

Grid and Theme Regression 3.1e

Aka: grid_regression.avx

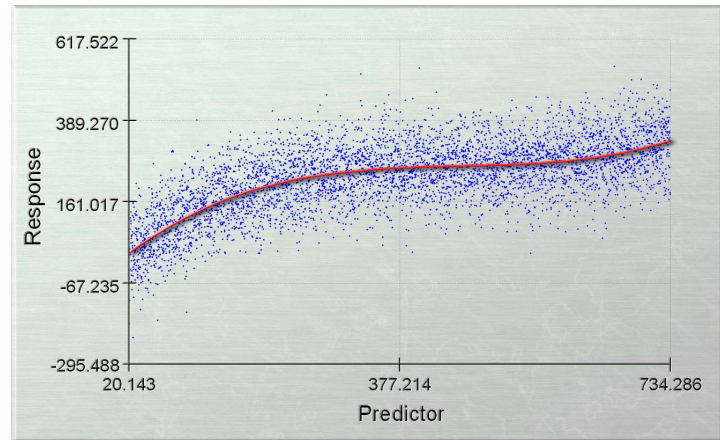
Last modified: August 29, 2006

Topics: Regression, R-Square, Slope, Scatterplot, ANOVA, SLR, Grid, P-value, Statistics, Distributions, t, F, logistic, normal, skewness, kurtosis, binomial, probability, critical, Model, Sequential Sums of Squares, Parameter

AUTHOR:

Jeff Jenness
Jenness Enterprises
3020 N. Schevene Blvd.
Flagstaff, AZ 86004
USA

jeffj@jennessent.com
<http://www.jennessent.com>
(928) 607-4638



DESCRIPTION: This extension allows you to conduct Simple and Multiple Linear Regression analyses for both tabular and grid data. Regression can be conducted with fields in a table or between grid datasets. This extension also gives you the ability to calculate a wider range of summary statistical data than that available in the standard ArcView interface, and the power to generate probability values and critical values from a wide range of statistical distributions.



Linear Regression on Themes and Tables: This function uses Least Squares methods to calculate the linear relationship between one or more independent Predictor variables and a dependent Response variable. Predictor and response values are drawn from fields in an ArcView table or a point/line/polygon theme feature attribute table.



Linear Regression on Grids: This function calculates the linear relationship between one or more grids of predictor values and a grid of response values, again using Least Squares methods.



Summary Statistics: From any numeric field in a table, this function will calculate the mean, standard error of the mean, confidence intervals, mode, minimum, 1st quartile, median, 3rd quartile, maximum, variance, standard deviation, average absolute deviation, skewness (normal and Fisher's G1), kurtosis (normal and Fisher's G2), number of records, number of null values, and total sum.



Probability Calculators: This function will allow you to calculate the probability, cumulative probability and inverse probability (i.e. given a cumulative probability, calculate the corresponding critical value) of a wide range of statistical distributions, including the Beta, Binomial, Cauchy, Chi-Square, Exponential, F, Logistic, LogNormal, Normal, Poisson, Student's T and Weibull distributions. This function is available as a general calculator that remains open until you are finished with it, or as a Table tool that performs the calculations on all selected records in a table.

Acknowledgments: Version 1 of this extension was originally developed for the Inland Water Resources and Aquaculture Service (FIRI) of the United Nations Food and Agriculture Organization (FAO), for use in a training program to teach managers how to utilize GIS technology when managing fisheries. See *Geographic Information Systems in fisheries management and planning. Technical manual*, by G. de Graaf, F.J.B. Marttin, J. Aguilar-Manjarrez & J. Jenness. FAO Fisheries Technical Paper No. 449. Rome. 162p. The author gratefully acknowledges the assistance of **Felix Martin** and **José Aguilar-Manjarrez** of FAO, and **Gertjan de Graaf** of Nefisco Foundation, for their assistance in developing the Grid Regression tools.

Certain tools (esp. the Field Statistics and the histogram) were originally developed by the author for the University of Arizona's Saguaro project (see <http://saguaro.geo.arizona.edu/>) and are included here with their permission. The author thanks **Larry Kendall** of the University of Arizona, and **Scott Walker** of Northern Arizona University, for their help in developing those tools and their willingness to share them. The Theme and Table Regression tools were originally developed for the FAO-FIRI African Water Resource Database project, and portions of this manual are adapted from the documentation for that project. The author gratefully acknowledges the assistance of **Joe Dooley** of Spatial Data Services & Mapping (Namibia), and **José Aguilar-Manjarrez** and **Claudia Riva** of FAO for their assistance in writing and editing that manual.

The Statistical Probability tools are almost identical to those in the author's *Statistics and Probability Tools* extension (see http://www.jennessent.com/arcview/stats_dist.htm) and are included because they enhance and complement the regression functions. The manual for that extension has also been adapted into this manual.

REQUIRES: ArcView 3.x, Spatial Analyst

This extension also requires that the file "avdlog.dll" be present in the ArcView/BIN32 directory (or \$AVBIN/avdlog.dll) and that the Dialog Designer extension be located in your ArcView/ext32 directory, which they usually are if you're running AV 3.1 or better. The Dialog Designer doesn't have to be loaded; it just has to be available. If you are running AV 3.0a, you can download the appropriate files for free from ESRI at:

<http://support.esri.com/index.cfm?fa=downloads.patchesServicePacks.viewPatch&PID=25&MetalD=483>

REVISIONS: See p. 93

Recommended Citation Format: For those who wish to cite this extension, the author recommends something similar to:

Jenness, Jeff. 2006. Grid and Theme Regression 3.1e (grid_regression.avx) extension for ArcView 3.x. Jenness Enterprises. Available at: <http://www.jennessent.com/arcview/regression.htm>.

Please let me know if you cite this extension in a publication (jeffj@jennessent.com). I will update the citation list to include any publications that I am told about.

Copyright © 2005 - 2006










Jeff Jenness is a wildlife biologist with the USFS Rocky Mountain Research Station in Flagstaff, Arizona. He received his BS and MS in Forestry, as well as an MA in Educational Psychology, from Northern Arizona University. Since 1990, his research has focused primarily on issues related to Mexican spotted owls in the southwestern United States. He has also become increasingly involved in GIS-based analyses, and in 2000 he started a GIS consulting business in which he specializes in developing GIS-based analytical tools. He has worked with universities, businesses and governmental agencies around the world, including a long-term contract with the United Nations Food and Agriculture Organization (FAO) for which he relocated to Rome, Italy for 3 months. His free ArcView tools have been downloaded from his website and the ESRI ArcScripts site over 150,000 times.










Table of Contents

GENERAL INSTRUCTIONS:	5
LINEAR REGRESSION:	6
History of Regression:	6
What This Extension Does:	6
Method of Least Squares:	7
Assumptions of Linear Regression:	8
Estimation Uncertainty:	8
R^2	8
Confidence Intervals	8
Different Datasets Require Different Models:	8
Response Follows Linear Relationship:	11
Response Follows Exponentially Increasing Curve:	12
Response Increases and Reaches Plateau:	13
Response Follows S-Shaped Curve:	14
Response Follows Inverse S-Shaped Curve:	15
LINEAR REGRESSION FOR THEMES AND TABLES:	16
Overview:	16
Defining a Model:	16
Additional Statistical Options:	17
Regression Report and Output Options:	18
Model and Parameter Estimates:	21
R-Squared and ANOVA Table:	22
Confidence Bands, Residuals and Predicted Values:	22
Predicting New Observations:	24
Performing Analyses on Different Subsets of Data:	24
LINEAR REGRESSION FOR GRIDS:	29
Restricting the Number of Points:	30
Setting Analysis Boundaries:	31
Technical Note:	32
Technical Note regarding Scatterplots from Grid Data:	32
GENERATING SCATTERPLOTS	33
Predicting New Observations:	35
Altering the Appearance of your Scatterplot:	35
Modifying Text Fonts and Sizes:	35
Modifying X-Axis Attributes:	36
Modifying Y-Axis Attributes:	37
Modifying Description:	38
Refreshing Scatterplot:	38
Adding additional components:	38
Adding your Scatterplot to a Layout:	40
Beware of File Accumulation:	41
DESCRIBING AND PREDICTING NEW OBSERVATIONS USING YOUR MODEL:	43
Describing your Model:	43
Predicting New Observations:	44
Why are the confidence intervals so large?	49

Beware of predicting outside the range of predictor values:	50
A WARNING ABOUT REGRESSION WITH SPATIAL DATA:.....	53
Additional Reading:	53
MANUALLY TRANSFORMING VARIABLES:	55
Transforming Variables in Themes and Tables	55
Transforming Grids:.....	56
FIELD SUMMARY STATISTICS:.....	57
Summary Statistics on a Theme:	57
Summary Statistics on a Field in a Table:	58
<i>Generating Statistics on Multiple Subsets of Data:.....</i>	<i>59</i>
<i>Generating Statistics on a Single Dataset:</i>	<i>60</i>
PROBABILITY DISTRIBUTION CALCULATORS:	65
Avenue Functions:.....	66
<i>Calculating Summary Statistics with Avenue.....</i>	<i>67</i>
Distribution Functions, Parameters and Usages.....	69
<i>Probability Density Functions:.....</i>	<i>69</i>
<i>Cumulative Distribution Functions:</i>	<i>75</i>
<i>Quantiles (also referred to as Inverse Density Functions or Percent Point Functions).</i>	<i>82</i>
TROUBLESHOOTING:	90
REVISIONS:.....	93
REFERENCES:.....	94
INDEX:	95
PLANNED MODIFICATIONS:	98

General Instructions:

- 1) Begin by placing the "**grid_regression.avx**" file into the ArcView extensions directory (../Av_gis30/Arcview/ext32/).
- 2) After starting ArcView, load the extension by clicking on **File --> Extensions...** , scrolling down through the list of available extensions, and then clicking on the checkbox next to the extension called "**Grid and Theme Regression.**"
- 3) This extension will add 4 buttons to your View button bar: 
 - a.  Linear Regression on Feature Themes
 - b.  Linear Regression on Grids
 - c.  Probability Calculator
 - d.  Field Statistics
- 4) This extension will also add 3 buttons to your Table button bar: 
 - a.  Probability Calculator
 - b.  Field Statistics
 - c.  Linear Regression on Table Fields
- 5) This extension creates 2 new document types in your project (*Reports* and *Scatterplots*), each of which contains tools for functions within that document type.

Report	
 Describe Model	 Predict new observations with model
Scatterplot	
 Describe Model	 Change Y-Axis Attributes of Scatterplot
 Predict new observations with model	 Change Description of Scatterplot
 Change Text Attributes of Scatterplot	 Refresh Scatterplot
 Change X-Axis Attributes of Scatterplot	

Linear Regression

There are many good texts that cover linear regression in great depth, and the author highly recommends Applied Regression Analysis (3rd ed.) by Draper and Smith (1998), and Applied Linear Statistical Models: Regression, Analysis of Variance and Experimental Design (4th ed.) by Neter et al. (1996) for a proper understanding of how it all works. I will touch on only the basic concepts here.

History of Regression:

Linear regression provides users with a powerful method for analyzing relationships between data. The history of this technique dates back to 1875 when Sir Francis Galton (1822-1911) observed that subsequent generations of pea plants tended to have less extreme sizes than the previous generations, noting that the offspring of exceptionally large peas tended to be smaller than their parent while the offspring of exceptionally small peas tended to be larger than their parent. The net effect of this phenomenon was to bring the pea population closer to the mean pea size, and Galton used the term “regression” to describe this effect. The natural tendencies of many biological processes are to move away from extreme events and therefore to “regress” to the mean.

Galton did not come up with the mathematical formulas and techniques currently used in most linear regression analyses, however. Galton based his analyses on medians and inter-quartile ranges, possibly not recognizing the statistical advantages of means and standard deviations (see Stanton 2001), but he did recognize the importance of the slope of the best-fitting line and the correlation between variables. In the late 19th and early 20th century, Karl Pearson (1857-1936) provided rigorous mathematical methods and proofs, incorporating the methods of least squares (originated by Adrien-Marie Legendre [1752-1833] and/or Carl Friedrich Gauss [1777-1855]) and the formulas for correlation from Auguste Bravais' (1811-1863), to give a solid mathematical foundation to regression (Denis 2000). The mathematical methods were further refined by R.A. Fisher (1890-1962) in the 1920s (Salsburg 2001).

What This Extension Does:

This ArcView extension has been specifically designed to allow a user to conduct simple linear regression analyses (with a single “independent” or “predictor” variable) or multiple linear regression (with multiple independent variables) and lets the user apply several possible transformations to the predictor variables. These types of regression allow the user to identify whether a dependent variable varies in a predictable way over a range of values of the independent variables. For example, regression analysis could tell a user whether fish stocks tend to rise or fall as nutrient levels in the water rise and fall, and can quantify the linear relationship that may exist between fish stocks and nutrient levels. In addition, the analysis also provides users with the probability that any relationship established is due solely to chance. Once such a relationship has been established, it can be used to explain how much of the variation in fish stocks is due to nutrient levels, and to predict what the fish stock might be at some particular nutrient level.

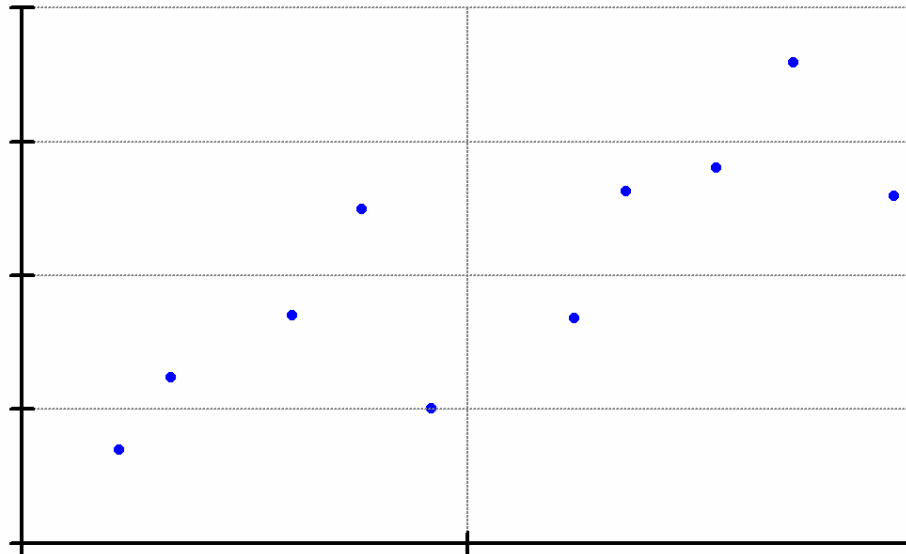
Simple linear regression refers to the most basic type of regression, with a single dependent variable and a single independent variable. Multiple linear regression refers to cases with multiple independent variables, including higher-order models (where the parameters are different exponential values of the same variable).

Despite the variable names (“Independent” and “Dependent”), regression analysis is not intended to demonstrate causal relationships between the dependent and predictor variables. Just because the dependent variable varies in a predictable way over different levels of the independent variables does not necessarily imply that the predictor variable causes the variation. In the above example, one cannot say with any certainty that nutrient levels cause the fish stocks to be at certain levels even if a strong correlation was found. It is possible that both variables may actually be influenced by some third variable, such as precipitation, population density, climatic factors, or even some combination of variables, and that fish stocks and nutrient levels both fluctuate in response to these other factors. True causal relationships

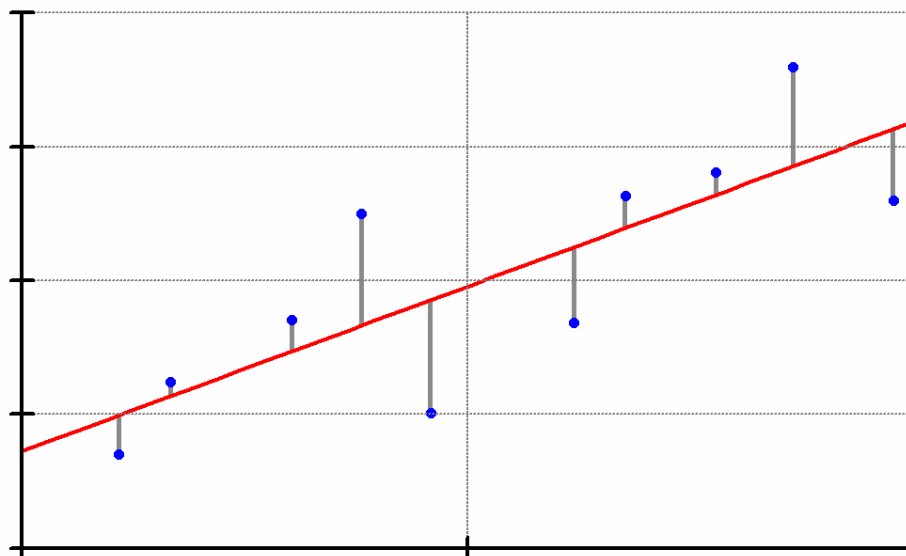
can only be established through controlled experiments where the causal relationship is being specifically tested and measured. However, this does not diminish the value of the correlational relationships that can be identified using regression.

Method of Least Squares:

This extension uses the Least Squares method to generate the best-fitting line for a dataset. This method minimizes the total squared deviation of each sample point from the best-fitting line. For example, given a sample dataset:



There is a unique best-fitting line that minimizes the total vertical distance from each sample point to that line. The best-fitting line is drawn through the cloud of points in the illustration below, and gray lines represent the distances from each point to that line. The distances represented by the gray lines are often referred to as *Errors* (ε_i) or *Residuals*.



This line is just one of an infinite number of possible lines. However, this line is unique in that, if the error value of each point is squared, and all squared errors are summed up, then this line has the smallest total error possible. Any other line drawn through these points would have a higher total error, and therefore

this line is considered the *best-fitting* line. If we wish to say that these points follow some linear pattern, then this line is the best descriptor of that linear pattern. If we wanted to predict what the Y-value of some new point would be, based on some X-value, then this line would be the best tool to use to make that prediction.

Assumptions of Linear Regression:

Like most statistical analyses, linear regression requires certain assumptions to be met in order for the output to be valid. These assumptions are:

- 1) The expected error $E(\varepsilon_i) = 0$, meaning that the regression line is the best predictor for the dependent variable. The Least Squares method produces a regression function that satisfies this assumption. This assumption also allows us to drop the error term (ε) from the estimated regression function.

$$\text{Population function: } Y = \beta_0 + \beta_1 X + \varepsilon$$

$$\text{Estimated Function: } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

- 2) The variance of ε_i is constant over the range of independent values. In other words, the cloud of points in the scatterplot should maintain approximately the same vertical spread about the regression line over the range of independent values.
- 3) The sample points are independent. You should not be able to predict one response value based solely on another response value. This assumption is often violated when regressing spatial data due to spatial autocorrelation, but this does not mean the regression is useless (see *A Warning About Regression with Spatial Data* on p. 53).

Estimation Uncertainty:

We generally like to know both the regression function that best fits our data, plus the uncertainty associated with the various parameter estimates. There are several measures of uncertainty that can be derived using least squares methods:

R²: Also known as the *Coefficient of Multiple Determination*, R² is a numeric measure of how much of the variation in the response variable (in the Y-axis) can be explained by variation in the predictor variable(s). For example, an R² = 0.8 would mean that 80% of the variation in the response variable can be explained by the predictor variable(s).

Some researchers prefer a modified form of R² which is standardized based on the sample size and degrees of freedom (*Adjusted R²*; see p. 22)

Confidence Intervals: All of the parameters in the regression equation are estimates based on the sample data. These terms are random variables, and the least squares method assumes they are normally-distributed random variables, and therefore we can estimate confidence intervals based on the variance of each parameter. When viewed on a scatterplot, these confidence intervals produce confidence bands above and below the actual regression line.

