



Real Options Valuation

QDM – QUANTITATIVE DATA MINER

User Manual



Johnathan Mun, Ph.D., MBA, MS, BS, CFC, CRM, FRM, MIFC



QDM 2010

This manual and the software described in it are furnished under license and may only be used or copied in accordance with the terms of the end user license agreement. Information in this document is provided for informational purposes only, is subject to change without notice, and does not represent a commitment as to merchantability or fitness for a particular purpose by Real Options Valuation, Inc.

No part of this manual may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording, for any purpose without the express written permission of Real Options Valuation, Inc.

Content based on copyrighted publications by Dr. Johnathan Mun, Founder and CEO, Real Options Valuation, Inc.

Written by Dr. Johnathan Mun.

Written, designed, and published in the United States of America.

To purchase additional copies of this document, contact Real Options Valuation, Inc., at the e-mail address below:

Admin@RealOptionsValuation.com

or visit www.realoptionsvaluation.com

© 2010 by Dr. Johnathan Mun. All rights reserved.

Microsoft® is a registered trademark of Microsoft Corporation in the U.S. and other countries.

Other product names mentioned herein may be trademarks and/or registered trademarks of the respective holders.

TABLE OF CONTENTS

INTRODUCTION	7
<i>Installation Requirements and Licensing Procedures</i>	<i>7</i>
QUICK GETTING STARTED: HANDS-ON EXERCISES WITH ROV-QDM	8
<i>Modeling Tab.....</i>	<i>8</i>
<i>Analytics Tab</i>	<i>8</i>
<i>Forecasts Tab.....</i>	<i>8</i>
<i>Charts Tab</i>	<i>9</i>
<i>Simulation Tab.....</i>	<i>9</i>
<i>Data Linking, Variable Mapping, SQL Scripting</i>	<i>10</i>
<i>Variables Management.....</i>	<i>11</i>
OVERVIEW OF THE QDM PROCESS	12
USING THE QDM SOFTWARE	15
<i>DATA</i>	<i>15</i>
<i>Data Variable Mapping.....</i>	<i>18</i>
<i>TIP: Saving into CSV File Format</i>	<i>19</i>
<i>TIP: Variable Management.....</i>	<i>23</i>
<i>RELATIONSHIPS.....</i>	<i>24</i>
<i>MODELING.....</i>	<i>26</i>
<i>ANALYTICS</i>	<i>28</i>
<i>FORECASTS</i>	<i>31</i>
<i>CHARTS.....</i>	<i>36</i>
<i>SIMULATION</i>	<i>39</i>
<i>REPORTS</i>	<i>46</i>
<i>ROV VALUATOR</i>	<i>47</i>
<i>ROV OPTIMIZER.....</i>	<i>50</i>
<i>Linking to Other Databases</i>	<i>55</i>
<i>Case One: Link to Oracle.....</i>	<i>55</i>
<i>Case Two: Link to User DSN.....</i>	<i>55</i>
TECHNICAL APPENDICES	57
<i>Mathematical Probability Distributions.....</i>	<i>57</i>
<i>Probability Density Functions, Cumulative Distribution Functions, and Probability Mass Functions</i>	<i>57</i>
<i>Discrete Distributions.....</i>	<i>58</i>
<i>Bernoulli or Yes/No Distribution</i>	<i>58</i>
<i>Binomial Distribution.....</i>	<i>59</i>
<i>Discrete Uniform.....</i>	<i>59</i>
<i>Geometric Distribution</i>	<i>60</i>
<i>Hypergeometric Distribution</i>	<i>60</i>
<i>Negative Binomial Distribution.....</i>	<i>62</i>
<i>Poisson Distribution</i>	<i>62</i>
<i>Continuous Distributions</i>	<i>63</i>
<i>Beta Distribution.....</i>	<i>63</i>
<i>Cauchy Distribution or Lorentzian or Breit-Wigner Distribution</i>	<i>64</i>

<i>Chi-Square Distribution</i>	64
<i>Exponential Distribution</i>	65
<i>Extreme Value Distribution or Gumbel Distribution</i>	65
<i>F Distribution or Fisher-Snedecor Distribution</i>	66
<i>Gamma Distribution (Erlang Distribution)</i>	67
<i>Logistic Distribution</i>	68
<i>Lognormal Distribution</i>	68
<i>Normal Distribution</i>	69
<i>Pareto Distribution</i>	70
<i>Student's t-Distribution</i>	70
<i>Triangular Distribution</i>	71
<i>Uniform Distribution</i>	72
<i>Weibull Distribution (Rayleigh Distribution)</i>	72
QUICK TECHNICAL DESCRIPTIONS OF MODELS	74
<i>ANOVA: Randomized Blocks N-Treatments, 1-Factor N-Treatments, 2-Way ANOVA</i>	74
<i>ARIMA</i>	75
<i>Auto ARIMA</i>	76
<i>Autocorrelation and Partial Autocorrelation</i>	76
<i>Autoeconometrics (Quick) and Autoeconometrics (Detailed)</i>	78
<i>Basic Econometrics and Custom Econometrics</i>	80
<i>Control Charts: C, NP, P, R, U, XMR</i>	80
<i>C Chart</i>	82
<i>NP Chart</i>	82
<i>P Chart</i>	82
<i>R Chart</i>	82
<i>U Chart</i>	82
<i>XMR Chart</i>	82
<i>Correlation (Linear, Nonlinear)</i>	82
<i>Cubic Spline</i>	83
<i>Data Descriptive Statistics</i>	83
<i>Deseasonalizing</i>	84
<i>Distributional Fitting</i>	84
<i>Exponential J Curve</i>	85
<i>Heteroskedasticity</i>	85
<i>Limited Dependent Variables: Logit, Probit, Tobit</i>	86
<i>Linear Interpolation</i>	87
<i>Logistic S Curve</i>	88
<i>Markov Chain</i>	88
<i>Multiple Regression (Linear Regression and Nonlinear Regression)</i>	88
<i>Nonlinear Extrapolation</i>	89
<i>Nonparametric Hypothesis Tests</i>	90
<i>Chi-Square Goodness of Fit</i>	90
<i>Chi-Square Independence</i>	90
<i>Chi-Square Population Variance</i>	91
<i>Friedman's Test</i>	91
<i>Kruskal-Wallis Test</i>	91
<i>Lilliefors Test</i>	92
<i>Runs Test</i>	92
<i>Wilcoxon Signed-Rank (One Var)</i>	92

<i>Wilcoxon Signed-Rank (Two Var)</i>	92
<i>Parametric Hypothesis Tests</i>	93
<i>One Variable (T)</i>	94
<i>One Variable (Z)</i>	94
<i>One-Variable (Z) Proportion</i>	94
<i>Two-Variable (T) Dependent</i>	95
<i>Two-Variable (T) Independent Equal Variance</i>	95
<i>Two-Variable (T) Independent Unequal Variance</i>	96
<i>Two-Variable (Z) Independent Means</i>	97
<i>Two-Variable (Z) Independent Proportions</i>	98
<i>Two-Variable (F) Variances</i>	99
<i>Principal Component Analysis</i>	99
<i>R-Square Computation</i>	99
<i>Seasonality</i>	101
<i>Segmentation Clustering</i>	101
<i>Stepwise Regression (Backward)</i>	101
<i>Stepwise Regression (Correlation)</i>	101
<i>Stepwise Regression (Forward)</i>	101
<i>Stepwise Regression (Forward-Backward)</i>	102
<i>Stochastic Process Estimations</i>	102
<i>Brownian Motion Random Walk Process</i>	102
<i>Mean-Reversion Process</i>	103
<i>Jump-Diffusion Process</i>	103
<i>Jump-Diffusion Process with Mean Reversion</i>	103
<i>Structural Break</i>	104
<i>Time-Series Analysis</i>	104
<i>Time-Series Analysis (Auto)</i>	104
<i>Time-Series Analysis (DES)</i>	105
<i>Time-Series Analysis (DMA)</i>	106
<i>Time-Series Analysis (HWA)</i>	106
<i>Time-Series Analysis (HWM)</i>	106
<i>Time-Series Analysis (SA)</i>	106
<i>Time-Series Analysis (SES)</i>	107
<i>Time-Series Analysis (SM)</i>	107
<i>Time-Series Analysis (SMA)</i>	107
<i>Trending and Detrending</i>	107
<i>Trend Line (Difference Detrended)</i>	108
<i>Trend Line (Exponential Detrended)</i>	108
<i>Trend Line (Exponential)</i>	108
<i>Trend Line (Linear Detrended)</i>	108
<i>Trend Line (Linear)</i>	108
<i>Trend Line (Logarithmic Detrended)</i>	108
<i>Trend Line (Logarithmic)</i>	108
<i>Trend Line (Moving Average Detrended)</i>	108
<i>Trend Line (Moving Average)</i>	108
<i>Trend Line (Polynomial Detrended)</i>	108
<i>Trend Line (Polynomial)</i>	108
<i>Trend Line (Power Detrended)</i>	108
<i>Trend Line (Power)</i>	108
<i>Trend Line (Rate Detrended)</i>	108
<i>Trend Line (Static Mean Detrended)</i>	108

Trend Line (Static Median Detrended).....	108
Volatility: GARCH Models.....	108
Volatility: GARCH and TGARCH.....	110
Volatility: GARCH-M.....	110
Volatility: EGARCH and EGARCH-T.....	111
Volatility: GJR GARCH and GJR TGARCH.....	111
Volatility: TGARCH and TGARCH-M.....	111
Volatility: Log Returns Approach	112
Yield Curve (Bliss).....	112
Yield Curve (Nelson-Siegel).....	112
APPENDIX: DATABASE SQL USE CASES AND EXAMPLES.....	113
SQL Conditional Use Cases.....	114
Use Case 1: Selection of Rows by Value	115
Use Case 2: Use of 'AND'	116
Use Case 3: Use of 'OR'	117
Use Case 4: Use of 'AND' and 'OR' together.....	118
Use Case 5: Use of 'IN'.....	119
Use Case 6: Use of 'BETWEEN'.....	120
Use Case 7: Use of 'LIKE'.....	121
Use Case 8: Simple Math Functions.....	122
Use Case 9: Nested Math Functions.....	123
Use Case 10: Use of 'Union' to Connect Commands.....	124
Use Case 11: Filtering Different Value Types.....	125
Use Case 12: Choosing the Top N Rows	126
Use Case 13: Use of 'NOT IN'	127
Use Case 14: Use of 'EXISTS'	128
Use Case 15: Use of Multiple Table.....	129
Use Case 16: Example Using 'AND'	130
Use Case 17: Example Using Wildcards with 'AND'	131
Use Case 18: Example Using 'Union' with Sorting	132
Use Case 19: Example Using Wildcards and Math	133
Use Case 20: Example Using Nested 'AND/OR' with Math	134
Use Case 21: Use of 'AND'	135
Use Case 22: Use of SQL Functions	136
Use Case 23: Use of 'GROUP BY'.....	137
Use Case 24: Use of 'DISTINCT'.....	138
Use Case 25: Use of 'ORDER BY'	139
Use Case 26: Selection by Dates with 'BETWEEN'	140

INTRODUCTION

Welcome to the **ROV QUANTITATIVE DATA MINER (QDM)** Software, brought to you by Real Options Valuation, Inc. This software application is used for analytical data crunching and modeling. It runs in the Windows environment and can be used to link to databases to download and run large datasets at extreme high speeds. This software comes in three separate modules. The first module is the main ROV Quantitative Data Miner (QDM) with about 150 methods for running Data Modeling, Analytics, Forecasting, Simulation, Data Computation, and Charts. The second module is the ROV Optimizer for running static, dynamic, and stochastic optimization at high speeds on a large number of decision variables. The third module is the ROV Valuator, with over 600 closed-form, partial differential, lattice and analytical models. For a detailed list of these methods and models, review the List of Models file located on the [Start | Programs | Real Options Valuation | ROV Quantitative Data Miner](#) shortcut folder.

Installation Requirements and Licensing Procedures

Follow the on-screen instructions to install the software. The software's minimum requirements are:

- Dual core processor or later
- Windows XP, Vista or Windows 7 (MAC OS requires a Windows emulator such as Parallels or VM)
- 100MB free space and 1GB RAM minimum (2–4GB recommended)
- Administrative rights to install software

A permanent or trial license is required to run the software for the first time. To obtain a trial or full corporate license, contact Real Options Valuation, Inc., at admin@realoptionsvaluation.com or call (925) 271-4438 or visit our website at www.realoptionsvaluation.com. Visit this website and click on DOWNLOAD to obtain the latest software release, or click on the FAQ link to obtain any updated information on licensing or installation issues and fixes.

If you have installed the software and have purchased a full license to use it, you will need to e-mail us your 8-digit Hardware Fingerprint so that we can generate a license for you. Once installation is complete, start the QDM software by going to [Start | Programs | Real Options Valuation | ROV Quantitative Data Miner | ROV Quantitative Data Miner](#). When starting the software for the first time, you will be given the Hardware Fingerprint (Figure 1) of your computer and be asked for a Name and Key combination to run the software. E-mail us your 8-digit Hardware Fingerprint so that we can generate a name and key license combination for you that is unique to your computer.



Figure 1 – ROV QDM Hardware Fingerprint

QUICK GETTING STARTED: HANDS-ON EXERCISES WITH ROV-QDM

Typically, the fastest way to get started using new software is not to read a detailed user manual. The best and most effective method is to do a few hands-on exercises using predefined examples. Assuming you already have QDM software installed, start the program and do a quick run through of the following exercises, and you will be on your way to mastering this software in no time. After that, you can spend all the time you want on this user manual to dig deep into the analytical methods get the total value that QDM has in store.

Modeling Tab

For practice, follow the steps below:

1. Click on *File | Examples | 03 Modeling*, go to the *Modeling* tab, click on and select the third model, *Custom Econometrics*, and do the following:
 - a. Click *Compute* and review the results.
 - b. Review the input parameters:
 - i. Dependent Variable: $\text{LN}(\$Y\$)$
 - ii. Independent Variables: $\$(X1)\2 ; $\text{LAG}(\$X2\$,1)$; $\$(X3)\$*\$(X4)\$$; $\$(X5)\$$
 - c. Notice how the “ $\$()\$$ ” is used around a variable name, how “ $;$ ” is used to separate variables, and the mathematical operators that can be used (*, LAG, ^, and so forth).
 - d. Add an extra variable X4 to the end of the variables list. You can do this two ways:
 - i. Manually type in $\$(X4)\$$ after the end, making sure to include a semicolon before this new variable, or
 - ii. Double-click on the X4 variable from the Variables list grid on the left
2. Double-click on some of the other Saved Models. Review the inputs in each model and the results and notice that double-clicking an existing model actually runs it as well.
3. Create and run your own models and review the technical section for details of each approach.

Analytics Tab

For practice, follow the steps below:

1. Click on *File | Examples | 04 Analytics*, go to the *Analytics* tab, click on and select the sixth model, *Data Descriptives*, and do the following:
 - a. Click *Compute* and review the results.
 - b. Add an extra variable (VAR2) to run by double-clicking on VAR2 and hitting *Compute*.
 - c. Click on *Data Descriptives* again and you will notice that your new VAR2 variable is gone. To make this VAR2 variable permanent, make sure *Data Descriptives* is still selected and click on *Edit*, then type in your new variable... and remember to save the example file (*File | Save*). You can now click on another saved analytical model and then click back on *Data Descriptives* and you will see that the new variable is saved.
 - d. Create your own analytical model, give it a unique name, and click *Add* to add it to the list of saved analytics.

Forecasts Tab

For practice, follow the steps below:

1. Click on *File | Examples | 05 Forecasting*, go to the *Simulation* tab, click on and select the model, *Time-Series Auto*, and do the following:
 - a. Click *Compute* and review the results in the Chart, Data, and Statistics subtabs.

- i. Identify which line corresponds to the historical data and which line is the back-fitting and forecast prediction.
 - ii. Copy or extract the data to Excel or some other software (Word, PowerPoint).
 2. Double-click on the *Detrend* models (choose any one of them) and identify what is going on here and what methodology is applied.
 3. Double-click on *Stochastic – GBM* for the geometric Brownian motion stochastic process.
 - a. Double-click on the same model a few additional times and notice what happens each time, what changes, and why.
 - b. Repeat the process but this time focus on the Data results subtab and notice how the results change each time.

Charts Tab

For practice, follow the steps below:

1. Click on *File | Examples | 06 Charts*, go to the *Charts* tab, click on and select the chart, *P Chart*, and do the following:
 - a. Click *Update* to view the *P-Control* chart (see the Quick Technical Discussions section for details of control chart types and what they represent).
 - b. Click on the *Chart Type* droplist and select other charts to view.
2. Create a basic new chart by doing the following:
 - a. Click on the *last empty row* in the *Created Charts* grid list.
 - b. Select the chart type, for example, *Standard 2D Line with Points*.
 - c. Select the Variable, for example, *Defective Units*.
 - d. Click *Add* and give it a name such as “*Test*”.
 - e. Double-click on *Test* (or the newly created chart name).
 - f. Select a different type of chart and see the results.
 - g. Play with some of the chart power tools by clicking on some of the icons and seeing the effects of each icon.
3. Create an overlay chart of several variables by doing the following:
 - a. Click on the *last empty row* in the *Created Charts* grid list.
 - b. Select the chart type, for example, *Standard 2D Bar*.
 - c. Select several Variables by holding down the CTRL key on your keyboard and clicking on several variables, for example, *Measurement 1*, *Measurement 2*, *Measurement 3*.
 - d. Click *Add* and give it a name such as “*Test 2*”.
 - e. Double-click on *Test 2* (or the newly created chart name).
 - f. Select a different type of chart and see the results.
4. Create a new Control Chart type by first reviewing the Quick Technical Discussions section of this manual to understand what the required inputs are for each type of control chart and replicating the steps above.

Simulation Tab

For practice, follow the steps below:

1. Click on *File | Examples | 07 Simulation*, go to the *Simulation* tab, click on and select the first model, *Addition*, and do the following:
 - a. Click on *Run Simulation* or *F9* on the keyboard.
 - b. Review each of the results subtabs: *Chart*, *Statistics*, *Chart Data*, and *Simulation Data*.
2. Select the *Addition* model and click *Run Simulation*, then do the following:
 - a. Double-click on the *Certainty* input box or select the default 100 value and type in 90 to compute the two-tailed 90% confidence interval on the simulated results. Try other inputs as well as long as the input values are between 0 and 100.
 - b. Select a different tail confidence such as Left Tail or Right Tail and enter in a certainty % value between 0% and 100%; explain what the results represent.

- c. Select a different tail confidence such as Left Tail or Right Tail and this time enter in the value to obtain the certainty percentile; explain what the results represent.
3. Select the *Brownian Motion* model and *Run Simulation*, then do the following:
 - a. Rerun the simulation a few times and pay attention to the results (e.g., look at the Statistics subtab and see what happens to the results each time a new simulation is run on the same model).
 - b. This time, check the box beside Enable Seed Value, rerun the simulation a few times, and notice what happens to the results (e.g., the results in the Statistics subtab); explain what happened and what seed values do to the analysis.
 - c. Notice that in the Brownian motion stochastic process, there are multiple steps in the process and each step has its own forecast chart. So in this case of multiple charts, you will see a new *Select Charts to Show* droplist where you can select the specific charts to show or to show all charts at once.
4. Select the *Normal (50, 5)* model and click *F9* or *Run Simulation* then do the following:
 - a. Select different chart types and explain what each of these mean:
 - i. Histogram
 - ii. Fitting and Histogram
 - iii. CDF and Histogram
 - iv. PDF and Histogram
 - v. Cumulative (CDF)
 - vi. Probability (PDF)
 - vii. Multiple CDF Overlay
 - viii. Multiple PDF Overlay
 - b. Change the number of *Bins* and *Decimals* to show on the chart and click *Run Simulation* to update the chart.
 - c. Change the *Bar Type* and *Bar Color* as well as the *Line Color* on the chart.
 - d. Change some of the advanced settings on the chart by using some of the chart power tool icons.

Data Linking, Variable Mapping, SQL Scripting

To practice learning how to set up new Groups and mapping new Variables or linking these variables to databases, follow the steps below. You can visualize Variables as individual columns of data, whereas Groups are simply a collection of variables.

1. In QDM, click on *File | New* to start a new example file.
 - a. Click on the menu *Variables | Group Management*, click *Add* on the left panel (*Existing Groups*) to create a new variable group (give it a name, e.g., *Main Group*, and a short description).
 - b. Click *Add* on the right panel (*Variables in Selected Group*).
 - c. Here you are provided a choice of 5 data input and linking methods. Click on each, one at a time, click *Next* to see how each method works, and click *Back* when done.
 - i. Data Link – allows you to link to existing databases (SQL, Oracle, and other ODBC compliant databases) and data files (CSV, text files, Excel files).
 - ii. Manual Input – you can manually enter in data, paste in data, or open text files with existing data.
 - iii. Data Compute – using other existing variables (e.g., other variables that already exist from data links, manual inputs, or other methods), you can perform mathematical and analytical computations to create a new variable.
 - iv. Set Simulation Assumption – use this to set simulation assumptions to run risk simulations.
 - v. Data Fitting – using raw data, perform statistical fitting to existing distributions to find the best fit, and use this best-fitting result as the input distribution for the purposes of running risk simulations.

- d. Let's try the Data Link approach to create a new variable:
 - i. Select the first option *Data Link* and click *Next*.
 - ii. Enter in a name for the new variable, for example, Variable X.
 - iii. In the *ODBC DSN Type*, click on the droplist and select *Connect to Excel*.
 - iv. Click *Open*, browse to the install path's Examples folder (e.g., c:\Program Files\Real Options Valuation\ROV Quantitative Data Miner\Examples), and select the file *Sample Data 1.xls*.
 - v. In the *Available Fields* input box on the left, under the *Driver = Microsoft Excel*, double-click on *Sheet1\$* to open the worksheet called Sheet 1.
 - vi. Either double-click on *Number* or single-click on *Number* and click on the >> button to select this variable.
 - vii. If you wish, you can enter in some SQL command in the *Condition* box or just click *Finish* to map this variable and click *OK*. For additional practice, you can come back to this step later and enter in some SQL commands by reviewing and following the example SQL use cases at the end of this user manual.
 - viii. Back in the main QDM interface, make sure you are in the Data tab. Here you can see the data linked into the grid.

Variables Management

For practice, follow the steps below:

1. In QDM, click on *File | Examples | 08 SQL on Data Mapping*.
 - a. While working through these examples, it is suggested that you also review, in parallel, the Appendix on SQL Conditional Use Cases. Note that the existing variables in this example model use the same SQL queries listed in that Appendix.
 - b. Click on *Variables | Group Management*, and in the Variables in the Selected Group grid, click on and select the "DL_GTE 100" (scroll down the list to select the last variable on the list) and click *EDIT | YES*.
 - c. Keep the default *Data Link* selection and click *Next*.
 - d. Review the Selected Fields on the right panel, and notice that it is a link to an Excel file (the name and location of the Excel file as well as the worksheet name and variable header Number). Also notice the SQL Condition statement at the bottom, of "Number > 100" indicating that the variable "DL_GTE 100" is from the Number variable and will return only all values that are greater than 100.
 - e. Click *Finish | OK* when done. Then go back to the main *DATA* tab and review the first column variable "DL_GTE 100" and notice that all the values are indeed greater than 100.
 - f. Review Use Case 1 from the Appendix on SQL Conditional Statements to understand the approach that was implemented in querying this variable.
2. Repeat with the rest of variables if desired, and each time match the variable SQL query with the use cases in the Appendix on SQL Conditional Statements and the resulting variable values in the Data tab.

OVERVIEW OF THE QDM PROCESS

Historical, contemporaneous, and predicted quantitative and qualitative data abound in the business world, impact business decisions and, ultimately, affect the profitability and survival of the corporation. Real Options Valuation, Inc.'s (ROV) Quantitative Data Miner (QDM) software incorporates multiple advanced analytical techniques and algorithms and compiles them in such a unique and novel way as to facilitate business decision and data analysis. It does so through an intelligent set of statistical and analytical tests and models to analyze and extract information that otherwise cannot be obtained manually. That is, instead of requiring the user to understand advanced statistics, financial modeling, and mathematics in order to know what analysis to run on some existing data or to have the ability to interpret the raw numerical results, this software runs the relevant analyses in an integrated stepwise process and provides detailed description in its reports, coupled with the numerical results and charts for easy interpretation.

The benefits of this software are many: a comprehensive set of advanced analytics for the purposes of forecasting, analyzing, and modeling datasets, linking to and from small or large databases, running complex analytical methods and algorithms at super speeds, creating and modeling portfolios and optimization of portfolio selections, and running hundreds of valuation models all in one place.

The QDM software can be used in a variety of settings and is not restricted to any specific industry or domain-specific application, as long as there exist data (QDM can handle small or large datasets) and multidimensional variables where the user is interested in modeling the relationships among these variables for the purposes of prediction and forecasting, as well as to understand the structure of the relationships and to obtain actionable business intelligence of a business or operation. For instance, in a corporation, one can run the revenues, profits, earnings per share, stock prices, and other financial variables against economic variables such as industrial production (there are hundreds of such industrial production variables in the United States), gross domestic product, inflation rates, unemployment rates, prices of commodities, interest rates, competitive environment, general pricing structure, market size, and so forth. The question is which of these economic variables can be used to determine the effects on the company's profitability, or perhaps provide leading indicators as to when a structural shift will occur (e.g., when a profitability downturn will occur ahead of or at some time lag after the economic downturn) so as to better prepare for these events. In the manufacturing sector, a company can forecast the total sales or model the total cost of manufacturing a complex product with many inputs to better determine the pricing structure for its customers in order to maximize its profits. In this scenario, the total cost or profit of a specific product is modeled using the price and cost of all its individual inputs. Many other such applications exist and can be similarly modeled using QDM.

This software comes in three separate modules. The first module is the main ROV Quantitative Data Miner (QDM) with about 150 methods for running Modeling, Analytics, Forecasting, Simulation, Data Computation, and Charts. The second module is the ROV Optimizer for running static, dynamic, and stochastic optimization at high speeds on a large number of decision variables. The third module is the ROV Valuator, with over 600 closed-form, partial differential, lattice and analytical models. For a detailed list of these methods and models, review the List of Models file located on the [Start](#) | [Programs](#) | [Real Options Valuation](#) | [ROV Quantitative Data Miner](#) shortcut folder.

The QDM software can be used one model or analytical method at a time, or an analytical modeling sequence can be executed. For the sake of introducing this software application, we will look at the entire sequence of events through this process overview, while in later sections of this user manual we will go through each of the methodologies and applications. In most cases, the typical user will load in the raw data through either manual input or database uploads and links, and selects one or several models and analytics to run on the data, without going through the entire data mining process.

Figure 2 illustrates the QDM (Quantitative Data Miner) process and method undertaken in the ROV QDM software. Starting with the user's own raw dataset [001], which includes dependent and independent variables—dependent variables are those variables we would like to predict (such as revenues, profits, stock prices, cost, price, etc.) using the independent variables (such as economic indices, quantity of raw materials, interest rates, etc.). Through any one of the five methods [002] explained in the subsequent sections, such as database and data file linking, manual data inputs, data computation, running Monte Carlo simulation, or distributional fitting, the initial dataset is obtained [003]. From this initial raw dataset, new variables or new groupings of the dataset [004] can be developed. For instance, new variables can be created using the existing dataset or the initial dataset can be sliced into different groups or chunks of data. This initial dataset together with the newly developed variables or groups will then be considered the secondary or expanded dataset [005]. From here, the process is to perform a quick preliminary relationship screening using correlation analysis [006] to determine which variables are not significant and do not correlate to the dependent variable. Then those variables that are statistically significant in the analysis will be saved as the filtered dataset [007], and advanced analytics are applied [008] on this tertiary filtered dataset. The advanced analytical techniques come in a variety of types [009] including applications in autoeconometrics (AE), multiple regression (MR), stepwise multiple regression (SWMR), custom econometric modeling (CEM), and many other techniques. These advanced analytics can be developed outside of the existing software and linked in to run in this step within the process. Based on the analytics, newly created variables [010] may sometimes be required and will be generated, and these new variables together with the previous tertiary filtered dataset constitute the final dataset [011] where additional modeling [012] can be run. These modeling approaches are discussed in more detail in subsequent sections. Using these relationship modeling approaches, forecasts [013] for the dependent variables can then be created, and charts [014] are run to show the relationships and forecast behavior of the variables. Detailed reports can be generated complete with the technical details of the modeling and analytics, and the data from the final dataset [015] can be extracted for further analysis in the future or in another software environment. Monte Carlo risk simulation [016] can then be run on the forecast predictions and additional reports and simulated data can be extracted [017].

The overview of the process can be summed up as a series of data mining processes. That is, starting with the original raw dataset with many variables, we perform multiple series of analytics in sequence, and after the results are obtained from each set of analytics, the data variables are filtered down or new variables that are more statistically significant are created in its place, and the process continues through multiple iterations until the final set of variables that are the most applicable remain. Using this final set of variables, the user can then perform a variety of actions, including forecasting and prediction, modeling the relationships among these variables, and so forth, and along each step of the process, data can be extracted, charts can be developed, and detailed analytical results are available if required.

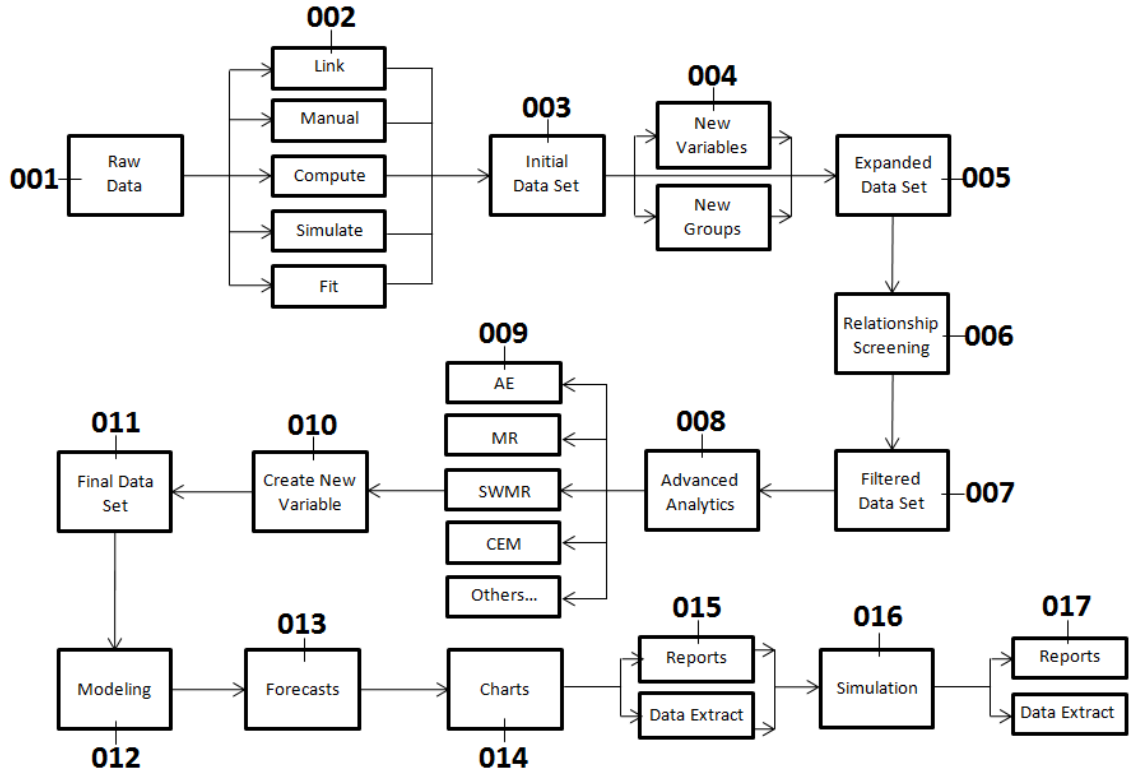


Figure 2 – ROV QDM Analytical Process

USING THE QDM SOFTWARE

The QDM software comes in a single easy-to-use user interface with multiple tabs and subtabs. The following text describes the functionality of each tab in the QDM software. You can also review some of the Example profiles available in the [FILE | EXAMPLES](#) menu item for predefined models and datasets.

Figure 3 illustrates the QDM software with the typical software menu items. In the case of QDM, the [FILE](#) menu allows users to create a new file, open an existing file, save the current file, save the current file as a different name, open predefined example files, or exit the QDM software. The [VARIABLES](#) menu allows users to access and manage existing groups and variables, or to create new groups or variables through one of the five methods discussed in more detail in later sections. The [TOOLS](#) menu launches the other two modules in QDM, namely, the ROV Optimizer and ROV Valuator. The [LANGUAGE](#) menu allows users to change among various translated languages such as English, Chinese, Japanese, Spanish, Portuguese, Italian, German, Korean, French, and others, where the QDM software will refresh its user interface with these new languages by reading a previously translated file. Finally, the [HELP](#) menu runs the user manual to the QDM software.

DATA

The first tab in QDM is the [DATA](#) tab. The first step in this tab is to load the user raw dataset by clicking on the [MAP VARIABLES](#) button, which will invoke the option to select one of five data loading methods (illustrated by figures 4 through 8), and from these loaded variables, the single dependent variable will have to be selected. In step three, creating the new variables is optional, that is, the user can skip this step entirely. However, sometimes, time-series data (i.e., data that follows time in a series, such as revenues for January, February, March, and so forth) require some additional analysis. A variable can be lagged some time periods; for instance, the entire time series is moved down a specific time period (e.g., the January revenue will be shifted down one period to match the February time period). So if we lagged the variable from 1 to 12 periods, we create 12 new variables, or lag one period, lag two periods, and all the way to lag twelve periods. Leading a variable goes the opposite way, that is, the January value now becomes last December's value, and so forth. The Time Index is simply a value starting at 1 and goes sequentially down all the way to the last row N (i.e., 1, 2, 3, ..., N), and the name of the variable will be set as Time, for future modeling use.

Next, we have three versions of Period to Period calculations, that is, the difference from one period to the next. For instance, suppose that the following data exist: 100, 120, 110...[then the period to period change every 1 period will be N/A, 20, -10; the period to period % change will be N/A, 20%, -8.33%; and the period to period relative return 031 will be N/A, 1.2000, 0.9167. Of course, this example shows a difference of 1 period, whereas the user can set as many periods back as required. So, for example, 12 periods means the calculation is the same as shown but instead of using the data of one period back, it computes the data back 12 periods and creates a new variable (it does not create 12 new variables, just one). The user can also create his or her own [CUSTOM VARIABLES](#), which is the same as clicking on the Map Variables in the first step, which will invoke an option to choose any one of the five methods to link or compute a new variable as shown in figures four through 8.

In step four, users can now slice or group the data into various times if required. This step is completely optional and will not be run if “Do not slice dataset” is selected. However, if “Group all rows but shift down M Periods N Times” is selected—for example, 3 and 3 for *M* and *N*, respectively—then the dataset, say there are 100 periods or rows, will then be grouped as follows: 1-100, 4-100, 7-100. Thus, three different new groups are created and the new variables’ names will have the same name as the original variable with a suffix of “GMSDNT,” where the acronym stands for “Group M Shift Down N Times” and *M* and *N* will be replaced as required, from 1 to 3 for *N* and 3 for *M*. Next, if “Group all rows but shift down M Periods N Times” is selected and let’s say the original data is again 1-100 rows, new variables of 1-12, 13-24, 25-36 will be created, and the new variable names will bear the same name as the original plus the aforementioned suffix of “GMRSDNT.” Finally, users can create their own custom groups by entering the rows to group, for example, 1-12, 1-24, 49-100 and so forth, separated either by commas or semicolons. From this point onwards, the dataset or chart can be seen in a grid of values or visualized as charts. When in the data grid mode, certain basic functions, such as the number of decimals to show and whether all data rows are shown (which will take a longer time to update if the dataset is very large) or only the first *N* rows (for faster data grid updating), and the initial dataset, including the groupings and new variables, can be extracted to Microsoft Excel for further analysis or to a flat text file for uploading into other databases and software applications. Clicking on the next button takes the user to the next tab as illustrated in Figure 9.

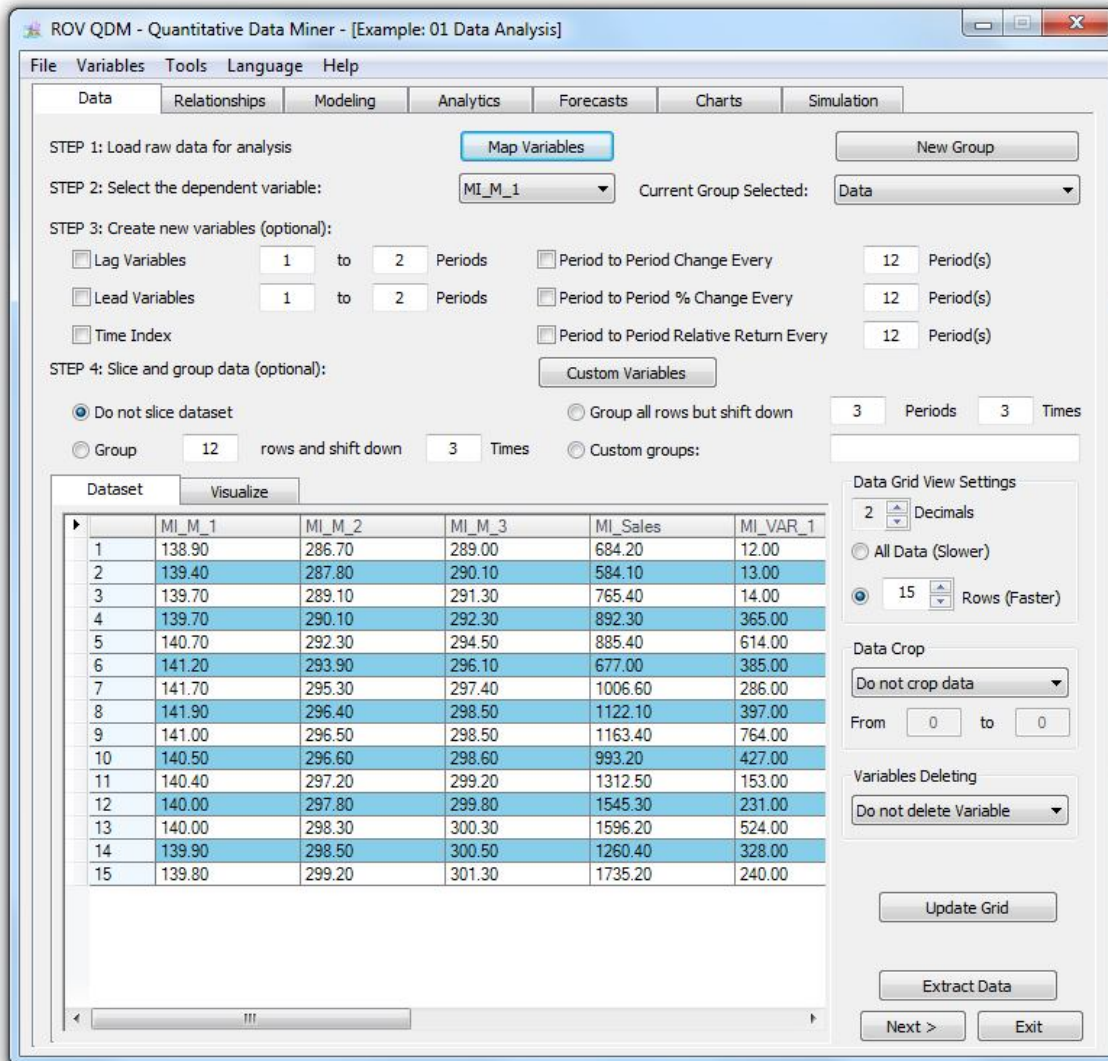


Figure 3 – ROV QDM Data Tab

Data Variable Mapping

Figure 4 illustrates how the various input parameters can be mapped to existing data. The icon graphic shows the five ways data can be obtained. These input methods include data link (linking to existing data files, databases, and other proprietary data sources); manual input (data are typed in or pasted in directly); data compute (existing data variables are first modified and analyzed before entering them as input variables); set assumption (creating any of the twenty-four statistical distributions to run simulations on); or model fitting (using existing raw data to find the best-fitting distribution assumption for simulation). Once one of these input methods is selected, the next step will allow the user to provide more details on the location of the data or its characteristics.

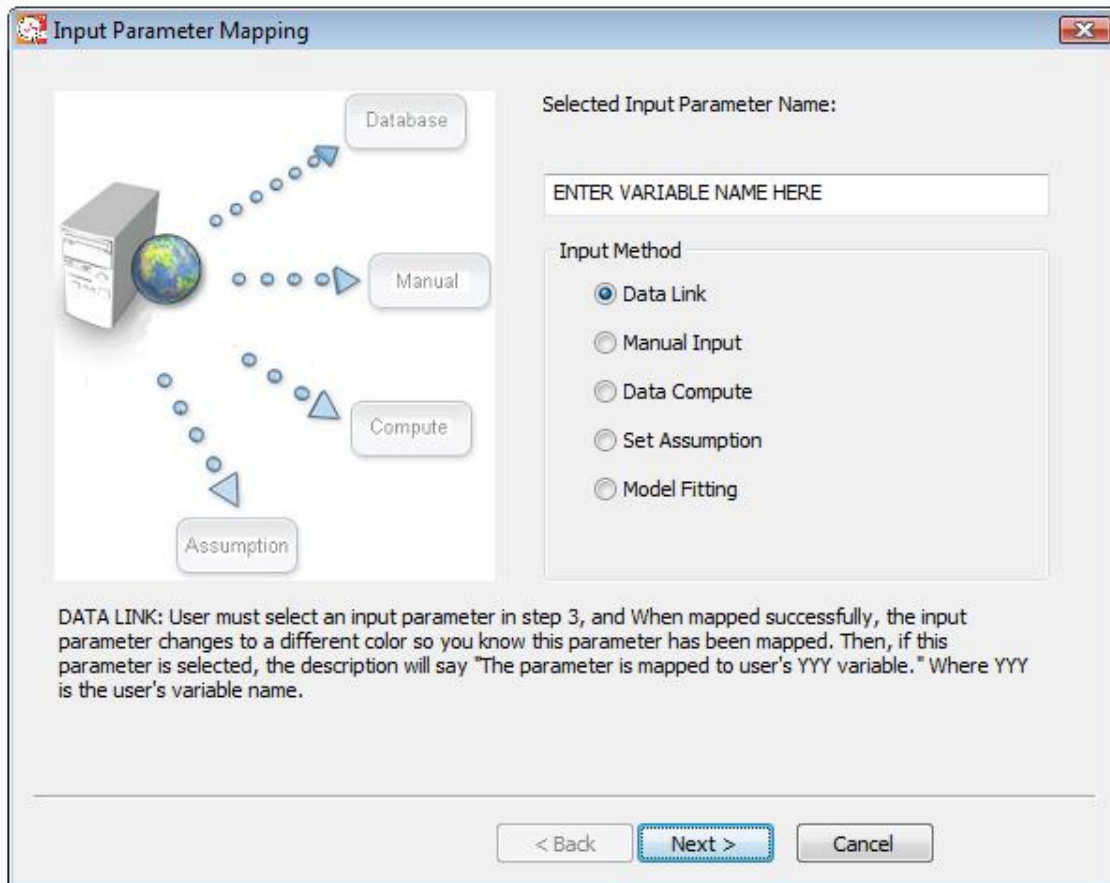


Figure 4 – ROV QDM Data Mapping

Figure 5 illustrates the data link process, and this method links to various databases and data types such as Excel, Oracle financial data model, SQL servers, flat files and other user-specific data files, databases, and file locations, where an existing database, data file, or data table can be opened to view the available fields or variables. For instance, if the data files have several worksheets or variable columns, the relevant variables can be selected and added (>>) or removed (<<) from the list of selected fields. The data can then be filtered using conditional SQL statements (see the Appendix on SQL Conditional Statements for examples of using SQL statements and queries in QDM). Clicking the OK button will create a new variable or list of variables depending on how many fields are selected in this method.

