Technical Guide for Simetar

By

James W. Richardson

Keith Schumann

Peter Feldman

College Station, Texas

January 2004

APPENDIX A TECHNICAL GUIDE TO SIMETAR

The Simetar Add-in to Excel contains more than 100 functions written in VBA to perform mathematical and statistical operations to data. The purpose of this Technical Appendix is to document most all of these functions.

Each Simetar function performs either a mathematical manipulation, a statistical test, or calculates a statistic used for data analysis, parameter estimation, simulation, or analysis of simulation results. Each function is documented in a standard format which is

- Brief description of the function
- The mathematical equation behind the function
- The name of the Simetar function and its parameter requirements
- The reference for the mathematical equation

This format of providing references for each of the functions is used throughout the Technical Appendix, even for simple statistics like mean and standard deviation.

Where possible Simetar uses existing Excel functions (e.g., mean, standard deviation, minimum, etc.), however, Simetar packages these functions into more complete tools and for that reason they are included here. In some cases we developed functions to replace Excel functions because Excel limited the size of the problems that can be handled, e.g., Excel's multiple regression would only handle 17 explanatory variables, far too few for a complex VAR model.

SIMULATION

The simulation features in Simetar enable the user to perform stochastic simulations in Excel. Stochastic simulations use the interactions of random variables in a system to analyze the uncertainty in that system and its performance under alternative situations. Simetar functions allow the user to define and estimate distributions for random variables and to randomly sample those distributions so probabilistic outcomes for the system can be modeled. Simetar also provides useful tools to analyze the probabilistic outcomes generated from a stochastic simulation.

Pseudo Random Number Generation

Stochastic simulations rely on the generation of random numbers. Computers cannot generate truly random numbers but can generate so called "pseudo" random numbers. A pseudo random number sequence is one that cannot be differentiated from a truly random number sequence using statistical tests and thus are referred to as random numbers. There are several algorithms used by computer programs to generate random numbers. All of these algorithms require an initial random number seed to generate a sequence of random numbers. One of the most common algorithms used is the congruential method. The steps of the congruential method are as follows

$$(1) \qquad R = \frac{S_i}{L}$$

$$(2) \qquad S_{i+1} = \left(aS_i + b\right) \bmod L$$

where,

R = U(0,1) pseudo random number
S = User defined random number seed
L = The largest possible integer that a computer can store
a = A fixed integer
b = A fixed integer

The congruential method requires an initial random number seed (S_1), which can be specified by the user in most computer programs, to generate the first random number (R_1), from equation (1). To generate the next random number in the sequence, a new seed is generated using equation (2). In equation (2) S_{i+1} is the new seed generated using the constants a, b, and L and the previous seed S_i. Each S_{i+1} from equation (2) is subsequently entered into equation (1) to generate a sequence of random numbers. Simetar uses the internal random number generator in VBA, which is a modified congruential algorithm.

Reference:

Feldman, Richard M. and Ciriaco Valdez-Flores (1996). *Applied Probability and Stochastic Processes*. Boston, Massachusetts PWS Publishing Company, pg 79-81.

Inverse Transform Method

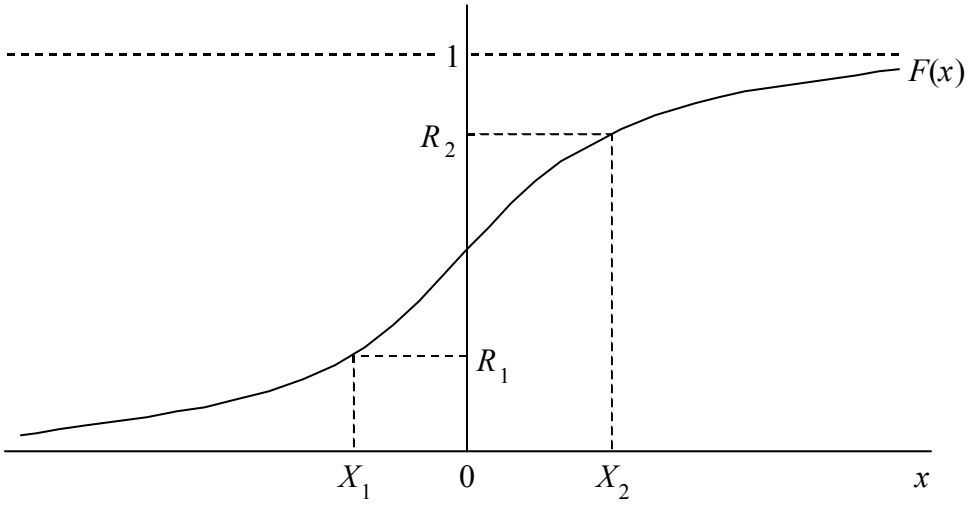
Random number algorithms generate uniform random numbers or U(0,1), however, we are not limited to using variables defined by uniform distributions. To generate a sample of random variates from a continuous distribution F, the inverse-transform method is used on the sample of U(0,1) numbers generated by the random number algorithm. The modified algorithm for generating random numbers from continuous distributions using the inverse-transform method returns a random variate X from $F(x)$ or

$$(3) \qquad X = F^{-1}\left(R\right)$$

where,

R = U(0,1) pseudo random number from equation (1)
$P\left(X \le x\right) = F\left(x\right)$ for any real number x

Intuitively the inverse transform method can be understood from its graphical representation with a CDF for a continuous distribution $F(x)$ and U(0,1) random numbers R_1 and R_2 used to generate variates X_1 and X_2



Simetar uses internal commands in VBA to perform the inverse-transform method on the U(0,1) random numbers generated by VBA to generate random numbers for numerous probability distributions.

Reference:

Law, Averill M. and W. David Kelton (2000). *Simulation Modeling and Analysis*. New York McGraw-Hill, Incorporated, pg 440-448.

Latin Hypercube Sampling

The most prevalent sampling technique for simulation modeling is Monte Carlo sampling. Monte Carlo sampling is the direct application of deviates generated from a random number generator and the inverse transform method to sample from a given probability distribution. When using Monte Carlo sampling it is necessary to use a large number of iterations to accurately recreate the desired probability distribution, or the problem of clustering may occur.

One way to avoid clustering is to use Latin Hypercube sampling. The Latin Hypercube sampling technique divides the cumulative density function of the given probability distribution into N equal intervals on the probability scale, where N is the number of iterations to simulate. The Latin Hypercube procedure then proceeds to randomly draw one value from each of the N intervals. This stratified sampling of the cumulative density function recreates the probability distribution accurately with fewer iterations than is required when using Monte Carlo sampling. Because of the gains in efficiency the simulation engine in Simetar uses Latin Hypercube sampling.

Reference:

McKay, M.D., W.J. Conover and R.J. Beckman (1979). "A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code." *Technometrics*. 21, 239-245.

Probability Distributions

Simetar has the capability of simulating several different types of probability distributions using both Simetar and Excel functions for distributions. The following is a brief description of the distribution functions available

Bernoulli(*p*)

Range	$\{0,1\}$
Mean	p
Variance	$p(p\text{-}1)$
Mass	$f(X=x\|p)=p^x(1-p)^{1-x};\quad x=0,1;\quad 0\le p\le 1$
Simetar	$=\text{BERNOULLI}(Conditional\_Probability, RandNumber)$

Beta(α_1,α_2)

Range	$[0,1]$
Mean	$\dfrac{\alpha_1}{\alpha_1+\alpha_2}$
Variance	$\dfrac{\alpha_1\alpha_2}{(\alpha_1+\alpha_2)^2(\alpha_1+\alpha_2+1)}$
Density	$f(x\|\alpha_1,\alpha_2)=\dfrac{1}{\text{B}(\alpha_1,\alpha_2)}x^{\alpha_1-1}(1-x)^{\alpha_2-1};\quad 0\le x\le 1;\quad \alpha_1,\alpha_2>0$
Simetar	$=\text{BETADIST}(\text{X},\text{Alpha},\text{Beta},A,B)$
	$=\text{BETAINV}(\text{Probability},\text{Alpha},\text{Beta},A,B)$

Binomial(*t*,*p*)

Range $\{0,1,\cdots,t\}$

Mean tp

Variance $tp(1-p)$

Mass $f(X=x|t,p)=\binom{t}{x}p^x(1-p)^{t-x};\quad x\in 0,1,\cdots,t$

Simetar $=\text{BINOMDIST}(\text{Number\_s},\text{Trials},\text{Probability\_s},\text{Cumulative})$

Empirical(*i*,*i*+1,…,*j*)

Range $\{i,i+1,\cdots,j\}$

Mean $\frac{\text{i}+\text{j}}{2}$

Variance $\frac{(\text{j}-\text{i}+1)^2-1}{12}$

Mass $f(X=x|i,i+1,\cdots,j)=\frac{1}{j-i+1};\quad x\in \text{i},\text{i}+1,\cdots,\text{j}$

Simetar $=\text{EMPIRICAL}(\text{Values},\textit{Probabilities},\textit{RandNumber})$

Exponential(β)

Range $[0,\infty]$

Mean β

Variance β^2

Density $f(x|\beta)=\frac{1}{\beta}e^{-x/\beta};\quad 0\le x\le\infty;\quad \beta>0$

Simetar $=\text{EXPONDIST}(\text{X},\text{Lamda},\text{Cumulative})$

Gamma(α,β)

Range $[0,\infty]$

Mean $\alpha\beta$

Variance $\alpha\beta^2$

Density $f(x|\alpha,\beta)=\dfrac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{-x/\beta};\quad 0\le x\le\infty;\quad \alpha,\beta>0$

Simetar = GAMMADIST(X, Alpha, Beta, Cumulative)

= GAMMAINV(Probability, Alpha, Beta)

Hypergeometric(*N*,*M*,*K*)