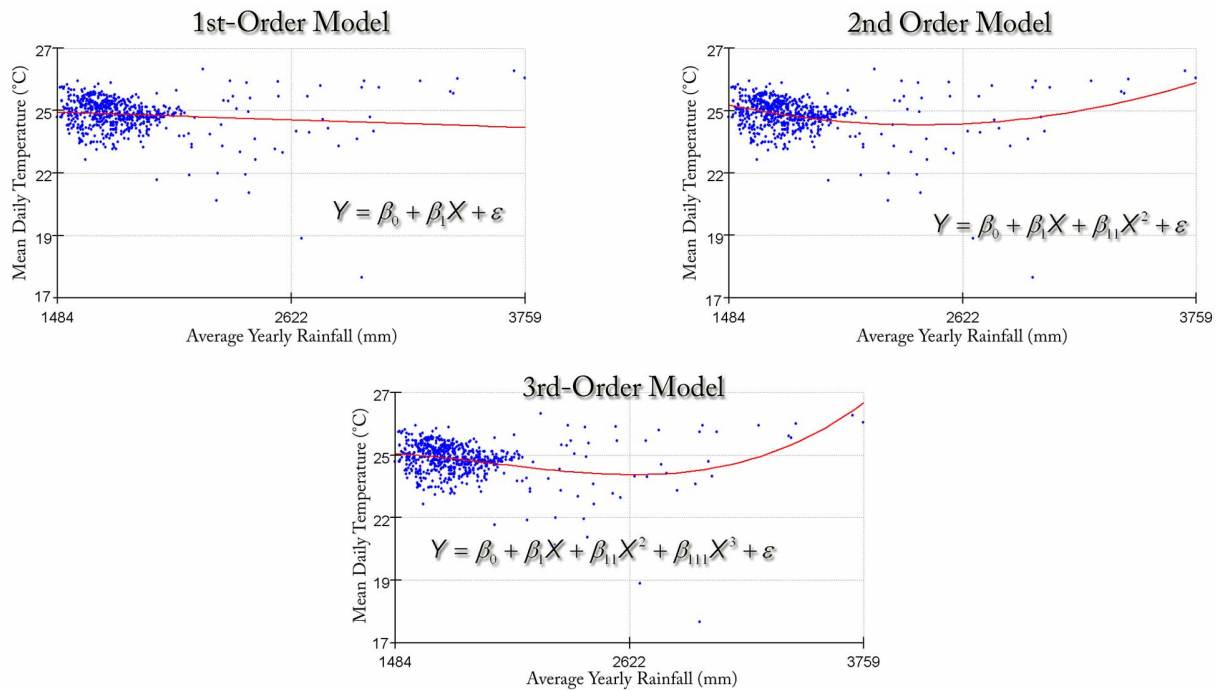
Examples of these concepts are presented on p. 21 of this manual.

Different Datasets Require Different Models:

The most common type of regression involves fitting a straight line to your data, similar to the illustration above. This is an easy relationship to display and explain and it often works well at describing the relationship between predictor and response variable. However, it is only appropriate if there really is a straight-line relationship in your data. Data often follow different patterns, and sometimes more complex models are more appropriate.

This extension allows you to automatically perform several transformations on your predictor variables to create more complex models. All of these options are considered “linear” regression even though only one model produces a straight line. They are considered “linear” because the parameter estimates are all linear. There are many other types of regression available, including nonlinear regression, which you can review at your leisure in a good regression textbook (Draper and Smith [1998] and Neter et al. [1996], just to name two of the author’s favorites).

Higher-order models in particular are curved, but linear regression can fit these models because the parameters (β values) are linear.



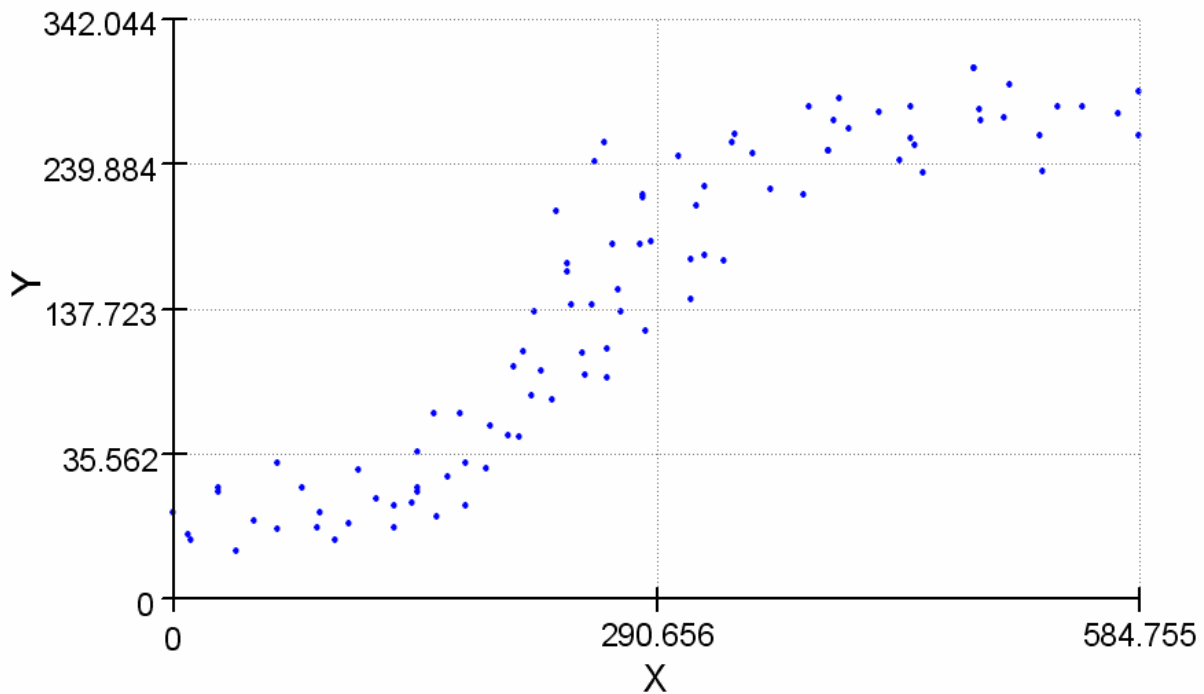
Terminology: There is a difference between the population model and the estimated regression fit. Assuming that there is a true relationship between variables, this relationship is described by the population β parameters in the model. However, we will never know what the actual population parameters are because we can never measure all members of a population. We must be satisfied with estimating those population parameters based on a sample of that population. In this manual, estimates of the population parameters will have little hats on them:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

Actual population parameters will not have hats:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

This extension allows you to generate linear models using inverse, natural log, exponential, 2nd, and 3rd order transformations of your predictor variables. For example, suppose you made a scatterplot of your data and it looked like the following:



There is a visually obvious trend here where Y increases gradually at low levels of X, then increases rapidly for a bit before leveling off at high levels of X. A straight-line relationship would only be accurate over short distances. In this case a more complex model might be more appropriate.

This extension allows you to build complex models with multiple predictor variables and multiple transformations. It would be worth your time to familiarize yourself with the way different models behave in order to decide which model might be most appropriate for your data. As a general guide, the following pages illustrate how 6 basic models perform with five simulated datasets.

This extension provides tools to build models based on 6 basic transformations (see p. 16). All of these transformations are applied to the predictor variable(s). If you would like to build a model using a transformation of the response variable, you can easily transform the variable yourself prior to running the regression analysis (see *Manually Transforming Data* on p. 55).

The 6 basic transformations include:

$$1^{\text{st}}\text{-order Model (straight line): } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$$\text{Full } 2^{\text{nd}}\text{-order Model (Quadratic Polynomial): } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$$

$$\text{Full } 3^{\text{rd}}\text{-order Model (Cubic Polynomial): } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2 + \hat{\beta}_{111} X^3$$

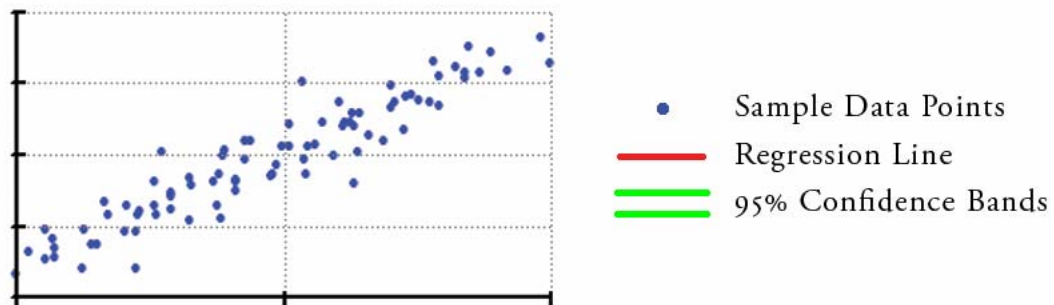
$$\text{Inverse: } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \frac{1}{X}$$

$$\text{Logarithmic: } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln X$$

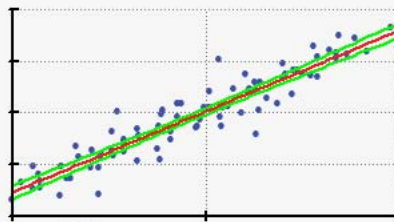
$$\text{Exponential: } \hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 e^X$$

This extension provides tools to build a model including any or all of these transformations.

Response Follows Linear Relationship:

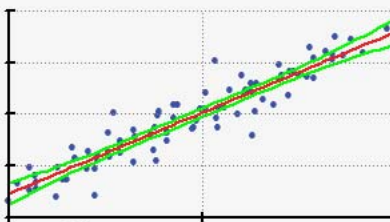


Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$



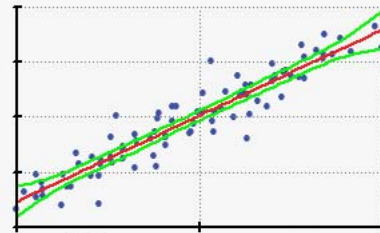
$R^2 = 0.887$
 $R^2 (\text{Adjusted}) = 0.885$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$



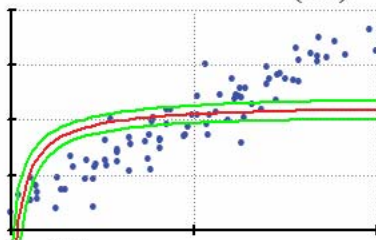
$R^2 = 0.887$
 $R^2 (\text{Adjusted}) = 0.884$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2 + \hat{\beta}_{111} X^3$



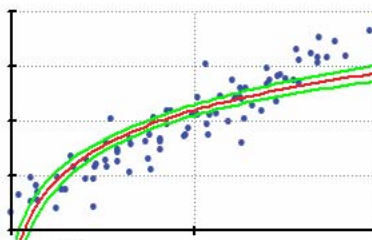
$R^2 = 0.887$
 $R^2 (\text{Adjusted}) = 0.882$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \left(\frac{1}{X} \right)$



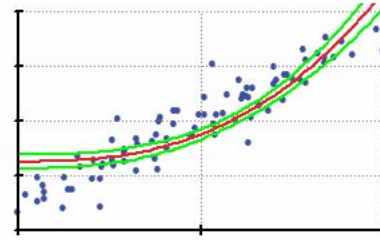
$R^2 = 0.323$
 $R^2 (\text{Adjusted}) = 0.315$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$



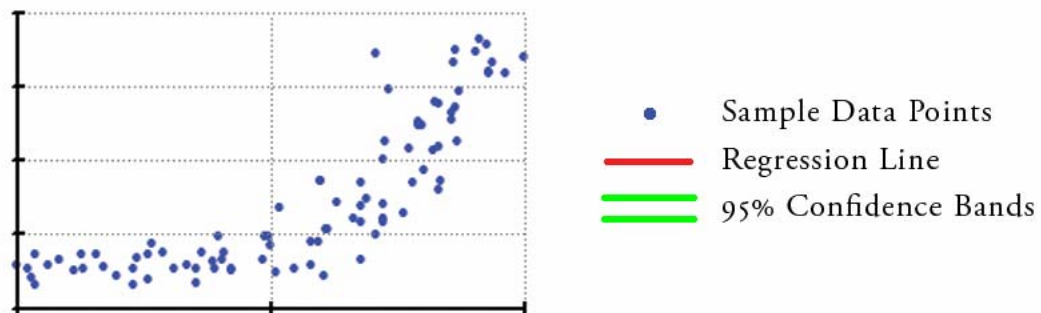
$R^2 = 0.751$
 $R^2 (\text{Adjusted}) = 0.748$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 e^X$

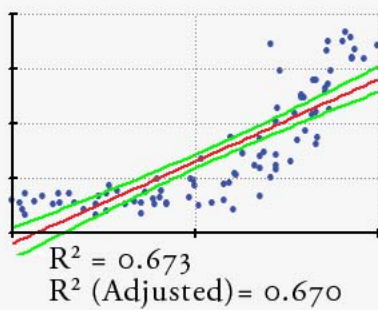


$R^2 = 0.761$
 $R^2 (\text{Adjusted}) = 0.758$

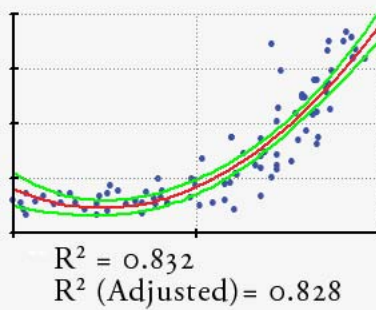
Response Follows Exponentially Increasing Curve:



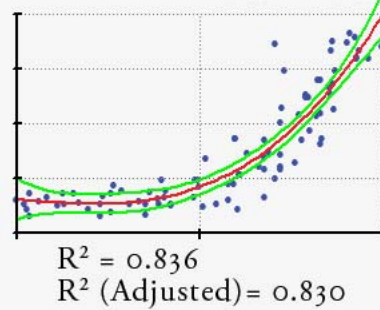
Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$



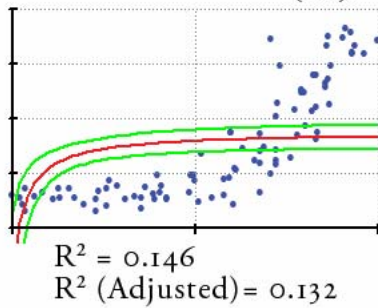
Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$



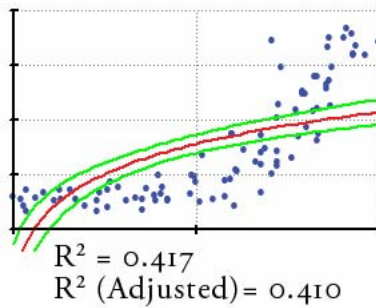
Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2 + \hat{\beta}_{111} X^3$



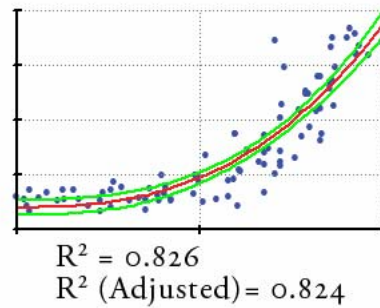
Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \left(\frac{1}{X} \right)$



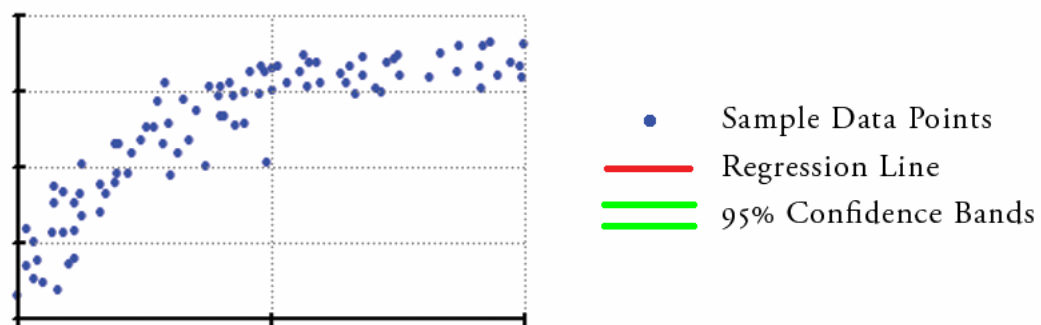
Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$



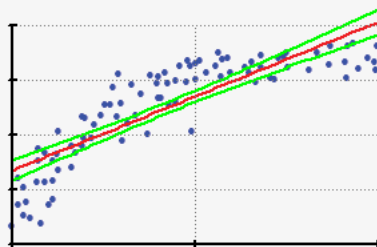
Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 e^X$



Response Increases and Reaches Plateau:

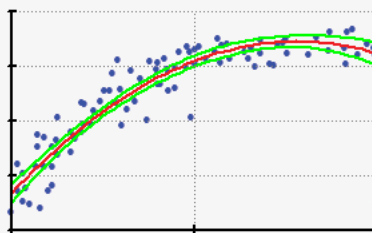


Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$



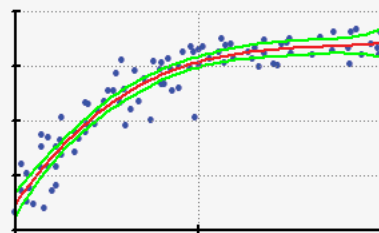
$R^2 = 0.723$
 $R^2 (\text{Adjusted}) = 0.720$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$



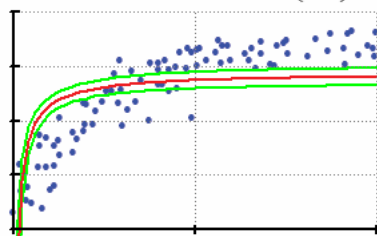
$R^2 = 0.880$
 $R^2 (\text{Adjusted}) = 0.877$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2 + \hat{\beta}_{111} X^3$



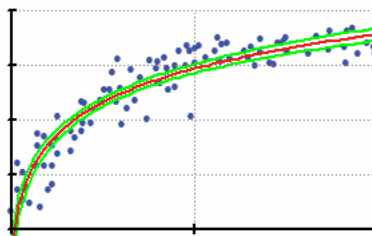
$R^2 = 0.888$
 $R^2 (\text{Adjusted}) = 0.884$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \left(\frac{1}{X} \right)$



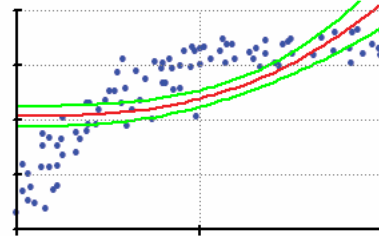
$R^2 = 0.447$
 $R^2 (\text{Adjusted}) = 0.440$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$



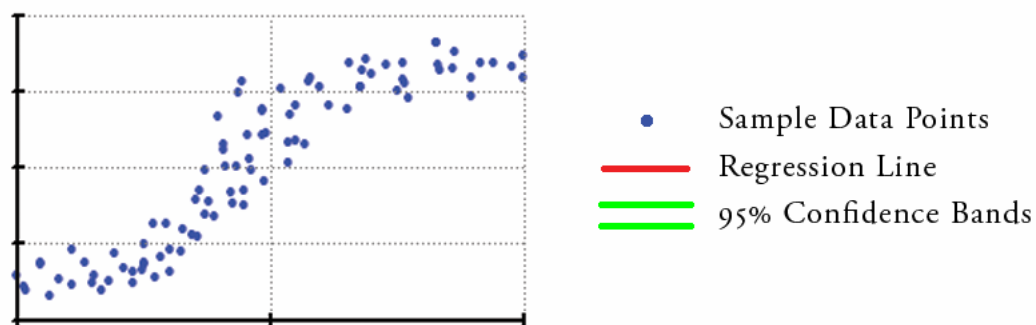
$R^2 = 0.855$
 $R^2 (\text{Adjusted}) = 0.854$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 e^X$

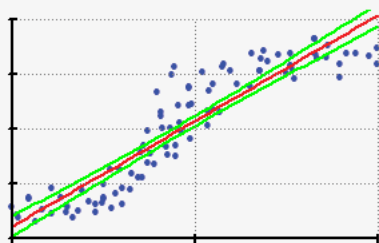


$R^2 = 0.402$
 $R^2 (\text{Adjusted}) = 0.395$

Response Follows S-Shaped Curve:

