraction. PValue is the usual p -value for the test of a single Y against Process. The FDR PValue is a p -value that is adjusted to guarantee a given false discovery rate (FDR). In this example, the false discovery rate is 0.05. The FDR PValues are plotted in blue and the PValues are plotted in red. The Rank Fraction ranks the FDR p -values from smallest to largest, in order of decreasing significance.

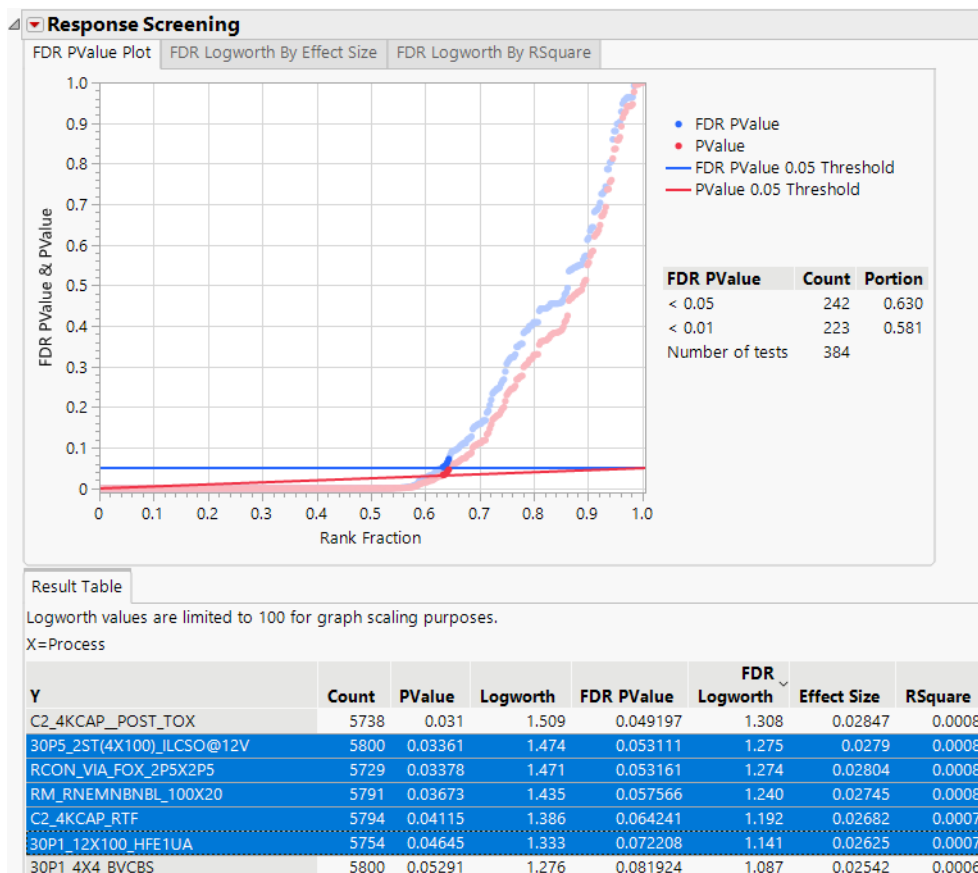
Both the horizontal blue line and the sloped red line on the plot are thresholds for significance. Tests with FDR p -values that fall below the blue line are significant at the 0.05 level when adjusted for the false discovery rate. Tests with ordinary p -values that fall below the red line are significant at the 0.05 level without adjusting for the false discovery rate. In this way, the plot enables you to read FDR significance from either set of p -values.

Figure 24.2 Response Screening Report for 384 Tests against Process



The FDR PValue Plot shows that more than 60% of the tests are significant. A handful of tests are significant using the unadjusted p -value, but not significant using the FDR p -value. These tests correspond to the red points that are above the red line, but below the blue line. You can also view these tests by scrolling down in the Result Table to the first FDR PValue that is greater than 0.05. The corresponding ordinary p -value is 0.03361.

Figure 24.3 Response Screening Report Tests



To identify the characteristics that are significantly different across Process, you can drag a rectangle around the appropriate points in the plot. This selects the rows corresponding to these points in the Result Table, where the names of the characteristics are given in the first column.

The Result Table contains 387 rows, one for each response measure in the Responses group. The response is given in the first column, called Y. Each response is tested against the effect in the X column, namely, Process. Note that the last three rows of the Result Table contain missing values. This is why the plot reports 384 tests. The response columns that correspond to the last three rows contain constant measurements, so no analyses are performed.

The remaining columns give information about the test of Y against X. Here the test is a Oneway Analysis of Variance. In addition to other information, the table gives the test's *p*-value, logworth, FDR (False Discovery Rate) *p*-value, and FDR logworth. Use this table to sort by the various statistics, select rows, or plot quantities of interest.

Notice that logworth and FDR logworth values that correspond to p -values of $1e-100$ or less are reported as 100, because you set Max Logworth to 100 in the launch window. Also, cells corresponding to FDR logworth values greater than two are colored with an intensity gradient.

See [“The Response Screening Report”](#) for more information about the report and PValues table.

Launch the Response Screening Platform

Launch the Response Screening platform by selecting **Analyze > Screening > Response Screening**.

Figure 24.4 Response Screening Launch Window

Fits a large number of Y by X combinations while controlling the False-Discovery Rate.

Select Columns

394 Columns

Enter column name

Lot ID

Wafer Number

Process

Die X

Die Y

Site

Start Time

Responses (387/0)

Cast Selected Columns into Roles

Y, Response

required

optional

X

required

optional

Grouping

optional

Weight

optional numeric

Freq

optional numeric

By

optional

Action

OK

Cancel

Remove

Recall

Help

☐ Common Y Scale

☐ Common X Scale

Means Comparisons

Each with Control

Ratio Adjustment

No Adjustment

☐ Missing is category

☐ Paired X and Y

☐ PValues Table on Launch

☐ No Report

Advanced Options

☐ Robust

☐ Cauchy

☐ Poisson Y

☐ Empirical Bayes Shrinkage

☐ Kappa

☐ Corr

Max Logworth

.

☐ Force X Categorical

☐ Force X Continuous

☐ Force Y Categorical

☐ Force Y Continuous

☐ Unthreaded

For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Launch Window Roles

Y, Response Identifies the response columns containing the measurements to be analyzed.

X Identifies the columns against which you want to test the responses.

Grouping Analyzes the rows assigned to each level of the specified column separately. All results are presented in a single table and report.

Weight Identifies a column whose values assign a weight to each row. These values are used as weights in the analysis. See *Fitting Linear Models*.

Freq Identifies a column whose values assign a frequency to each row. These values enable you to account for pre-summarized data. See *Fitting Linear Models*.

By For each level of the specified column, analyzes the corresponding Ys and Xs and presents the results in separate tables and reports.

Launch Window Options

Common Y Scale (Applies only when there is more than one continuous response.) Notifies the platform that all of the continuous responses are on the same scale. This is necessary to compare mean differences or slopes. If this option is selected and there is at least one categorical X variable, the Logworth by Difference plot and the Means Differences report are shown by default in the Response Screening report window. If not selected, the Logworth by Difference plot is accessible through the Response Screening red triangle menu.

Common X Scale (Applies only when there is more than one continuous X variable.) Notifies the platform that all of the continuous X variables are on the same scale. This is necessary to compare the slopes of different X variables. If both the Common X Scale and Common Y Scale options are selected, the Logworth by Slope plot is shown by default in the Response Screening report window. If not selected, the Logworth by Slope plot is accessible through the Response Screening red triangle menu.

Means Comparisons (Applies only when there is at least one continuous response and at least one categorical X variable.) Specifies how the means of different levels of an X variable are compared.

Each with Control For each categorical X variable, compares each level with a control group level. The default control group level is the first level.

Tip: You can add a Control Level column property to the factor column to specify a control group that is different from the default control group. See *Using JMP*.

All Combinations For each categorical X variable, compares all possible level combinations.

Ratio Adjustment Provides options to add 0.5 to cell counts when calculating risk ratios, odds ratios, and risk differences. This adjustment avoids problems that arise from dividing by zero, if any of the cell counts are zero. See Haldane (1956). Use the ratio adjustment options if you have at least one nominal response and at least one categorical predictor.

No Adjustment No adjustment is made to the cell counts.

Add 0.5 when any zero Adds 0.5 to cells that have a count of zero when calculating the risk ratio, odds ratio, and risk difference.

Add 0.5 always Adds 0.5 to all cell counts when calculating the risk ratio, odds ratio, and risk difference.

Missing is category For any categorical X variable, treats missing values on X as a category.

Paired X and Y Performs tests only for Y columns paired with X columns according to their order in the **Y, Response** and **X** lists. The first Y is paired with the first X, the second Y with the second X, and so on.

Note: The following features are not supported if the Paired X and Y option is selected: Show Means Differences, Practical Differences and Equivalences, Show Slopes, and 2-by-M reports.

PValues Table on Launch Creates a data table for the p -values and individual model fit statistics. This data table is linked to the Result Table in the Response Screening report.

No Report Suppresses the report window. Use this option to run Save commands to obtain results without the report window appearing.

Advanced Options

Robust (Robust results shown only for continuous responses.) Fits regression and ANOVA models using the Huber M-estimation method (Huber and Ronchetti 2009). This method estimates parameters by minimizing the Huber loss function. The Huber loss function penalizes outliers, which reduces the influence of the outliers on tests. If there are no outliers, the estimates are close to the least squares estimates. When you select the Robust option, several Robust measures and model fit statistics columns are added to the Result

Table. See *Basic Analysis* for more information about Huber M-estimation. For an example, see [“Example of Robust Fit”](#).

Caution: The Robust option can increase processing time.

Cauchy (Cauchy results shown only for continuous responses.) Estimates parameters using maximum likelihood and a Cauchy link function. This estimation method assumes that the errors have a Cauchy distribution, which has fatter tails than the normal distribution, resulting in a reduced emphasis on outliers. This option can be useful if you have a large proportion of outliers in your data. However, if your data are close to normal with only a few outliers, this option can lead to incorrect inferences. When you select the Cauchy option, several Cauchy measures and model fit statistics columns are added to the Result Table.

Poisson Y Fits each Y response as a count having a Poisson distribution. The test is performed only when the response is continuous and the X variable is categorical. This option is appropriate when your responses are counts.

Empirical Bayes Shrinkage (Applicable only if the Common Y Scale option is selected.) Shrinks residual variance estimates towards an estimated prior mode, borrowing strength across all estimates. This is useful when screening many continuous responses that are on a common scale. By applying the prior, there is an increase in the degrees of freedom and sum of squares error for each individual fit. This leads to a slight increase in power for the tests. See Smyth (2004).

Kappa (Applies only if Y and X are both categorical and have the same number of levels.) Calculates Kappa, a measure of agreement between Y and X. Results are reported in a column called Kappa in the Result Table.

Corr Calculates the Pearson product-moment correlation. Results are reported in a column called Corr in the Result Table. For categorical variables, the correlation is calculated in terms of the indices defined by the value ordering.

If X and Y are both binary, the Pearson product-moment calculation gives Spearman's Rho and Kendall's Tau-b. Otherwise, a value of Corr that is large in magnitude indicates an association; a Corr value that is small in magnitude does not preclude an association.

Max Logworth Controls the scale of plots involving logworth values ($-\log_{10}$ of p -values). Logworth values that exceed Max Logworth are plotted as Max Logworth to prevent extreme scales in logworth plots. See [“Example of the Max Logworth Option”](#) for an example.

Force X Categorical Ignores the modeling type and treats all X columns as categorical.

Force X Continuous Ignores the modeling type and treats all X columns as continuous.

Force Y Categorical Ignores the modeling type and treats all Y columns as categorical.

Force Y Continuous Ignores the modeling type and treats all Y columns as continuous.

Unthreaded Suppresses multithreading used for computational speed.

The Response Screening Report

The Response Screening report consists of a section of plots and a section of reports.

- [“Response Screening Plots”](#)
- [“Response Screening Report Tables”](#)

Tip: To change the layout of the plot section or the report section, right-click the top of the desired section and select **Set Style**. The default layout is a tabbed report for up to five plots. When there are more than five plots the default is a combo layout. The combo layout uses a drop down plot selector.

Response Screening Plots

The default plots included in the Response Screening report window depend on the selections made in the launch window. The plots focus on False Discovery Rate (FDR) statistics. See [“Statistical Details for the Response Screening Platform”](#). The following list describes all of the possible plots and when they are available.

FDR PValue Plot Plots the FDR p -values and the standard p -values on the vertical axis and the Rank Fraction on the horizontal axis. This plot indicates how much adjustment is being applied by showing how many of the tests are significant with and without adjustment for the false discovery rate. The Rank Fraction ranks the PValues in order of decreasing significance. FDR PValues are plotted in blue and PValues in red. A blue horizontal line shows the 0.05 significance level. A red increasing line provides an FDR threshold for unadjusted p -values. A p -value falls below the red line precisely when the FDR-adjusted p -value falls below the blue line. This enables you to read significance relative to the FDR from either the adjusted or unadjusted p -values. There is also a legend next to the plot that indicates the number of tests and the portion of significant tests at both the 0.05 and 0.01 alpha levels.

FDR Logworth by Effect Size Plots the FDR logworth on the vertical axis and the effect size on the horizontal axis. This plot helps graphically visualize very small p -values, which are associated with large effects. When transformed to the logworth ($-\log_{10}(p\text{-value})$) scale, highly significant p -values have large logworth values and nonsignificant p -values have low logworth values. A logworth value of zero corresponds to a nonsignificant p -value of

1. Any logworth value above 2 corresponds to a p -value below 0.01. There is a reference line on the plot at logworth = 2 for quick visualization.

Generally, larger effects lead to more significant p -values and larger logworth values. However, this relationship is not necessarily strong because significance also depends on the error variance. In fact, large logworth values can be associated with small effects, and small logworth values can be associated with large effects, because of the size of the error variance. The FDR Logworth by Effect Size plot enables you to explore this relationship.

FDR Logworth by RSquare (Available only for continuous Y variables.) Plots the FDR logworth on the vertical axis and RSquare values on the horizontal axis. Larger logworth values tend to be associated with larger RSquare values, but this relationship also depends on the number of observations.

FDR Logworth by Kappa Plot (Available only if the Kappa option is selected in the launch window and there are X and Y columns that are categorical with the same number of levels.) Plots the FDR logworth on the vertical axis and Kappa on the horizontal axis.

FDR Logworth by Corr Plot (Available only if the Corr option is selected in the launch window and there are categorical X and Y columns.) Plots the FDR logworth on the vertical axis and the correlation on the horizontal axis.

Logworth by Difference (Available by default only if there is at least one continuous response, at least one categorical X variable, and the Common Y Scale option is selected in the launch window.) Plots the logworth on the vertical axis and the mean difference of the horizontal axis. Means differences are calculated for each level versus a control level. The default control level is the first level. If there is only one variable and it has more than two levels, there are check boxes that enable you to select which levels to show on the plot.

If a point is not significant (FDR p -value ≥ 0.01) it is colored gray. If a point is significant, FDR p -value < 0.01 , it is colored red if the mean difference is positive and blue if the mean difference is negative.

Note: This plot is also called a Volcano plot.

Logworth by Slope (Available by default only if there is at least one continuous response, at least one continuous X variable, and both the Common Y Scale and Common X scale options are selected in the launch window.) Plots the FDR logworth difference on the vertical axis and the slope on the horizontal axis. The slopes are calculated from the simple linear regression between each response and each continuous variable.

There is a reference line to indicate significance. The reference line is located where the FDR logworth is 2, which corresponds to an FDR p -value of 0.01. Note that any logworth value above 2 corresponds to a p -value below 0.01. If a point is not significant (FDR p -value ≥ 0.01) it is colored gray. If a point is significant (FDR p -value < 0.01) it is colored red if the slope is positive and blue if the slope is negative.

Note: This plot is also called a Volcano plot.

Practical Difference Plots the maximum of the FDR logworth difference and the FDR logworth equivalence on the vertical axis. The horizontal axis depends on specifications made in the Practical Differences and Equivalences dialog. If a portion is specified, the horizontal axis is the difference divided by the difference to detect. If a specific difference is specified, the horizontal axis is the difference.

Robust FDR PValue Plot (Available only if the Robust option is selected in the launch window.) Plots the robust FDR p -values and the standard p -values on the vertical axis and the Robust Rank Fraction on the horizontal axis. This is similar to the FDR PValue Plot, but uses the robust values instead of the standard values.

Robust FDR Logworth by FDR Logworth Plot (Available only if the Robust option is selected in the launch window.) Plots the robust FDR logworth on the vertical axis and the FDR logworth on the horizontal axis.

Robust FDR Logworth by Effect Size (Available only if the Robust option is selected in the launch window.) Plots the robust FDR logworth on the vertical axis and the effect size on the horizontal axis. This is similar to the FDR Logworth by Effect Size plot, but uses the robust FDR logworth instead of the standard FDR logworth.

Cauchy FDR PValue Plot (Available only if the Cauchy option is selected in the launch window.) Plots the Cauchy FDR p -values and the standard p -values on the vertical axis and the Cauchy Rank Fraction on the horizontal axis. This is similar to the FDR PValue Plot, but uses the Cauchy values instead of the standard values.

Cauchy FDR Logworth by FDR Logworth Plot (Available only if the Cauchy option is selected in the launch window.) Plots the Cauchy FDR logworth on the vertical axis and the FDR logworth on the horizontal axis.

Cauchy FDR Logworth by Effect Size (Available only if the Cauchy option is selected in the launch window.) Plots the Cauchy FDR logworth on the vertical axis and the effect size on the horizontal axis. This is similar to the Cauchy Logworth by Effect Size plot, but uses the Cauchy FDR logworth instead of the standard FDR logworth.

Response Screening Report Tables

The default tables included in the Response Screening report window depend on the selections made in the launch window. The Result Table is always shown and the 2 by M Results table is shown only when there are one or more nominal responses and one or more categorical predictors.

Result Table

The Result Table in the Response Screening report contains a row for each pair of Y and X variables. The columns of the table contain measures and model fit statistics that are specific to the selected fit and Y and X modeling types.

Group (Appears only if there is a grouping variable.) The level of the grouping column.

Y The specified response columns.

X The specified factor columns.

Count The number of rows used for testing, or the corresponding sum of the Freq or Weight variable.

PValue The p -value for the significance test corresponding to the pair of Y and X variables. For more information about Fit Y by X statistics, see *Basic Analysis*.

Logworth The quantity $-\log_{10}(p\text{-value})$. This transformation adjusts p -values to provide an appropriate scale for graphing. A value that exceeds 2 is significant at the 0.01 level ($-\log_{10}(0.01) = 2$).

FDR PValue The False Discovery Rate p -value calculated using the Benjamini-Hochberg technique. This technique adjusts the p -values to control the false discovery rate for multiple tests. If there is no Group variable, the set of multiple tests includes all tests displayed in the table. If there is a Group variable, the set of multiple tests consists of all tests conducted for each level of the Group variable. For more information about the FDR correction, see Benjamini and Hochberg (1995). For more information about the false discovery rate, see [“Statistical Details for the Response Screening Platform”](#).

FDR Logworth The quantity $-\log_{10}(\text{FDR PValue})$. This is the statistic used for plotting and assessing significance. Note that small p -values result in high FDR logworth values. Cells corresponding to FDR logworth values greater than two (p -values less than 0.01) are colored with an intensity gradient.

Effect Size Indicates the extent to which response values differ across the levels or values of X. Effect sizes are scale invariant.