TIP: Saving into CSV File Format

As the CSV file format is the most commonly used format (this file format is compatible with most databases as a means to upload or download data), here are some getting started notes on creating the CSV data file. It is always advisable to change a flat text data file into CSV as it has more features and the data can be viewed quickly and easily. To convert a text file to CSV, within Excel, do a File, Open and open the text file (go through the data file filter with space or tab delimited). Then Save As the file into CSV Comma Delimited.

When manipulating CSV files, make sure that you do not add rows or values or type in data at the bottom (after the end of the dataset) because whatever happens at the bottom of the CSV file is saved even if you have deleted the cell values. If you have done some computations at the bottom, select the rows and perform a DELETE ROW(s) to eliminate all residual items that will be saved in the CSV (because deleted cells are assumed to contain empty values). Doing the DELETE ROW is critical otherwise the SQL upload will include empty elements and the computed values might be incorrect.

It is also good practice that the first row of the data has the Variable name. Note that Variable names can have spaces and special characters for Risk Modeler to work. Nonetheless, in some other databases, special characters and spaces might not be allowed and you need to be aware of this limitation when creating your dataset. Therefore, it is always safer to not add spaces and special characters as variable names (e.g., do not use things such as @, %, #, &, / and so forth).

If the first row of data has an integer value (e.g., 0, 1, 2, etc.), then make sure that it has decimals associated with the value. Sometimes certain databases using MySQL and SQL scripts may identify that as a string instead of a value. It is always a good idea to double-check this. You can always change the number of decimals in CSV when editing in Excel. Just add a few decimals simply as a precaution for database manipulation when uploading and downloading files.

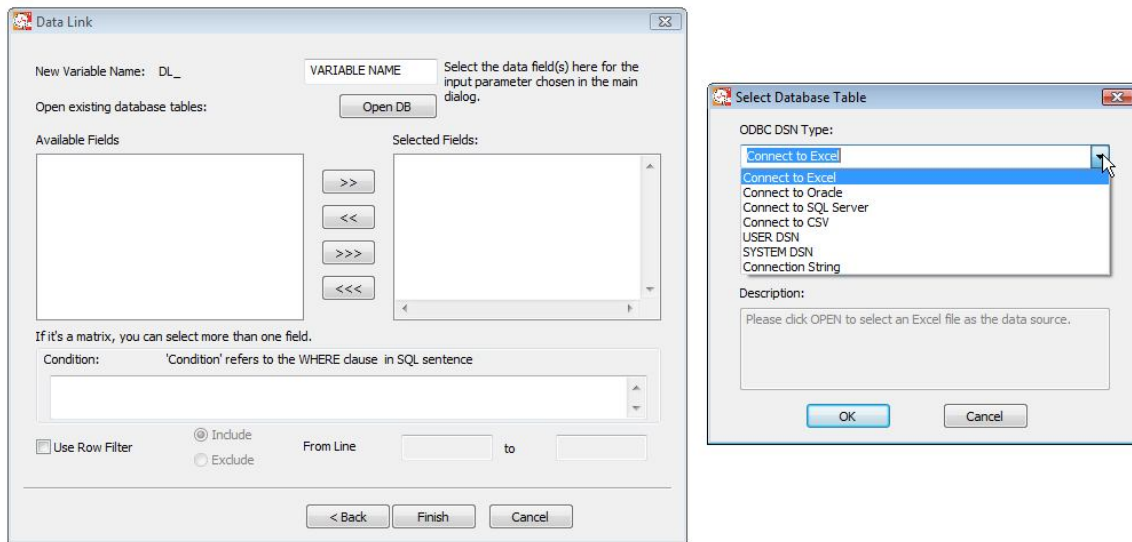


Figure 5 – ROV QDM Data Linking

Figure 6 illustrates the manual input process method where data can be entered or pasted in manually as a matrix, array, or sequence. Users can enter in a unique variable name for the new data variable and select if a single value is replicated for every record in the variable, or whether unique data is uploaded from a flat data file or manually typed in or pasted into a textbox. The text file data upload and clipboard paste functions are also available, and once the OK button is clicked, the new variable will be created.

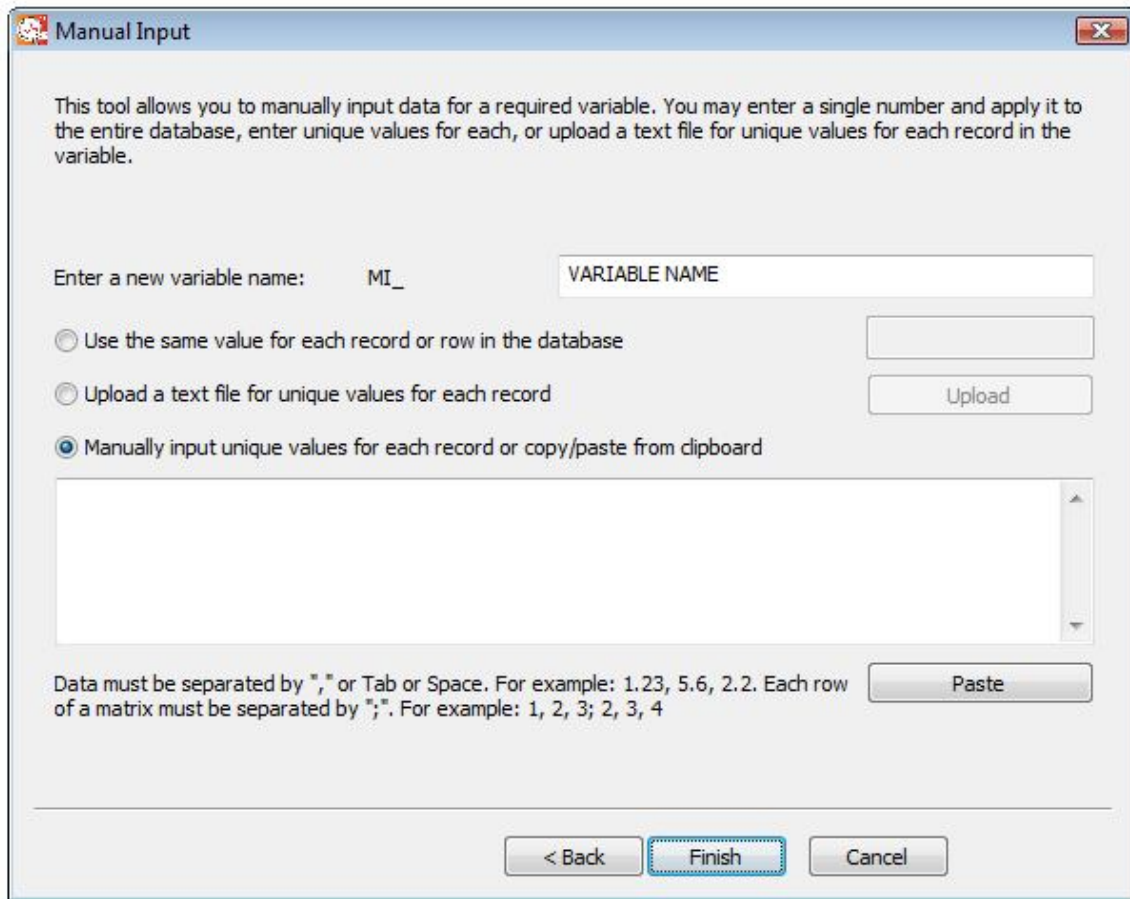


Figure 6 – ROV QDM Manual Data Input

Figure 7 illustrates the data computation method process in which a new variable is created that can be made to depend on existing variables through some computation created by entering some numerical expression. This data computation method can parse mathematical functions as illustrated in this figure, including multiple mathematical, statistical, and financial functions, and applied to numerical inputs typed in directly or using existing data variables and a numerical and functional keypad. When the OK button is clicked, the new variable will be created.

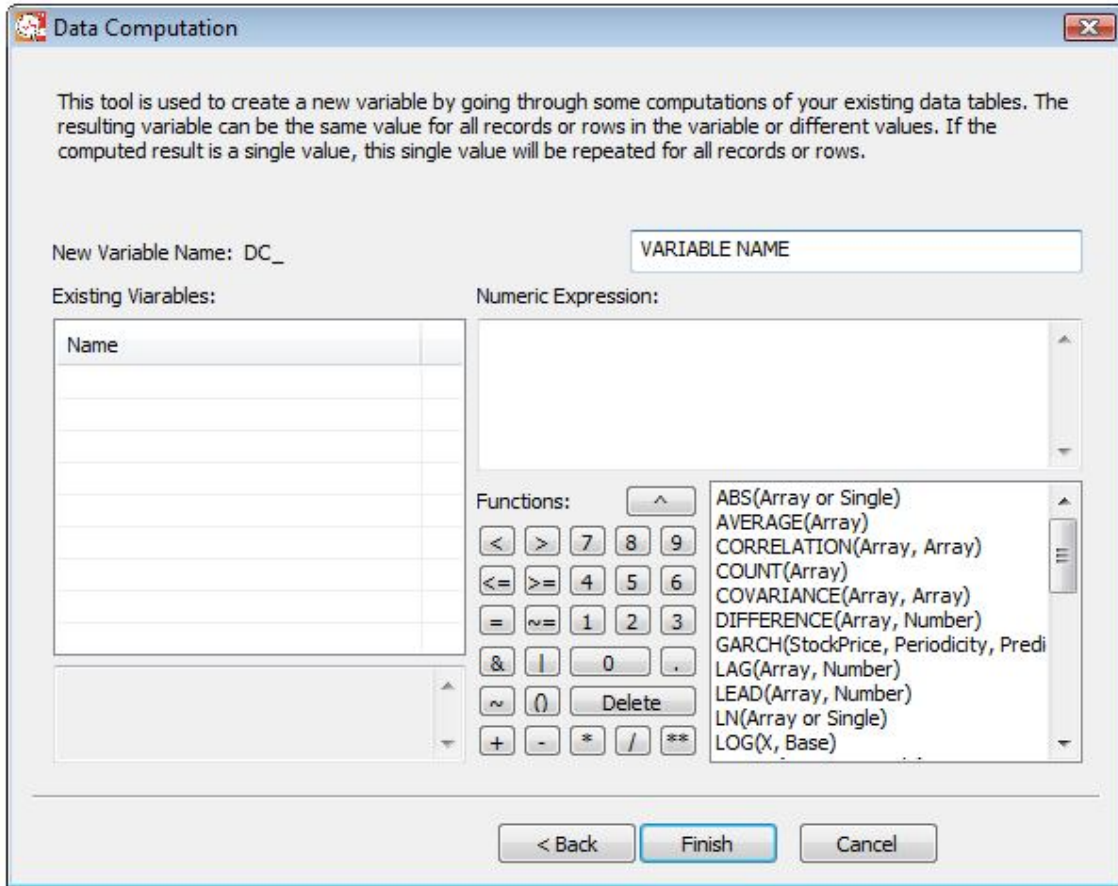


Figure 7 – ROV QDM Data Compute

Figure 8 illustrates the set simulation assumptions method on a new variable, where, when no data points exist or when the variable is known to follow some prescribed statistical and mathematical distribution, can be set and a simulation of thousands to millions of values can be generated. Depending on the distribution selected, different input parameters will be required (see the later section on Mathematical Probability Distributions for the technical details). When the OK button is clicked, the new variable will be created.

The following lists the Data Compute methods currently supported in QDM:

- Absolute Values
- Average
- Correlation
- Count
- Covariance
- Difference
- GARCH
- Lag
- Lead
- LN
- Log

- Max
- Median
- Min
- Mode
- Power
- Rank Ascending
- Rank Descending
- Relative Returns
- Relative LN Returns
- Semi-Standard Deviation (Upper)
- Semi-Standard Deviation (Lower)
- Standard Deviation (Sample)
- Standard Deviation (Population)
- Sum
- Variance (Sample)
- Variance (Population)
- Volatility

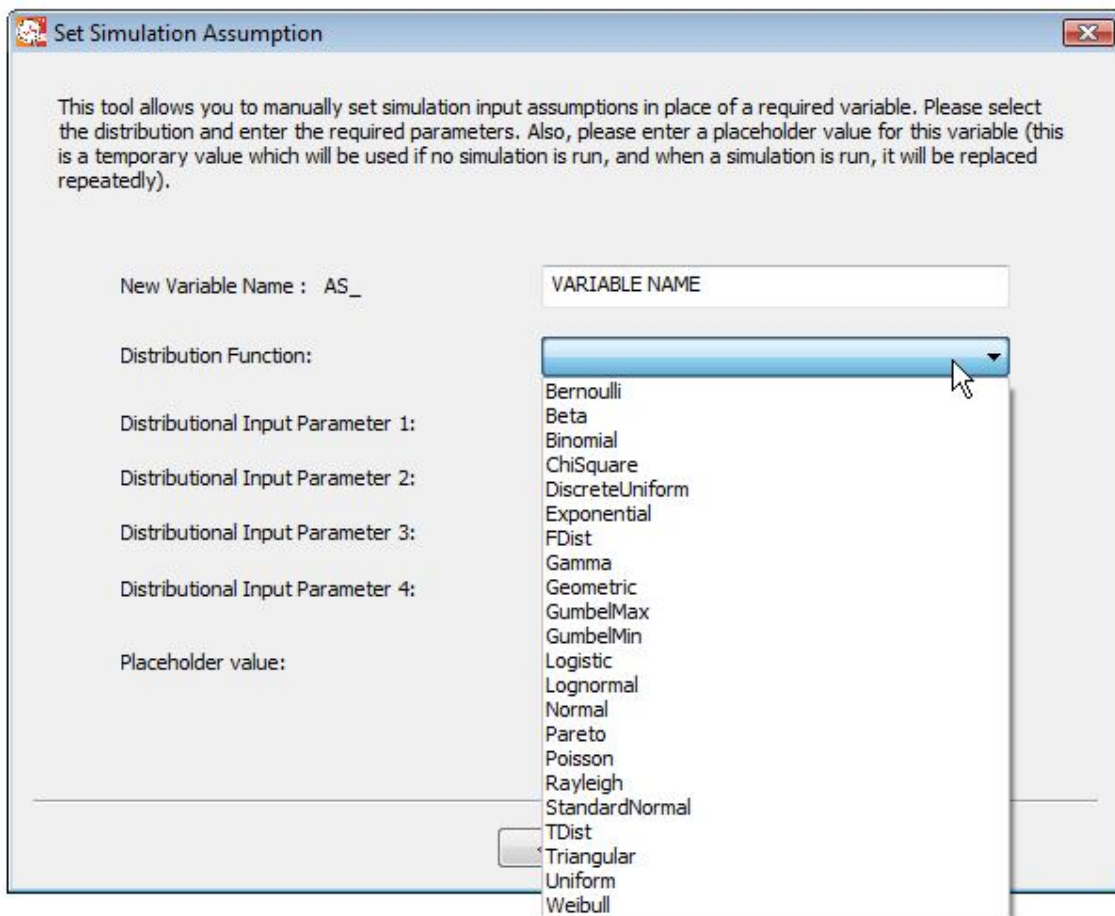


Figure 8 – ROV QDM Set Simulation Assumption

Figure 9 illustrates the data-fitting method where thousands of existing data points can be fitted to a distributional assumption such that Monte Carlo simulations can be run on this variable. To create this new variable, the user has to determine whether to use a continuous or discrete distribution (continuous variables take on any value, such as 1.2534, -102.24, and so forth, whereas discrete variables can take on only integer values such as 1, 200, -5, and so forth). The user then selects whether the data already exists in a database field location and selects the relevant field from the database, or to upload from an existing text file by clicking on the upload button or manually input the data values by pasting directly into the text box or clicking on the paste button. When the OK button is clicked, the new variable will be created.

One important note is that for any of these five methods of data linking, the approaches can be interconnected. For instance, the user can first use the database linking to create a new variable, and then perform some computations on this existing variable to create a new variable, which will then be used in the data-fitting routines, and so forth. This interchangeability approach provides the user with significant amounts of flexibility in generating new variables as required.

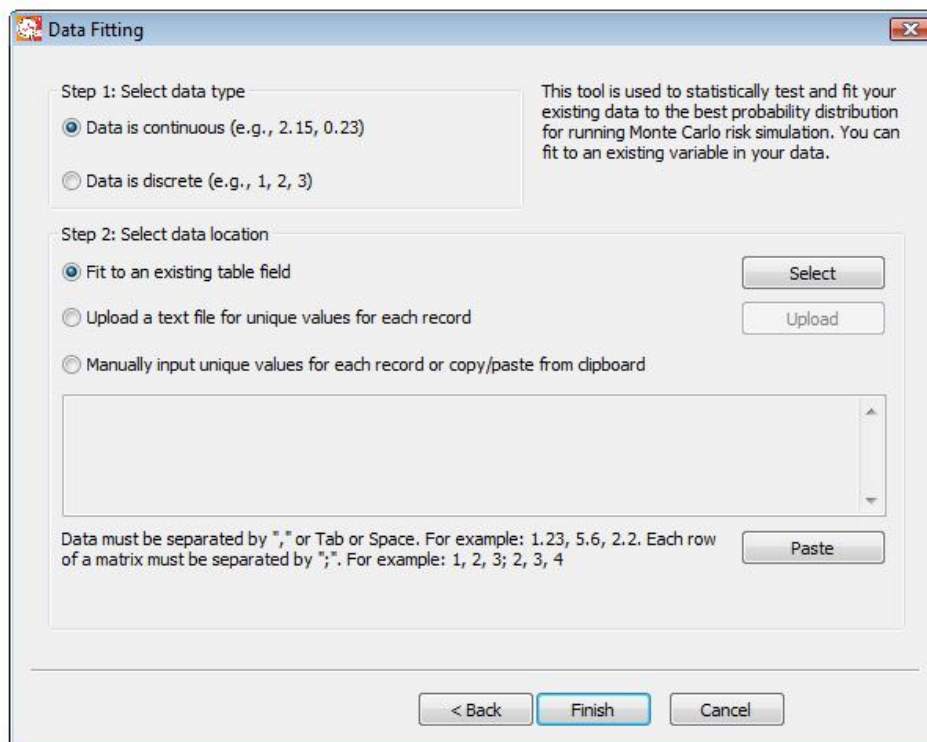


Figure 9 – ROV QDM Data Fitting

TIP: Variable Management

As a power user of the ROV Modeler software, the Variable Management tool is indispensable. You can click on the menu item Variable and select Variable Management to show the list of previously mapped variables. Using this approach, you can Add, Edit, or Delete any existing variables. The power of this variable management is evident in the Data Compute examples above, where you can link in as many variables as you wish from a dataset or database and then perform subsequent manipulations as desired. By using this combination of data linking, data variable management, and data compute, you can essentially control the sequence of events and manipulate the data as required, before they are used in the model.

RELATIONSHIPS

Figure 10 illustrates the second modeling tab, **RELATIONSHIPS**. It is used to reveal the relationships among each and every one of the independent variables, including the newly created and newly grouped variables, to the dependent variable. The independent variables are then presented individually, with the variable names as well as the resulting linear and nonlinear correlations and R-Squared values with respect to the dependent variable (the subsequent sections in this document detail the multiple regression analysis and correlation approach used in the software). The dependent variable itself is not presented in this results grid. The results grid allows the user to select only those specific variables that show the best relationship in order to filter out the erroneous independent variables with little to no relationship to the dependent variable for the next analysis step. The results in the grid can be sorted by linear or nonlinear correlations (from high to low descending or low to high ascending order), and the selection filter can be automated to pick the top N variables, or all variables can be selected at once, or variables can be selected manually. Further, the computed R-Square values can be obtained one of two ways: using a simple linear correlation's results and squaring it (this approach can be computed quickly) or using a regression analysis to obtain the R-Square (this approach takes a slightly longer computation time as more detailed calculations are required). To facilitate viewing the results in the grid, users can optionally elect to highlight variables with correlations above a specific absolute value cutoff point, or R-Squares above a specific value. Finally, specific data groups that were created in the previous Data tab will be listed here and available for selection. Depending on the group(s) selected, where one or more groups can be selected at once or all groups can be selected (+) or deselected (-), the results grid will show only the relevant variables within the selected group(s). The values in the results grid can be copied and pasted into another software application such as Microsoft Word, or the user can navigate back to the first tab or continue on to the next tab when the tasks are completed in this step.

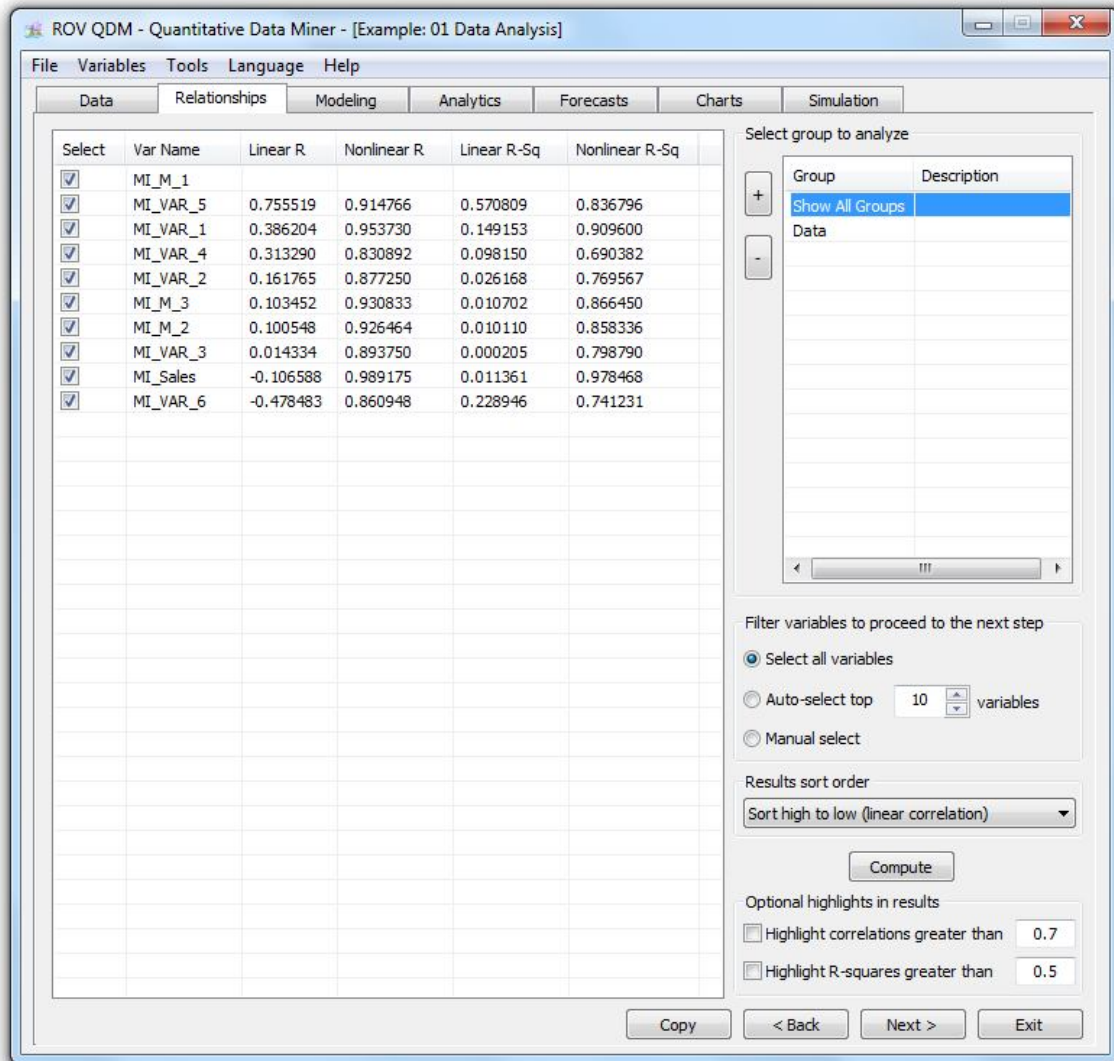


Figure 10 – ROV QDM Relationships Tab

MODELING

Figure 11 illustrates the **MODELING** tab. To follow along with the examples below, we suggest you start QDM and click on **FILE | EXAMPLES | 03 MODELING** to open the relevant sample profile discussed. Note that this example file already has preexisting data variables mapped and created models.

In this section, depending on the **Group** selected, the variables that are members of that group will be listed in the **Variable** list. Next, select the **Modeling Method** to run from the droplist, such as custom econometric model, autocorrelation, partial autocorrelation, heteroskedasticity, seasonality, structural shift, and so forth (see the section on Quick Technical Discussions on Models for details of each of these methods). Based on the modeling method selected, a short **Description** of the method as well as a sample set of required input parameters are shown. At this point, you can type in the **Model Inputs** required for the model chosen or double-click on any one of the variables. Continue to complete the required inputs. Click **Compute** to run the model and the **Results** will be displayed. You can keep creating new models in this tab by typing in a model **Name** and clicking **Add**. You can also **Edit** or **Delete** any existing model as required, and the **Saved Models** section shows the updated list.

As multiple models can be performed in this step, you can provide unique names for each of the models and additional inputs in the models can be entered as required, depending on the model selected. The list of existing models can be recalled by clicking on the saved models list and can be run by either clicking **Compute** or simply double-clicking on the model name. The results can be updated or **Copied** to another software application as required, and if **Update** is selected, the results will be shown in the results work area. A detailed **Report** for this step can be created (see the section on reports creation for details). You can select the results and right-click to copy the contents, use CTRL+C on the keyboard to copy the contents to clipboard, or click on **Data Extract** to extract the results into a Microsoft Word document. Clicking on the **Next** button will continue the process to the next step or Analytics tab, or clicking on the **Back** button will return the user back to the previous Relationships tab. Finally, you can access the **ROV Optimizer** or **ROV Valuator** from the two buttons at the bottom or through the **Tools** menu.

The following lists the modeling methods that are currently available in ROV's QDM software:

- Autoeconometrics (Detailed)
- Autoeconometrics (Quick)
- Custom Econometric Model
- Deseasonalize
- Limited Dependent Variables (Logit)
- Limited Dependent Variables (Probit)
- Limited Dependent Variables (Tobit)
- Linear Regression
- Nonlinear Regression
- Principal Component Analysis
- Stepwise Regression (Correlation)
- Stepwise Regression (Forward)
- Stepwise Regression (Backward)
- Stepwise Regression (Forward-Backward)
- ROV Compiler EXE Model

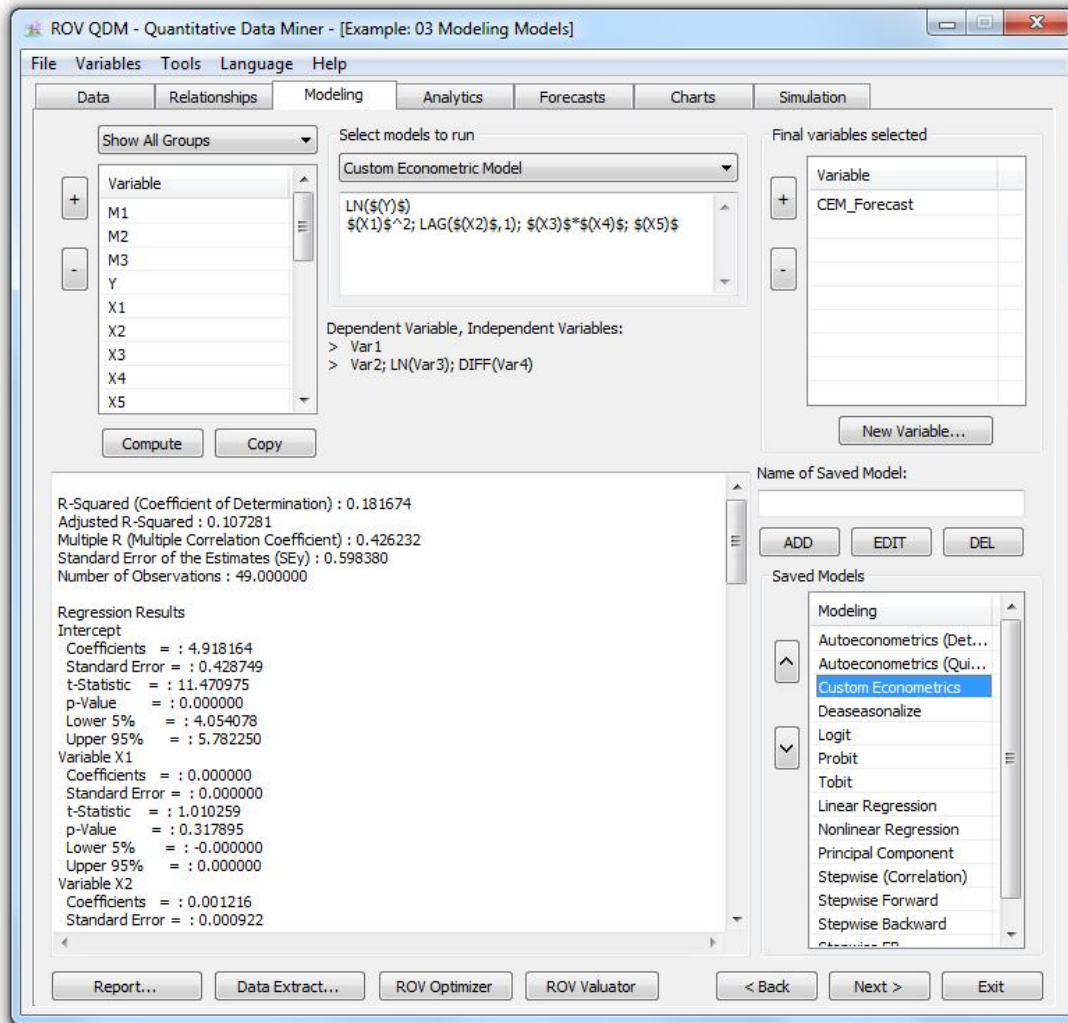


Figure 11 – ROV QDM Modeling Tab

Hands-on Exercises in Modeling

For practice, follow the steps below:

4. Click on *File | Examples | 03 Modeling*, click on and select the third model, *Custom Econometrics*, and do the following:
 - a. Click *Compute* and review the results.
 - b. Review the input parameters:
 - i. Dependent Variable: $\text{LN}(\$Y)\$$
 - ii. Independent Variables: $\$(X1)\$^2; \text{LAG}(\$X2)\$,1; \$(X3)\$*\$(X4)\$; \$(X5)\$$
 - c. Notice how the “ $\$()\$$ ” is used around a variable name, how “ $;$ ” is used to separate variables, and the mathematical operators that can be used ($*$, LAG , $^$, and so forth).
 - d. Add an extra variable X4 to the end of the variables list. You can do this two ways:
 - i. Manually type in $\$(X4)\$$ after the end, making sure to include a semicolon before this new variable, or
 - ii. Double-click on the X4 variable from the Variables list grid on the left
5. Double-click on some of the other Saved Models. Review the inputs in each model and the results and notice that double-clicking an existing model actually runs it as well.
6. Create and run your own models and review the technical section for details of each approach

ANALYTICS

Figure 12 illustrates the **ANALYTICS** tab. To follow along with the examples below, we suggest you start QDM and click on **FILE | EXAMPLES | 04 ANALYTICS** to open the relevant sample profile discussed. Note that this example file already has preexisting data variables mapped and created models.

In the **ANALYTICS** tab, you can select the data **Group** to analyze. Depending on the group selected, the **Variables** that are members of that group will be listed, where one or more variables can be selected, or all variables can either be selected (+) or deselected (–) at once. Next, the **Analysis** type can be chosen such as autoeconometrics, linear or nonlinear multiple regression analysis, principal component analysis, create your own econometric models, and so forth (see the section on Quick Technical Discussions on Models for more details). The results can be updated or **Copied** to another software application as required, and if **Compute** is selected, the results will be shown in the results work area. As multiple analyses can be performed in this step, the user can provide unique names for each of the analysis and the list of existing analyses can be recalled where you can **Add** or **Delete** an analysis as required. A detailed **Report** for this step can be created or the data for the final variables can be extracted into Microsoft Excel or as a flat text file for use in another software or database application. Clicking on the **Next** button will continue the process to the next step or Forecasts tab, or clicking on the **Back** button will return the user back to the previous Modeling tab.

The following lists the analytical techniques that are available in ROV's QDM:

- ANOVA: Randomized Blocks Multiple Treatments
- ANOVA: Single Factor Multiple Treatments
- ANOVA: Two Way Analysis
- Autocorrelation & Partial Autocorrelation
- Correlation (Linear, Nonlinear)
- Data Descriptive Statistics
- Distributional Fitting
- Heteroskedasticity
- Nonparametric: Chi-Square Goodness of Fit
- Nonparametric: Chi-Square Independence
- Nonparametric: Chi-Square Population Variance
- Nonparametric: Friedman's Test
- Nonparametric: Kruskal-Wallis Test
- Nonparametric: Lilliefors Test
- Nonparametric: Runs Test
- Nonparametric: Wilcoxon Signed-Rank (One Var)
- Nonparametric: Wilcoxon Signed-Rank (Two Var)
- Parametric: One Variable (T) Mean
- Parametric: One Variable (Z) Mean
- Parametric: One Variable (Z) Proportion
- Parametric: Two Variable (T) Dependent Means
- Parametric: Two Variable (T) Independent Equal Variance
- Parametric: Two Variable (T) Independent Unequal Variance
- Parametric: Two Variable (Z) Independent Means

- Parametric: Two Variable (Z) Independent Proportions
- Parametric: Two Variable (F) Variances
- Seasonality
- Segmentation Clustering
- Structural Break
- ROV Compiler EXE Model

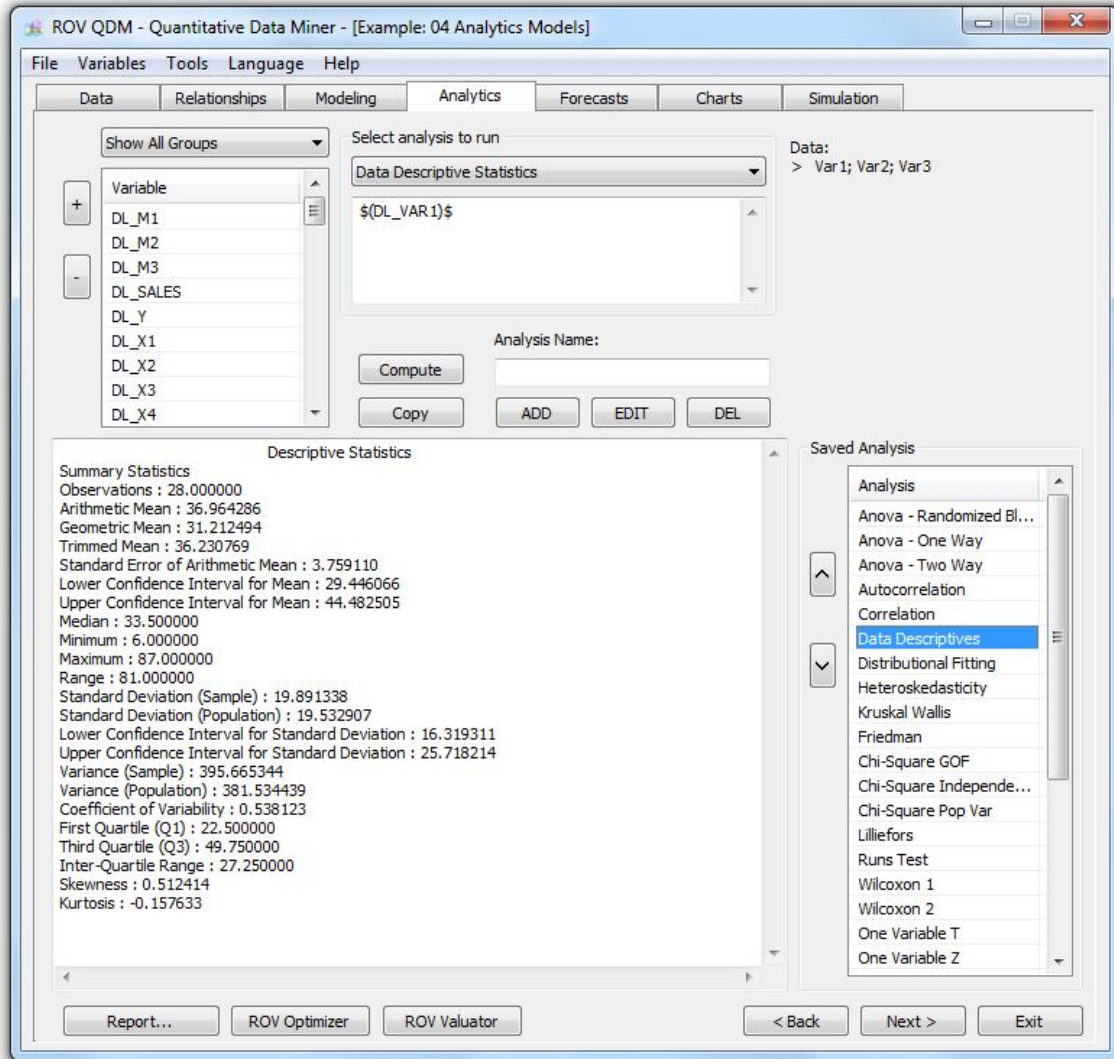


Figure 12 – ROV QDM Analytics Tab

Hands-on Exercises in Analytics

For practice, follow the steps below:

1. Click on *File | Examples | 04 Analytics*, click on and select the sixth model, *Data Descriptives*, and do the following:
 - a. Click *Compute* and review the results.
 - b. Add an extra variable (VAR2) to run by double-clicking on *VAR2* and hitting *Compute*
 - c. Click on *Data Descriptives* again and you will notice that your new VAR2 variable is gone. To make this VAR2 variable permanent, make sure *Data Descriptives* is still selected and click on *Edit*, then type in your new variable... and remember to save the example file (*File | Save*). You can now click on another saved analytical model and then click back on *Data Descriptives* and you will notice the new variable is saved.
 - d. Create your own analytical model, give it a unique name, and click *Add* to add it to the list of saved analytics.

FORECASTS

Figures 13, 14, and 15 illustrate the **FORECASTS** tab. To follow along with the examples below, we suggest you start QDM and click on **FILE | EXAMPLES | 05 FORECASTING** to open the relevant sample profile discussed. Note that this example file already has preexisting data variables mapped and created models.

In the **FORECASTS** tab, depending on the **Group** selected, the **Variables** that are members of that group will be listed, where one or more variables can be selected. Next, the prediction or **Forecast** procedure can be chosen such as AUTO ARIMA, ARIMA, econometric model, multiple regression, time-series forecasting, or other predictive algorithms, and so forth (see the section on Quick Technical Discussions on Models for details). Depending on the forecast model selected from the droplist, a short description of the model and its required inputs will be displayed. As multiple models can be performed in this step, the user can provide unique names for each of the forecast models or new models can be **Added** or **Deleted** as required and the list of existing models can be recalled as required. Further, the list of forecast models can be rearranged by moving a specific model **Up** or **Down** in the list using the respective model icons. The **Results** from the forecast method selected are displayed as a **Chart** of the actual and predicted values, **Data** of the forecast values, or detailed **Statistics** of the forecast results. The results can then be updated or **Copied** to another software application as required, and if **Compute** is selected, the forecast will be recomputed and the results shown. The numerical forecast results are also available in a results grid complete with the time period listing, actual data used, forecast fit, and estimated errors in prediction. A detailed report for this step can be created when the **Report** button is selected. Clicking on the **Next** button will continue the process to the next step or Charts tab, or clicking on the **Back** button will return the user back to the previous Analytics tab.

In the **Chart** results subtab (Figure 13), you will see an overlay of two or more charts (e.g., double-click on the **Time-Series Auto** model and you will see a chart of the historical data and another chart overlaid on it that performs a backcast or back-fitting of these historical data and a forecast of the future). Here, you can click on **Show Values** to see the actual data points on the charts, **Show Legend** to identify the corresponding colors of the charts, or any of the numerous chart control icons (you can rotate the chart, change the colors of background or charts, shift, zoom, and apply other changes to the chart).

In the **Data** results subtab (Figure 14), you can obtain the data points behind the charts and forecast results (e.g., double-click on the **ARIMA** example model to run it and go to the Data results subtab and you will be able to **Copy** these data to Excel, which includes the original historical data points, the back-fitted and forecast values, as well as the prediction errors). From here, you can also change the number of **Decimals** to show on the grid.

In the **Statistics** results subtab (Figure 15), you can obtain the detailed statistical analysis of the forecast method applied, including its goodness-of-fit, historical, and forecast results; accuracy; errors; and details of the model applies (e.g., double-click on the **Autoeconometrics (Quick)** example model to run it and go to the Statistics results subtab and you will see the details of all the models run and be able to **Copy** these results to clipboard and paste them in any other software of your choice).

The following are the forecasting methods currently supported by ROV's QDM software:

- ARIMA
- Auto ARIMA
- Auto Econometrics (Quick)
- Auto Econometrics (Detailed)
- Basic Econometrics
- Cubic Spline
- Exponential J Curve
- Linear Interpolation
- Logistic S Curve
- Markov Chain
- Multiple Regression (Linear)
- Multiple Regression (Nonlinear)
- Stochastic Processes (Geometric Brownian Motion)
- Stochastic Processes (Exponential Brownian Motion)
- Stochastic Processes (Jump Diffusion)
- Stochastic Processes (Mean Reversion)
- Stochastic Processes (Mean Reversion with Jump Diffusion)
- Time-Series Analysis (Auto)
- Time-Series Analysis (Single Moving Average)
- Time-Series Analysis (Double Moving Average)
- Time-Series Analysis (Single Exponential Smoothing)
- Time-Series Analysis (Double Exponential Smoothing)
- Time-Series Analysis (Seasonal Additive)
- Time-Series Analysis (Seasonal Multiplicative)
- Time-Series Analysis (Holt-Winter's Additive)
- Time-Series Analysis (Holt-Winter's Multiplicative)
- Trend Line (Linear)
- Trend Line (Exponential)
- Trend Line (Logarithmic)
- Trend Line (Moving Average)
- Trend Line (Polynomial)
- Trend Line (Power)
- Trend Line (Linear Detrended)
- Trend Line (Difference Detrended)
- Trend Line (Exponential Detrended)
- Trend Line (Logarithmic Detrended)
- Trend Line (Moving Average Detrended)
- Trend Line (Polynomial Detrended)
- Trend Line (Power Detrended)
- Trend Line (Rate Detrended)
- Trend Line (Static Mean Detrended)
- Trend Line (Static Median Detrended)
- Volatility: Log Returns Approach
- Volatility: GARCH
- Volatility: GARCH-M
- Volatility: EGARCH
- Volatility: EGARCH-T
- Volatility: GJR GARCH
- Volatility: GJR TGARCH
- Volatility: TGARCH
- Volatility: TGARCH-M
- Yield Curve (Bliss)
- Yield Curve (Nelson-Siegel)

Hands-on Exercises in Forecasts

For practice, follow the steps below:

1. Click on *File | Examples | 05 Forecasting*, click on and select the model, *Time-Series Auto*, and do the following:
 - a. Click *Compute* and review the results in the Chart, Data, and Statistics subtabs.
 - i. Identify which line corresponds to the historical data and which line is the back-fitting and forecast prediction.
 - ii. Copy or extract the data to Excel or some other software (Word, PowerPoint).
2. Double-click on the *Detrend* models (chose any one of them) and identify what is going on here and what methodology is applied.
3. Double-click on *Stochastic – GBM* for the geometric Brownian motion stochastic process.
 - a. Double-click on the same model a few additional times and notice what happens each time, what changes, and why.
 - b. Repeat the process, but this time focus on the Data results subtab and notice how the results change each time.