Range $\{\max(0, \mathrm{N}+\mathrm{M}-\mathrm{K}), \min[N,K]\}$

Mean $\dfrac{KM}{N}$

Variance $\dfrac{KM}{N}\dfrac{(N-M)(N-K)}{N(N-1)}$

Mass $f(X=x|N,M,K)=\dfrac{\binom{M}{x}\binom{N-M}{K-x}}{\binom{N}{K}};\quad x\in 0,1,\cdots,K;$

$M-(N-K)\le x\le M;\quad N,M,K\ge 0$

Simetar

= HYPGEOMDIST(Sample_s, Number_sample, Population_s, Number_pop)

Lognormal(μ,σ^2)

Range $[0,\infty]$

Mean $e^{\mu+\sigma^2/2}$

Variance $e^{2\mu+\sigma^2}\left(e^{\sigma^2}-1\right)$

Density $f(x|\mu,\sigma^2)=\dfrac{1}{x\sqrt{2\pi\sigma^2}}e\dfrac{-(\ln x-\mu)^2}{2\sigma^2};\quad 0\le x\le\infty;$

$-\infty\le\mu\le\infty;\quad \sigma^2>0$

Simetar = LOGINV(Probability, Mean, Standard_dev)

= LOGNORMDIST(X, Mean, Standard_dev)

Normal(μ,σ^2)

Range	$[-\infty,\infty]$
Mean	μ
Variance	σ^2
Density	$f(x\|\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)};\quad -\infty\le x,\mu\le\infty;\quad \sigma^2>0$
Simetar	$=\text{NORM}(\text{Mean},\text{StanDev},\textit{RandNumber})$
	$=\text{NORMDIST}(\text{X},\text{Mean},\text{Standard\_ dev},\text{Cumulative})$
	$=\text{NORMINV}(\text{Probability},\text{Mean},\text{Standard\_ dev})$
	$=\text{NORMSDIST}(\text{Z})$
	$=\text{NORMSINV}(\text{Probability})$

Poisson(λ)

Range	$\{0,1,\cdots\}$
Mean	λ
Variance	λ
Mass	$f(X=x\|\lambda)=\frac{e^{-\lambda}\lambda^x}{x!};\quad x\in 0,1,\cdots;\quad 0\le\lambda<\infty$
Simetar	$=\text{POISSON}(\text{X},\text{Mean},\text{Cumulative})$

Uniform(a,b)

Range	$[a,b]$
Mean	$\frac{a+b}{2}$
Variance	$\frac{(b-a)^2}{12}$
Density	$f(x\|a,b)=\frac{1}{b-a};\quad a\le x\le b$
Simetar	$=\text{UNIFORM}(\textit{Lower\_Value},\textit{Upper\_Value})$

Weibull(α,β)

Range $[0,\infty]$

Mean $\frac{\beta^2}{\alpha}\Gamma\left(\frac{1}{\alpha}\right)$

Variance $\frac{\beta^2}{\alpha}\left\{2\Gamma\left(\frac{2}{\alpha}\right)-\frac{1}{\alpha}\left[\Gamma\left(\frac{1}{\alpha}\right)\right]^2\right\}$

Density $f(x|\alpha,\beta)=\alpha\beta^{-\alpha}x^{\alpha-1}e^{-(x/\beta)^{\alpha}};\quad 0\le x\le\infty;\quad \alpha,\beta>0$

Simetar = WEIBULL(X, Alpha, Beta, Cumulative)

Reference:

Law, Averill M. and W. David Kelton (2000). *Simulation Modeling and Analysis*. New York McGraw-Hill, Incorporated, pg 299-326

Correlating Random Deviates

Computer based random number algorithms generate independent series of pseudo random numbers, however, independent deviates are not suitable for modeling random variables with multivariate probability distributions. An effective method of modeling the relationships between random variables in a simulation model is the correlation of computer generated random numbers. The correlation coefficients from correlated random deviates are not statistically different than the correlation coefficients from the multivariate distributions used to generate the deviates.

In Simetar, random deviates are correlated using the Choleski decomposition of the specified correlation matrix. The Choleski decomposition is an algorithm for the square root method of factoring a positive definite matrix S_{nxn} into an upper triangular matrix T_{nxn} such the $S=TT'$. To correlate random deviates a factored correlation matrix T is multiplied with an nx1 column vector of independent standard normal deviates yielding an nx1 column vector of correlated standard normal deviates. A mathematical description of this procedure is as follows

$$CSND_{nx1} = T_{nxn} \cdot ISND_{nx1}$$

$$\begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & t_{22} & \cdots & t_{1n} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & t_{nn} \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ \vdots \\ i_n \end{bmatrix}$$

where,

CSND = an *nx*1 column vector of correlated standard normal deviates distributed N(0,1)

T = an *nxn* factored correlation matrix

ISND = an *nx*1 column vector of independent standard normal deviates

The row number of each of the correlated standard normal deviates corresponds to the row number of the correlation matrix and must be applied to the random variable associated with that row in the correlation matrix. The correlated standard normal deviates can be converted to uniform deviates and used to simulate any distribution by applying the inverse-transform method.

Simetar has functions that allow the user to correlate random numbers and return either correlated standard normal deviates or correlated uniform standard deviates. The Simetar functions for correlating random deviates are as follows

=CSND(Matrix_Range,*NormalDeviate_Range*,*MatrixRow*)

=CUSD(Matrix_Range,*NormalDeviate_Range*,*MatrixRow*)

Nesting the CUSD function inside any of the distribution functions previously specified will return correlated values for the given distribution.

Reference:

Richardson, James W., Steven L. Klose and Allan W. Gray (2000). "An Applied Procedure for Estimating and Simulating Multivariate Empirical (MVE) Probability Distributions In Farm-Level Risk Assessment and Policy Analysis." *Journal of Agricultural and Applied Economics*. 32, 299-315.

REGRESSION ANALYSIS

Ordinary least squares (OLS) is a technique commonly used in quantitative modeling for analysis and prediction. The general specification of the OLS model is as follows

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \cdots + \beta_k X_{kt} + \varepsilon_t$$

where,

Y_t = observation t of the dependent variable
X_{kt} = observation t of the *kth* independent or explanatory variable
β_t = parameter estimated for the *kth* explanatory variable X_{kt}
ε_t = error or difference between the observed value and the predicted value for observation t of the dependent variable Y_t

It is also useful to present the matrix formulation for an OLS model

$$Y = X\beta + \varepsilon$$

where,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_t \end{bmatrix} \qquad X = \begin{bmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & & \vdots \\ 1 & X_{1t} & \cdots & X_{kt} \end{bmatrix} \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \qquad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \\ \varepsilon_t \end{bmatrix}$$

The estimated OLS equation is represented as,

$$y = Xb + e$$

which defines a vector of residuals,

$$e = y - Xb$$

The least squares principle in estimating β, is to select b such that the residual sum of squares, $e'e$, is minimized. The application of the least squares principle in a quantitative model is called regression analysis

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, pg 69-72.

Adjusted R-Squared

The Adjusted R-Squared is similar to the R-Squared, however, the Adjusted R-Squared takes into account the number of independent variables in the regression. The Adjusted R-Squared is useful when comparing the fit of two equations with the same dependent variable but a different number of explanatory variables. The Adjusted R-Squared test statistic is calculated as follows

$$\overline{R}^2 = 1 - \frac{RSS/(n-k+1)}{TSS/(n-1)}$$

where,

$$k+1 = \text{total number of parameters}$$

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, pg 74.

Akaike Information Criterion

The Akaike Information Criterion is used in the selection of regressors. A penalty for increasing the number of regressors is added to a transformation of the minimum residual sum of squares. The Akaike Information Criterion is calculated as follows

$$AIC = \ln\frac{RSS}{n} + \frac{2(k+1)}{n}$$

In Simetar the Akaike Information Criterion is calculated using the function

=AIC(Depend_values,Indep_values,Constant,Regression_type,*Restriction_vector*).

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, pg 74.

Box-Cox Transformation

In order for the underlying assumptions of many statistical procedures to be valid the data must be normally distributed. The Box-Cox transformation is a common power transformation used to normalize data and is defined as

$$y^{\lambda} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(y) & \text{if } \lambda = 0 \end{cases}$$

The determination of λ is made using a maximum likelihood estimation. For the transformed data the log-likelihood function is given by

$$\ell(\mu,\sigma^2,\lambda) = (\lambda-1)\sum_{i=1}^{n}\ln(|y_i|) - \frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(y_i^{(\lambda)}-\mu\right)$$

Maximizing over (μ,σ^2) for a fixed value of λ provides the equation

$$L_{MAX}(\lambda) = (\lambda-1)\sum_{i=1}^{n}\ln(|y_i|) - \frac{n}{2}\ln(2\pi\hat{\sigma}^2(\lambda)) - \frac{n}{2}$$

where,

$$\hat{\sigma}^2(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(\lambda)} - \bar{y}_i^{(\lambda)}\right)^2$$

Using these equations, λ is estimated by iteratively approximating the λ that maximizes the equation

$$L^*_{MAX}(\lambda) = (\lambda - 1)\sum_{i=1}^{n} \ln(|y_i|) - \frac{n}{2}\ln(\hat{\sigma}^2(\lambda))$$

In Simetar the Box-Cox Transformation is calculated with the function

=BOXCOX(DataRange,PowerValue,ShiftToPlus)

Reference:

Box, G. E. P. and D. R. Cox (1964). "An Analysis of Transformations." *Journal of the Royal Statistical Society. Series B (Methodological)*. 26, 211-252.