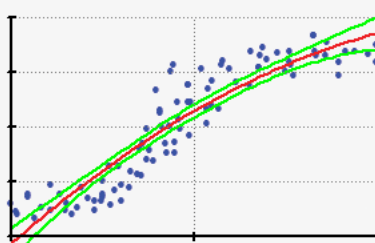


Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$



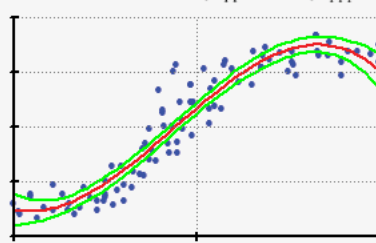
$R^2 = 0.838$
 $R^2 \text{ (Adjusted)} = 0.836$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$



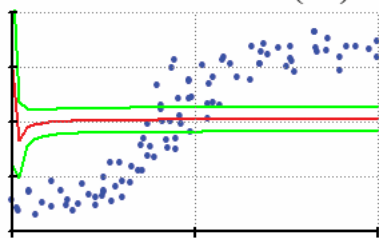
$R^2 = 0.856$
 $R^2 \text{ (Adjusted)} = 0.853$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2 + \hat{\beta}_{111} X^3$



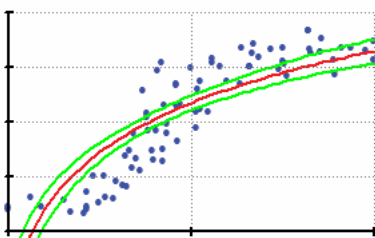
$R^2 = 0.903$
 $R^2 \text{ (Adjusted)} = 0.900$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \left(\frac{1}{X} \right)$



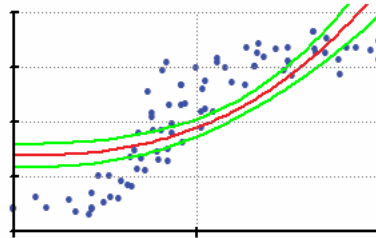
$R^2 = 0.013$
 $R^2 \text{ (Adjusted)} = 0.002$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$



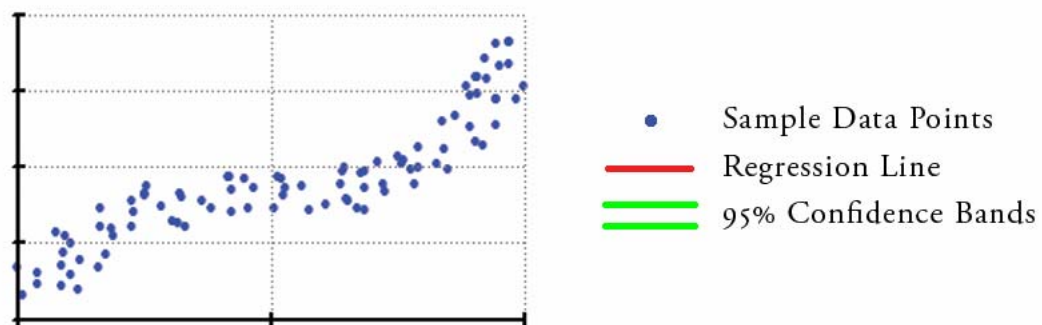
$R^2 = 0.692$
 $R^2 \text{ (Adjusted)} = 0.688$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 e^X$

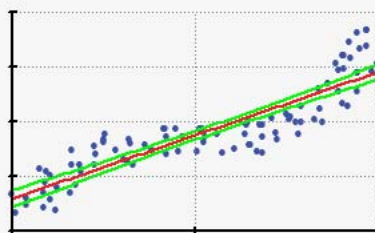


$R^2 = 0.578$
 $R^2 \text{ (Adjusted)} = 0.572$

Response Follows Inverse S-Shaped Curve:

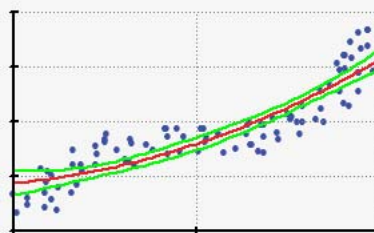


Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$



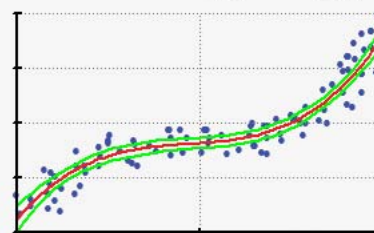
$R^2 = 0.792$
 $R^2 \text{ (Adjusted)} = 0.790$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2$



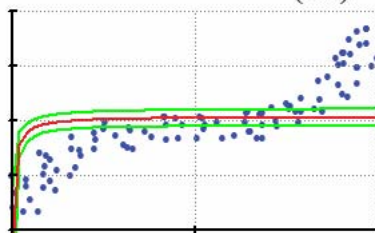
$R^2 = 0.815$
 $R^2 \text{ (Adjusted)} = 0.810$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_{11} X^2 + \hat{\beta}_{111} X^3$



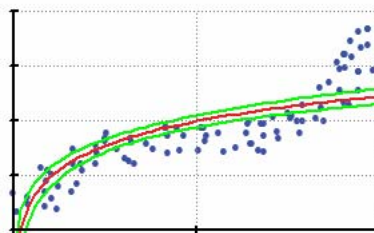
$R^2 = 0.886$
 $R^2 \text{ (Adjusted)} = 0.882$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \left(\frac{1}{X} \right)$



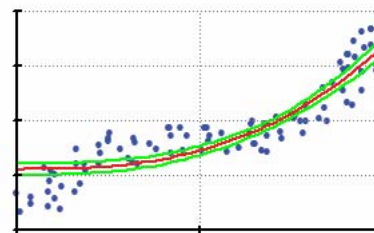
$R^2 = 0.149$
 $R^2 \text{ (Adjusted)} = 0.140$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \ln(X)$




$R^2 = 0.610$
 $R^2 \text{ (Adjusted)} = 0.606$

Model: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 e^X$

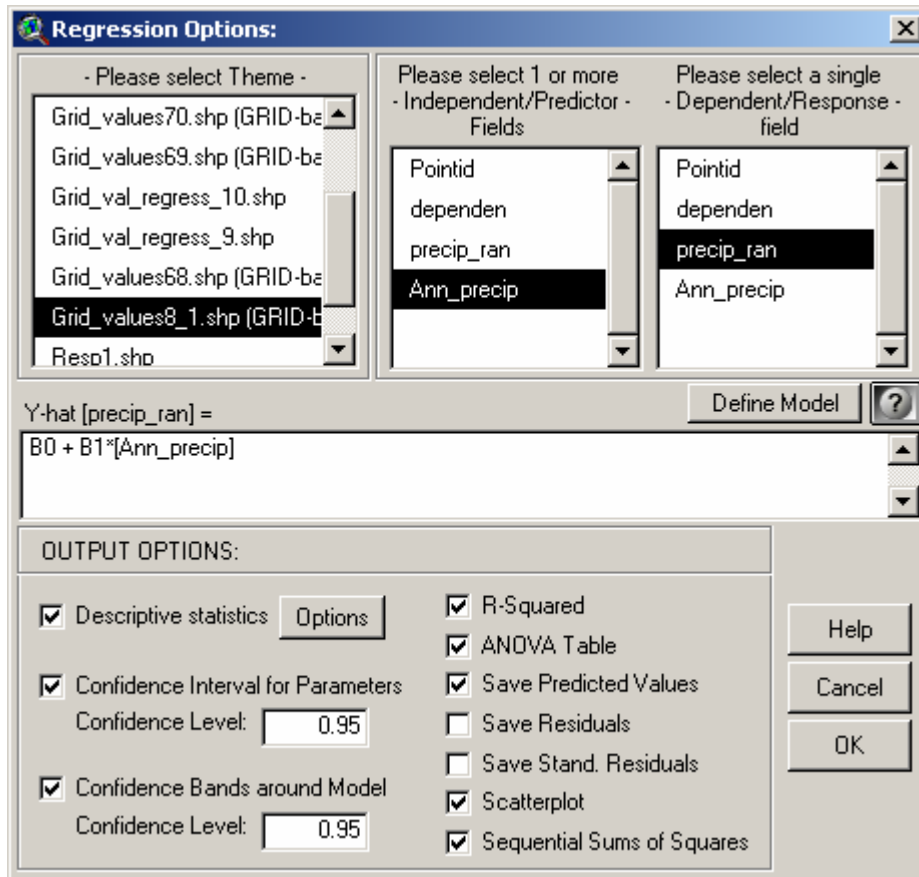


$R^2 = 0.810$
 $R^2 \text{ (Adjusted)} = 0.808$

Linear Regression for Themes and Tables:

The Theme and Table Regression Tool is opened by clicking on the  button in either the view or table button bar. This tool allows users to conduct linear regression analyses between numeric fields in a table and examine the values in these fields for any correlation. The regression tool will either analyze all of the records in a table or only those records that lie within the currently selected set.

Overview:



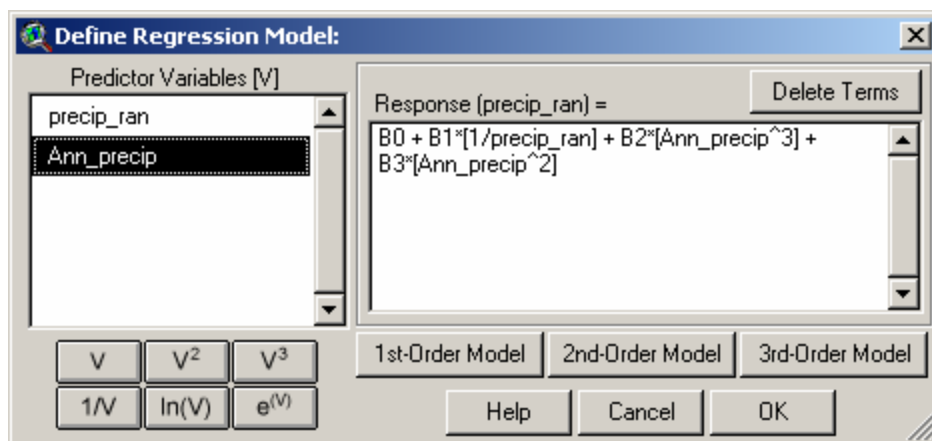
If this tool is opened from a View, the user will see a list of themes in the listbox on the left. When the user selects a theme, the Independent and Dependent Variable listboxes will fill with lists of that theme's numeric attribute fields. If this tool is opened from a Table, the "Theme" listbox will only show the current table and the variable listboxes will automatically fill with the current numeric fields.

Output options associated with this tool are extensive and include the basic R^2 ; an ANOVA table with both the F-values and P-values; basic and standardized residuals; a large variety of descriptive statistics; parameter estimate variability; confidence intervals; predicted values; a scatterplot; and a table of sequential sums of squares.

Defining a Model:

Linear regression finds the best-fitting line by fitting the estimated parameters to some pre-specified model. These models can take many forms and this extension provides a means to fit some of the more popular ones.

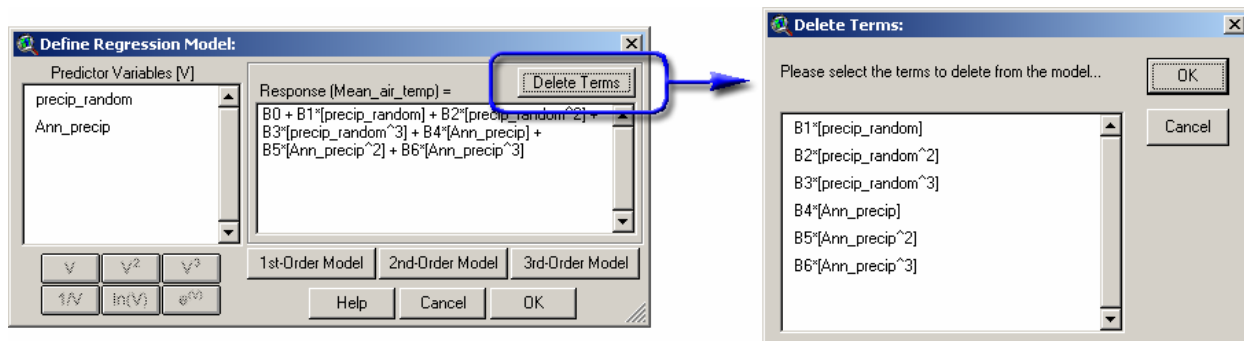
If you simply wish to fit your data to a set of predictor variables, just select them in the list of Independent/Predictor fields. The model will appear in the textbox window as you select the values. If you wish to develop a more sophisticated model, first select all the independent variables you wish to include and then click the “Define Model” button to open the model builder dialog:



To add a term to the model, first select that term in the list of Predictor Variables on the left, then click the appropriate button to add the term with the associated transformation. At this point, you can generate models with 1st, 2nd and 3rd-order terms, and inverse, natural log and exponential transformations. The Buttons for 1st-order Model, 2nd-order Model and 3rd-order Model will automatically generate a full model containing all predictor variables at all levels. For example, clicking on the 3rd-order Model button with the two predictor variables above would automatically generate the following model:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1[\text{precip_ran}] + \hat{\beta}_{11}[\text{precip_ran}]^2 + \hat{\beta}_{111}[\text{precip_ran}]^3 + \hat{\beta}_2[\text{Ann_precip}] + \hat{\beta}_{22}[\text{Ann_precip}]^2 + \hat{\beta}_{222}[\text{Ann_precip}]^3$$

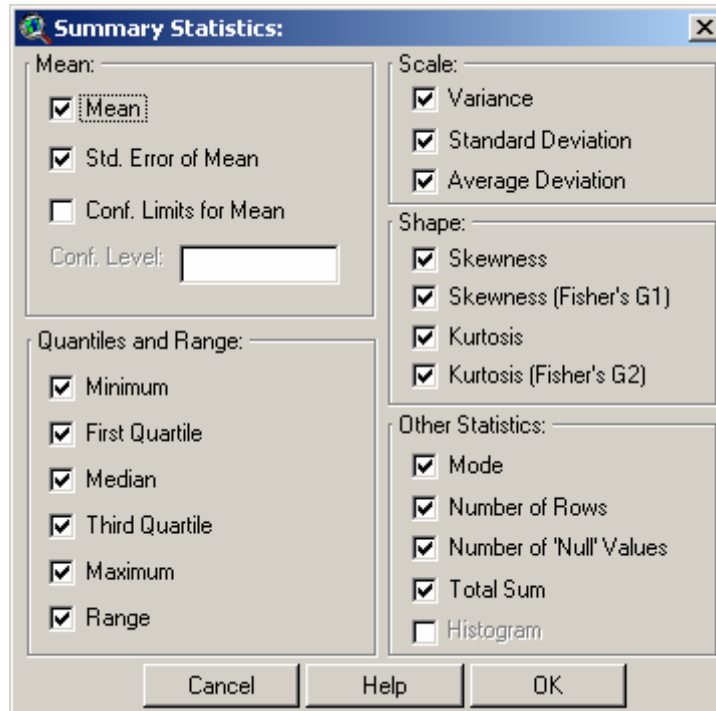
You can delete existing terms by clicking the “Delete Terms” button.



IMPORTANT: You should take care to make sure that any transformations you perform are appropriate for your data. For example, do not attempt to perform a natural log transformation on data that contain negative numbers, which is impossible to do without using imaginary numbers. This regression tool does not work with imaginary numbers, and you will likely get an error message stating that you encountered a singular matrix error (see *Troubleshooting* on p. 90 for more details).

Additional Statistical Options:

This tool provides a wide range of statistical output associated with simple linear regression, as well as general descriptive statistics on the dependent and independent variables. The general descriptive statistics are available by clicking on the “Descriptive Statistics” box and then clicking the “Options” button:



Brief definitions and descriptions of the functions used to derive the above summary statistics for the regression tool are provided in the discussion of Field Summary Statistics. At this point, the “Histogram” function is disabled for regression analyses.

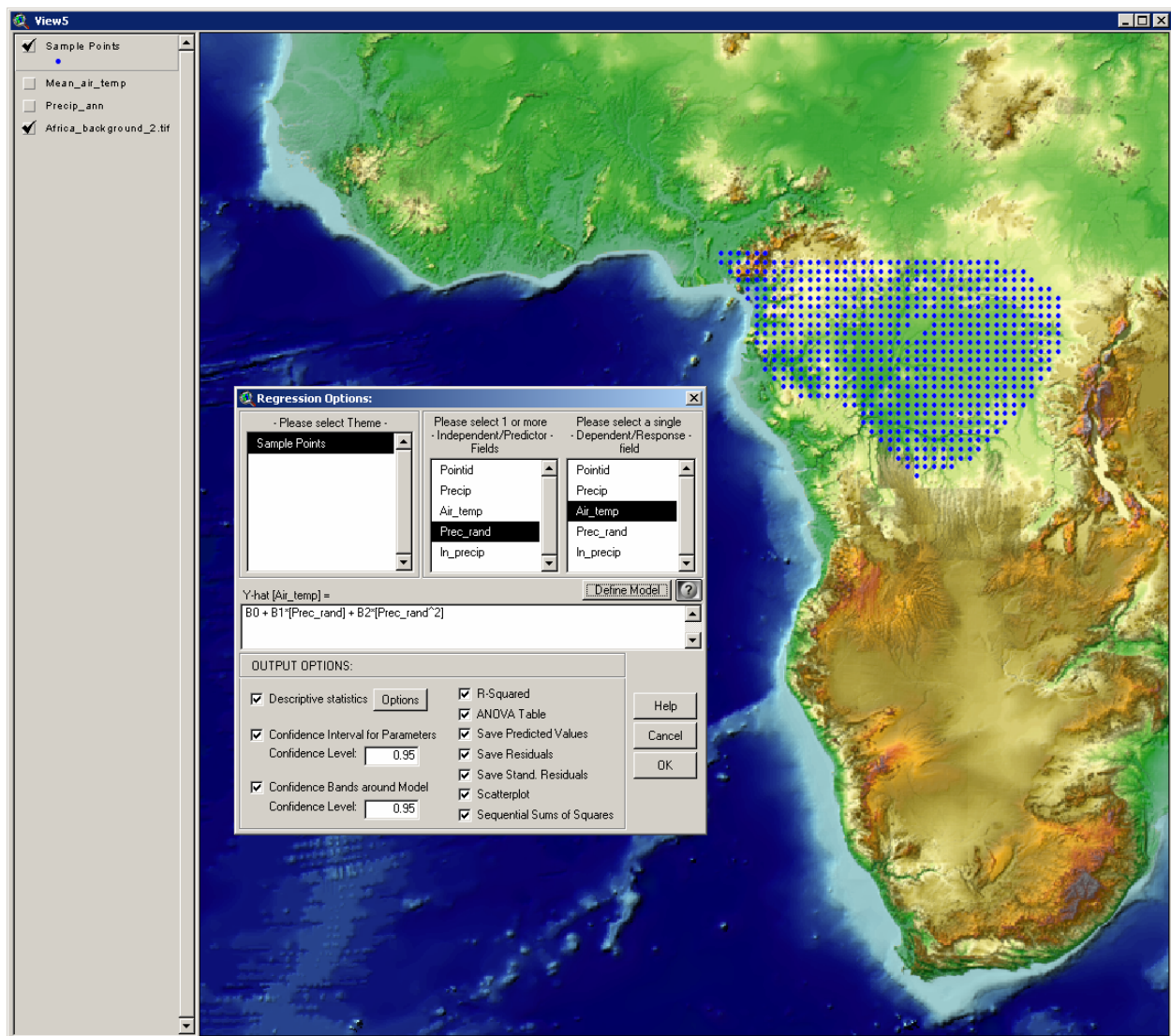
Regression Report and Output Options:

In all cases, this tool will produce a regression report detailing all the output options that were selected. This report will automatically be saved to the hard drive and opened in a text window for the user to review.