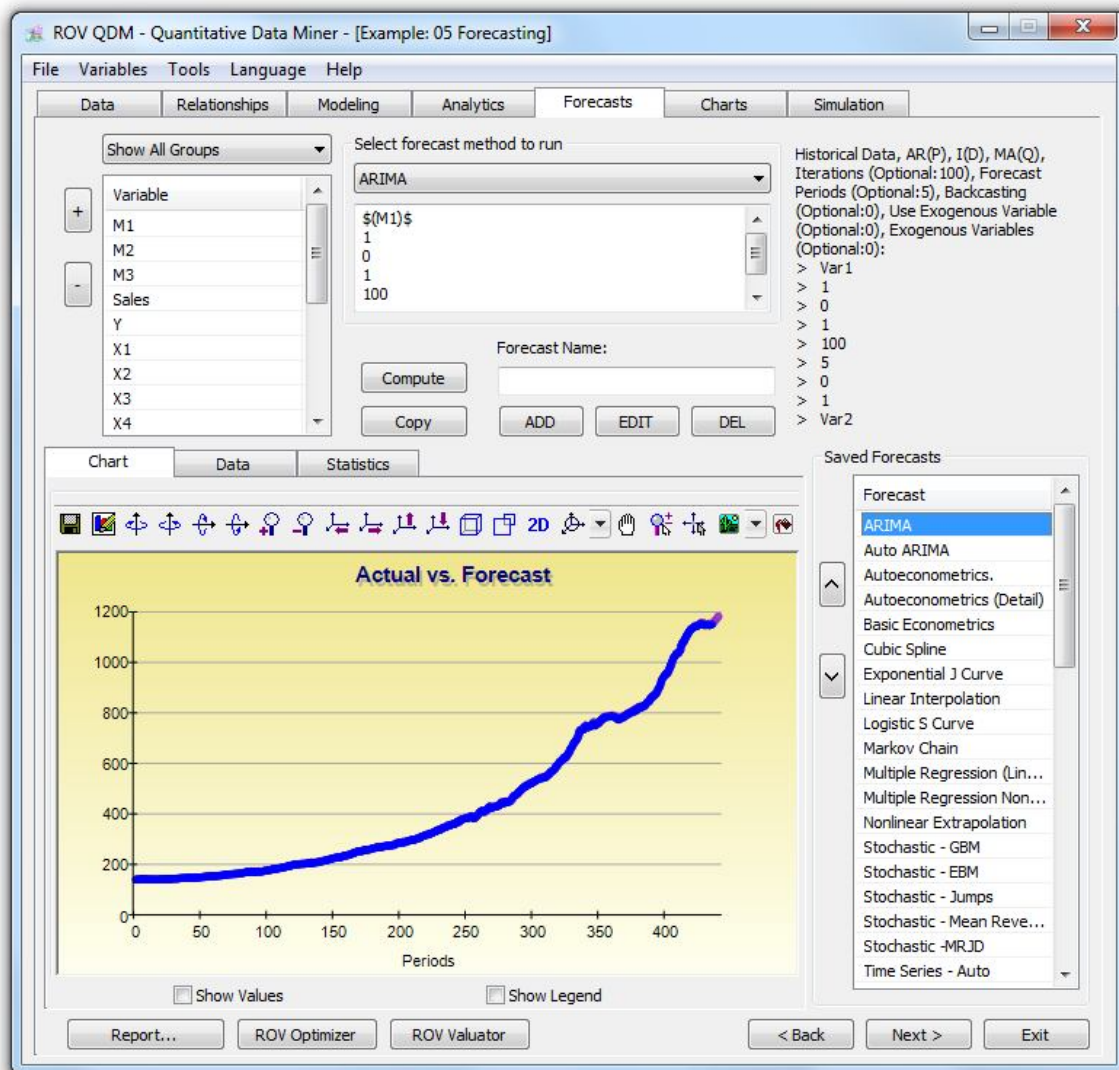


Figure 13 – ROV QDM Forecasts Tab: Chart

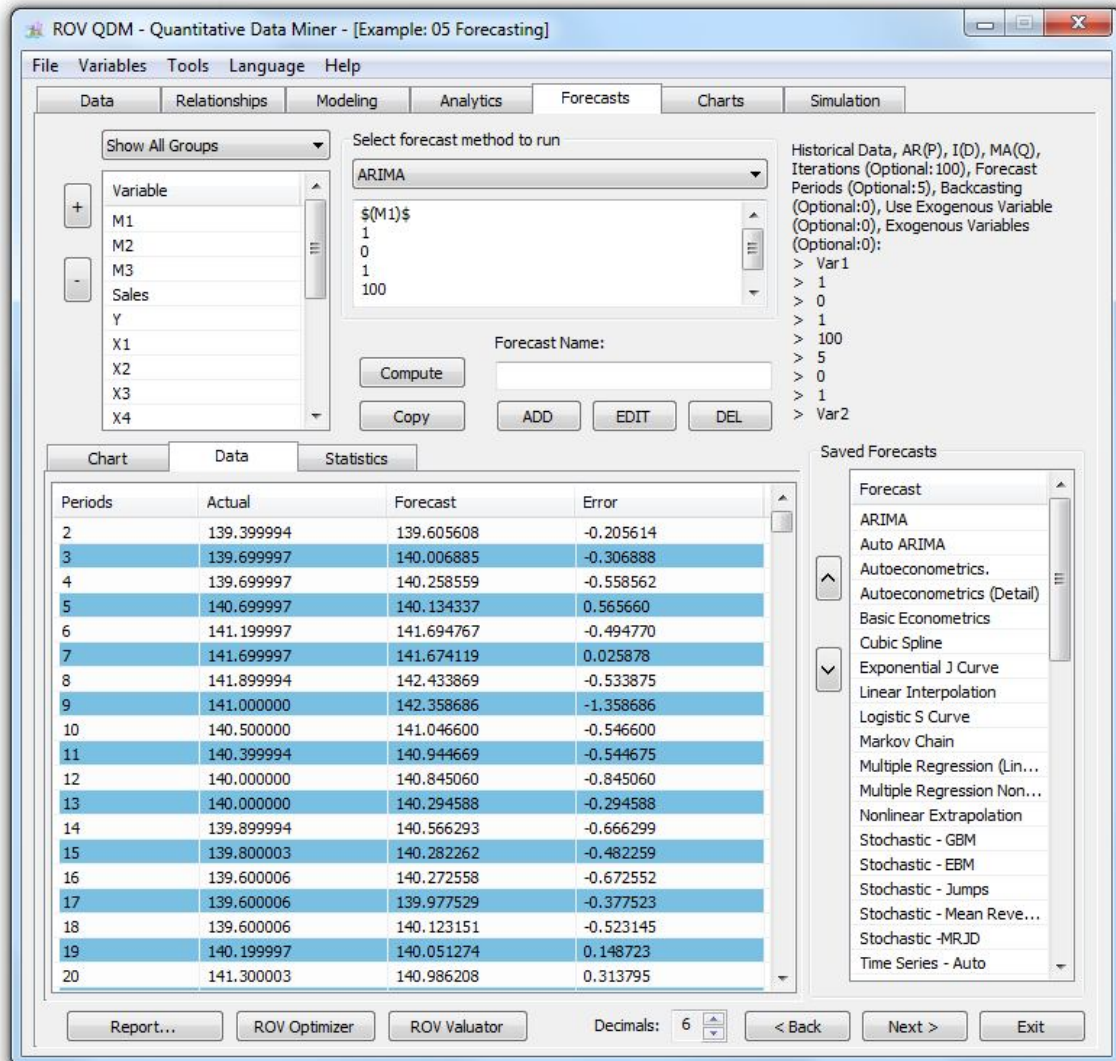


Figure 14 – ROV QDM Forecasts Tab: Data

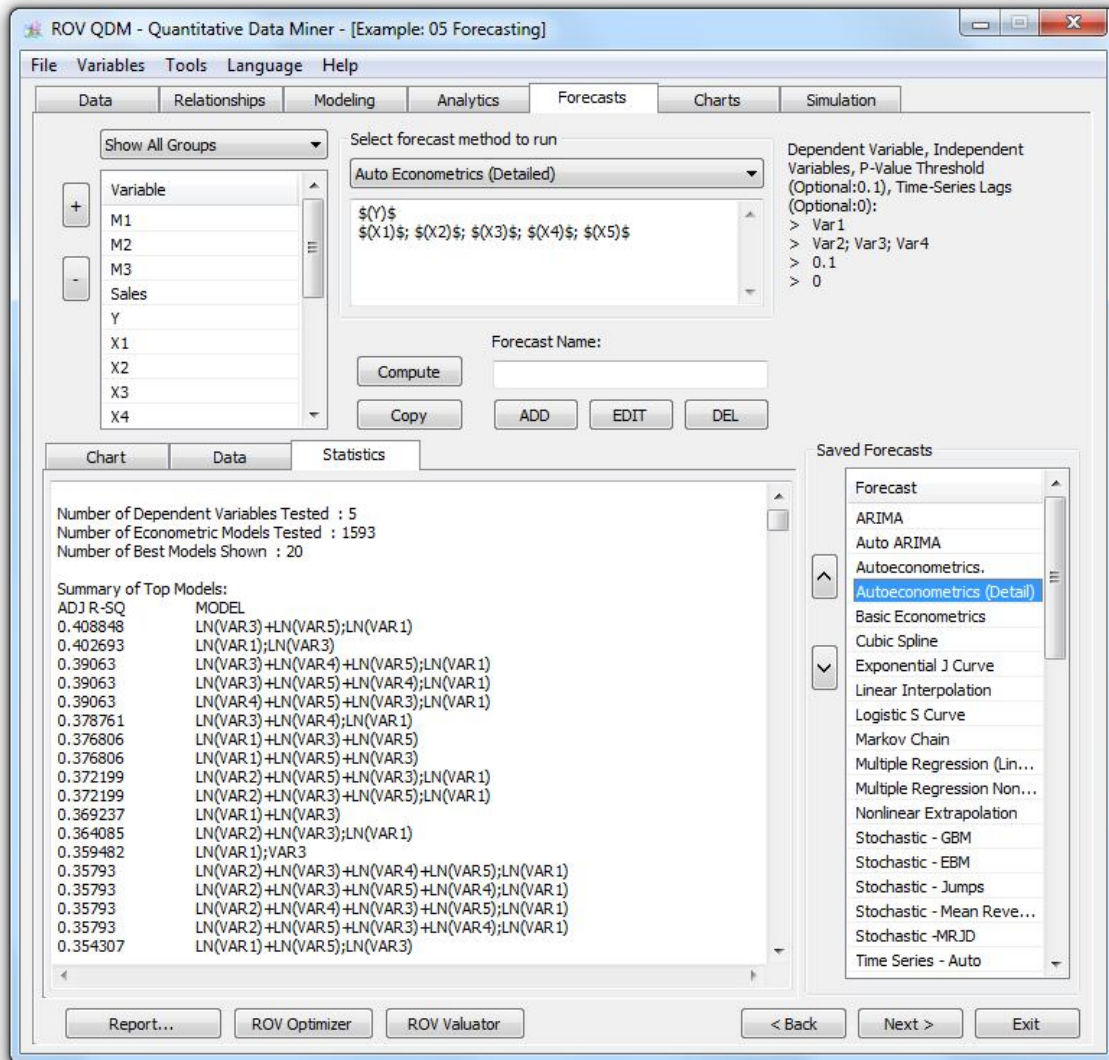


Figure 15 – ROV QDM Forecasts Tab: Statistics

CHARTS

Figure 16 illustrates the **CHARTS** tab. To follow along with the examples below, we suggest you start QDM and click on **FILE | EXAMPLES | 06 CHARTS** to open the relevant sample profile discussed. Note that this example file already has preexisting data variables mapped and created models.

Depending on the **Group** selected, the **Variables** that are members of that group will be listed, where in this charting analysis module one or more variables can be selected at once. New charts can be **Added** or an existing chart can be **Deleted** where the list of previously created and saved charts can be recalled at any time to view the results and can be reordered by selecting a created chart on the list and clicking on the **Up** and **Down** buttons. See the hands-on exercises below for more step-by-step details on how to create new charts. Further, as each new chart is added, the **Chart Type** (e.g., 3D line chart, 2D line chart, 3D bar chart, 2D bar chart, and others) can be selected, a **Chart Title** can be saved together with the chart (the titles will be displayed in the list of saved and created charts), **Chart Notes** can also be saved as a reminder of what the chart is all about, and the chart can then be displayed when the **Update** button is depressed. The chart will be shown in the chart area complete with a series of **Chart Power Tools** that allow the user to change certain look and feel features of the chart such as the background color, chart color, 2D look, 3D rotation, chart shifts, and others. In addition, the chart's **x-axis** and **y-axis** can be automatically computed or can be manually entered to create a custom chart. Here, you can click on **Show Values** to see the actual data points on the charts and on **Show Legend** to identify the corresponding colors of the charts. A detailed report for this step can be created when the **Report** button is selected. The chart can also be **Copied** as a jpeg image to be pasted into another software application as required. Clicking on the **Next** button will continue the process to the next step or Simulation tab, or clicking on the **Back** button will return the user back to the previous Forecasts tab.

The following are the types of charts and charting techniques supported by the current version of ROV's QDM software:

- Standard 2D Line
- Standard 3D Line
- Standard 2D Bar
- Standard 3D Bar
- Standard 2D Area
- Standard 3D Area
- Standard 2D Point
- Standard 3D Point
- Standard 2D Scatter
- Standard 3D Scatter
- Control Chart: P
- Control Chart: NP
- Control Chart: U
- Control Chart: C
- Control Chart: X
- Control Chart: R

- Control Chart: XMR

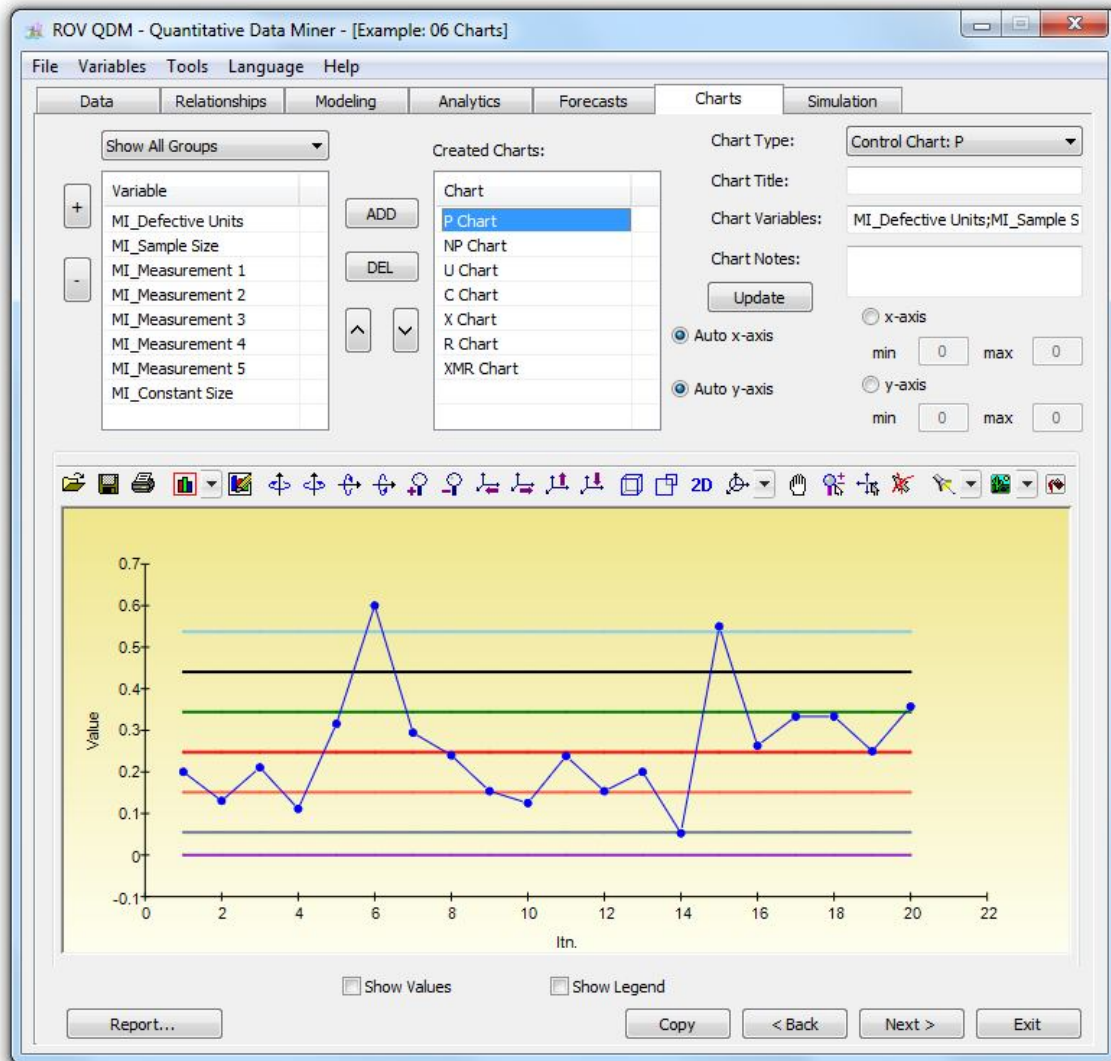


Figure 16 – ROV QDM Charts Tab

Hands-on Exercises in Charts

For practice, follow the steps below:

1. Click on *File | Examples | 06 Charts*, click on and select the chart, *P Chart*, and do the following:
 - a. Click *Update* to view the *P-Control* chart (see the Quick Technical Discussions section for details of control chart types and what they represent).
 - b. Click on the *Chart Type* droplist and select other charts to view.
2. Create a basic new chart by doing the following:
 - a. Click on the *last empty row* in the *Created Charts* grid list.
 - b. Select the chart type, for example, *Standard 2D Line with Points*.
 - c. Select the Variable, for example, *Defective Units*.
 - d. Click *Add* and give it a name such as "*Test*".
 - e. Double-click on *Test* (or the newly created chart name).
 - f. Select a different type of chart and see the results.
 - g. Play with some of the chart power tools by clicking on some of the icons and seeing the effects of each icon.
3. Create an overlay chart of several variables by doing the following:
 - a. Click on the *last empty row* in the *Created Charts* grid list.
 - b. Select the chart type, for example, *Standard 2D Bar*.
 - c. Select several Variables by holding down the CTRL key on your keyboard and clicking on several variables, for example, *Measurement 1, Measurement 2, Measurement 3*.
 - d. Click *Add* and give it a name such as "*Test 2*".
 - e. Double-click on *Test 2* (or the newly created chart name).
 - f. Select a different type of chart and see the results.
4. Create a new Control Chart type by first reviewing the Quick Technical Discussions section of this manual to understand what the required inputs are for each type of control chart and replicating the steps above.

SIMULATION

Figure 17 illustrates the **SIMULATION** tab. To follow along with the examples below, we suggest you start QDM and click on **FILE | EXAMPLES | 07 SIMULATION** to open the relevant sample profile discussed. Note that this example file already has preexisting data variables mapped and created models.

Depending on the **Simulation Charts** selected, new **Models with Simulation** can be **Added** or **Deleted** as a new chart as required and the list of existing simulation charts can be recalled as required (i.e., models with at least one input set as a simulation assumption). The results from the simulation profile selected are displayed as either detailed analytics results or in a graphical chart depending on which chart type is selected. Users can also add in a custom chart title and include any notes on what the chart represents. Clicking on the **Run Simulation** button will run the simulation (see Figure 17 for details). The generated chart is by default a histogram, which charts all the frequency of occurrences of values as a vertical bar chart, and depending on the **Chart Type** selected, additional overlay charts such as a best-fitting probability density function or PDF chart can also be constructed and viewed overlaid on the same chart. There are also a series of chart control icons available that allow the user to change certain look and feel features of the chart such as the background color, chart color, 2D look, 3D rotation, chart shifts, and others. There is an interactive control section at the top-right corner of the chart that returns the probabilities and confidence levels of the simulated results. For instance, users can choose the tail type to compute (such as two-tails, left-tail $<$, left-tail \leq , right-tail $>$, or right-tail \geq), enter in the relevant values for these tails to obtain the **Certainty** percentage, or, alternatively, enter in the Certainty values to obtain the relevant **Tail Values**. A detailed report for this step can be created when the **Report** button is selected, and the resulting chart can also be **Copied** to another software application as required. Clicking on the **back** button will return the user back to the previous Charts tab.

Figure 17 illustrates a sample simulation run where the number of Monte Carlo risk simulation **Trials** can be set, with the additional option of setting a simulation **Seed Value** where if selected, the simulated values will be reproducible every time, versus completely random sequences if not set. Clicking on the **Run Simulation** button or hitting **F9** on the keyboard will execute the simulation procedure whereas Cancel will bring the user back to the Simulation tab.

There is a **Statistics** results subtab (Figure 18) that includes the computations of trials (number of simulation trials run), mean (arithmetic average), median (50th percentile), standard deviation (statistical computation of the square root of the average squared distance from the mean), variance (square of the standard deviation), coefficient of variation (standard deviation divided by the mean), skew (measure of third degree directionality), kurtosis (measure of peakedness of the distribution), minimum (smallest value), maximum (largest value), and range (maximum minus minimum).

The **Chart Data** results subtab (Figure 19) returns the data required to recreate the chart(s) displayed. For instance, if Histogram and CDF is chosen as the chart to display, then the chart data returned will include the Histogram data and CDF data used to plot the charts. Finally, the **Simulation Data** results subtab (Figure 20) returns the actual simulated raw data.

The following statistical and mathematical distributions are supported as input assumptions in the current ROV's QDM software for running simulations:

- Bernoulli Distribution
- Beta Distribution
- Binomial Distribution
- Chi-Square Distribution
- Discrete Uniform Distribution
- Exponential Distribution
- F Distribution
- Gamma Distribution
- Gumbel Min Distribution
- Gumbel Max Distribution
- Logistic Distribution
- Lognormal Distribution
- Normal Distribution
- Pareto Distribution
- Poisson Distribution
- Rayleigh Distribution
- Standard Normal Distribution
- T-Distribution
- Triangular Distribution
- Uniform Distribution
- Weibull Distribution

Hands-on Exercises in Simulation

For practice, follow the steps below:

1. Click on *File | Examples | 07 Simulation*, click on and select the first model, *Addition*, and do the following:
 - a. Click on *Run Simulation* or *F9* on the keyboard.
 - b. Review each of the results subtabs: *Chart*, *Statistics*, *Chart Data*, and *Simulation Data*.
2. Select the *Addition* model and click *Run Simulation*, then do the following:
 - a. Double-click on the *Certainty* input box or select the default 100 value and type in *90* to compute the two-tailed 90% confidence interval on the simulated results. Try other inputs as well as long as the input values are between 0 and 100.
 - b. Select a different tail confidence such as *Left Tail* or *Right Tail* and enter in a certainty % value between 0% and 100%. Explain what the results represent.
 - c. Select a different tail confidence such as *Left Tail* or *Right Tail* and this time enter in the value to obtain the certainty percentile. Explain what the results represent.
3. Select the *Brownian Motion* model and *Run Simulation*, then do the following:
 - a. Rerun the simulation a few times and pay attention to the results (e.g., look at the *Statistics* subtab and see what happens to the results each time a new simulation is run on the same model).
 - b. This time, check the box beside *Enable Seed Value*, rerun the simulation a few times, and notice what happens to the results (e.g., the results in the *Statistics* subtab). Explain what happened and what seed values do to the analysis.
 - c. Notice that in the *Brownian motion* stochastic process, there are multiple steps in the process and each step has its own forecast chart. So in this case of multiple charts, you will see a new *Select Charts to Show* droplist where you can select the specific charts to show or to show all charts at once.
4. Select the *Normal (50, 5)* model and click *F9* or *Run Simulation* then do the following:
 - a. Select different chart types and explain what each of these mean:
 - i. Histogram
 - ii. Fitting and Histogram
 - iii. CDF and Histogram
 - iv. PDF and Histogram
 - v. Cumulative (CDF)
 - vi. Probability (PDF)
 - vii. Multiple CDF Overlay
 - viii. Multiple PDF Overlay
 - b. Change the number of *Bins* and *Decimals* to show on the chart and click *Run Simulation* to update the chart.
 - c. Change the *Bar Type* and *Bar Color* as well as the *Line Color* on the chart.
 - d. Change some of the advanced settings on the chart by using some of the chart power tool icons.

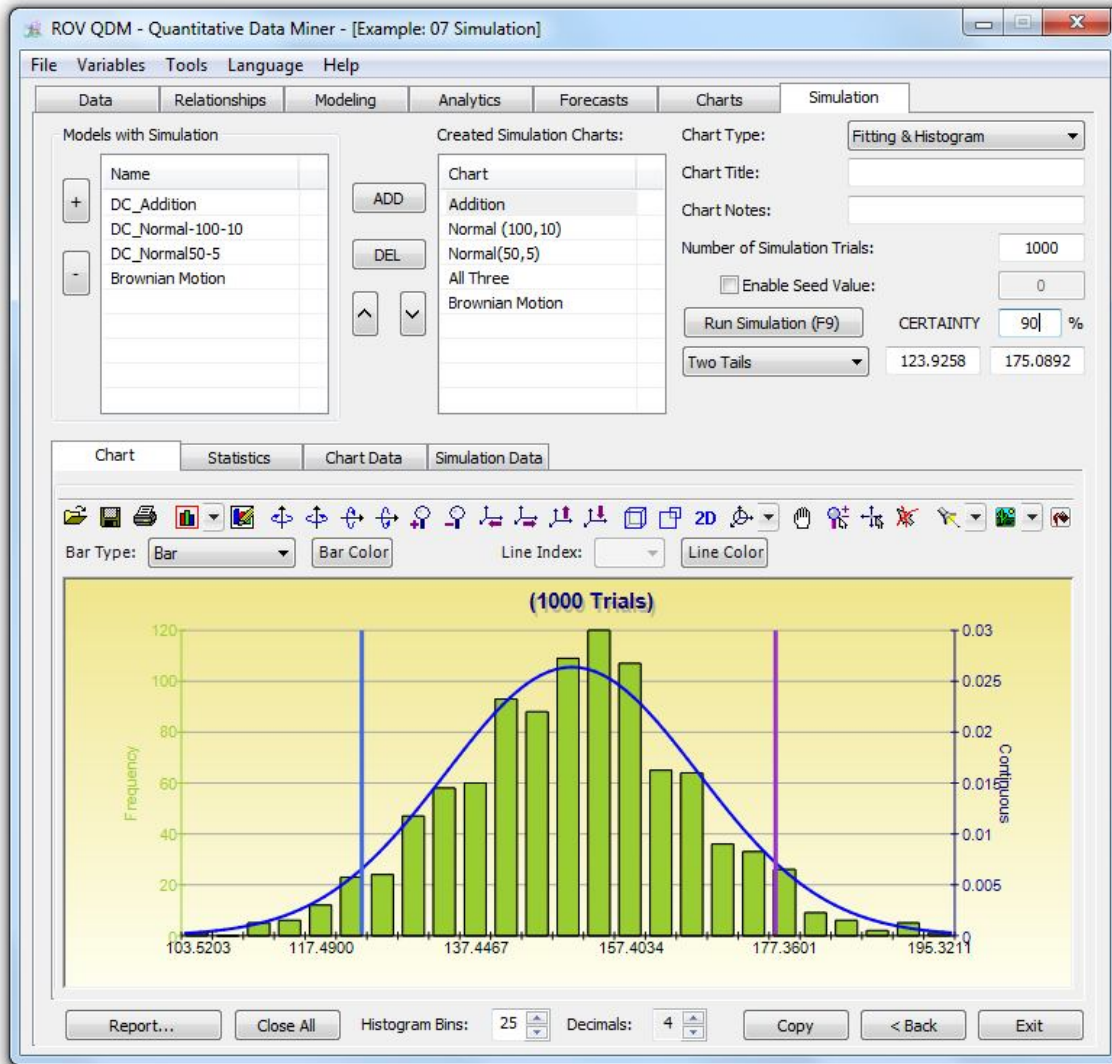


Figure 17 – ROV QDM Simulation Tab: Charts

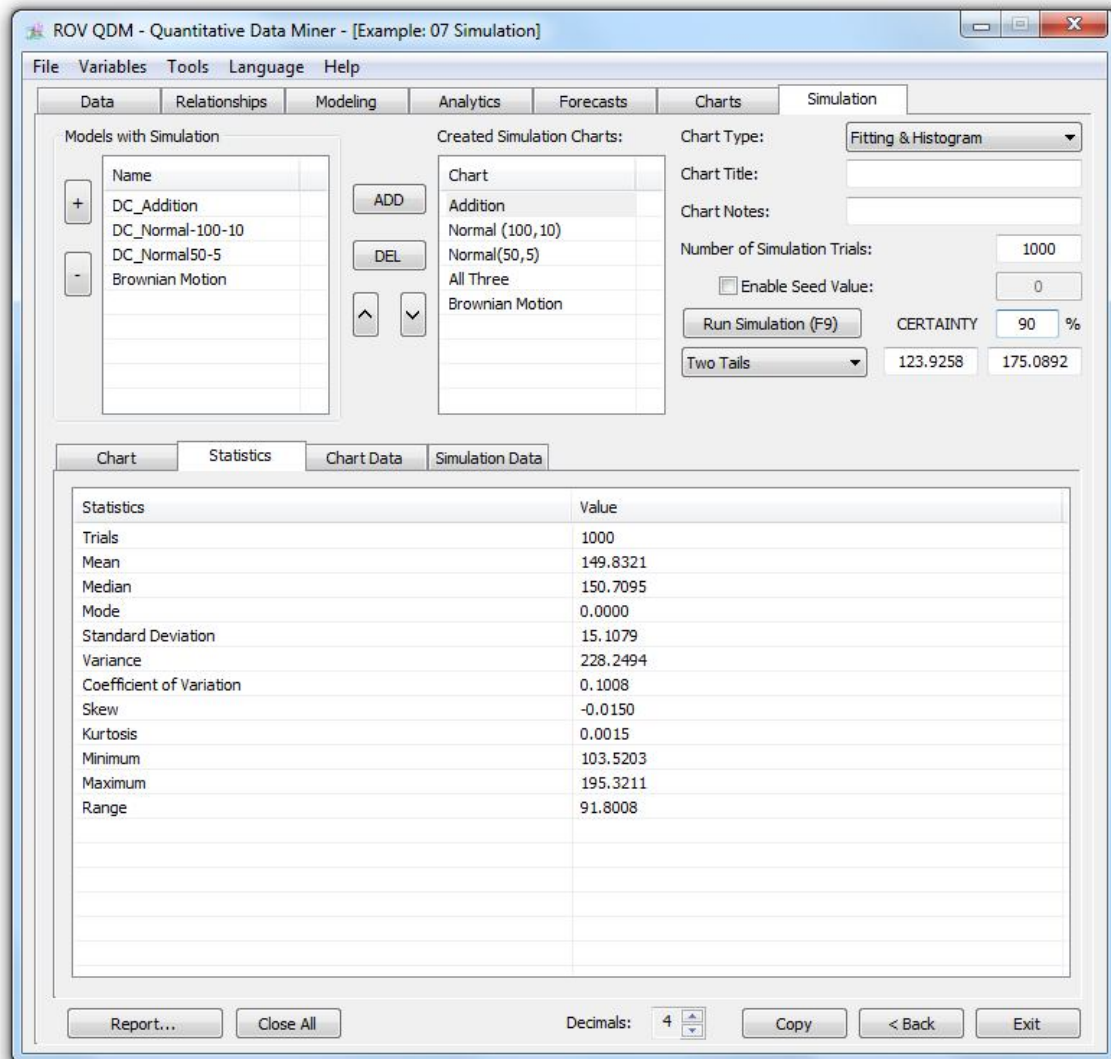


Figure 18 – ROV QDM Simulation Tab: Statistics

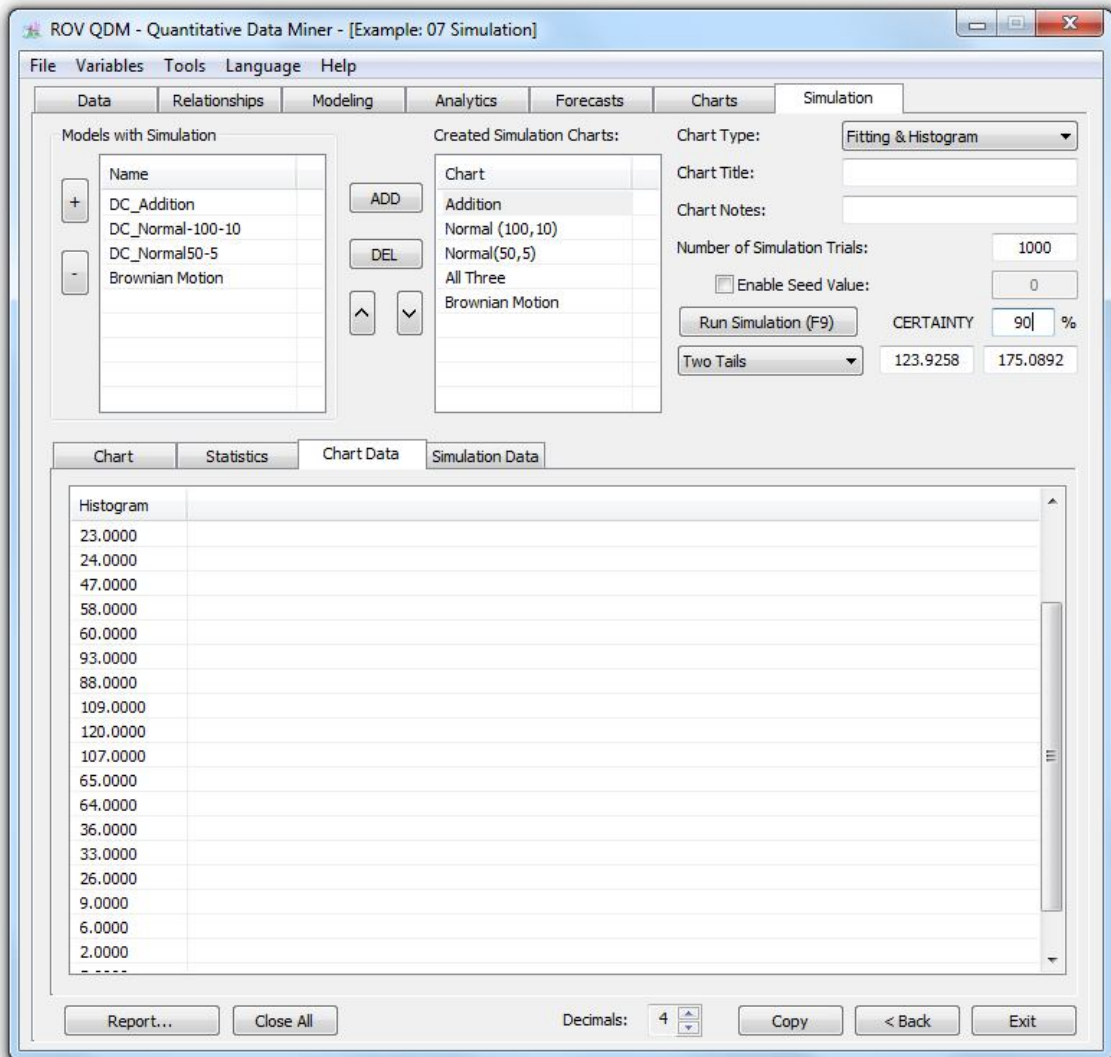


Figure 19 – ROV QDM Simulation Tab: Chart Data

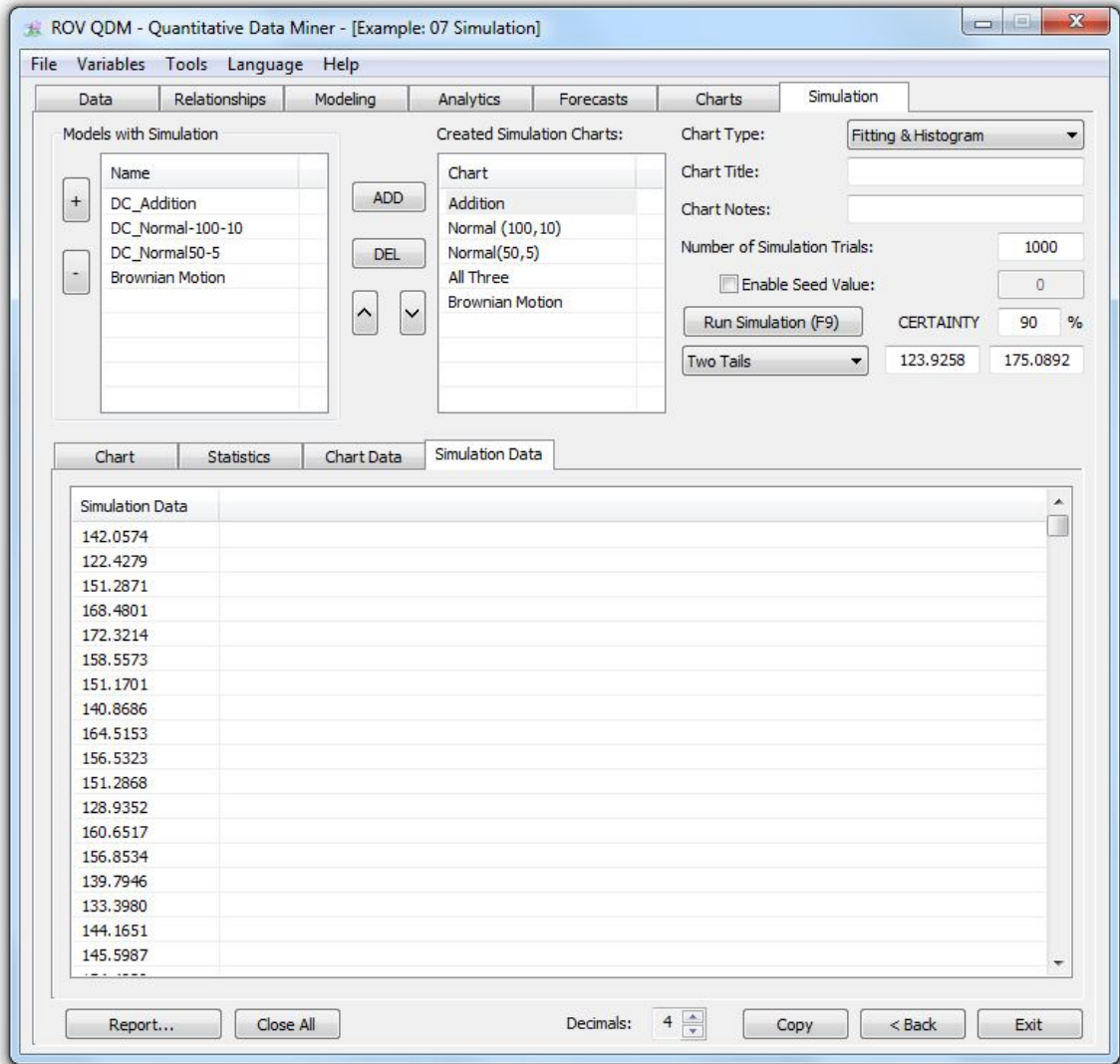


Figure 20 – ROV QDM Simulation Tab: Simulation Data

REPORTS

Figure 21 illustrates the Report Creator, where there are seven types of reports that can be generated based on the selections made in the QDM software: In the Data section of QDM, users can choose whether to extract all original primary raw data or secondary data based on the newly created groups. In the Relationships section, the resulting correlations and R-square results or the list of variables that made it through the third round of filtering can be extracted. Under the Analytics section, the detailed results of the analytical runs can be extracted or only the variables' name list of the final selected variables can be extracted. Next, the Modeling detailed results can be extracted, and the Forecasts on various variables can also be extracted as a report or raw data. All Charts can be copied and pasted into the Windows clipboard for pasting into other software such as Microsoft PowerPoint, or the formatted data behind each of the charts can be extracted and pasted into another software application. The results from Simulation runs can also be extracted, including the simulation histogram charts or the actual simulated data points. The selected items in the Report Creator will be generated once the Run command is selected, or the entire selection is canceled otherwise.

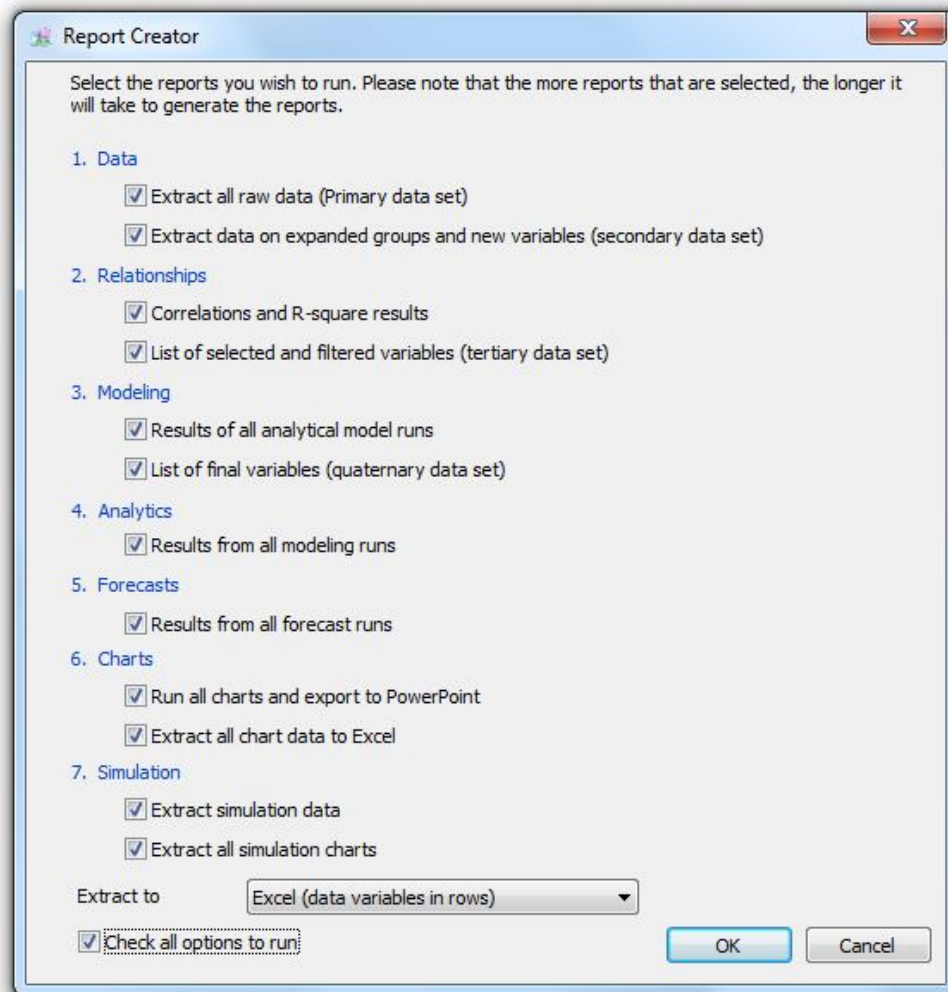


Figure 21 – ROV QDM Create Reports

ROV VALUATOR

ROV Valuator is the application of over 600+ advanced analytical functions. It affords hundreds of models in different categories from which the user can select. The user can input the required data for the selected model and this application will return the computed results very quickly. This module is useful for valuing derivative instruments, debt instruments, exotic options, and options-embedded instruments, as well as multiple types of financial models. The 600+ advanced models are categorized into the following groups of applications:

- Advanced Math Functions
- Basic Finance Models
- Basic Options Models
- Bond Math, Options, Pricing, and Yields
- Credit Risk Analysis
- Delta Gamma Hedging
- Exotic Options and Derivatives
- Financial Ratios
- Forecasting, Extrapolation, and Interpolation
- Probability Distributions
- Put-Call Parity and Option Sensitivities
- Real Options Analysis
- Value at Risk, Volatility, Portfolio Risk and Return

ROV Valuator **[AR]** (see the accompanying screen shots) is used to perform quick computations from simple and basic models to advanced analytical models, and can handle single point values or a series of values. After installing the software, start ROV Valuator. Simply select the model type in the Model Category **[AS]** box and select the model of interest in the Model Selection **[AT]** box. The required input parameters will then be listed. Single point inputs (e.g., 10 or 10.4532) will be in the single input parameters area **[AU]**, whereas multiple data requirements will be shown in the multiple series input parameters area **[AV]**. When entering a single series of multiple data points, use commas or spaces to separate the values (e.g., a time series of 6 months of interest rates can be entered either as 0.12, 0.124, 0.112, 0.1, 0.09, 0.16 or simply as 0.12 0.124 0.112 0.1 0.09 0.16). Hit COMPUTE and the analysis is run and the results are returned **[AW]**.