- When Y is continuous, the effect size is the square root of the average sum of squares from the hypothesis test divided by a robust estimate of the response standard deviation. If the interquartile range (IQR) is nonzero and $\text{IQR} > \text{range}/20$, the standard deviation estimate is $\text{IQR}/1.3489795$. Otherwise, the sample standard deviation is used.
- When Y is categorical and X is continuous, the effect size is the square root of the average ChiSquare value for the whole model test.
- When Y and X are both categorical, the effect size is the square root of the average Pearson ChiSquare value.

Rank Fraction The rank of the FDR Logworth expressed as a fraction of the number of tests. If the number of tests is m , the largest FDR Logworth value has Rank Fraction $1/m$, and the smallest has Rank Fraction 1. The Rank Fraction is used in plotting the PValues and FDR PValues in rank order of decreasing significance.

RSquare (Appears only when Y is continuous.) The coefficient of determination, which measures the proportion of total variation explained by the model.

Kappa (Appears only when specified in the launch window. Available only for categorical Y and X with the same number of levels.) A measure of agreement between Y and X.

Corr (Appears only when specified in the launch window.) The Pearson product-moment correlation. For categorical variables, the correlation is calculated in terms of the indices defined by the value ordering.

The following columns are added to the Result Table when the Robust option is selected in the launch window. The Robust option applies only when Y is continuous, so Robust column cells are empty when Y is categorical.

Robust PValue The p -value for the significance test corresponding to the pair of Y and X variables using a robust fit.

Robust Logworth The quantity $-\log_{10}(\text{Robust PValue})$.

Robust FDR PValue The False Discovery Rate calculated for the Robust PValues using the Benjamini-Hochberg technique. If there is no Group variable, the multiple test adjustment applies to all tests displayed in the table. If there is a Group variable, the multiple test adjustment applies to all tests conducted for each level of the Group variable.

Robust FDR Logworth The quantity $-\log_{10}(\text{Robust FDR PValue})$.

Robust Rank Fraction The rank of the Robust FDR Logworth expressed as a fraction of the number of tests.

Robust Chisq The chi-square value associated with the robust test.

Robust Sigma The robust estimate of the error standard deviation.

Robust Outlier Portion The portion of the values whose distance from the robust mean exceeds three times the Robust Sigma.

The following columns are added to the Result Table when the Cauchy option is selected in the launch window. The Cauchy option applies only when Y is continuous, so Cauchy column cells are empty when Y is categorical.

Cauchy PValue The p -value for the significance test corresponding to the pair of Y and X variables using a Cauchy fit.

Cauchy Logworth The quantity $-\log_{10}(\text{Cauchy PValue})$.

- Cauchy FDR PValue** The False Discovery Rate calculated for the Cauchy PValues using the Benjamini-Hochberg technique. If there is no Group variable, the multiple test adjustment applies to all tests displayed in the table. If there is a Group variable, the multiple test adjustment applies to all tests conducted for each level of the Group variable.
- Cauchy FDR Logworth** The quantity $-\log_{10}(\text{Cauchy FDR PValue})$.
- Cauchy Rank Fraction** The rank of the Cauchy FDR Logworth expressed as a fraction of the number of tests.
- Cauchy ChiSquare** The chi-square value associated with the test from the Cauchy fit.
- Cauchy Sigma** The estimate of the error standard deviation from the Cauchy fit.
- Cauchy Outlier Portion** The portion of the values whose distance from the mean of the Cauchy fit exceeds three times the Cauchy Sigma.

2 by M Results

The 2 by M table contains results that you can use to investigate the relationship between a two level categorical response and a categorical predictor. For each response and predictor pair, the relative risk, risk difference, and odds ratio are reported by default. Right-click on the table and select **Columns** to add counts, confidence intervals, labels, or the FDR Logworth values to the table. The first level is set as the baseline and each of the other levels is compared to it.

Note: The 2 by M Results report is available only when there are one or more categorical responses with two levels and one or more categorical predictors.

Means Differences

Each row in the Means Differences table compares a response across two levels of a categorical factor. Use the Means Comparisons option in the launch window to specify whether to compare each level to a control level or to compare all possible level combinations. The table indicates the X variables and the levels that are compared.

The Difference column in the Means Differences table contains the estimated difference in the response means across the two levels. If the Robust option is selected, robust estimates of the means are used. There are also columns for the FDR p-value and the FDR logworth, which are obtained from testing whether or not the difference between the two means is zero.

Note: The Means Differences table is available only when there is one or more continuous responses and one or more categorical predictors.

Response Screening Platform Options

The Response Screening red triangle menu contains options to customize the display and to compute and save calculated data.

Show Plots Shows or hides the plots in the report window.

Show Report Tables Shows or hides the result tables in the report window.

Show Means Differences (Available only if there is at least one continuous response and at least one categorical X variable.) Adds the Logworth by Difference plot and the Means Differences report to the Response Screening report window. See [“Logworth by Difference”](#) and [“Means Differences”](#).

Note: This option is not available if Common Y Scale is selected in the launch window because the plot and table are already included in the report.

Show Slopes Adds the Logworth by Slope plot to the Response Screening report window. See [“Logworth by Slope”](#).

Note: This option is not available if Common Y Scale and Common X Scale are also selected in the launch window because the plot is already included in the report.

Practical Differences and Equivalences Performs tests for practical differences and equivalences of the means across two levels of a categorical variable. Use the Means Comparisons option in the launch window to specify whether to compare each level to a control level or to compare all possible level combinations. Instead of testing that the difference between means is zero, a test of practical significance tests whether the difference in means exceeds a specified practical difference. Equivalence tests enable you to determine whether two levels have essentially the same effect, from a practical perspective, on the response. In other words, an equivalence test tests whether the difference is smaller than the practical difference.

When you select this option, you specify the size of the difference that you consider worth detecting. There are two options to specify the practical difference:

Portion If you assign a Spec Limit property to the Y variable, the practical difference is computed as the difference between the lower and upper specification limits multiplied by the specified portion. If you do not assign a Spec Limit property to the Y variable, the practical difference is computed as $6\hat{\sigma}$ multiplied by the specified portion, where $\hat{\sigma}$ is a robust estimate of the standard deviation. If the interquartile range (IQR) is nonzero and the $IQR > \text{range}/20$, the standard deviation estimate is $\hat{\sigma} = (IQR)/(1.3489795)$. Otherwise, the sample standard deviation is used as $\hat{\sigma}$. By default, the portion is 0.1.

Actual Difference Specifies the practical difference for all comparison tests. This option is appropriate if all of the responses are on the same scale.

Once you click OK, a Practical Difference plot and a Practical Differences table are added to the report window. For information about the plot, see [“Practical Difference”](#). The Practical Differences table indicates the X variables and the levels that are compared. There are columns for the actual difference between levels, the specified practical difference to detect, and the p-values from the practical difference and equivalences tests. There is also a column that indicates whether the levels are different, equivalent, or inconclusive based on the test results.

Fit Selected Items For selected relationships, adds the appropriate Fit Y by X reports to the Response Screening report. You can select relationships by selecting points in the plots, rows in the Result Table, or if opened, rows in the PValue data table.

Select Columns Selects the columns in the original data table that correspond to rows that you select in the Result Table, to points that you select in plots in the Response Screening report window, or if opened, rows in the PValues table. Select the rows or points first, then select Select Columns. The corresponding columns in the data table are selected. You can select columns corresponding to additional rows in the PValues table or points in plots by first selecting them and then selecting Select Columns again. To select columns corresponding to different rows or points, first clear the current column selection in the original data table.

Select Where Opens the Select Where window. You can select specific responses in the Result Table that correspond to a particular condition by using the Comparison menu and Value text box. For example, you can select all responses such that Effect Size > 0.80. After you click OK, the responses are selected in the Result Table.

Tip: You can also access the Select Where window by right-clicking anywhere in the Result Table.

Save Tables Shows a submenu of options to create new data tables.

Save PValues Creates the PValues data table. See [“Response Screening Platform Options”](#).

Save Means (Available for continuous Y variables and categorical X variables.) Creates a data table with the counts, means, and standard deviations for each level of the categorical variable. If the Robust option is selected, the robust mean is included. See [“Means Data Table”](#).

Save Means Differences (Available for continuous Y variables and categorical X variables.) Creates a new data table for tests of comparisons across the levels of the categorical variables. This contains the columns that are in the Means Differences report table as well as additional columns. If practical differences tests have been specified, these test results are included in the new data table. See [“Means Differences Data Table”](#). For an example, see [“Example of Tests of Practical Significance and Equivalence”](#).

Save 2 by M (Available for categorical X variables and categorical Y variables with two levels.) Creates a data table of the information in the 2 by M Results report, as well as other test statistics. The data table contains the following columns groups:

Cell Counts Contains columns of the cell counts that typically appear in a contingency table.

Odds Ratios Contains columns for the estimate, the lower and upper confidence intervals, and the significance status of the odds ratio.

Relative Risks Contains columns for the estimate, the lower and upper confidence intervals, and the significance status of the relative risk.

Risk Differences Contains columns for the estimate, the lower and upper confidence intervals, and the significance status of the risk difference.

Test Statistics Contains several columns of test statistics.

Fisher Exacts Contains several columns of right tailed, left tailed, and two tailed Fisher’s Exact tests.

For more information about the results in the 2 by M table, see *Basic Analysis*.

Save Columns (Available for continuous Y variables.) Shows submenu of options to save new columns to the original data table.

Save Std Residuals Saves a new group of columns to the original data table and places these in a column group call Residual Group. For each continuous Y and categorical X, a column is constructed containing the residuals divided by their estimated standard deviation. In other words, the column contains standardized residuals. The column is defined by a formula.

If the Robust option is selected, standardized residual columns are constructed using robust fits and robust estimates.

Save Outlier Indicator Saves a new group of columns to the original data table and places these in a column group call Outlier Group. Save Outlier Indicator is most effective when you have selected the Robust option.

For each continuous Y and categorical X, a column that indicates outliers is constructed. An outlier is a point whose distance to the predicted value exceeds three times an estimate of sigma. In other words, an outlier is a point whose standardized residual exceeds three. The column is defined by a formula.

If the Robust option is selected, robust fits and robust estimates are used. An outlier is a point whose distance to the predicted value exceeds three times the robust estimate of sigma.

The Cluster Outliers script is added to the original data table. The script shows outliers on a hierarchical cluster plot of the data.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

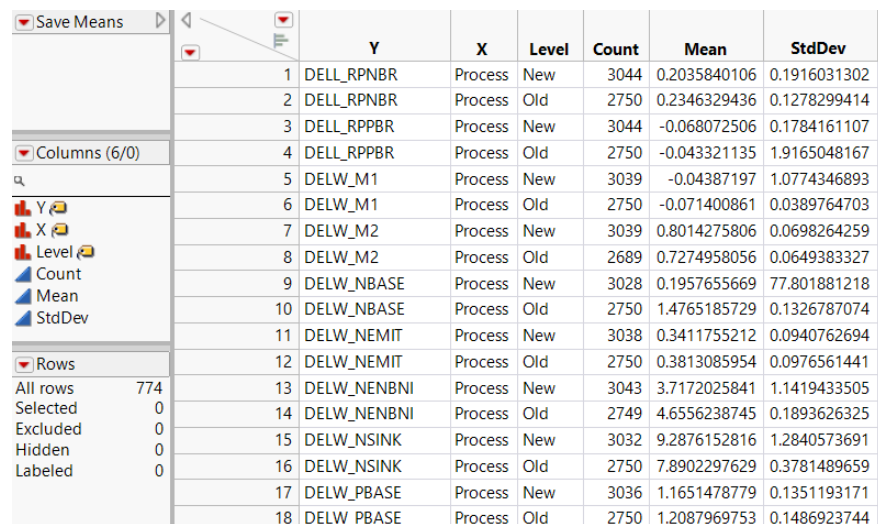
Save Script Contains options that enable you to save a script that reproduces the report to several destinations.

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.

Means Data Table

The Means data table is created after you select the Save Means option in the Response Screening red triangle menu. The data table contains a row for each combination of response and X level. For the Probe.jmp sample data table there are 387 response variables, each tested against Process at two levels. Therefore the Means table contains $387 \times 2 = 774$ rows.

Figure 24.5 Means Data Table



	Y	X	Level	Count	Mean	StdDev
1	DELL_RPNBR	Process	New	3044	0.2035840106	0.1916031302
2	DELL_RPNBR	Process	Old	2750	0.2346329436	0.1278299414
3	DELL_RPPBR	Process	New	3044	-0.068072506	0.1784161107
4	DELL_RPPBR	Process	Old	2750	-0.043321135	1.9165048167
5	DELW_M1	Process	New	3039	-0.04387197	1.0774346893
6	DELW_M1	Process	Old	2750	-0.071400861	0.0389764703
7	DELW_M2	Process	New	3039	0.8014275806	0.0698264259
8	DELW_M2	Process	Old	2689	0.7274958056	0.0649383327
9	DELW_NBASE	Process	New	3028	0.1957655669	77.801881218
10	DELW_NBASE	Process	Old	2750	1.4765185729	0.1326787074
11	DELW_NEMIT	Process	New	3038	0.3411755212	0.0940762694
12	DELW_NEMIT	Process	Old	2750	0.3813085954	0.0976561441
13	DELW_NENBNI	Process	New	3043	3.7172025841	1.1419433505
14	DELW_NENBNI	Process	Old	2749	4.6556238745	0.1893626325
15	DELW_NSINK	Process	New	3032	9.2876152816	1.2840573691
16	DELW_NSINK	Process	Old	2750	7.8902297629	0.3781489659
17	DELW_PBASE	Process	New	3036	1.1651478779	0.1351193171
18	DELW_PBASE	Process	Old	2750	1.2087969753	0.1486923744

Save Means

Columns (6/0)

Y
X
Level
Count
Mean
StdDev

Rows

All rows 774
Selected 0
Excluded 0
Hidden 0
Labeled 0

The Means data table includes the following columns:

Y The continuous response variables.

X The categorical variables.

Level The level of the categorical X variable.

Count The count of values in the corresponding Level.

Mean The mean of the Y variable for the specified Level.

StdDev The standard deviation of the Y variable for the specified Level.

Robust Mean The robust M-estimate of the mean. Appears when you select the Robust option on the launch window.

Means Differences Data Table

The Means Differences data table is created after you select the Save Means Differences option in the Response Screening red triangle menu. The data table provides results for standard difference tests. If the practical differences and equivalences option has been specified, both tests of practical difference and tests of practical equivalence are also provided in the new data table. Each row compares a response across two levels of a categorical factor. Results of the pairwise comparisons are color-coded to facilitate interpretation. See [“Practical Difference”](#) for a description of how the practical difference is specified. See [“Example of Tests of Practical Significance and Equivalence”](#) for an example.

Y The continuous response variables.

X The categorical variables.

Level The level of the categorical X variable.

-Level The level of the categorical X variable being compared to the initial Level.

This Mean The mean of the response variable for the level in the Level column. If the Robust option is selected, robust estimates of the means are used.

Base Mean The mean of the response variable for the level in the -Level column. If the Robust option is selected, robust estimates of the means are used.

Difference The estimated difference in means across the two levels, calculated as This Mean - Base Mean.

Std Err Diff The standard error of the difference in means. This is a robust estimate if the Robust option is selected.

Difference L95 The lower limit of the 95% confidence interval for the difference between the means.

Difference U95 The upper limit of the 95% confidence interval for the difference between the means.

p-Value The p -value for the usual Student's t test for a pairwise comparison. This is the robust version of the t test when the Robust option is selected. Tests that are significant at the 0.05 level are highlighted.

Logworth The quantity $-\log_{10}(p\text{-value})$. This transformation adjusts p -values to provide an appropriate scale for graphing.

FDR p-Value The FDR p -value for the Student's t test for a pairwise comparison. See ["Statistical Details for the Response Screening Platform"](#).

FDR Logworth The quantity $-\log_{10}(\text{FDR } p\text{-Value})$.

Difference to Detect The difference in means that is considered to be of practical interest. This is specified when you select the Practical Differences and Equivalences red triangle menu option. See ["Practical Differences and Equivalences"](#).

Practical Dif PValue The p -value for a test of whether the absolute value of the mean difference in Y between Level and -Level is less than or equal to the Difference to Detect. A small p -value indicates that the absolute difference exceeds the Difference to Detect. This indicates that Level and -Level account for a difference that is of practical consequence.

Practical Equiv PValue Uses the Two One-Sided Tests (TOST) method to test for a practical difference between the means (Schuirmann [1987](#)). The Difference to Detect specifies a

threshold difference for which smaller differences are considered practically equivalent. One-sided t tests are constructed for two null hypotheses: the true difference exceeds the Difference to Detect; the true difference is less than the negative of the Difference to Detect. If both tests reject, this indicates that the absolute difference in the means falls within the Difference to Detect. Therefore, the groups are considered practically equivalent.

The Practical Equivalence PValue is the largest p -value obtained on the one-sided t tests. A small Practical Equiv PValue indicates that the mean response for Level is equivalent, in a practical sense, to the mean for -Level.

FDR Practical Dif PValue The FDR p -value for the practical difference test for a pairwise comparison. See [“Statistical Details for the Response Screening Platform”](#).

FDR Practical Equiv PValue The FDR p -value for the TOST test for a pairwise comparison. See [“Statistical Details for the Response Screening Platform”](#).

Practical Result A description of the results of the tests for practical difference and equivalence. Values are color-coded to help identify significant results.

- Different (Pink): Indicates that the absolute difference is significantly greater than the practical difference.
- Equivalent (Green): Indicates that the absolute difference is significantly within the practical difference.
- Inconclusive (Gray): Indicates that neither the test for practical difference nor the test for practical equivalence is significant.

The PValues Data Table

If you select the PValues Tables on Launch option in the Response Screening launch window, a PValues data table is created. This data table contains the same information as the Result Table that is in the Response Screening report. The PValues data table also contains the additional, model specific columns:

YMean The mean of Y.

SSE Appears when Y is continuous. The sum of squares for error.

DFE Appears when Y is continuous. The degrees of freedom for error.

MSE Appears when Y is continuous. The mean squared error.

F Ratio Appears when Y is continuous. The F Ratio for the analysis of variance or regression test.

Intercept Appears when Y and X are both continuous. The intercept of the regression model relating the corresponding pair of Y and X variables.

Slope Appears when Y and X are both continuous. The slope of the regression model relating the corresponding pair of Y and X variables.

DF Appears when Y and X are both categorical. The degrees of freedom for the ChiSquare test.

LR Chisq Appears when Y and X are both categorical. The value of the Likelihood Ratio ChiSquare statistic.

Robust Elapsed Time (Appears only when Robust is specified in the launch window.) Computing time, in seconds, required to create the Robust report.

PValues Data Table Scripts

Relevant scripts are saved to the Response Screening PValues data table. All but one of these reproduce plots provided in the report. When you select rows in the PValues table, the Fit Selected Items script produces the appropriate Fit Y by X analyses.