Covariance Matrix for Estimated Coefficients

The Covariance Matrix for Estimated Coefficients returns a matrix with the variance of the coefficients on the diagonal and the covariance of coefficients as the off diagonal elements. The covariance of the coefficients is a measure of the linear relationship between two coefficients. The Covariance Matrix for Estimated Coefficients can be written as follows

$$\text{cov}(b) = \begin{bmatrix} \text{var}(b_0) & \text{cov}(b_0, b_1) & \cdots & \text{cov}(b_1, b_k) \\ \text{cov}(b_1, b_0) & \text{var}(b_1) & \cdots & \text{cov}(b_1, b_k) \\ \vdots & \vdots & & \vdots \\ \text{cov}(b_k, b_0) & \text{cov}(b_k, b_1) & \cdots & \text{var}(b_k) \end{bmatrix}$$

and is calculated as

$$\text{cov}(b) = \sigma^2 (X'X)^{-1}$$

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 224-226.

Coefficient of Variation for the Regression

The coefficient of variation for the regression is a measure of the average error relative to the actual mean of the dependent variable. The coefficient of variation for the regression is calculated as follows

$$cv(reg) = \frac{s}{\bar{Y}} \times 100$$

where,

s = standard error of the regression

Confidence Interval

The dependant variable in an ordinary least squares estimation is distributed as

$$y \sim N\left(Y, \sigma_y\right)$$

Using the distribution of y a confidence interval can be estimated giving a lower and upper bound on y for which there is 100(1-α)% confidence that the true value Y lies in the estimated interval. The confidence interval for y is estimated as

$$CI_{U,L} = y \pm t_{n-k,\alpha/2} s_y$$

where,

CI_U = upper bound of the confidence interval

CI_L = lower bound of the confidence interval

s_y = standard error of the estimate, $\hat{\sigma}_y$

Reference:

Bowerman, Bruce L. and Richard T. O'Connell (1993). *Forecasting and Time Series an Applied Approach*. Pacific Grove, California Duxbury, pg 166-171.

Durbin-Watson Test Statistic

The Durbin-Watson test statistic is a measure of first-order autocorrelation in the model. The Durbin-Watson test statistic is calculated as follows

$$DW = \frac{\sum_{t=2}^{n} \left(e_t - e_{t-1}\right)^2}{\sum_{t=1}^{n} e_t^2}$$

where,

$DW \cong 2$ indicates error terms are not autocorrelated

$DW < 2$ indicates positively autocorrelated error terms

$DW > 2$ indicates negatively autocorrelated error terms

In Simetar the Durbin-Watson test statistic is calculated using the function

=DW(Depend_Values,Indep_Values,Constant,Regression_Type,*Restriction_Vector*).

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, pg 179-182.

Elasticity at the Mean for Estimated Coefficients

The elasticity at the mean for the estimated coefficient is a measure of the percent change in the dependent variable with respect to a percent change in a given explanatory variable. The elasticity at the mean for the estimated coefficient is calculated as

$$\eta_{x,y} = \beta_i \frac{\bar{x}_i}{\bar{y}}$$

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 172-175.

F-Test

The F-test for a multiple regression tests joint hypotheses about the elements of the column vector of true parameters β. The F-test is calculated

$$F = \frac{(b_s - \beta_s)'\left[\text{cov}(b_s)\right]^{-1}(b_s - \beta_s)}{K}$$

where,

b_s = Column vector of estimated coefficients excluding b_0
β_s = Column vector of hypothesized values for the true parameters excluding β_0
K = Number of elements in the vectors b_s and β_s

and F is distributed $F_{\alpha,[(K),(T-K+1)]}$. In Simetar the null and alternative hypotheses are

$$H_0: \beta_s = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \text{and} \qquad H_1: \beta_s \neq 0$$

The null hypothesis is rejected at the 1-α level if the test statistic F is great than the critical value $F_{[(K),(T-K+1)]}$.

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 365-368.

Goldfeld-Quandt Test

The Goldfeld-Quandt Test is used to test for heteroskedacity and is calculated as follows

$$GQ = \frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2}$$

where,

$\hat{\sigma}_1^2$ = the variance of the first partition of data

$\hat{\sigma}_2^2$ = the variance of the second partion of data

In Simetar the partitions for the Goldfeld-Quandt test are assumed to occur at the midpoint of the data with half of the observations in the first partition and half of the observations in the second partition. The null hypothesis is that the variances of each partition are equal. Reject H_0 if GQ is greater than the 1-α quantile from the F distribution with $(T_1 - K_1),(T_2 - K_2)$ degrees of freedom, where T is the number of observations in the given partition and K is the number of coefficients estimated for each partition.

In Simetar the Goldfeld-Quandt test statistic is calculated using the function

=GQ(Depend_Values,Indep_Values,Constant,Regression_Type,*Restriction_Vector*).

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 494-495.

Mean Error Measures of Forecasting

The Mean Absolute Percent Error is a measure of a model's forecasting ability. The formulas used for calculating these statistics are as follows

$$MAPE = 100\% \times \frac{\sum_{t=1}^{N} \left| E_t / Y_t \right|}{N}$$

where,

E_t = the forecast error for time period t

In general it is desirable to minimize the Mean Absolute Percent Error when selecting the best forecasting model.

In Simetar the Mean Absolute Percent Error is calculated using the function

=MAPE(Residuals,Y_Values).

Reference:

Albright, S. Christian, Wayne L. Winston, and Christopher J. Zappe (2000). *Managerial Statistics*. Pacific Grove, California Duxbury, pg 842-843.

Observational Diagnostics

Simetar includes a suite of tools that perform observational Diagnostics for multiple regression models. Observational Diagnostics are used to assess the quality and reliability of the data used to estimate the model. In Simetar Observational Diagnostics are calculated using the function

=DFBETA(Depend_Values,Indep_Values,Constant,*Restriction_Vector*,*Obs_RestVector*)

Covariance Ratio

The Covariance Ratio shows the sensitivity of the covariance matrix to the deletion of rows from the *X* and *Y* matrices in the regression model. The ratio compares the determinant of covariance matrix for the row deleted model with the determinant of covariance matrix for full model and is calculation as follows

$$COVRatio = \frac{\left| s^2(i)\left[X^T(i)X(i) \right]^{-1} \right|}{\left| s^2\left(X^T X \right)^{-1} \right|}$$

where,

$s^2(i)$ = estimated variance with the *i*th row deleted

X = matrix of independent variables for full model

$X(i)$ = matrix of independent variables for model with the *i*th row deleted

A *COVRatio* near one indicates little to no change in the covariance matrix from the full model to the row deleted model. If the absolute value of the *COVRatio* is not within the bounds of the calculated critical values the *ith* row is considered to be significant. The critical value is calculated as

$$COVRatio_{cr} = 1 \pm \frac{3k}{n-i}$$

Belsley, David A., Edwin Kuh, and Roy E. Welsch (1980). *Regression Diagnostics*. John Wiley and Sons, Incorporated, New York, pg 22-24.

DFBetas

DFBetas are calculated to determine the influence of a single observation on the strength of the overall model. DFBetas measure the effect of removing a row of data from the calculation of the regression. DFBetas are calculated as

$$DFBetas_{ij} = \frac{b_j - b_j(i)}{s(i)\sqrt{\left(X^T X\right)^{-1}_{jj}}}$$

where,

b_j = the *j*th component of the column vector *b* of estimated coefficients for β

$b_j(i)$ = the estimated coefficient b_j when the *i*th row of *Y* and *X* are removed

$s(i)$ = the estimated variance when the *i*th row of *Y* and *X* are removed

If the absolute value of the $DFBeta_{ij}$ is larger than the calculated critical value the *jth* observation is considered to be significant. The critical value is calculated as

$$DFBeta_{cr} = k/\sqrt{n-i}$$

where,

k = number of estimated coeficients

n = number of observations

i = number of rows deleted

Belsley, David A., Edwin Kuh, and Roy E. Welsch (1980). *Regression Diagnostics*. John Wiley and Sons, Incorporated, New York, pg 11-14.

DFFit(s)

The DFFit(s) are a measure of the fit of the regression equation when the ith row is deleted from the matrices of independent and dependent variables. DFFit uses the estimated standard error from the full model and is calculated as

$$DFFit = \frac{h_i e_i}{1 - h_i}$$

where,

$$e_i = \text{residuals}$$
$$h_i = \text{leverage}$$

An alternative calculation for measuring the fit is the *DFFits*, which uses the estimation of the standard error from the row deleted model and is calculated as

$$DFFits = \left[\frac{h_i}{1 - h_i}\right]^{1/2} \frac{e_i}{s(i)\sqrt{1 - h_i}}$$

If the absolute value of the *DFFit* is larger than the calculated critical value the *ith* row is considered to be significant. The critical value is calculated as

$$DFFit_{cr} = 2\sqrt{\frac{k}{n - i}}$$

Belsley, David A., Edwin Kuh, and Roy E. Welsch (1980). *Regression Diagnostics*. John Wiley and Sons, Incorporated, New York, pg 14-16.

Leverage

The least squares projection matrix or the hat matrix determines the predicted values of a regression model. The diagonal elements of the hat matrix or Leverage can be used to measure the effect that the individual observations of the dependant variable have on the corresponding estimation of that observation. The hat matrix is calculated as

$$H = X\left(X^T X\right)^{-1} X^T$$

which has the following relationship to the estimation of the dependant variable

$$\hat{y} \equiv Xb = Hy$$

Leverage measures the influence of the actual value of a data point for the dependent variable on the predicted value. If the absolute value of the Leverage is larger than the calculated critical value the *ith* observation is considered to be significant. The critical value for determining Leverage is calculated as

$$2k/(n-i)$$

Belsley, David A., Edwin Kuh, and Roy E. Welsch (1980). *Regression Diagnostics*. John Wiley and Sons, Incorporated, New York, pg 16-18.