Sometimes, certain models, such as the Value at Risk model using the standard correlation method, require different columns of data and a correlation matrix. For instance, the goal is to compute the portfolio VaR using this model, where there are 3 asset classes, each with their own amounts, specific daily volatility for each asset class, and a square correlation matrix among these asset classes. In such a situation, the amounts and volatility inputs will have to be entered as a single column (hit ENTER at the end of entering a value to create a new line designating a new asset class. or use the semicolon as a line separator **[AX]**) and the correlation matrix will be separated by commas for the same row with different columns, and semicolons for different rows **[AY]**. This ROV Valuator module does not allow the user to link to various databases or simulate. To do so, use the ROV QDM module instead. Many of the same models exist in both places. The ROV Valuator module is used to quickly obtain results without having to link to databases and so forth.

AR

AS

AU

AV

AW

ROV Valuator - [C:\Program Files\Real Options Valuation\Risk Modeler\ModuleDefaultValue.xml]

File Languages

Model Category: [All Categories]

Model Selection: [ARIMA]

Model Description: Forecasts time-series variables using the Box-Jenkins autoregressive integrated moving average model.

Single Input Parameters:

P	1	D	0	Q	1
Max Iteration	1000	Forecasts	5	Backcast	0
Input7		Input8		Input9	
Input10		Input11		Input12	
Input13		Input14		Input15	

Multiple Series Input Parameters (Values are COMMA separated, Rows are SEMICOLON separated):

Time-Series Data	Exogenous Data	Input3	Input4	Input5
138.90;139.40;139.70	286.70;287.80;289.10			
8.80;220.00;222.00;2	1.30;660.50;668.80;6			
50;503.20;508.30;510	865.10;1877.00;1895.503.90;3504.10;3507.			

Results:

Regression Statistic: 0.999929 ; Adjusted R-Squared: 0.999929 ; Multiple R: 0.999965 ; Standard Error of the Estimates (SEy): 279.697750 ; Observations: 425

Compute Exit

ROV Valuator - [C:\Program Files\Real Options Valuation\Risk Modeler\ModuleDefaultValue.xml]

File Languages

Model Category: [Value at Risk, Volatility, Portfolio Risk and Returns]

Model Selection: [Portfolio Risk]

Model Description: Computes the portfolio risk given individual asset allocations and variance-covariance matrix

Single Input Parameters:

Input1		Input2		Input3	
Input4		Input5		Input6	
Input7		Input8		Input9	
Input10		Input11		Input12	
Input13		Input14		Input15	

Multiple Series Input Parameters (Values are COMMA separated, Rows are SEMICOLON separated):

Asset Allocations	Covariances	Input3	Input4	Input5
0.25; 0.25; 0.5;	1234,2234,3223; 2334,3322,4454; 2334,3345,3332;			

Results:

Compute Exit

To get started learning how to use this tool, click on File menu and select Load Sample Inputs. Then, select a model category and select a model of interest. You will see the sample inputs loaded. You can then click on Compute to obtain the results. Use these sample inputs as a guide to get started with your modeling needs.

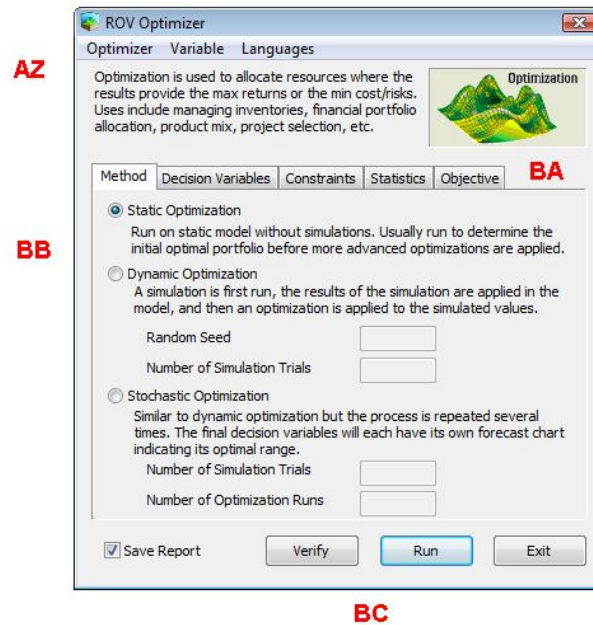
Just as in ROV Modeler, you can customize the list of models that appear in the ROV Valuator, as well as the descriptions for each model. Simply go to the installation path (e.g., c:\program files\real options valuation\risk modeler), look for files "ROV Custom Valuator (English).xml", and select the correct files depending on your language. This XML file controls the user interface names and descriptions. You can edit this file directly using an XML editor or using Notepad (Start, Programs, Accessories, Notepad, and then drag the XML file and drop it into Notepad to edit it). In the XML file, there are several things you can do, including:

- You can delete an entire category starting from `<category>` to `</category>`
- You can delete a specific function inside a category from `<function>` to `</function>`
- You can change anything in the "category name", "displayname" and "desc" description for the model
- You cannot and should not change the "function name", "type" and "param_type" values
- You can but should not change the "var name" of the model (you run the risk that sample values loaded might not have a valid value)
- You can rearrange the location of the models and categories to have certain models and categories appear first or appear later
- Instead of deleting models, try commenting them out using the "open triangular bracket, apostrophe and two dashes" and "two dashes and close triangular bracket" such that if you need the models again later, they are available
- You can also create your own category of models using the examples in this document, with your own favorite list of models...

ROV OPTIMIZER

ROV Optimizer is an advanced optimization module that can be used to optimize portfolios and to find optimal investment decisions and optimal project selections for a corporation, a bank, an investment firm, a manufacturer, an R&D outfit, and many others. The decision variables can be discrete, continuous, integer, or binary, and the objective function can be linear or nonlinear. In addition, ROV Optimizer allows the user to link to existing data tables to run simulations, find the best-fitting models, and couple these techniques with optimization. The technical details of optimization fall outside the scope of this document. For more details and examples, see *Modeling Risk: Applying Monte Carlo Simulation, Real Options Analysis, Stochastic Forecasting, and Portfolio Optimization, Second Edition*, by Dr. Johnathan Mun (Wiley Finance 2010).

Here is a simple example of how to use the ROV Optimizer [AZ] (see the accompanying screen shots). (We also suggest you click on the File menu and select Examples to load some predefined models to learn how the models can be set up.) When you install the ROV Optimizer, you can open and see the UI of the software. “Method”, “Decision Variables”, “Constraints” will show in front of the user. Choose the “Method” [BA] tab and select “Static Optimization” [BB]. Again, for details on the differences among static, dynamic and stochastic optimization, contact our technical support department, review Dr. Mun’s book cited above, or attend one of Real Options Valuation, Inc.’s training seminars.



Next, click on the Decision Variables tab [BA] and hit ADD to add some variables. For instance, we have 4 different variables [BD] (Asset1 to Asset4), and each asset can be set to take continuous, integer, binary, or discrete values [BE]. For our simple illustration, set the variables to all be Continuous between 0.10 and 0.40 (i.e., only asset allocations between 10% and 40% are allowed). Keep adding 4 different asset classes as decision variables.

Next, click on the Constraints tab and select ADD [BF]. Then, in the expressions input box, enter in the constraint (you can double-click on the list of variables and the variable string will be transferred up to the expressions box). In our simple example, the total decision variable values must sum to 1.0 (i.e., the total allocation of asset classes must total 100% in an investment portfolio) [BG]. You can also create an Efficient Frontier by adding the Frontier Variables [BH]. Again, for details on efficient frontiers, review the previously cited modeling risk book by Dr. Mun.

The image displays three screenshots of the ROV Optimizer software interface, illustrating the steps to add constraints and frontier variables.

Top Screenshot: Shows the 'Decision Variable Properties' dialog box. The 'Decision Name' is 'Asset4' and the 'Initial Value' is '0.000000'. The 'Decision Type' is set to 'Continuous (e.g., 1.15, 2.35, 10.55)'. The 'Lower Bound' is '0.1' and the 'Upper Bound' is '0.4'. A red label 'BE' is placed to the right of the dialog box.

Middle Screenshot: Shows the 'Constraints Properties' dialog box. The 'Expression' is '\$(Asset1)\$+\$(Asset2)\$+\$(Asset3)\$+\$(Asset4)\$=1'. A red label 'BG' is placed to the left of the dialog box.

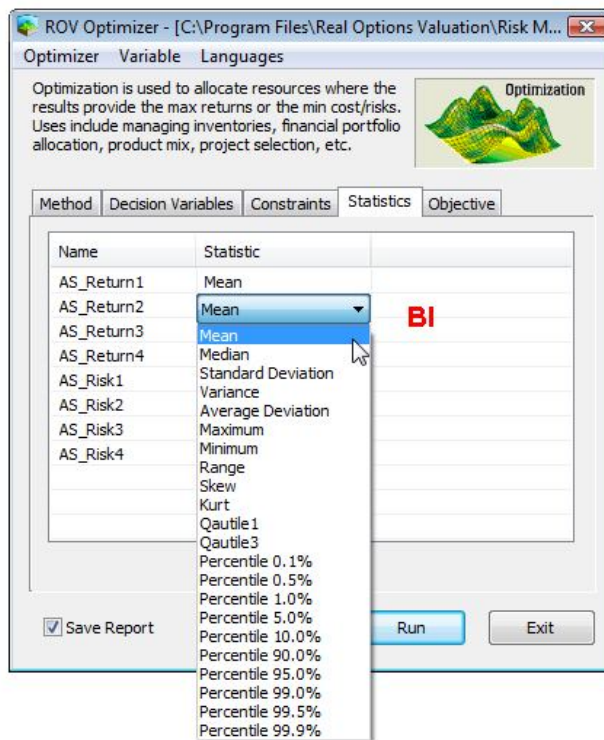
Bottom Screenshot: Shows the 'Frontier Variable Properties' dialog box. The 'Name' is 'EfficientFrontier', the 'From' value is '0.1', the 'To' value is '0.9', and the 'Steps' value is '0.05'. A red label 'BH' is placed to the right of the dialog box.

Background Screenshot: Shows the main ROV Optimizer window. The 'Decision Variables' tab is active, displaying a table with columns 'Name', 'Type', 'Rules', and 'Starting Value'. The table contains four rows for 'Asset1', 'Asset2', 'Asset3', and 'Asset4', all of type 'Continuous' with rules '0.100000 to 0.400000' and starting values of '0.250000'. A red label 'BD' is placed in the table area. The 'Constraints' tab is also visible, showing an 'Add' button.

In addition, if you are using static optimization, you can skip the Statistics tab; the statistics tab is important when you are running a dynamic or stochastic optimization, when some of the variables are mapped to probability distributions and simulations will be run before and after the optimization **[BI]**.

Next, select the Objective tab **[BJ]** and select if you wish to run Maximization or Minimization on your objective. In addition, enter in the relevant objective expression as outlined below. You can double-click on the list of Variables to bring the variable name string to the objective expression input box. When completed, click on RUN to obtain the results of your optimization, or you can first click on Verify to test if the model has been set up correctly.

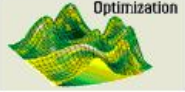
The optimization results **[BK]** will appear if the optimization model is set up correctly. The results will show the number of iterations, the specific model configuration, the parameters, the initial and optimized results of the objective and decision variables, the technical analytics (Lagrange multipliers, Hessian matrices, and others), and an optimization objective chart.



ROV Optimizer - [C:\Program Files\Real Options Valuation\Risk M...

Optimizer Variable Languages

Optimization is used to allocate resources where the results provide the max returns or the min cost/risks. Uses include managing inventories, financial portfolio allocation, product mix, project selection, etc.



Method Decision Variables Constraints Statistics Objective

Optimization Objective

Maximize the value of objective **BJ**

Minimize the value of objective

Objective Expression

$(\$(\text{Asset1})\$(\text{AS_Return1})\$(\text{Asset2})\$(\text{AS_Return2})\$(\text{Asset3})\$(\text{AS_Return3})\$(\text{Asset4})\$(\text{AS_Return4})) / \sqrt{(\text{AS_Risk1})\$(\text{Asset1})\$(\text{AS_Risk2})\$(\text{Asset2})\$(\text{AS_Risk3})\$(\text{Asset3})\$(\text{AS_Risk4})\$(\text{Asset4})\$(\text{AS_Risk4})\$(\text{Asset4})\$(\text{AS_Risk2})\$(\text{Asset4})\$(\text{AS_Risk2})}$

Variables

Name
AS_Return1
AS_Return2
AS_Return3
AS_Return4
AS_Risk1
AS_Risk2
AS_Risk3
AS_Risk4
Asset1
Asset2

Save Report Verify Run Exit

Result

Risk optimizer Report: Date Sun Nov 30 20:21:24 2008
 Problem Title: 63 Risk Optimizer - Portfolio Optimization (Sharpe Ratio)
 Number of variables is 4
 Number of functions is 2

Objective function will be Maximized

Itn No.	Objective Function	Binding Constrs	Super Basics	Infeas Constrs	Norm of Red. Grad	Hessian Cond. No.	Step Size	Degen Step
0	1.4971	1	3	0	0.48	1	0	
1	1.5286	1	3	0	0.27	8.6	0.086	
2	1.5408	1	3	0	0.013	4.5	0.059	

No. Name Initial Value Final Value Status Distance from Nearest Bound Lagrange Multiplier

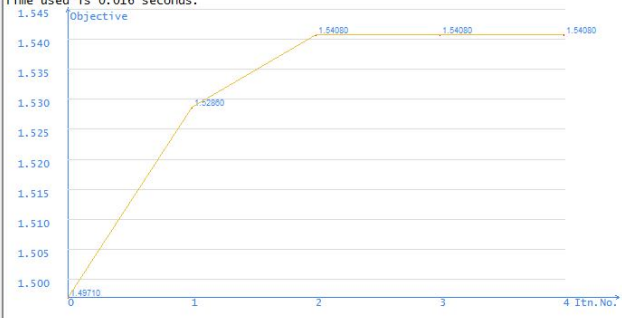
1	G				1e-010 :U	-3.5749e-006
2	G	1.4971	1.5408	UpperBnd Objective		

Variables:

No.	Name	Initial Value	Final Value	Status	Distance from Nearest Bound	Reduced Gradient
1	X	0.25	0.31006	Basic	0.08994 :U	
2	X	0.25	0.19338	SupBasic	0.09338 :L	-5.33e-007
3	X	0.25	0.18835	SupBasic	0.08835 :L	-1.08e-006
4	X	0.25	0.30821	supBasic	0.09179 :U	8.47e-006

Maximized objective function value is 1.54083
 Termination: INFORM = 0. Number of function evaluations 37
 Kuhn-Tucker conditions are satisfied to within 8.5e-006 for the current variable values.
 Relative change in the objective function value is 1.2e-005 for the last iteration.

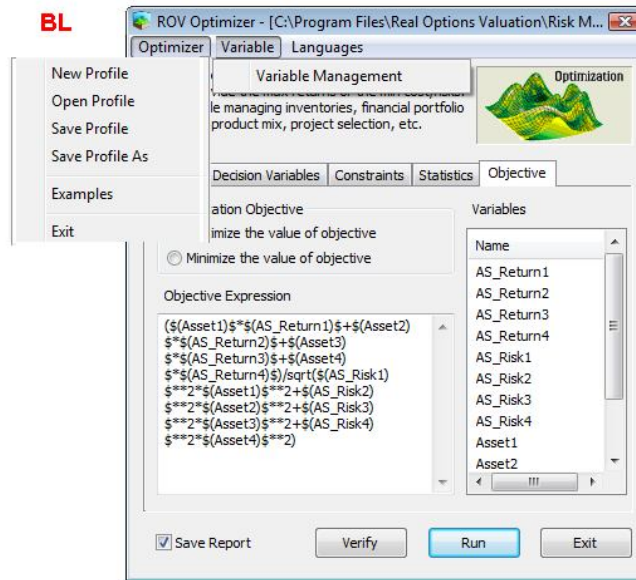
Time used is 0.016 seconds.



Itn. No.	Objective
0	1.49715
1	1.52860
2	1.54080
3	1.54080
4	1.54080

OK

There are also two important functionalities in ROV Optimizer, available in the File menu, including Examples and Variable Management under the Variable menu item **[BL]**. The Variable Management tool allows you to Add, Edit, or Delete variables. For instance, by clicking on ADD, the familiar Input Parameter Mapping tool appears, allowing you to link, compute, paste, simulate, or fit existing data for use in the optimization process. Finally, if Dynamic or Stochastic Optimization is selected, and if the variables have risk simulation assumptions associated with them, you can then access the Statistics tab, whereby you can make use of the simulated statistical properties to run optimization on.

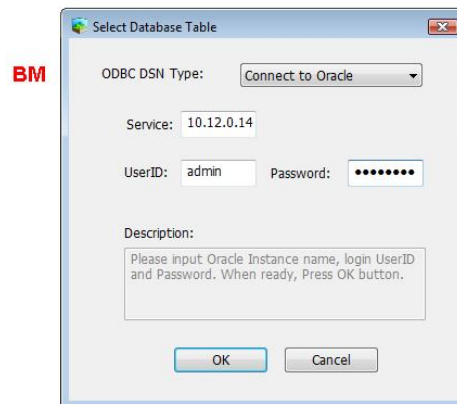


Linking to Other Databases

ROV Risk Modeler can link to different types of data using ODBC standard. When mapping to the database, you can select the Data Link in the method input. Click on Next and type in a name in the New Variable Name. Then select Open DB to open a database type and you can select your data source in terms of different data types that ROV Risk Modeler can connect to, including CSV, Excel, SQL Server, Oracle, User DSN, System DSN, and Connection Strings, with the ODBC data source standard.

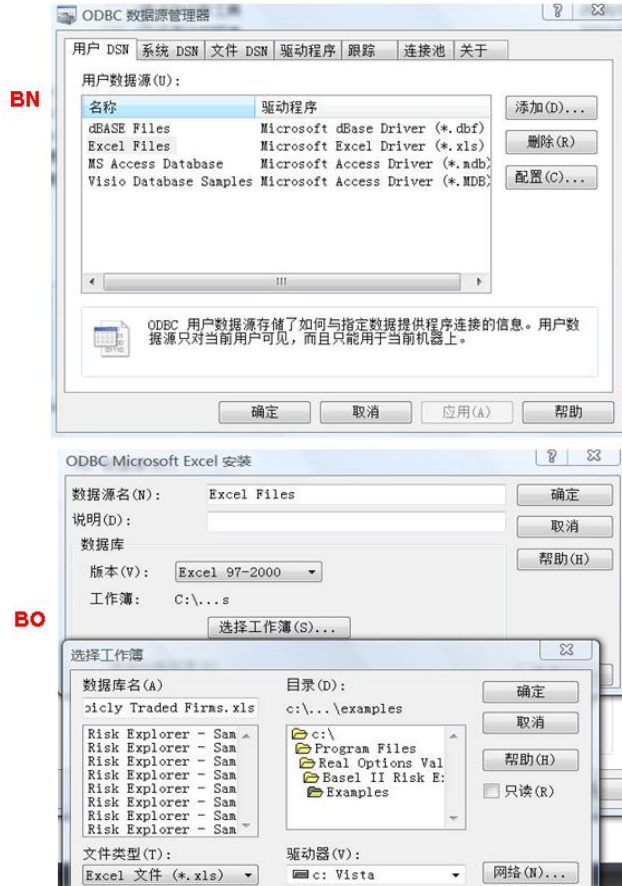
Case One: Link to Oracle

When you choose ODBC DSN as Connect to Oracle, input the local IP address of the Database Server and the relevant User ID and Password to log in **[BM]**. You can then find the available fields (variables) that can be selected. You can also write SQL sentences in the Condition box until the right variables and values are linked to the ROV Risk Modeler. It is important to notice that the database components must be Oracle version 7.3 or higher.



Case Two: Link to User DSN

Before you choose ODBC DSN as User DSN, you must set the DSN to a certain file in the first step. Click Start, select Control Panel, and select Management Tools where you can see the Data Source (ODBC) selection **[BN]**. Choose the User DSN tag, click Excel Files, and then click Configuration. In the new dialog, click on Choose Workshop, find an existing Excel file, and click OK **[BO]**. You can now return to the ROV Risk Modeler, map a variable using Data Link, click on Open DB and User DSN, choose Excel Files, and a list of tables will be listed. You can now map the existing table data to the Selected Fields.



TECHNICAL APPENDICES

Mathematical Probability Distributions

This section demonstrates the mathematical models and computations used in creating the Monte Carlo simulations. In order to get started with simulation, one first needs to understand the concept of probability distributions. To begin to understand probability, consider this example: You want to look at the distribution of nonexempt wages within one department of a large company. First, you gather raw data—in this case, the wages of each nonexempt employee in the department. Second, you organize the data into a meaningful format and plot the data as a frequency distribution on a chart. To create a frequency distribution, you divide the wages into group intervals and list these intervals on the chart's horizontal axis. Then you list the number or frequency of employees in each interval on the chart's vertical axis. Now you can easily see the distribution of nonexempt wages within the department. You can chart this data as a probability distribution. A probability distribution shows the number of employees in each interval as a fraction of the total number of employees. To create a probability distribution, you divide the number of employees in each interval by the total number of employees and list the results on the chart's vertical axis.

Probability distributions are either discrete or continuous. *Discrete probability distributions* describe distinct values, usually integers, with no intermediate values and are shown as a series of vertical bars. A discrete distribution, for example, might describe the number of heads in four flips of a coin as 0, 1, 2, 3, or 4. *Continuous probability distributions* are actually mathematical abstractions because they assume the existence of every possible intermediate value between two numbers; that is, a continuous distribution assumes there is an infinite number of values between any two points in the distribution. However, in many situations, you can effectively use a continuous distribution to approximate a discrete distribution even though the continuous model does not necessarily describe the situation exactly.

Probability Density Functions, Cumulative Distribution Functions, and Probability Mass Functions

In mathematics and Monte Carlo simulation, a probability density function (PDF) represents a *continuous* probability distribution in terms of integrals. If a probability distribution has a density of $f(x)$, then intuitively the infinitesimal interval of $[x, x + dx]$ has a probability of $f(x) dx$. The PDF therefore can be seen as a smoothed version of a probability histogram; that is, by providing an empirically large sample of a continuous random variable repeatedly, the histogram using very narrow ranges will resemble the random variable's PDF. The probability of the interval between $[a, b]$ is given by $\int_a^b f(x) dx$, which means

that the total integral of the function f must be 1.0. It is a common mistake to think of $f(a)$ as the probability of a . This is incorrect. In fact, $f(a)$ can sometimes be larger than 1—consider a uniform distribution between 0.0 and 0.5. The random variable x within this distribution will have $f(x)$ greater than 1. The probability in reality is the function $f(x)dx$ discussed previously, where dx is an infinitesimal amount.

The cumulative distribution function (CDF) is denoted as $F(x) = P(X \leq x)$ indicating the probability of X taking on a less than or equal value to x . Every CDF is monotonically increasing, is continuous from the right, and at the limits, has the following properties: $\lim_{x \rightarrow -\infty} F(x) = 0$ and $\lim_{x \rightarrow +\infty} F(x) = 1$. Further, the CDF is related to the PDF by $F(b) - F(a) = P(a \leq X \leq b) = \int_a^b f(x)dx$, where the PDF function f is the derivative of the CDF function F .

In probability theory, a probability mass function, or PMF, gives the probability that a *discrete* random variable is exactly equal to some value. The PMF differs from the PDF in that the values of the latter, defined only for continuous random variables, are not probabilities; rather, its integral over a set of possible values of the random variable is a probability. A random variable is discrete if its probability distribution is discrete and can be characterized by a PMF. Therefore, X is a discrete random variable if

$$\sum_u P(X = u) = 1 \text{ as } u \text{ runs through all possible values of the random variable } X.$$

Discrete Distributions

Following is a detailed listing of the different types of probability distributions that can be used in Monte Carlo simulation.

Bernoulli or Yes/No Distribution

The Bernoulli distribution is a discrete distribution with two outcomes (e.g., head or tails, success or failure, 0 or 1). The Bernoulli distribution is the binomial distribution with one trial and can be used to simulate Yes/No or Success/Failure conditions. This distribution is the fundamental building block of other more complex distributions. For instance:

Binomial distribution: Bernoulli distribution with higher number of n total trials and computes the probability of x successes within this total number of trials.

Geometric distribution: Bernoulli distribution with higher number of trials and computes the number of failures required before the first success occurs.

Negative binomial distribution: Bernoulli distribution with higher number of trials and computes the number of failures before the x th success occurs.

The mathematical constructs for the Bernoulli distribution are as follows:

$$P(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases}$$

or

$$P(x) = p^x (1-p)^{1-x}$$

$$\text{mean} = p$$

$$\text{standard deviation} = \sqrt{p(1-p)}$$

$$\text{skewness} = \frac{1-2p}{\sqrt{p(1-p)}}$$

$$\text{excess kurtosis} = \frac{6p^2 - 6p + 1}{p(1-p)}$$

The probability of success (p) is the only distributional parameter. Also, it is important to note that there is only one trial in the Bernoulli distribution, and the resulting simulated value is either 0 or 1. The input requirements are such that

Probability of Success > 0 and < 1 (that is, $0.0001 \leq p \leq 0.9999$).

Binomial Distribution

The binomial distribution describes the number of times a particular event occurs in a fixed number of trials, such as the number of heads in 10 flips of a coin or the number of defective items out of 50 items chosen. The three conditions underlying the binomial distribution are:

- For each trial, only two outcomes are possible that are mutually exclusive.
- The trials are independent—what happens in the first trial does not affect the next trial.
- The probability of an event occurring remains the same from trial to trial.

The mathematical constructs for the binomial distribution are as follows:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{(n-x)}$$

for $n > 0$; $x = 0, 1, 2, \dots, n$; and $0 < p < 1$

$$\text{mean} = np$$

$$\text{standard deviation} = \sqrt{np(1-p)}$$

$$\text{skewness} = \frac{1-2p}{\sqrt{np(1-p)}}$$

$$\text{excess kurtosis} = \frac{6p^2 - 6p + 1}{np(1-p)}$$

The probability of success (p) and the integer number of total trials (n) are the distributional parameters. The number of successful trials is denoted x . It is important to note that probabilities of success (p) of 0 or 1 are trivial conditions and do not require any simulations and, hence, are not allowed in the software. The input requirements are such that Probability of Success > 0 and < 1 (i.e., $0.0001 \leq p \leq 0.9999$), the Number of Trials ≥ 1 or positive integers and ≤ 1000 (for larger trials, use the normal distribution with the relevant computed binomial mean and standard deviation as the normal distribution's parameters).

Discrete Uniform

The discrete uniform distribution is also known as the *equally likely outcomes* distribution, where the distribution has a set of N elements, and each element has the same probability. This distribution is related to the uniform distribution but its elements are discrete and not continuous.

The mathematical constructs for the discrete uniform distribution are as follows:

$$P(x) = \frac{1}{N}$$

$$\text{mean} = \frac{N+1}{2} \text{ ranked value}$$

$$\text{standard deviation} = \sqrt{\frac{(N-1)(N+1)}{12}} \text{ ranked value}$$

skewness = 0 (that is, the distribution is perfectly symmetrical)

$$\text{excess kurtosis} = \frac{-6(N^2 + 1)}{5(N-1)(N+1)} \text{ ranked value}$$

The input requirements are such that Minimum < Maximum and both must be integers (negative integers and zero are allowed).

Geometric Distribution

The geometric distribution describes the number of trials until the first successful occurrence, such as the number of times you need to spin a roulette wheel before you win. The three conditions underlying the geometric distribution are:

- The number of trials is not fixed.
- The trials continue until the first success.
- The probability of success is the same from trial to trial.

The mathematical constructs for the geometric distribution are as follows:

$$P(x) = p(1-p)^{x-1} \text{ for } 0 < p < 1 \text{ and } x = 1, 2, \dots, n$$

$$\text{mean} = \frac{1}{p} - 1$$

$$\text{standard deviation} = \sqrt{\frac{1-p}{p^2}}$$

$$\text{skewness} = \frac{2-p}{\sqrt{1-p}}$$

$$\text{excess kurtosis} = \frac{p^2 - 6p + 6}{1-p}$$

The probability of success (p) is the only distributional parameter. The number of successful trials simulated is denoted x , which can only take on positive integers. The input requirements are such that Probability of success > 0 and < 1 (i.e., $0.0001 \leq p \leq 0.9999$). It is important to note that probabilities of success (p) of 0 or 1 are trivial conditions and do not require any simulations and, hence, are not allowed in the software.

Hypergeometric Distribution

The hypergeometric distribution is similar to the binomial distribution in that both describe the number of times a particular event occurs in a fixed number of trials. The difference is that binomial distribution

trials are independent, whereas hypergeometric distribution trials change the probability for each subsequent trial and are called *trials without replacement*. For example, suppose a box of manufactured parts is known to contain some defective parts. You choose a part from the box, find it is defective, and remove the part from the box. If you choose another part from the box, the probability that it is defective is somewhat lower than for the first part because you have removed a defective part. If you had replaced the defective part, the probabilities would have remained the same, and the process would have satisfied the conditions for a binomial distribution.

The three conditions underlying the hypergeometric distribution are:

- The total number of items or elements (the population size) is a fixed number, a finite population. The population size must be less than or equal to 1,750.
- The sample size (the number of trials) represents a portion of the population.
- The known initial probability of success in the population changes after each trial.

The mathematical constructs for the hypergeometric distribution are as follows:

$$P(x) = \frac{\frac{(N_x)!}{x!(N_x - x)!} \frac{(N - N_x)!}{(n - x)!(N - N_x - n + x)!}}{\frac{N!}{n!(N - n)!}}$$

for $x = \text{Max}(n - (N - N_x), 0), \dots, \text{Min}(n, N_x)$

$$\text{mean} = \frac{N_x n}{N}$$

$$\text{standard deviation} = \sqrt{\frac{(N - N_x)N_x n(N - n)}{N^2(N - 1)}}$$

skewness =

$$\frac{(N - 2N_x)(N - 2n)}{N - 2} \sqrt{\frac{N - 1}{(N - N_x)N_x n(N - n)}}$$

$$\text{excess kurtosis} = \frac{V(N, N_x, n)}{(N - N_x) N_x n(-3 + N)(-2 + N)(-N + n)} \text{ where}$$

$$\begin{aligned} V(N, N_x, n) = & (N - N_x)^3 - (N - N_x)^5 + 3(N - N_x)^2 N_x - 6(N - N_x)^3 N_x \\ & + (N - N_x)^4 N_x + 3(N - N_x) N_x^2 - 12(N - N_x)^2 N_x^2 + 8(N - N_x)^3 N_x^2 + N_x^3 \\ & - 6(N - N_x) N_x^3 + 8(N - N_x)^2 N_x^3 + (N - N_x) N_x^4 - N_x^5 - 6(N - N_x)^3 N_x \\ & + 6(N - N_x)^4 N_x + 18(N - N_x)^2 N_x n - 6(N - N_x)^3 N_x n + 18(N - N_x) N_x^2 n \\ & - 24(N - N_x)^2 N_x^2 n - 6(N - N_x)^3 n - 6(N - N_x) N_x^3 n + 6N_x^4 n + 6(N - N_x)^2 n^2 \\ & - 6(N - N_x)^3 n^2 - 24(N - N_x) N_x n^2 + 12(N - N_x)^2 N_x n^2 + 6N_x^2 n^2 \\ & + 12(N - N_x) N_x^2 n^2 - 6N_x^3 n^2 \end{aligned}$$

The number of items in the population (N), the number of trials sampled (n), and number of items in the population that have the successful trait (N_x) are the distributional parameters. The number of successful trials is denoted x . The input requirements are such that Population ≥ 2 and integer, Trials > 0 and integer.

Successes > 0 and integer, Population > Successes

Trials < Population and Population < 1750.

Negative Binomial Distribution

The negative binomial distribution is useful for modeling the distribution of the number of trials until the r th successful occurrence, such as the number of sales calls you need to make to close a total of 10 orders. It is essentially a *superdistribution* of the geometric distribution. This distribution shows the probabilities of each number of trials in excess of r to produce the required success r .

The three conditions underlying the negative binomial distribution are:

- The number of trials is not fixed.
- The trials continue until the r th success.
- The probability of success is the same from trial to trial.

The mathematical constructs for the negative binomial distribution are as follows:

$$P(x) = \frac{(x+r-1)!}{(r-1)!x!} p^r (1-p)^x$$

for $x = r, r+1, \dots$; and $0 < p < 1$

$$\text{mean} = \frac{r(1-p)}{p}$$

$$\text{standard deviation} = \sqrt{\frac{r(1-p)}{p^2}}$$

$$\text{skewness} = \frac{2-p}{\sqrt{r(1-p)}}$$

$$\text{excess kurtosis} = \frac{p^2 - 6p + 6}{r(1-p)}$$

Probability of success (p) and required successes (r) are the distributional parameters. Where the input requirements are such that Successes required must be positive integers > 0 and < 8000, Probability of success > 0 and < 1 (i.e., $0.0001 \leq p \leq 0.9999$). It is important to note that probabilities of success (p) of 0 or 1 are trivial conditions and do not require any simulations and, hence, are not allowed in the software.

Poisson Distribution

The Poisson distribution describes the number of times an event occurs in a given interval, such as the number of telephone calls per minute or the number of errors per page in a document.

The three conditions underlying the Poisson distribution are:

- The number of possible occurrences in any interval is unlimited.
- The occurrences are independent. The number of occurrences in one interval does not affect the number of occurrences in other intervals.
- The average number of occurrences must remain the same from interval to interval.

The mathematical constructs for the Poisson are as follows:

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!} \text{ for } x \text{ and } \lambda > 0$$

$$\text{mean} = \lambda$$

$$\text{standard deviation} = \sqrt{\lambda}$$

$$\text{skewness} = \frac{1}{\sqrt{\lambda}}$$

$$\text{excess kurtosis} = \frac{1}{\lambda}$$

Rate (λ) is the only distributional parameter, and the input requirements are such that Rate > 0 and ≤ 1000 (i.e., 0.0001 ≤ rate ≤ 1000).

Continuous Distributions

Beta Distribution

The beta distribution is very flexible and is commonly used to represent variability over a fixed range. One of the more important applications of the beta distribution is its use as a conjugate distribution for the parameter of a Bernoulli distribution. In this application, the beta distribution is used to represent the uncertainty in the probability of occurrence of an event. It is also used to describe empirical data and predict the random behavior of percentages and fractions, as the range of outcomes is typically between 0 and 1. The value of the beta distribution lies in the wide variety of shapes it can assume when you vary the two parameters, alpha and beta. If the parameters are equal, the distribution is symmetrical. If either parameter is 1 and the other parameter is greater than 1, the distribution is J-shaped. If alpha is less than beta, the distribution is said to be positively skewed (most of the values are near the minimum value). If alpha is greater than beta, the distribution is negatively skewed (most of the values are near the maximum value). The mathematical constructs for the beta distribution are as follows:

$$f(x) = \frac{(x)^{(\alpha-1)}(1-x)^{(\beta-1)}}{\left[\frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \right]} \text{ for } \alpha > 0; \beta > 0; x > 0$$

$$\text{mean} = \frac{\alpha}{\alpha + \beta}$$

$$\text{standard deviation} = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(1 + \alpha + \beta)}}$$

$$\text{skewness} = \frac{2(\beta - \alpha)\sqrt{1 + \alpha + \beta}}{(2 + \alpha + \beta)\sqrt{\alpha\beta}}$$

$$\text{excess kurtosis} = \frac{3(\alpha + \beta + 1)[\alpha\beta(\alpha + \beta - 6) + 2(\alpha + \beta)^2]}{\alpha\beta(\alpha + \beta + 2)(\alpha + \beta + 3)} - 3$$

Alpha (α) and beta (β) are the two distributional shape parameters, and Γ is the gamma function. The two conditions underlying the beta distribution are:

- The uncertain variable is a random value between 0 and a positive value.
- The shape of the distribution can be specified using two positive values.

Input requirements:

Alpha and beta > 0 and can be any positive value.

Cauchy Distribution or Lorentzian or Breit-Wigner Distribution

The Cauchy distribution, also called the Lorentzian or Breit-Wigner distribution, is a continuous distribution describing resonance behavior. It also describes the distribution of horizontal distances at which a line segment tilted at a random angle cuts the x-axis.

The mathematical constructs for the cauchy or Lorentzian distribution are as follows:

$$f(x) = \frac{1}{\pi} \frac{\gamma / 2}{(x - m)^2 + \gamma^2 / 4}$$

The cauchy distribution is a special case because it does not have any theoretical moments (mean, standard deviation, skewness, and kurtosis) as they are all undefined. Mode location (m) and scale (γ) are the only two parameters in this distribution. The location parameter specifies the peak or mode of the distribution, while the scale parameter specifies the half-width at half-maximum of the distribution. In addition, the mean and variance of a cauchy or Lorentzian distribution are undefined. In addition, the cauchy distribution is the Student's t-distribution with only 1 degree of freedom. This distribution is also constructed by taking the ratio of two standard normal distributions (normal distributions with a mean of zero and a variance of one) that are independent of one another. The input requirements are such that Location can be any value, whereas Scale > 0 and can be any positive value.

Chi-Square Distribution

The chi-square distribution is a probability distribution used predominantly in hypothesis testing and is related to the gamma distribution and the standard normal distribution. For instance, the sums of independent normal distributions are distributed as a chi-square (χ^2) with k degrees of freedom:

$$Z_1^2 + Z_2^2 + \dots + Z_k^2 \sim \chi_k^2$$

The mathematical constructs for the chi-square distribution are as follows:

$$f(x) = \frac{2^{-k/2}}{\Gamma(k/2)} x^{k/2-1} e^{-x/2} \quad \text{for all } x > 0$$

mean = k

standard deviation = $\sqrt{2k}$

$$\text{skewness} = 2\sqrt{\frac{2}{k}}$$

$$\text{excess kurtosis} = \frac{12}{k}$$

Γ is the gamma function. Degrees of freedom k is the only distributional parameter.

The chi-square distribution can also be modeled using a gamma distribution by setting the shape parameter = $\frac{k}{2}$ and scale = $2S^2$ where S is the scale. The input requirements are such that Degrees of freedom > 1 and must be an integer < 1000 .

Exponential Distribution

The exponential distribution is widely used to describe events recurring at random points in time, such as the time between failures of electronic equipment or the time between arrivals at a service booth. It is related to the Poisson distribution, which describes the number of occurrences of an event in a given interval of time. An important characteristic of the exponential distribution is the “memoryless” property, which means that the future lifetime of a given object has the same distribution regardless of the time it existed. In other words, time has no effect on future outcomes. The mathematical constructs for the exponential distribution are as follows:

$$f(x) = \lambda e^{-\lambda x} \quad \text{for } x \geq 0; \lambda > 0$$

$$\text{mean} = \frac{1}{\lambda}$$

$$\text{standard deviation} = \frac{1}{\lambda}$$

skewness = 2 (this value applies to all success rate λ inputs)

excess kurtosis = 6 (this value applies to all success rate λ inputs)

Success rate (λ) is the only distributional parameter. The number of successful trials is denoted x .

The condition underlying the exponential distribution is:

The exponential distribution describes the amount of time between occurrences.

Input requirements: Rate > 0 and ≤ 300

Extreme Value Distribution or Gumbel Distribution

The extreme value distribution (Type 1) is commonly used to describe the largest value of a response over a period of time, for example, in flood flows, rainfall, and earthquakes. Other applications include the breaking strengths of materials, construction design, and aircraft loads and tolerances. The extreme value distribution is also known as the Gumbel distribution.

The mathematical constructs for the extreme value distribution are as follows:

$$f(x) = \frac{1}{\beta} z e^{-z} \text{ where } z = e^{\frac{x-m}{\beta}}$$

for $\beta > 0$; and any value of x and m

$$\text{mean} = m + 0.577215 \beta$$

$$\text{standard deviation} = \sqrt{\frac{1}{6} \pi^2 \beta^2}$$

$$\text{skewness} = \frac{12\sqrt{6}(1.2020569)}{\pi^3} = 1.13955 \text{ (this applies for all values of mode and scale)}$$

$$\text{excess kurtosis} = 5.4 \text{ (this applies for all values of mode and scale)}$$

Mode (m) and scale (β) are the distributional parameters. There are two standard parameters for the extreme value distribution: mode and scale. The mode parameter is the most likely value for the variable (the highest point on the probability distribution). The scale parameter is a number greater than 0. The larger the scale parameter, the greater the variance. The input requirements are such that Mode can be any value and Scale > 0 .

F Distribution or Fisher-Snedecor Distribution

The F distribution, also known as the Fisher-Snedecor distribution, is another continuous distribution used most frequently for hypothesis testing. Specifically, it is used to test the statistical difference between two variances in analysis of variance tests and likelihood ratio tests. The F distribution with the numerator degree of freedom n and denominator degree of freedom m is related to the chi-square distribution in that:

$$\frac{\chi_n^2 / n}{\chi_m^2 / m} \sim F_{n,m}$$

$$f(x) = \frac{\Gamma\left(\frac{n+m}{2}\right) \left(\frac{n}{m}\right)^{n/2} x^{n/2-1}}{\Gamma\left(\frac{n}{2}\right) \Gamma\left(\frac{m}{2}\right) \left[x\left(\frac{n}{m}\right) + 1\right]^{(n+m)/2}}$$

or

$$\text{mean} = \frac{m}{m-2}$$

$$\text{standard deviation} = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)} \text{ for all } m > 4$$

$$\text{skewness} = \frac{2(m+2n-2)}{m-6} \sqrt{\frac{2(m-4)}{n(m+n-2)}}$$

excess kurtosis =

$$\frac{12(-16+20m-8m^2+m^3+44n-32mn+5m^2n-22n^2+5mn^2)}{n(m-6)(m-8)(n+m-2)}$$

The numerator degree of freedom n and denominator degree of freedom m are the only distributional parameters. The input requirements are such that Degrees of freedom numerator and degrees of freedom denominator both > 0 integers.

Gamma Distribution (Erlang Distribution)

The gamma distribution applies to a wide range of physical quantities and is related to other distributions: lognormal, exponential, Pascal, Erlang, Poisson, and chi-square. It is used in meteorological processes to represent pollutant concentrations and precipitation quantities. The gamma distribution is also used to measure the time between the occurrence of events when the event process is not completely random. Other applications of the gamma distribution include inventory control, economic theory, and insurance risk theory.

The gamma distribution is most often used as the distribution of the amount of time until the r th occurrence of an event in a Poisson process. When used in this fashion, the three conditions underlying the gamma distribution are:

- The number of possible occurrences in any unit of measurement is not limited to a fixed number.
- The occurrences are independent. The number of occurrences in one unit of measurement does not affect the number of occurrences in other units.
- The average number of occurrences must remain the same from unit to unit.