The Response Screening Personality in Fit Model

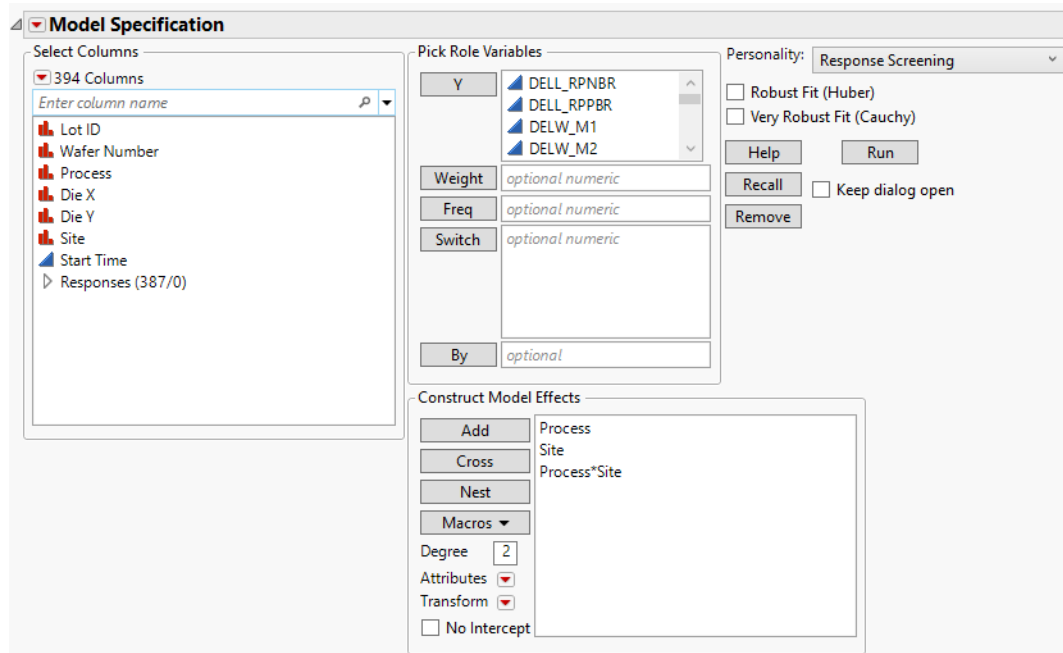
If you are interested in univariate tests against linear model effects, you can fit the Response Screening personality in the Fit Model platform. The report and tables produced test all responses against all model effects. An example using the Response Screening personality is given in [“Example of the Response Screening Personality”](#).

- [“Launch Response Screening in Fit Model”](#)
- [“The Fit Response Screening Report”](#)
- [“Fit Response Screening Options”](#)

Launch Response Screening in Fit Model

Launch Fit Response Screening by selecting **Analyze > Fit Model** and selecting **Response Screening** from the Personality list.

Figure 24.6 Response Screening from the Fit Model Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

The following column role is unique to the Response Screening personality:

Switch Specifies columns to be switched, one at a time, into the model.

There are two options for robust estimation:

Robust Fit (Huber) Specifies robust (Huber) estimation to down weight outliers for continuous responses. If there are no outliers, these estimates are close to the least squares estimates. Note that this option increases processing time.

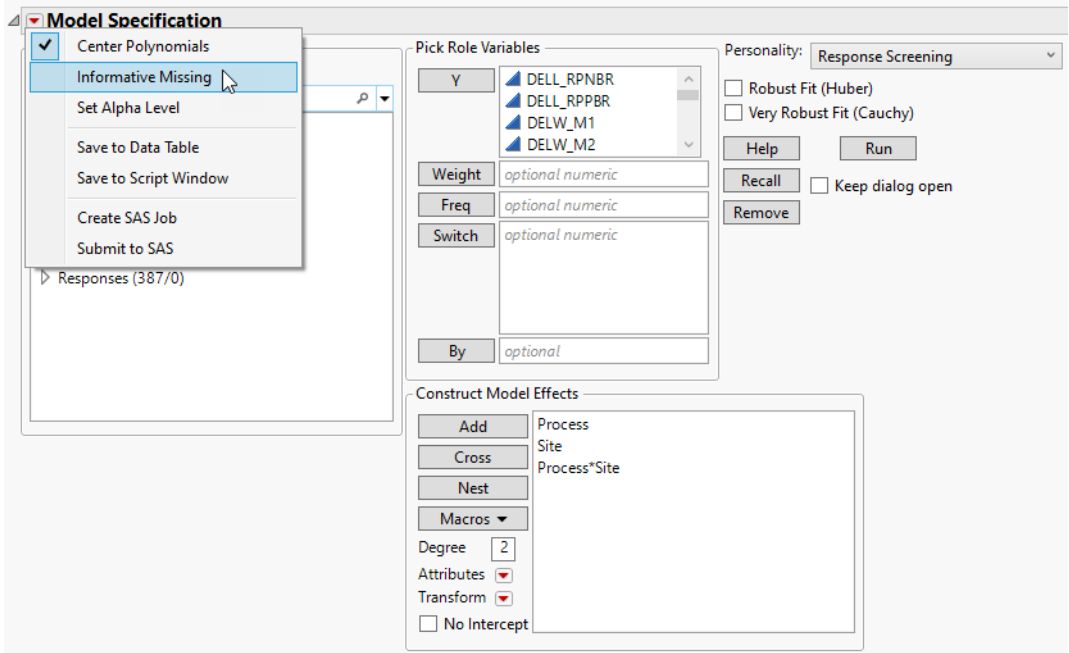
Very Robust Fit (Cauchy) Specifies very robust (Cauchy) estimation to down weight outliers for continuous responses. Assumes that the errors have a Cauchy distribution. A Cauchy distribution has fatter tails than the normal distribution, resulting in a reduced emphasis on outliers. This option can be useful if you have a large proportion of outliers in your data. However, if your data are close to normal with only a few outliers, this option can

lead to incorrect inferences. The Cauchy option estimates parameters using maximum likelihood and a Cauchy link function.

Tip: If both robust options are selected, the platform uses only Cauchy estimation.

The **Informative Missing** option provides a coding system for missing values. The Informative Missing coding allows estimation of a predictive model despite the presence of missing values. It is useful in situations where missing data are informative. Select this option from the Model Specification red triangle menu.

Figure 24.7 Informative Missing Option



For more information about the Fit Model window, see *Fitting Linear Models*.

The Fit Response Screening Report

When you launch Response Screening through the Fit Model platform, the Fit Response Screening report contains an Effect Tests table and two plots. The plots shown are the FDR PValue Plot for Effects and the FDR Logworth by Effect Size plot. Both plots are interpreted in the same way as for the platform itself. See [“Response Screening Plots”](#).

Effect Tests Table

The Effect Tests table contains a row for each pair consisting of a Y variable and a model Effect. If you select the Robust Fit option on the launch window, the models are fit using Huber M-estimation. If you specified a By variable, an Effect Tests table is created for each level of the By variable. The table includes the following columns:

Y The specified response columns.

Switch (Available when Switch columns are specified in the launch.) Specifies the column that is included in the model. This identifies which model the row is describing. When the Effect column contains “Switch”, the Switch column specifies the effect.

Tip: Click on the Switch column header to sort the Effect Tests table by Switch. This enables you to more clearly see the effects within each model tested. For multiple Ys, sort on Switch and then on Y to arrange the effects within models for each response.

Effect The specified model effects.

FRatio The test statistic for a test of the Effect. This is the value found in the Effect Tests report in Least Squares Fit.

Chi Square (Available only when a robust fit is specified in the launch.) The test statistic for a test of the Effect.

PValue The p -value for the significance test corresponding to the FRatio. See *Fitting Linear Models* for more information about Effect Tests.

Logworth The quantity $-\log_{10}(p\text{-value})$. 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 PValue The False Discovery Rate p -value calculated using the Benjamini-Hochberg technique. This technique adjusts the p -values to control the false discovery rate for multiple tests. For more information about the FDR correction, see Benjamini and Hochberg (1995). For more information about the false discovery rate, see [“Statistical Details for the Response Screening Platform”](#) or Westfall et al. (2011).

FDR Logworth The quantity $-\log_{10}(\text{FDR PValue})$. This is the best statistic for plotting and assessing significance. Note that small p -values result in high FDR logworth values.

Rank Fraction (Not shown by default.) The rank of the FDR Logworth expressed as a fraction of the number of tests. If the number of tests is m , the largest FDR logworth value has Rank Fraction $1/m$, and the smallest has Rank Fraction 1. Equivalently, the Rank Fraction ranks the p -values in increasing order, as a fraction of the number of tests. The Rank Fraction is used in plotting the PValues and FDR PValues in rank order of decreasing significance.

Effect Size Indicates the extent to which response values differ across the levels or values of X . Effect sizes are scale invariant.

- When Y is continuous, the effect size is the square root of the average sum of squares from the hypothesis test divided by a robust estimate of the response standard deviation. If the interquartile range (IQR) is nonzero and $IQR > range/20$, the standard deviation estimate is $IQR/1.3489795$. Otherwise the sample standard deviation is used.
- When Y is categorical and X is continuous, the effect size is the square root of the average chi-Square value for the whole model test.
- When Y and X are both categorical, the effect size is the square root of the average Pearson ChiSquare.
- When a robust fit is specified in the launch, the effect size is calculated as the square root of $(\text{chi-square test statistic}/n)$, where n is the number of observations.

Test DF The degrees of freedom for the effect test.

Fit Response Screening Options

The Fit Response Screening red triangle menu contains options to save the calculated data.

Effect Tests Shows or hides the Effect Tests table. See [“Effect Tests Table”](#).

Effect Plots Shows or hides the FDR PValue Plot for Effects and the FDR Logworth by Effect Size plot.

Overall Report Shows or hides the Overall Fit table, which contains a row for each Y variable. For each Y , the columns in the table summarize information about the model fit. If you select the Robust Fit option on the launch window, the models are fit using Huber M-estimation. The table contains the following columns:

Y The specified response columns.

RSquare The multiple correlation coefficient.

RMSE The Root Mean Square Error.

Count The number of observations (or sum of the Weight variable).

Overall FRatio The test statistic for model fit from the Analysis of Variance report in Least Squares Fit.

Overall PValue The p -value for the overall test of model significance.

Overall Logworth The logworth of the p -value for the overall test of model significance.

Overall FDR PValue The overall p -value adjusted for the false discovery rate. (See [“The Response Screening Report”](#).)

Overall FDR Logworth The logworth of the Overall FDR PValue.

Overall Rank Fraction (Not shown by default.) The rank of the Overall FDR Logworth expressed as a fraction of the number of tests. If the number of tests is m , the largest Overall FDR Logworth value has Rank Fraction $1/m$, and the smallest has Rank Fraction 1.

Overall Plots Shows or hides the Overall FDR PValue Plot and the FDR Logworth by RSquare plot.

Select Effects Where Opens the Select Where window. You can select specific responses in the Effects Tests table that correspond to a particular condition by using the Comparison menu and Value text box. For example, you can select all effects such that Effect Size > 0.80. After you click OK, the responses are selected in the Result Table.

Tip: You can also access the Select Where window by right-clicking anywhere in the Effect Tests table.

Select Responses for Selected Effects Selects response columns in the original data table that correspond to the selected effects in the Effect Tests table.

Save Effect Tests Saves the Effect Tests table to a new data table. The Effect Tests data table contains a table variable called Original Data that gives the name of the data table that was used for the analysis. If you specified a By variable the Original Data variable also gives the By variable and its level.

Save Overall Fit Saves the Overall Fit table to a new data table. The Overall Fit data table contains a table variable called Original Data that gives the name of the data table that was used for the analysis. If you specified a By variable the Original Data variable also gives the By variable and its level.

Save Estimates Opens a data table in which each row corresponds to a response and the columns correspond to the model terms. The entries are the parameter estimates obtained by fitting the specified model. This data table also contains a table variable called Original Data that gives the name of the data table that was used for the analysis. If you specified a By variable, JMP creates an estimates table for each level of the By variable, and the Original Data variable gives the By variable and its level.

Save Prediction Formula Adds columns to the original data table containing prediction equations for all responses.

Save Least Squares Means Opens a data table where each row corresponds to a response and a combination of effect settings. The row contains the least squares mean and standard error for that combination of settings.

Model Dialog Opens a window containing the model dialog that you have run to obtain the given report.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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

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

Additional Examples of Response Screening

The section contains examples using the Response Screening platform.

- [“Example of Tests of Practical Significance and Equivalence”](#)
- [“Example of the Max Logworth Option”](#)
- [“Example of Robust Fit”](#)
- [“Example of the Response Screening Personality”](#)
- [“Example of a Volcano Plot”](#)

Example of Tests of Practical Significance and Equivalence

Use the Response Screening platform to test for practical differences.

1. Select **Help > Sample Data Folder** and open Probe.jmp.
2. Select **Analyze > Screening > Response Screening**.
3. Select the Responses column group and click **Y, Response**.
4. Select Process and click **X**.
5. Click **OK**.

6. Click the Response Screening red triangle and select **Practical Differences and Equivalences**.
7. Type 0.15 in the box next to Enter a portion of Spec Range or $6 \times \text{Sigma}$.
8. Click **OK**.
9. Click the Response Screening red triangle and select **Save Tables > Save Means Differences**.

Figure 24.8 shows a portion of the data table. For each response in Y, the corresponding row gives information about tests of the New and the Old levels of Process.

Figure 24.8 Means Differences Table, Partial View

	Y	FDR PValue	FDR Logworth	Difference to detect	Practical Dif PValue	Practical Equiv PValue	FDR Practical Dif PValue	FDR Practical Equiv PValue	Practical Result
1	DELL_RPNBR	2.39467e-12	11.620754237	0.0990979657	1	6.407405e-55	1	9.611108e-55	Equivalent
2	DELL_RPPBR	0.5449559037	0.2636386382	0.7959556871	1	3.07185e-104	1	2.35918e-103	Equivalent
3	DELW_M1	0.2459236742	0.6091996613	0.4687012219	1	1.160101e-98	1	4.283449e-98	Equivalent
4	DELW_M2	0	325.14036157	0.0469881911	1.350706e-50	1	1.440753e-49	1	Different
5	DELW_NBASE	0.4570716144	0.3400157489	33.792861061	1	1.28251e-102	1	7.53792e-102	Equivalent
6	DELW_NEMIT	4.065967e-55	54.390836173	0.0565587625	1	3.954083e-11	1	4.804962e-11	Equivalent
7	DELW_NENBNI	0	342.83731846	0.5760161527	1.142115e-59	1	1.289918e-58	1	Different
8	DELW_NSINK	0	528.07666335	0.7148733981	6.27067e-150	1	8.02645e-149	1	Different
9	DELW_PBASE	1.227409e-30	29.911010679	0.0864253619	1	2.094533e-30	1	2.77345e-30	Equivalent
10	DELW_PCOLL	0	2733.3647756	0.78359742	0	1	0	1	Different
11	DELW_PEMIT	0.7257260746	0.1392272729	1.2176717085	1	1.10189e-106	1	1.51116e-105	Equivalent
12	DELW_PSINK	7.709735e-40	39.112960512	1.5705731641	1	1.83559e-22	1	2.381305e-22	Equivalent
13	DELW_RPNBR	3.391334e-23	22.469629432	0.2309922348	1	7.505304e-38	1	1.040446e-37	Equivalent
14	DELW_RPPBR	4.583412e-26	25.338811147	0.6465385194	1	1.402789e-34	1	1.896729e-34	Equivalent
15	DELW_SICR	0.1667193458	0.7780140026	0.4016073762	1	1.042102e-96	1	3.51024e-96	Equivalent
16	M1_COMB_VG...	0.4548355067	0.3421456395	13.77484916	1	8.16456e-103	1	4.97649e-102	Equivalent
17	M1_TRENCH_V...	0.0005193703	3.2845228853	3.872532637	1	4.896901e-80	1	1.062379e-79	Equivalent
18	M2/M1_CAP_V...	0.6866587979	0.1632590111	2.2331316201	1	2.36705e-106	1	2.93209e-105	Equivalent
19	M2_COMB_BB...	0.999486039	0.0002232678	2.8980258306	1	1.08191e-110	1	6.5754e-109	Equivalent
20	M2_COMB_VG...	0.3827958663	0.4170327607	6.527621621	1	1.91844e-101	1	6.56606e-101	Equivalent
21	NISO_TUB-TRE...	1.38575e-163	162.85831566	6.0414047752	0.0000469484	0.9999530516	0.000400626	1	Different
22	NISO_TUB-TUB...	0.5509658664	0.2588753059	2.4675993783	1	1.70833e-104	1	1.49091e-103	Equivalent
23	PS_RPNBR	0.0000493267	4.3069182395	353.51614685	1	1.545634e-77	1	3.157041e-77	Equivalent

Because specification limits are not saved as column properties in Probe.jmp, JMP calculates a value of the practical difference for each response. The practical difference portion of 0.15 that you specified is multiplied by an estimate of the 6σ range of the response. This value is used in testing for practical difference and equivalence. It is shown in the Difference to detect column.

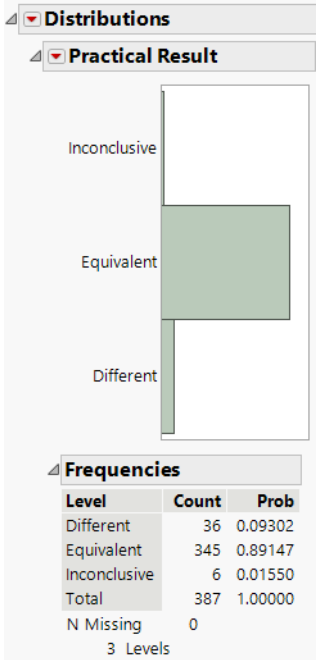
The FDR PValue column shows responses whose p -values indicate significance. The FDR Practical Diff PValue and FDR Practical Equiv PValue columns give the p -values for tests of practical difference and practical equivalence. Note that many columns show *statistically* significant differences, but do not show *practically* significant differences.

10. Display the Means Differences data table and select **Analyze > Distribution**.
11. Select Practical Result and click **Y, Columns**.
12. Click **OK**.

Figure 24.9 shows the distribution of results for practical significance. Only 36 tests are different, as determined by testing for the specified practical difference. For 6 of the

responses, the tests were inconclusive. You cannot tell whether the responses result in a practical difference across Process.

Figure 24.9 Distribution of Practical Significance Results



The 36 responses can be selected for further study by clicking on the corresponding bar in the plot.

Example of the Max Logworth Option

In this example, use the Max Logworth option in the Response Screening platform to control the logworth scale from being distorted by very large values. When data sets have a large number of observations, *p*-values can be very small. Logworth values provide a useful way to study *p*-values graphically in these cases. In some data sets the *p*-values are so small that the logworth scale is distorted by very large values.

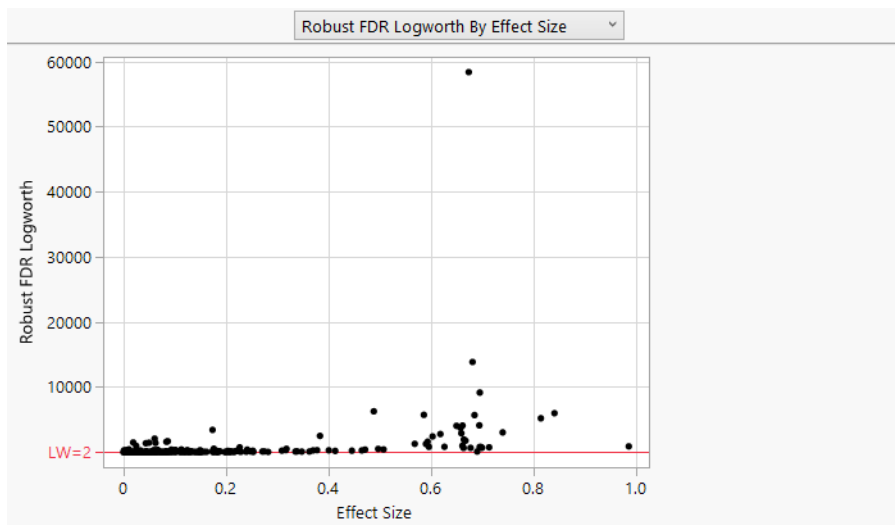
1. Select **Help > Sample Data Folder** and open Probe.jmp.
2. Select **Analyze > Screening > Response Screening**.
3. In the Response Screening Launch window, select the Responses column group and click **Y, Response**.
4. Select Process and click **X**.
5. Click the gray triangle next to Advanced Options.