Studentized Residuals

In order to better detect problematic data it is common to standardize residuals by dividing the residuals by their estimated standard error. The estimated variance of the residuals is commonly calculated as

$$\text{var}(\hat{e}_i) = s^2(1-h_i)$$

where,

$\hat{e}_i$ = estimated residuals
s^2 = estimated variance of the model
h_i = leverage

In an observational diagnostics modeling environment the estimated variance can be calculated as

$$s^2(i) = \frac{1}{n-k-1}\sum_{j\neq i}\left[y_j - x_j b(i)\right]^2$$

where,

$s^2(i)$ = estimated variance with the *i*th row deleted

y = dependant variable

x = independant variable

n = number of observations

k = number of parameters estimated

Using this new calculation for variance standardized residuals are referred to as Studentized Residuals and are calculated as

$$e_i^* = \frac{e_i}{s(i)\sqrt{1-h_i}}$$

In many modeling situations the Studentized Residuals are distributed as a *t* distribution. Using a large sample approximation the critical value for the Studentized Residuals can be approximated as two. If the absolute value of the Studentized Residual is larger than two value the *ith* observation is considered to be significant.

Reference:

Belsley, David A., Edwin Kuh, and Roy E. Welsch (1980). *Regression Diagnostics*. John Wiley and Sons, Incorporated, New York, pg 18-20.

Partial Correlation Coefficient

The Partial Correlation Coefficient is a measure of the linear relationship between two variables while holding a third variable constant. The Partial Correlation Coefficient is calculated as follows

$$\rho_{yx.t} = \frac{\rho_{yx} - \rho_{yt}\rho_{xt}}{\sqrt{(1-\rho_{xt}^2)(1-\rho_{yt}^2)}}$$

where,

ρ_{yx} = the correlation coefficient of y and x

ρ_{yt} = the correlation coefficient of y and t

ρ_{xt} = the correlation coefficient of x and t

In Simetar the Partial Correlation Coefficient is calculated using the function

=PARTIALCORREL(Depend_Values,Indep_Values,Indep_Index,Indep_Remove, Restriction_Vector).

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, pg 211-213.

Prediction Interval

The standard deviation σ_Y of the true parameter Y is estimated as the standard error of the estimate s_y, which can be decomposed algebraically into two components as

$$s_y = s\sqrt{h_t}$$

The point estimate of $\sigma_{(Y_t - y_t)}$ is called the standard error of the prediction error and is calculated as

$$s_{(Y_t - y_t)} = s\sqrt{1 + h_t}$$

Using the standard error of the prediction error a 100(1-α)% prediction interval can be estimated for the forecasted value of y in time t. The prediction interval is calculated as

$$PI_{U,L} = y_t \pm t_{n-k,\alpha/2} s\sqrt{1 + h_t}$$

where,

PI_U = upper bound of the prediction interval
PI_L = lower bound of the prediction interval

Reference:

Bowerman, Bruce L. and Richard T. O'Connell (1993). *Forecasting and Time Series an Applied Approach*. Pacific Grove, California Duxbury, pg 166-171.

R-Squared

The R-Squared test measures the proportion of the variance in the dependent variable Y attributable to the variance in the independent variables X. The R-Squared test statistic is calculated as follows

$$R^2 = 1 - \frac{e'e}{y'y - n\overline{Y}^2}$$

where,

$y'y - n\overline{Y}^2$ = Total Sum of Squares (TSS)

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, pg 74.

Rho

The most common procedure for modeling a system with autocorrelation is a first-order autoregressive process or an AR(1). In an AR(1) process the error in time *t* is lagged on the error in *t-1* which yields the equation

$$e_t = \rho e_{t-1} + \varepsilon_t$$

where,

e_t = Error term in time *t* from a regression model $y = Xb + e$

ρ = Parameter rho that determines the properties of e_t

ε_t = Independent disturbances for the AR(1) process

The parameter rho can be calculated from the regression equation $y=Xb+e$ as

$$\rho = \frac{e_t - \varepsilon_t}{e_{t-1}}$$

where,

$\rho \cong 0$ indicates error terms are not autocorrelated

$\rho < 0$ indicates negatively autocorrelated error terms

$\rho > 0$ indicates positively autocorrelated error terms

In Simetar the parameter rho is calculated using the function

=RHO(Depend_Values,Indep_Values,Constant,Regression_Type,*Restriction_Vector*).

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 536-538.

Schwarz Information Criterion

The Schwarz Criterion is used in the selection of lags for an AR(*p*) process. A penalty for increasing the number of lags is added to a transformation of the minimum residual sum of squares. The Schwarz Criterion is calculated as follows

$$SC = \ln\frac{RSS}{n} + \frac{k}{n}\ln n$$

In Simetar the Schwarz Criterion is calculated using the function

=ARSCHWARZ(Y_Values,Constant,NumDifferences).

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, pg 74.

Semi-Partial Correlation Coefficient

The Semi-Partial Correlation Coefficient is a measure of the linear relationship between two variables controlling for the effect that a third variable has on only one of the other variables. The Semi-Partial Correlation Coefficient is calculated as follows

$$\rho_{y(x.t)} = \frac{\rho_{yx} - \rho_{yt}\rho_{xt}}{\sqrt{\left(1-\rho_{xt}^2\right)}}$$

where,

ρ_{yx} = the correlation coefficient of y and x
ρ_{yt} = the correlation coefficient of y and t
ρ_{xt} = the correlation coefficient of x and t

In Simetar the parameter Semi-Partial Correlation Coefficient is calculated using the function

=SEMIPARTIALCORREL(Depend_Values,Indep_Values,Indep_Index,Indep_Remove, Restriction_Vector).

Reference:

Kerlinger, Fred N. and Elazar J. Pedhazur (1973). *Multiple Regression in Behavioral Research*. NewYork Holt, Rinehart and Winston, Incorporated, pg 92-93.

Standard Error of the Coefficient

The standard error option in Simetar's simple regression tool returns the standard error for the estimated coefficient b_1. The coefficients estimated with OLS are distributed

$$b \sim N\left(\beta, \sigma^2 / \sum x^2\right)$$

where,

$$\sqrt{\sigma^2 / \sum x^2} = \text{standard error of the coefficient}$$

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, 25-26

Standard Error of the Regression

The standard error of the regression is the square root of s^2, an unbiased estimate for disturbance variance, σ^2. The standard error of the regression is calculated as follows

$$s = \sqrt{\frac{e'e}{n-k+1}}$$

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated pg 89.

T-test

The t-test is a measure of the difference between the estimated parameter *b* and a hypothesized value of the true parameter β. The t-test is calculated

$$t = \frac{b_i - \beta_i}{\sqrt{\sigma^2 / \sum x_i^2}}$$

where *t* is distributed $t_\alpha(n\text{-}2)$. In Simetar the null and alternative hypotheses are

$$H_0: \beta_i = 0 \qquad \text{and} \qquad H_1: \beta_i \neq 0$$

The null hypothesis is rejected at the 1-α level if the test statistic t is great than the critical value $t_\alpha(n\text{-}2)$.

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 363-365.

CORRELATION

In quantitative modeling it is often important to summarize the relationship between variables. The correlation matrix tool in Simetar is a useful way of summarizing the relationship between two variables both parametrically and non-parametrically.

Covariance

The Covariance measures the direction and strength of the linear relationship between two variables. The Covariance is calculated as follows

$$\operatorname{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

The value of the Covariance is dependent on the units of the variables. The Covariance of a variable against itself is the variance.

The Covariance is calculated using the function

=COVAR(Array1,Array2)

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 42-46.

Hypothesis Test for Spearman's Rho

The hypothesis test for Spearman's Rho is a measure of the difference between the estimated rank correlation coefficient ρ and 0. Simetar uses the normal approximation of the hypothesis test for Spearman's Rho, which is calculated as follows

$$z_p = |\rho|\sqrt{n-1}$$

where z is distributed z_p. In Simetar the null and alternative hypotheses are

$$H_0: \beta_i = 0 \qquad \text{and} \qquad H_1: \beta_i \neq 0$$

The null hypothesis is rejected at the p level if the test statistic z is great than the critical value z_p.

Reference:

Conover, W. J. (1999). *Practical Nonparametric Statistics*. New York John Wiley and Sons, Incorporated, pg 314-319.

Pearson's Product Moment Correlation Coefficient

Pearson's Product Moment Correlation Coefficient measures the direction and strength of the linear relationship between two variables. Pearson's Product Moment Correlation Coefficient is calculated as follows

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

One of the advantages of Pearson's Product Moment Correlation Coefficient is that it is a unitless quantity that falls between –1 and +1.

Pearson's Product Moment Correlation Coefficient is in Simetar calculated using the function

=CORREL(Array1,Array2).

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 42-46.

Rank Correlation Coefficient (Spearman's Rho)

The Rank Correlation Coefficient given by Spearman (1904) is an alternative to Pearson's product moment correlation coefficient. Unlike Pearson's Product Moment Correlation Coefficient, Spearman's Rho does not depend on the bivariate distribution of *(X, Y)*. Spearman's Rho requires bivariate data that is of at least an ordinal scale. Data used for calculating Spearman's Rho is ranked and the test statistic is calculated as follows

$$\rho = \frac{\sum_{i=1}^{n} R(X_i)R(Y_i) - n\left(\frac{n+1}{2}\right)^2}{\sqrt{\left(\sum_{i=1}^{n} R(X_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)}\sqrt{\left(\sum_{i=1}^{n} R(Y_i)^2 - n\left(\frac{n+1}{2}\right)^2\right)}}$$

where,

$R(X_i)$ = the rank of the ith observation of the random variable X
$R(Y_i)$ = the rank of the ith observation of the random variable Y
n = the total number of observations

In Simetar Spearman's Rho is calculated using the function

=RANKCORREL(Array1,Array2).

Reference:

Conover, W. J. (1999). *Practical Nonparametric Statistics*. New York John Wiley and Sons, Incorporated, pg 314-319.

T-test for the Correlation Coefficient

The t-Test for the Correlation Coefficient is a measure of the difference between the estimated correlation coefficient ρ and 0. The t-test is calculated

$$t = \rho\sqrt{\frac{n-2}{1-\rho^2}}$$

where t is distributed $t_\alpha(n\text{-}2)$. The null and alternative hypotheses are

$$H_0: \beta_i = 0 \qquad \text{and} \qquad H_1: \beta_i \neq 0$$

The null hypothesis is rejected at the 1-α level if the test statistic t is greater than the critical value $t_\alpha(n\text{-}2)$.