The mathematical constructs for the gamma distribution are as follows:

$$f(x) = \frac{\left(\frac{x}{\beta}\right)^{\alpha-1} e^{-\frac{x}{\beta}}}{\Gamma(\alpha)\beta} \quad \text{with any value of } \alpha > 0 \text{ and } \beta > 0$$

$$\text{mean} = \alpha\beta$$

$$\text{standard deviation} = \sqrt{\alpha\beta^2}$$

$$\text{skewness} = \frac{2}{\sqrt{\alpha}}$$

$$\text{excess kurtosis} = \frac{6}{\alpha}$$

Shape parameter alpha (α) and scale parameter beta (β) are the distributional parameters, and Γ is the gamma function. When the alpha parameter is a positive integer, the gamma distribution is called the Erlang distribution, used to predict waiting times in queuing systems, where the Erlang distribution is the sum of independent and identically distributed random variables each having a memoryless exponential distribution. Setting n as the number of these random variables, the mathematical construct of the Erlang distribution is:

$$f(x) = \frac{x^{n-1} e^{-x}}{(n-1)!} \quad \text{for all } x > 0 \text{ and all positive integers of } n, \text{ where the input requirements are such that}$$

Scale Beta > 0 and can be any positive value, Shape Alpha ≥ 0.05 and any positive value, and Location can be any value.

Logistic Distribution

The logistic distribution is commonly used to describe growth, that is, the size of a population expressed as a function of a time variable. It also can be used to describe chemical reactions and the course of growth for a population or individual.

The mathematical constructs for the logistic distribution are as follows:

$$f(x) = \frac{e^{-\frac{\mu-x}{\alpha}}}{\alpha \left[1 + e^{-\frac{\mu-x}{\alpha}} \right]^2} \text{ for any value of } \alpha \text{ of } \mu$$

mean = μ

$$\text{standard deviation} = \sqrt{\frac{1}{3} \pi^2 \alpha^2}$$

skewness = 0 (this applies to all mean and scale inputs)

excess kurtosis = 1.2 (this applies to all mean and scale inputs)

Mean (μ) and scale (α) are the distributional parameters. There are two standard parameters for the logistic distribution: mean and scale. The mean parameter is the average value, which for this distribution is the same as the mode, because this distribution is symmetrical. The scale parameter is a number greater than 0. The larger the scale parameter, the greater the variance.

Input requirements:

Scale > 0 and can be any positive value

Mean can be any value

Lognormal Distribution

The lognormal distribution is widely used in situations where values are positively skewed, for example, in financial analysis for security valuation or in real estate for property valuation, and where values cannot fall below zero. Stock prices are usually positively skewed rather than normally (symmetrically) distributed. Stock prices exhibit this trend because they cannot fall below the lower limit of zero but might increase to any price without limit. Similarly, real estate prices illustrate positive skewness and are lognormally distributed as property values cannot become negative.

The three conditions underlying the lognormal distribution are:

- The uncertain variable can increase without limits but cannot fall below zero.
- The uncertain variable is positively skewed, with most of the values near the lower limit.
- The natural logarithm of the uncertain variable yields a normal distribution.

Generally, if the coefficient of variability is greater than 30 percent, use a lognormal distribution. Otherwise, use the normal distribution.

The mathematical constructs for the lognormal distribution are as follows:

$$f(x) = \frac{1}{x\sqrt{2\pi} \ln(\sigma)} e^{-\frac{[\ln(x)-\ln(\mu)]^2}{2[\ln(\sigma)]^2}}$$

for $x > 0$; $\mu > 0$ and $\sigma > 0$

$$\text{mean} = \exp\left(\mu + \frac{\sigma^2}{2}\right)$$

$$\text{standard deviation} = \sqrt{\exp(\sigma^2 + 2\mu)[\exp(\sigma^2) - 1]}$$

$$\text{skewness} = \left[\sqrt{\exp(\sigma^2) - 1}\right](2 + \exp(\sigma^2))$$

$$\text{excess kurtosis} = \exp(4\sigma^2) + 2\exp(3\sigma^2) + 3\exp(2\sigma^2) - 6$$

Mean (μ) and standard deviation (σ) are the distributional parameters. The input requirements are such that Mean and Standard deviation are both > 0 and can be any positive value. By default, the lognormal distribution uses the arithmetic mean and standard deviation. For applications for which historical data are available, it is more appropriate to use either the logarithmic mean and standard deviation or the geometric mean and standard deviation.

Normal Distribution

The normal distribution is the most important distribution in probability theory because it describes many natural phenomena, such as people's IQs or heights. Decision makers can use the normal distribution to describe uncertain variables such as the inflation rate or the future price of gasoline.

The three conditions underlying the normal distribution are:

- Some value of the uncertain variable is the most likely (the mean of the distribution).
- The uncertain variable could as likely be above the mean as it could be below the mean (symmetrical about the mean).
- The uncertain variable is more likely to be in the vicinity of the mean than further away.

The mathematical constructs for the normal distribution are as follows:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for all values of } x \text{ and } \mu \text{ while } \sigma > 0$$

$$\text{mean} = \mu$$

$$\text{standard deviation} = \sigma$$

$$\text{skewness} = 0 \text{ (this applies to all inputs of mean and standard deviation)}$$

$$\text{excess kurtosis} = 0 \text{ (this applies to all inputs of mean and standard deviation)}$$

Mean (μ) and standard deviation (σ) are the distributional parameters. The input requirements are such that Standard deviation > 0 and can be any positive value and Mean can be any value.

Pareto Distribution

The Pareto distribution is widely used for the investigation of distributions associated with such empirical phenomena as city population sizes, the occurrence of natural resources, the size of companies, personal incomes, stock price fluctuations, and error clustering in communication circuits.

The mathematical constructs for the Pareto are as follows:

$$f(x) = \frac{\beta L^\beta}{x^{(1+\beta)}} \text{ for } x > L$$

$$\text{mean} = \frac{\beta L}{\beta - 1}$$

$$\text{standard deviation} = \sqrt{\frac{\beta L^2}{(\beta - 1)^2 (\beta - 2)}}$$

$$\text{skewness} = \sqrt{\frac{\beta - 2}{\beta} \left[\frac{2(\beta + 1)}{\beta - 3} \right]}$$

$$\text{excess kurtosis} = \frac{6(\beta^3 + \beta^2 - 6\beta - 2)}{\beta(\beta - 3)(\beta - 4)}$$

Location (L) and shape (β) are the distributional parameters.

There are two standard parameters for the Pareto distribution: location and shape. The location parameter is the lower bound for the variable. After you select the location parameter, you can estimate the shape parameter. The shape parameter is a number greater than 0, usually greater than 1. The larger the shape parameter, the smaller the variance and the thicker the right tail of the distribution. The input requirements are such that Location > 0 and can be any positive value while Shape ≥ 0.05 .

Student's t-Distribution

The Student's t-distribution is the most widely used distribution in hypothesis test. This distribution is used to estimate the mean of a normally distributed population when the sample size is small and to test the statistical significance of the difference between two sample means or confidence intervals for small sample sizes.

The mathematical constructs for the t-distribution are as follows:

$$f(t) = \frac{\Gamma[(r+1)/2]}{\sqrt{r\pi} \Gamma[r/2]} (1 + t^2/r)^{-(r+1)/2}$$

mean = 0 (this applies to all degrees of freedom r except if the distribution is shifted to another nonzero central location)

$$\text{standard deviation} = \sqrt{\frac{r}{r-2}}$$

skewness = 0

$$\text{excess kurtosis} = \frac{6}{r-4} \text{ for all } r > 4$$

where $t = \frac{x - \bar{x}}{s}$ and Γ is the gamma function.

Degree of freedom, r , is the only distributional parameter. The t-distribution is related to the F-distribution as follows: the square of a value of t with r degrees of freedom is distributed as F with 1 and r degrees of freedom. The overall shape of the probability density function of the t-distribution also resembles the bell shape of a normally distributed variable with mean 0 and variance 1, except that it is a bit lower and wider or is leptokurtic (fat tails at the ends and peaked center). As the number of degrees of freedom grows (say, above 30), the t-distribution approaches the normal distribution with mean 0 and variance 1. The input requirements are such that Degrees of freedom ≥ 1 and must be an integer.

Triangular Distribution

The triangular distribution describes a situation where you know the minimum, maximum, and most likely values to occur. For example, you could describe the number of cars sold per week when past sales show the minimum, maximum, and usual number of cars sold.

The three conditions underlying the triangular distribution are:

- The minimum number of items is fixed.
- The maximum number of items is fixed.
- The most likely number of items falls between the minimum and maximum values, forming a triangular-shaped distribution, which shows that values near the minimum and maximum are less likely to occur than those near the most-likely value.

The mathematical constructs for the triangular distribution are as follows:

$$f(x) = \begin{cases} \frac{2(x - \text{Min})}{(\text{Max} - \text{Min})(\text{Likely} - \text{Min})} & \text{for } \text{Min} < x < \text{Likely} \\ \frac{2(\text{Max} - x)}{(\text{Max} - \text{Min})(\text{Max} - \text{Likely})} & \text{for } \text{Likely} < x < \text{Max} \end{cases}$$

$$\text{mean} = \frac{1}{3}(\text{Min} + \text{Likely} + \text{Max})$$

$$\text{standard deviation} = \sqrt{\frac{1}{18}(\text{Min}^2 + \text{Likely}^2 + \text{Max}^2 - \text{MinMax} - \text{MinLikely} - \text{MaxLikely})}$$

$$\text{skewness} = \frac{\sqrt{2}(\text{Min} + \text{Max} - 2\text{Likely})(2\text{Min} - \text{Max} - \text{Likely})(\text{Min} - 2\text{Max} + \text{Likely})}{5(\text{Min}^2 + \text{Max}^2 + \text{Likely}^2 - \text{MinMax} - \text{MinLikely} - \text{MaxLikely})^{3/2}}$$

$$\text{excess kurtosis} = -0.6$$

Minimum (Min), most likely (Likely), and maximum (Max) are the distributional parameters, and the input requirements are such that $\text{Min} \leq \text{Most Likely} \leq \text{Max}$ and can take any value, and $\text{Min} < \text{Max}$ and can take any value.

Uniform Distribution

With the uniform distribution, all values fall between the minimum and maximum and occur with equal likelihood.

The three conditions underlying the uniform distribution are:

- The minimum value is fixed.
- The maximum value is fixed.
- All values between the minimum and maximum occur with equal likelihood.

The mathematical constructs for the uniform distribution are as follows:

$$f(x) = \frac{1}{Max - Min}$$

for all values such that $Min < Max$

$$\text{mean} = \frac{Min + Max}{2}$$

$$\text{standard deviation} = \sqrt{\frac{(Max - Min)^2}{12}}$$

skewness = 0

excess kurtosis = -1.2 (this applies to all inputs of Min and Max)

Maximum value (Max) and minimum value (Min) are the distributional parameters. The input requirements are such that $Min < Max$ and can take any value.

Weibull Distribution (Rayleigh Distribution)

The Weibull distribution describes data resulting from life and fatigue tests. It is commonly used to describe failure time in reliability studies as well as the breaking strengths of materials in reliability and quality control tests. Weibull distributions are also used to represent various physical quantities, such as wind speed. The Weibull distribution is a family of distributions that can assume the properties of several other distributions. For example, depending on the shape parameter you define, the Weibull distribution can be used to model the exponential and Rayleigh distributions, among others. The Weibull distribution is very flexible. When the Weibull shape parameter is equal to 1.0, the Weibull distribution is identical to the exponential distribution. The Weibull location parameter lets you set up an exponential distribution to start at a location other than 0.0. When the shape parameter is less than 1.0, the Weibull distribution becomes a steeply declining curve. A manufacturer might find this effect useful in describing part failures during a burn-in period.

The mathematical constructs for the Weibull distribution are as follows:

$$f(x) = \frac{\alpha}{\beta} \left[\frac{x}{\beta} \right]^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^\alpha}$$

$$\text{mean} = \beta \Gamma(1 + \alpha^{-1})$$

$$\text{standard deviation} = \beta^2 [\Gamma(1 + 2\alpha^{-1}) - \Gamma^2(1 + \alpha^{-1})]$$

$$\text{skewness} = \frac{2\Gamma^3(1+\beta^{-1}) - 3\Gamma(1+\beta^{-1})\Gamma(1+2\beta^{-1}) + \Gamma(1+3\beta^{-1})}{[\Gamma(1+2\beta^{-1}) - \Gamma^2(1+\beta^{-1})]^{3/2}}$$

excess kurtosis =

$$\frac{-6\Gamma^4(1+\beta^{-1}) + 12\Gamma^2(1+\beta^{-1})\Gamma(1+2\beta^{-1}) - 3\Gamma^2(1+2\beta^{-1}) - 4\Gamma(1+\beta^{-1})\Gamma(1+3\beta^{-1}) + \Gamma(1+4\beta^{-1})}{[\Gamma(1+2\beta^{-1}) - \Gamma^2(1+\beta^{-1})]^2}$$

Location (L), shape (α), and scale (β) are the distributional parameters, and Γ is the Gamma function. The input requirements are such that Scale > 0 and can be any positive value, and Shape ≥ 0.05 and Location can take on any value.

QUICK TECHNICAL DESCRIPTIONS OF MODELS

The following are quick technical descriptions of each of the models and methodologies available in the QDM main module. Whenever appropriate, several models are discussed as a group. This technical appendix only provides high-level explanations of these models and methods. For hands-on applications, run QDM, open one of the many examples provided, and review how the data are setup and the models run.

ANOVA: Randomized Blocks N-Treatments, 1-Factor N-Treatments, 2-Way ANOVA

The One-Way ANOVA for Single Factor with Multiple Treatments is an extension of the two-variable t-test, looking at multiple variables simultaneously. The ANOVA is appropriate when the sampling distribution is assumed to be approximately normal. ANOVA can be applied to only the two-tailed hypothesis test. A two-tailed hypothesis tests the null hypothesis such that the population means of each treatment is statistically identical to the rest of the group, which means that there is no effect among the different treatment groups. The alternative hypothesis is such that the real population means are statistically different from one another when tested using the sample dataset. To illustrate, suppose that three different drug indications ($T = 3$) were developed and tested on 100 patients each ($N = 100$). The One-Way ANOVA can be applied to test if these three drugs are all equally effective statistically. If the calculated p-value is less than or equal to the significance level used in the test, then reject the null hypothesis and conclude that there is a significant difference among the different Treatments. Otherwise, the Treatments are all equally effective.

The One-Way Randomized Block ANOVA is appropriate when the sampling distribution is assumed to be approximately normal and when there exists a Block variable for which ANOVA will Control (i.e., Block the effects of this variable by controlling it in the experiment). ANOVA can be applied to only the two-tailed hypothesis test. This analysis can test for the effects of both the Treatments as well as the effectiveness of the Control or Block variable. If the calculated p-value for the Treatment is less than or equal to the significance level used in the test, then reject the null hypothesis and conclude that there is a significant difference among the different Treatments. If the calculated p-value for the Block variable is less than or equal to the significance level used in the test, then reject the null hypothesis and conclude that there is a significant difference among the different Block variables. To illustrate, suppose that three different headlamp designs ($T = 3$) were developed and tested on 4 groups of volunteer drivers grouped by their age ($B = 4$). The One-Way Randomized Block ANOVA can be applied to test if these three headlamps are all equally effective statistically when tested using the volunteers' driving test grades. Otherwise, the Treatments are all equally effective. This test can determine if the differences occur because of the Treatment (that the type of headlamp will determine differences in driving test scores) or from the Block or controlled variable (that age may yield different driving abilities).

The Two-Way ANOVA is an extension of the Single Factor and Randomized Block ANOVA by simultaneously examining the effects of two factors on the dependent variable, along with the effects of interactions between the different levels of these two factors. Unlike the randomized block design, this model examines the interactions between different levels of the factors, or independent variables. In a two-factor experiment, interaction exists when the effect of a level for one factor depends on which level

of the other factor is present. There are three sets of null and alternate hypotheses to be tested in the two-way analysis of variance.

The first test is on the first independent variable, where the null hypothesis is that no level of the first factor has an effect on the dependent variable. The alternate hypothesis is that there is at least one level of the first factor having an effect on the dependent variable. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

The second test is on the second independent variable, where the null hypothesis is that no level of the second factor has an effect on the dependent variable. The alternate hypothesis is that there is at least one level of the second factor having an effect on the dependent variable. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

The third test is on the interaction of both the first and second independent variables, where the null hypothesis is that there are no interacting effects between levels of the first and second factors. The alternate hypothesis is that there is at least one combination of levels of the first and second factors having an effect on the dependent variable. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

ARIMA

One very powerful advanced times-series forecasting tool is the ARIMA or *Auto Regressive Integrated Moving Average* approach, which assembles three separate tools into a comprehensive model. The first tool segment is the autoregressive or “AR” term, which corresponds to the number of lagged value of the residual in the unconditional forecast model. In essence, the model captures the historical variation of actual data to a forecasting model and uses this variation or residual to create a better predicting model. The second tool segment is the integration order or the “I” term. This integration term corresponds to the number of differencings the time series to be forecasted goes through to make the data stationary. This element accounts for any nonlinear growth rates existing in the data. The third tool segment is the moving average or “MA” term, which is essentially the moving average of lagged forecast errors. By incorporating this lagged forecast errors term, the model in essence learns from its forecast errors or mistakes and corrects for them through a moving average calculation. The ARIMA model follows the Box-Jenkins methodology with each term representing steps taken in the model construction until only random noise remains. Also, ARIMA modeling uses correlation techniques in generating forecasts. ARIMA can be used to model patterns that may not be visible in plotted data. In addition, ARIMA models can be mixed with exogenous variables, but make sure that the exogenous variables have enough data points to cover the additional number of periods to forecast. Finally, be aware that ARIMA cannot and should not be used to forecast stochastic processes or time-series data that are stochastic in nature—use the Stochastic Process module to forecast instead.

There are many reasons why an ARIMA model is superior to common time-series analysis and multivariate regressions. The common finding in time-series analysis and multivariate regression is that the error residuals are correlated with their own lagged values. This serial correlation violates the

standard assumption of regression theory that disturbances are not correlated with other disturbances. The primary problems associated with serial correlation are:

- Regression analysis and basic time-series analysis are no longer efficient among the different linear estimators. However, as the error residuals can help to predict current error residuals, we can take advantage of this information to form a better prediction of the dependent variable using ARIMA.
- Standard errors computed using the regression and time-series formula are not correct and are generally understated. If there are lagged dependent variables set as the regressors, regression estimates are biased and inconsistent but can be fixed using ARIMA.

Autoregressive Integrated Moving Average or ARIMA(p,d,q) models are the extension of the AR model that uses three components for modeling the serial correlation in the time-series data. The first component is the autoregressive (AR) term. The AR(p) model uses the p lags of the time series in the equation. An AR(p) model has the form: $y_t = a_1 y_{t-1} + \dots + a_p y_{t-p} + e_t$. The second component is the integration (d) order term. Each integration order corresponds to differencing the time series. I(1) means differencing the data once; I(d) means differencing the data d times. The third component is the moving average (MA) term. The MA(q) model uses the q lags of the forecast errors to improve the forecast. An MA(q) model has the form: $y_t = e_t + b_1 e_{t-1} + \dots + b_q e_{t-q}$. Finally, an ARMA(p,q) model has the combined form: $y_t = a_1 y_{t-1} + \dots + a_p y_{t-p} + e_t + b_1 e_{t-1} + \dots + b_q e_{t-q}$.

Auto ARIMA

This tool provides analyses identical to the ARIMA module except that the Auto-ARIMA module automates some of the traditional ARIMA modeling by automatically testing multiple permutations of model specifications and returns the best-fitting model. Running the Auto-ARIMA module is similar to running regular ARIMA forecasts, with the differences being that the P, D, Q inputs are no longer required and that different combinations of these inputs are automatically run and compared.

Autocorrelation and Partial Autocorrelation

One very simple approach to test for autocorrelation is to graph the time series of a regression equation's residuals. If these residuals exhibit some cyclicity, then autocorrelation exists. Another more robust approach to detect autocorrelation is the use of the Durbin-Watson statistic, which estimates the potential for a first-order autocorrelation. The Durbin-Watson test also identifies model misspecification, that is, if a particular time-series variable is correlated to itself one period prior. Many time-series data tend to be autocorrelated to their historical occurrences. This relationship can be due to multiple reasons, including the variables' spatial relationships (similar time and space), prolonged economic shocks and events, psychological inertia, smoothing, seasonal adjustments of the data, and so forth.

The Durbin-Watson statistic is estimated by the ratio of the sum of the squares of the regression errors for one period prior, to the sum of the current period's errors:

$$DW = \frac{\sum (\varepsilon_t - \varepsilon_{t-1})^2}{\sum \varepsilon_t^2}$$

There is a Durbin-Watson critical statistic table at the end of the book that provides a guide as to whether a statistic implies any autocorrelation.

Another test for autocorrelation is the Breusch-Godfrey test, where for a regression function in the form of:

$$Y = f(X_1, X_2, \dots, X_k)$$

Estimate this regression equation and obtain its errors ε_t . Then, run the secondary regression function in the form of:

$$Y = f(X_1, X_2, \dots, X_k, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-p})$$

Obtain the R-squared value and test it against a null hypothesis of no autocorrelation versus an alternate hypothesis of autocorrelation, where the test statistic follows a chi-square distribution of p degrees of freedom:

$$R^2(n-p) \sim \chi_{df=p}^2$$

Fixing autocorrelation requires the application of advanced econometric models including the applications of ARIMA (as described above) or ECM (Error Correction Models). However, one simple fix is to take the lags of the dependent variable for the appropriate periods, add them into the regression function, and test for their significance, as for instance:

$$Y_t = f(Y_{t-1}, Y_{t-2}, \dots, Y_{t-p}, X_1, X_2, \dots, X_k)$$

In interpreting the results of an Autoeconometrics and Auto ARIMA model, most of the specifications are identical to the multivariate regression analysis. However, there are several additional sets of results specific to the econometric analysis. The first is the addition of Akaike Information Criterion (AIC) and Schwarz Criterion (SC), which are often used in ARIMA model selection and identification. That is, AIC and SC are used to determine if a particular model with a specific set of p , d , and q parameters is a good statistical fit. SC imposes a greater penalty for additional coefficients than the AIC but, generally, the model with the lowest AIC and SC values should be chosen. Finally, an additional set of results called the autocorrelation (AC) and partial autocorrelation (PAC) statistics are provided in the ARIMA report.

For instance, if autocorrelation AC(1) is nonzero, it means that the series is first order serially correlated. If AC dies off more or less geometrically with increasing lags, it implies that the series follows a low-order autoregressive process. If AC drops to zero after a small number of lags, it implies that the series follows a low-order moving average process. In contrast, PAC measures the correlation of values that are k periods apart after removing the correlation from the intervening lags. If the pattern of autocorrelation can be captured by an autoregression of order less than k , then the partial autocorrelation at lag k will be close to zero. The Ljung-Box Q-statistics and their p-values at lag k are also provided, where the null hypothesis being tested is such that there is no autocorrelation up to order k . The dotted lines in the plots of the autocorrelations are the approximate two standard error bounds. If the autocorrelation is within these bounds, it is not significantly different from zero at approximately the 5% significance level. Finding the

right ARIMA model takes practice and experience. These AC, PAC, SC, and AIC are highly useful diagnostic tools to help identify the correct model specification. Finally, the ARIMA parameter results are obtained using sophisticated optimization and iterative algorithms, which means that although the functional forms look like those of a multivariate regression, they are not the same. ARIMA is a much more computationally intensive and advanced econometric approach.

Autoeconometrics (Quick) and Autoeconometrics (Detailed)

This section explains the Autoeconometrics methodology. Using the user's selected variable's data, we take these X variables and create in memory $LN(X_i)$, $LAG(X_i, N)$, $LN(LAG(X_i, N))$, $DIFF(X_i)$, $RATE(X_i)$ and the cross products $X_i X_j$, where $LN(X_i)$ is the natural logarithm of some variable X_i , while $LAG(X_i, N)$ is lag of the variable X_i for N periods. $DIFF(X_i)$ is the first difference (i.e., the new value at period 2 is period 2's X value less period 1's X value). $X_i X_j$ is variable X_i times variable X_j , and $RATE(X_i)$ is the first level ratio (i.e., the new value at period 2 is period 2's X value divided by period 1's X value).

We then run the analysis:

- Run the basic econometrics routines using Y on all these X variables created.
- Then, look at the list of p-values, take out the variable with the highest p-value (as long as it is above the user input P-Value Threshold), and rerun the analysis without this variable. The intercept's p-value is not considered.
- Continue running and eliminating each variable one at a time until all remaining variables have p-values under or at this threshold.
- Report the results of the final model where all p-values are under this threshold.

When running the analysis, here are some things to make the run go faster:

- Take all the Y and X values into memory.
- In memory, create new variables such as $LN(X_i)$, $LAG(X_i)$, $DIFF(X_i)$, $RATE(X_i)$, and so forth, based on the list generated previously.
- If the original variable has negative values, we do not do the LN for this variable. The same applies when $X_i X_j$ is negative: we do not compute the LN for it.
- When running, we only need the p-values in memory. So, there is no need to run entire econometrics routine and this will make things run faster.
- We only show the detailed report of the final result.

If there is a problem when running the econometrics analysis when starting the first model with all the variables, we do a bypass procedure:

- If error exists when running all of the variables the first time, skip and do this: Calculate the correlation between Y and each of the X variables (i.e., Correlation of Y to X_1 , Y to X_2 , Y to X_n). Then, eliminate the lowest absolute value of the correlation. So, if the lowest is variable X_n , we eliminate it and then run the econometrics analysis and repeat this step if required.
- If the user selects the checkbox for Autoregressive $AR(p)$ and puts in some value (only positive integers are allowed and by default this is set to 1 and unchecked), we simply add in the list of

functions the value $LAG(Y,N)$ where we lag the Y dependent variable at N number of periods. This N periods is user's $AR(p)$ value entered.

At the bottom of the user interface, we have a droplist where we have:

- *Dependent (Y)*: this just means we use the regular Y data selected by user.
- *LN(Y) Dependent*: we take the $LN(Y)$ as the dependent variable when running the analysis.
- *DIFF(Y) Dependent*: we take the difference in Y , i.e., $DIFF(Y)$ at period 2 is Y at period 2 less Y at period 1.
- *RATE(Y) Dependent*: we take the rate ratio in Y , i.e., $DIFF(Y)$ period 2 is Y at period 2 divided by Y at period 1.

If we have, for example, three variables that the user links in ($X1$, $X2$, and $X3$), we want to get all the combinations such that they include *TIME*, the original variables, the *LN* of these variables, the *LAG* of these variables, *LN* of the *LAGS*, *DIFF* of the variables, and the combinatorial multiplication of these variables (two at a time). See below for a simple example.

If there are three variables $X1$, $X2$, $X3$, the combinations list is:

TIME, $X1$; $X2$; $X3$; $LN(X1)$; $LN(X2)$; $LN(X3)$; $LAG(X1,N)$; $LAG(X2,N)$; $LAG(X3,N)$; $LN(LAG(X1,N))$; $LN(LAG(X2,N))$; $LN(LAG(X3,N))$; $DIFF(X1)$; $DIFF(X2)$; $DIFF(X3)$; $RATE(X1)$; $RATE(X2)$; $RATE(X3)$; $LN(RATE(X1))$; $LN(RATE(X2))$; $LN(RATE(X3))$; $X1*X2$; $X1*X3$; $X2*X3$; $LN(X1*X2)$; $LN(X1*X3)$; $LN(X2*X3)$

and possibly adding two more variables, $LAG(Y,N)$ and $LN(LAG(Y,N))$, if Autoregressive $AR(p)$ is chosen.

If five variables $X1$, $X2$, $X3$, $X4$, $X5$, the combinations list is:

TIME; $X1$; $X2$; $X3$; $X4$; $X5$; $LN(X1)$; $LN(X2)$; $LN(X3)$; $LN(X4)$; $LN(X5)$; $LAG(X1,N)$; $LAG(X2,N)$; $LAG(X3,N)$; $LAG(X4,N)$; $LAG(X5,N)$; $LN(LAG(X1,N))$; $LN(LAG(X2,N))$; $LN(LAG(X3,N))$; $LN(LAG(X4,N))$; $LN(LAG(X5,N))$; $DIFF(X1)$; $DIFF(X2)$; $DIFF(X3)$; $DIFF(X4)$; $DIFF(X5)$; $RATE(X1)$; $RATE(X2)$; $RATE(X3)$; $RATE(X4)$; $RATE(X5)$; $LN(RATE(X1))$; $LN(RATE(X2))$; $LN(RATE(X3))$; $LN(RATE(X4))$; $LN(RATE(X5))$; $X1*X2$; $X1*X3$; $X1*X4$; $X1*X5$; $X2*X3$; $X2*X4$; $X2*X5$; $X3*X4$; $X3*X5$; $X4*X5$; $LN(X1*X2)$; $LN(X1*X3)$; $LN(X1*X4)$; $LN(X1*X5)$; $LN(X2*X3)$; $LN(X2*X4)$; $LN(X2*X5)$; $LN(X3*X4)$; $LN(X3*X5)$; $LN(X4*X5)$

and possibly adding two more variables, $LAG(Y,N)$ and $LN(LAG(Y,N))$, if Autoregressive $AR(p)$ is chosen

As a quick check, the total number of variables on each list is $[7*X+1] + 2(X!/(2!*(X-2)!))$. So, in the case of 5 X variables, we have $7*5+1 + 2(5!/(2!*(5-2)!)) = 35+1+20 = 56$ combinations. That is, the $7*5+1$ is the regular variables and the *LN* and *LAG/DIFF* functions. The $2(5!/(2!*(5-2)!))$ portion is for the interacting variables $X1*X2$ and $LN(X1*X2)$ portion.

The previous few paragraphs detail the heuristics of the algorithm and illustrate the complexity of the approach. With an example of 7 independent X variables, there are a total of over 4 million model permutations and combinations that will be generated using this algorithm. The user's selected or pasted data is loaded into memory such that the algorithm can run quickly in a virtual environment. The data are first checked for validity and integrity by looking at various issues such as micronumerosity where the

number of independent variables generated exceeds the total number of rows of data, creating an error in the procedure, or multicollinearity, where the independent variables are highly correlated to one another, returning an error in the regression analysis model. The data are also checked for any alphanumerical inputs or missing or invalid data. If the data pass all these checks, they will be entered into memory for the next step in the process. Using the data, the algorithm determines how many independent variables exist and initiates the generation of all the unique intermediate variables such as the natural logarithm, the first difference, lagged values, and so forth. The proprietary algorithm is then run to enumerate in detail all possible combinations and permutations of models required. The unique variables in these enumerated models are then identified and matched against the list generated previously and the actual data of these revised variables are computed and stored in temporary memory. Each of the enumerated models is then run where each of the unique model's results is stored in memory and the running list of best models is maintained in memory. This list of best models is selected based on two criteria. The first is that all models are selected and ranked based on the adjusted R-square or regular R-square coefficient. The second is that all of the variables' p-values have to be below the statistical significance threshold of 0.10. At the end of running all combinations and permutations of models, the list of best models is shown and ranked by the adjusted R-square or regular R-square, and the detailed regression analysis results are shown for these best models.

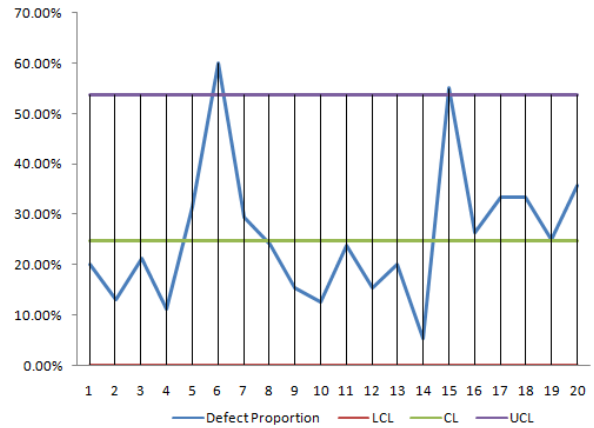
Basic Econometrics and Custom Econometrics

Econometrics refers to a branch of business analytics, modeling, and forecasting techniques for modeling the behavior or forecasting certain business, financial, economic, physical science, and other variables. Running the Basic Econometrics models is similar to regular regression analysis except that the dependent and independent variables are allowed to be modified before a regression is run. The report generated is the same as in the Multiple Regression method, and the interpretations are identical to those in a multiple regression analysis.

Control Charts: C, NP, P, R, U, XMR

Sometimes the specification limits are not set; instead, statistical control limits are computed based on the actual data collected (e.g., the number of defects in a manufacturing line). For instance, in the figure below, we see 20 sample experiments or samples taken at various times of a manufacturing process. The number of samples taken varied over time, and the number of defective parts were also gathered. The upper control limit (UCL) and lower control limit (LCL) are computed, as are the central line (CL) and other sigma levels. The resulting chart is called a control chart, and if the process is out of control, the actual defect line will be outside of the UCL and LCL lines.

Subgroup	Defective Units	Sample Size	Defect Proportion	LCL	CL	UCL
1	5	25	20.00%	0.00%	24.76%	53.71%
2	3	23	13.04%	0.00%	24.76%	53.71%
3	4	19	21.05%	0.00%	24.76%	53.71%
4	2	18	11.11%	0.00%	24.76%	53.71%
5	6	19	31.58%	0.00%	24.76%	53.71%
6	12	20	60.00%	0.00%	24.76%	53.71%
7	5	17	29.41%	0.00%	24.76%	53.71%
8	6	25	24.00%	0.00%	24.76%	53.71%
9	4	26	15.38%	0.00%	24.76%	53.71%
10	3	24	12.50%	0.00%	24.76%	53.71%
11	5	21	23.81%	0.00%	24.76%	53.71%
12	4	26	15.38%	0.00%	24.76%	53.71%
13	5	25	20.00%	0.00%	24.76%	53.71%
14	1	19	5.26%	0.00%	24.76%	53.71%
15	11	20	55.00%	0.00%	24.76%	53.71%
16	5	19	26.32%	0.00%	24.76%	53.71%
17	6	18	33.33%	0.00%	24.76%	53.71%
18	6	18	33.33%	0.00%	24.76%	53.71%
19	4	16	25.00%	0.00%	24.76%	53.71%
20	5	14	35.71%	0.00%	24.76%	53.71%

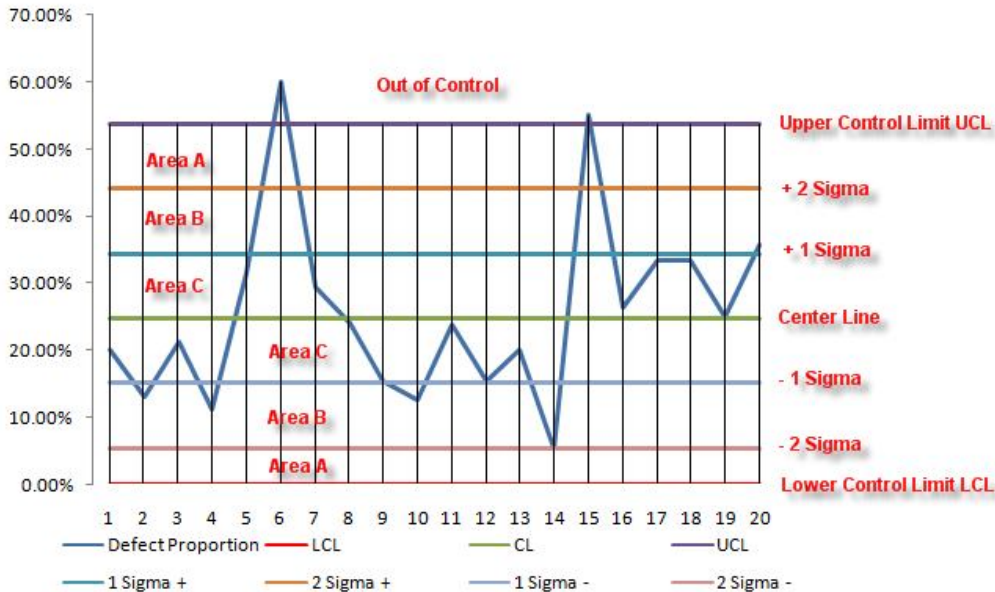


In the interpretation of a control chart, by adding in the ± 1 and ± 2 sigma lines, we can divide the control charts into several areas or zones, as illustrated in the figure below. The following are rules of thumb that typically apply to control charts to determine if the process is out of control:

- If one point is beyond Area A
- If two out of three consecutive points are in Area A or beyond
- If four out of five consecutive points are in Area B or beyond
- If eight consecutive points are in Area C or beyond

Additionally, a potential structural shift can be detected if any one of the following occurs:

- At least 10 out of 11 sequential points are on one side of the CL
- At least 12 out of 14 sequential points are on one side of the CL
- At least 14 out of 17 sequential points are on one side of the CL
- At least 16 out of 20 sequential points are on one side of the CL



C Chart

A C chart is used under the following conditions: when the variable of interest is an attribute (e.g., defective or nondefective), the data collected are in total number of defects (actual count in units), and there are multiple measurements in a sample experiment; when multiple experiments are run and the average number of defects of the collected data is of interest; and when the number of samples collected in each experiment are the same.

NP Chart

An NP chart is used under the following conditions: when the variable of interest is an attribute (e.g., defective or nondefective), the data collected are in proportions of defects (or number of defects in a specific sample), and there are multiple measurements in a sample experiment; when multiple experiments are run and the average proportion of defects of the collected data is of interest; and when the number of samples collected in each experiment is constant for all experiments.

P Chart

A P chart is used under the following conditions: when the variable of interest is an attribute (e.g., defective or nondefective), the data collected are in proportions of defects (or number of defects in a specific sample), and there are multiple measurements in a sample experiment; when multiple experiments are run and the average proportion of defects of the collected data is of interest; and when the number of samples collected in each experiment differs.

R Chart

An R chart is used when the variable has raw data values, there are multiple measurements in a sample experiment, multiple experiments are run, and the range of the collected data is of interest.

U Chart

A U chart is used under the following conditions: when the variable of interest is an attribute (e.g., defective or nondefective), the data collected are in total number of defects (actual count in units) and there are multiple measurements in a sample experiment; when multiple experiments are run and the average number of defects of the collected data is of interest; and when the number of samples collected in each experiment differs.

XMR Chart

An X_mR chart: used when the variable has raw data values and is a single measurement taken in each sample experiment, multiple experiments are run, and the actual value of the collected data is of interest.

Correlation (Linear, Nonlinear)

The Correlation module lists the Pearson's product moment correlations (commonly referred to as the Pearson's R) between variable pairs. The correlation coefficient ranges between -1.0 and $+1.0$ inclusive. The sign indicates the direction of association between the variables while the coefficient indicates the magnitude or strength of association. The Pearson's R only measures a linear relationship and is less effective in measuring nonlinear relationships.

A hypothesis t-test is performed on the Pearson's R and the p-values are reported. If the calculated p-value is less than or equal to the significance level used in the test, then reject the null hypothesis and conclude that there is a significant correlation between the two variables in question. Otherwise, the correlation is not statistically significant.

Finally, a Spearman Rank-Based Correlation is also included. The Spearman's R first ranks the raw data then performs the correlation calculation, which allows it to better capture nonlinear relationships. The Pearson's R is a parametric test and the underlying data are assumed to be normally distributed; hence, the t-test can be applied. However, the Spearman's R is a nonparametric test, and where no underlying distributions are assumed, the t-test cannot be applied.

Cubic Spline

Sometimes there are missing values in a time-series dataset. For instance, interest rates for years 1 to 3 may exist, followed by years 5 to 8, and then year 10. Spline curves can be used to interpolate the missing years' interest rate values based on the data that exist. Spline curves can also be used to forecast or extrapolate values of future time periods beyond the time period of available data. The data can be linear or nonlinear. The Known X values represent the values on the x-axis of a chart (in our example, this is Years of the known interest rates, and, usually, the x-axis values are the those that are known in advance such as time or years) and the Known Y values represent the values on the y-axis (in our case, the known Interest Rates). The y-axis variable is typically the variable you wish to interpolate missing values from or extrapolate the values into the future.

Data Descriptive Statistics

Almost all distributions can be described within 4 moments (some distributions require one moment, while others require two moments, and so forth). Descriptive statistics quantitatively captures these moments. The first moment describes the location of a distribution (i.e., mean, median, and mode) and is interpreted as the expected value, expected returns, or the average value of occurrences.

The second moment measures a distribution's spread or width and is frequently described using measures such as standard deviations, variances, quartiles, and inter-quartile ranges. Standard deviation is a popular measure indicating the average deviation of all data points from their mean. It is a popular measure as it is frequently associated with risk (higher standard deviations meaning a wider distribution, higher risk, or wider dispersion of data points around the mean value) and its units are identical to the units in the original dataset.

Skewness is the third moment in a distribution. Skewness characterizes the degree of asymmetry of a distribution around its mean. Positive skewness indicates a distribution with an asymmetric tail extending toward more positive values. Negative skewness indicates a distribution with an asymmetric tail extending toward more negative values.

Kurtosis characterizes the relative peakedness or flatness of a distribution compared to the normal distribution. It is the fourth moment in a distribution. A positive kurtosis value indicates a relatively peaked distribution. A negative kurtosis indicates a relatively flat distribution. The kurtosis measured here has been centered to zero (certain other kurtosis measures are centered on 3.0). While both are equally

valid, centering across zero makes the interpretation simpler. A high positive kurtosis indicates a peaked distribution around its center and leptokurtic or fat tails. This indicates a higher probability of extreme events (e.g., catastrophic events, terrorist attacks, stock market crashes) than is predicted in a normal distribution.

Deseasonalizing

This model deseasonalizes and detrends your original data to take out any seasonal and trending components. In forecasting models, the process eliminates the effects of accumulating datasets from seasonality and trend to show only the absolute changes in values and to allow potential cyclical patterns to be identified by removing the general drift, tendency, twists, bends, and effects of seasonal cycles of a set of time-series data. For example, a detrended dataset may be necessary to see a more accurate account of a company's sales in a given year more clearly by shifting the entire dataset from a slope to a flat surface to better show the underlying cycles and fluctuations.

Many time-series data exhibit seasonality where certain events repeat themselves after some time period or seasonality period (e.g., ski resorts' revenues are higher in winter than in summer, and this predictable cycle will repeat itself every winter). Seasonality periods represent how many periods would have to pass before the cycle repeats itself (e.g., 24 hours in a day, 12 months in a year, 4 quarters in a year, 60 minutes in an hour, etc.). This tool deseasonalizes and detrends your original data to take out any seasonal components. A seasonal index greater than 1 indicates a high period or peak within the seasonal cycle and a value below 1 indicates a dip in the cycle.

Distributional Fitting

Another powerful simulation tool is distributional fitting; that is, which distribution does an analyst or engineer use for a particular input variable in a model? What are the relevant distributional parameters? If no historical data exist, then the analyst must make assumptions about the variables in question. One approach is to use the Delphi method, where a group of experts are tasked with estimating the behavior of each variable. For instance, a group of mechanical engineers can be tasked with evaluating the extreme possibilities of a spring coil's diameter through rigorous experimentation or guesstimates. These values can be used as the variable's input parameters (e.g., uniform distribution with extreme values between 0.5 and 1.2). When testing is not possible (e.g., market share and revenue growth rate), management can still make estimates of potential outcomes and provide the best-case, most-likely case, and worst-case scenarios, whereupon a triangular or custom distribution can be created. The null hypothesis (H_0) being tested is such that the fitted distribution is the same distribution as the population from which the sample data to be fitted comes. Thus, if the computed p-value is lower than a critical alpha level (typically 0.10 or 0.05), then the distribution is the wrong distribution. Conversely, the higher the p-value, the better the distribution fits the data. Roughly, you can think of p-value as a percentage explained; for example, if the p-value is 0.9727, then, setting the resulting distribution explains about 97.27% of the variation in the data, indicating an especially good fit. The data were from a 1,000-trial simulation in Risk Simulator based on a normal distribution with a mean of 100 and a standard deviation of 10. Because only 1,000 trials were simulated, the resulting distribution is fairly close to the specified distributional parameters, and in this case, about a 97.27% precision.