6. Select the **Robust** check box.
7. Click **OK**.

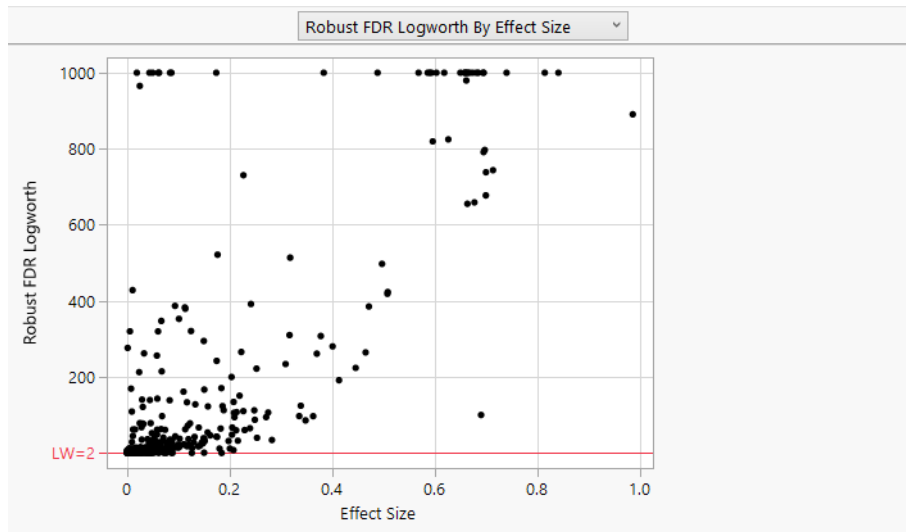
The analysis is numerically intensive and might take some time to complete.

8. Click the Response Screening plot title list and select **Robust FDR Logworth by Effect Size**.
The detail in the plot is difficult to see, because of the one large Robust FDR Logworth value of about 58,000. To ensure that your graphs show sufficient detail, you can set a maximum value of the logworth.

Figure 24.10 Robust FDR Logworth vs. Effect Size, Max Logworth Not Set



9. Click the Response Screening red triangle and select **Redo > Relaunch Analysis**.
10. Click the gray triangle next to Advanced Options.
11. Type 1000 in the box next to Max Logworth.
12. Click **OK**.
The analysis might take some time to complete.
13. Click the Response Screening plot title list and select **Robust FDR Logworth by Effect Size**.
Now the detail in the plot is apparent. Robust FDR Logworth values greater than 1000 are plotted at 1000.

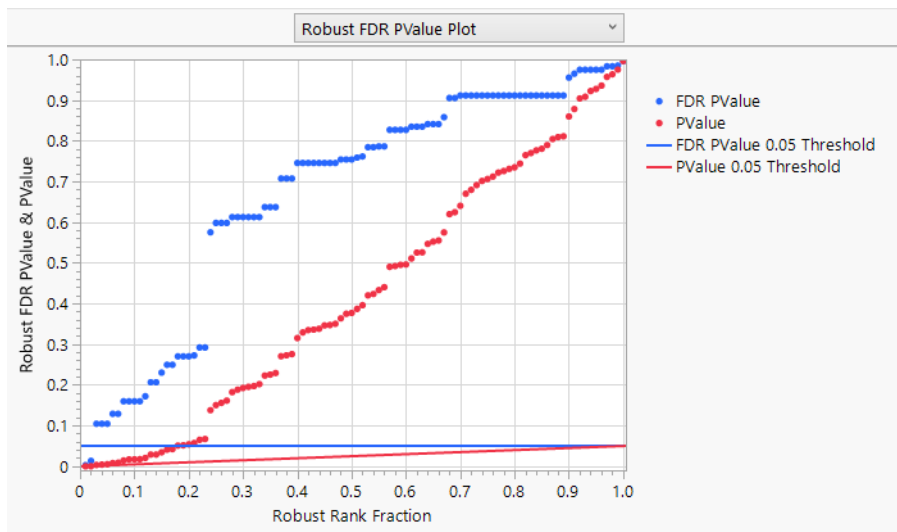
Figure 24.11 Robust FDR Logworth vs. Effect Size, Max Logworth = 1000

Example of Robust Fit

Use the Robust option in the Response Screening platform to analyze responses with a lot of outliers. It is important to use a robust fit because outliers can inflate the error variance of a traditional test, which makes it more difficult to see significant effects.

1. Select **Help > Sample Data Folder** and open *Drosophila Aging.jmp*.
2. Select **Analyze > Screening > Response Screening**.
3. Select all of the continuous columns and click **Y, Response**.
4. Select line and click **X**.
5. Click the gray triangle next to Advanced Options.
6. Check **Robust**.
7. Click **OK**.

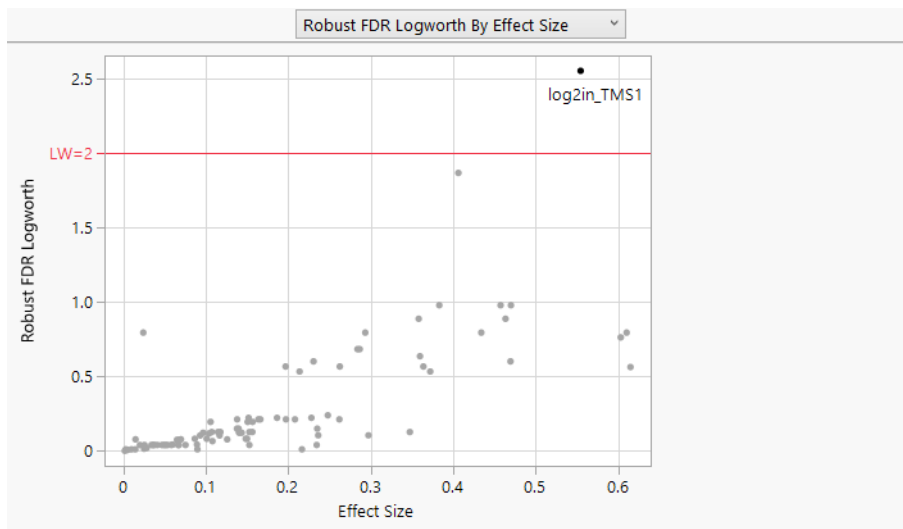
Figure 24.12 Robust FDR PValue Plot for Drosophila Data



Note that a number of tests are significant using the unadjusted robust p -values, as indicated by the red points that are less than 0.05. However, only two tests are significant according to the robust FDR p -values. These two points are more easily identified in a plot that shows FDR logworth values.

8. Click Response Screening plot title list and select **Robust FDR Logworth by Effect Size**.
9. Click the point with a Robust FDR Logworth value that exceeds 2, right click and select **Row Label**.

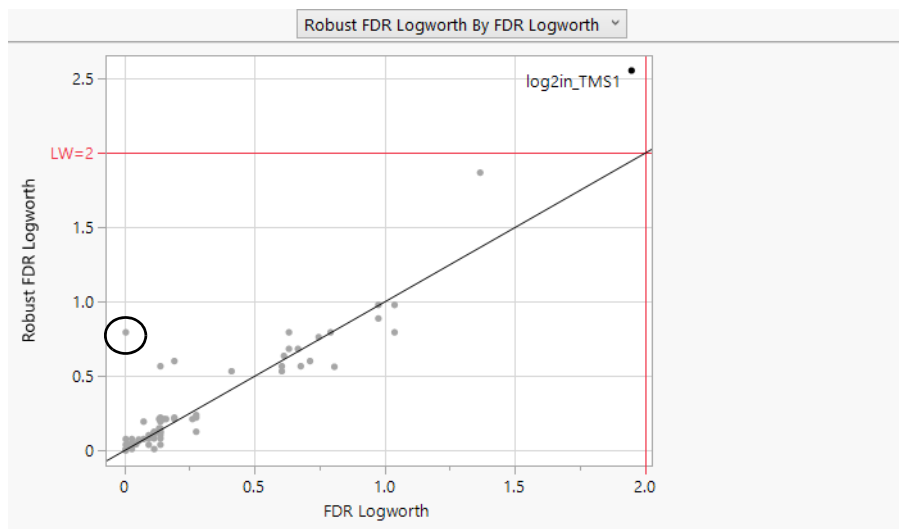
Figure 24.13 Robust Logworth by Effect Size for Drosophila Data



Points above the red line at 2 have significance levels below 0.01 ($-\log_{10}(0.01) = 2$). A horizontal line at about 1.3 corresponds to a 0.05 significance level.

10. Click the Response Screening plot title list and select **Robust FDR Logworth by FDR Logworth**.

Figure 24.14 Robust Logworth by Logworth for Drosophila Data



If the robust test for a response were identical to the usual test, its corresponding point would fall on the diagonal line in Figure 24.14. The circled point in the plot does not fall near the line, because it has a robust logworth value that exceeds its logworth value.

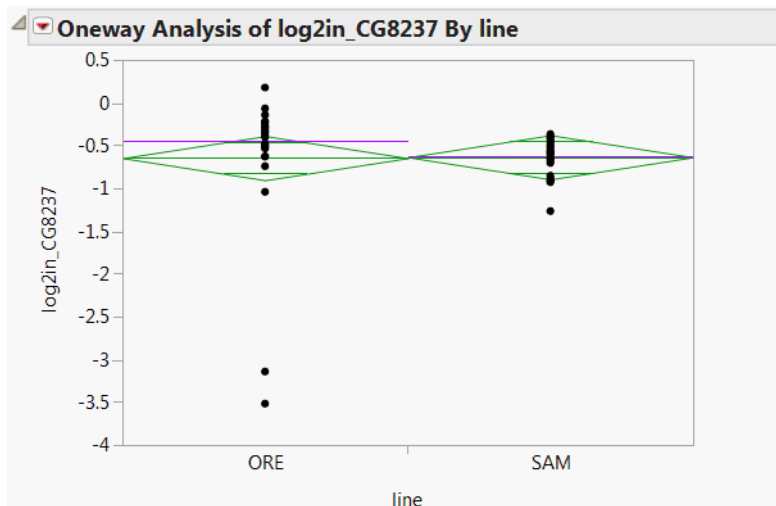
11. Click to select this point in the plot.
12. Find the row for this point in the Result Table.

Note that the response, log2in_CG8237 has PValue 0.9568 and Robust PValue 0.0176.

13. Click the Response Screening red triangle and select **Fit Selected Items**.

A Fit Selected Items report is displayed containing a Oneway Analysis for the response log2in_CG8237. The plot shows two outliers for the ORE line (Figure 24.15). These outliers indicate why the robust test and the usual test give disparate results. The outliers inflate the error variance for the non-robust test, which makes it more difficult to see a significant effect. In contrast, the robust fit down-weights these outliers, thereby reducing their contribution to the error variance.

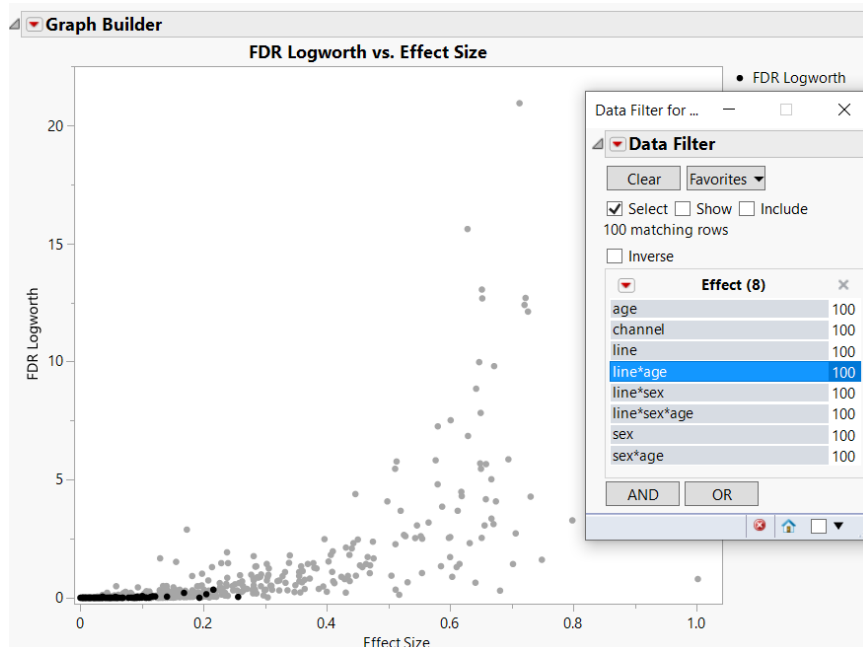
Figure 24.15 Oneway Analysis for log2in_CG8237



Example of the Response Screening Personality

This example uses the Response Screening personality in Fit Model to study tests of multiple responses against linear model effects.

1. Select **Help > Sample Data Folder** and open Drosophila Aging.jmp.
2. Select **Analyze > Fit Model**.
3. Select all the continuous columns and click **Y**.
4. Select channel and click **Add**.
5. Select line, sex, and age and select **Macros > Full Factorial**.
6. Select **Response Screening** from the Personality list.
7. Click **Run**.
8. Click the Fit Response Screening red triangle and select Save Effect Tests.
9. Run the FDR Logworth by Effect Size script in the Effect Tests data table.
10. Select **Rows > Data Filter**.
11. In the Data Filter window, select Effect and click .
12. In the Data Filter, click each of the model effects sequentially while you view the selected points in the FDR Logworth vs. Effect Size plot.

Figure 24.16 FDR Logworth vs. Rank Fraction Plot with line*age Tests Selected


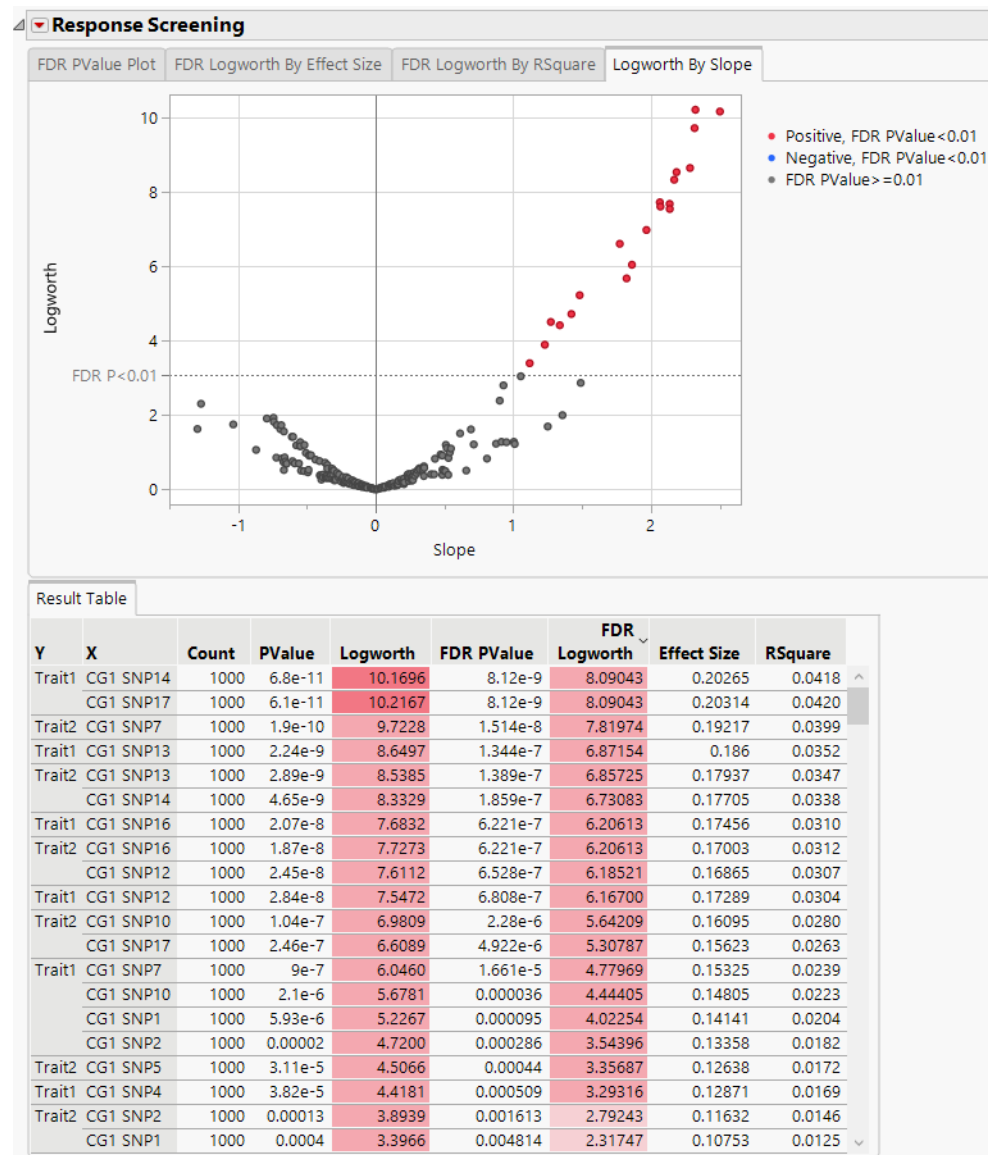
Keep in mind that logworth values that exceed 2 are significant at the 0.01 level. The Data Filter helps you see that, with the exception of *sex* and *channel*, the model effects are rarely significant at the 0.01 level. The points for tests of the *line*age* interaction effect are selected. None of these are significant at the 0.01 level.

Example of a Volcano Plot

Use the Response Screening platform to determine which SNPs are associated with four different traits. A SNP is a type of genetic variant. Volcano plots are helpful in visualizing the results from a large number of hypothesis tests that are of the same type. In this case, all of the responses and variables are continuous, so all of the tests are from a simple linear regression.

1. Select **Help > Sample Data Folder > Life Sciences** and open *Genotypes Pedigree.jmp*.
2. Select **Analyze > Screening > Response Screening**.
3. Select Trait1 through Trait4 and click **Y, Response**.
4. Select the Markers column group and click **X**.
5. Select **Common Y Scale** and **Common X Scale**.
6. Click **OK**.
7. Select the Logworth By Slope tab.

Figure 24.17 Logworth By Slope Plot



The colored points in the Logworth By Slope plot, also known as a volcano plot, indicate tests that are significant at the 0.01 level while controlling for the false discovery rate. In this example, there are only red points in the plot. This indicates that all of the significant X variables have positive slopes in the simple linear regressions. Therefore, for all SNPs that are significantly associated with a trait, a higher SNP value leads to a higher Trait value. The Result Table shows the results from the individual tests, with the most significant tests at the top. All of the top tests involve Trait 1 and Trait 2. You can conclude that this set of SNPs have higher association with Trait 1 and Trait 2 than the other two traits.

Statistical Details for the Response Screening Platform

All of the Response Screening plots involve p -values for tests conducted using the False Discovery Rate technique described in Benjamini and Hochberg (1995). See also Westfall et al. (2011). This method assumes that the p -values are independent and uniformly distributed.

The Response Screening platform uses the following procedure to control the false discovery rate at level α :

1. Conduct the m hypothesis tests of interest to obtain p -values $p_1, p_2, \dots, p_m$.
2. Rank the p -values from smallest to largest. Denote these by $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$.
3. Find the largest p -value for which $p_{(i)} \leq (i/m)\alpha$. Suppose this first p -value is the k^{th} largest, $p_{(k)}$.
4. Reject the k hypotheses associated with p -values less than or equal to $p_{(k)}$.

This procedure ensures that the false discovery rate does not exceed α .