Reference:

Vose, David (2000). *Risk Analysis*. New York John Wiley and Sons, Limited, pg 53-55.

MATRIX OPERATIONS

Deriving the solution to linear sets of equations is often required in quantitative modeling. A useful methodology in the solution of these types of problems is the arrangement of equations in matrices. Simetar contains several useful features to assist the user in manipulating matrices for the solution of linear systems.

Determinant of a Matrix

For any square matrix A there exists a scalar representation of A denoted det(A). The determinant is used for determining whether or not a matrix is singular. If det(A) = 0

then *A* is singular. A singular matrix contains two or more rows that are linearly dependent. A system of equations that are linearly dependent does not have a unique solution.

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 84-107.

Factoring a Matrix using the Choleski Decomposition

The Choleski Decomposition is an algorithm for the square root method of factoring a positive definite matrix, *S*, as an upper triangular matrix *T* such that *S* = *TT'*. The detailed description of the matrices is as follows

$$S = TT'$$

$$\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix} = \begin{bmatrix} t_{11} & 0 & \cdots & 0 \\ t_{12} & t_{22} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ t_{1p} & t_{2p} & \cdots & 0 \end{bmatrix} \begin{bmatrix} t_{11} & t_{12} & \cdots & t_{1p} \\ 0 & t_{22} & \cdots & t_{2p} \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & t_{pp} \end{bmatrix}$$

The Simetar function for factoring a positive definite matrix using the Choleski Decomposition is

=MSQRT(Matrix,StartRow,EndRow)

Reference:

Graybill, Franklin A. (1976). *Theory and Application of the Linear Model*. North Scituate, Massachusetts Duxbury Press, pg 231.

Inner Product

The inner product of any given vectors *x* and *y* on a vector space *V* assigns a real number to the vectors *x* and *y*. The inner product is calculated as

$$\langle x, y \rangle = x^T y$$

The inner product can also be calculated with a vector *w* of weights calculated as

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i w_i$$

In Simetar the inner product is calculated using the function

=MIP(Matrix1,Matrix2).

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 217.

Inverse of a Matrix

The multiplicative inverse of any real number a is denoted as a^{-1} where

$$aa^{-1} = 1$$

There also exists a multiplicative inverse of any nonsingular square matrix A such that

$$AA^{-1} = I$$

where,

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$ is called the identity matrix

The matrix A^{-1} is referred to as the inverse of A.

The Inverse of a matrix can be calculated using the function

=MINVERSE(Array)

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 48-49.

Kronecker Product

The Kronecker Product is a specialized form of matrix multiplication. For two matrices A_{mxn} and B_{pxq} the Kronecker Product $A \otimes B$ yields an $MP \times NQ$ matrix calculated as follows

$$A_{m\times n}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{nn} \end{bmatrix} \qquad B_{m\times n}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1q} \\ b_{21} & b_{22} & \cdots & b_{2q} \\ \vdots & \vdots & & \vdots \\ b_{p1} & b_{p2} & \cdots & b_{pq} \end{bmatrix}$$

$$A\otimes B=\begin{bmatrix} a_{11}b_{11} & a_{11}b_{12} & \cdots & a_{11}b_{1q} & a_{12}b_{11} & a_{12}b_{12} & \cdots & a_{12}b_{1q} & \cdots & a_{1n}b_{11} & a_{1n}b_{12} & \cdots & a_{1n}b_{1q} \\ a_{1}b_{21} & a_{11}b_{22} & \cdots & a_{11}b_{2q} & a_{12}b_{21} & a_{12}b_{22} & \cdots & a_{12}b_{2q} & \cdots & a_{1n}b_{21} & a_{1n}b_{22} & \cdots & a_{1n}b_{2q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ a_{11}b_{p1} & a_{11}b_{p2} & \cdots & a_{11}b_{pq} & a_{12}b_{p1} & a_{12}b_{p2} & \cdots & a_{12}b_{pq} & \cdots & a_{1n}b_{p1} & a_{1n}b_{p2} & \cdots & a_{1n}b_{pq} \\ a_{21}b_{11} & a_{21}b_{12} & \cdots & a_{21}b_{1q} & a_{22}b_{11} & a_{22}b_{12} & \cdots & a_{22}b_{1q} & \cdots & a_{2n}b_{11} & a_{2n}b_{12} & \cdots & a_{2n}b_{1q} \\ a_{21}b_{21} & a_{21}b_{22} & \cdots & a_{21}b_{2q} & a_{22}b_{21} & a_{22}b_{22} & \cdots & a_{22}b_{2q} & \cdots & a_{2n}b_{21} & a_{2n}b_{22} & \cdots & a_{2n}b_{2q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ a_{21}b_{p1} & a_{21}b_{p2} & \cdots & a_{21}b_{pq} & a_{22}b_{p1} & a_{22}b_{p2} & \cdots & a_{22}b_{pq} & \cdots & a_{2n}b_{p1} & a_{2n}b_{p2} & \cdots & a_{2n}b_{pq} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ a_{m1}b_{11} & a_{m1}b_{12} & \cdots & a_{m1}b_{1q} & a_{m2}b_{11} & a_{m2}b_{12} & \cdots & a_{m2}b_{1q} & \cdots & a_{mn}b_{11} & a_{mn}b_{12} & \cdots & a_{mn}b_{1q} \\ a_{21}b_{21} & a_{22}b_{22} & \cdots & a_{m1}b_{2q} & a_{m2}b_{21} & a_{m2}b_{22} & \cdots & a_{m2}b_{2q} & \cdots & a_{mn}b_{21} & a_{mn}b_{22} & \cdots & a_{mn}b_{2q} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots & & \vdots & \vdots & & \vdots \\ a_{m1}b_{p1} & a_{m1}b_{p2} & \cdots & a_{m1}b_{pq} & a_{m2}b_{p1} & a_{m2}b_{p2} & \cdots & a_{m2}b_{pq} & \cdots & a_{mn}b_{p1} & a_{mn}b_{p2} & \cdots & a_{mn}b_{pq} \end{bmatrix}$$

where,

$A\otimes B$ = Kronecker Product of matrix A and matrix B

In Simetar the inner product is calculated using the function

=MKRON(Matrix1,Matrix2).

Reference:

Bronson, Richard (1989). *Theory and Problems of Matrix Operations*. New York McGraw-Hill, Incorporated, pg 165.

Matrix Multiplication

A matrix A can be multiplied to a matrix B represented AB if the number of columns in A equals the number of rows in B. The matrix AB is calculated as

$$A_{n\times m}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \qquad B_{m\times n}=\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

$$AB_{n\times n}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}\begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & & \vdots \\ b_{m1} & b_{m2} & \cdots & b_{mn} \end{bmatrix}$$

$$=\begin{bmatrix} a_{11}b_{11}+a_{12}b_{21}+\cdots+a_{1m}b_{m1} & a_{11}b_{12}+a_{12}b_{22}+\cdots+a_{1m}b_{m2} & \cdots & a_{11}b_{1n}+a_{12}b_{2n}+\cdots+a_{1m}b_{mn} \\ a_{21}b_{11}+a_{22}b_{21}+\cdots+a_{2m}b_{m1} & a_{21}b_{12}+a_{22}b_{22}+\cdots+a_{2m}b_{m2} & \cdots & a_{21}b_{1n}+a_{22}b_{2n}+\cdots+a_{2m}b_{mn} \\ \vdots & \vdots & & \vdots \\ a_{n1}b_{n1}+a_{n2}b_{21}+\cdots+a_{nm}b_{m1} & a_{n1}b_{12}+a_{n2}b_{22}+\cdots+a_{nm}b_{m2} & \cdots & a_{n1}b_{1n}+a_{n2}b_{2n}+\cdots+a_{nm}b_{mn} \end{bmatrix}$$

Matrices can be multiplied in Simetar using the function

=MMULT(Array1,Array2)

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 38-40.

Matrix Orthogonalization

Simetar uses the Gram-Schmidt Process to calculate an orthonormal basis Q for a given square matrix A. The properties of an orthonormal basis are

$$X^T X = I$$
$$X^T = X^{-1}$$
$$\langle Qx, Qy \rangle = \langle x, y \rangle$$
$$\|Qx\|_2 = \|x\|_2$$

In Simetar the orthonormal basis is calculated using the function

=MORTH(Matrix1).

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 241-243.

Norm of a Vector

For any vector v the norm of v is given as

$$\|v\| = \sqrt{\langle v, v \rangle}$$

where,

$\|v\|$ = norm of v

$\langle v, v \rangle$ = innerproduct of the vector v

In Simetar the Norm of a Vector is calculated using the function

=MNORM(Matrix1).

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 222-224.

Rank of a Matrix

The rank of any given matrix A is the dimension of the row space of A. The dimension of the row space of A is the number of rows in the row echelon form of A that have nonzero entries. Determining the rank of a matrix indicates whether or not the row vectors in the matrix are linearly dependent.

In Simetar the Rank of a Matrix is calculated using the function

=MRANK(Matrix1)

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 154.

Row Echelon Form Matrix

One method for solving a linear system of equations is to arrange the system as a matrix and reduce the matrix to row echelon form. Three conditions must hold for a matrix to be in row echelon form:

1. The first nonzero entry in every row is a 1.

2. If the *kth* row does not consist entirely of zeros, then the number of leading zeros in row *k* must be greater than the number of zeros in row *k*-1.
3. Any row that consists entirely of zero must be below all rows with nonzero entries.

In Simetar the row echelon form of a matrix is calculated using the function

=MREDUCE(Matrix1).

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 154.