Exponential J Curve

The J-curve, or exponential growth curve, is one where the growth of the next period depends on the current period's level and the increase is exponential. This phenomenon means that the values will increase significantly over time, from one period to another. This model is typically used in forecasting biological growth and chemical reactions over time.

Heteroskedasticity

Several tests exist to test for the presence of heteroskedasticity. These tests also are applicable for testing misspecifications and nonlinearities. The simplest approach is to graphically represent each independent variable against the dependent variable as illustrated earlier. Another approach is to apply one of the most widely used models, the White's test, where the test is based on the null hypothesis of no heteroskedasticity against an alternate hypothesis of heteroskedasticity of some unknown general form. The test statistic is computed by an auxiliary or secondary regression, where the squared residuals or errors from the first regression are regressed on all possible (and nonredundant) cross products of the regressors. For example, suppose the following regression is estimated:

$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \varepsilon_t$$

The test statistic is then based on the auxiliary regression of the errors (ε):

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 X + \alpha_2 Z + \alpha_3 X^2 + \alpha_4 Z^2 + \alpha_5 XZ + v_t$$

The nR^2 statistic is the White's test statistic, computed as the number of observations (n) times the centered R-squared from the test regression. White's test statistic is asymptotically distributed as a χ^2 with degrees of freedom equal to the number of independent variables (excluding the constant) in the test regression.

The White's test is also a general test for model misspecification, because the null hypothesis underlying the test assumes that the errors are both homoskedastic and independent of the regressors, and that the linear specification of the model is correct. Failure of any one of these conditions could lead to a significant test statistic. Conversely, a nonsignificant test statistic implies that none of the three conditions is violated. For instance, the resulting F-statistic is an omitted variable test for the joint significance of all cross products, excluding the constant. One method to fix heteroskedasticity is to make it homoskedastic by using a weighted least squares (WLS) approach. For instance, suppose the following is the original regression equation:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

Further suppose that X_2 is heteroskedastic. Then transform the data used in the regression into:

$$Y = \frac{\beta_0}{X_2} + \beta_1 \frac{X_1}{X_2} + \beta_2 + \beta_3 \frac{X_3}{X_2} + \frac{\varepsilon}{X_2}$$

The model can be redefined as the following WLS regression:

$$Y_{WLS} = \beta_0^{WLS} + \beta_1^{WLS} X_1 + \beta_2^{WLS} X_2 + \beta_3^{WLS} X_3 + v$$

Alternatively, the Park's test can be applied to test for heteroskedasticity and to fix it. The Park's test model is based on the original regression equation, uses its errors, and creates an auxiliary regression that takes the form of:

$$\ln e_i^2 = \beta_1 + \beta_2 \ln X_{k,i}$$

Suppose β_2 is found to be statistically significant based on a t-test; then heteroskedasticity is found to be present in the variable $X_{k,i}$. The remedy, therefore, is to use the following regression specification:

$$\frac{Y}{\sqrt{X_k^{\beta_2}}} = \frac{\beta_1}{\sqrt{X_k^{\beta_2}}} + \frac{\beta_2 X_2}{\sqrt{X_k^{\beta_2}}} + \frac{\beta_3 X_3}{\sqrt{X_k^{\beta_2}}} + \varepsilon$$

Limited Dependent Variables: Logit, Probit, Tobit

Limited Dependent Variables describe the situation where the dependent variable contains data that are limited in scope and range, such as binary responses (0 or 1), truncated, ordered, or censored data. For instance, given a set of independent variables (e.g., age, income, education level of credit card or mortgage loan holders), we can model the probability of default using maximum likelihood estimation (MLE). The response or dependent variable Y is binary, that is, it can have only two possible outcomes that we denote as 1 and 0 (e.g., Y may represent presence/absence of a certain condition, defaulted/not defaulted on previous loans, success/failure of some device, answer yes/no on a survey, etc.). We also have a vector of independent variable regressors X , which are assumed to influence the outcome Y . A typical ordinary least squares regression approach is invalid because the regression errors are heteroskedastic and non-normal, and the resulting estimated probability estimates will return nonsensical values of above 1 or below 0. MLE analysis handles these problems using an iterative optimization routine to maximize a log likelihood function when the dependent variables are limited.

A logit, or logistic regression, is used for predicting the probability of occurrence of an event by fitting data to a logistic curve. It is a generalized linear model used for binomial regression, and like many forms of regression analysis, it makes use of several predictor variables that may be either numerical or categorical. MLE applied in a binary multivariate logistic analysis is used to model dependent variables to determine the expected probability of success of belonging to a certain group. The estimated coefficients for the logit model are the logarithmic odds ratios, and cannot be interpreted directly as probabilities. A quick computation is first required and the approach is simple.

Specifically, the logit model is specified as $Estimated Y = LN[P_i/(1-P_i)]$ or, conversely, $P_i = EXP(Estimated Y)/(1+EXP(Estimated Y))$, and the coefficients β_i are the log odds ratios. So, taking the antilog or $EXP(\beta_i)$, we obtain the odds ratio of $P_i/(1-P_i)$. This means that with an increase in a unit of β_i , the log odds ratio increases by this amount. Finally, the rate of change in the probability $dP/dX = \beta_i P_i(1-P_i)$. The Standard Error measures how accurate the predicted Coefficients are, and the t-Statistics are the ratios of each predicted Coefficient to its Standard Error and are used in the typical regression hypothesis test of the significance of each estimated parameter. To estimate the probability of success of belonging to a certain group (e.g., predicting if a smoker will develop chest complications given the amount smoked per year), simply compute the *Estimated Y* value using the MLE coefficients. For example, if the model is $Y = 1.1 + 0.005(Cigarettes)$, then someone smoking 100 packs per year has an *Estimated Y* of $1.1 + 0.005(100) = 1.6$. Next, compute the inverse antilog of the odds ratio by doing $EXP(Estimated Y)/[1 + EXP(Estimated Y)]$

$= \text{EXP}(1.6)/(1+ \text{EXP}(1.6)) = 0.8320$. So, such a person has an 83.20% chance of developing some chest complications in his lifetime.

A probit model (sometimes also known as a normit model) is a popular alternative specification for a binary response model, which employs a probit function estimated using maximum likelihood estimation and is called probit regression. The probit and logistic regression models tend to produce very similar predictions where the parameter estimates in a logistic regression tend to be 1.6 to 1.8 times higher than they are in a corresponding probit model. The choice of using a probit or logit is entirely up to convenience, and the main distinction is that the logistic distribution has a higher kurtosis (fatter tails) to account for extreme values. For example, suppose that house ownership is the decision to be modeled, and this response variable is binary (home purchase or no home purchase) and depends on a series of independent variables X_i such as income, age, and so forth, such that $I_i = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$, where the larger the value of I_i , the higher the probability of home ownership. For each family, a critical I^* threshold exists, where if exceeded, the house is purchased, otherwise, no home is purchased, and the outcome probability (P) is assumed to be normally distributed, such that $P_i = \text{CDF}(I)$ using a standard normal cumulative distribution function (CDF). Therefore, use the estimated coefficients exactly like that of a regression model and, using the *Estimated Y* value, apply a standard normal distribution (you can use Excel's *NORMSDIST* function or Risk Simulator's Distributional Analysis tool by selecting Normal distribution and setting the mean to be 0 and standard deviation to be 1). Finally, to obtain a probit or probability unit measure, set $I_i + 5$ (because whenever the probability $P_i < 0.5$, the estimated I_i is negative, due to the fact that the normal distribution is symmetrical around a mean of zero).

The tobit model (censored tobit) is an econometric and biometric modeling method used to describe the relationship between a non-negative dependent variable Y_i and one or more independent variables X_i . A tobit model is an econometric model in which the dependent variable is censored; that is, the dependent variable is censored because values below zero are not observed. The tobit model assumes that there is a latent unobservable variable Y^* . This variable is linearly dependent on the X_i variables via a vector of β_i coefficients that determine their interrelationships. In addition, there is a normally distributed error term U_i to capture random influences on this relationship. The observable variable Y_i is defined to be equal to the latent variables whenever the latent variables are above zero, and Y_i is assumed to be zero otherwise. That is, $Y_i = Y^*$ if $Y^* > 0$ and $Y_i = 0$ if $Y^* = 0$. If the relationship parameter β_i is estimated by using ordinary least squares regression of the observed Y_i on X_i , the resulting regression estimators are inconsistent and yield downward biased slope coefficients and an upward biased intercept. Only MLE would be consistent for a tobit model. In the tobit model, there is an ancillary statistic called sigma, which is equivalent to the standard error of estimate in a standard ordinary least squares regression, and the estimated coefficients are used the same way as a regression analysis.

Linear Interpolation

Sometimes interest rates or any type of time-dependent rates may have missing values. For instance, the Treasury rates for Years 1, 2, and 3 exist, and then jump to Year 5, skipping Year 4. We can, using linear interpolation (i.e., we assume the rates during the missing periods are linearly related), determine and “fill in” or interpolate their values. In contrast, the cubic spline polynomial interpolation and extrapolation model is used to “fill in the gaps” of missing values (interpolation) and to forecast outside of the known values (extrapolation) when the underlying structure is nonlinear. For example, we can use apply this approach to spot yields and term structure of interest rates whereby the model can be used to both interpolate missing data points within a time series of interest rates (as well as other macroeconomic

variables such as inflation rates and commodity prices or market returns) and also used to extrapolate outside of the given or known range, useful for forecasting purposes.

Logistic S Curve

The S-curve, or logistic growth curve, starts off like a J-curve, with exponential growth rates. Over time, the environment becomes saturated (e.g., market saturation, competition, overcrowding), the growth slows, and the forecast value eventually ends up at a saturation or maximum level. The S-curve model is typically used in forecasting market share or sales growth of a new product from market introduction until maturity and decline, population dynamics, growth of bacterial cultures, and other naturally occurring variables.

Markov Chain

A Markov chain exists when the probability of a future state depends on a previous state and when linked together forms a chain that reverts to a long-run steady state level. This Markov approach is typically used to forecast the market share of two competitors. The required inputs are the starting probability of a customer in the first store (the first state) will return to the same store in the next period versus the probability of switching to a competitor's store in the next state.

Multiple Regression (Linear Regression and Nonlinear Regression)

This section demonstrates the mathematical models and computations used in creating the general regression equations, which take the form of $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$ where β_0 is the intercept, β_i are the slope coefficients, and ε is the error term. The Y term is the dependent variable and the X terms are the independent variables, where these X variables are also known as the regressors. The dependent variable is named as such as it *depends* on the independent variable; for example, sales revenue depends on the amount of marketing costs expended on a product's advertising and promotion, making the dependent variable sales and the independent variable marketing costs. An example of a bivariate regression, where there is only a single Y and a single X variable, is seen as simply inserting the best-fitting line through a set of data points in a two-dimensional plane. In other cases, a multivariate regression can be performed, where there are multiple or k number of independent X variables or regressors and where the best-fitting line will be within a $k + 1$ dimensional plane.

Fitting a line through a set of data points in a multidimensional scatter plot may result in numerous possible lines. The *best-fitting line* is defined as the single unique line that minimizes the total vertical errors, that is, the sum of the absolute distances between the actual data points (Y_i) and the estimated line \hat{Y} . To find the best-fitting unique line that minimizes the errors, a more sophisticated approach is applied, using multivariate regression analysis. Regression analysis therefore finds the unique best-fitting line by requiring that the total errors be minimized, or by calculating

$$\text{Min} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Only one unique line will minimize this sum of squared errors as shown in the equation above. The errors (vertical distances between the actual data and the predicted line) are squared to avoid the negative errors from canceling out the positive errors. Solving this minimization problem with respect to the slope and intercept requires calculating first derivatives and setting them equal to zero:

$$\frac{d}{d\beta_0} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0 \text{ and } \frac{d}{d\beta_1} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = 0$$

which yields the simple bivariate regression's set of least squares equations:

$$\beta_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n X_i Y_i - \frac{\sum_{i=1}^n X_i \sum_{i=1}^n Y_i}{n}}{\sum_{i=1}^n X_i^2 - \frac{\left(\sum_{i=1}^n X_i\right)^2}{n}}$$

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

For multivariate regression, the analogy is expanded to account for multiple independent variables, where $Y_i = \beta_1 + \beta_2 X_{2,i} + \beta_3 X_{3,i} + \varepsilon_i$ and the estimated slopes can be calculated by:

$$\hat{\beta}_2 = \frac{\sum Y_i X_{2,i} \sum X_{3,i}^2 - \sum Y_i X_{3,i} \sum X_{2,i} X_{3,i}}{\sum X_{2,i}^2 \sum X_{3,i}^2 - \left(\sum X_{2,i} X_{3,i}\right)^2}$$

$$\hat{\beta}_3 = \frac{\sum Y_i X_{3,i} \sum X_{2,i}^2 - \sum Y_i X_{2,i} \sum X_{2,i} X_{3,i}}{\sum X_{2,i}^2 \sum X_{3,i}^2 - \left(\sum X_{2,i} X_{3,i}\right)^2}$$

This set of results can be summarized using matrix notations: $[X'X]^{-1}[X'Y]$. In running multivariate regressions, great care must be taken to set up and interpret the results. For instance, a good understanding of econometric modeling is required (e.g., identifying regression pitfalls such as structural breaks, multicollinearity, heteroskedasticity, autocorrelation, specification tests, nonlinearities, etc.) before a proper model can be constructed. Therefore, the present software includes some advanced econometrics approaches that are based on the principles of multiple regression outlined above.

Nonlinear Extrapolation

Extrapolation involves making statistical forecasts by using historical trends that are projected for a specified period of time into the future. It is only used for time-series forecasts. For cross-sectional or mixed panel data (time series with cross-sectional data), multivariate regression is more appropriate. This methodology is useful when major changes are not expected, that is, when causal factors are expected to remain constant or when the causal factors of a situation are not clearly understood. It also helps discourage the introduction of personal biases into the process. Extrapolation is fairly reliable, relatively simple, and inexpensive. However, extrapolation, which assumes that recent and historical trends will continue, produces large forecast errors if discontinuities occur within the projected time period; that is, pure extrapolation of time series assumes that all we need to know is contained in the historical values of the series being forecasted. If we assume that past behavior is a good predictor of future behavior, extrapolation is appealing. This makes it a useful approach when all that is needed are many short-term forecasts. This methodology estimates the $f(x)$ function for any arbitrary x value by interpolating a smooth nonlinear curve through all the x values and, using this smooth curve, extrapolating future x values beyond the historical dataset. The methodology employs either the polynomial functional form or the

rational functional form (a ratio of two polynomials). Typically, a polynomial functional form is sufficient for well-behaved data, but rational functional forms are sometimes more accurate (especially with polar functions, i.e., functions with denominators approaching zero).

Nonparametric Hypothesis Tests

Nonparametric techniques make no assumptions about the specific shape or distribution from which the sample is drawn. This lack of assumptions is different from the other hypotheses tests such as ANOVA or t-tests (parametric tests) where the sample is assumed to be drawn from a population that is normally or approximately normally distributed. If normality is assumed, the power of the test is higher due this normality restriction. However, if flexibility on distributional requirements is needed, then nonparametric techniques are superior. In general, nonparametric methodologies provide the following advantages over other parametric tests:

- Normality or approximate normality does not have to be assumed.
- Fewer assumptions about the population are required; that is, nonparametric tests do not require that the population assume any specific distribution.
- Smaller sample sizes can be analyzed.
- Samples with nominal and ordinal scales of measurement can be tested.
- Sample variances do not have to be equal, which is required in parametric tests.

However, several caveats are worthy of mention:

- Compared to parametric tests, nonparametric tests use data less efficiently.
- The power of the test is lower than that of the parametric tests.

Therefore, if all the required assumptions are satisfied, it is better to use parametric tests. However, in reality, it may be difficult to justify these distributional assumptions or small sample sizes may exist, requiring the need for nonparametric tests. Thus, nonparametric tests should be used when the data are nominal or ordinal, or when the data are interval or ratio but the normality assumption is not met. The following lists each of the nonparametric tests available for use in the software.

Chi-Square Goodness of Fit

The chi-square test for goodness of fit is used to examine if a sample dataset could have been drawn from a population having a specified probability distribution. The probability distribution tested here is the normal distribution. The null hypothesis tested is such that the sample is randomly drawn from the normal distribution, versus the alternate hypothesis that the sample is not from a normal distribution. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

Chi-Square Independence

The chi-square test for independence examines two variables to see if there is some statistical relationship between them. This test is not used to find the exact nature of the relationship between the two variables, but to simply test if the variables could be independent of each other. The null hypothesis tested is such that the variables are independent of each other, versus the alternate hypothesis that the variables are not independent of each other.

The chi-square test looks at a table of observed frequencies and a table of expected frequencies. The amount of disparity between these two tables are calculated and compared with the chi-square test statistic. The observed frequencies reflect the cross-classification for members of a single sample, and the table of expected frequencies is constructed under the assumption that the null hypothesis is true.

Chi-Square Population Variance

The chi-square test for population variance is used for hypothesis testing and confidence interval estimation for a population variance. The population variance of a sample is typically unknown, and hence the need for quantifying this confidence interval. The population is assumed to be normally distributed.

Friedman's Test

The Friedman test is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample datasets to be analyzed. This method is the extension of the Wilcoxon Signed-Rank test for paired samples. The corresponding parametric test is the Randomized Block Multiple Treatment ANOVA, but unlike the ANOVA, the Friedman test does not require that the dataset be randomly sampled from normally distributed populations with equal variances. The Friedman test uses a two-tailed hypothesis test where the null hypothesis is such that the population medians of each treatment are statistically identical to the rest of the group, that is, there is no effect among the different treatment groups. The alternative hypothesis is such that the real population medians are statistically different from one another when tested using the sample dataset, that is, the medians are statistically different and thus there is a statistically significant effect among the different treatment groups. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

Kruskal-Wallis Test

The Kruskal-Wallis test is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample datasets to be analyzed. This method is the extension of the Wilcoxon Signed-Rank test by comparing more than two independent samples. The corresponding parametric test is the One-Way ANOVA, but unlike the ANOVA, the Kruskal-Wallis does not require that the dataset be randomly sampled from normally distributed populations with equal variances. The Kruskal-Wallis test is a two-tailed hypothesis test where the null hypothesis is such that the population medians of each treatment are statistically identical to the rest of the group, that is, there is no effect among the different treatment groups. The alternative hypothesis is such that the real population medians are statistically different from one another when tested using the sample dataset, that is, the medians are statistically different and thus there is a statistically significant effect among the different treatment groups. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

The benefit of the Kruskal-Wallis test is that it can be applied to ordinal, interval, and ratio data while ANOVA is only applicable for interval and ratio data. Also, the Friedman Test can be run with fewer data points. To illustrate, suppose that three different drug indications ($T = 3$) were developed and tested on

100 patients each ($N = 100$). The Kruskal-Wallis test can be applied to test if these three drugs are all equally effective statistically. If the calculated p-value is less than or equal to the significance level used in the test, then reject the null hypothesis and conclude that there is a significant difference among the different treatments. Otherwise, the treatments are all equally effective.

Lilliefors Test

The Lilliefors test is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample datasets to be analyzed. This test evaluates the null hypothesis of whether the data sample was drawn from a normally distributed population, versus an alternate hypothesis that the data sample is not normally distributed. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis. This test relies on two cumulative frequencies: one derived from the sample dataset and the second, from a theoretical distribution based on the mean and standard deviation of the sample data. An alternative to this test is the chi-square test for normality. The chi-square test requires more data points to run compared to the Lilliefors test.

Runs Test

The Runs test is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample datasets to be analyzed. This test evaluates the randomness of a series of observations by analyzing the number of runs it contains. A run is a consecutive appearance of one or more observations that are similar. The null hypothesis tested is whether the data sequence is random, versus the alternate hypothesis that the data sequence is not random. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

Wilcoxon Signed-Rank (One Var)

The single-variable Wilcoxon Signed-Rank test is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample datasets to be analyzed. This method looks at whether a sample dataset could have been randomly drawn from a particular population whose median is being hypothesized. The corresponding parametric test is the one-sample t-test, which should be used if the underlying population is assumed to be normal, providing a higher power on the test. The Wilcoxon Signed-Rank test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test. If the calculated Wilcoxon statistic is outside the critical limits for the specific significance level in the test, reject the null hypothesis and conclude that the true population median is not equal to (two-tail test), less than (left-tail test), or greater than (right-tail test) the hypothesized median based on the sample tested. Otherwise, the true population median is statistically similar to the hypothesized median.

Wilcoxon Signed-Rank (Two Var)

The Wilcoxon Signed-Rank test for paired variables is a form of nonparametric test, which makes no assumptions about the specific shape of the population from which the sample is drawn, allowing for smaller sample datasets to be analyzed. This method looks at whether the median of the differences

between the two paired variables are equal. This test is specifically formulated for testing the same or similar samples before and after an event (e.g., measurements taken before a medical treatment are compared against those measurements taken after the treatment to see if there is a difference). The corresponding parametric test is the two-sample t-test with dependent means, which should be used if the underlying population is assumed to be normal, providing a higher power on the test. The Wilcoxon Signed-Rank test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test.

To illustrate, suppose that a new engine design is tested against an existing engine design to see if there is a statistically significant difference between the two. The paired-variable Wilcoxon Signed-Rank Test can be applied. If the calculated Wilcoxon statistic is outside the critical limits for the specific significance level in the test, reject the null hypothesis and conclude that the difference between the true population medians is not equal to (two-tail test), less than (left-tail test), or greater than (right-tail test) the hypothesized median difference based on the sample tested. Otherwise, the true population median is statistically similar to the hypothesized median.

Parametric Hypothesis Tests

A hypothesis test is a statistical test of the characteristics a population by testing a small sample collected. In most cases, the population to be studied might be too large, difficult, or expensive to be completely sampled (e.g., all 100 million registered voters in the United States in a particular election), hence a smaller sample (e.g., a random sample of 1,100 voters from 20 cities) is collected and the sample statistics are tabulated. Then, using hypothesis tests, the characteristics of the entire population can be inferred from this small sample. XStatistics allows the user to test one-variable, two-variable, and multiple-variable hypotheses tests.

To perform a hypothesis test, first set up the null hypothesis (H_0) and the alternate hypothesis (H_a). Here are some quick rules:

- Always set up the alternate hypothesis first, then the null hypothesis
- The alternate hypothesis always has the following signs: $>$ or $<$ or \neq
- The null hypothesis always has the following signs: \geq or \leq or $=$
- If the alternate hypothesis is \neq , then it's a two-tailed test; if $<$, then it's a left (one) tail; and if $>$, then it's a right- (one) tailed test

Then, collect the sample data, run the appropriate hypothesis tests, and make the relevant conclusions about the population based on the sample data collected. That is, if the p-value is less than the significance level (the significance level α is selected by the user and is usually 0.10, 0.05, or 0.01) tested, reject the null hypothesis and accept the alternate hypothesis.

Two-Tailed Hypothesis Test

A two-tailed hypothesis tests the null hypothesis such that the population median of the sample dataset is statistically identical to the hypothesized median. The alternative hypothesis is that the real population median is statistically different from the hypothesized median when tested using the sample dataset. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis

and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

Right-Tailed Hypothesis Test

A right-tailed hypothesis tests the null hypothesis such that the population median of the sample dataset is statistically less than or equal to the hypothesized median. The alternative hypothesis is that the real population median is statistically greater than the hypothesized median when tested using the sample dataset. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

Left-Tailed Hypothesis Test

A left-tailed hypothesis tests the null hypothesis such that the population median of the sample dataset is statistically greater than or equal to the hypothesized median. The alternative hypothesis is that the real population median is statistically less than the hypothesized median when tested using the sample dataset. If the calculated p-value is less than or equal to the alpha significance value, then reject the null hypothesis and accept the alternate hypothesis. Otherwise, if the p-value is higher than the alpha significance value, do not reject the null hypothesis.

One Variable (T)

The one-variable t-test of means is appropriate when the population standard deviation is not known but the sampling distribution is assumed to be approximately normal (the t-test is used when the sample size is less than 30). This t-test can be applied to three types of hypothesis tests—a two-tailed test, a right-tailed test, and a left-tailed test—to examine if the population mean is equal to, less than, or greater than the hypothesized mean based on the sample dataset. If the calculated p-value is less than or equal to the significance level in the test, then reject the null hypothesis and conclude that the true population mean is not equal to (two-tail test), less than (left-tail test), or greater than (right-tail test) the hypothesized mean based on the sample tested. Otherwise, the true population mean is statistically similar to the hypothesized mean.

One Variable (Z)

The one-variable Z-test is appropriate when the population standard deviation is known and the sampling distribution is assumed to be approximately normal (this applies when the number of data points exceeds 30). This Z-test can be applied to three types of hypothesis tests—a two-tailed test, a right-tailed test, and a left-tailed test—to examine if the population mean is equal to, less than, or greater than the hypothesized mean based on the sample dataset. If the calculated p-value is less than or equal to the significance level in the test, then reject the null hypothesis and conclude that the true population mean is not equal to (two-tail test), less than (left-tail test), or greater than (right-tail test) the hypothesized mean based on the sample tested. Otherwise, the true population mean is statistically similar to the hypothesized mean.

One-Variable (Z) Proportion

The one-variable Z-test for proportions is appropriate when the sampling distribution is assumed to be approximately normal (this applies when the number of data points exceeds 30, and when the number of

data points, N , multiplied by the hypothesized population proportion mean, P , is greater than or equal to 5, $NP \geq 5$). The data used in the analysis have to be proportions and be between 0 and 1. This Z-test can be applied to three types of hypothesis tests—a two-tailed test, a right-tailed test, and a left-tailed test—to examine if the population mean is equal to, less than, or greater than the hypothesized mean based on the sample dataset. If the calculated p-value is less than or equal to the significance level in the test, then reject the null hypothesis and conclude that the true population mean is not equal to (two-tail test), less than (left-tail test), or greater than (right-tail test) the hypothesized mean based on the sample tested. Otherwise, the true population mean is statistically similar to the hypothesized mean.

Two-Variable (T) Dependent

The two-variable dependent t-test is appropriate when the population standard deviation is not known but the sampling distribution is assumed to be approximately normal (the t-test is used when the sample size is less than 30). In addition, this test is specifically formulated for testing the same or similar samples before and after an event (e.g., measurements taken before a medical treatment are compared against those measurements taken after the treatment to see if there is a difference). This t-test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test.

Suppose that a new heart medication was administered to 100 patients ($N = 100$) and the heart rates before and after the medication was administered were measured. The two dependent variables t-test can be applied to test if the new medication is effective by testing to see if there is a statistically different "before and after" averages. The dependent variables test is used here because there is only a single sample collected (the same patients' heartbeats were measured before and after the new drug administration).

The two-tailed null hypothesis tests that the true population's mean of the difference between the two variables is zero, versus the alternate hypothesis that the difference is statistically different from zero. The right-tail null hypothesis test is such that the differences in the population means (first mean less second mean) is statistically less than or equal to zero (which is identical to saying that mean of the first sample is less than or equal to the mean of the second sample). The alternative hypothesis is that the real populations' mean difference is statistically greater than zero when tested using the sample dataset (which is identical to saying that the mean of the first sample is greater than the mean of the second sample). The left-tail null hypothesis test is such that the differences in the population means (first mean less second mean) is statistically greater than or equal to zero (which is identical to saying that the mean of the first sample is greater than or equal to the mean of the second sample). The alternative hypothesis is that the real populations' mean difference is statistically less than zero when tested using the sample dataset (which is identical to saying that the mean of the first sample is less than the mean of the second sample). If the calculated p-value is less than or equal to the significance level in the test, then reject the null hypothesis and conclude that the true population difference of the population means is not equal to (two-tail test), less than (left-tail test), or greater than (right-tail test) zero based on the sample tested. Otherwise, the true population mean is statistically similar to the hypothesized mean.

Two-Variable (T) Independent Equal Variance

The two-variable t-test with equal variances is appropriate when the population standard deviation is not known but the sampling distribution is assumed to be approximately normal (the t-test is used when the

sample size is less than 30). In addition, the two independent samples are assumed to have similar variances.

For illustration, suppose that a new engine design is tested against an existing engine design to see if there is a statistically significant difference between the two. The t-test on two (independent) variables with equal variances can be applied. This test is used here because there are two distinctly different samples collected (new engine and existing engine), but the variances of both samples are assumed to be similar (the means may or may not be similar, but the fluctuations around the mean are assumed to be similar).

This t-test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test. A two-tailed hypothesis tests the null hypothesis, H_0 , such that the populations' mean difference (HMD) between the two variables is statistically identical to the hypothesized mean differences. If HMD is set to zero, this is the same as saying that the first mean equals the second mean. The alternative hypothesis is that the difference between the real population means is statistically different from the hypothesized mean differences when tested using the sample dataset. If HMD is set to zero, this is the same as saying that the first mean does not equal the second mean.

A right-tailed hypothesis tests the null hypothesis, H_0 , such that the population mean differences between the two variables is statistically less than or equal to the hypothesized mean differences. If HMD is set to zero, this is the same as saying that the first mean is less than or equal to the second mean. The alternative hypothesis is that the real difference between population means is statistically greater than the hypothesized mean differences when tested using the sample dataset. If HMD is set to zero, this is the same as saying that the first mean is greater than the second mean.

A left-tailed hypothesis tests the null hypothesis, H_0 , such that the differences between the population means of the two variables is statistically greater than or equal to the hypothesized mean differences. If HMD is set to zero, this is the same as saying that the first mean is greater than or equal to the second mean. The alternative hypothesis is that the real difference between population means is statistically less than the hypothesized mean difference when tested using the sample dataset. If HMD is set to zero, this is the same as saying that the first mean is less than the second mean.

If the calculated p-value is less than or equal to the significance level in the test, then reject the null hypothesis and conclude that the true population difference of the population means is not equal to (two-tail test), less than (left-tail test), or greater than (right-tail test) HMD based on the sample tested. Otherwise, the true difference of the population means is statistically similar to the HMD.

Two-Variable (T) Independent Unequal Variance

The two-variable t-test with unequal variances (the population variance of sample 1 is expected to be different from the population variance of sample 2) is appropriate when the population standard deviation is not known but the sampling distribution is assumed to be approximately normal (the t-test is used when the sample size is less than 30). In addition, the two independent samples are assumed to have similar variances.

To illustrate, suppose that a new customer relationship management (CRM) process is being evaluated for its effectiveness, and the customer satisfaction rankings between two hotels (one with and the other without CRM implemented) are collected. The t-test on two (independent) variables with unequal variances can be applied. This test is used here because there are two distinctly different samples collected (customer survey results of two different hotels) and the variances of both samples are assumed to be dissimilar (due to the difference in geographical location, plus the demographics and psychographics of the customers are different on both properties).

This t-test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test. A two-tailed hypothesis tests the null hypothesis H_0 such that the population mean differences between the two variables is statistically identical to the hypothesized mean differences. If HMD is set to zero, this is the same as saying that the first mean equals the second mean. The alternative hypothesis is that the real difference between the population means is statistically different from the hypothesized mean differences when tested using the sample dataset. If HMD is set to zero, this is the same as saying that the first mean does not equal the second mean.

A right-tailed hypothesis tests the null hypothesis, H_0 , such that the difference between the two variables' population means is statistically less than or equal to the hypothesized mean differences. If HMD is set to zero, this is the same as saying that the first mean is less than or equal to the second mean. The alternative hypothesis is that the real populations' mean difference is statistically greater than the hypothesized mean differences when tested using the sample dataset. If HMD is set to zero, this is the same as saying that the first mean is greater than the second mean.

A left-tailed hypothesis tests the null hypothesis H_0 , such that the difference between the two variables' population means is statistically greater than or equal to the hypothesized mean differences. If HMD is set to zero, this is the same as saying that the first mean is greater than or equal to the second mean. The alternative hypothesis is that the real difference between population means is statistically less than the hypothesized mean difference when tested using the sample dataset. If HMD is set to zero, this is the same as saying that the first mean is less than the second mean.

If the calculated p-value is less than or equal to the significance level in the test, then reject the null hypothesis and conclude that the true population difference of the population means is not equal to (two-tail test), less than (left-tail test), or greater than (right-tail test) the hypothesized mean based on the sample tested. Otherwise, the true difference of the population means is statistically similar to the hypothesized mean.

Two-Variable (Z) Independent Means

The two-variable Z-test is appropriate when the population standard deviations are known for the two samples, and the sampling distribution of each variable is assumed to be approximately normal (this applies when the number of data points of each variable exceeds 30).

To illustrate, suppose that market research was conducted on two different markets, the sample collected is large (N must exceed 30 for both variables), and the researcher is interested in testing whether there is a statistically significant difference between the two markets. Further suppose that such a market survey has been performed many times in the past and the population standard deviations are known. A two

independent variables Z-test can be applied because the sample size exceeds 30 on each market and the population standard deviations are known.

This Z-test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test. A two-tailed hypothesis tests the null hypothesis, H_0 , such that the difference between the two population means is statistically identical to the hypothesized mean. The alternative hypothesis is that the real difference between the two population means is statistically different from the hypothesized mean when tested using the sample dataset.

A right-tailed hypothesis tests the null hypothesis, H_0 , such that the difference between the two population means is statistically less than or equal to the hypothesized mean. The alternative hypothesis is that the real difference between the two population means is statistically greater than the hypothesized mean when tested using the sample dataset.

A left-tailed hypothesis tests the null hypothesis, H_0 , such that the difference between the two population means is statistically greater than or equal to the hypothesized mean. The alternative hypothesis is that the real difference between the two population means is statistically less than the hypothesized mean when tested using the sample dataset.

Two-Variable (Z) Independent Proportions

The two-variable Z-test on proportions is appropriate when the sampling distribution is assumed to be approximately normal (this applies when the number of data points of both samples exceeds 30). Further, the data should all be proportions and be between 0 and 1.

To illustrate, suppose that a brand research was conducted on two different headache pills, the sample collected is large (N must exceed 30 for both variables), the researcher is interested in testing whether there is a statistically significant difference between the proportion of headache sufferers of both samples using the different headache medication. A two-independent-variable Z-test for proportions can be applied because the sample size exceeds 30 on each market and the data collected are proportions.

This Z-test can be applied to three types of hypothesis tests: a two-tailed test, a right-tailed test, and a left-tailed test. A two-tailed hypothesis tests the null hypothesis, H_0 , that the difference in the population proportion is statistically identical to the hypothesized difference (if the hypothesized difference is set to zero, the null hypothesis tests if the population proportions of the two samples are identical). The alternative hypothesis is that the real difference in population proportions is statistically different from the hypothesized difference when tested using the sample dataset.

A right-tailed hypothesis tests the null hypothesis, H_0 , that the difference in the population proportion is statistically less than or equal to the hypothesized difference (if the hypothesized difference is set to zero, the null hypothesis tests if population proportion of Sample 1 is equal to or less than the population proportion of Sample 2). The alternative hypothesis is that the real difference in population proportions is statistically greater than the hypothesized difference when tested using the sample dataset.

A left-tailed hypothesis tests the null hypothesis, H_0 , that the difference in the population proportion is statistically greater than or equal to the hypothesized difference (if the hypothesized difference is set to

zero, the null hypothesis tests if population proportion of Sample 1 is equal to or greater than the population proportion of Sample 2). The alternative hypothesis is that the real difference in population proportions is statistically less than the hypothesized difference when tested using the sample dataset.

Two-Variable (F) Variances

The two-variable F-test analyzes the variances from two samples (the population variance of Sample 1 is tested with the population variance of Sample 2 to see if they are equal) and is appropriate when the population standard deviation is not known but the sampling distribution is assumed to be approximately normal.

The measurement of variation is a key issue in Six Sigma and quality control applications. In this illustration, suppose that the variation or variance around the units produced in a manufacturing process is compared to another process to determine which process is more variable and, hence, less predictable in quality.

This F-test can typically be applied to a single hypothesis test: a two-tailed test. A two-tailed hypothesis tests the null hypothesis, H_0 , such that the population variance of the two variables is statistically identical. The alternative hypothesis is that the population variances are statistically different from one another when tested using the sample dataset.

If the calculated p-value is less than or equal to the significance level in the test, then reject the null hypothesis and conclude that the true population variances of the two variables are not statistically equal to one another. Otherwise, the true population variances are statistically similar to each other.

Principal Component Analysis

Principal component analysis, or PCA, makes multivariate data easier to model and summarize. To understand PCA, suppose we start with n variables that are unlikely to be independent of one another, such that changing the value of one variable will change another variable. PCA modeling will replace the original n variables with a new set of m variables which are less than n but are uncorrelated to one another, while at the same time, each of these m variables is a linear combination of the original n variables, so that majority of the variation can be accounted for just using fewer explanatory variables. The mathematical properties can be summarized as $Y_k / \sum_{i=1}^n Y_i$, where R is the correlation matrix for the n

independent variables, the coefficients of the dependent variables corresponding to the k principal component are the elements of the eigenvector corresponding to the k largest eigenvalue, λ_k of the correlation matrix R . All of the eigenvalues are assumed to be real and non-negative as R is positive semidefinite. Therefore, starting from a large database of variables, we are now able to sort through and reduce the number of variables to several factors that will account for the majority of the observed variations in the independent variables, such that the linear combinations identified will have some natural interpretation.

R-Square Computation

In order to determine the best-fitting model, we apply several goodness-of-fit statistics to provide a glimpse into the accuracy and reliability of the estimated regression model. They usually take the form of

a t-statistic, F-statistic, R-squared statistic, adjusted R-squared statistic, Durbin-Watson statistic, Akaike Criterion, Schwarz Criterion, and their respective probabilities.

The R-squared (R^2), or coefficient of determination, is an error measurement that looks at the percent variation of the dependent variable that can be explained by the variation in the independent variable for a regression analysis. The coefficient of determination can be calculated by:

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2} = 1 - \frac{SSE}{TSS}$$

where the coefficient of determination is one less the ratio of the sums of squares of the errors (SSE) to the total sums of squares (TSS). In other words, the ratio of SSE to TSS is the unexplained portion of the analysis, thus, one less the ratio of SSE to TSS is the explained portion of the regression analysis.

The estimated regression line is characterized by a series of predicted values (\hat{Y}); the average value of the dependent variable's data points is denoted \bar{Y} ; and the individual data points are characterized by Y_i . Therefore, the total sum of squares, that is, the total variation in the data or the total variation about the average dependent value, is the total of the difference between the individual dependent values and its average (the total squared distance of $Y_i - \bar{Y}$). The explained sum of squares, the portion that is captured by the regression analysis, is the total of the difference between the regression's predicted value and the average dependent variable's dataset (seen as the total squared distance of $\hat{Y} - \bar{Y}$). The difference between the total variation (TSS) and the explained variation (ESS) is the unexplained sums of squares, also known as the sums of squares of the errors (SSE).

Another related statistic, the adjusted coefficient of determination, or the adjusted R-squared (\bar{R}^2), corrects for the number of independent variables (k) in a multivariate regression through a degrees of freedom correction to provide a more conservative estimate:

$$\bar{R}^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2 / (k - 2)}{\sum_{i=1}^n (Y_i - \bar{Y})^2 / (k - 1)} = 1 - \frac{SSE / (k - 2)}{TSS / (k - 1)}$$

The adjusted R-squared should be used instead of the regular R-squared in multivariate regressions because every time an independent variable is added into the regression analysis, the R-squared will increase, indicating that the percent variation explained has increased. This increase occurs even when nonsensical regressors are added. The adjusted R-squared takes the added regressors into account and penalizes the regression accordingly, providing a much better estimate of a regression model's goodness of fit.

Other goodness-of-fit statistics include the t-statistic and the F-statistic. The former is used to test if *each* of the estimated slope and intercept(s) is statistically significant, that is, if it is statistically significantly different from zero (therefore making sure that the intercept and slope estimates are statistically valid). The latter applies the same concepts but simultaneously for the entire regression equation including the

intercept and slopes. Using the previous example, the following illustrates how the t-statistic and F-statistic can be used in a regression analysis

Seasonality

Many time-series data exhibit seasonality where certain events repeat themselves after some time period or seasonality period (e.g., ski resorts' revenues are higher in winter than in summer, and this predictable cycle will repeat itself every winter). Seasonality periods represent how many periods would have to pass before the cycle repeats itself (e.g., 24 hours in a day, 12 months in a year, 4 quarters in a year, 60 minutes in an hour, etc.), but sometimes other seasonal periods exist that are not entirely evident by simply looking at the data or the variable. This seasonality test looks at your time-series data to determine the best-fitting seasonality periodicity for the data. Using this seasonality, you can now adjust for seasonal effects using the Deseasonalize Data tool or using the Time Series Analysis tool to provide a better forecast. Various seasonality periods (i.e., the number of periods per cycle) are tested and the results are presented below. The best-fitting seasonality periodicity is listed first (ranked by the lowest RMSE error measure) and all the relevant error measurements are included for comparison: root mean squared error (RMSE), mean squared error (MSE), mean absolute deviation (MAD), and mean absolute percentage error (MAPE).

Segmentation Clustering

A final analytical technique of interest is that of segmentation clustering. That is, taking the original dataset, we run some internal algorithms (a combination of k-means hierarchical clustering and other method of moments in order to find the best-fitting groups or natural statistical clusters) to statistically divide or segment the original dataset into two groups. Clearly, you can segment this dataset into as many groups as you wish. This technique is valuable in a variety of settings including marketing (market segmentation of customers into various customer relationship management groups etc.), physical sciences, engineering, and others.

Stepwise Regression (Backward)

In the backward method, run a regression with Y on all X variables and reviewing each variable's p-value, systematically eliminate the variable with the largest p-value. Then run a regression again, repeating each time until all p-values are statistically significant.

Stepwise Regression (Correlation)

In the correlation method, the dependent variable (Y) is correlated to all the independent variables (X), and starting with the X variable with the highest absolute correlation value, a regression is run. Then subsequent X variables are added until the p-values indicate that the new X variable is no longer statistically significant. This approach is quick and simple but does not account for interactions among variables, and an X variable, when added, will statistically overshadow other variables.

Stepwise Regression (Forward)

In the forward method, we first correlate Y with all X variables, run a regression for Y on the highest absolute value correlation of X, and obtain the fitting errors. Then, correlate these errors with the remaining X variables and choose the highest absolute value correlation among this remaining set and run

another regression. Repeat the process until the p-value for the latest X variable coefficient is no longer statistically significant and then stop the process.

Stepwise Regression (Forward-Backward)

In the forward and backward method, apply the forward method to obtain three X variables, then apply the backward approach to see if one of them needs to be eliminated because it is statistically insignificant. Repeat the forward method and then the backward method until all remaining X variables are considered.

Stochastic Process Estimations

A stochastic process is nothing but a mathematically defined equation that can create a series of outcomes over time, outcomes that are not deterministic in nature; that is, an equation or process that does not follow any simple discernible rule such as price will increase X percent every year or revenues will increase by this factor of X plus Y percent. A stochastic process is by definition nondeterministic, and one can plug numbers into a stochastic process equation and obtain different results every time. For instance, the path of a stock price is stochastic in nature, and one cannot reliably predict the exact stock price path with any certainty. However, the price evolution over time is enveloped in a process that generates these prices. The process is fixed and predetermined, but the outcomes are not. Hence, by stochastic simulation, we create multiple pathways of prices, obtain a statistical sampling of these simulations, and make inferences on the potential pathways that the actual price may undertake given the nature and parameters of the stochastic process used to generate the time series. Four stochastic processes are included in Risk Simulator's Forecasting tool, including Geometric Brownian motion or random walk, which is the most common and prevalently used process due to its simplicity and wide-ranging applications. The other three stochastic processes are the mean-reversion process, jump-diffusion process, and a mixed process.

The interesting thing about stochastic process simulation is that historical data is not necessarily required; that is, the model does not have to fit any sets of historical data. Simply compute the expected returns and the volatility of the historical data or estimate them using comparable external data or make assumptions about these values.