The p -values adjusted for the false discovery rate, denoted $p_{(i), FDR}$, are computed as follows:

$$p_{(i), FDR} = \begin{cases} p_{(m)} & \text{for } i = m \\ \min \left[p_{(i+1), FDR}, \left(\frac{m}{i} \right) p_{(i)} \right] & \text{for } i = m-1, \dots, 1 \end{cases}$$

If a hypothesis has an FDR-adjusted p -value that falls below α , then it is rejected by the procedure.

Chapter **25**

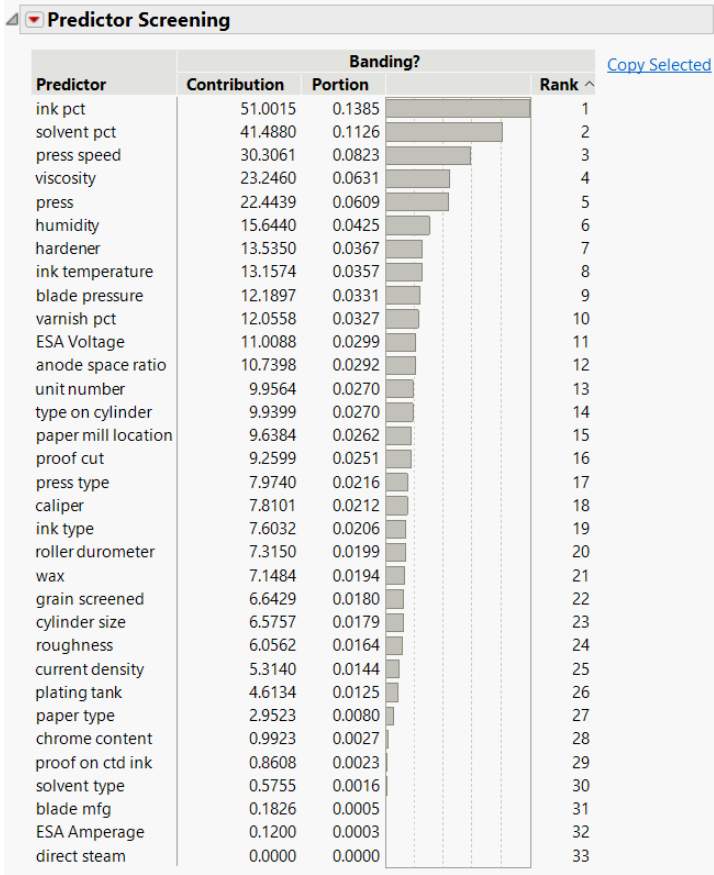
Predictor Screening

Screen Many Predictors for Significant Effects

The Predictor Screening platform provides a method of screening many predictors for their ability to predict an outcome. This is useful in the analysis of large data sets, where hundreds to thousands of measurements on a part, process, or sample are taken requires innovative approaches. For example, predictor screening can be used to help identify biomarkers from thousands tested in samples from patients with and without a condition to predict the condition.

Predictor screening differs from response screening. Response screening tests factors one at a time as a predictor of the response. Predictor screening uses bootstrap forest partitioning to evaluate the contribution of predictors on the response. The partition models are built on multiple predictors. Predictor screening can identify predictors that might be weak alone but strong when used in combination with other predictors. See [“Response Screening”](#).

Figure 25.1 Example of a Predictor Screening Report



Contents

Overview of the Predictor Screening Platform	530
Example of Predictor Screening.....	530
Launch the Predictor Screening Platform.....	532
The Predictor Screening Report.....	533
Predictor Screening Platform Options	533

Overview of the Predictor Screening Platform

Predictor screening is useful for identifying significant predictors from a large number of candidates. Suppose you had hundreds of Xs and needed to determine which of those were most significant as predictors of an outcome.

The predictor screening platform uses a bootstrap forest to identify potential predictors of your response. For each response, a bootstrap forest model using 100 decision trees is built. The column contributions to the bootstrap forest model for each predictor are ranked. Because the bootstrap forest method involves a random component, column contributions can differ when you rerun the report. For more information about decision trees, see [“Partition Models”](#).

Example of Predictor Screening

You are interested in understanding what factors, or features, are most likely to predict a response. Consider a data table containing measurements from machinery in a rotogravure printing business. The data set contains 539 records and 38 variables.

1. Select **Help > Sample Data Folder** and open Bands Data.jmp.
2. Select **Analyze > Screening > Predictor Screening**.
3. Select Banding? as **Y, Response**.
4. Select the grouped columns grain screened to chrome content and click **X**.
5. Click **OK**.

Figure 25.2 Ranked Column Contributions

Predictor Screening				
Predictor	Banding?			Rank ^
	Contribution	Portion		
ink pct	51.0015	0.1385		1
solvent pct	41.4880	0.1126		2
press speed	30.3061	0.0823		3
viscosity	23.2460	0.0631		4
press	22.4439	0.0609		5
humidity	15.6440	0.0425		6
hardener	13.5350	0.0367		7
ink temperature	13.1574	0.0357		8
blade pressure	12.1897	0.0331		9
varnish pct	12.0558	0.0327		10
ESA Voltage	11.0088	0.0299		11
anode space ratio	10.7398	0.0292		12
unit number	9.9564	0.0270		13
type on cylinder	9.9399	0.0270		14
paper mill location	9.6384	0.0262		15
proof cut	9.2599	0.0251		16
press type	7.9740	0.0216		17
caliper	7.8101	0.0212		18
ink type	7.6032	0.0206		19
roller durometer	7.3150	0.0199		20
wax	7.1484	0.0194		21
grain screened	6.6429	0.0180		22
cylinder size	6.5757	0.0179		23
roughness	6.0562	0.0164		24
current density	5.3140	0.0144		25
plating tank	4.6134	0.0125		26
paper type	2.9523	0.0080		27
chrome content	0.9923	0.0027		28
proof on ctd ink	0.8608	0.0023		29
solvent type	0.5755	0.0016		30
blade mfg	0.1826	0.0005		31
ESA Amperage	0.1200	0.0003		32
direct steam	0.0000	0.0000		33

Note: Because this analysis is based on the Bootstrap Forest method that has a random selection component, your results can differ slightly from those in Figure 25.2. See “Bootstrap Forest”.

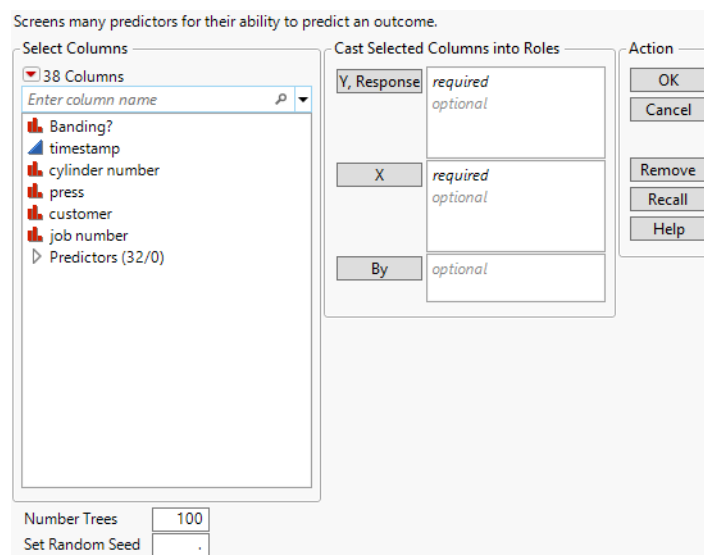
The columns are sorted and ranked in order of contribution in the bootstrap forest model. Predictors with the highest contributions are strong candidate predictors for the response of interest.

Tip: Click on a column heading to change how the columns are sorted in the table.

Launch the Predictor Screening Platform

Launch the Predictor Screening platform by selecting **Analyze > Screening > Predictor Screening**.

Figure 25.3 Predictor Screening Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Y, Response The response columns.

X Predictor columns.

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.

Number of Trees The number of decision trees in the bootstrap forest model.

Set Random Seed Sets the seed for the starting values used in the fitting procedures. Set Random Seed is useful if you want to reproduce an analysis. If you set a random seed and save the script, the seed is automatically saved in the script.

The Predictor Screening Report

The Predictor Screening report shows the list of predictors with their respective contributions and rank. Predictors with the highest contributions are likely to be important in predicting Y.

The Contribution column shows the contribution of each predictor to the bootstrap forest model. The Portion column in the report shows the percent contribution of each variable.

You can select the important predictors in the Predictor Screening report. Selecting the important predictors selects the corresponding columns in the data table, enabling you to easily enter these columns into the launch windows for modeling platforms. The Copy Selected button enables you to copy the selected columns to the clipboard so that you can paste them into a model as well. In this way, the Predictor Screening enhances the modeling process.

Predictor Screening Platform Options

The Predictor Screening red triangle menu has the following options:

Save Results Saves the predictor ranking table from the report to a new data table.

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.

Chapter 26

JMP[®] PRO Association Analysis

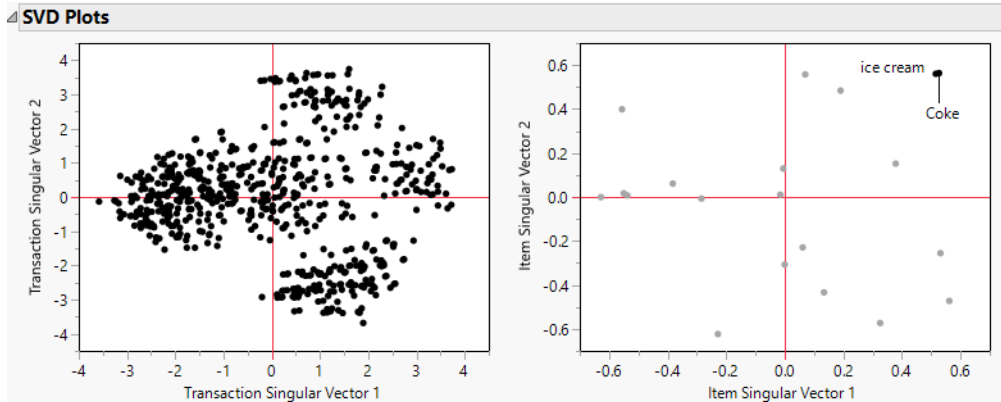
Perform Market Basket Analysis

The Association Analysis platform is available only in JMP Pro.

Association analysis enables you to identify items that have an affinity for each other. It is frequently used to analyze transactional data (also called *market baskets*) to identify items that often appear together in transactions. For example, grocery stores and online merchants use association analysis to strategically organize and recommend products that tend to be purchased together.

Association analysis is also used for identifying dependent or associated events. For example, you can identify car parts that seem to fail around the same time. In this application, car inspections are treated as the market baskets and you analyze the associations among groups of faulty parts found in each inspection.

Figure 26.1 Example of Singular Value Decomposition Plots



Contents

Overview of the Association Analysis Platform	537
Example of the Association Analysis Platform	538
Launch the Association Analysis Platform	540
Data Format	541
The Association Analysis Report	542
Frequent Item Sets	542
Rules	542
Association Analysis Platform Options	543
Overview of Singular Value Decomposition	544
SVD Report	545
SVD Report Options	546
Overview of Topic Analysis	547
Topic Analysis Report	547
Topic Analysis Report Options	548
Additional Example of the Association Analysis Platform	549
Statistical Details for the Association Analysis Platform	551
Statistical Details for Frequent Item Set Generation	551
Statistical Details for Performance Measures	552



Overview of the Association Analysis Platform

The Association Analysis platform identifies connections among groups of items in an independent event or *transaction*. In association analysis, an *item* is the basic object of interest. For example, an item could be a product, a web page, or a service. An *item set* is a list of one or more items.

The relationship between two item sets is defined by an *association rule*. An association rule consists of a *condition* item set and a *consequent* item set. *Antecedents* are the individual items in the condition item set. Association analysis identifies association rules, which predict that a consequent item set will be in a transaction, given that the condition item set is already in the transaction. Some association rules are stronger, and therefore more useful, than others. The following three performance measures describe the strength of an association rule:

- *Support* is the proportion of transactions in which an item set appears. A high value for support indicates that the item set occurs frequently.
- *Confidence* is the proportion of transactions that contain the consequent item set, given that the condition item set is in the transaction. Confidence measures the strength of implication, or the predictive power, of an association rule. Note that confidence in association analysis is not related to the concept of confidence intervals.
- *Lift* is the ratio of an association rule's confidence to its expected confidence, assuming that the condition and consequent item sets appear in transactions independently. Lift measures how much the consequent item set depends on the presence of the condition item set. The minimum value for lift is 0.
 - A lift ratio less than 1 indicates that the condition and consequent repel each other, because they occur together less frequently than one would expect by chance alone.
 - A lift ratio close to 1 indicates that the consequent occurs at the same rate in transactions that contain the condition as one would expect from chance alone.
 - A lift ratio greater than 1 indicates that the consequent item set has an affinity for the condition item set. The consequent item set occurs more often with the condition item set than one would expect by chance alone.

For more information about these performance measures, see [“Statistical Details for Performance Measures”](#).

The Association Analysis platform also enables you to perform singular value decomposition. Singular value decomposition (SVD) groups similar transactions and also groups similar items using a matrix reducing methodology that is different from what is used in association analysis. Use the SVD methodology to gain insights that complement what you learn from association analysis.

For more information about association analysis, see Hastie et al. (2009) and Shmueli et al. (2010). For more information about singular value decomposition, see Jolliffe (2002).



Example of the Association Analysis Platform

This example uses transactional data reported by a grocery store. For items purchased by 1001 customers, you want to explore the associations among items in order to identify patterns in consumer behavior.

1. Select **Help > Sample Data Folder** and open Grocery Purchases.jmp.
2. Select **Analyze > Screening > Association Analysis**.
3. Select Product and click **Item**.
4. Select Customer ID and click **ID**.

Note: Each of the 1001 customers is assigned a unique custom ID.

5. Click **OK**.

By default, the Rules report is sorted in decreasing order by Confidence. However, association rules with extremely high confidence also tend to have a higher number of items in the condition set. Since you want to view association rules with smaller condition sets, sort the report by Confidence, but in increasing order.

6. Right-click in the Rules report and select **Sort By Column**.

The Select Columns window appears.

7. Select Confidence and then check the Ascending option.
8. Click **OK**.

Figure 26.2 Association Analysis Report

Rules			
Rule		Confidence	Lift
Condition	Consequent		
apples	sardines	40%	1.357
Coke	turkey	40%	1.422
sardines	chicken, Heineken	40%	2.002
herring, olives	ham, turkey	40%	3.006
ice cream	turkey	40%	1.424
artichoke	baguette, Heineken	40%	1.547
ham	corned beef, herring	40%	1.648
ham	corned beef, olives	40%	1.703
peppers	sardines	41%	1.371
sardines	peppers	41%	1.371
Coke	Heineken, sardines	41%	2.23
sardines	Coke, Heineken	41%	2.332
avocado, Heineken	crackers, ham	41%	2.707
soda	baguette, herring	41%	1.631
corned beef	ham	41%	1.343
Heineken, soda	bourbon, crackers, olives	41%	3.152
artichoke, Heineken	avocado, baguette	41%	1.921
Heineken, herring	avocado, baguette	41%	1.924

9. Scroll down the report to where the Confidence values are 58%.

There is an entry in the Rules report table that indicates that 58% of customers who bought an avocado also bought an artichoke. The value of Lift is 1.908, indicating that there is a likely dependency. You want to verify that avocados and artichokes occur in a significant portion of transactions.

10. Click the disclosure icon next to Frequent Item Sets.

Figure 26.3 Frequent Item Sets Report

Frequent Item Sets		
Item Set	Support	N Items
{Heineken}	60%	1
{crackers}	49%	1
{herring}	49%	1
{olives}	47%	1
{bourbon}	40%	1
{baguette}	39%	1
{corned beef}	39%	1
{crackers, Heineken}	37%	2
{avocado}	36%	1
{soda}	32%	1
{chicken}	31%	1
{apples}	31%	1
{ice cream}	31%	1
{artichoke}	30%	1
{ham}	30%	1
{Coke}	30%	1
{peppers}	30%	1
{sardines}	30%	1
{Heineken, herring}	29%	2
{turkey}	28%	1
{baguette, Heineken}	26%	2
{Heineken, soda}	26%	2
{herring, olives}	26%	2
{artichoke, Heineken}	25%	2

The Frequent Item Sets report shows that 36% of customers purchased avocados. The Rules report in [Figure 26.2](#) shows that 58% of these customers also bought artichokes. Because of the large proportion of customers who follow this behavior, the grocery store management might use this information to strategically locate avocados and artichokes.

You also decide to look at the association rules with the highest lift.

11. Right-click in the Rules report table and select **Sort By Column**.

The Select Columns window appears.

12. Select Lift and click **OK**.

The Rules table is sorted by decreasing values of lift. Notice that the second association rule has a lift of 5.642 and 83% confidence. You want to verify that both the condition set, {chicken, ice cream}, and the consequent item set, {Coke, sardines}, have adequate support.

13. Right-click in the Frequent Item Sets report and select **Sort By Column**.

The Select Columns window appears.

14. Select Item Set and then check the Ascending option.

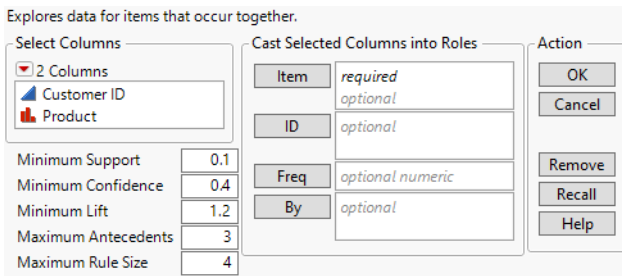
15. Click **OK**.

The Frequent Item Sets table is sorted alphabetically by item set. Scroll through the list to see that the condition item set, {chicken, ice cream}, has 14% support. The consequent item set, {Coke, sardines}, has 15% support. This association rule has high lift, but represents fewer transactions than the first association rule that you examined.

JMP^{PRO} Launch the Association Analysis Platform

Launch the Association Analysis platform by selecting **Analyze > Screening > Association Analysis**.

Figure 26.4 Association Analysis Launch Window



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Item The categorical column or columns that contain the item data to be analyzed. If the column has the Multiple Response modeling type, each of the multiple responses in each row is treated as an item. If multiple columns are specified, each column in each row is treated as an item.

ID The column that identifies the transaction to which an item belongs. The ID role is required, unless the Item role is filled by a column with the Multiple Response modeling type or the Item role is filled by multiple columns.

Note: If an ID role is specified, all rows with the same ID are combined into one transaction.

Freq The column that identifies a frequency for the transactions in the Item column. For example, if a transaction has a frequency of 3, that is treated as three identical transactions that contain the items in the Item column.

Note: The Freq role is ignored if the Item role is filled by a column that does not have the Multiple Response modeling type.

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.

Minimum Support Specifies a minimum value for the proportion of occurrences of an item set. This value must be between 0 and 1. Only item sets with support equal to or exceeding this value are considered in the analysis.

Minimum Confidence Specifies a minimum value for the proportion of occurrences that a consequent item set occurs within transactions that contain the conditional item set. This value must be between 0 and 1. Only association rules with confidence equal to or exceeding this value appear in the report.

Minimum Lift Specifies a minimum dependency ratio. Lift values must be 0 or greater. Only association rules with lift equal to or exceeding this value appear in the report.

Maximum Antecedents Specifies the maximum number of items in the condition item set. Association rules with more than this number of items in the condition set are not considered in the analysis.

Maximum Rule Size Specifies the maximum number of items that appear in the union of the condition and consequent item sets. Association rules with more than this combined number of items are not considered in the analysis. The default value is 4.