Trace of a Square Matrix

The Trace of a Square Matrix A is the sum of the diagonal elements in A and denoted as $tr(A)$.

$$tr(A) = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} \quad \text{for all } i = j$$

where,

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

$i = ith$ column of the matrix A

$j = jth$ row of the matrix A

In Simetar the Trace of a Square Matrix is calculated using the function

=MTRACE(SquareMatrix).

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 288.

Transpose of a Matrix

The transpose of any n x m matrix A is an m x n matrix A^T represented as

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

$$A^T = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{n1} \\ a_{12} & a_{22} & \cdots & a_{n2} \\ \vdots & \vdots & & \vdots \\ a_{1m} & a_{2m} & \cdots & a_{nm} \end{bmatrix}$$

The Transpose of a Matrix can be calculated using the function

=TRANSPOSE(Array)

Reference:

Leon, Steven J. (1998). *Linear Algebra with Applications*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 50.

STOCHASTIC DOMINANCE AND CERTAINTY EQUIVALENCE

Stochastic Dominance is a methodology based on expected utility maximization that has been developed to analyze risky alternatives. Risky alternatives are evaluated with Stochastic Dominance using the weakest possible assumption about the decision maker's expected utility function. However, as the degree of Stochastic Dominance increases the assumption on risk preference become stronger.

Certainty Equivalence

Certainty Equivalence is the minimum amount of money a decision maker would require as a lump sum payment to forgo a risky alternative, thus the decision maker is indifferent between the certainty equivalent and the future payoff of the risky alternative. The value of the certainty for any given risky alternative is dependent upon the expected utility function of the decision maker and the decision maker's level of risk aversion. Due to the difficulty in measuring a decision maker's expected utility function it is common to assume an exponential utility function. The formula for calculating Certainty Equivalence with an exponential utility function is

$$E(U) = \sum_i p_i \left(-e^{-RAC(X_i + \omega)} \right)$$

$$CE = \frac{\ln\left(E(U)\right)}{RAC} - \omega$$

where,

$$CE = \text{certainty equivalence}$$
$$E(U) = \text{expected utility}$$
$$RAC = \text{risk aversion coefficient}$$
$$\omega = \text{initial wealth}$$

Another commonly assumed form for a decision maker's expected utility function is the power utility function. The formula for calculating Certainty Equivalence with a power utility function is

$$E(U) = \sum_i p_i \left(X_i + \omega\right)^{(1-RAC)}$$
$$CE = E(U)^{(1/(1-RAC))} - \omega$$

where,

$$CE = \text{certainty equivalence}$$
$$E(U) = \text{expected utility}$$
$$RAC = \text{risk aversion coefficient}$$
$$\omega = \text{initial wealth}$$

For both assumptions of the form of the expected utility function the Certainty Equivalence can be analyzed at varying levels of risk aversion. The value of the Risk Aversion Coefficient can be interpreted as

$$RAC < 0 \text{ risk loving}$$
$$RAC = 0 \text{ risk indifferent}$$
$$RAC > 0 \text{ risk averse}$$

A stronger attitude toward risk is inferred as the absolute value of the RAC increases. In Simetar Certainty Equivalence is calculated using the function

=CERTEQ(DataList,Risk_Aversion_Coefficient,Utility_Function,*Beginning_Wealth*)

Reference:

Clemen, Robert T. (1991). *Making Hard Decisions*. Boston PWS-Kent Publishing Company, pg 371-375.

Confidence Premiums

Confidence Premiums are the lower and upper bounds of certainty equivalence on a given risky alternative. The upper and lower bounds of certainty equivalence are determined by using the upper and lower bounds on the risk aversion coefficient that a decision maker may have. Confidence premiums give you a range of values for which a decision maker would be indifferent between receiving that value as a lump sum and receiving the futures payoff of a risky alternative.

Reference:

Mjelde, James W. and Mark J. Cochran (1988). "Obtaining Lower and Upper Bounds on the Value of Seasonal Climate Forecasts as a Function of Risk Preferences." *Western Journal of Agricultural Economics*. 12, 285-293.

First, Second, and Third Degree Stochastic Dominance

The assumption made on the expected utility function for First Degree Stochastic Dominance is that the decision maker has a positive marginal utility. For two alternatives *A* and *B*, *A* is first degree stochastic dominant over *B* if

$$F_A(x) \leq F_B(x) \text{ for all } x \text{ with at least one strict inequality}$$

where,

$$F_A(x) = \text{cumulative density function of alternative } A$$
$$F_B(x) = \text{cumulative density function of alternative } B$$

The additional assumption made on the expected utility function for Second Degree Stochastic Dominance is that the decision maker is risk averse for all values of *x*, meaning the decision maker's expected utility function is positive but has a decreasing slope. For two alternatives *A* and *B*, *A* is first degree stochastic dominant over *B* if

$$\int_{-\infty}^{x^*} F_A(x)dx \leq \int_{-\infty}^{x^*} F_B(x)dx \text{ for all } x^* \text{ with at least one strict inequality}$$

Third Degree Stochastic Dominance makes the additional assumption that the risk aversion coefficient is decreasing with income or wealth.

Reference:

Hardaker, J. Brian, Ruud B.M. Huirne, and Jock R. Anderson (1997). *Coping With Risk in Agriculture*. New York CAB International, pg 146-149.

Risk Aversion Coefficient

One of the basic assumptions of expected utility theory is that given no risk more wealth is preferred to less wealth indicated by

$$U^{(1)}(W) > 0$$

where,

$$U^{(i)}(W) = ith \text{ derivate of the utility function for wealth}$$

When risk is introduced a decision maker's attitude toward risk can be represented by the second derivative of the utility function for wealth.

$$U^{(2)}(W) < 0 \text{ risk averse}$$
$$U^{(2)}(W) = 0 \text{ risk indifferent}$$
$$U^{(2)}(W) > 0 \text{ risk loving}$$

The measure of a decision maker's risk aversion coefficient is called the risk aversion coefficient and defined as

$$RAC = -\frac{U^{(2)}(W)}{U^{(1)}(W)}$$

Reference:

Hardaker, J. Brian, Ruud B.M. Huirne, and Jock R. Anderson (1997). *Coping With Risk in Agriculture*. New York CAB International, pg 96-99.

COMPARATIVE TESTS FOR DATA

Quantitative modeling often involves the statistical analysis of data. Simetar includes comparative tests for data to evaluate the statistical properties of a given data series with another data series or a hypothesized statistical assumption.

Anderson-Darling Test for Normality

The Anderson-Darling Test for Normality is based on the more general Anderson-Darling goodness of fit test. The Anderson-Darling Test measures the weighted distance between the empirical distribution of a function and the distribution of the hypothesized function. The Anderson-Darling Test statistic is defined as

$$A_n^2 = \left(-\left\{ \sum_{i=1}^{n} (2i-1)\left[\ln \hat{F}(X_i) + \ln\left(1 - \hat{F}(X_{n+1-i})\right)\right]\right\} \Big/ n \right) - n$$

where,

$\hat{F}(x)$ = cumulative density function for the hypothesized distribution

X_i = ith order statistic of the empirical distribution

n = total number of observations

In Simetar the H_0 for the Anderson-Darling test is that the data are normally distributed. Reject H_0 at a level of significance α if it the test statistic A_n is greater than the 1-α quantile of the Anderson-Darling tables, which can be found in a statistics textbook.

In Simetar the Anderson-Darling Test for Normality is calculated using the function

=NORMAD(Data_Range)

Reference:

Law, Averill M. and W. David Kelton (1991). *Simulation Modeling and Analysis*. New York McGraw-Hill, Incorporated, pg 368-369.

ANOVA

Analysis of Variance (ANOVA) is a statistical technique used to test the homogeneity hypothesis. The homogeneity hypothesis is specified as

$$H_0: \mu_1 = \mu_2 = \cdots = \mu_t \quad \text{vs.} \quad H_1: \mu_1 \neq \mu_2 \neq \cdots \neq \mu_t, \text{ for at least one } \mu_i$$

In Simetar ANOVA is calculated using the function

=ANOVA(DataRange,...)

Degrees of Freedom

The degrees of freedom are calculated for the total model as N-1, for the treatments as t-1, and for the error as N-t.

Reference:

Tamhane, Ajit C. and Dorothy D. Dunlop (2000). *Statistics and Data Analysis From Elementary to Intermediate*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 360-361.

F Test

The test statistic for the homogeneity hypothesis in an ANOVA framework is the F statistic. The F statistic is calculated as

$$F_0 = \frac{MST}{MSE}$$

The null hypothesis is rejected with probability of Type I error α if

$$F_0 > F_{\alpha,(t-1),(N-t)}$$

where,

$F_{\alpha,(t-1),(N-t)}$ = critical values from the F distribution for $P\left(F_0 \geq F_{\alpha,(t-1),(N-t)}\right) = \alpha$

Reference:

Tamhane, Ajit C. and Dorothy D. Dunlop (2000). *Statistics and Data Analysis From Elementary to Intermediate*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 360-361.

Mean Square

The mean square is calculated for the treatments and the error. The mean square of the treatments is calculated as

$$MST = \frac{SST}{t-1}$$

The mean square error is an estimate for the experimental error variance and is calculated as

$$MSE = \frac{SSE}{N-t}$$

Reference:

Tamhane, Ajit C. and Dorothy D. Dunlop (2000). *Statistics and Data Analysis From Elementary to Intermediate*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 360-361.

Sum of Squares

The sum of squares can be broken down and viewed in three components, total sum of squares, treatment sum of squares, and error sum of squares. Total sum of squares measures the variability of the observations with respect to the overall mean and is calculated as

$$SSTot = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{..}\right)^2$$

where,

t = number of treatments
r = number of observations in the ith treatment
y_{ij} = jth observation in the ith treatment
$\bar{y}_{..}$ = overall mean for all treatments

Treatment sum of squares measures the variability between the treatment means and the overall mean and is calculated as

$$SSTrt = r\sum_{i=1}^{t}\left(\bar{y}_{i.} - \bar{y}_{..}\right)^2$$

where,

$\bar{y}_{i.}$ = mean for the ith treatment

Error sum of squares measures the variability between the observations within a treatment and the treatment mean and is calculated as

$$SSE = \sum_{i=1}^{t}\sum_{j=1}^{r}\left(y_{ij} - \bar{y}_{i.}\right)^2$$

Reference:

Tamhane, Ajit C. and Dorothy D. Dunlop (2000). *Statistics and Data Analysis From Elementary to Intermediate*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 360-361.