Brownian Motion Random Walk Process

The Brownian motion random walk process takes the form of $\frac{\delta S}{S} = \mu(\delta t) + \sigma\varepsilon\sqrt{\delta t}$ for regular options simulation, or a more generic version takes the form of $\frac{\delta S}{S} = (\mu - \sigma^2 / 2)\delta t + \sigma\varepsilon\sqrt{\delta t}$ for a geometric process. For an exponential version, we simply take the exponentials, and as an example, we have

$$\frac{\delta S}{S} = \exp[\mu(\delta t) + \sigma\varepsilon\sqrt{\delta t}],$$

where we define

- S as the variable's previous value
- δS as the change in the variable's value from one step to the next
- μ as the annualized growth or drift rate
- σ as the annualized volatility

To estimate the parameters from a set of time-series data, the drift rate and volatility can be found by setting μ to be the average of the natural logarithm of the relative returns $\ln \frac{S_t}{S_{t-1}}$ while σ is the standard deviation of all $\ln \frac{S_t}{S_{t-1}}$ values.

Mean-Reversion Process

The following describes the mathematical structure of a mean-reverting process with drift: $\frac{\delta S}{S} = \eta(\bar{S}e^{\mu(\delta t)} - S)\delta t + \mu(\delta t) + \sigma\varepsilon\sqrt{\delta t}$. In order to obtain the rate of reversion and long-term rate, using the historical data points, run a regression such that $Y_t - Y_{t-1} = \beta_0 + \beta_1 Y_{t-1} + \varepsilon$ and we find $\eta = -\ln[1 + \beta_1]$ and $\bar{S} = -\beta_0 / \beta_1$,

where we define

- η as the rate of reversion to the mean
- \bar{S} as the long-term value the process reverts to
- Y as the historical data series
- β_0 as the intercept coefficient in a regression analysis
- β_1 as the slope coefficient in a regression analysis

Jump-Diffusion Process

A jump-diffusion process is similar to a random walk process but there is a probability of a jump at any point in time. The occurrences of such jumps are completely random but their probability and magnitude are governed by the process itself.

$$\frac{\delta S}{S} = \eta(\bar{S}e^{\mu(\delta t)} - S)\delta t + \mu(\delta t) + \sigma\varepsilon\sqrt{\delta t} + \theta F(\lambda)(\delta t) \text{ for a jump diffusion process}$$

where we define

- θ as the jump size of S
- $F(\lambda)$ as the inverse of the Poisson cumulative probability distribution
- λ as the jump rate of S

The jump size can be found by computing the ratio of the postjump to the prejump levels, and the jump rate can be imputed from past historical data. The other parameters are found the same way as above.

Jump-Diffusion Process with Mean Reversion

This model is essentially a combination of all three models discussed above (geometric Brownian motion with mean-reversion process and a jump-diffusion process).

Structural Break

A structural break tests whether the coefficients in different datasets are equal, and this test is most commonly used in time-series analysis to test for the presence of a structural break. A time-series dataset can be divided into two subsets and each subset is tested on each other and on the entire dataset to statistically determine if, indeed, there is a break starting at a particular time period. The structural break test is often used to determine whether the independent variables have different impacts on different subgroups of the population, such as to test if a new marketing campaign, activity, major event, acquisition, divestiture, and so forth, have an impact on the time-series data. Suppose the dataset has 100 time-series data points. You can set various breakpoints to test, for instance, data points 10, 30, and 51 (this means that three structural break tests will be performed on the following dataset: data points 1-9 compared with 10-100; data points 1-29 compared with 30-100; and 1-50 compared with 51-100, to see if, indeed, at the start of data point 10, 30, and 51 there is a break in the underlying structure). A one-tailed hypothesis test is performed on the null hypothesis (H_0) such that the two data subsets are statistically similar to one another, that is, there is no statistically significant structural break. The alternative hypothesis (H_a) is that the two data subsets are statistically different from one another, indicating a possible structural break. If the calculated p-values are less than or equal to 0.01, 0.05, or 0.10, this means that the hypothesis is rejected, which implies that the two data subsets are statistically significantly different at the 1%, 5%, and 10% significance levels. High p-values indicate there is no statistically significant structural break.

Time-Series Analysis

Time-series forecasting decomposes the historical data into the baseline, trend, and seasonality, if any. The models then apply an optimization procedure to find the alpha, beta, and gamma parameters for the baseline, trend, and seasonality coefficients and then recombine them into a forecast. In other words, this methodology first applies a “backcast” to find the best-fitting model and best-fitting parameters of the model that minimizes forecast errors, and then proceeds to “forecast” the future based on the historical data that exist. This process, of course, assumes that the same baseline growth, trend, and seasonality hold going forward. Even if they do not, say, when there exists a structural shift (e.g., company goes global or has a merger, spin-off), the baseline forecasts can be computed and then the required adjustments can be made to the forecasts.

Time-Series Analysis (Auto)

There are eight models in the time-series analysis or time-series decomposition method for forecasting, and in each model, there are input parameters such as the historical base level (α), trend (β), and seasonality (γ), and each input will need to be calibrated in the model against the historical data provided. Selecting this automatic approach will allow the user to initiate an automated process in methodically selecting the best input parameters in each model and ranking the forecast models from best to worst by looking at their goodness-of-fit results and error measurements.

Several different types of errors can be calculated for time-series forecast methods, including the mean-squared error (MSE), root mean-squared error (RMSE), mean absolute deviation (MAD), and mean absolute percent error (MAPE). MSE is an absolute error measure that squares the errors (the difference between the actual historical data and the forecast-fitted data predicted by the model) to keep the positive and negative errors from canceling out each other. This measure also tends to exaggerate large errors by weighting the large errors more heavily than smaller errors by squaring them, which can help

when comparing different time-series models. MSE is calculated by simply taking the average of the $Error^2$. RMSE is the square root of MSE and is the most popular error measure, also known as the *quadratic loss function*. RMSE can be defined as the average of the absolute values of the forecast errors and is highly appropriate when the cost of the forecast errors is proportional to the absolute size of the forecast error. MAD is an error statistic that averages the distance (absolute value of the difference between the actual historical data and the forecast-fitted data predicted by the model) between each pair of actual and fitted forecast data points. MAD is calculated by taking the average of the $|Error|$ values, and it is most appropriate when the cost of forecast errors is proportional to the absolute size of the forecast errors. MAPE is a relative error statistic measured as an average percent error of the historical data points, and it is most appropriate when the cost of the forecast error is more closely related to the percentage error than to the numerical size of the error. This error estimate is calculated by taking the

average of the $\left| \frac{Y_t - \hat{Y}_t}{Y_t} \right|$ computations, where Y_t is the historical data at time t , while \hat{Y}_t is the fitted or

predicted data point at time t using this time-series method. Finally, an associated measure is the Theil's U statistic, which measures the naivety of the model's forecast. That is, if the Theil's U statistic is less than 1.0, then the forecast method used provides an estimate that is statistically better than guessing. The following provides the mathematical details of each error estimate.

RMSE	79.00	$RMSE = \sqrt{\frac{\sum_{i=1}^n (Error^2)_i}{n}} = \sqrt{MSE}$ $MSE = \frac{\sum_{i=1}^n (Error^2)_i}{n} = RMSE^2$ $MAD = \frac{\sum_{i=1}^n Error _i}{n}$ $MAPE = \frac{\sum_{i=1}^n \left \frac{Y_t - \hat{Y}_t}{Y_t} \right _i}{n}$	$Theil's U = \sqrt{\frac{\sum_{i=1}^n \left[\frac{\hat{Y}_t - Y_t}{Y_{t-1}} \right]_i^2}{\sum_{i=1}^n \left[\frac{Y_t - Y_{t-1}}{Y_{t-1}} \right]_i^2}}$
MSE	6241.27		
MAD	63.00		
MAPE	20.80%		
Theil's U	0.80		

Time-Series Analysis (DES)

The approach to use when the data exhibit a trend but no seasonality is the double exponential-smoothing (DES) method. Double exponential smoothing applies single exponential smoothing twice, once to the original data and then to the resulting single exponential smoothing data. An alpha (α) weighting parameter is used on the first or single exponential smoothing (SES), while a beta (β) weighting parameter is used on the second or double exponential smoothing. This approach is useful when the historical data series is not stationary. The forecast is calculated using the following:

$$DES_t = \beta (SES_t - SES_{t-1}) + (1 - \beta) DES_{t-1}$$

$$SES_t = \alpha Y_t + (1 - \alpha)(SES_{t-1} + DES_{t-1})$$

Time-Series Analysis (DMA)

The double moving average method smoothes out past data by performing a moving average on a subset of data that represents a moving average of an original set of data. That is, a second moving average is performed on the first moving average. The second moving average application captures the trending effect of the data. An example 3-month double moving average and forecast value uses the following:

$$\text{Forecast} = 2MA_{1,t} - MA_{2,t} + \frac{2}{m-1} [MA_{1,t} - MA_{2,t}]$$

where the forecast value is twice the amount of the first moving average (MA_1) at time t , less the second moving average estimate (MA_2) plus the difference between the two moving averages multiplied by a correction factor (two divided into the number of months in the moving average, m , less one).

Time-Series Analysis (HWA)

When both seasonality and trend exist, more advanced models are required to decompose the data into their base elements: a base case level (L) weighted by the alpha parameter (α); a trend component (b) weighted by the beta parameter (β); and a seasonality component (S) weighted by the gamma parameter (γ). Several methods exist but the two most common are the Holt-Winters additive seasonality and Holt-Winters multiplicative seasonality methods.

$$\text{Level } L_t = \alpha(Y_t - S_{t-s}) + (1 - \alpha)(L_{t-1} + b_{t-1})$$

$$\text{Trend } b_t = \beta(L_t - L_{t-1}) + (1 - \beta)(b_{t-1})$$

$$\text{Seasonality } S_t = \gamma(Y_t - L_t) + (1 - \gamma)(S_{t-s})$$

$$\text{Forecast } F_{t+m} = L_t + mb_t + S_{t+m-s}$$

Time-Series Analysis (HWM)

When both seasonality and trend exist, more advanced models are required to decompose the data into their base elements: a base case level (L) weighted by the alpha parameter (α); a trend component (b) weighted by the beta parameter (β); and a seasonality component (S) weighted by the gamma parameter (γ). Several methods exist but the two most common are the Holt-Winters additive seasonality and Holt-Winters multiplicative seasonality methods.

$$\text{Level } L_t = \alpha(Y_t / S_{t-s}) + (1 - \alpha)(L_{t-1} + b_{t-1})$$

$$\text{Trend } b_t = \beta(L_t - L_{t-1}) + (1 - \beta)(b_{t-1})$$

$$\text{Seasonality } S_t = \gamma(Y_t / L_t) + (1 - \gamma)(S_{t-s})$$

$$\text{Forecast } F_{t+m} = (L_t + mb_t)S_{t+m-s}$$

Time-Series Analysis (SA)

If the time-series data have no appreciable trend but exhibit seasonality, then the additive seasonality and multiplicative seasonality methods apply. The additive seasonality model breaks the historical data into a level (L) or base-case component as measured by the alpha parameter (α), and a seasonality (S) component measured by the gamma parameter (γ). The resulting forecast value is simply the addition of this base case level to the seasonality value.

$$\text{Level } L_t = \alpha(Y_t - S_{t-s}) + (1 - \alpha)(L_{t-1})$$

$$\text{Seasonality } S_t = \gamma(Y_t - L_t) + (1 - \gamma)(S_{t-s})$$

$$\text{Forecast } F_{t+m} = L_t + S_{t+m-s}$$

Time-Series Analysis (SES)

The second approach to use when no discernable trend or seasonality exists is the single exponential-smoothing method. This method weights past data with exponentially decreasing weights going into the past; that is, the more recent the data value, the greater its weight. This weighting largely overcomes the limitations of moving averages or percentage-change models. The weight used is termed the *alpha* measure. The method uses the following model:

$$ESF_t = \alpha Y_{t-1} + (1 - \alpha)ESF_{t-1}$$

where the exponential smoothing forecast (ESF_t) at time t is a weighted average between the actual value one period in the past (Y_{t-1}) and last period's forecast (ESF_{t-1}), weighted by the alpha parameter (α).

Time-Series Analysis (SM)

Similarly, the seasonality multiplicative model requires the alpha and gamma parameters. The difference being that the model is multiplicative, for example, the forecast value is the multiplication between the base case level and seasonality factor.

$$\text{Level } L_t = \alpha(Y_t / S_{t-s}) + (1 - \alpha)(L_{t-1})$$

$$\text{Seasonality } S_t = \gamma(Y_t / L_t) + (1 - \gamma)(S_{t-s})$$

$$\text{Forecast } F_{t+m} = L_t S_{t+m-s}$$

Time-Series Analysis (SMA)

The single moving average is applicable when time-series data with no trend and seasonality exist. The approach simply uses an average of the actual historical data to project future outcomes. This average is applied consistently moving forward, hence the term *moving average*.

The value of the moving average (*MA*) for a specific length (n) is simply the summation of actual historical data (Y) arranged and indexed in time sequence (i).

$$MA_n = \frac{\sum_{i=1}^n Y_i}{n}$$

Trending and Detrending

This tool detrends your original data to take out any trending components. In forecasting models, the process removes the effects of accumulating datasets from seasonality and trend to show only the absolute changes in values and to allow potential cyclical patterns to be identified after removing the general drift, tendency, twists, bends, and effects of seasonal cycles of a set of time-series data. For example, a detrended dataset may be necessary to discover a company's true financial health—one may detrend increased sales around Christmas time to see a more accurate account of a company's sales in a given year more clearly by shifting the entire dataset from a slope to a flat surface to better see the underlying cycles and fluctuations. The resulting charts show the effects of the detrended data against the original dataset, and the statistics reports show the percentage of the trend that was removed based on each detrending method employed, as well as the actual detrended dataset. The following lists the trend line analysis methods used for forecasting and detrending methods for identifying cycles in data:

Trend Line (Difference Detrended)

Trend Line (Exponential Detrended)

Trend Line (Exponential)

Trend Line (Linear Detrended)

Trend Line (Linear)

Trend Line (Logarithmic Detrended)

Trend Line (Logarithmic)

Trend Line (Moving Average Detrended)

Trend Line (Moving Average)

Trend Line (Polynomial Detrended)

Trend Line (Polynomial)

Trend Line (Power Detrended)

Trend Line (Power)

Trend Line (Rate Detrended)

Trend Line (Static Mean Detrended)

Trend Line (Static Median Detrended)

Volatility: GARCH Models

The generalized autoregressive conditional heteroskedasticity (GARCH) model is used to model historical and forecast future volatility levels of a marketable security (e.g., stock prices, commodity prices, oil prices, etc.). The dataset has to be a time series of raw price levels. GARCH will first convert the prices into relative returns and then run an internal optimization to fit the historical data to a mean-reverting volatility term structure, while assuming that the volatility is heteroskedastic in nature (changes over time according to some econometric characteristics). All other variations below are based on the original GARCH model, but account for specific idiosyncrasies of the data. The typical volatility forecast situation requires $P = 1$, $Q = 1$; Periodicity = number of periods per year (12 for monthly data, 52 for weekly data, 252 or 365 for daily data); Base = minimum of 1 and up to the periodicity value; and Forecast Periods = number of annualized volatility forecasts you wish to obtain.

GARCH models are used mainly in analyzing financial time-series data to ascertain their conditional variances and volatilities. These volatilities are then used to value the options as usual, but the amount of historical data necessary for a good volatility estimate remains significant. Usually, several dozen—and even up to hundreds—of data points are required to obtain good GARCH estimates. GARCH is a term that incorporates a family of models that can take on a variety of forms, known as GARCH(p,q), where p and q are positive integers that define the resulting GARCH model and its forecasts. In most cases for financial instruments, a GARCH(1,1) is sufficient and is most generally used. For instance, a GARCH (1,1) model takes the form of:

$$y_t = x_t\gamma + \varepsilon_t$$

$$\sigma_t^2 = \omega + \alpha\varepsilon_{t-1}^2 + \beta\sigma_{t-1}^2$$

where the first equation's dependent variable (y_t) is a function of exogenous variables (x_t) with an error term (ε_t). The second equation estimates the variance (squared volatility σ_t^2) at time t , which depends on a historical mean (ω); news about volatility from the previous period, measured as a lag of the squared residual from the mean equation (ε_{t-1}^2); and volatility from the previous period (σ_{t-1}^2). The exact modeling specification of a GARCH model is beyond the scope of this book. Suffice it to say that detailed knowledge of econometric modeling (model specification tests, structural breaks, and error estimation) is required to run a GARCH model, making it less accessible to the general analyst. Another problem with GARCH models is that the model usually does not provide a good statistical fit. That is, it is impossible to predict the stock market and, of course, equally if not harder to predict a stock's volatility over time. Note that the GARCH function has several inputs as follow:

- *Time-Series Data.* The time series of data in chronological order (e.g., stock prices). Typically, dozens of data points are required for a decent volatility forecast.
- *Periodicity.* A positive integer indicating the number of periods per year (e.g., 12 for monthly data, 252 for daily trading data, etc.), assuming you wish to annualize the volatility. For getting periodic volatility, enter 1.
- *Predictive Base.* The number of periods (the time-series data) back to use as a base to forecast volatility. The higher this number, the longer the historical base is used to forecast future volatility.
- *Forecast Period.* A positive integer indicating how many future periods beyond the historical stock prices you wish to forecast.
- *Variance Targeting.* This variable is set as False by default (even if you do not enter anything here) but can be set as True. False means the omega variable is automatically optimized and computed. The suggestion is to leave this variable empty. If you wish to create mean-reverting volatility with variance targeting, set this variable as True.
- *P.* The number of previous lags on the mean equation.
- *Q.* The number of previous lags on the variance equation.

There are several GARCH models available in this software, including EGARCH, EGARCH-T, GARCH-M, GJR-GARCH, GJR-GARCH-T, IGARCH, and T-GARCH. For the GARCH-M models, the conditional variance equations are the same in the six variations but the mean questions are different and assumption on z_t can be either normal distribution or t-distribution. The estimated parameters for GARCH-M with normal distribution are those five parameters in the mean and conditional variance equations. The estimated parameters for GARCH-M with the t-distribution are those five parameters in the mean and conditional variance equations plus another parameter, the degrees of freedom for the t-distribution. In contrast, for the GJR models, the mean equations are the same in the six variations and the differences are that the conditional variance equations and the assumption on z_t can be either a normal distribution or t-distribution. The estimated parameters for EGARCH and GJR-GARCH with normal distribution are those four parameters in the conditional variance equation. The estimated parameters for GARCH, EARCH, and GJR-GARCH with t-distribution are those parameters in the conditional variance equation plus

the degrees of freedom for the t-distribution The detailed theoretical specifics of a GARCH model are outside the purview of this user manual.

The accompanying tables list some of the GARCH specifications used in Risk Simulator with two underlying distributional assumptions: one for normal distribution and the other for the t-distribution. For the GARCH-M models, the conditional variance equations are the same in the six variations but the mean questions are different and assumption on z_t can be either normal distribution or t-distribution. The estimated parameters for GARCH-M with normal distribution are those five parameters in the mean and conditional variance equations. The estimated parameters for GARCH-M with the t-distribution are those five parameters in the mean and conditional variance equations plus another parameter, the degrees of freedom for the t-distribution. In contrast, for the GJR models, the mean equations are the same in the six variations and the differences are that the conditional variance equations and the assumption on z_t can be either a normal distribution or t-distribution. The estimated parameters for EGARCH and GJR-GARCH with normal distribution are those four parameters in the conditional variance equation. The estimated parameters for GARCH, EARCH, and GJR-GARCH with t-distribution are those parameters in the conditional variance equation plus the degrees of freedom for the t-distribution.

Volatility: GARCH and TGARCH

	$z_t \sim$ Normal Distribution	$z_t \sim$ T-Distribution
GARCH	$y_t = x_t \gamma + \varepsilon_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$

Volatility: GARCH-M

	$z_t \sim$ Normal Distribution	$z_t \sim$ T-Distribution
GARCH-M Variance in Mean Equation	$y_t = c + \lambda \sigma_t^2 + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$	$y_t = c + \lambda \sigma_t^2 + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$
GARCH-M Standard Deviation in Mean Equation	$y_t = c + \lambda \sigma_t + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$	$y_t = c + \lambda \sigma_t + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$
GARCH-M Log Variance in Mean Equation	$y_t = c + \lambda \ln(\sigma_t^2) + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$	$y_t = c + \lambda \ln(\sigma_t^2) + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$

Volatility: EGARCH and EGARCH-T

	$z_t \sim \text{Normal Distribution}$	$z_t \sim \text{T-Distribution}$
EGARCH	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\ln(\sigma_t^2) = \omega + \beta \cdot \ln(\sigma_{t-1}^2) + \alpha \left[\left \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right - E(\varepsilon_t) \right] + r \frac{\varepsilon_{t-1}}{\sigma_{t-1}}$ $E(\varepsilon_t) = \sqrt{\frac{2}{\pi}}$	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\ln(\sigma_t^2) = \omega + \beta \cdot \ln(\sigma_{t-1}^2) + \alpha \left[\left \frac{\varepsilon_{t-1}}{\sigma_{t-1}} \right - E(\varepsilon_t) \right] + r \frac{\varepsilon_{t-1}}{\sigma_{t-1}}$ $E(\varepsilon_t) = \frac{2\sqrt{v-2} \Gamma((v+1)/2)}{(v-1)\Gamma(v/2)\sqrt{\pi}}$

Volatility: GJR GARCH and GJR TGARCH

	$z_t \sim \text{Normal Distribution}$	$z_t \sim \text{T-Distribution}$
GJR-GARCH	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + r \varepsilon_{t-1}^2 d_{t-1} + \beta \sigma_{t-1}^2$ $d_{t-1} = \begin{cases} 1 & \text{if } \varepsilon_{t-1} < 0 \\ 0 & \text{otherwise} \end{cases}$	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + r \varepsilon_{t-1}^2 d_{t-1} + \beta \sigma_{t-1}^2$ $d_{t-1} = \begin{cases} 1 & \text{if } \varepsilon_{t-1} < 0 \\ 0 & \text{otherwise} \end{cases}$

Volatility: TGARCH and TGARCH-M

	$z_t \sim \text{T-Distribution}$
GARCH-M Variance in Mean Equation	$y_t = c + \lambda \sigma_t^2 + \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$
GARCH	$y_t = \varepsilon_t$ $\varepsilon_t = \sigma_t z_t$ $\sigma_t^2 = \omega + \alpha \varepsilon_{t-1}^2 + \beta \sigma_{t-1}^2$

Volatility: Log Returns Approach

The *Logarithmic Cash Flow Returns or Logarithmic Stock Price Returns Approach* calculates the volatility using the individual future cash flow estimates, comparable cash flow estimates, or historical prices, generating their corresponding logarithmic relative returns. Starting with a series of forecast future cash flows or historical prices, convert them into relative returns. Then take the natural logarithms of these relative returns. The standard deviation of these natural logarithm returns is the *periodic volatility* of the cash flow series. The resulting periodic volatility from the sample dataset must be annualized.

No matter what the approach used, the periodic volatility estimate used in a real options or financial options analysis has to be an *annualized volatility*. Depending on the periodicity of the raw cash flow or stock price data used, the volatility calculated should be converted into annualized values using $\sigma\sqrt{P}$, where P is the number of periods in a year and σ is the periodic volatility. For instance, if the calculated volatility using monthly cash flow data is 10%, the annualized volatility is $10\%\sqrt{12} = 35\%$. Similarly, P is 365 (or about 250 if accounting for trading days and not calendar days) for daily data, 4 for quarterly data, 2 for semiannual data, and 1 for annual data.

Yield Curve (Bliss)

Several alternative methods exist for estimating the term structure of interest rates and the yield curve. Some are fully specified stochastic term structure models, while others are simply interpolation models. The former include the CIR and Vasicek models, while the latter are interpolation models such as the Bliss or Nelson approach. We now look at the Bliss interpolation model (BIM) for generating the term structure of interest rates and yield curve estimation. Some econometric modeling techniques are required to calibrate the values of several input parameters in this model. The Bliss approach modifies the Nelson-Siegel method by adding an additional generalized parameter. Virtually any yield curve shape can be interpolated using these models, which are widely used at banks around the world.

Yield Curve (Nelson-Siegel)

The Nelson-Siegel (NS) is an interpolation model for generating the term structure of interest rates and yield curve estimation. Some econometric modeling techniques are required to calibrate the values of several input parameters in this model. Just like the Bliss model, the NS model is purely an interpolation model, with four estimated parameters. If properly modeled, it can be made to fit almost any yield curve shape. Calibrating the inputs in the NS model requires facility with econometric modeling and error optimization techniques. Typically, if some interest rates exist, a better approach is to use the spline interpolation method.

APPENDIX: DATABASE SQL USE CASES AND EXAMPLES

SQL Conditional Use Cases

The following are some common use cases in which a large dataset can be screened, cleaned, and filtered to return the required rows of data for analysis in ROV Quantitative Data Miner. Each use case shows a quick summary of the problem to be solved, accompanied by the XLS sample data file, QDM sample profile and model name, as well as some simple screen shots to illustrate the existing dataset, the approach taken, and the results. Clearly, the sample datasets are intentionally kept small to facilitate the learning experience, but the same approaches and techniques illustrated in this use case document are applicable for all dataset sizes. To follow along, start QDM and click on [File >> Examples >> 08 SQL on Data Mapping](#). In the first Data tab, you will see the results from the SQL commands. To view these commands, click on [Variables >> Group Management](#), select the Variable you wish to review, and click [Edit](#) to see how the Data Link and SQL commands work.

Below is a quick summary of the key items in these use case examples:

Variable > Value obtains rows above a specific threshold value.

Variable > 80 AND Variable < 100 allows you to append with AND to create multiple filters.

Variable < 80 OR Variable > 100 allows you to select data with OR condition filters.

(Variable > 80 AND Variable < 90) OR (Variable > 100) allows you to append nested AND/OR.

Variable IN ('aaa', 'ccc') allows you to match rows with certain strings in the data.

Variable BETWEEN 80 AND 100 allows a selection of values between two numbers.

Variable LIKE '%AN%' uses wildcard matching % long strings and characters including spaces.

Variable LIKE '_AN' allows wildcard matching of a single character (_).

Variable1 / Variable2, Variable1 * Variable2, Variable1 + Variable2 runs computations.

(Y/100 + Z /10)/ 3 > X OR (Z - Y/100) > X allows combinations of OR with computations.

X < 4 UNION SELECT X FROM [first\$] WHERE X > 10 allows union of multiple queries.

ISNUMERIC(Variable) allows the selection of numerical values only.

1 = 2 UNION SELECT TOP 5 [first\$].Z FROM [first\$] allows choosing top few rows by incorporating the union and top functions.

NOT X IN (SELECT TOP 5 [first\$].[X] FROM [first\$]) selects data not in the first few top rows.

EXISTS (SELECT [first\$].Z FROM [first\$] WHERE Z>75) checks to see if the query returns any values or if not, returns an empty set.

Variable1 IN (SELECT [second\$].[A] FROM [second\$]) combines multiple data tables.

NOTES: **`Long Variable Names`** use back-tick to apply long variable names and regular ticks (apostrophe) for values (e.g., **`Country of Origin`** = 'United States').

Union will always sort the results of the first column in ascending order.

Use Case 1: Selection of Rows by Value

Situation: In a large dataset, we can use the conditional statement to select the rows with specific values (e.g., greater than a required threshold).

SQL Statement: `Variable > Value`

Example: `Number > 100`

Example Variable (see example file 08 SQL on Data Mapping): `GTE 100`

Example Data File: `Sample Data 1.xls`

Note: You can use `>=`, `<=`, `>`, `<` inequalities

	A	B	C	D	E		A	B	C	D	E
1	Number	String	Money	DATE	Mix	16	110.851872	ooo	¥83.37	4/13/1900	103.05
2	102.885141	aaa	¥108.16	4/13/1900	102.88	17	102.680179	ppp	¥82.69	4/5/1900	102.87
3	84.705038	bbb	¥100.89	4/13/1900	aaa	18	69.971395	qqq	¥92.11	4/21/1900	*
4	92.160695	ccc	¥108.16	4/8/1900	122.26	19	93.250333	rrr	¥98.85	4/6/1900	89.48
5	102.185995	ddd	¥104.63	3/31/1900	108.46	20	96.570041	sss	¥104.34	3/26/1900	&&
6	91.309775	eee	¥111.91	4/28/1900	100.64	21	97.653884	ttt	¥101.73	4/4/1900	108.71
7	126.086623	fff	¥99.98	3/27/1900	99.74	22	75.886155	uuu	¥115.73	3/29/1900	108.09
8	108.029949	ggg	¥91.16	4/3/1900	110.64	23	107.151940	vvv	¥86.08	4/5/1900	95.85
9	88.869916	hhh	¥100.39	4/27/1900	83.52	24	103.529863	www	¥114.24	3/28/1900	95.75
10	95.844675	iii	¥108.17	4/4/1900	106.53	25	108.222820	xxx	¥100.62	4/20/1900	()
11	100.831152	jjj	¥115.29	4/2/1900	##@##	26	106.491195	yyy	¥106.88	4/16/1900	108.62
12	107.798552	kkk	¥98.99	4/1/1900	107.64	27	80.822858	zzz	¥117.50	4/16/1900	100.64
13	111.168206	lll	¥78.03	4/19/1900	96.88	28	91.103886	abc	¥103.53	4/8/1900	95.56
14	93.873964	mmm	¥109.57	4/7/1900	101.14	29	92.807726	def	¥85.72	4/14/1900	103.20
15	100.974688	nnn	¥104.37	4/16/1900	113.02	30	107.605145	ghi	¥84.81	4/4/1900	%%%%
						31	94.677514	jkm	¥101.05	4/13/1900	FCCC

Dataset	DL_GTE 100	DL_80 to 100	DL_80- or 100+	DL_80-90 or 100+
1	102.885141	84.705038	102.885141	102.885141
2	102.185995	92.160695	102.185995	84.705038
3	126.086623	91.309775	126.086623	102.185995
4	108.029949	88.869916	108.029949	126.086623
5	100.831152	95.844675	100.831152	108.029949
6	107.798552	93.873964	107.798552	88.869916
7	111.168206	93.250333	111.168206	100.831152
8	100.974688	96.570041	100.974688	107.798552
9	110.851872	97.653884	110.851872	111.168206
10	102.680179	80.822858	102.680179	100.974688
11	107.151940	91.103886	107.151940	110.851872
12	103.529863	92.807726	75.886155	102.680179
13	108.222820	94.677514	107.151940	107.151940
14	106.491195		103.529863	103.529863
15	107.605145		108.222820	108.222820
16			106.491195	106.491195
17			107.605145	80.822858
18				107.605145

Use Case 2: Use of 'AND'

Situation: Use 'AND' to connect two or more conditions together, if all conditions are "TRUE", then the data is selected.

SQL Statement: condition AND condition AND condition AND...

Example: Number > 80 AND Number < 100

Example Variable (see example file 08 SQL on Data Mapping): 80 to 100

Example Data File: Sample Data 1.xls

	A	B	C	D	E		A	B	C	D	E
1	Number	String	Money	DATE	Mix	16	110.851872	ooo	¥83.37	4/13/1900	103.05
2	102.885141	aaa	¥108.16	4/13/1900	102.88	17	102.680179	ppp	¥82.69	4/5/1900	102.87
3	84.705038	bbb	¥100.89	4/13/1900	aaa	18	69.971395	qqq	¥92.11	4/21/1900	*
4	92.160695	ccc	¥108.16	4/8/1900	122.26	19	93.250333	rrr	¥98.85	4/6/1900	89.48
5	102.185995	ddd	¥104.63	3/31/1900	108.46	20	96.570041	sss	¥104.34	3/26/1900	&&
6	91.309775	eee	¥111.91	4/28/1900	100.64	21	97.653884	ttt	¥101.73	4/4/1900	108.71
7	126.086623	fff	¥99.98	3/27/1900	99.74	22	75.886155	uuu	¥115.73	3/29/1900	108.09
8	108.029949	ggg	¥91.16	4/3/1900	110.64	23	107.151940	vvv	¥86.08	4/5/1900	95.85
9	88.869916	hhh	¥100.39	4/27/1900	83.52	24	103.529863	www	¥114.24	3/28/1900	95.75
10	95.844675	iii	¥108.17	4/4/1900	106.53	25	108.222820	xxx	¥100.62	4/20/1900	()
11	100.831152	jjj	¥115.29	4/2/1900	#&@#&#	26	106.491195	yyy	¥106.88	4/16/1900	108.62
12	107.798552	kkk	¥98.99	4/1/1900	107.64	27	80.822858	zzz	¥117.50	4/16/1900	100.64
13	111.168206	lll	¥78.03	4/19/1900	96.88	28	91.103886	abc	¥103.53	4/8/1900	95.56
14	93.873964	mmm	¥109.57	4/7/1900	101.14	29	92.807726	def	¥85.72	4/14/1900	103.20
15	100.974688	nnn	¥104.37	4/16/1900	113.02	30	107.605145	ghi	¥84.81	4/4/1900	%%%%%
						31	94.677514	jkm	¥101.05	4/13/1900	FCCC

Dataset	DL_GTE 100	DL_80 to 100	DL_80- or 100+	DL_80-90 or 100+
1	102.885141	84.705038	102.885141	102.885141
2	102.185995	92.160695	102.185995	84.705038
3	126.086623	91.309775	126.086623	102.185995
4	108.029949	88.869916	108.029949	126.086623
5	100.831152	95.844675	100.831152	108.029949
6	107.798552	93.873964	107.798552	88.869916
7	111.168206	93.250333	111.168206	100.831152
8	100.974688	96.570041	100.974688	107.798552
9	110.851872	97.653884	110.851872	111.168206
10	102.680179	80.822858	102.680179	100.974688
11	107.151940	91.103886	69.971395	110.851872
12	103.529863	92.807726	75.886155	102.680179
13	108.222820	94.677514	107.151940	107.151940
14	106.491195		103.529863	103.529863
15	107.605145		108.222820	108.222820
16			106.491195	106.491195
17			107.605145	80.822858
18				107.605145

Use Case 3: Use of 'OR'

Situation: Use 'OR' to connect two or more conditions together, once a condition is "TRUE", the data is selected even when other conditions are "FALSE".

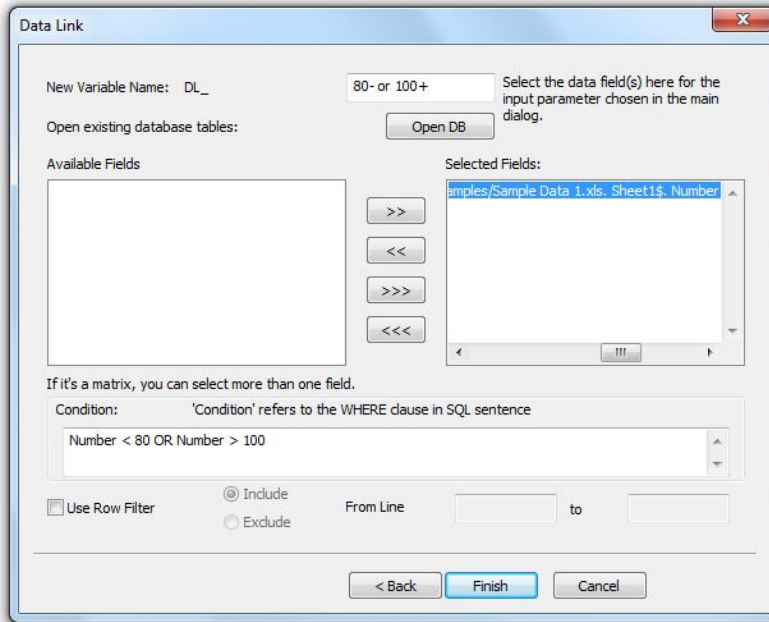
SQL Statement: condition OR condition OR condition OR...

Example: Number < 80 OR Number > 100

Example Variable (see example file 08 SQL on Data Mapping): 80- or 100+

Example Data File: [Sample Data 1.xls](#)

	A	B	C	D	E		A	B	C	D	E
1	Number	String	Money	DATE	Mix	16	110.851872	ooo	¥83.37	4/13/1900	103.05
2	102.885141	aaa	¥108.16	4/13/1900	102.88	17	102.680179	ppp	¥82.69	4/5/1900	102.87
3	84.705038	bbb	¥100.89	4/13/1900	aaa	18	69.971395	qqq	¥92.11	4/21/1900	*
4	92.160695	ccc	¥108.16	4/8/1900	122.26	19	93.250333	rrr	¥98.85	4/6/1900	89.48
5	102.185995	ddd	¥104.63	3/31/1900	108.46	20	96.570041	sss	¥104.34	3/26/1900	&&
6	91.309775	eee	¥111.91	4/28/1900	100.64	21	97.653884	ttt	¥101.73	4/4/1900	108.71
7	126.086623	fff	¥99.98	3/27/1900	99.74	22	75.886155	uuu	¥115.73	3/29/1900	108.09
8	108.029949	ggg	¥91.16	4/3/1900	110.64	23	107.151940	vvv	¥86.08	4/5/1900	95.85
9	88.869916	hhh	¥100.39	4/27/1900	83.52	24	103.529863	www	¥114.24	3/28/1900	95.75
10	95.844675	iii	¥108.17	4/4/1900	106.53	25	108.222820	xxx	¥100.62	4/20/1900	()
11	100.831152	jjj	¥115.29	4/2/1900	#3@#3#	26	106.491195	yyy	¥106.88	4/16/1900	108.62
12	107.798552	kkk	¥98.99	4/1/1900	107.64	27	80.822858	zzz	¥117.50	4/16/1900	100.64
13	111.168206	lll	¥78.03	4/19/1900	96.88	28	91.103886	abc	¥103.53	4/8/1900	95.56
14	93.873964	mmm	¥109.57	4/7/1900	101.14	29	92.807726	def	¥85.72	4/14/1900	103.20
15	100.974688	nnn	¥104.37	4/16/1900	113.02	30	107.605145	ghi	¥84.81	4/4/1900	%%%
						31	94.677514	jkm	¥101.05	4/13/1900	FCCC



Dataset	Visualize			
	DL_GTE 100	DL_80 to 100	DL_80- or 100+	DL_80-90 or 100+
1	102.885141	84.705038	102.885141	102.885141
2	102.185995	92.160695	102.185995	84.705038
3	126.086623	91.309775	126.086623	102.185995
4	108.029949	88.869916	108.029949	126.086623
5	100.831152	95.844675	100.831152	108.029949
6	107.798552	93.873964	107.798552	88.869916
7	111.168206	93.250333	111.168206	100.831152
8	100.974688	96.570041	100.974688	107.798552
9	110.851872	97.653884	110.851872	111.168206
10	102.680179	80.822858	102.680179	100.974688
11	107.151940	91.103886	69.971395	110.851872
12	103.529863	92.807726	75.886155	102.680179
13	108.222820	94.677514	107.151940	107.151940
14	106.491195		103.529863	103.529863
15	107.605145		108.222820	108.222820
16			106.491195	106.491195
17			107.605145	80.822858
18				107.605145

Use Case 4: Use of 'AND' and 'OR' together

Situation: We can use 'AND' and 'OR' together to build a complex query command.

SQL Statement: condition AND condition OR condition...

Example: (Number > 80 AND Number < 90) OR (Number > 100)

Example Variable (see example file 08 SQL on Data Mapping): 80-90 or 100+

Example Data File: Sample Data 1.xls

Note: you can group commands using parenthesis ().

	A	B	C	D	E		A	B	C	D	E
1	Number	String	Money	DATE	Mix	16	110.851872	ooo	¥83.37	4/13/1900	103.05
2	102.885141	aaa	¥108.16	4/13/1900	102.88	17	102.680179	ppp	¥82.69	4/5/1900	102.87
3	84.705038	bbb	¥100.89	4/13/1900	aaa	18	69.971395	qqq	¥92.11	4/21/1900	*
4	92.160695	ccc	¥108.16	4/8/1900	122.26	19	93.250333	rrr	¥98.85	4/6/1900	89.48
5	102.185995	ddd	¥104.63	3/31/1900	108.46	20	96.570041	sss	¥104.34	3/26/1900	&&
6	91.309775	eee	¥111.91	4/28/1900	100.64	21	97.653884	ttt	¥101.73	4/4/1900	108.71
7	126.086623	fff	¥99.98	3/27/1900	99.74	22	75.886155	uuu	¥115.73	3/29/1900	108.09
8	108.029949	ggg	¥91.16	4/3/1900	110.64	23	107.151940	vvv	¥86.08	4/5/1900	95.85
9	88.869916	hhh	¥100.39	4/27/1900	83.52	24	103.529863	www	¥114.24	3/28/1900	95.75
10	95.844675	iii	¥108.17	4/4/1900	106.53	25	108.222820	xxx	¥100.62	4/20/1900	()
11	100.831152	jjj	¥115.29	4/2/1900	¥\$@#%#	26	106.491195	yy	¥106.88	4/16/1900	108.62
12	107.798552	kkk	¥98.99	4/1/1900	107.64	27	80.822858	zzz	¥117.50	4/16/1900	100.64
13	111.168206	lll	¥78.03	4/19/1900	96.88	28	91.103886	abc	¥103.53	4/8/1900	95.56
14	93.873964	mmm	¥109.57	4/7/1900	101.14	29	92.807726	def	¥85.72	4/14/1900	103.20
15	100.974688	nnn	¥104.37	4/16/1900	113.02	30	107.605145	ghi	¥84.81	4/4/1900	%%%%
						31	94.677514	jkm	¥101.05	4/13/1900	FCCC

Data Link

New Variable Name: DL_ 80-90 or 100+ Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields: Samples/Sample Data 1.xls, Sheet1\$. Number

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

(Number > 80 AND Number < 90) OR (Number > 100)

Use Row Filter Include From Line to Exclude

< Back Finish Cancel

Dataset	Visualize					
		DL_GTE 100	DL_80 to 100	DL_80- or 100+	DL_80-90 or 100+	DL_String If
1		102.885141	84.705038	102.885141	102.885141	102.885141
2		102.185995	92.160695	102.185995	84.705038	84.705038
3		126.086623	91.309775	126.086623	102.185995	92.160695
4		108.029949	88.869916	108.029949	126.086623	108.029949
5		100.831152	95.844675	100.831152	108.029949	88.869916
6		107.798552	93.873964	107.798552	88.869916	95.844675
7		111.168206	93.250333	111.168206	100.831152	
8		100.974688	96.570041	100.974688	107.798552	
9		110.851872	97.653884	110.851872	111.168206	
10		102.680179	80.822858	102.680179	100.974688	
11		107.151940	91.103886	69.971395	110.851872	
12		103.529863	92.807726	75.886155	102.680179	
13		108.222820	94.677514	107.151940	107.151940	
14		106.491195		103.529863	103.529863	
15		107.605145		108.222820	108.222820	
16				106.491195	106.491195	
17				107.605145	80.822858	
18				107.605145		

Use Case 5: Use of 'IN'

Situation: Use 'IN' command to specific a value (or multiple values) to match.

SQL Statement: Variable IN ('value1', 'value2'...)

Example: String IN ('aaa', 'ccc')

Example Variable (see example file 08 SQL on Data Mapping): String IN

Example Data File: [Sample Data 2.xls](#)

Note: If the values filtered are strings, use 'quotes'.