Note: You can use the minimum support, maximum antecedent, and maximum rule size options in the launch window to reduce computational time for large data sets. For more information about these measures, see [“Statistical Details for the Association Analysis Platform”](#).

Data Format

There are three data formats that are accepted in the Association Analysis platform:

- You can specify a single item response in each row in the Item role and identify which items are included in each transaction using the ID role. In this format, the Item and ID roles are required and the Freq role is ignored.
- You can specify a Multiple Response modeling type column for the Item role. In this format, the Item role is required, the ID and Freq roles are optional. If an ID role is specified, all rows with the same ID are combined into one transaction. If a Freq role is specified, the frequency applies to the transaction, not the specific items in the transaction.
- You can specify multiple item response columns in the Item role. In this format, the ID and Freq roles are optional. If an ID role is specified, all rows with the same ID are combined into one transaction. If a Freq role is specified, the frequency applies to the transaction, not the specific items in the transaction.



The Association Analysis Report

By default, the Association Analysis report contains the following reports:

- “Frequent Item Sets”
- “Rules”

Tip: To order the contents of a table in a report by any of its columns, right-click in the table and select **Sort by Column**.



Frequent Item Sets

The Frequent Item Sets report in the Association Analysis platform lists item sets in decreasing order of support. The listed item sets meet the Minimum Support value that you specified in the launch window. Each item set is considered as a conditional and as a consequent item set to form association rules. The table contains the following columns:

Item Set The item sets that are considered as conditional or consequent sets for the association rules.

Support The proportion of transactions in which all of the items in the Item Set occur.

N Items The number of items in the Item Set.



Rules

The Rules report in the Association Analysis platform shows a table of association rules that are sorted in decreasing order of confidences. Only association rules that meet the Minimum Support, Minimum Confidence, Minimum Lift, Maximum Antecedents, and Maximum Rule Size requirements that you specified in the launch window appear in this report.

The Rules report table contains the following columns:

Rule The association rules formed by combining Condition and Consequent item sets.

Condition The item set that is thought to influence the presence of a Consequent item set within transactions.

Consequent The item set whose presence is thought to be influenced by the presence of a Condition item set.

Confidence The proportion of transactions that contain the Consequent item set, given that the condition item set is in the transaction. Confidence measures the strength of implication, or the predictive power, of an association rule.

Lift The ratio of an association rule's confidence to its expected confidence, assuming that the condition and consequent item sets appear in transactions independently. Lift measures how much the Consequent item set depends on the presence of the Condition item set. The minimum value for lift is 0.

- A lift ratio less than 1 indicates that the Condition and Consequent item sets repel each other, because they occur together less frequently than one would expect by chance alone.
- A lift ratio close to 1 indicates that the Consequent item set occurs at the same rate in transactions that contain the Condition item set as one would expect from chance alone.
- A lift ratio greater than 1 indicates that the Consequent item set has an affinity for the Condition item set. The Consequent item set occurs more often with the Condition item set than one would expect by chance alone.

Association Analysis Platform Options

The Association Analysis red triangle menu contains the following options:

Transaction Listing Shows or hides a table listing each Transaction ID value and the items included in that transaction. The table is sorted by the Transaction ID column.

Frequent Item Sets Shows or hides a list of item sets whose support exceeds the Minimum Support value specified in the launch window. See [“Frequent Item Sets”](#).

Rules Shows or hides a table of association rules that meet the Minimum Support, Minimum Confidence, Minimum Lift, Maximum Antecedents, and Maximum Rule Size requirements specified in the launch window. See [“Rules”](#).

SVD Shows or hides a report of a partial singular value decomposition (SVD) of the incidence matrix for the items. This decomposition reduces the incidence matrix into a user-specified number of dimensions for analysis. The SVD report also contains options to visualize the singular vectors, cluster transactions and items, save out singular vectors, and perform a rotated SVD analysis. See [“Overview of Singular Value Decomposition”](#).

See *Using JMP* 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.

Platform Preferences Contains options that enable you to view the current platform preferences or update the platform preferences to match the settings in the current JMP report.

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.



Overview of Singular Value Decomposition

Singular value decomposition (SVD) complements association analysis by providing another method to identify items that have an affinity for each other. Singular value decomposition of the transaction item matrix reduces the matrix to a manageable number of dimensions, thereby enabling you to group similar transactions and similar items. The SVD analysis is equivalent to performing principal components analysis (PCA) on a correlation matrix.

The *transaction item matrix* is a matrix for which each row corresponds to a transaction and each column corresponds to an item. The entries of the matrix are zeros and ones. If an item occurs in a transaction, the corresponding row and column entry is one. Otherwise, the row and column entry is zero. Because the transaction item matrix usually contains more values of zero than one, it is called a *sparse matrix*.

The partial singular value decomposition approximates the column-standardized transaction item matrix using three matrices: $\mathbf{U}$, $\mathbf{S}$, and $\mathbf{V}'$. The relationship between these matrices is defined as follows:

$$\text{Transaction Item Matrix} \approx \mathbf{U} * \mathbf{S} * \mathbf{V}'$$

Define $nTran$ as the number of transactions (rows) in the transaction item matrix, $nItem$ as the number of items (columns) in the transaction item matrix, and $nVec$ as the specified number of singular vectors. Note that $nVec$ must be less than or equal to $\min(nTran, nItem)$. It follows that $\mathbf{U}$ is an $nTran$ by $nVec$ matrix that contains the left singular vectors of the transaction item matrix. $\mathbf{S}$ is a diagonal matrix of dimension $nVec$. The diagonal entries in $\mathbf{S}$ are the singular values in the transaction item matrix. $\mathbf{V}'$ is an $nVec$ by $nItem$ matrix. The rows in $\mathbf{V}'$ (or columns in $\mathbf{V}$) are the right singular vectors.

The right singular vectors capture connections among different items with similar functions or topic areas. If three items tend to appear in the same transactions, the SVD is likely to produce a singular vector in $\mathbf{V}'$ with large values for those three items. The $\mathbf{U}$ singular vectors represent the transactions projected into this new item space.

The SVD also captures indirect connections. If two items never appear together in the same transaction, but they generally appear in transactions with another third item, the SVD is able to capture some of that connection. If two transactions have no items in common but contain items that are connected in the dimension-reduced space, they map to similar vectors in the SVD plots.

The SVD transforms transaction data into a fixed-dimensional vector space, making it amenable to clustering, classification, and regression techniques. The Save options enable you to export this vector space to be analyzed in other JMP platforms.

The transaction item matrix is centered, scaled, and divided by $nTran$ minus 1 before the singular value decomposition is carried out. This analysis is equivalent to a PCA of the correlation matrix of the transaction item matrix. The SVD implementation takes advantage of the sparsity of the transaction item matrix.

SVD Report

The SVD option in the Association Analysis platform produces two SVD plots and a table of the singular values from the singular value decomposition.

SVD Plots

The first plot contains a point for each transaction. For a given transaction, the point that is plotted is defined by the transaction's values in the first two singular vectors (the first two columns of the $\mathbf{U}$ matrix) multiplied by the diagonal singular values matrix ($\mathbf{S}$). This plot is equivalent to the Score Plot in the Principal Components platform. Each point in this plot represents a transaction. In the Transaction SVD plot, points that are visibly grouped together indicate transactions with a similar composition.

The second plot contains a point for each item. For a given item, the point that is plotted is defined by the item's values in the first two singular vectors (the first two rows of the $\mathbf{V}'$ matrix) multiplied by the diagonal singular values matrix ($\mathbf{S}$). This plot is equivalent to the Loadings Plot in the Principal Components platform. Each point in this plot represents an item. In the Item SVD plot, items that are visibly grouped together indicate items that have similar functions or topic areas.

JMP[®] PRO Singular Values

Below the transaction and item SVD plots, a table of the singular values appears. These are the diagonal entries of the **S** matrix in the singular value decomposition of the transaction item matrix. The Singular Values table also contains a column of corresponding eigenvalues for the equivalent principal components analysis. Like in the Principal Components platform, there are columns for the percent and cumulative percent of variation explained by each eigenvalue (or singular value). You can use the Cum Percent column to decide what percent of variance from the transaction item matrix you want to preserve, and then use the corresponding number of singular vectors.

JMP[®] PRO SVD Report Options

The SVD red triangle menu in the Association Analysis platform contains the following options:

SVD Scatterplot Matrix Shows or hides a scatterplot matrix of the item and transaction singular value decomposition vectors. You must select the size of the scatterplot matrix when you select this option. This scatterplot matrix enables you to visualize more than the first two dimensions of the singular value decomposition.

Topic Analysis, Rotated SVD Performs a varimax rotated partial singular value decomposition of the item transaction matrix to produce groups of items called topics. You can select this option multiple times to find different numbers of topics. See [“Overview of Topic Analysis”](#).

Cluster Items Shows or hides a hierarchical clustering analysis of the items in the data. The following options are available to the right of the dendrogram:

Set Clusters Specifies the number of clusters that are colored in the dendrogram and used in the Save Clusters option. Points in the Item SVD plot are also colored by cluster.

Save Clusters Saves a data table that contains the items and their cluster assignments. You must set the number of clusters before clicking Save Clusters.

For more information about clustering and dendrograms, see *Multivariate Methods*.

Cluster Transactions Shows or hides a hierarchical clustering analysis of the transactions in the data. The following options are available to the right of the dendrogram:

Set Clusters Specifies the number of clusters that are colored in the dendrogram and used in the Save Clusters option. Points in the Transaction SVD plot are also colored by cluster.

Save Clusters Saves a data table that contains the transactions and their cluster assignments. You must set the number of clusters before clicking Save Clusters.

For more information about clustering and dendrograms, see *Multivariate Methods*.

Save Transaction Singular Vectors Saves a user-specified number of singular vectors from the transaction singular value decomposition as columns to a new data table where each row corresponds to a transaction. The first two saved columns represent the points plotted in the transaction SVD plot. See [“Overview of Singular Value Decomposition”](#).

Save Item Singular Vectors Saves a user-specified number of singular vectors from the items singular value decomposition as columns to a new data table where each row corresponds to an item. The first two saved columns represent the points plotted in the item SVD plot. See [“Overview of Singular Value Decomposition”](#).

Overview of Topic Analysis

The Topic Analysis, Rotated SVD option in the Association Analysis platform performs a varimax rotation on the partial singular value decomposition (SVD) of the transaction item matrix. See [“Overview of Singular Value Decomposition”](#). You must specify a number of rotated singular vectors, which corresponds to the number topics that you want to retain from the transaction item matrix. After you specify a number of topics, the Topic Analysis report appears.

Topic analysis is equivalent to a rotated principal component analysis (rotated PCA). The varimax rotation takes a set of singular vectors and rotates them to make them point more directly in the coordinate directions (toward the items). This rotation makes the vectors help explain the composition of the transactions as each rotated vector orients toward a set of items. Negative values indicate a repulsion force. The items with negative values occur in a topic less frequently compared to the items with positive values.

Topic Analysis Report

The Topic Analysis report in the Association Analysis platform shows the items that have the largest loadings in each topic after rotation. There are additional reports that show the components of the rotated singular value decomposition.

The Top Loadings by Topic report shows a table for each topic. The items in each table are the ones that have the largest loadings in absolute value for each topic. Each table is sorted in descending order by the absolute value of the loading. These tables can be used to determine conceptual themes that correspond to each topic.

The Topic Analysis report also contains the following reports:

Topic Loadings Contains a matrix of the loadings across topics for each item. This matrix is equivalent to the factor loading matrix in a rotated PCA.

Word Clouds by Topic Contains a matrix of word clouds, one for each topic.

Topic Scores Contains a matrix of transaction scores for each topic. Transactions with higher scores in a topic are more likely to be associated with that topic.

Topic Scores Plots Shows the topic scores for all transactions in one-dimensional scatterplots. Negative values indicate transactions that are negatively associated with a topic. Use these plots to explore the distribution of transactions within each topic.

Variance Explained by Each Topic Contains a table of the variance explained by each topic. The table also contains columns for the percent and cumulative percent of the variation explained by each topic.

Rotation Matrix Contains the rotation matrix for the varimax rotation.

Topic Analysis Report Options

The Topic Analysis red triangle menu in the Association Analysis platform contains the following options:

Topic Scatterplot Matrix Shows or hides a scatterplot matrix of the rotated singular value decomposition vectors.

Display Options Contains options to show or hide content that appears in the Topic Analysis report. See [“Topic Analysis Report”](#).

Rename Topics Shows a dialog that enables you to add descriptive names for one or more of the topics.

Save Transaction Topic Vectors Saves a user-specified number of singular vectors from the rotated singular value decomposition as columns to the data table.

Save Item Topic Vectors Saves the topic vectors as columns to a new Item Topic Scores data table.

Remove Removes the Topic Analysis report from the SVD report.

JMP PRO Additional Example of the Association Analysis Platform

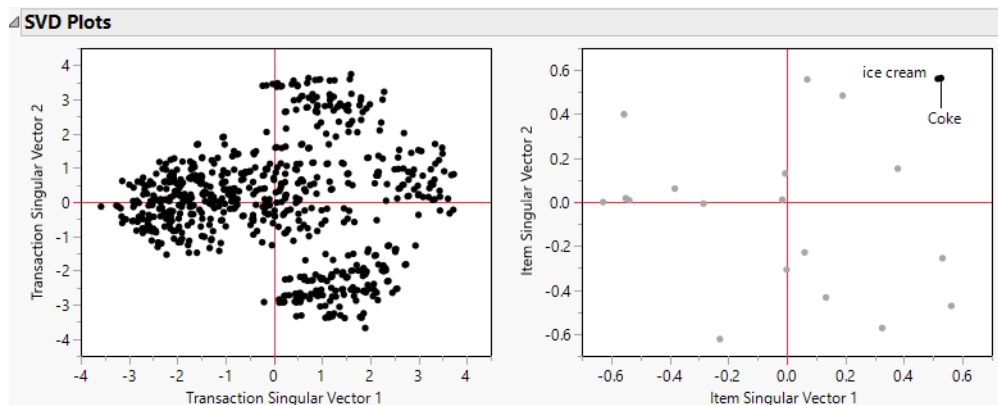
In this example, you use singular value decomposition of the transaction item matrix to analyze transactional data reported by a grocery store. The data contains information for items purchased by 1001 customers.

1. Select **Help > Sample Data Folder** and open Grocery Purchases.jmp.
2. Select **Analyze > Screening > Association Analysis**.
3. Select Product and click **Item**.
4. Select Customer ID and click **ID**.

Note: Each of the 1001 customers is assigned a unique custom ID.

5. Click **OK**.
6. Click the Association Analysis red triangle and select **SVD**.
7. Click **OK**.

Figure 26.5 SVD Plots



The transaction SVD plot suggests that there might be two or three groups of transactions. In the upper right corner of the item SVD plot, notice that the points that represent Coke and ice cream overlap. The proximity of these two items indicates that there is a strong affinity between them.

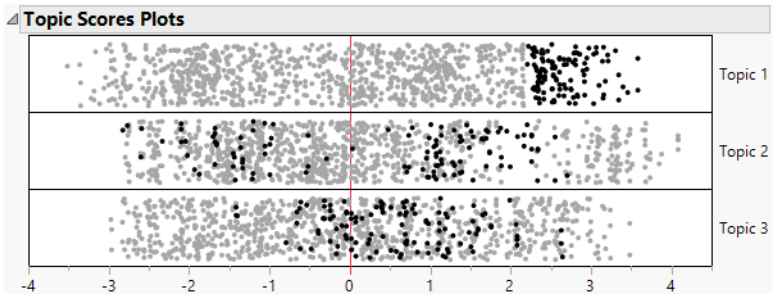
8. Click the SVD red triangle and select **Topic Analysis, Rotated SVD**.
 9. Type 3 next to Number of topics (rotated singular vectors) and click **OK**.
- The Topic Items and Topic Scores reports appear.

Figure 26.6 Topic Items Report

Top Loadings by Topic					
Topic 1		Topic 2		Topic 3	
Item	Loading	Item	Loading	Item	Loading
avocado	-0.7191	Coke	0.7017	crackers	0.7050
olives	0.7142	ice cream	0.6945	apples	-0.6437
baguette	-0.6089	herring	-0.6550	soda	0.6411
turkey	0.5697	sardines	0.6122	Heineken	0.6409
artichoke	-0.4889	chicken	0.5158	sardines	-0.3618
bourbon	0.4860	corned beef	-0.4257	steak	-0.3468
Heineken	-0.4329	steak	-0.3395	corned beef	-0.3409
corned beef	0.3836	olives	-0.3035	bourbon	0.3340
Coke	0.3096	ham	-0.3022		

Three groups, or topics, are created and shown in the Topic Items report. The first items listed in the Topic Item tables represent the primary items for that group. For example, Topic 1 is a group that is identified primarily by transactions that do not contain avocados, but do contain olives.

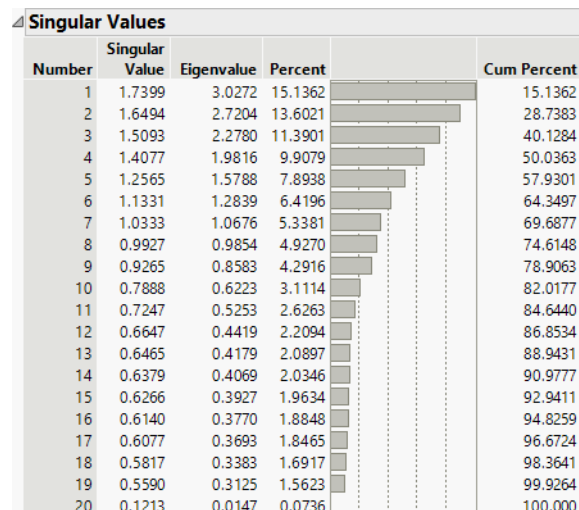
Figure 26.7 Topic Scores



The topic scores that are assigned to each of the 1001 transactions are plotted in the Topic Scores report. Select groups of points for a topic to see how those transactions relate to other topics. For example, transactions with very high values on Topic 1 tend to have low values on Topics 2 and 3.

- 10. Open the Singular Values report.

Figure 26.8 Singular Values Table



As seen in [Figure 26.8](#), the first two singular values explain only about 30% of the variability in the grocery store data. Additional dimensions might be required to explain a sufficient amount of variability.

JMP PRO Statistical Details for the Association Analysis Platform

This section contains statistical details for these topics:

- [“Statistical Details for Frequent Item Set Generation”](#)
- [“Statistical Details for Performance Measures”](#)

JMP PRO Statistical Details for Frequent Item Set Generation

The Association Analysis platform uses the *Apriori algorithm* to reduce computational time when generating frequent item sets. The Apriori algorithm leverages the fact that an item set’s support is never larger than the support of its subsets. The platform generates larger item sets from combinations of smaller item sets that meet the minimum support level. In addition, the platform does not generate item sets that exceed either the specified maximum number of antecedents or the maximum rule size. These options are useful when working with large data sets, because the total possible number of rules increases exponentially with the number of items. For more information about the Apriori algorithm, see Agrawal and Srikant (1994).

JMP[®] PRO Statistical Details for Performance Measures

This section defines the performance measures used in the Association Analysis platform. Denote an association rule with condition item set X and consequent item set Y by $X \Rightarrow Y$. Hahsler (2015) contains a collection of performance measures used in association analysis, including support, confidence, and lift.

JMP[®] PRO Support

Support is the proportion of transactions in which an item set occurs. Support can also be viewed as the probability that a transaction contains an item set.

The support S of a condition item set X is defined as follows:

$$\text{supp}(X) = \frac{N_X}{N} = \text{Pr}(X)$$

where:

N_X is the number of transactions that contain the item set X

N is the total number of transactions.