Chi-squared Test for Normality

The Chi-Squared Test for Normality is based on the more general Chi-Squared test for goodness of fit, which measures how well the observed frequency of a random variable compares to the frequency of the hypothesized distribution. The Chi-Squared test statistic is calculated as follows

$$\chi^2 = \sum_i \frac{\{O(i) - E(i)\}^2}{E(i)}$$

where,

$E(i)$ = the expected number of observations in the ith histogram bin when H_0 is true

$O(i)$ = the number of observations in the ith histogram bin

j = the total number of bins selected

In Simetar the H_0 for the Chi-Squared test is that the data are normally distributed. The parameters of the hypothesized normal distribution, mean and standard deviation, are estimated in Simetar using the observed data. Reject H_0 if χ^2 is greater than the 1-α quantile from the chi-squared distribution with j-3 degrees of freedom (in the generalized Chi-Squared test j-1-k degrees of freedom are used, where k is the number of parameters estimated for the hypothesized distribution).

In Simetar the Chi-Squared Test for Normality is calculated using the function

=NORMCHI(Data_Range,Intervals)

Reference:

Conover, W. J. (1999). *Practical Nonparametric Statistics*. New York John Wiley and Sons, Incorporated, pg 239-243.

Chi-squared Test on Variance

The hypothesis test in Simetar to determine if the sample variance of a data series is equal to a hypothesized parameter variance is the chi-squared test. The chi-squared test is calculated as

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

where,

n = number of observations

s^2 = observed sample variance

σ_0^2 = hypothesized value for the parameter variance

In Simetar the two sided chi-squared test is used with H_0 that the sample variance is equal to the hypothesized parameter variance. The H_0 is rejected at a level of significance 1-α if the chi-squared test is greater than $\chi^2_{n-1,\alpha/2}$ or if the chi-squared test is less than $\chi^2_{n-1,1-\alpha/2}$.

Reference:

Tamhane, Ajit C. and Dorothy D. Dunlop (2000). *Statistics and Data Analysis From Elementary to Intermediate*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 256-257.

F-test

The F-test is used to determine whether or not two sets of data have statistically different variances. The F-test is calculated as

$$F = \frac{\sigma^2_{\text{larger}}}{\sigma^2_{\text{smaller}}}$$

In Simetar the H_0 for the F-test is that the sample variances of the two data series are not statistically different. Reject H_0 at a level of significance α if it the test statistic is greater than the 1-α quantile of the F distribution.

Reference:

Albright, S. Christian, Wayne L. Winston, and Christopher J. Zappe (2000). *Managerial Statistics*. Pacific Grove, California Duxbury, pg 516-517.

Kolmogorov-Smirnov Test for Normality

The Kolmogorov-Smirnov Test for Normality is based on the more general Kolmogorov-Smirnov test for goodness of fit. The Kolmogorov-Smirnov test compares the empirical distribution of a given function to the hypothesized distribution. A comparison of the maximum vertical distance between the cumulative distribution of the data and the cumulative distribution of the hypothesized function is used to calculate the test statistic. The Kolmogorov-Smirnov test statistic is calculated as follows

$$D_n = \max\left[\left|F_n(x) - F(x)\right|\right]$$

where,

$F_n(x) = i / n$

$F(x)$ = the cumulative distribution function of the hypothesized distribution

n = the total number of observations in the data

In Simetar the H_0 for the Kolmogorov-Smirnov test is that the data are normally distributed. The parameters of the hypothesized normal distribution, mean and standard deviation, are estimated in Simetar using the observed data. Reject H_0 at a level of significance α if it the test statistic D_n is greater than the 1-α quantile of the Kolmogorov-Smirnov tables, which can be found in a statistics textbook.

In Simetar the Kolmogorov-Smirnov Test for Normality is calculated using the function

=NORMKS(Data_Range,Adjusted_KS)

Reference:

Law, Averill M. and W. David Kelton (1991). *Simulation Modeling and Analysis*. New York McGraw-Hill, Incorporated, pg 363-367.

Kurtosis

Kurtosis is a measure of the density under the peak of a functions probability density function. Kurtosis as calculated as

$$K = \frac{\sum_{i=1}^{n} (x_i - \mu)^4 p_i}{\sigma^4}$$

It is common to compare Kurtosis values to the Kurtosis of a normal distribution where

$K = 3$ normal distribution

$K < 3$ flatter than a normal distribution

$K > 3$ higher peak than a normal distribution

Kurtosis is calculated in Simetar using the function

=KURT(Number1,*Number2*)

Reference:

Vose, David (2000). *Risk Analysis*. New York John Wiley and Sons, Limited, pg 35-36.

Shapiro-Wilk Test for Normality

The Shapiro-Wilk test for normality is a measure of the straightness of the normal probability plot. The Shapiro-Wilk test is calculated as

$$W = \frac{\left[\sum_{i=1}^{k} a_i \left(X^{(n-i+1)} - X^{(i)} \right)\right]^2}{\sum_{i=1}^{n} \left(X_i - \overline{X} \right)^2}$$

where,

a_i = table value coefficient
$X^{(i)}$ = ith order statistic
$\overline{X}$ = mean
n = total number of observations
k = approximately $n/2$

The H_0 for the Shapiro-Wilk Test is that the data are normally distributed. The H_0 is rejected at the level of significance α if the test statistic W is less than or equal to W_α quantile given by a table of quantiles for the Shapiro-Wilk Test statistic.

In Simetar the Shapiro-Wilk Test is calculated using the function

=NORMSW(Data_Range)

Conover, W. J. (1999). *Practical Nonparametric Statistics*. New York John Wiley and Sons, Incorporated, pg 450-453.

Skewness

Skewness is a measure of the symmetry of a distribution. Skewness is calculated as

$$S = \frac{\sum_{i=1}^{n} \left(x_i - \mu \right)^3 p_i}{\sigma^3}$$

Skewness values give an indication to the density under the tails of a distribution. A normal distribution has a Skewness of 0 since both the right and left tails of the probability distribution function for a normal distribution have the same density. A negative Skewness indicates that a probability distribution function has more density

under the left tail than the right and is thus said to be skewed to the left. A positive Skewness indicates that a probability distribution function has more density under the right tail than the left and is thus said to be skewed to the right.

Skewness is calculated in Simetar using the function

=SKEW(Number1,*Number2*)

Reference:

Vose, David (2000). *Risk Analysis*. New York John Wiley and Sons, Limited, pg 35-36.

T-test on Mean

The hypothesis test in Simetar to determine if the sample mean of a data series is equal to a hypothesized parameter mean is the t-test. The t-test is calculated as

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

where,

$\bar{x}$ = observed sample mean
μ_0 = hypothesized value for the parameter mean
s = observed sample standard deviation
n = number of observations

In Simetar the two sided t-test is used with H_0 that the sample mean is equal to the hypothesized parameter mean. Reject H_0 at a level of significance 1-α if the absolute value of the t-test is greater than $t_{n-1,\alpha/2}$ from the students t distribution.

Reference:

Tamhane, Ajit C. and Dorothy D. Dunlop (2000). *Statistics and Data Analysis From Elementary to Intermediate*. Upper Saddle River, New Jersey Prentice Hall, Incorporated, pg 252-253.

Two Sample T-test

The two sample t-test is used to determine whether or not two sets of data have statistically different means. The two sample t-test is calculated as

$$t - value = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

In Simetar the H_0 for the 2 sample t-test is that the sample means of the two data series are not statistically different. Reject H_0 at a level of significance α if it the test statistic is greater than the 1-α quantile of the student *t* distribution.

Reference:

Albright, S. Christian, Wayne L. Winston, and Christopher J. Zappe (2000). *Managerial Statistics*. Pacific Grove, California Duxbury, pg 447-453.

TIME SERIES

Time Series analysis is a specialized application of the ordinary least squares solution to a linear problem. In time series analysis lagged values of the dependent variable are used as explanatory variables to describe the system. One of the underlying principles behind time series analysis is that the historical values of a variable provide a means of forecasting the future values of that variable. When only lagged values of a single variable are used in a time series analysis the model is called a univariate time series analysis. A system of time series equations that also uses a priori information is called a vector auto regression

Dickey-Fuller Test

A desirable property for a time series analysis is for the data to be stationary. A series of data y_t are said to be stationary if

1. the mean of the data is a finite constant

$$E[y_t] = \mu \quad \text{for all } t$$

2. the variance of the dada is a finite constant

$$\text{var}(y_t) = \sigma_y^2 \quad \text{for all } t$$

3. the covariance of the data for any two lags is a finite constant

$$\text{cov}(y_t, y_{t+k}) = E[(y_t - \mu)(y_{t+k} - \mu)] = \gamma_k \quad \text{for all } t$$

The Dickey-Fuller and Augmented Dickey Fuller tests are used to determine if data are stationary. For a time series equation

$$y_t = \alpha + \beta t + \rho y_{t-1} + \varepsilon_t$$

the Dickey Fuller test is calculated as follows

$$\Delta y_t = \alpha + \beta t + (\rho - 1) y_{t-1} + \varepsilon_t$$

where,

$$\Delta y_t = y_t - y_{t-1}$$

The null hypothesis of the Dickey-Fuller test is that the data series is not stationary or that $\rho_a = 1$. Critical values for the Dickey-Fuller test can be found in a book of statistical tables. The Augmented Dickey-Fuller test is used to test if data are stationary for an AR(p) process greater than 1. The Augmented Dickey-Fuller is calculated as

$$\Delta y_t = \alpha + \beta t + (\rho - 1) y_{t-1} + \sum_{i=1}^{n} \rho_i \Delta y_{t-i} + \varepsilon_t$$

In Simetar the Dickey-Fuller test and the Augmented Dickey-Fuller test is calculated using the function

=DF(Y_Values,Time_Trend,NumLagDiffs,NumDifferences)

Reference:

Griffiths, William E., R. Carter Hill, and George G. Judge (1993). *Learning and Practicing Econometrics*. New York John Wiley and Sons, Incorporated, pg 697-700.