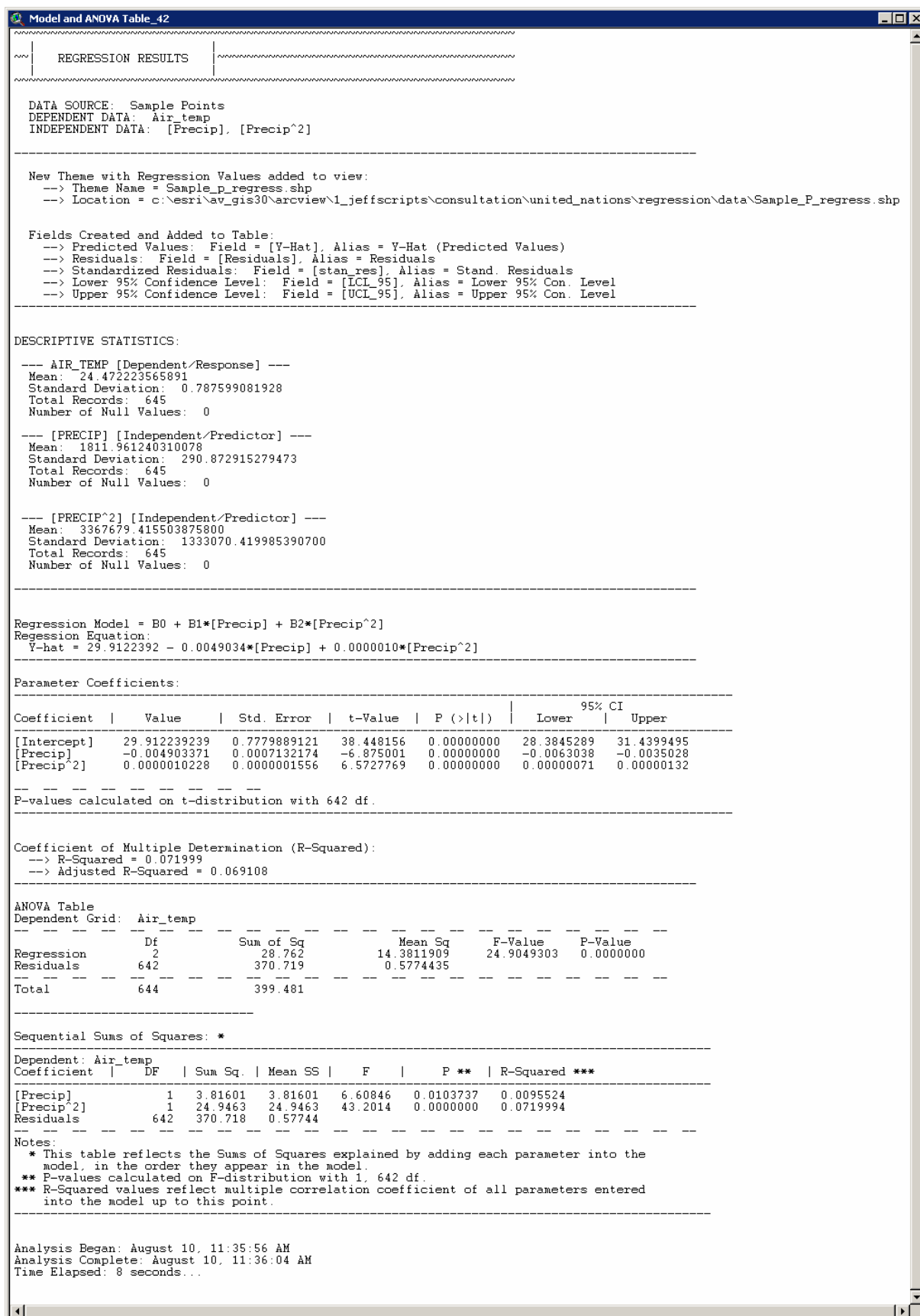
Optionally, a user may also choose to generate a scatterplot illustrating the regression relationship. If so, this scatterplot will open in a new document type called 'Reg. Plots', which is likely located beneath your 'Reports' document type in your project window. The scatterplot will be named 'Scatterplot of [Response Field Name] over [Predictor Field Name]'. Points are generated from the Response and Predictor values by plotting the Predictor value of each feature on the X-axis and the Response value of that feature on the Y-axis. The point data and regression line are presented as themes in the Scatterplot view. Scatterplots are described in detail on p. 33.

If the user elects to generate confidence bands, predicted values, residuals or standardized residuals, then the tool will also produce a new theme in the active view which will be identical to the input theme except that it will also contain fields for these additional values. The name of this new theme will be the same as the input theme, appended by “_regress”. The input theme will never be altered by using this tool, and any predicted values, residuals or confidence levels for the selected set will be added to this new “regression” theme.

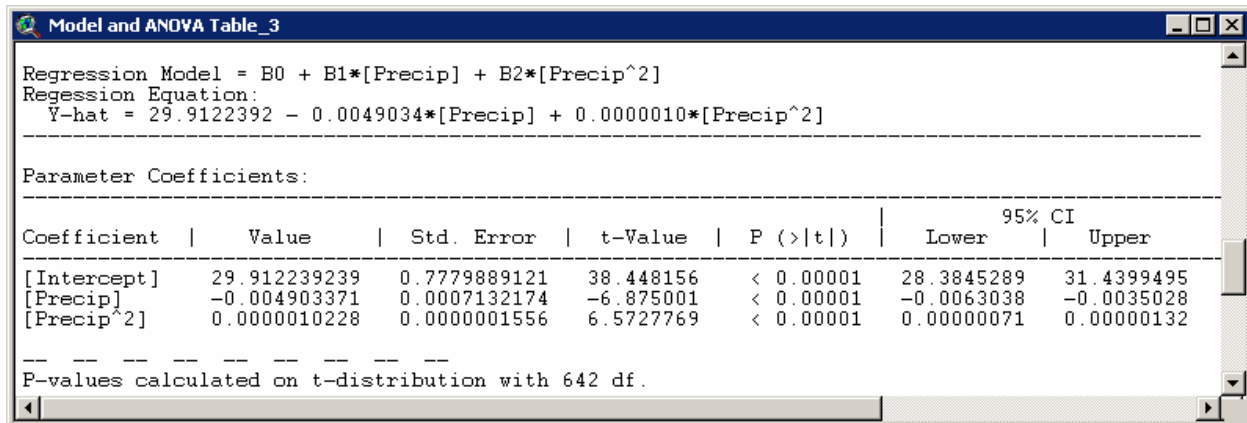
The various output options of this tool can be illustrated by running a sample regression analysis investigating a potential relationship between *Mean Annual Air Temperature* [Air_temp] and *Mean Annual Precipitation* [Precip] for a region along the west coast in central Africa. A scatterplot will also be created, as will the full range of output options by setting up the regression tool as follows:



As soon as the tool finishes, the new “regression” theme will be added to the view, a scatterplot will open, and a regression report will appear. The regression equation, as well as any additional statistics you choose to generate, will open in a new document type called a 'Report', and which should be located beneath your 'Layout' document type in your Project window. These reports will be saved when you save your ArcView project.



Model and Parameter Estimates:



Regression Model = B0 + B1*[Precip] + B2*[Precip^2]
Regression Equation:
Y-hat = 29.9122392 - 0.0049034*[Precip] + 0.0000010*[Precip^2]

Parameter Coefficients:

Coefficient	Value	Std. Error	t-Value	P (> t)	Lower 95% CI	Upper
[Intercept]	29.912239239	0.7779889121	38.448156	< 0.00001	28.3845289	31.4399495
[Precip]	-0.004903371	0.0007132174	-6.875001	< 0.00001	-0.0063038	-0.0035028
[Precip^2]	0.0000010228	0.0000001556	6.5727769	< 0.00001	0.00000071	0.00000132

--- --
P-values calculated on t-distribution with 642 df.

In all cases the regression report will include the regression equation, containing the parameters that best fit the data to the model. In simple linear regression, this equation will be based on the straight line model:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where “Y” is the dependent variable, “X” is the independent variable, β_1 is the parameter estimate of that variable (i.e. the slope) and β_0 is the y-intercept of the line. The example illustrated in the report window above fits the data to a 2nd-order polynomial model:

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$$

and therefore parameters must be estimated for the *y-intercept* (β_0), *Precipitation* (β_1) and *Precipitation*² (β_{11}).

The entire regression equation can then be used to predict new values of *y* based on values of *x*. For example, a user might be interested in predicting what the mean annual air temperature might be if an area received 2600 mm/yr of precipitation. Based on this equation, the best estimate of mean annual air temperature would be calculated as:

$$\begin{aligned} y &= 29.9122392 - 0.0049034(2600) + 0.000001(2600^2) \\ &= 23.9 \text{ Degrees Celsius} \end{aligned}$$

For more information on using your model to predict new observations, please review the discussion on *Predicting New Observations* on p. 43.

Confidence Intervals for Parameter Estimates: This option produces a variety of statistics regarding the parameter estimates, including several measures of the variability and uncertainty of each estimate. Selecting this option will provide the user with values for the parameter, the parameter standard deviation, and the upper and lower confidence limits based on the confidence level specified.

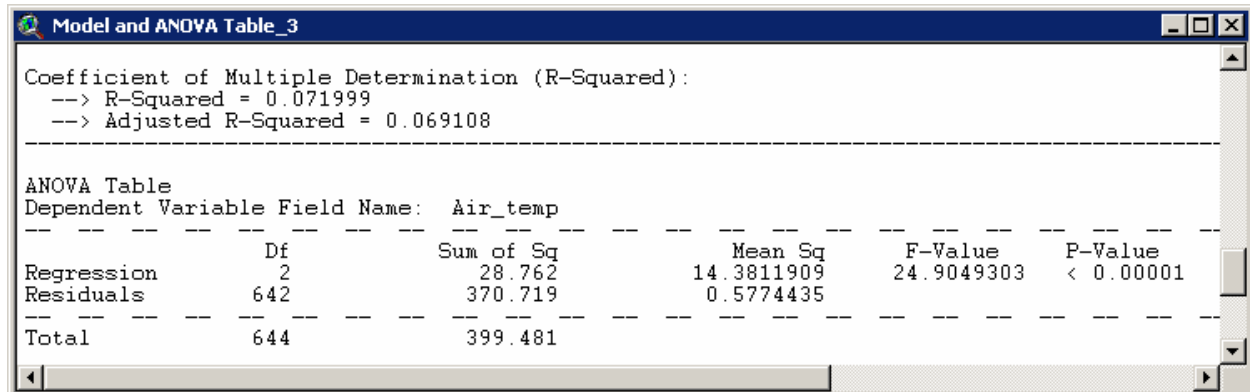
One important use of these statistics is to confirm whether there really is a relationship between the independent and dependent variables. If the parameter were equal to 0 (i.e. a perfectly flat slope), then the dependent variable would not change at all as the independent variable changed and therefore there would be no relationship between them.

This value is considered an “estimate” because it is generated from a sample of precipitation values rather than the full population of all possible precipitation values. We would like to know the true population regression parameter value but it is rarely possible to measure the entire population, and therefore we have to accept an estimate of the slope based on a sample of the data. This is one of the fundamental foundations of statistics.

The confidence limits of the parameter estimate tell us how confident we are about our estimate. Our 95% confidence intervals should be interpreted to mean that, if we took an infinite number of samples of

air temperature and annual precipitation, then the true population regression parameter (the one that we are really interested in) would lie between the upper and lower confidence limits 95% of the time. In this case, we can take this to mean that there is approximately a 95% chance that the true population parameter for β_1 (*Precipitation*) lies within the interval [-0.0063, -0.0035], and our best estimate of it is -0.0049. Likewise, our best estimate of β_{11} (*Precipitation*²) is 0.000001, and there is a 95% chance that the true population parameter lies within the interval [0.00000071, 0.00000132]. If either confidence interval contained 0 within its bounds, then this would be evidence that there may not be any relationship at all between this variable and the annual air temperature.

R-Squared and ANOVA Table:



Model and ANOVA Table_3

Coefficient of Multiple Determination (R-Squared):
 --> R-Squared = 0.071999
 --> Adjusted R-Squared = 0.069108

ANOVA Table
 Dependent Variable Field Name: Air_temp

	Df	Sum of Sq	Mean Sq	F-Value	P-Value
Regression	2	28.762	14.3811909	24.9049303	< 0.00001
Residuals	642	370.719	0.5774435		
Total	644	399.481			

R-squared: Also called the *Coefficient of Determination*, this value is a measure of how much of the variability in the dependent variable can be explained by the variability in the independent variables. In the above example, an R^2 value of 0.072 indicates that only 7.2% of the variability in the dependent variable *Annual Precipitation* can be explained by variation in the independent variable *Mean Annual Air Temperature*. This tells a user that they should be very hesitant about predicting annual air temperature in a particular area based only on the mean precipitation in that area, because precipitation appears to have little influence over air temperature.

Adjusted R-Squared: Some people prefer a variant of R^2 that attempts to standardize R^2 values from different analyses so they may be more easily compared (see Draper and Smith [1998:139-140] for a discussion of this concept). This extension provides both R^2 and Adjusted R^2 . Adjusted R^2 is calculated as:

$$R_A^2 = 1 - \left(1 - R^2\right) \left(\frac{n-1}{n-p}\right)$$

where n = sample size

p = degrees of freedom due to regression

ANOVA Table: Also called an *Analysis of Variance* table, this table provides a breakdown of the various components of the regression relationship as well as an estimate of the confidence that a true linear relationship exists between the two variables. The *P-Value* reflects the probability that the relationship examined is not linear at all, but that it is rather simply an artifact of random chance. In this case, the *P-value* < 0.00001 indicates that the chances are extremely remote that the relationship is due to chance, and therefore it can be concluded that there is indeed strong evidence of a linear relationship between the two variables.

Confidence Bands, Residuals and Predicted Values:

Confidence Bands: It is often wise to include some measure of the uncertainty of any statistical output and this applies to regression as well as to most other statistical analyses. Although the plotted

regression line is the best estimate of the relationship, there will always be some uncertainty unless every possible combination of dependent and independent variables are sampled for all locations for all time.

The confidence bands in this case reflect the upper and lower confidence levels for the regression line over different levels of the independent variable. Since a confidence level of 95% was used, the results should be interpreted to mean "If identical regression relationships were developed an infinite number of times, based on an infinite number of random samples of the respective variable populations, then the true regression line will lie within the confidence bands 95% of the time."

It can also be observed from the scatterplot above that the confidence bands tend to diverge from the regression line at higher levels of annual precipitation. This is because the regression relationship is strongest when the sample points are close to the means of the input variables. The confidence bands will always be closest to the regression line at the mean value of the Independent variable, and diverge as one moves away from that mean.

Confidence band values will also be added to the new "regression" theme attribute table, in fields labeled "LCL" (for Lower Confidence Limit) and "UCL" (for Upper Confidence Limit):

W's ID	W's name	Appec_mean	Elev_mean	model	resids	res_stand	LCL	UCL
1	Coastal Drainage	746.9035	117.6459	659.6537375835	87.2497624165	0.1297009281	625.4230686098	693.8844065572
2	Oued Sedienne	660.1838	219.2642	672.4504137821	-12.2666137821	-0.0182349057	642.0465809831	702.8542465811
3	Garaet El Ichkeul	675.0000	6.0354	645.5987550767	29.4012449233	0.0437063511	606.6705149502	684.5269952032
4	Wadi Fartot	675.0000	71.1535	653.7990028703	21.2009971297	0.0315162922	617.6640098843	689.9339958563
5	Oued Bou Namoussa	600.3897	283.7665	680.5731145866	-80.1834145866	-0.1191964654	652.2922273041	708.8540018691
6	Oued Medjerda	513.7298	112.7502	659.0372276865	-145.3074276865	-0.2160064131	624.6100172499	693.4644381231
7	Coastal Drainage	427.8265	219.3318	672.4589265727	-244.6324265727	-0.3636577554	642.0574578091	702.8603953363
8	Coastal Drainage	449.5599	91.0955	656.3102761032	-206.7503761032	-0.3073442828	621.0022092318	691.6183429746
9	Oued Radjeria	593.1285	294.3992	681.9120783266	-88.7835783266	-0.1319810185	653.9533798324	709.8707768208
10	Oued Sebaou	515.4745	495.6404	707.2541522859	-191.7796522859	-0.2850895887	683.4268160249	731.0814885469
11	Oued Seybouse	536.3246	624.6031	723.4942774760	-187.1696774760	-0.2782366416	699.8410842289	747.1474707231

Residuals and Predicted Values: Along with confidence bands, the new "regression" theme attribute table also contains fields for the predicted values, residuals and standardized residuals. The "model" field holds values for the predicted value of *Mean Annual Precipitation* based on the regression equation and that record's *Mean Elevation* value. The *Residuals* field [resids] holds values reflecting how much the measured *Mean Annual Precipitation* deviated from the predicted model value. The *Standardized Residuals* field [res_stand] standardizes these residuals values by converting them to Z-scores, making it easy to identify extreme outliers.

Sequential Sums of Squares: You may be interested in how much each successive term contributes to your model. The *sequential sums of squares* option allows you to calculate the proportion of the variance explained by the model as each term is added to the model. For example, you may wish to know whether a particular variable was worth including in the model at all. In the illustration below, both terms contributed significantly to the final model but [Precip^2] had a much greater contribution than [Precip].

Sequential Sums of Squares: *						
Dependent: Air_temp						
Coefficient	DF	Sum Sq.	Mean SS	F	P **	R-Squared ***
[Precip]	1	3.81601	3.81601	6.60846	0.0103737	0.0095524
[Precip^2]	1	24.9463	24.9463	43.2014	< 0.00001	0.0719994
Residuals	642	370.718	0.57744			



Notes:

- * This table reflects the Sums of Squares explained by adding each parameter into the model, in the order they appear in the model.
- ** P-values calculated on F-distribution with 1, 642 df.
- *** R-Squared values reflect multiple correlation coefficient of all parameters entered into the model up to this point.

You may also wish to check how the model performs if the terms are entered in a different order. You may use the *Define Regression Model* dialog (see p. 16) to enter terms in any order you wish.

Predicting New Observations:

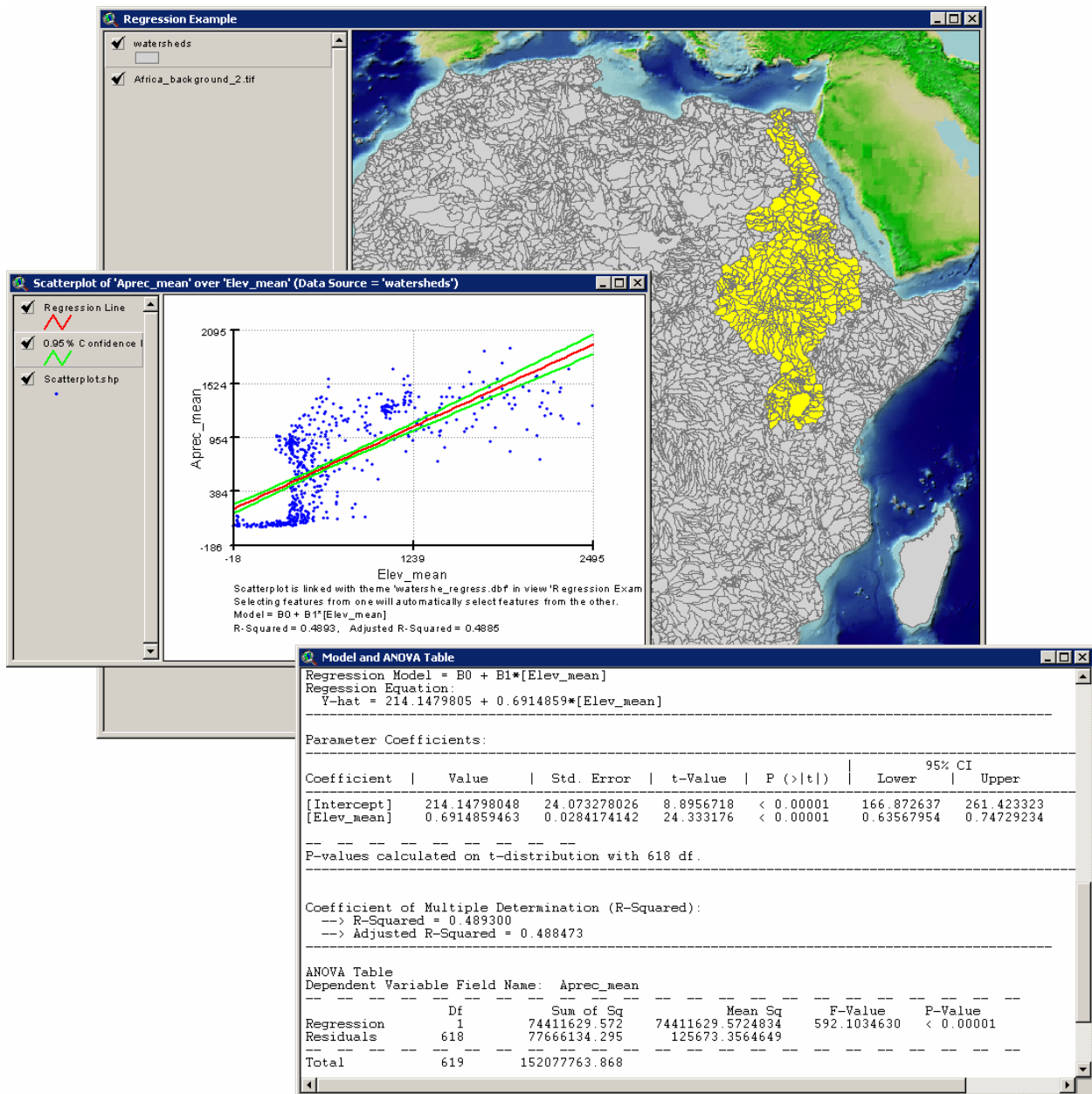
The report document includes 2 buttons in the button bar which allow you to use your model to predict new observations. These functions are described in detail on p. 43, but briefly they are:

- 1)  Describe your model: This function generates a report with all the values you will need if you wish to predict new observations in some other software.
- 2)  Predict New Observations: This function predicts new observations using either specified predictor values, a table of predictor values, or predictor grids.