Support for an association rule is defined as follows:

$$\text{supp}(X \Rightarrow Y) = \text{supp}(X \cup Y) = \text{Pr}(X \cap Y)$$

In this case, support is equivalent to the probability that a transaction contains both item sets X and Y .

For both the item set and association rule cases, support ranges from 0% to 100%.

JMP[®] PRO Confidence

Confidence is the proportion of transactions that contain the consequent item set, given that the transaction contains the condition item set. Confidence can also be viewed as the conditional probability that a transaction contains the consequent item set, given that the transaction contains the condition item set.

$$\text{conf}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X)} = \frac{\text{Pr}(X \cap Y)}{\text{Pr}(X)} = \text{Pr}(Y|X)$$

Confidence ranges from 0% to 100%. An association rule with a confidence of 0% has a consequent item set that does not appear in any transaction with the condition item set. A confidence of 100% indicates that every transaction that contains the condition item set also contains the consequent item set.

Note: Confidence in association analysis is not related to the concept of confidence intervals.

JMP PRO Lift

Lift measures dependency between X and Y .

$$\text{lift}(X \Rightarrow Y) = \frac{\text{supp}(X \cup Y)}{\text{supp}(X) \times \text{supp}(Y)} = \frac{\text{Pr}(X \cap Y)}{\text{Pr}(X)\text{Pr}(Y)}$$

The numerator for lift is the proportion of transactions where X and Y occur jointly. The denominator is an estimate of the expected joint occurrence of X and Y , assuming that they occur independently.

Lift ranges from 0 to ∞ . A lift value of 1 indicates that X and Y jointly occur in transactions with the frequency that would be expected by chance alone. Increasing lift values suggest that Y occurs more often than expected when X is present.

Note: The lift for association rule $X \Rightarrow Y$ is equal to the lift for association rule $Y \Rightarrow X$.

Chapter 27

Process History Explorer

Identify Problem Components in Complex Process Histories

Process History Explorer is useful in complex manufacturing situations, such as in semiconductor factories, where a defective tool or route combination might lead to poor yield or reduced quality. There might be hundreds of steps that a unit passes through that involve many different tools. It can be difficult to determine the exact combination of factors that leads to the poor yield.

Process History Explorer organizes all of the production tracking data and enables you to identify factors that seem to be associated with poor yield. You can then perform stepwise regression. You can analyze the time that a unit spends waiting between steps or tools in the process and see how that waiting time is related to yield. You can also analyze the effect on yield of transitions from one tool to another.

Note: The Process History Explorer platform is experimental. It is subject to change in future releases.

Figure 27.1 Process History Explorer Report

Process History Explorer						
1250 Units, using ID columns: Lot, Wafer						
125 Steps, using columns: Layer, Operation						
156250 Operations, using time columns: TimeIn, TimeOut						
195 levels in X columns: Tool, Route						
1 response: Yield						
Response Yield						
Levels with Lowest Yields						
Term	Level	Count	Mean	Std Error	Std Dev	
Tool	PlanarTool008	1000	0.75299	0.00188	0.05952	
Tool	DepTool003	1375	0.75577	0.00168	0.06245	
Tool	EtchTool019	750	0.76192	0.0021	0.05749	
Tool	EtchTool012	500	0.76203	0.0026	0.05807	
Tool	EtchTool034	850	0.76279	0.00199	0.0581	
Tool	PlanarTool021	1050	0.7637	0.00186	0.06028	
Tool	PlanarTool030	1025	0.76539	0.00184	0.05894	
Tool	EtchTool016	900	0.76566	0.00191	0.05734	
Tool	EtchTool009	925	0.76598	0.00197	0.05994	
Tool	Dep2Tool018	1525	0.76766	0.00156	0.06088	
Tool	Dep2Tool014	1825	0.76813	0.00138	0.05893	
Tool	DepTool019	1650	0.76839	0.00143	0.05794	
Tool	EtchTool020	1000	0.76909	0.00177	0.05586	
Tool	Dep2Tool015	1475	0.76948	0.0015	0.05775	
Tool	PhotoTool010	3525	0.76996	0.00098	0.05822	
Tool	Dep2Tool013	1825	0.7705	0.00141	0.06025	
Tool	DepTool012	1750	0.77063	0.00145	0.06051	
Tool	Dep2Tool006	1400	0.77063	0.00158	0.05905	
Tool	PlanarTool028	900	0.77103	0.00202	0.06071	
Tool	Dep2Tool011	1600	0.77124	0.00145	0.05797	
Tool	EtchTool010	850	0.77125	0.00213	0.06208	

Contents

Example of Process History Explorer..... 557

Launch the Process History Explorer Platform 559

 Data Format 560

The Process History Explorer Report..... 560

Process History Explorer Platform Options 561

Example of Process History Explorer

Use Process History Explorer to investigate the yields of wafers that have gone through a complex manufacturing process.

1. Select **Help > Sample Data Folder** and open Quality Control > Lot Wafer History.jmp and Quality Control > Lot Wafer Yield.jmp.

Lot Wafer History.jmp contains step-by-step history of 25 wafers in 50 lots. Lot Wafer Yield.jmp contains the yield for each wafer in each lot.

2. From the Lot Wafer History.jmp data table, select **Analyze > Screening > Process History Explorer**.
3. Select Lot and Wafer and click **ID**.
4. Select Tool and Route and click **X, Process**.
5. Select Layer and Operation and click **Step**.
6. Select TimeIn and TimeOut and click **Timestamp**.
7. Click **OK**.

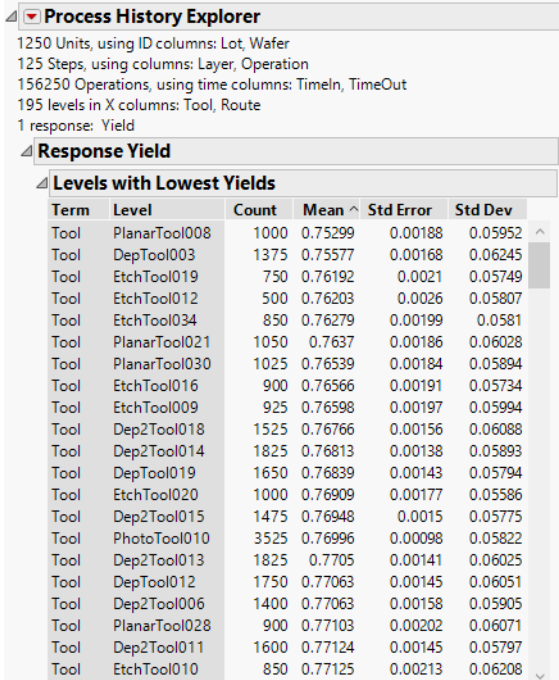
A window appears that enables you to select the corresponding yield table.

8. Select Lot Wafer Yield and click **OK**.

A window appears that enables you to select the yield column.

9. Select Yield and click **OK**.
10. Click the Process History Explorer red triangle and select **Levels with Lowest Yield**.

Figure 27.2 Process History Explorer Report

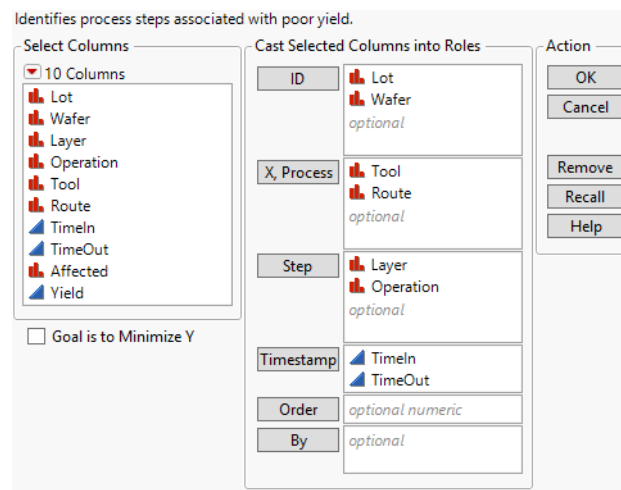


The report contains information about the process, including the number of units, steps, operations, and levels in the X columns. The Levels with Lowest Yields table contains counts and yield information for the 195 levels in the X columns. Notice that the PlanarTool008 tool had the lowest yield and 1,000 units went through that tool.

Launch the Process History Explorer Platform

Launch the Process History Explorer platform by selecting **Analyze > Screening > Process History Explorer**.

Figure 27.3 Process History Explorer Launch



For more information about the options in the Select Columns red triangle menu, see *Using JMP*.

Launch Window Roles

ID Specifies one or more columns that identify a unit in the process.

Note: The ID columns must have corresponding columns in the yield table.

X, Process Specifies one or more columns that identify the components or specific details that make up the steps of the process.

Step Optional columns that specify the steps and sub-steps that a unit goes through. For example, the layer and the type of operation within the layer.

Timestamp Specifies the time of the operation. The time can be specified as a single time point or as a pair of time points that indicate the times that the operation started and ended.

Order Optional column that specifies the order that a batch of units went through the components specified in the X, Process role.

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.

Launch Window Options

Goal is to Minimize Y Specifies that the goal of the analysis is to minimize the response. By default, the analysis seeks to maximize the response.

Note: The response is specified in a second data table, not in the data table used to launch the platform.

Data Format

The data for Process History Explorer must be specified in two data tables. The first table is used to launch the platform and contains the steps and details of the process. The second table contains the response, or yield, for each level of the ID columns.

The sample data tables Quality Control > Lot Wafer History and Quality Control > Lot Wafer Yield provide an example of how the data are specified in two data tables.

The Process History Explorer Report

The initial Process History Explorer Report provides a summary of the information in the data table(s).

- The number of Units, determined by the ID columns. If there is more than one ID column, this is the number of unique combinations.
- The number of Steps, determined by the Step columns. If there is more than one Step column, this is the number of unique combinations.
- The number of Operations, determined by the Timestamp columns. This is the same as the number of rows in your data table.
- The number of levels, determined by the X, Process columns. If there is more than one X, Process column, this is the sum of the number of levels in each column.
- The number of responses, determined by the Yield columns.

Select options from the Process History Explorer red triangle menu to continue your analysis. See [“Process History Explorer Platform Options”](#).

Process History Explorer Platform Options

The Process History Explorer red triangle menu contains the following options:

Levels with Lowest Yield Shows or hides a table of summary statistics for the yield for each level of each X, Process variable. The summary statistics provided for each level are the number of operations, and the mean, standard error, and standard deviation of the yield. By default, the table is sorted in ascending order by mean yield.

Levels with Lowest Yield with Time Filter Shows or hides a table of summary statistics for the time intervals of the top 50 levels with the lowest yields. The routine searches for a time interval for each of the top 50 levels in the Levels with Lowest Yield table that minimizes the mean yield within the interval. The interval must contain at least 25% of the operations in that level. The summary statistics provided for each level are the number of operations in the identified interval, the mean yield in the identified interval, the timestamps for the first and last operations in the level (First Time and Last Time), and the timestamps for the first and last operations in the identified interval (RunStart Time and RunEnd Time).

Stepwise Regression Shows a Stepwise Regression window that enables you to specify options for a stepwise regression analysis. You can use stepwise regression to find problems in the process, where a problem is something that causes your yield to decrease. At each step in the regression, a term enters according to which term has the largest negative yield. This helps you identify process steps that most impact the yield. Stepwise Regression provides a table that shows the order in which terms entered the model at each step of the regression.

Note: If the Goal is to Minimize Y option is selected in the launch window, Stepwise Regression adjusts to detect terms that cause the yield to increase.

Goal Specifies the criterion used in the stepwise regression model. The Biggest individual difference option identifies the terms that individually cause the yield to change the most. The Biggest total difference option identifies the terms that cause the biggest total difference, which is the difference multiplied by the number of units affected. The Most predictive option produces a traditional stepwise regression model.

Time Filtering Enables you to use the starting time of a component in the stepwise regression model. If you choose the None option, the times are averaged over the entire time period. If you choose the Starting Time option, the component is allowed to fail at a particular starting time. The starting time is the time that maximizes the negative yield affect. If you choose the Starting Time option, the Stepwise Regression report table contains a Start Time column.

X Transform Enables you to use the regular count data or the $\text{Log}(\text{Count} + 1)$ data for the X, Process variables in the stepwise regression. The $\text{Log}(\text{Count} + 1)$ option transforms the variables in the regression such that high frequency counts have less of a linear influence in the regression. If you choose the None option, high frequency counts have a linear influence in the regression.

Number of Regression Steps The number of steps in the Stepwise Regression analysis. The default is ten.

Figure 27.4 Stepwise Regression Window

Specify Stepwise Options

Goal

☒ Biggest individual difference
☐ Biggest total difference
☐ Most predictive

Time Filtering

☒ None
☐ Starting Time

X Transform

☒ None
☐ Log(Count+1)

Number of Regression Steps

10

OK Cancel

Waiting Time Analysis (Available only when two Timestamp variables are specified.) Each unique ID can go through only one level of a step at a time. In addition, each step can process only one ID at a time. Waiting Time Analysis provides a table of waiting time summary statistics for each ID. In this table, waiting time is the amount of time between when an ID is done with one step and when it starts another step. Additional Waiting Time Analysis tables are also provided. These tables contain waiting time summary statistics for each ID within the levels of each Step variable. In these tables, waiting time is the amount of time between when an ID is done with a particular level of a Step variable and when it goes through that same level again. The tables provide the following statistics:

Waiting Sum The sum of the waiting times for the given combination of ID and Step variable levels.

Waiting Count The number of waiting times for the given combination of ID and Step variable levels.

Waiting Mean The mean of the waiting times for the given combination of ID and Step variable levels.

Waiting Max The maximum waiting time for the given combination of ID and Step variable levels.

Transition Analysis Shows a report that helps you identify transitions from one component to another that are problematic for the resulting yield. This report can help you identify alignment and registration issues in the process. For example, some levels of the process column might be out of alignment with respect to other levels of the process column. If there is a transition problem, you would expect that items that repeat a specific level to have better yields than items that go through different levels of the process column.

When you select this option, you must choose one or more X, Process columns and one or more levels in each of those columns to include in the analysis. For each column that you specify for the analysis, a report appears that contains four tables. The first shows yields and counts for transitions from a specific level of the column to another specific level of the column. The second shows yields and counts for transitions away from specific levels of the column. The third shows yields and counts for transitions to specific levels of the column. The fourth shows yields and counts for transitions where the unit repeats a specific level of the column.

Order Interaction (Available only if an Order variable is specified in the launch window.) Shows a report of a one-way analysis of the yield on the Order column. The report contains a table for each X, Process column. In each table, the levels of the process column are listed in descending order of logworth of the corresponding analysis of variance model. The tables also contain the *p*-values and a heat map for each level of the process column.

Save Count Table Saves a new data table that contains counts of how many units went through each level of the X, Process variables. The count table has a row for each unique unit, determined by the ID values.

Save Log Count Table Saves a new data table that is the same format as the Count table, except with entries of $\text{Log}(\text{Count} + 1)$.

Save Waiting Time (Available only when two Timestamp variables are specified.) Saves each table in the Waiting Time Analysis report to a separate data table.

Appendix **A**

Statistical Details

Predictive and Specialized Modeling

This appendix discusses measures of fit, the types of cross validation available in each platform, and the predicted probabilities in tree methods.

Contents

Entropy RSquare 567

Validation in JMP Modeling 567

 Validation Column Role 568

 K-Fold and Holdback Validation 570

Predicted Probabilities in Tree Methods 572

Entropy RSquare

The Entropy RSquare statistic is computed for the training set and for the validation and test sets if validation is used. For the training set, Entropy RSquare is computed as follows:

- The specified model is fit using the training set.
- Predicted probabilities for each level of the response are obtained, based on the model.
- Using these predicted probabilities, the likelihood is computed for observations in the training set. Call this $Likelihood\_Full_{Training}$.
- The reduced model (no predictors) is fit using the training set.
- The predicted probabilities for the levels of the response from the reduced model are used to compute the likelihood for observations in the training set. Call this quantity $Likelihood\_Reduced_{Training}$.
- The Entropy RSquare for the training set is:

$$\text{Entropy RSquare}_{Training} = 1 - \frac{\log(Likelihood\_Full_{Training})}{\log(Likelihood\_Reduced_{Training})}$$

The Entropy RSquare for the validation and test sets are computed in a manner analogous to the Entropy RSquare for the training set.

Validation in JMP Modeling

Data can be partitioned into sets before modeling to avoid overfitting and to select a good predictive model. This process uses part of the original data to estimate parameters and uses the rest of the data to tune or evaluate the parameters, or do both. In JMP Pro, you can partition the data into two or three sets in the following ways:

Train and Evaluate Partitions the data into two sets, called Training and Validation. The training set is used to estimate the model parameters. The validation set is used to independently evaluate the performance of the fitted model.

Train and Tune Partitions the data into two sets, called Training and Validation. The training set is used to estimate the model parameters. The validation set is used in the model fitting algorithm to tune the model parameters and ultimately choose a model with good predictive ability. There is no independent model evaluation done in this case.

Train, Tune, and Evaluate Partitions the data into three sets, called Training, Validation, and Test. The training set is used to estimate the model parameters. The validation set is used in the model fitting algorithm to tune the model parameters and ultimately choose a

model with good predictive ability. The test set is then used to independently evaluate the performance of the fitted model.

JMP[®] PRO Validation Column Role

One way to create data partitions is to use the Validation Column role. The Validation Column role uses the column’s values to divide the data into parts. The column is assigned using the Validation role in the platform’s launch window. For information about how to create a validation column, see [“Make Validation Column”](#).

The Validation Column role is available only in JMP Pro. For JMP, see [“Excluded Rows as Validation Holdback”](#).

Caution: The use of a validation column is platform specific. Different platforms use the levels of the validation column differently. See notes in [Table A.1](#).

Table A.1 Validation Column by Platform

Platform	Train & Evaluate	Train & Tune	Train, Tune, & Evaluate	Notes
<i>Fit Model</i>				
Fit Least Squares	Yes	No	No	If there are more than three levels, the validation column is ignored.
Stepwise Regression	No	Yes	Yes	If there are more than three levels, K-Fold Cross-Validation is used.
Logistic Regression	Yes	No	No	If there are more than three levels, the validation column is ignored.
Generalized Regression	No	Yes	Yes	If there are more than three levels, K-Fold Cross-Validation is used.
Partial Least Squares	No	Yes	Yes	If there are more than three levels, K-Fold Cross-Validation is used.
<i>Predictive Models</i>				
Neural	No	Yes	Yes	If there are more than three levels, K-Fold Cross-Validation is used.
Partition	No	Yes	Yes	If there are more than three levels, the platform only uses rows with the three smallest values.

Platform	Train & Evaluate	Train & Tune	Train, Tune, & Evaluate	Notes
Bootstrap Forest	No	Yes	Yes	If there are more than three levels, the platform only uses rows with the three smallest values.
Boosted Tree	No	Yes	Yes	If there are more than three levels, the platform only uses rows with the three smallest values.
K Nearest Neighbors	No	Yes	Yes	If there are more than three levels, the platform only uses rows with the three smallest values.
Naive Bayes	No	Yes	Yes	If there are more than three levels, the platform only uses rows with the three smallest values.
Support Vector Machines	Yes	Yes	Yes	If there are more than three levels, K-Fold Cross-Validation is used.
<i>Specialized Models</i>				
Functional Data Explorer	Yes	No	No	Must be created as a Grouped Random validation column. If there are more than two levels, the smallest value defines the training set and all other values define the validation set.
<i>Multivariate Models</i>				
Discriminant	Yes	Yes	Yes	If there are more than three levels, the platform only uses rows with the three smallest values.
Partial Least Squares	No	Yes	Yes	If there are more than three levels, K-Fold Cross-Validation is used.
Uplift	No	Yes	Yes	If there are more than three levels, the platform only uses rows with the three smallest values.