Exponential Smoothing

Exponential smoothing is an approach to simple time series forecasting that addresses concerns with the moving average model. A weighted average of the past observations is used in exponential smoothing to make projections with more weight being given to the most recent information. The simple exponential smoothing model is

$$S_t = \alpha Y_t + (1 - \alpha) S_{t-1}$$

where,

S_t = exonentially smoothed forecast at time t
Y_t = observed data point at time t
S_{t-1} = exonentially smoothed forecast at time $t-1$
α = smoothing constant

A more complex exponential smoothing model includes a trend value and is specified as

$$S_t = \alpha Y_t + (1-\alpha)(S_{t-1} + T_{t-1})$$
$$T_{t-1} = \beta(Y_{t-1} - Y_{t-2}) + (1-\beta)T_{t-2}$$

where,

T_{t-1} = exponentially smoothed trend at time $t-1$
T_{t-2} = actual trend value at time $t-2$
β = smoothing constant for the trend

In Simetar Exponential Smoothing is calculated using the function

=EWMA(Data_Range,DampFactor,TrendFactor,No_Forecast)

Reference:

Albright, S. Christian, Wayne L. Winston, and Christopher J. Zappe (2000). *Managerial Statistics*. Pacific Grove, California Duxbury, pg 876-887.

Impulse Response Function

The Impulse Response Function is used in a vector autoregression to determine the affects of a shock to the system, i.e. the affect on y_t of increasing e_t by one unit. Taking the equations for a two series VAR of order p

$$\begin{bmatrix} X_1(0) \\ X_2(0) \end{bmatrix} = \begin{pmatrix} \phi_{11}(1) & \phi_{12}(1) \\ \phi_{21}(1) & \phi_{22}(1) \end{pmatrix} \begin{bmatrix} X_1(-1) \\ X_2(-1) \end{bmatrix} + \cdots + \begin{pmatrix} \phi_{11}(p) & \phi_{12}(p) \\ \phi_{21}(p) & \phi_{22}(p) \end{pmatrix} \begin{bmatrix} X_1(-p) \\ X_2(-p) \end{bmatrix} + \begin{bmatrix} \varepsilon_1(0) \\ \varepsilon_2(0) \end{bmatrix}$$

This system can be shocked in time period 0 to yield

$$\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{pmatrix} \phi_{11}(1) & \phi_{12}(1) \\ \phi_{21}(1) & \phi_{22}(1) \end{pmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \cdots + \begin{pmatrix} \phi_{11}(p) & \phi_{12}(p) \\ \phi_{21}(p) & \phi_{22}(p) \end{pmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

The system can be moved ahead systematically one time period at a time to determine its Impulse Response Function.

In Simetar the Impulse Response Function returns the number of time periods a system takes to stabilize. The Impulse Response Function is calculated using the function

=IMPULSE(Y_Values,Lags,Impulse_Periods,NumDifferences,Error_Correction).

Reference:

Bessler, David A. (1984). "An Analysis of Dynamic Economic Relationships An Application to the U.S. Hog Market." *Canadian Journal of Agricultural Economics*. 32, 109-124.

Likelihood Ratio Test

The Likelihood Ratio Test for a vector autoregressive process tests the number of lags that should be used in the system. The Likelihood Ratio Test is calculated as

$$L(k) = (T - C)\left(\ln\left|\Omega_k\right| - \ln\left|\Omega_{k+1}\right|\right)$$

where,

T = number of usable observations

C = small sample correctrion

Ω_k = covariance matrix for the VAR of k lags

The null hypothesis for the Likelihood Ratio test is that all parameters at lag *k+1* are zero. The null hypothesis is rejected at the level of significance α if it the test statistic is greater than the 1-α quantile of the chi-squared distribution

In Simetar the Likelihood Ratio Test is calculated using the function

=LRT(Y_Values,Lags,Constant,NumDifferences,Error_Correction).

Reference:

Sims, Christopher A. (1980). "Macroeconomics and Reality." *Econometrica*. 48, 1-48.

Partial Autocorrelation Coefficient

The Partial Autocorrelation Coefficient of order k measures the strength of correlation among pairs of entries in the time series while removing the effects of all autocorrelations below order k. The Partial Autocorrelation Coefficient is calculated by solving the system of Yule-Walker equations

$$\begin{bmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{k-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{k-2} \\ \vdots & \vdots & \vdots & & \vdots \\ \rho_{k-1} & \rho_{k-2} & \rho_{k-3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} \Phi_{k1} \\ \Phi_{k2} \\ \vdots \\ \Phi_{kk} \end{bmatrix} = \begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_k \end{bmatrix}$$

where,

ρ_k = autocorrelation coefficient
ϕ_{kk} = partial autocorrelation coefficient

In Simetar the Partial Autocorrelation Coefficient is calculated using the function

=PAUTOCORR(Y_Values,Lags,NumDifferences).

Reference:

Box, George P. and Gwilym M. Jenkins (1976). *Time Series Analysis Forecasting and Control*. San Francisco Holden-Day, Incorporated, pg 64-65.

Sample Autocorrelation Coefficient

The Sample Autocorrelation Coefficient measures the dependence times series values at one time on the value at another time. The Autocorrelation Coefficient is calculated as

$$r_k = \frac{\sum (x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum (x_t - \bar{x})^2}$$

In Simetar the Sample Autocorrelation Coefficient is calculated using the function

=AUTOCORR(Y_Values,Lags,Differences).

Reference:

Johnston, Jack and John DiNardo (1997). *Econometric Methods*. New York The McGraw-Hill Companies, Incorporated, pg 215-220.

OPTIMIZATION

Golden Section

The Golden Section algorithm is used to find the optimum of a single variable function. An upper and lower bound must be specified to find the optimum of a function using the Golden Section algorithm. Between these bounds the function must be unimodal to insure that the bounded global optimum is found. Given these criteria the Golden Section algorithm uses a ratio of two parts known as the "Golden Section" to specify a pair of points between the upper and lower bounds of the function. Each pair of

points is evaluated to determine which point is closer to the optimum. Based on this evaluation the bounds are adjusted and the procedure is repeated until the difference between the bounds is zero.

The computational procedure used in Simetar to find the maximum of a function f(x) using the Golden Section algorithm is outlined as follows

Define

$f(x)$ = function for which the maximum value will be found

$f(X_L)$ = lower bound of $f(x)$

$f(X_U)$ = upper bound of $f(x)$

$\alpha = \frac{3-\sqrt{5}}{2}$, derived from the golden section

$\beta = 1-\alpha$

ε = desired precision for the optimum

Given $f(x)$, X_L, X_U, and ε

$\Delta = X_U - X_L$

$X_1 = X_L + \alpha \times \Delta$

$X_2 = X_L + \beta \times \Delta$

$F_1 = f(X_1)$

$F_2 = f(X_2)$

While $\Delta > \varepsilon$

If $F_1 < F_2$ **Then**

$X_L = X_1$

$X_1 = X_2$

$\Delta = X_U - X_L$

$X_2 = X_L + \beta \times \Delta$

$F_1 = f(X_1)$

$F_2 = f(X_2)$

Else

$X_U = X_2$

$X_2 = X_1$

$\Delta = X_U - X_L$

$X_1 = X_L + \alpha \times \Delta$

$F_1 = f(X_1)$

$F_2 = f(X_2)$

End If

End While

If $F_1 < F_2$ **Return** $(X_U + X_1)/2$

Else **Return** $(X_L + X_2)/2$

In Simetar the Golden Section optimization is calculated using the function

=OPT(FormulaRef,Constraint,ChangeVariable,LowerOrGuess,UpperBound)

Reference:

Vanderplaats, Garret N. (1984). *Numerical Optimization Techniques for Engineering Design with Applications*. New York McGraw-Hill, Incorporated, pg 41-49.

Newton-Raphson Method

The Newton-Raphson Method is used to find the root of a single variable function. Motivation for the Newton-Raphson method comes from the Taylor series expansion the function f(x) at a point δ.

$$f(x+\delta) \approx f(\delta) + f'(\delta) + \frac{f''(\delta)}{2}\delta^2 + \cdots$$

For $f(x + \delta) = 0$

$$\delta = -\frac{f(x)}{f'(x)}$$

with this consideration the iterative formula for the Newton-Raphson method is

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}$$

The computational procedure used in Simetar to find the value of a function f(x) using the Newton-Raphson method is outlined as follows

Define

$f(x)$ = function for which the root or a specific value will be found

δ = desired value of $f(x)$

θ = initial guess

dx = desired precision for the optimum

Given $f(x)$, δ, θ, and dx

$x_1 = \theta$

Do

$$x_0 = x_1$$

$$x_L = x_0 - \frac{dx}{2}$$

$$x_U = x_0 + \frac{dx}{2}$$

$$f'(x_0) = \frac{f(x_U) - f(x_L)}{dx}$$

$$x_1 = x_0 - \frac{f(x_0) - \delta}{f'(x_0)}$$

Loop While $| x_1 - x_0 | > dx$

Return x_1

In Simetar the Newton Raphson optimization is calculated using the function

=OPT(FormulaRef,Constraint,ChangeVariable,LowerOrGuess,UpperBound)

Reference:

Judd, Kenneth L. (1998). *Numerical Methods in Economics*. Cambridge Massachusetts The MIT Press, pg 96-97.