Performing Analyses on Different Subsets of Data:

As was mentioned earlier, one of the strong points of this Regression tool is that a user can restrict the analysis to a subset of features by selecting those features prior to analysis. If any features are selected, then the tool will only operate on those selected features. If no features are selected, then the tool will operate on all features in the theme.

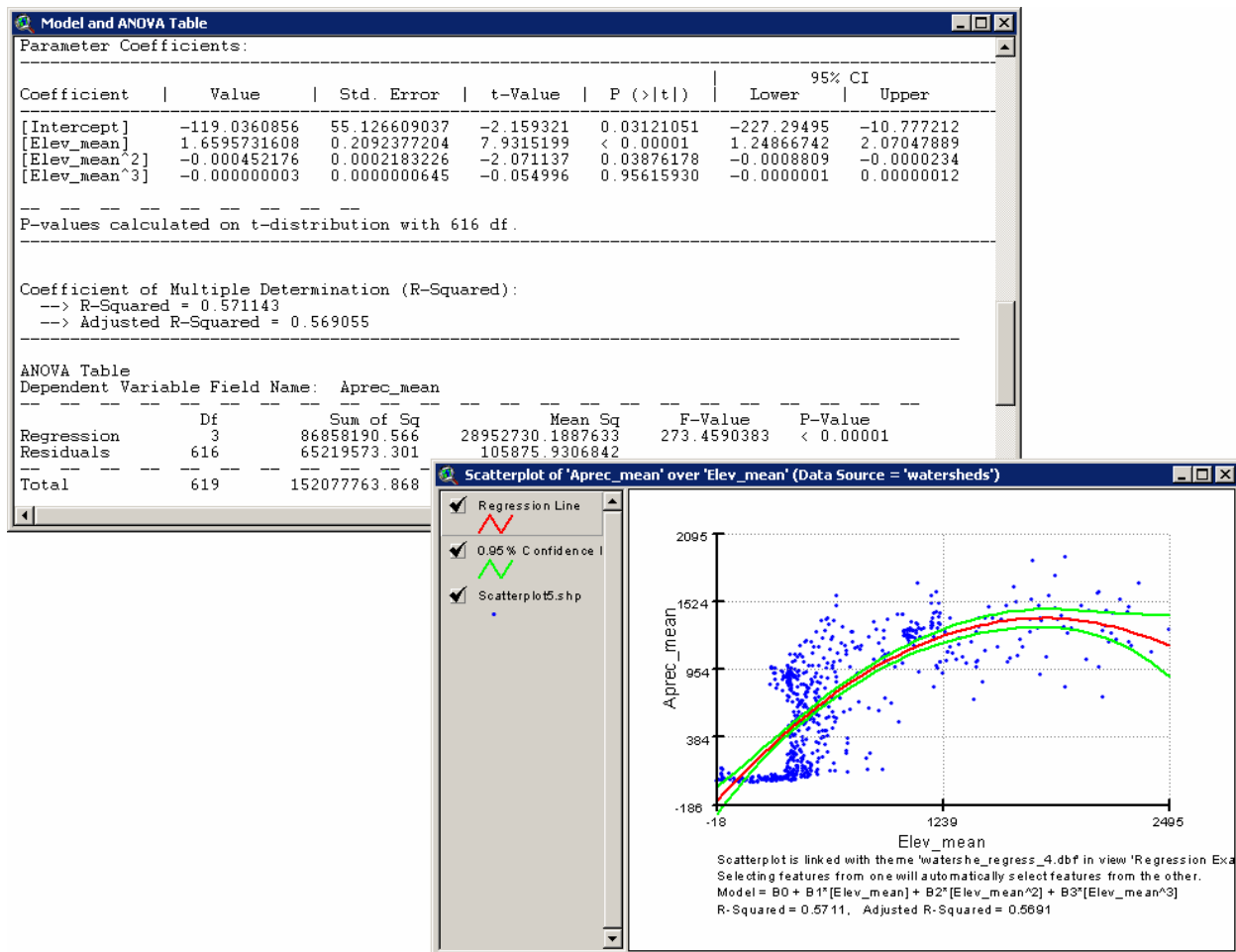
If, for example, a user was really only interested in phenomena occurring within those watersheds that comprise the Nile River basin, then they could select those watersheds and run the Regression tool on that subset.



Interestingly, in this example there is a reasonably strong regression relationship where the annual precipitation appears to have a strong correlation with elevation, although visual examination of the scatterplot indicates that a linear model may not be the best choice for this dataset. The evidence for a linear relationship is very strong with a $P\text{-Value} < 0.00001$, and the R^2 -value of 0.49 implies that 49% of the variation in precipitation can be explained by variation in elevation. In general, this analysis shows that annual precipitation in the Nile river basin is reasonably well correlated with elevation, and that higher elevations tend to get more precipitation than lower elevations.

Visual examination of the scatterplot shows that the relationship does not appear to be linear over the full range of elevation values. Lower elevations tend to have very low precipitation levels, and the precipitation becomes much more variable at approximately 400 meters. This suggests that it might be interesting to run the regression twice; once for watersheds < 400 meters and again for watersheds > 400 meters.

A different model would have given us a higher R^2 , but this dataset appears to follow markedly different patterns over different ranges of Elevation and therefore it would probably make more sense to run the analysis separately on the different ranges.



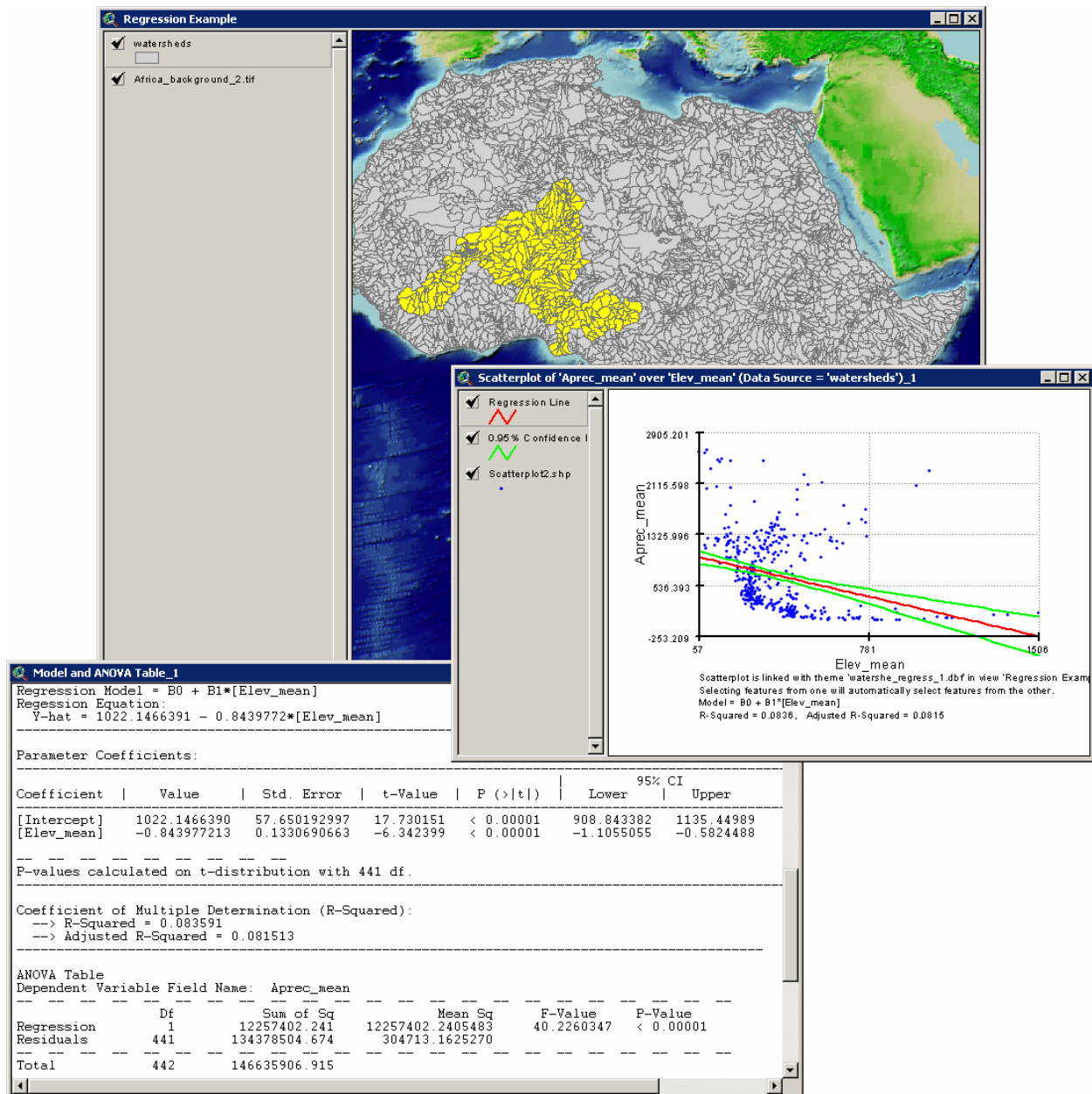
By the way, this example does not intend to suggest that you should keep running new regression analyses looking for the best model that fits your data! Although it is tempting to try several different models and use the best one, this process is called data-snooping and it defeats the purpose of statistical analysis and random sampling. Ideally you should have a hypothesis already in mind before you decide on a model, and you should design your model to test that hypothesis.

Keep in mind that almost any set of data will have some statistically unlikely quality to it somewhere, which will occur simply by chance. Perhaps a large random sample of people will have an unusually large proportion who were born on a Thursday, or who have a first name starting with an "A". If you dig far enough, you can probably find something statistically unusual in the dataset. However, just because such an artifact exists in your data does not imply that such a phenomenon really exists in the population. It may point to something interesting to look at in future research, but it is poor statistical practice to use that artifact when drawing conclusions about the data.

If we were to try many different models before deciding which one best fit our random sample of data, this would be equivalent to conducting a study on income levels, surveying a random sample of people to find out their income, noticing incidentally that the people in our sample coincidentally happened to be born mostly on Thursdays, and concluding from our study that most people are born on Thursday. Our conclusion would have nothing to do with our original question, would likely be wrong, and would (deservedly) open us up to ridicule from the scientific and general community.

If you are unsure about what model might be appropriate for your research question, a good strategy would be to conduct a pilot study with a smaller sample of data prior to the main study. Use the pilot data to decide which model would be most appropriate, then get a new sample of data for the main study.

Back on topic, by selecting the Niger river basin, the analysis can quickly be re-run on a different region:



The results of this analysis are also interesting in that they differ dramatically from the relationship established for the Nile river basin. The evidence for a linear relationship is still very strong with a P -Value < 0.00001, but the R^2 -value is very low at 0.08, meaning that elevation appears to have little influence on annual precipitation in this region. Interestingly, the regression relationship takes a different direction than was seen in the Nile megabasin, in that here the mean annual precipitation decreases as elevation increases.

Visual examination of the scatterplot again indicates that the relationship might be better modeled with a more complex model.


There are several excellent texts available that discuss regression in exhaustive detail. For those who are interested, the author recommends:

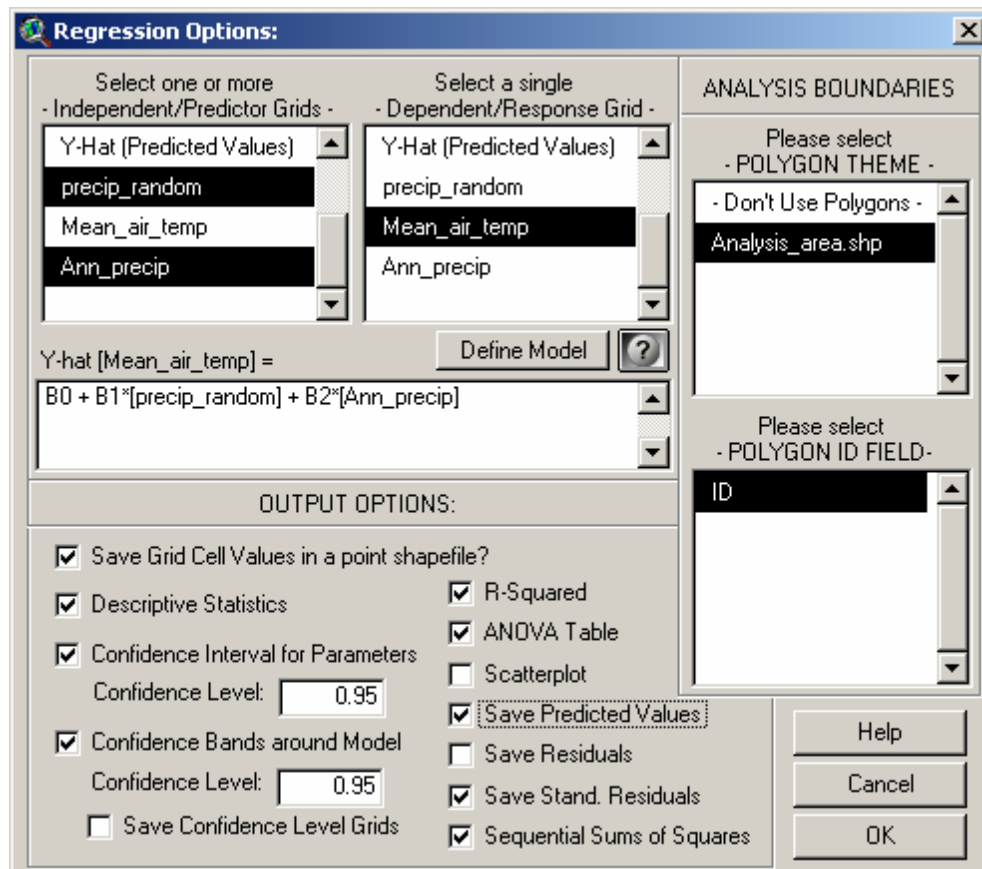
Draper, Norman R. and Smith, Harry. (1998) Applied Regression Analysis. 3rd ed. New York: John Wiley & Sons, Inc.; 706 pages. (Wiley Series in Probability and Statistics).

Neter, John; Wasserman, William; Nachtschiem, Christopher J.; and Kutner, Michael H. (1996) Applied linear statistical models: regression, analysis of variance and experimental design. 4th ed. Burr Ridge, Illinois: McGraw-Hill/Irwin; 1408 pages.

Linear Regression for Grids:

This function analyzes the linear relationship between one or more independent predictor grids and a dependent response grid, producing a grid of predicted values and similar outputs to those discussed in *Linear Regression for Themes and Tables* (p. 16). In this case, the predicted values, residuals and confidence levels will be generated as separate grids.

Click the  button in the view button bar to open the Regression Options window:



The "Regression Options" dialog box is shown. It has a title bar with a magnifying glass icon and the text "Regression Options:". The dialog is divided into several sections. On the left, under "Select one or more - Independent/Predictor Grids -", there is a list box containing "precip_random", "Mean_air_temp", and "Ann_precip". To its right, under "Select a single - Dependent/Response Grid -", there is a list box containing "Y-Hat (Predicted Values)", "precip_random", "Mean_air_temp", and "Ann_precip". Below these list boxes, there is a text field for the regression equation: "Y-hat [Mean_air_temp] = B0 + B1*[precip_random] + B2*[Ann_precip]". To the right of this field is a "Define Model" button and a help icon. On the far right, under "ANALYSIS BOUNDARIES", there are two sections: "Please select - POLYGON THEME -" with a list box containing "- Don't Use Polygons -" and "Analysis_area.shp", and "Please select - POLYGON ID FIELD -" with a list box containing "ID". At the bottom, under "OUTPUT OPTIONS:", there are several checkboxes: "Save Grid Cell Values in a point shapefile?" (checked), "Descriptive Statistics" (checked), "Confidence Interval for Parameters" (checked), "Confidence Bands around Model" (checked), "Save Confidence Level Grids" (unchecked), "R-Squared" (checked), "ANOVA Table" (checked), "Scatterplot" (unchecked), "Save Predicted Values" (checked), "Save Residuals" (unchecked), "Save Stand. Residuals" (checked), and "Sequential Sums of Squares" (checked). The "Confidence Level" for both the interval and bands is set to "0.95". At the bottom right are "Help", "Cancel", and "OK" buttons.

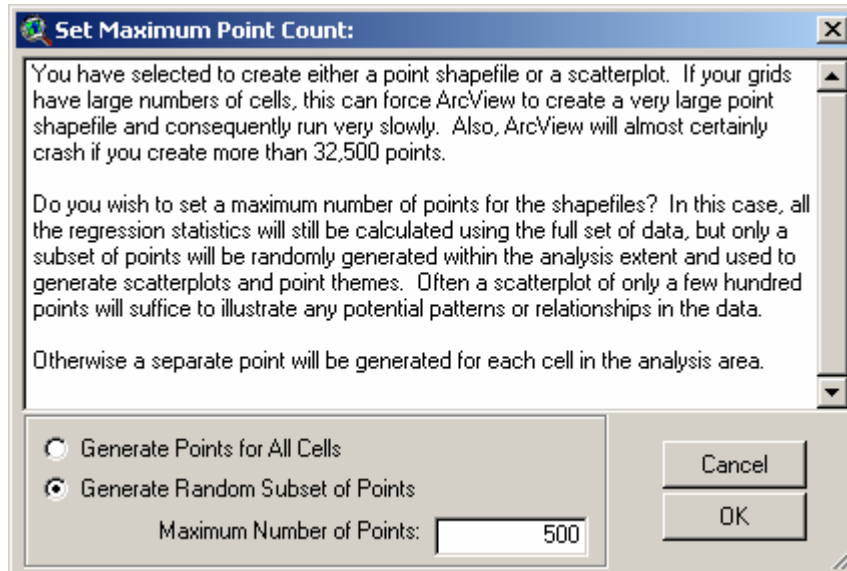
Select your grids of independent and dependent values from the appropriate listboxes and choose your output options. All selected output will appear in a report document identical to that described on p. 18. The various output options will produce the following data:

1. **Save Grid Cell Values in a Point Shapefile?** This will produce a point shapefile where each point represents the cell center of the regression grid. The attribute table will contain fields for all the output values that are selected (i.e. confidence bands, residuals, etc.)
2. **Descriptive Statistics:** Statistics include the cell count, mean, minimum, maximum, range, standard deviation, variance and sum of both the independent and dependent grid. If either grid is an Integer grid, then statistics will also include the majority (i.e. mode), minority and median values.
3. **Confidence Interval for Parameters:** Generates a range of parameter values that meet a specified confidence level. Confidence intervals are often preferred over simple predicted values because they convey how certain you are about the true parameter. See the discussion of parameter confidence intervals on page 21 for a description and example.