K-Fold and Holdback Validation

The Validation Column role provides a framework for partitioning data into cross validation sets. In addition, some JMP platforms also support K-Fold and various types of Holdback validation.

K-Fold Cross-Validation Divides the original data into K subsets. In turn, each of the K sets is used to validate the model fit on the rest of the data, fitting a total of K models. The model that produces the best validation statistic is chosen as the final model, and the fold that is not used in the building of that model provides the validation set performance statistics shown in the report.

Note: For some platforms, you must specify K-Fold Cross-Validation in the model control panel. For other platforms, you must specify K-Fold Cross-Validation in the platform launch window. For still other platforms, you must specify K-Fold Cross-Validation through a validation column that contains more than three levels.

Random Validation Holdback (Available as a launch option for specific platforms.) Randomly divides the original data into the training and validation sets. A test set can also be included. You can specify the proportions of the original data to use in each set.

Leave-One-Out Validation Holdback (Available as an option for specific platforms.) Repeatedly fits the model leaving out one observation at a time. Leave-one-out validation is also known as the jackknife procedure.

Excluded Rows as Validation Holdback Uses the excluded rows in the data table as a validation holdback set. For JMP Pro, this option is available by selecting in the platform preferences.

Note: For platforms that support using excluded rows as a validation holdback set, the excluded rows are used only when there is no validation column or validation proportion specified in the launch window.

Table A.2 K-Fold and Holdback Validation by Platform

Platform	Excluded Rows as Validation Holdback	Random Validation Holdback	Leave-One-Out Holdback	K-Fold Cross-Validation
<i>Fit Model</i>				
Fit Least Squares	No	No	No	No

Platform	Excluded Rows as Validation Holdback	Random Validation Holdback	Leave-One-Out Holdback	K-Fold Cross-Validation
Stepwise Regression	No	No	No	Yes (for continuous response models only)
Logistic Regression	No	No	No	No
JMP PRO Generalized Regression	No	Yes	Yes	Yes (though the model controls)
JMP PRO Partial Least Squares	No	Yes	Yes	Yes (through the model controls)
<i>Predictive Model</i>				
Neural	Yes	Yes	No	Yes (through model launch or validation column)
Partition	Yes	Yes	No	Yes (select the option in the platform preferences)
JMP PRO Bootstrap Forest	Yes	Yes	No	No
JMP PRO Boosted Tree	Yes	Yes	No	No
JMP PRO K Nearest Neighbors	Yes	Yes	No	No
JMP PRO Naive Bayes	Yes	Yes	No	No
JMP PRO Support Vector Machines	No	Yes	No	Yes (through model launch)
<i>Specialized Models</i>				
JMP PRO Functional Data Explorer	No	No	No	No
<i>Multivariate Models</i>				

Platform	Excluded Rows as Validation Holdback	Random Validation Holdback	Leave-One-Out Holdback	K-Fold Cross-Validation
Discriminant	Optional	No	No	No
Partial Least Squares	No	Yes	Yes	Yes (through model launch or validation column)
 Uplift	No	Yes		No

Predicted Probabilities in Tree Methods

The predicted probabilities for the Decision Tree and Bootstrap Forest methods are calculated as described below by the Prob statistic.

For categorical responses in Decision Tree, the Show Split Prob command shows the following statistics:

- Rate** The proportion of observations at the node for each response level.
- Prob** The predicted probability for that node of the tree. The method for calculating Prob for the i^{th} response level at a given node is defined as follows:

$$\text{Prob}_i = \frac{n_i + \text{prior}_i}{\sum (n_i + \text{prior}_i)}$$

where the summation is across all response levels; n_i is the number of observations at the node for the i^{th} response level; and prior_i is the prior probability for the i^{th} response level, calculated as follows:

$$\text{prior}_i = \lambda p_i + (1 - \lambda) P_i$$

where p_i is the prior_i from the parent node, P_i is the Prob_i from the parent node, and λ is a weighting factor currently set at 0.9.

The method for calculating Prob assures that the predicted probabilities are always nonzero.

Appendix **B**

References

The following sources are referenced in *Predictive and Specialized Modeling*.

- Agrawal, R., and Srikant, R. (1994). "Fast Algorithms for Mining Association Rules." In *Proceedings of the 20th VLDB Conference*. Santiago, Chile: IBM Almaden Research Center. Accessed July 5, 2016. <https://web.stanford.edu/class/cs345d-01/rl/ar-mining.pdf>.
- Bates, D. M., and Watts, D. G. (1988). *Nonlinear Regression Analysis and Its Applications*. New York: John Wiley & Sons.
- Benford, F. (1938). "The law of anomalous numbers." *Proceedings of the American philosophical society*, 551–572.
- Benjamini, Y., and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society, Series B* 57:289–300.
- Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time Series Analysis: Forecasting and Control*. 3rd ed. Englewood Cliffs, NJ: Prentice-Hall.
- Candes, E. J., Li, X., Ma, Y., and Wright, J. (2009). "Robust Principal Component Analysis?" *Journal of the ACM*, 58:1–37.
- Cleveland, W. S. (1994). *Visualizing Data*, Summit, NJ: Hobart Press.
- Conover, W. J. (1999). *Practical Nonparametric Statistics*. 3rd ed. New York: John Wiley & Sons.
- Cureton, E. E. (1967). "The Normal Approximation to the Signed-Rank Sampling Distribution when Zero Differences are Present." *Journal of the American Statistical Association* 62: 319, 1068–1069.
- Donoho, D. L. (1995). "De-noising by soft-thresholding." *IEEE transactions on information theory*, 41(3), 613–627.
- Efron, B (1981). "Nonparametric standard errors and confidence intervals." *Canadian Journal of Statistics*, 9:139–172.
- Geladi, P., MacDougall, D., & Martens, H. (1985). "Linearization and scatter-correction for near-infrared reflectance spectra of meat." *Applied spectroscopy*, 39(3), 491–500.
- Hahsler, M. (2015). "A Probabilistic Comparison of Commonly Used Interest Measures for Association Rules." Accessed September 17, 2018. <https://mhahsler.github.io/arules/docs/measures>.
- Haldane, J. B. (1956). "The estimation and significance of the logarithm of a ratio of frequencies." *Ann Hum Genet*, 20, 309–311.
- Hand, D. J., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining*. Cambridge, MA: MIT Press.

- Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer-Verlag.
- Hawkins D. M., and Kass G. V. (1982). "Automatic Interaction Detection." In *Topics in Applied Multivariate Analysis*, edited by D. M. Hawkins, 267–300. Cambridge: Cambridge University Press.
- Huber, P. J., and Ronchetti, E. M. (2009). *Robust Statistics*. 2nd ed. New York: John Wiley & Sons.
- Hyndman, R. J., Koehler, A. B., Ord, J. K., and Snyder, R. D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Berlin: Springer-Verlag.
- Jolliffe, I. T. (2002). *Principal Component Analysis*. New York: Springer-Verlag.
- Kass, G. V. (1980). "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Journal of the Royal Statistical Society, Series C* 29:119–127
- Lehman, E. L. (2006). *Nonparametrics: Statistical Methods Based on Ranks*. 2nd ed. New York: Springer.
- Lin, Z., Chen, M., and Ma, Y. (2013). The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055.
- Mason, R. L., and Young, J. C. (2002). *Multivariate Statistical Process Control with Industrial Applications*. Philadelphia: SIAM.
- McCullagh, P., and Nelder, J. A. (1989). *Generalized Linear Models*. 2nd ed. London: Chapman & Hall.
- Nagelkerke, N. J. D. (1991). "A Note on a General Definition of the Coefficient of Determination." *Biometrika* 78:691–692.
- Nason, G. P. (Ed.). (2008). *Wavelet methods in statistics with R*. New York, NY: Springer New York.
- Nelder, J. A., and Wedderburn, R. W. M. (1972). "Generalized Linear Models." *Journal of the Royal Statistical Society, Series A* 135:370–384.
- Paixão, P., Gouveia, L. F., Silva, N., & Morais, J. A. (2017). "Evaluation of dissolution profile similarity—Comparison between the f_2 , the multivariate statistical distance and the f_2 bootstrapping methods." *European Journal of Pharmaceutics and Biopharmaceutics*, 112, 67–74.
- Parker, R. J. (2015). *Efficient Computational Methods for Large Spatial Data*. Ph.D. diss., Department of Statistics, North Carolina State University. Accessed June 30, 2016. <https://repository.lib.ncsu.edu/ir/bitstream/1840.16/10572/1/etd.pdf>.
- Platt, J. (1998). *Sequential minimal optimization: A fast algorithm for training support vector machines*. Technical Report MST-TR-98-14, Microsoft Research. <https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-98-14.pdf>.
- Qian, P. Z., Huaiqing, W., and Wu, C. F. (2012). "Gaussian process models for computer experiments with qualitative and quantitative factors." *Technometrics* 50:383–396.
- Ramsay, J. O., and Silverman, B. W. (2005). *Functional Data Analysis*. 2nd ed. New York: Springer.
- Ratkowsky, D. A. (1990). *Handbook of Nonlinear Regression Models*. New York: Marcel Dekker.

- Sall, J. (2002). "Monte Carlo Calibration of Distributions of Partition Statistics." SAS Institute Inc., Cary, NC. Accessed July 29, 2015.
<https://www.jmp.com/content/dam/jmp/documents/en/white-papers/montecarlocal.pdf>.
- Santer, T., Williams, B., and Notz, W. (2003). *The Design and Analysis of Computer Experiments*. New York: Springer-Verlag.
- SAS Institute Inc. (2020). *SAS/ETS 15.2 User's Guide*. Cary, NC: SAS Institute Inc.
<https://go.documentation.sas.com/api/docsets/etsug/15.2/content/etsug.pdf>.
- Savitzky, A., and Golay, M. J. (1964). "Smoothing and differentiation of data by simplified least squares procedures." *Analytical Chemistry* 36(8), 1627–1639.
- Schäfer, J., and Strimmer, K. (2005). "A Shrinkage Approach to Large-Scale Covariance Matrix Estimation and Implications for Functional Genomics." *Statistical Applications in Genetics and Molecular Biology* 4, Article 32.
- Schuirman, D. J. (1987). "A Comparison of the Two One-sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability." *Journal of Pharmacokinetics and Biopharmaceutics* 15:657–680.
- Shiskin, J., Young, A. H., and Musgrave, J. C. (1967). *The X-11 Variant of the Census Method II Seasonal Adjustment Program*. Technical Report 15, US Department of Commerce, Bureau of the Census.
- Shmueli, G., Patel, N. R., and Bruce, P. C. (2010). *Data Mining For Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner*. 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Shmueli, G., Bruce, P. C., Stephens M. L., and Patel, N. R. (2017). *Data Mining For Business Intelligence: Concepts, Techniques, and Applications with JMP Pro*. Hoboken, NJ: John Wiley & Sons.
- Smyth, Gordon K. (2004). "Linear models and empirical bayes methods for assessing differential expression in microarray experiments." *Statistical applications in genetics and molecular biology* 3(1).
- Westfall, P. H., Tobias, R. D., and Wolfinger, R. D. (2011). *Multiple Comparisons and Multiple Tests Using SAS*. 2nd ed. Cary, NC: SAS Institute Inc.

Appendix **C**

Technology License Notices

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

All Rights Reserved.

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

NEIL HODGSON DISCLAIMS ALL WARRANTIES WITH REGARD TO THIS SOFTWARE, INCLUDING ALL IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS, IN NO EVENT SHALL NEIL HODGSON BE LIABLE FOR ANY SPECIAL, INDIRECT OR CONSEQUENTIAL DAMAGES OR ANY DAMAGES WHATSOEVER RESULTING FROM LOSS OF USE, DATA OR PROFITS, WHETHER IN AN ACTION OF CONTRACT, NEGLIGENCE OR OTHER TORTIOUS ACTION, ARISING OUT OF OR IN CONNECTION WITH THE USE OR PERFORMANCE OF THIS SOFTWARE.

- Progress® Telerik® UI for WPF: Copyright © 2008-2019 Progress Software Corporation. All rights reserved. Usage of the included Progress® Telerik® UI for WPF outside of JMP is not permitted.
- ZLIB Compression Library is Copyright © 1995-2005, Jean-Loup Gailly and Mark Adler.
- Made with Natural Earth. Free vector and raster map data @ naturalearthdata.com.
- Packages is Copyright © 2009–2010, Stéphane Sudre (s.sudre.free.fr). All rights reserved.

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

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

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

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

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS “AS IS” AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES

(INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

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

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

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of OpenLink Software Inc. nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

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

- This program, “bzip2”, the associated library “libbzip2”, and all documentation, are Copyright © 1996–2019 Julian R Seward. All rights reserved.

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

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. The origin of this software must not be misrepresented; you must not claim that you wrote the original software. If you use this software in a product, an acknowledgment in the product documentation would be appreciated but is not required.
3. Altered source versions must be plainly marked as such, and must not be misrepresented as being the original software.

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

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

Julian Seward, jseward@acm.org

bzip2/libbzip2 version 1.0.8 of 13 July 2019

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

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

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and / or other materials provided with the distribution.
- Neither the name of Florian Reuter nor the names of its contributors may be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT OWNER OR CONTRIBUTORS BE LIABLE FOR ANY DIRECT, INDIRECT, INCIDENTAL, SPECIAL, EXEMPLARY, OR CONSEQUENTIAL DAMAGES (INCLUDING, BUT NOT LIMITED TO, PROCUREMENT OF SUBSTITUTE GOODS OR SERVICES; LOSS OF

USE, DATA, OR PROFITS; OR BUSINESS INTERRUPTION) HOWEVER CAUSED AND ON ANY THEORY OF LIABILITY, WHETHER IN CONTRACT, STRICT LIABILITY, OR TORT (INCLUDING NEGLIGENCE OR OTHERWISE) ARISING IN ANY WAY OUT OF THE USE OF THIS SOFTWARE, EVEN IF ADVISED OF THE POSSIBILITY OF SUCH DAMAGE.

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

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

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

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

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

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

- Regarding the decompression algorithm used for UNIX files:

Copyright © 1985, 1986, 1992, 1993

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

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

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

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

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

- Snowball is Copyright © 2001, Dr Martin Porter, Copyright © 2002, Richard Boulton. All rights reserved.

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

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

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

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

THIS SOFTWARE IS PROVIDED BY THE COPYRIGHT HOLDERS AND CONTRIBUTORS "AS IS" AND ANY EXPRESS OR IMPLIED WARRANTIES, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF MERCHANTABILITY AND FITNESS FOR A PARTICULAR PURPOSE ARE DISCLAIMED. IN NO EVENT SHALL THE COPYRIGHT HOLDER 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.

- Pako is Copyright © 2014–2017 by Vitaly Puzrin and Andrei Tuputcyn.

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 THE AUTHORS OR COPYRIGHT HOLDERS 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.

- HDF5 (Hierarchical Data Format 5) Software Library and Utilities are Copyright 2006–2015 by The HDF Group. NCSA HDF5 (Hierarchical Data Format 5) Software Library and Utilities Copyright 1998-2006 by the Board of Trustees of the University of Illinois. All rights reserved. DISCLAIMER: THIS SOFTWARE IS PROVIDED BY THE HDF GROUP AND THE CONTRIBUTORS “AS IS” WITH NO WARRANTY OF ANY KIND, EITHER EXPRESSED OR IMPLIED. In no event shall The HDF Group or the Contributors be liable for any damages suffered by the users arising out of the use of this software, even if advised of the possibility of such damage.
- agl-aglfn technology is Copyright © 2002, 2010, 2015 by Adobe Systems Incorporated. All Rights Reserved.

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

- Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
- Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
- Neither the name of Adobe Systems Incorporated 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 HOLDER 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.

- dmlc/xgboost is Copyright © 2019 SAS Institute.

Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- libzip is Copyright © 1999–2019 Dieter Baron and Thomas Klausner.

This file is part of libzip, a library to manipulate ZIP archives. The authors can be contacted at <libzip@nih.at>.

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

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. The names of the authors may not be used to endorse or promote products derived from this software without specific prior written permission.

THIS SOFTWARE IS PROVIDED BY THE AUTHORS “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 AUTHORS 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.

- OpenNLP 1.5.3, the pre-trained model (version 1.5 of en-parser-chunking.bin), and dmlc/xgboost Version .90 are licensed under the Apache License 2.0 are Copyright © January 2004 by Apache.org.

You may reproduce and distribute copies of the Work or Derivative Works thereof in any medium, with or without modifications, and in Source or Object form, provided that You meet the following conditions:

- You must give any other recipients of the Work or Derivative Works a copy of this License; and
 - You must cause any modified files to carry prominent notices stating that You changed the files; and
 - You must retain, in the Source form of any Derivative Works that You distribute, all copyright, patent, trademark, and attribution notices from the Source form of the Work, excluding those notices that do not pertain to any part of the Derivative Works; and
 - If the Work includes a “NOTICE” text file as part of its distribution, then any Derivative Works that You distribute must include a readable copy of the attribution notices contained within such NOTICE file, excluding those notices that do not pertain to any part of the Derivative Works, in at least one of the following places: within a NOTICE text file distributed as part of the Derivative Works; within the Source form or documentation, if provided along with the Derivative Works; or, within a display generated by the Derivative Works, if and wherever such third-party notices normally appear. The contents of the NOTICE file are for informational purposes only and do not modify the License. You may add Your own attribution notices within Derivative Works that You distribute, alongside or as an addendum to the NOTICE text from the Work, provided that such additional attribution notices cannot be construed as modifying the License.
 - You may add Your own copyright statement to Your modifications and may provide additional or different license terms and conditions for use, reproduction, or distribution of Your modifications, or for any such Derivative Works as a whole, provided Your use, reproduction, and distribution of the Work otherwise complies with the conditions stated in this License.
- LLVM is Copyright © 2003–2019 by the University of Illinois at Urbana-Champaign. Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at:
<http://www.apache.org/licenses/LICENSE-2.0>
Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS” BASIS, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.
 - clang is Copyright © 2007–2019 by the University of Illinois at Urbana-Champaign. Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at:
<http://www.apache.org/licenses/LICENSE-2.0>
Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS”, WITHOUT WARRANTIES OR CONDITIONS OF

ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- lld is Copyright © 2011–2019 by the University of Illinois at Urbana-Champaign.

Licensed under the Apache License, Version 2.0 (the “License”); you may not use this file except in compliance with the License. You may obtain a copy of the License at:

<http://www.apache.org/licenses/LICENSE-2.0>

Unless required by applicable law or agreed to in writing, software distributed under the License is distributed on an “AS IS”, WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied. See the License for the specific language governing permissions and limitations under the License.

- libcurl is Copyright © 1996–2021, Daniel Stenberg, daniel@haxx.se, and many contributors, see the THANKS file. All rights reserved.

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

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 OF THIRD PARTY RIGHTS. IN NO EVENT SHALL THE AUTHORS OR COPYRIGHT HOLDERS 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 a copyright holder shall not be used in advertising or otherwise to promote the sale, use or other dealings in this Software without prior written authorization of the copyright holder.

- On the Windows operating system, JMP utilizes the OpenBLAS library. OpenBLAS is licensed under the 3-clause BSD license. Full license text follows: Copyright © 2011-2015, The OpenBLAS Project

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

1. Redistributions of source code must retain the above copyright notice, this list of conditions and the following disclaimer.
2. Redistributions in binary form must reproduce the above copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution.
3. Neither the name of the OpenBLAS project 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 OPENBLAS PROJECT 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.