	A	B
1	Number	String
2	102.885141	aaa
3	84.705038	aaa
4	92.160695	aaa
5	102.185995	bbb
6	91.309775	bbb
7	126.086623	bbb
8	108.029949	ccc
9	88.869916	ccc
10	95.844675	ccc
11	100.831152	ddd
12	70.121340	ddd
13	65.121212	ddd

Data Link

New Variable Name: DL_ String IN Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields: Samples/Sample Data 2.xls, Sheet1\$. Number

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence
String IN ('aaa','ccc')

Use Row Filter Include From Line: to: Exclude

< Back Finish Cancel

Dataset	DL_80- or 100+	DL_80-90 or 100+	DL_String IN	DL_BETWEEN	DL_LIKE
	102.885141	102.885141	102.885141	84.705038	1500.000000
	102.185995	84.705038	84.705038	92.160695	250.000000
	126.086623	102.185995	92.160695	91.309775	300.000000
	108.029949	126.086623	108.029949	88.869916	
	100.831152	108.029949	88.869916	95.844675	
	107.798552	88.869916	95.844675		
	111.168206	100.831152			
	100.974688	107.798552			
	110.851872	111.168206			
	102.680179	100.974688			
	69.971395	110.851872			
	75.886155	102.680179			
	107.151940	107.151940			
	103.529863	103.529863			
	108.222820	108.222820			
	106.491195	106.491195			
	107.605145	80.822858			
		107.605145			

Use Case 6: Use of 'BETWEEN'

Situation: Using 'BETWEEN' selects data within a specific range.

SQL Statement: Variable BETWEEN 'value1' AND 'value2'

Example: Number BETWEEN 80 AND 100

Example Variable (see example file 08 SQL on Data Mapping): BETWEEN model

Example Data File: Sample Data 2.xls

	A	B
1	Number	String
2	102.885141	aaa
3	84.705038	aaa
4	92.160695	aaa
5	102.185995	bbb
6	91.309775	bbb
7	126.086623	bbb
8	108.029949	ccc
9	88.869916	ccc
10	95.844675	ccc
11	100.831152	ddd
12	70.121340	ddd
13	65.121212	ddd

Data Link

New Variable Name: DL_ BETWEEN Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields:

Number BETWEEN 80 AND 100

Condition: 'Condition' refers to the WHERE clause in SQL sentence

Number BETWEEN 80 AND 100

Use Row Filter Include From Line: to: Exclude

< Back Finish Cancel

Dataset	Visualize	DL_80- or 100+	DL_80-90 or 100+	DL_String IN	DL_BETWEEN	DL_LIKE
		102.885141	102.885141	102.885141	84.705038	1500.000000
		102.185995	84.705038	84.705038	92.160695	250.000000
		126.086623	102.185995	92.160695	91.309775	300.000000
		108.029949	126.086623	108.029949	88.869916	
		100.831152	108.029949	88.869916	95.844675	
		107.798552	88.869916	95.844675		
		111.168206	100.831152			
		100.974688	107.798552			
		110.851872	111.168206			
		102.680179	100.974688			
		69.971395	110.851872			

Use Case 7: Use of 'LIKE'

Situation: The 'LIKE' condition allows you to use wildcards in the Where clause, allowing you to perform pattern matching.

SQL Statement:

The patterns that you can choose from are:

% allows you to match any string of any length (including zero length)

_ allows you to match on a single character

Example: store_name LIKE '%AN%'

Example Variable (see example file 08 SQL on Data Mapping): LIKE

Example Data File: Sample Data 3.xls

	A	B	C
1	Store_Name	Number	Date
2	Los Angeles	1500.00	8/1/2008
3	San Diego	250.00	5/1/2008
4	San Francisco	300.00	2008/6/31
5	Boston	700.00	4/23/2008

Data Link

New Variable Name: DL_ LIKE Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields: Sample Data 3.xls. Store_Information\$. Number

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

Store_Name LIKE (%an%)

Use Row Filter Include From Line to Exclude

< Back Finish Cancel

Dataset	Visualize	DL_String IN	DL_BETWEEN	DL_LIKE	DL_X	DL_Y
		102.885141	84.705038	1500.000000	9.560000	3000.000000
		84.705038	92.160695	250.000000	10.780000	2600.000000
		92.160695	91.309775	300.000000		
		108.029949	88.869916			
		88.869916	95.844675			
		95.844675				

Use Case 8: Simple Math Functions

Situation: Basic mathematical functions can be applied on variables.

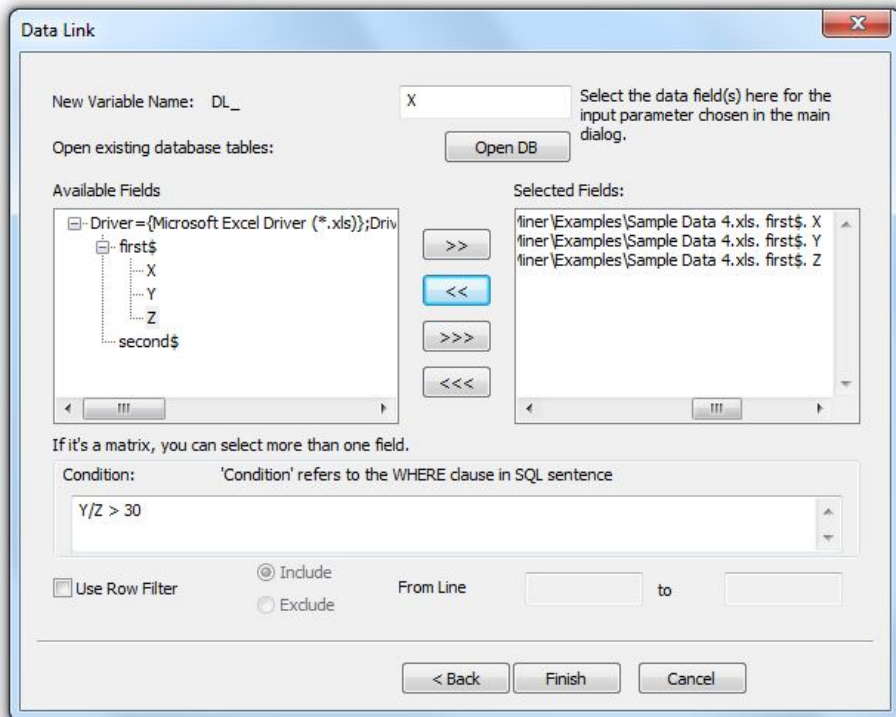
SQL Statement: Variable1 / Variable2, Variable1 * Variable2, Variable1 + Variable2 ...

Example: $Y/Z > 30$

Example Variable (see example file 08 SQL on Data Mapping): X, Y, Z

Example Data File: Sample Data 4.xls

	A	B	C
1	X	Y	Z
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00



DL_LIKE	DL_X	DL_Y	DL_Z	DL_NESTED
1500.000000	9.560000	3000.000000	65.000000	3.450000
250.000000	10.780000	2600.000000	35.000000	3.780000
300.000000				6.440000
				7.120000
				9.560000
				2.180000
				7.660000
				8.930000

Use Case 9: Nested Math Functions

Situation: The math functions can be very complex (just like in any mathematical equation).

Example: $(Y/100 + Z/10)/3 > X$ OR $(Z - Y/100) > X$

Example Variable (see example file 08 SQL on Data Mapping): NESTED

Example Data File: Sample Data 4.xls

	A	B	C
1	X	Y	Z
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00

Data Link

New Variable Name: DL_ NESTED Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields:

tliner/Examples/Sample Data 4.xls, first\$. X

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

$(Y/100 + Z/10)/3 > X$ OR $(Z - Y/100) > X$

Use Row Filter Include From Line to

Exclude

< Back Finish Cancel

DL_X	DL_Z	DL_NESTED	DL_UNION	DL_ISNUMERIC
9.560000	65.000000	3.450000	2.180000	3.450000
10.780000	35.000000	3.780000	3.450000	3.780000
		6.440000	3.780000	6.440000
		7.120000	10.780000	7.120000
		9.560000		2.180000
		2.180000		10.780000
		7.660000		8.930000
		8.930000		

Use Case 10: Use of 'Union' to Connect Commands

Situation: 'Union' is a very important command to connect two or more query results together. When creating complex commands, divide the entire command into small pieces and apply 'Union'.

SQL Statement: `CONDITION1 UNION SELECT COLUMN FROM TABLENAME WHERE CONDITION2`

Example: `X < 4 UNION SELECT X FROM [first$] WHERE X > 10`

Example Variable (see example file 08 SQL on Data Mapping): UNION

Example Data File: Sample Data 4.xls

Note: Using Union can sometimes sort the resulting dataset.

	A	B	C
1	X	Y	Z
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00

DL_NESTED	DL_UNION	DL_ISNUMERIC	DL_NOT IN	DL_EXISTS
3.45	2.18	3.45	2.18	3.45
3.78	3.45	3.78	7.66	3.78
6.44	3.78	6.44	10.78	6.44
7.12	10.78	7.12	8.93	7.12
9.56		2.18		9.56
2.18		10.78		2.18
7.66		8.93		7.66
8.93				10.78
				8.93

Use Case 11: Filtering Different Value Types

Situation: If a column of data has mixed numbers and strings or other value types, we can filter in numerical data by applying the 'ISNUMERIC' command.

SQL Statement: ISNUMERIC(Variable)

Example: ISNUMERIC(Number)

Example Variable (see example file 08 SQL on Data Mapping): ISNUMERIC

Example Data File: Sample Data 5.xls

	A
1	Number
2	3.45
3	3.78
4	6.44
5	7.12
6	AaA
7	2.18
8	BBB
9	10.78
10	8.93

Data Link

New Variable Name: DL_ ISNUMERIC Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields: Examples/Sample Data 5.xls, first\$. Number

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence
ISNUMERIC(Number)

Use Row Filter Include From Line to Exclude

< Back Finish Cancel

DL_NESTED	DL_UNION	DL_ISNUMERIC	DL_NOT IN	DL_EXISTS
3.450000	2.180000	3.450000	2.180000	3.450000
3.780000	3.450000	3.780000	7.660000	3.780000
6.440000	3.780000	6.440000	10.780000	6.440000
7.120000	10.780000	7.120000	8.930000	7.120000
9.560000		2.180000		9.560000
2.180000		10.780000		2.180000
7.660000		8.930000		7.660000
8.930000				10.780000
				8.930000

Use Case 12: Choosing the Top N Rows

Situation: To select the top N rows in a table, use 'UNION' and 'TOP' commands together.

SQL Statement: TOP N * FROM TABLE_NAME

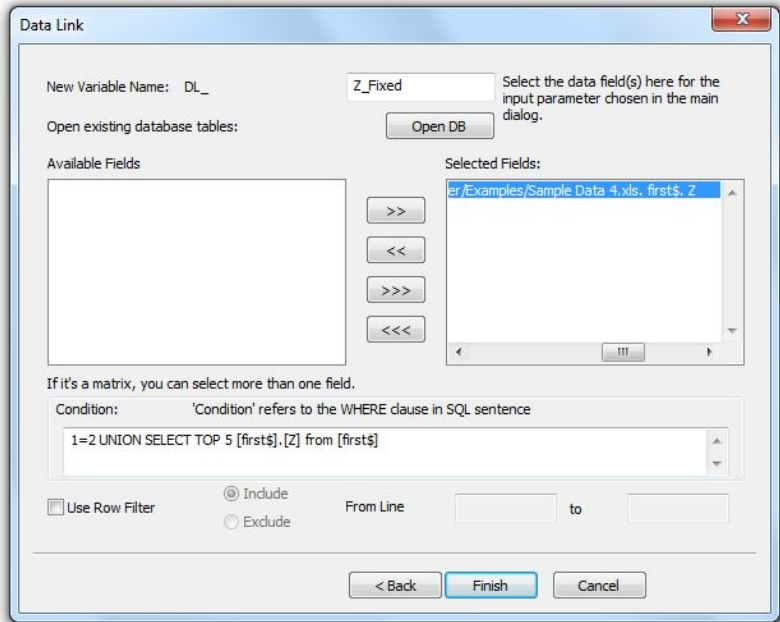
Example: 1 = 2 UNION SELECT TOP 5 [first\$].Z FROM [first\$]

Caution: The second SELECT command's selected rows must be the same with the selected rows in list box. '1 = 2' means Forever FALSE, so make the first select condition has no result.

Example Variable (see example file 08 SQL on Data Mapping): Z FIXED

Example Data File: Sample Data 4.xls

	A	B	C
1	X	Y	Z
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00



Result:

Dataset	Visualize			
DL_X_Fixed	DL_Y_Fixed	DL_Z_Fixed	DL_Y_Fixed_Fix...	DL_X_Fixed_Fix...
3.45	250.00	45.00	1500.00	3.45
3.78	300.00	50.00		
6.44	700.00	55.00		
7.12	1500.00	65.00		
9.56	3000.00			

Use Case 13: Use of 'NOT IN'

Situation: 'NOT IN' is used to filter out values obtained from the next condition command. If the column's value is unique, it can be used to obtain values from a range of rows.

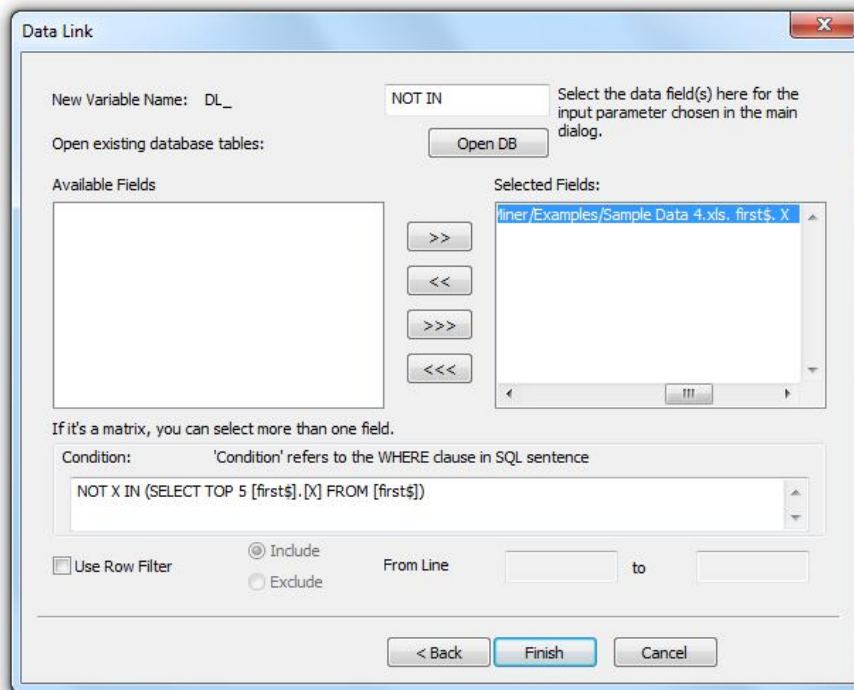
SQL Statement: NOT Variable IN (command)

Example: NOT X IN (SELECT TOP 5 [first\$].[X] FROM [first\$])

Example Variable (see example file 08 SQL on Data Mapping): NOT IN

Example Data File: Sample Data 4.xls

	A	B	C
1	X	Y	Z
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00



Dataset	Visualize	DL_NESTED	DL_UNION	DL_ISNUMERIC	DL_NOT IN	DL_EXISTS
		3.450000	2.180000	3.450000	2.180000	3.450000
		3.780000	3.450000	3.780000	7.660000	3.780000
		6.440000	3.780000	6.440000	10.780000	6.440000
		7.120000	10.780000	7.120000	8.930000	7.120000
		9.560000		2.180000		9.560000
		2.180000		10.780000		2.180000
		7.660000		8.930000		7.660000
		8.930000				10.780000
						8.930000

Use Case 14: Use of 'EXISTS'

Situation: 'EXISTS' simply tests whether the inner query returns any rows. If it does, then the outer query proceeds. If not, the outer query does not execute, and the entire SQL statement returns nothing.

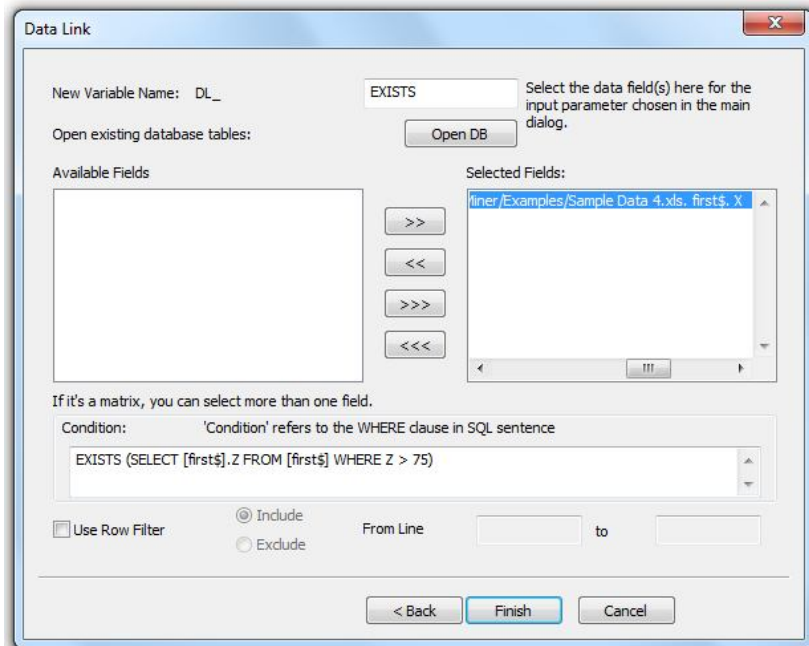
SQL Statement: EXISTS (SELECT * FROM "table_name2" WHERE [Condition])

Example: EXISTS (SELECT [first\$].Z FROM [first\$] WHERE Z>75)

Example Variable (see example file 08 SQL on Data Mapping): EXISTS

Example Data File: [Sample Data 4.xls](#)

	A	B	C
1	X	Y	Z
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00



Dataset	DL_NESTED	DL_UNION	DL_ISNUMERIC	DL_NOT IN	DL_EXISTS
	3.450000	2.180000	3.450000	2.180000	3.450000
	3.780000	3.450000	3.780000	7.660000	3.780000
	6.440000	3.780000	6.440000	10.780000	6.440000
	7.120000	10.780000	7.120000	8.930000	7.120000
	9.560000		2.180000		9.560000
	2.180000		10.780000		2.180000
	7.660000		8.930000		7.660000
	8.930000				10.780000
					8.930000

Use Case 15: Use of Multiple Table

Situation: Use the 'SELECT' command to connect multiple tables for matching elements.

SQL Statement: Variable1 IN (SELECT Variable2 FROM Table_Name2 WHERE Condition2)

Example: X IN (SELECT [second\$].[A] FROM [second\$])

Example Variable (see example file 08 SQL on Data Mapping): SELECT

Example Data File: Sample Data 4.xls

	A	B	C
1	X	Y	Z
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00

FIRST TABLE

	A
1	A
2	3.45
3	3.78
4	8.93
5	6.66

SECOND TABLE

Data Link

New Variable Name: DL_ SELECT Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields: Miner/Examples/Sample Data 4.xls. first\$. X

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence
 X IN (SELECT [second\$].[A] FROM [second\$])

Use Row Filter Include From Line to Exclude

< Back Finish Cancel

Dataset	Visualize			
DL_SELECT	DL_MULTIPLE...	DL_SOUNDS...	DL_WILDCARD...	DL_NESTED...
3.450000	1.000000	1.000000	6.000000	4.000000
3.780000	3.000000	3.000000	9.000000	6.000000
8.930000	4.000000	5.000000	12.000000	7.000000
	5.000000	14.000000	15.000000	8.000000
	6.000000	15.000000	17.000000	9.000000
	7.000000	17.000000		10.000000
	8.000000			11.000000
	10.000000			12.000000
	11.000000			15.000000
	12.000000			17.000000
	15.000000			18.000000
	16.000000			
	17.000000			
	18.000000			
	19.000000			

Use Case 16: Example Using 'AND'

Situation: Select the student's numbers for those who passed every test.

Example Variable (see example file 08 SQL on Data Mapping): **MULTIPLE AND**

Example Data File: **Sample Data 6.xls**

	A	B	C	D	E	F	G
1	Students No	Name	Age	Math	English	Biology	Geography
2	1	John	16	95	66	83	76
3	2	Tom	15	67	78	55	89
4	3	Jerry	16	93	67	92	87
5	4	Bob	17	88	88	97	92
6	5	Alexandra	16	77	98	89	68
7	6	William	18	78	100	100	70
8	7	Lily	15	96	79	87	89
9	8	Rose	16	91	84	79	90
10	9	Jack	14	99	57	92	93
11	10	Vivi	18	94	77	86	96
12	11	Vicky	15	87	65	95	75
13	12	Babala	15	99	97	95	96
14	13	Chris	17	76	57	87	98
15	14	Amanda	16	56	78	95	90
16	15	Alice	16	89	77	98	100
17	16	Amy	15	83	67	66	91
18	17	Annie	17	96	87	92	91
19	18	Cindy	16	78	89	92	85
20	19	Cora	17	67	82	83	89
21	20	Ella	18	67	65	86	56

Data Link

New Variable Name: DL_ MULTIPLE AND Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields

Selected Fields:

files/Sample Data 6.xls, first\$. Students No

>>
<<
>>>
<<<

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

Math > 60 AND English > 60 AND Biology > 60 AND Geography > 60

Use Row Filter
 Include
 From Line to

 Exclude

< Back Finish Cancel

Dataset	Visualize
DL_SELECT	DL_MULTIPLE...
3.450000	1.000000
3.780000	3.000000
8.930000	4.000000
	5.000000
	6.000000
	7.000000
	8.000000
	10.000000
	11.000000
	12.000000
	15.000000
	16.000000
	17.000000
	18.000000
	19.000000

Use Case 17: Example Using Wildcards with 'AND'

Situation: Select the student's numbers whose Names begin with 'A' or 'J' and Age older than 16.

Example Variable (see example file 08 SQL on Data Mapping): SOUNDS LIKE model

Example Data File: Sample Data 6.xls

	A	B	C	D	E	F	G
1	Students No	Name	Age	Math	English	Biology	Geography
2	1	John	16	95	66	83	76
3	2	Tom	15	67	78	55	89
4	3	Jerry	16	93	67	92	87
5	4	Bob	17	88	88	97	92
6	5	Alexandra	16	77	98	89	68
7	6	William	18	78	100	100	70
8	7	Lily	15	96	79	87	89
9	8	Rose	16	91	84	79	90
10	9	Jack	14	99	57	92	93
11	10	Vivi	18	94	77	86	96
12	11	Vicky	15	87	65	95	75
13	12	Babala	15	99	97	95	96
14	13	Chris	17	76	57	87	98
15	14	Amanda	16	56	78	95	90
16	15	Alice	16	89	77	98	100
17	16	Amy	15	83	67	66	91
18	17	Annie	17	96	87	92	91
19	18	Cindy	16	78	89	92	85
20	19	Cora	17	67	82	83	89
21	20	Ella	18	67	65	86	56

Data Link

New Variable Name: DL_ SOUNDS LIKE Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields: tples/Sample Data 6.xls. first\$. Students No

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

(Name LIKE 'J%' OR Name LIKE 'A%') AND (Age > 15)

Use Row Filter Include From Line to Exclude

< Back Finish Cancel

DL_SELECT	DL_MULTIPLE...	DL_SOUNDS...	DL_WILDCARD...	DL_NESTED...
3.450000	1.000000	1.000000	6.000000	4.000000
3.780000	3.000000	3.000000	9.000000	6.000000
8.930000	4.000000	5.000000	12.000000	7.000000
	5.000000	14.000000	15.000000	8.000000
	6.000000	15.000000	17.000000	9.000000
	7.000000	17.000000		10.000000
	8.000000			11.000000
	10.000000			12.000000
	11.000000			15.000000
	12.000000			17.000000
	15.000000			18.000000
	16.000000			
	17.000000			
	18.000000			
	19.000000			

Use Case 18: Example Using 'Union' with Sorting

Situation: Select the top 5 highest scores in Geography.

Example Variable (see example file 08 SQL on Data Mapping): **Students No, Geography**

Example Data File: **Sample Data 6.xls**

	A	B	C	D	E	F	G
1	Students No	Name	Age	Math	English	Biology	Geography
2	1	John	16	95	66	83	76.000000
3	2	Tom	15	67	78	55	88.601113
4	3	Jerry	16	93	67	92	87.485345
5	4	Bob	17	88	88	97	92.185350
6	5	Alexandra	16	77	98	89	68.000000
7	6	William	18	78	100	100	70.188680
8	7	Lily	15	96	79	87	89.425400
9	8	Rose	16	91	84	79	89.604188
10	9	Jack	14	99	57	92	92.732209
11	10	Vivi	18	94	77	86	96.048082
12	11	Vicky	15	87	65	95	75.000000
13	12	Babala	15	99	97	95	96.000000
14	13	Chris	17	76	57	87	98.260693
15	14	Amanda	16	56	78	95	89.711824
16	15	Alice	16	89	77	98	99.578803
17	16	Amy	15	83	67	66	91.170687
18	17	Annie	17	96	87	92	91.370947
19	18	Cindy	16	78	89	92	85.000000
20	19	Cora	17	67	82	83	89.131081
21	20	Ella	18	67	65	86	56.000000

Data Link

New Variable Name: DL_ Geography Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables:

Available Fields:

- first\$
 - Students No
 - Name
 - Age
 - Math
 - English
 - Biology
 - Geography

Selected Fields:

- bles/Sample Data 6.xls.first\$. Geography
- bles/Sample Data 6.xls.first\$. Students No

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

1 = 0 UNION (SELECT TOP 5 Geography FROM [first\$] ORDER BY Geography DESC)

Use Row Filter Include From Line to Exclude

DL_DISTINCT	DL_ORDER BY	DL_Geography	DL_Students No	DL_X_Fixed
111.000000	230.000000	92.732209	9.000000	3.450000
222.000000	250.000000	96.000000	10.000000	3.780000
333.000000	300.000000	96.048082	12.000000	6.440000
444.000000	700.000000	98.260693	13.000000	7.120000
555.000000	1000.000000	99.578803	15.000000	9.560000
	1500.000000			
	2100.000000			
	2600.000000			
	3000.000000			

Use Case 19: Example Using Wildcards and Math

Situation: Select the students whose names contain the character 'A' and the average score is greater than 85.

Example Variable (see example file 08 SQL on Data Mapping): WILDCARD MATH

Example Data File: Sample Data 6.xls

	A	B	C	D	E	F	G	H
1	Students No	Name	Age	Math	English	Biology	Geography	Average
2	1	John	16	95	66	83	76	80.00
3	2	Tom	15	67	78	55	89	72.15
4	3	Jerry	16	93	67	92	87	84.90
5	4	Bob	17	88	88	97	92	91.26
6	5	Alexandra	16	77	98	89	68	82.98
7	6	William	18	78	100	100	70	86.88
8	7	Lily	15	96	79	87	89	87.86
9	8	Rose	16	91	84	79	90	85.55
10	9	Jack	14	99	57	92	93	85.13
11	10	Vivi	18	94	77	86	96	88.31
12	11	Vicky	15	87	65	95	75	80.53
13	12	Babala	15	99	97	95	96	96.60
14	13	Chris	17	76	57	87	98	79.57
15	14	Amanda	16	56	78	95	90	79.79
16	15	Alice	16	89	77	98	100	91.04
17	16	Amy	15	83	67	66	91	76.79
18	17	Annie	17	96	87	92	91	91.59
19	18	Cindy	16	78	89	92	85	86.17
20	19	Cora	17	67	82	83	89	80.44
21	20	Ella	18	67	65	86	56	68.50

Data Link X

New Variable Name: DL_ WILDCARD MATH Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields

Selected Fields:

es/Sample Data 6.xls, first\$, Students No

>>
<<
>>>
<<<

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

Name LIKE '%a%' AND (Math + English + Biology + Geography)/4 > 85

Use Row Filter
 Include
 From Line to

< Back
Finish
Cancel

Dataset	Visualize	DL_MULTIPLE...	DL_SOUNDS...	DL_WILDCARD...	DL_NESTED...	DL_FUNCTIONS
		1.000000	1.000000	6.000000	4.000000	3150.000000
		3.000000	3.000000	9.000000	6.000000	
		4.000000	5.000000	12.000000	7.000000	
		5.000000	14.000000	15.000000	8.000000	
		6.000000	15.000000	17.000000	9.000000	
		7.000000	17.000000		10.000000	
		8.000000			11.000000	
		10.000000			12.000000	
		11.000000			15.000000	
		12.000000			17.000000	
		15.000000			18.000000	
		16.000000				
		17.000000				
		18.000000				
		19.000000				

Use Case 20: Example Using Nested 'AND/OR' with Math

Question: Select the students who have an average score between 85 and 95 when the student's age is ≥ 16 or has an average score higher than 80 when the student's age is < 16 .

Example Variable (see example file 08 SQL on Data Mapping): NESTED AND OR

Example Data File: Sample Data 6.xls

	A	B	C	D	E	F	G	H
1	Students No	Name	Age	Math	English	Biology	Geography	Average
2	1	John	16	95	66	83	76	80.00
3	2	Tom	15	67	78	55	89	72.15
4	3	Jerry	16	93	67	92	87	84.90
5	4	Bob	17	88	88	97	92	91.26
6	5	Alexandra	16	77	98	89	68	82.98
7	6	William	18	78	100	100	70	86.88
8	7	Lily	15	96	79	87	89	87.86
9	8	Rose	16	91	84	79	90	85.55
10	9	Jack	14	99	57	92	93	85.13
11	10	Vivi	18	94	77	86	96	88.31
12	11	Vicky	15	87	65	95	75	80.53
13	12	Babala	15	99	97	95	96	96.60
14	13	Chris	17	76	57	87	98	79.57
15	14	Amanda	16	56	78	95	90	79.29
16	15	Alice	16	89	77	98	100	91.04
17	16	Amy	15	83	67	66	91	76.79
18	17	Annie	17	96	87	92	91	91.59
19	18	Cindy	16	78	89	92	85	86.17
20	19	Cora	17	67	82	83	89	80.44
21	20	Ella	18	67	65	86	56	68.50

(Age ≥ 16 AND ((Math + English + Biology + Geography)/4 BETWEEN 85 AND 95)) OR (Age < 16 AND ((Math + English + Biology + Geography)/4 > 80))

Data Link

New Variable Name: DL_

Open existing database tables:

Available Fields:

Selected Fields: des/Sample Data 6.xls. first\$. Students No

Condition: 'Condition' refers to the WHERE clause in SQL sentence

(Age ≥ 16 AND ((Math + English + Biology + Geography)/4 BETWEEN 85 AND 95)) OR (Age < 16 AND ((Math + English + Biology + Geography)/4 > 80))

Use Row Filter Include Exclude

From Line: to

Dataset	Visualize	DL_MULTIPLE...	DL_SOUNDS...	DL_WILDCARD...	DL_NESTED...	DL_FUNCTIONS
1.000000		1.000000		6.000000	4.000000	3150.000000
3.000000		3.000000		9.000000	6.000000	
4.000000		5.000000		12.000000	7.000000	
5.000000		14.000000		15.000000	8.000000	
6.000000		15.000000		17.000000	9.000000	
7.000000		17.000000			10.000000	
8.000000					11.000000	
10.000000					12.000000	
11.000000					15.000000	
12.000000					17.000000	
15.000000					18.000000	

Use Case 21: Use of 'AND'

Situation: The purpose of the 'AND' command is to combine the results of two queries.

SQL Statement: [SQL Statement 1] AND [SQL Statement 2]

Example: $X < 4$ AND $Y > 1000$

Example Variable (see example file 08 SQL on Data Mapping): X FIXED FIXED, Y FIXED FIXED

Example Data File: Sample Data 4.xls

	X	Y	Z
1			
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00

Data Link

New Variable Name: DL_ X_Fixed_Fixed Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields Selected Fields:

er/Examples/Sample Data 4.xls.first\$. X

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

X < 4 AND Y > 1000

Use Row Filter Include Exclude From Line to

< Back Finish Cancel

Dataset	Visualize				
DL_X_Fixed	DL_Y_Fixed	DL_Z_Fixed	DL_Y_Fixed_Fix...	DL_X_Fixed_Fix...	
3.45	250.00	45.00	1500.00	3.45	
3.78	300.00	50.00			
6.44	700.00	55.00			
7.12	1500.00	65.00			
9.56	3000.00				

Use Case 22: Use of SQL Functions

Situation: SQL has several arithmetic functions, they are 'AVG', 'COUNT', 'MAX', 'MIN', 'SUM'. They are useful when we have to do some function with the result.

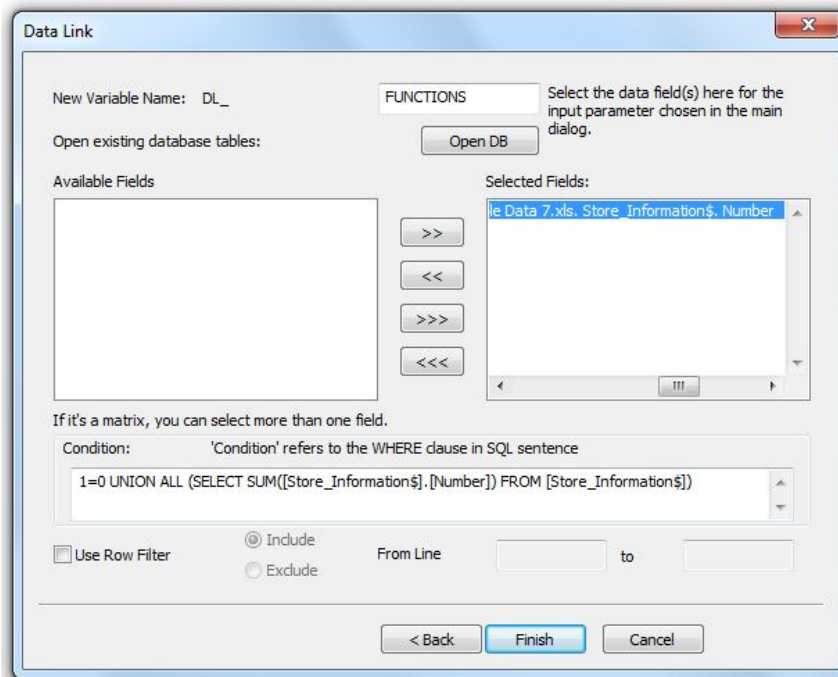
SQL Statement: `SELECT "function type"("column_name") FROM "table_name"`

Example: `1 = 0 UNION ALL (SELECT SUM([Store_Information$].[Number]) FROM [Store_Information$])`

Example Variable (see example file 08 SQL on Data Mapping): FUNCTIONS model

Example Data File: Sample Data 7.xls

1	Store_Name	Number	Date
2	Los Angeles	1500.00	2008/8/1
3	San Diego	250.00	2008/5/1
4	San Francisco	300.00	2008/6/31
5	Boston	700.00	2008/4/23
6	Los Angeles	400.00	2008/6/1



Dataset	Visualize	DL_GROUP BY	DL_DISTINCT	DL_ORDER BY	DL_Geography
DL_FUNCTIONS	3150.000000	700.000000	111.000000	230.000000	92.732209
		1900.000000	222.000000	250.000000	96.000000
		250.000000	333.000000	300.000000	96.048082
		300.000000	444.000000	700.000000	98.260693
			555.000000	1000.000000	99.578803
				1500.000000	
				2100.000000	
				2600.000000	
				3000.000000	

Use Case 23: Use of 'GROUP BY'

Situation: In Use Case 22 we use 'SUM' to compute the total number of all stores, but what do we do if we want to compute each store's number? We can accomplish this by using 'GROUP BY'.

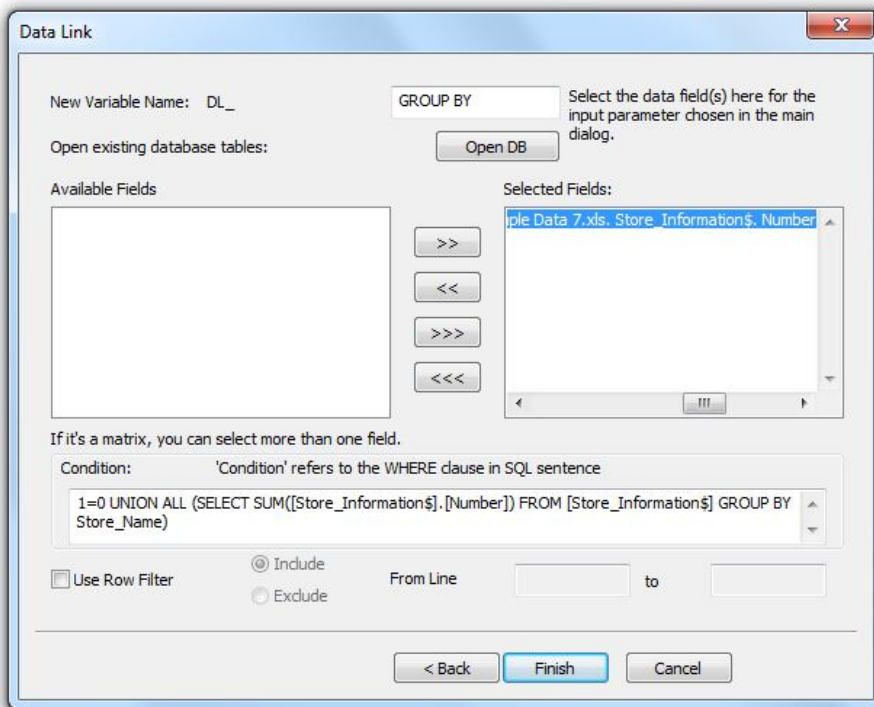
SQL Statement: SELECT "column_name1", SUM("column_name2") FROM "table_name" GROUP BY "column_name1"

Example: 1 = 0 UNION ALL (SELECT SUM([Store_Information\$].[Number]) FROM [Store_Information\$] GROUP BY Store_Name)

Example Variable (see example file 08 SQL on Data Mapping): GROUP BY

Example Data File: Sample Data 7.xls

1	Store_Name	Number	Date
2	Los Angeles	1500.00	2008/8/1
3	San Diego	250.00	2008/5/1
4	San Francisco	300.00	2008/6/31
5	Boston	700.00	2008/4/23
6	Los Angeles	400.00	2008/6/1



DL_FUNCTIONS	DL_GROUP BY	DL_DISTINCT	DL_ORDER BY	DL_Geography
3150.000000	700.000000	111.000000	230.000000	92.732209
	1900.000000	222.000000	250.000000	96.000000
	250.000000	333.000000	300.000000	96.048082
	300.000000	444.000000	700.000000	98.260693
		555.000000	1000.000000	99.578803
			1500.000000	
			2100.000000	
			2600.000000	
			3000.000000	

Use Case 24: Use of 'DISTINCT'

Situation: When in a column where some values are similar and you don't want to show them, use the 'DISTINCT' command for showing unique values.

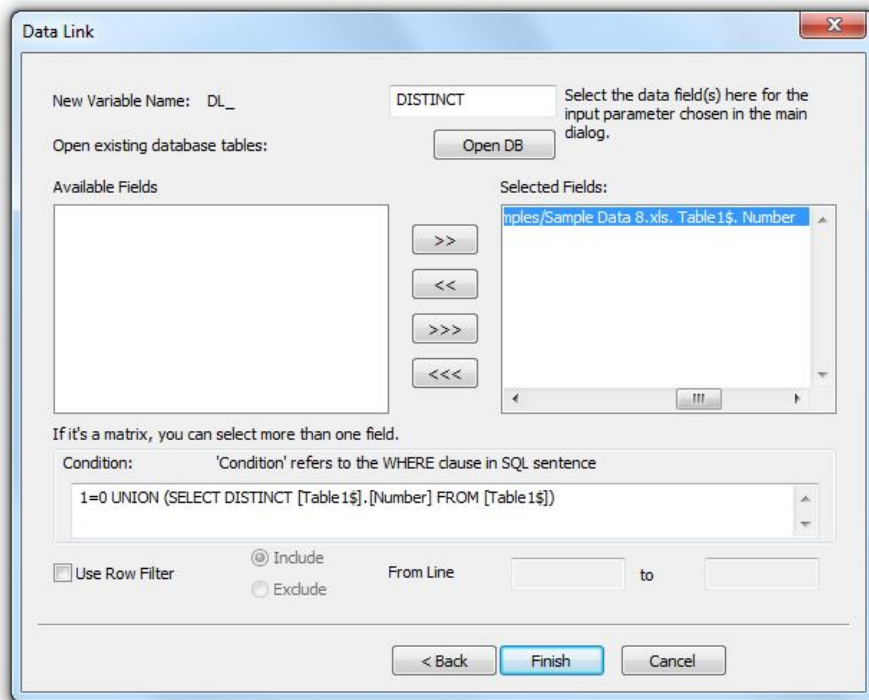
SQL Statement: SELECT DISTINCT Variable FROM Table_name

Example: 1 = 0 UNION (SELECT DISTINCT [Table1\$].[Number] FROM [Table1\$])

Example Variable (see example file 08 SQL on Data Mapping): **DISTINCT**

Example Data File: [Sample Data 8.xls](#)

	Number
1	
2	111
3	111
4	111
5	222
6	222
7	222
8	333
9	444
10	555
11	444
12	222



DL_FUNCTIONS	DL_GROUP BY	DL_DISTINCT	DL_ORDER BY	DL_Geography
3150.000000	700.000000	111.000000	230.000000	92.732209
	1900.000000	222.000000	250.000000	96.000000
	250.000000	333.000000	300.000000	96.048082
	300.000000	444.000000	700.000000	98.260693
		555.000000	1000.000000	99.578803
			1500.000000	
			2100.000000	
			2600.000000	
			3000.000000	

Use Case 25: Use of 'ORDER BY'

Situation: When you need to list the data in a particular order, use the 'ORDER BY' command.

SQL Statement: `SELECT "column_name" FROM "table_name" [WHERE "condition"] ORDER BY "column_name" [ASC, DESC]`

Example: `Number > 80 AND Number < 100`

Example Variable (see example file 08 SQL on Data Mapping): ORDER BY

Example Data File: Sample Data 4.xls

1	X	Y	Z
2	3.45	1500.00	50.00
3	3.78	250.00	45.00
4	6.44	300.00	55.00
5	7.12	700.00	55.00
6	9.56	3000.00	65.00
7	2.18	230.00	75.00
8	7.66	2100.00	80.00
9	10.78	2600.00	35.00
10	8.93	1000.00	40.00

Data Link

New Variable Name: DL_ ORDER BY Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables: Open DB

Available Fields: Selected Fields: Miner/Examples/Sample.Data.4.xls.first\$.Y

If it's a matrix, you can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence
1=1 ORDER BY Y ASC

Use Row Filter Include From Line to Exclude

< Back Finish Cancel

DL_FUNCTIONS	DL_GROUP BY	DL_DISTINCT	DL_ORDER BY	DL_Geography
3150.000000	700.000000	111.000000	230.000000	92.732209
	1900.000000	222.000000	250.000000	96.000000
	250.000000	333.000000	300.000000	96.048082
	300.000000	444.000000	700.000000	98.260693
		555.000000	1000.000000	99.578803
			1500.000000	
			2100.000000	
			2600.000000	
			3000.000000	

Use Case 26: Selection by Dates with 'BETWEEN'

Situation: 'BETWEEN' can be used in a Date variable but requires a special format to use.

SQL Statement: BETWEEN #date1# AND #date2#

Example: DATE BETWEEN #1905/7/1# AND #1905/7/5#

Example Data File: Sample Data 9.xls and Sample Data 10.csv

	A	B	C	D
1	Normal (Multi)	Uniform	Binomial	DATE
2	87.53	45.29	6	7/1/1905
3	abc	45.29	6	7/2/1905
4	99.66	46.94	6	7/3/1905
5	108.75	45.96	6	7/4/1905
6	108.75	##45.96	6	7/5/1905

Data Link

New Variable Name: DL_ Select the data field(s) here for the input parameter chosen in the main dialog.

Open existing database tables:

Available Fields (variables):

Selected Fields:

If it's a matrix, we can select more than one field.

Condition: 'Condition' refers to the WHERE clause in SQL sentence

Use Row Filter Include Exclude From Line to

Result:

Result

Selecting Dates in CSV

Data Extract :
87.530000;
1. #QNAN0;
99.660000;
108.750000;
108.750000;