4. *Confidence Bands around Model*: Generates a region in which the true regression line probably lies, based on a specified confidence level. Remember that any statistical analysis on a sample only estimates a true population parameter, and therefore regression analysis only estimates the true linear relationship. Therefore these confidence bands provide a more realistic estimate of the regression relationship than the basic regression line. See the discussion of confidence bands on page 22 for a description and example.
5. *R-Squared*: A measure of how well correlated the two variables are, and can be interpreted to mean the proportion of the variation in the dependent variable that can be explained by variation in the independent variable (see page 22).
6. *ANOVA Table*: Short for Analysis of Variance, this table reports several details about the regression relationship including the probability that a true linear relationship exists (see page 22).
7. *Scatterplot*: The scatterplot provides a visual illustration of the linear relationship. The plot includes a point for each pair of independent and dependent values, with the Independent variable plotted along the X-axis and the Dependent variable value plotted along the Y-axis. **IMPORTANT**: This scatterplot can only be developed if a single predictor variable is used, although that single variable can have multiple orders or transformations. See page 33 for an example and discussion.
8. *Save Predicted Values*: The predicted value reflects the Y-coordinate of the regression line at any particular independent variable value or combination of variables, and is therefore the expected value of the dependent variable based on the regression model. For grid regression, these predicted values are saved in a new grid.
9. *Save Residuals*: The residuals are equal to the observed value minus the predicted value of the dependent variable, and reflect how much the dependent values deviate from the model. These are also saved in a new grid.
10. *Save Standardized Residuals*: Standardized residuals reflect the residual values transformed into Z-scores, with Mean = 0 and Standard Deviation = 1. Standardizing the residuals makes it easier to identify outlying values. These values are also saved in a new grid.
11. *Sequential Sums of Squares*: This function generates a table showing how well the regression equation fits the data as each term is entered into the model. This table can be used to assess how relatively important each variable is to the final model (see p. 23 for an example).

Restricting the Number of Points:

Two of the output options (*Save Grid Cell Values* and *Scatterplot*) potentially will generate a large number of points which may cause ArcView to run very slowly and possibly even crash (See the discussion on “GRD ERROR – Syntax error at or near symbol NL” in *Troubleshooting* on p. 90). Therefore this extension allows you to set a maximum number of randomly-generated points distributed over the analysis area:



IMPORTANT: Choosing this option will have no effect on the statistics generated! All statistics will be based on the full dataset. The number of points in the point dataset and scatterplot will just be smaller.

Setting Analysis Boundaries:

In many cases a user may not wish to conduct a regression analysis using the entire grid extent. More often the user will wish to restrict the analysis to a particular region (perhaps to within a country or watershed). The user has two ways available to set up an analysis boundary:

Using a Grid Mask: A grid mask is a grid that is used only to define what areas should be used in any analysis. All analyses will be restricted to those regions where the mask grid does not have a "No Data" value. Users can designate any grid as an analysis mask by opening a view, clicking the "Analysis" menu item, and then clicking "Properties..." to open the Analysis Properties window:

Click the drop-down box next to “Analysis Mask” and choose the grid to use.

Using a Polygon Theme: Users can also use a polygon theme to delineate analysis boundaries. In this case, the tool will conduct separate regression analyses for each selected polygon in the polygon theme. All analyses will be reported in the final report window, with each analysis identified by the polygon ID value.

Technical Note:

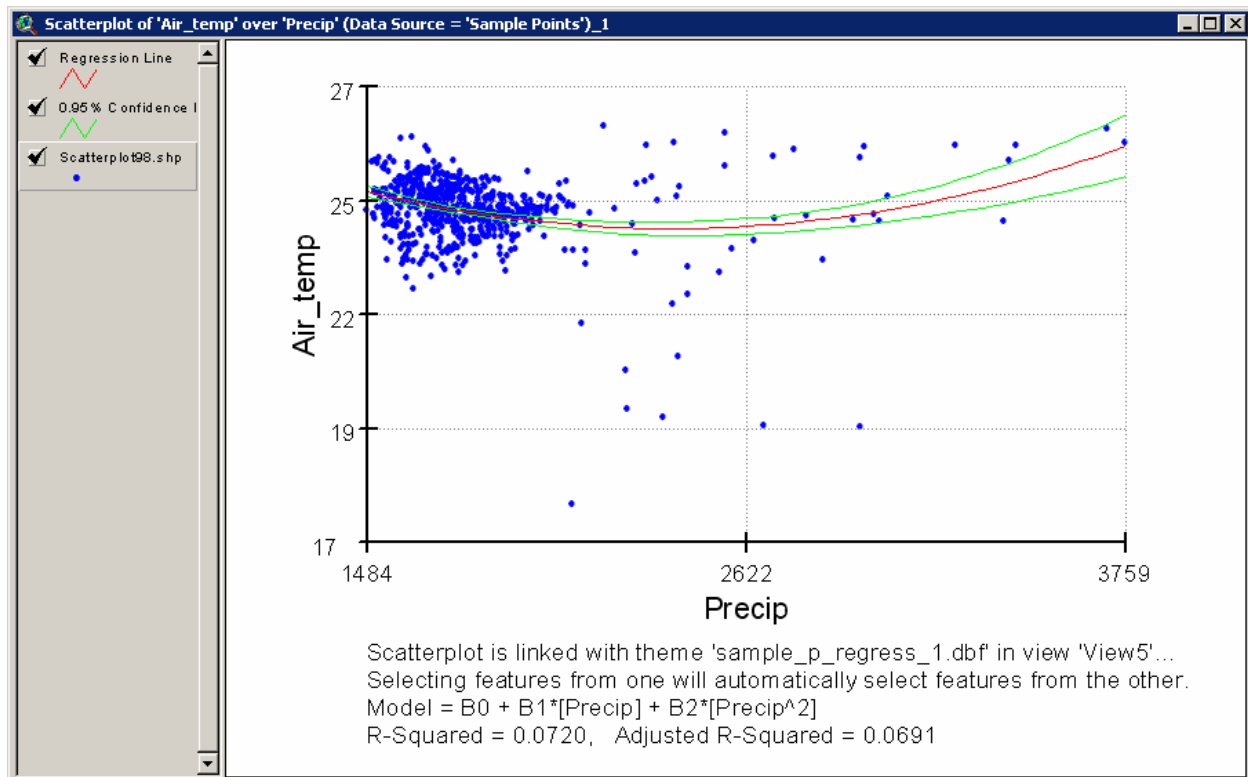
This tool only analyzes the regions of each grid that overlay each other and excludes all areas in which any grid has “No-Data” values. If the grids have different cell sizes, this tool uses the larger of the cell sizes for the analysis. If any mask grids were set prior to the analysis, those masks will remain in place during the analysis such that any areas excluded by the mask will also be excluded in this regression.


Technical Note regarding Scatterplots from Grid Data:

You may notice that the grid cell values intersected by your points in your view do not match up with the values recorded in the point attribute table. This can happen if your grids have different cell sizes or origins, and does not mean that the regression used incorrect values. As described in the *Technical Note* above, this extension uses the maximum cell size of all your input grids as the analysis cell size so grids with smaller cell sizes will be automatically resampled to larger cell sizes in the course of the regression. The values in your point attribute table reflect the resampled grid cell values.

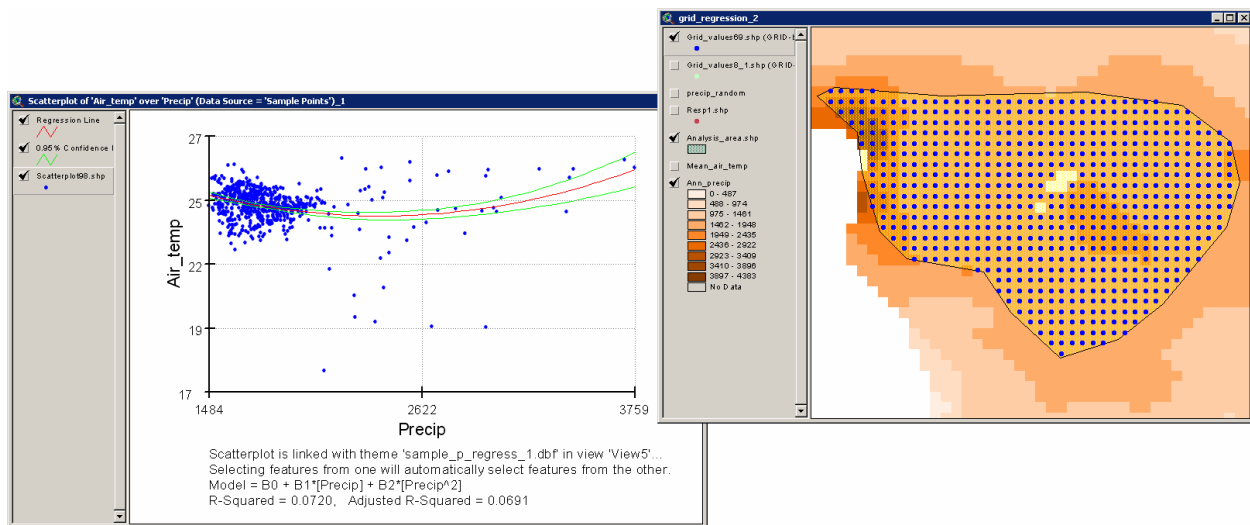
Generating Scatterplots

It is good statistical practice to analyze the usefulness of a regression analysis using all the factors of the relationship described previous sections, as well as to review a scatterplot of the data and regression line. In the example below there is extremely strong evidence of a linear relationship given the very small *P-value*, but the R^2 value shows that the linear relationship really does not explain the behavior of the dependent variable very well. Looking at the scatterplot of the output in the figure below illustrates the fact that the relationship is not strong. Furthermore, the air temperature variance appears to be change dramatically at about 2050 mm of precipitation, violating one of the assumptions of regression (see p. 8) and suggesting that data above and below that breakpoint should be analyzed differently.

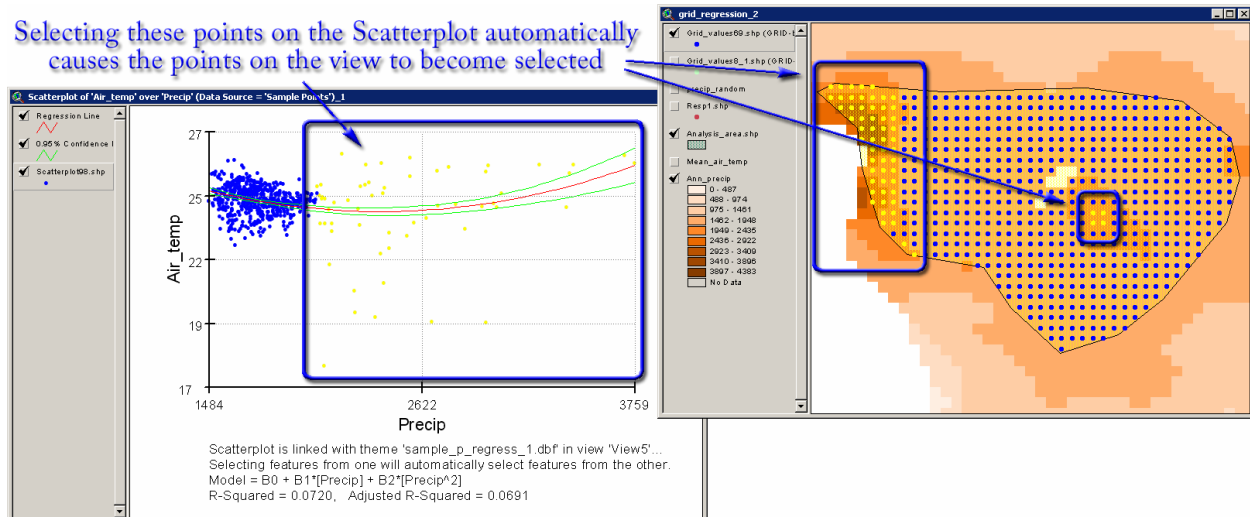


In cases where a user may wish to identify a particular feature in the scatterplot above, the user can click the Identify tool  to click on any of the points and view the attribute values.

Linked Scatterplot: The scatterplot will automatically be linked to the source data when it is generated, allowing you to quickly and easily query scatterplot points to find out where they lay on the landscape. For example, suppose our regression on the points below produced the corresponding scatterplot:



It appears that there is a dense cluster of points in the upper left corner of the scatterplot, and then a diffuse cloud of points over the rest of the scatterplot. We may wonder if there is some geographic component that may explain why these diffuse points are so distinct. If we select the points in the scatterplot, the corresponding points in the view will automatically select also:



Based on this quick analysis, we see that the diffuse points are clustered at the western edge of the analysis region, with an island in the middle. This suggests that there is some geographic component to mean annual air temperature that seems to be related to proximity to the western coast.

IMPORTANT: Please note that this extension only generates scatterplots if you use a single dependent variable in your model. The model may have multiple orders of this variable and still produce scatterplots, but this extension cannot calculate a scatterplot if there are more than 2 dimensions to the data. The following models may all be used to generate scatterplots:

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \beta_{111} X^3 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \beta_2 \ln X + \varepsilon$$

The following models may not be used to generate scatterplots:



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

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \beta_2 Z + \varepsilon$$

$$Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \beta_2 Z + \beta_{22} Z^2 + \varepsilon$$

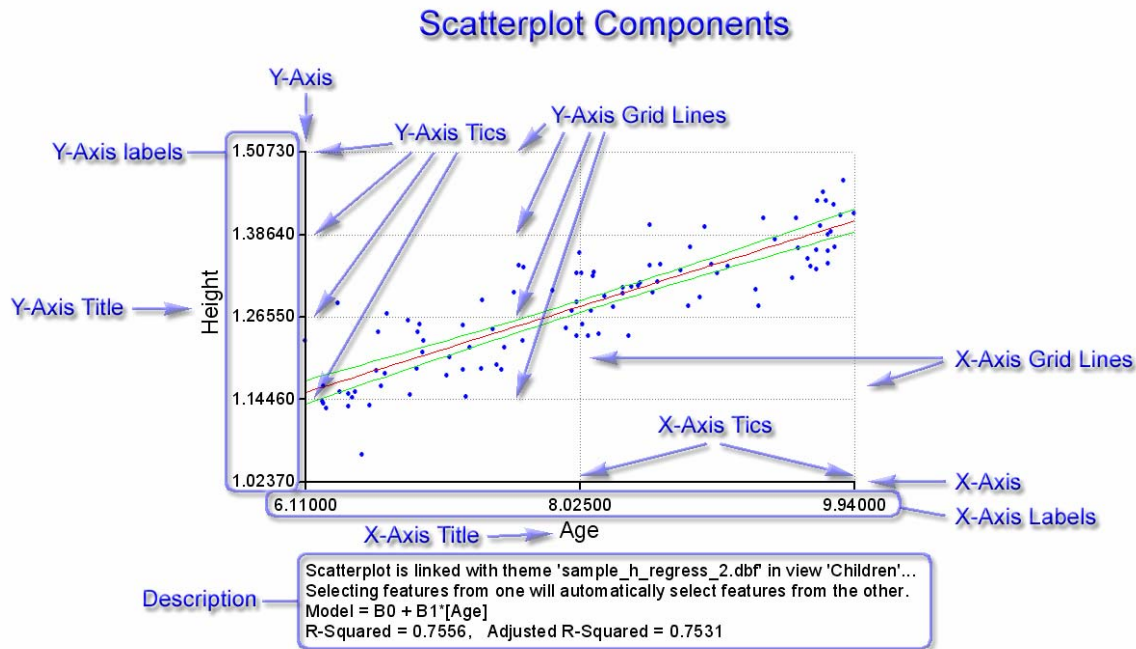
Predicting New Observations:

The scatterplot includes 2 buttons in the button bar which allow you to use your model to predict new observations. These functions are described in detail on p. 43, but briefly they are:


- 1)  Describe your model: This function generates a report with all the values you will need if you wish to predict new observations in some other software.
- 2)  Predict New Observations: This function predicts new observations using either specified predictor values, a table of predictor values, or predictor grids.

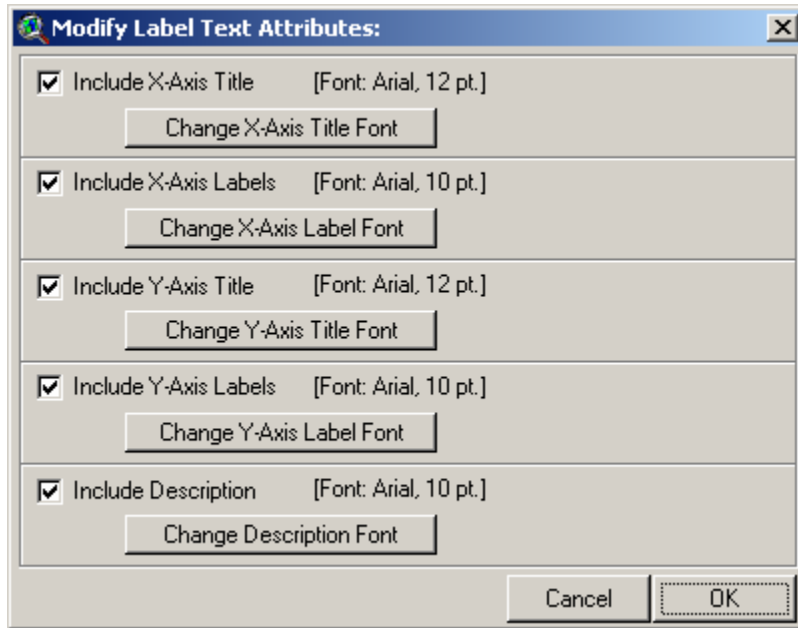
Altering the Appearance of your Scatterplot:

Scatterplots are also useful for adding to reports and manuscripts so this extension adds a few functions to enhance the appearance. All components of the scatterplot can be turned on and off and all text can be set to specific fonts and sizes. All functions are available from buttons and tools in your Scatterplot button and tool bar.

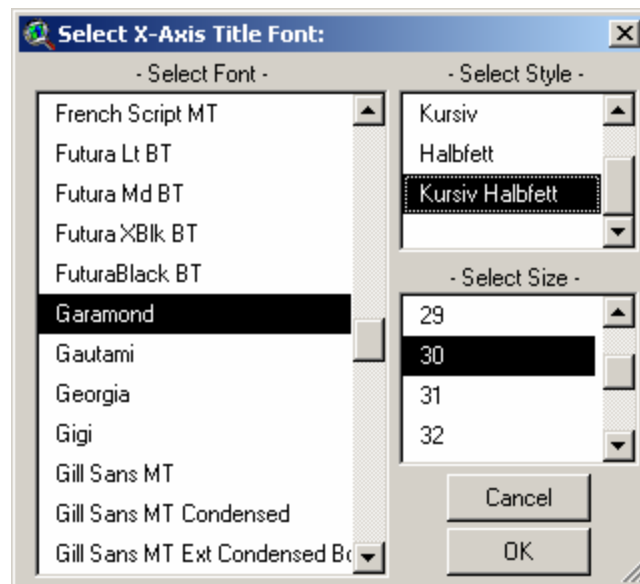


Modifying Text Fonts and Sizes:

Click the  button to begin modifying font attributes for all text in the scatterplot. The "Modify Text Label Attributes" dialog will open, listing all current fonts and whether they are currently turned on:




If you wish to change any of these fonts, simply click the appropriate button and specify the new font and size:

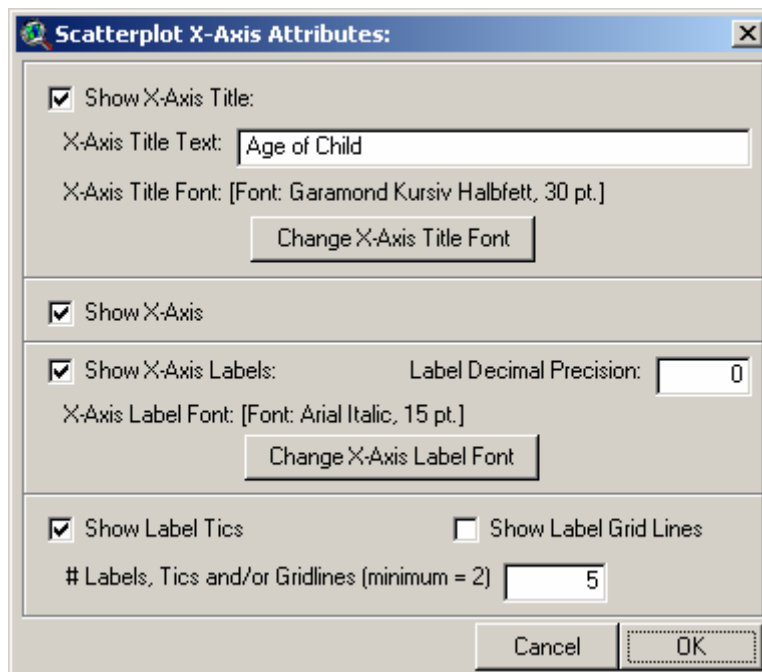


Note that you can also specify whether the various text components should be shown on the scatterplot. These text attributes can also be modified individually using the X-axis, Y-axis and Description buttons.

Modifying X-Axis Attributes:

This function allows you to set your X-Axis Title text and font, your X-axis label decimal precision and font, your number of X-axis sub-divisions, and whether you want any of these components turned off. Click

the  button to open the “Scatterplot X-Axis Attributes” window:



Scatterplot X-Axis Attributes:


☒ Show X-Axis Title:
 X-Axis Title Text:
 X-Axis Title Font: [Font: Garamond Kursiv Halbfett, 30 pt.]

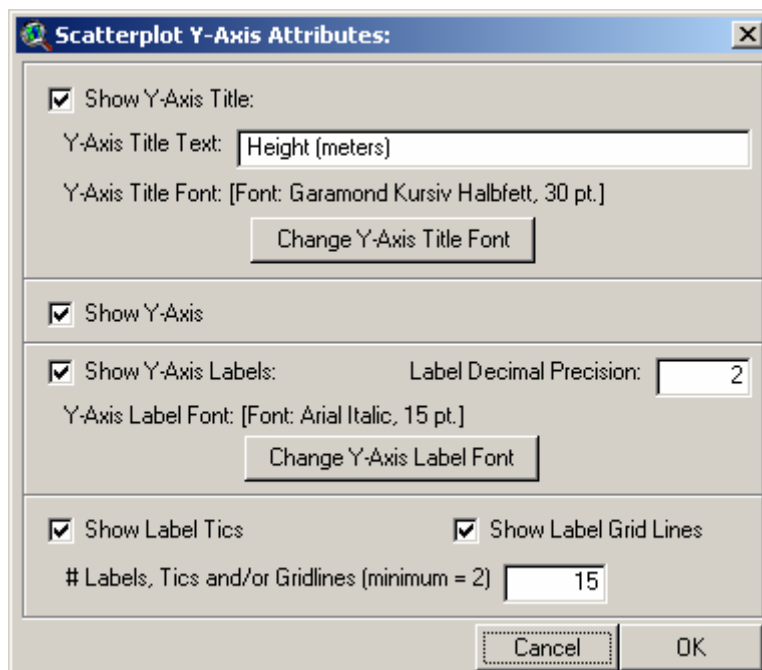
☒ Show X-Axis

☒ Show X-Axis Labels: Label Decimal Precision:
 X-Axis Label Font: [Font: Arial Italic, 15 pt.]

☒ Show Label Tics ☐ Show Label Grid Lines
 # Labels, Tics and/or Gridlines (minimum = 2)

Modifying Y-Axis Attributes:

This function allows you to set your Y-Axis Title text and font, your Y-axis label decimal precision and font, your number of Y-axis sub-divisions, and whether you want any of these components turned off. Click the  button to open the “Scatterplot Y-Axis Attributes” window:



Scatterplot Y-Axis Attributes:


☒ Show Y-Axis Title:
 Y-Axis Title Text:
 Y-Axis Title Font: [Font: Garamond Kursiv Halbfett, 30 pt.]

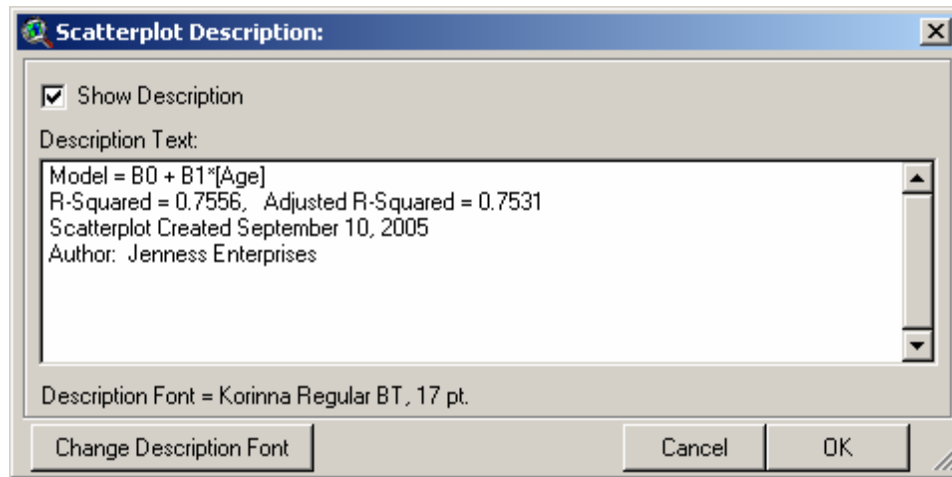
☒ Show Y-Axis

☒ Show Y-Axis Labels: Label Decimal Precision:
 Y-Axis Label Font: [Font: Arial Italic, 15 pt.]


☒ Show Label Tics ☒ Show Label Grid Lines
 # Labels, Tics and/or Gridlines (minimum = 2)

Modifying Description:

This function allows you to modify the description text and font, plus whether you want the description included with the scatterplot. Click the  button to open the “Scatterplot Description” window:



Refreshing Scatterplot:

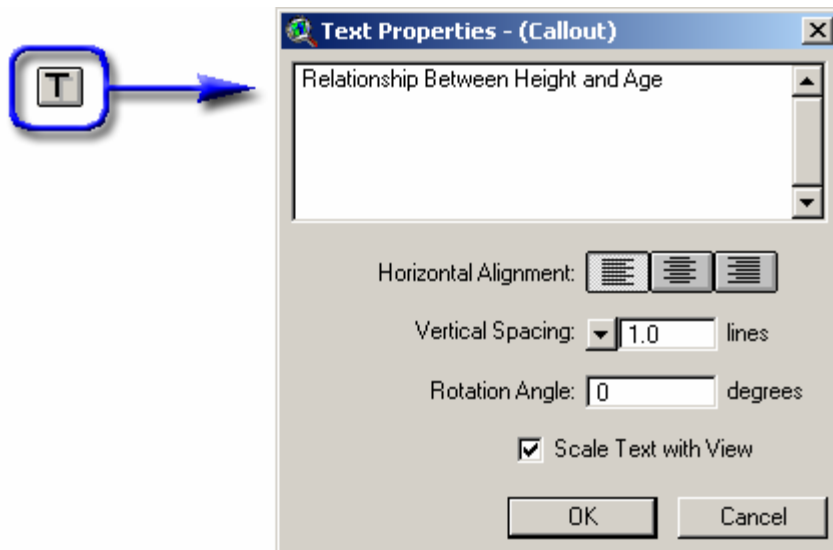
If you resize your scatterplot by dragging on a corner, or if you simply want to regenerate your scatterplot for whatever reason, click the  button to refresh it. The scatterplot will recreate all the graphic elements and redraw it to fit the current window size.

Adding additional components:

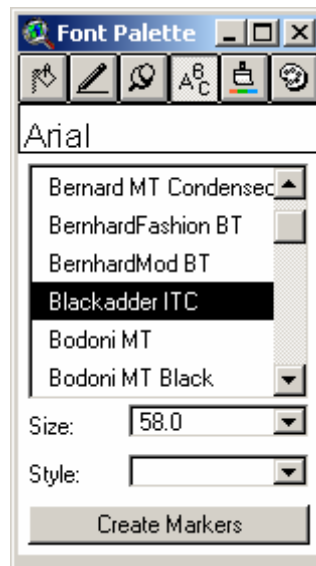
The Scatterplot document is based on the View document and shares many of its characteristics. For example, you can turn off the scatterplot points, regression line or confidence bands in the same way you turn off a theme in a view. You can also modify the symbology by double-clicking on them in the Table of Contents and using the standard ArcView legend editor.

Also as with views, you can add any graphic components you wish using the standard graphic tools in the tool bar. These are the same tools that are available in the View toolbar, except that they are arranged individually rather than in drop-down tool menus.

For example, if you wished to add a title with shadow effects to your scatterplot, use on the “drop-shadow text” tool to add the text.

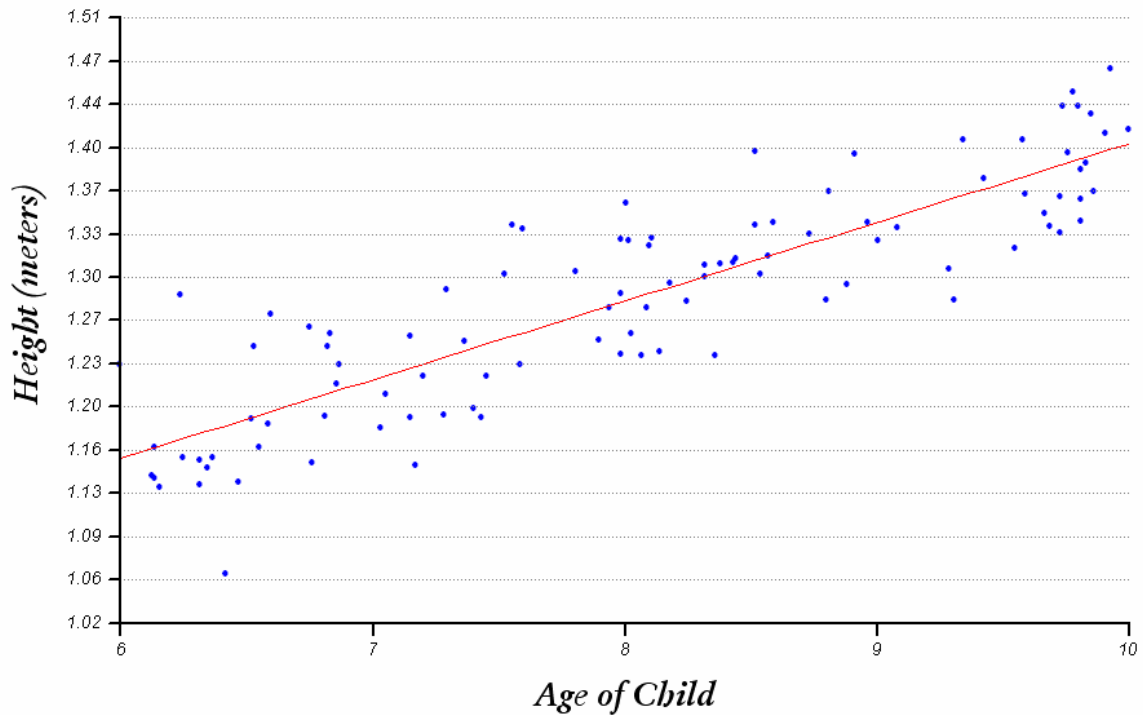


Next, click [Control]-P to open up the standard ArcView Symbol window (this window is also available in the “Window” menu) and set your font attributes:



Based on all the modifications in the examples above, the new scatterplot would look like this:


Relationship Between Height and Age

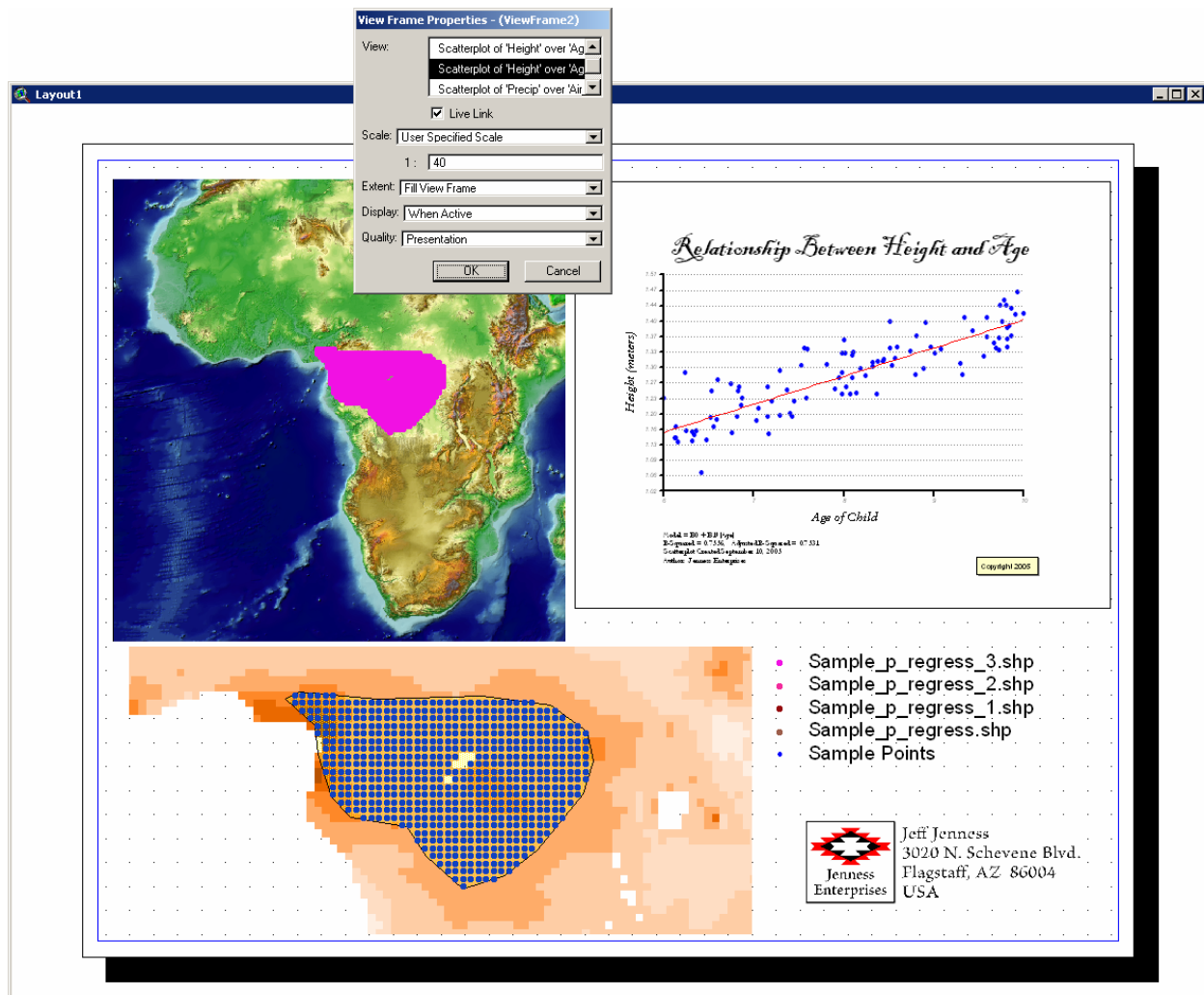


Model = $B_0 + B_1 \cdot \text{Age}$
R-Squared = 0.7556, Adjusted R-Squared = 0.7531
Scatterplot Created September 10, 2005
Author: Jenness Enterprises

Copyright 2005

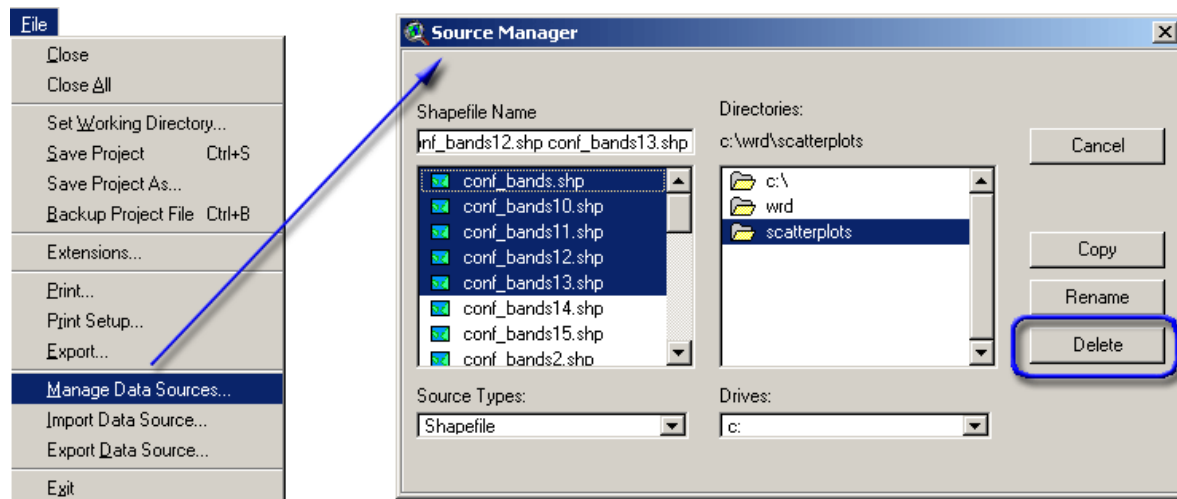
Adding your Scatterplot to a Layout:

Because the scatterplot is based on the View document type, it can be added to layouts in the same manner that Views are added to layouts. Simply click the  button in the layout, draw your box to hold the scatterplot, and then look for the scatterplot document in the list of views:



Beware of File Accumulation:


IMPORTANT: This process potentially generates 3 shapefiles every time a scatterplot is generated. Over time, this means you can accumulate a lot of files. We recommend that users periodically review the files in your project work directory and delete the ones that are not being used anymore. The easiest way to delete shapefiles is to use the "Manage Data Sources" menu item in the View "File" menu. Click that menu item, select the shapefiles you would like to delete, and click the "Delete" button. This function will automatically delete all the multiple files that make up each shapefile, and it will check to see if the file is currently in use in your project before deleting it.

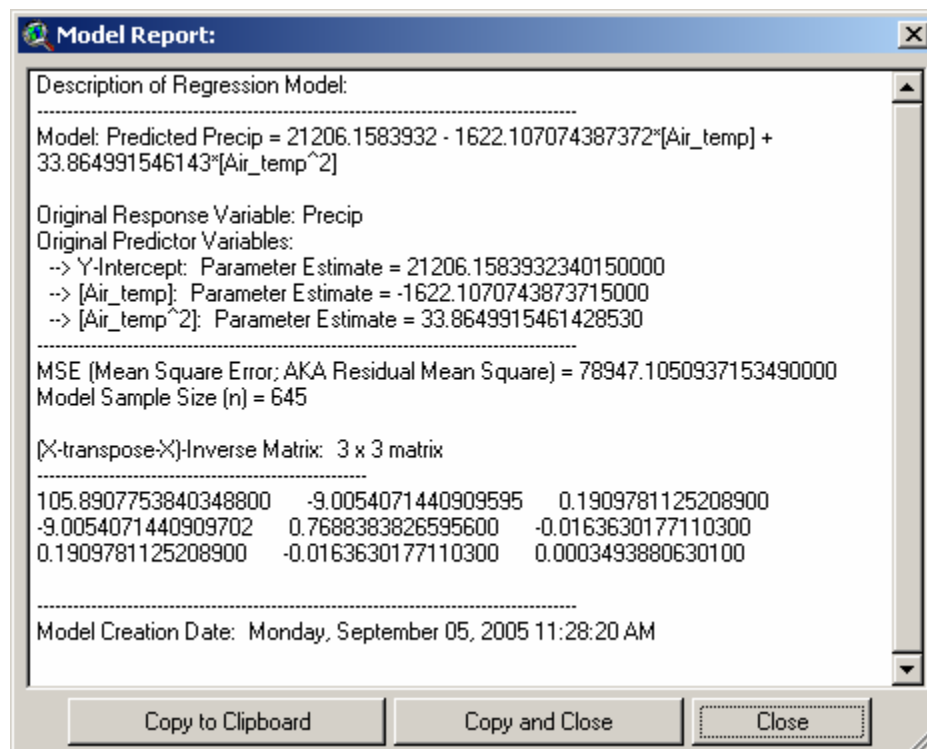


Describing and Predicting New Observations Using Your Model:

Describing your Model:

The regression report includes all the parameter estimates and model characteristics that are typically necessary to evaluate and describe your regression model. If you wish to predict new observations, you only need to plug in the appropriate predictor values into the equation (or use tools available in this extension; see below). However, if you wish to include a measure of the uncertainty along with the new observation, then you will need additional data not available in the regression report.

From either the Report or Scatterplot window, click the  button to view a description report of the model. This report will describe the model used to generate that report or scatterplot:



To predict new observations, you will only need the parameter estimates. To calculate the standard deviation or confidence intervals around those new predictions, you will also need the Mean Square Error (MSE), the sample size (n) and the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix (denoted as "[X-transpose-X]-Inverse") calculated from the original dataset. From Neter et al. (1996:235), the confidence limits around a predicted new observation are calculated as:

$$\hat{Y}_{new} \pm t_{(1-\alpha/2; n-p)} s\{\text{pred}\}$$

where \hat{Y}_{new} = new predicted observation

$t_{(1-\alpha/2; n-p)}$ = t -value at confidence level $(1 - \alpha)$, with $(n - p)$ degrees of freedom

n = original sample size

p = number of parameters, including y-intercept

$s\{\text{pred}\}$ = standard deviation of new observation

The standard deviation of the new observation is based on the mean square error of the original model, the original predictor values and the predictor values of the new observation:


$$s\{\text{pred}\} = \sqrt{MSE \left(1 + \mathbf{X}'_{new} (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_{new} \right)}$$

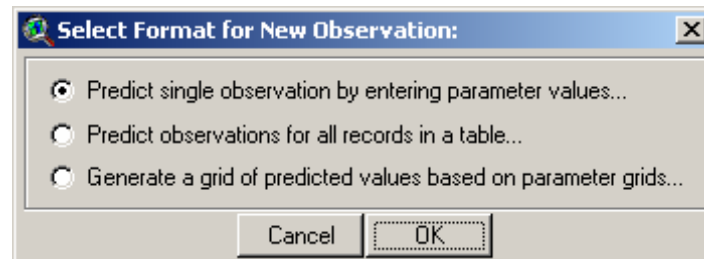
where \mathbf{X}'_{new} = Vector of new predictor variable values

Predicting New Observations:

This extension includes functions to help you predict new observations using your regression model, including confidence intervals around those predicted observations. Predictions may be made in 3 ways:

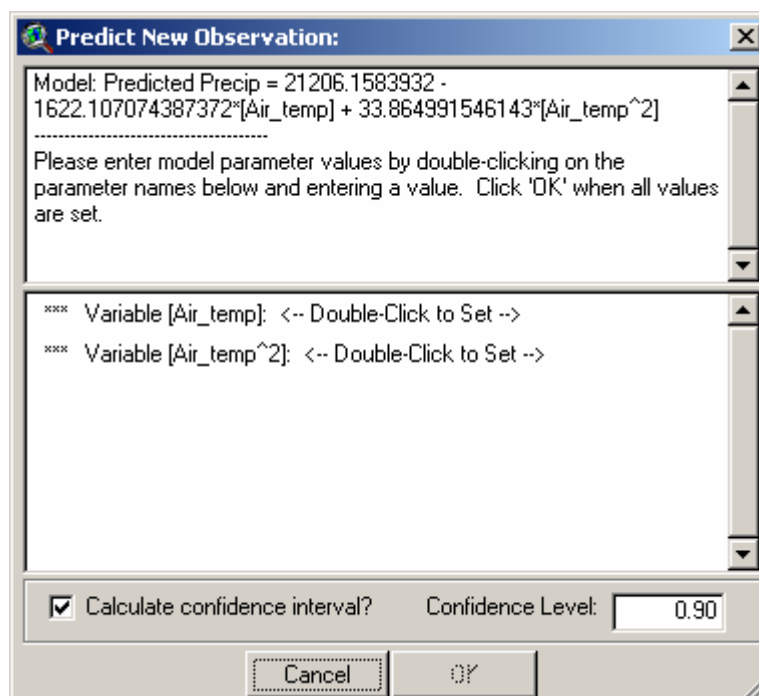
- 1) Predict a single new observation by entering the predictor variable values.
- 2) Predict a series of new observations using a table of predictor values.
- 3) Generate a grid of new observations using a set of predictor grids.

These options are available by clicking the  button on either the Report or Scatterplot document. The predictions will be based on the model associated with that report or scatterplot. After clicking the button, you will be asked which type of new observations you would like to predict:



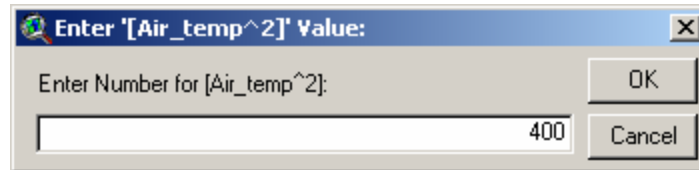
Predicting a Single New Observation:

Select the “Predict single observation by entering parameter values...” option and click ‘OK’. You will next be asked to identify your predictor variable values:

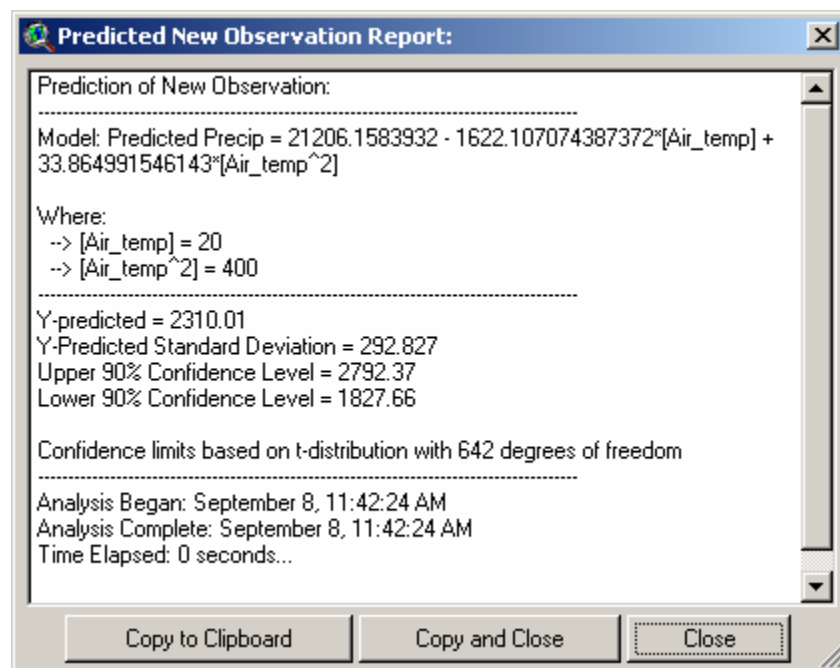


The top of the window describes the model that will be used, and the central portion lists the variables that must be assigned values. If you wish to generate a confidence interval around the predicted new observation, check the box at the bottom and enter a confidence level.

Assign values to your variables by double-clicking the variable name in the list. Another window will appear asking you to enter the number, and afterwards the main dialog will indicate that you have assigned a value to that variable:



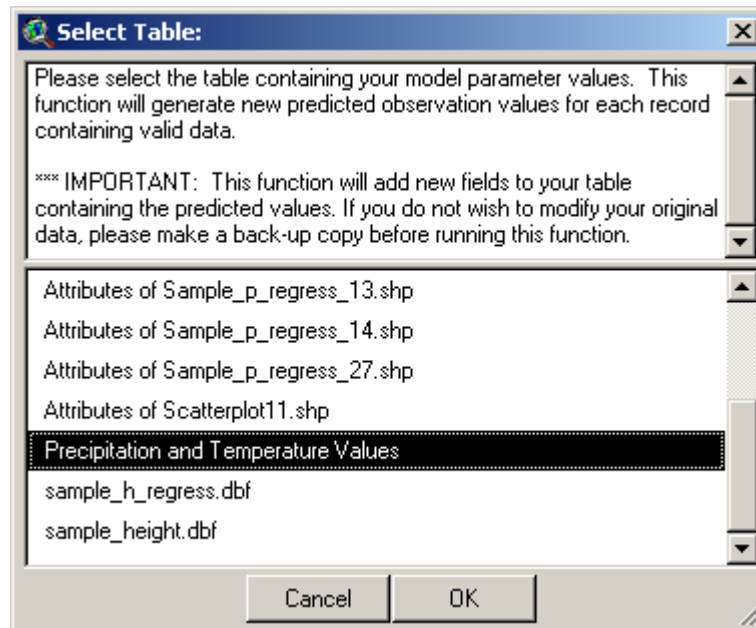
After you have entered all your values, click 'OK' on the main dialog to calculate the predicted new observation:



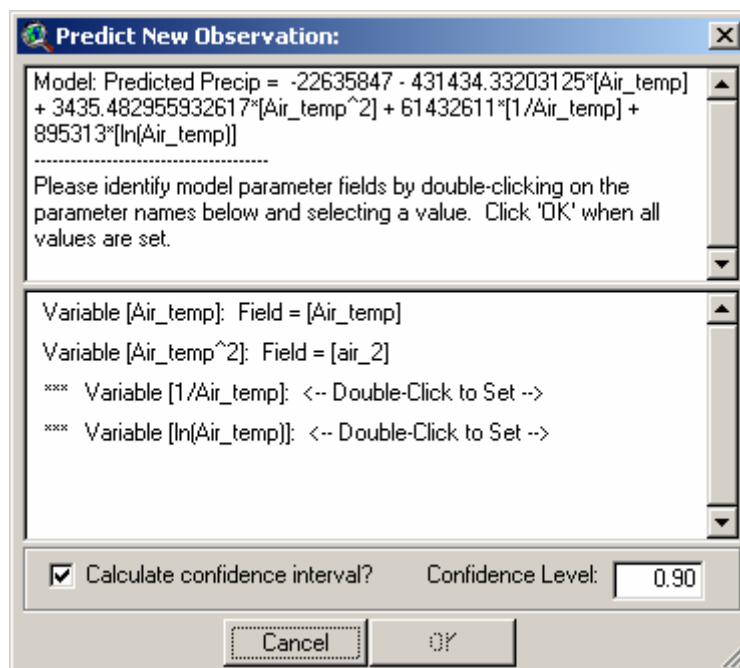
Predicting New Observations for All Records in a Table:

IMPORTANT: This function will add new fields to your table containing your predicted new observations and the standard deviation of each new observation, plus (if desired) fields for the upper and lower confidence limit for that observation. If you wish to leave your original data unmodified, make a backup copy of it before running this function.

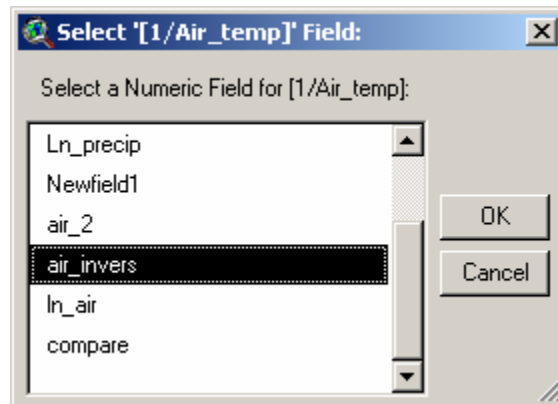
Select the "Predict observations for all records in a table..." option and click 'OK'. You will next be asked to identify the table containing your predictor variable fields:



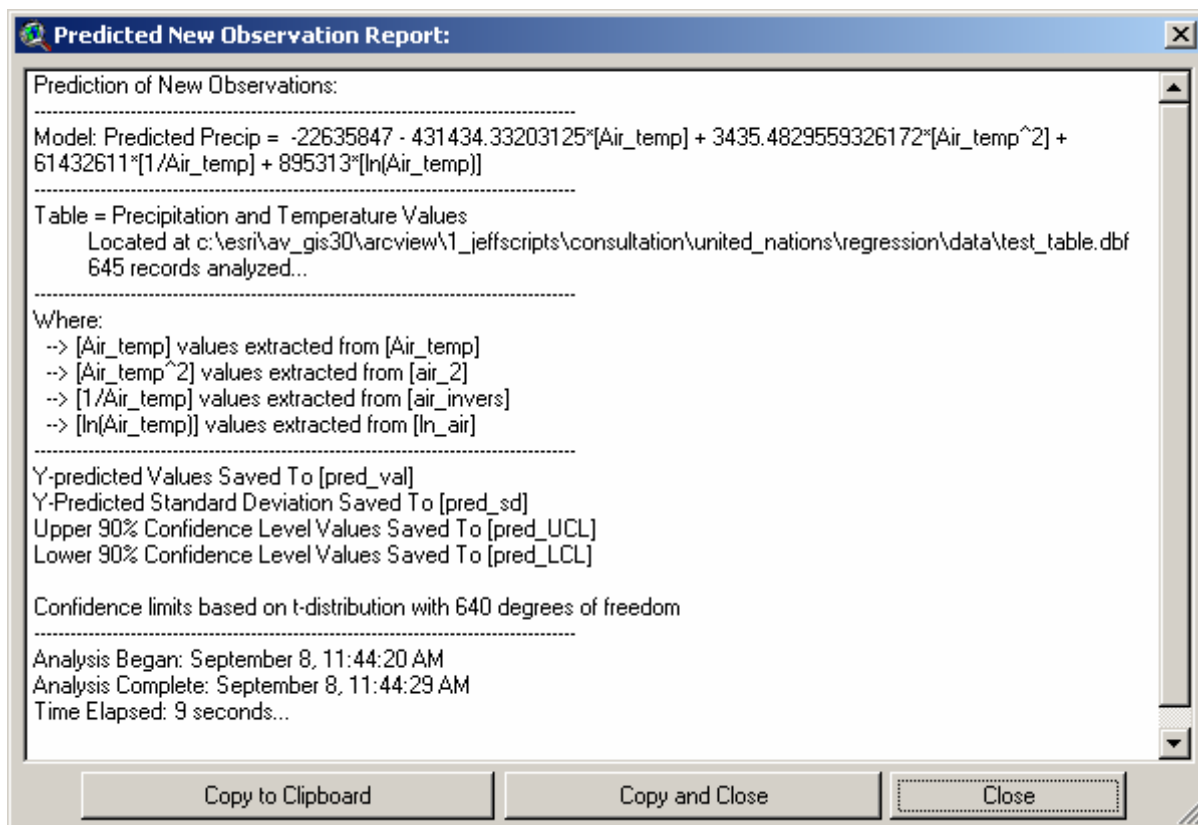
Click 'OK' and specify which fields correspond with each predictor variable. Note that even if your model contains multiple transformations of a single variable, your table must contain separate fields for each transformation.



Simply double-click on the predictor variable name and select the appropriate field from the list:



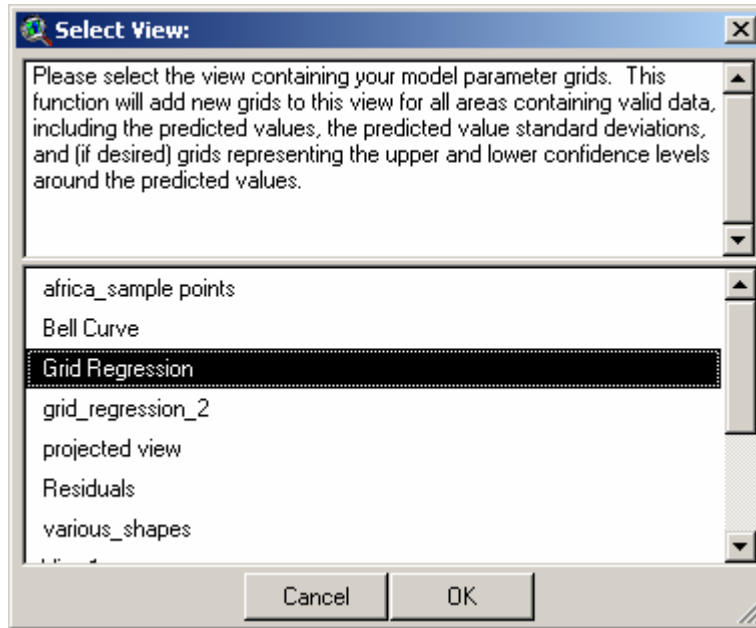
After you have specified all your predictor variable fields, click 'OK' on the main dialog to calculate the predicted new observations:



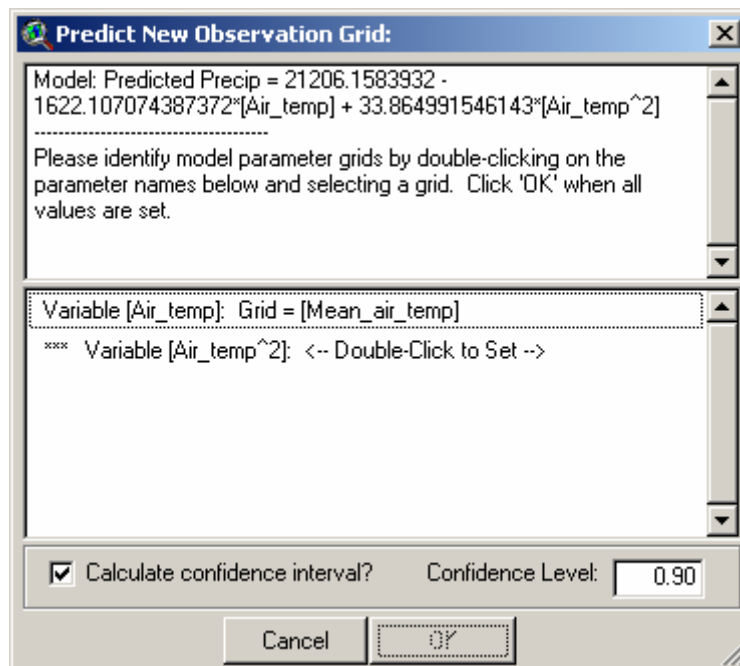
Generating Grids of New Observations:

This function will add new grid themes to a view containing your predicted new observations and the standard deviation of each new observation, plus (if desired) grids of upper and lower confidence limits for each new observation.

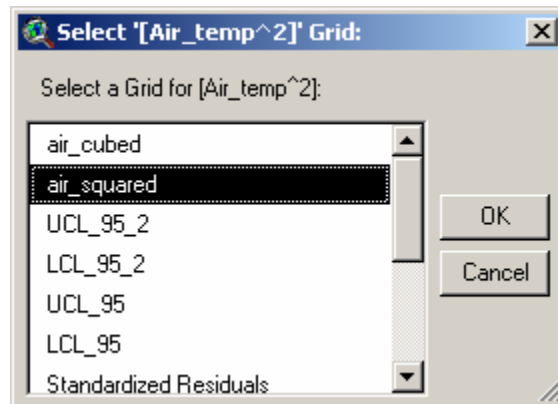
Select the "Generate a grid of predicted values based on parameter grids..." option and click 'OK'. You will next be asked to identify the view containing your predictor variable grids:



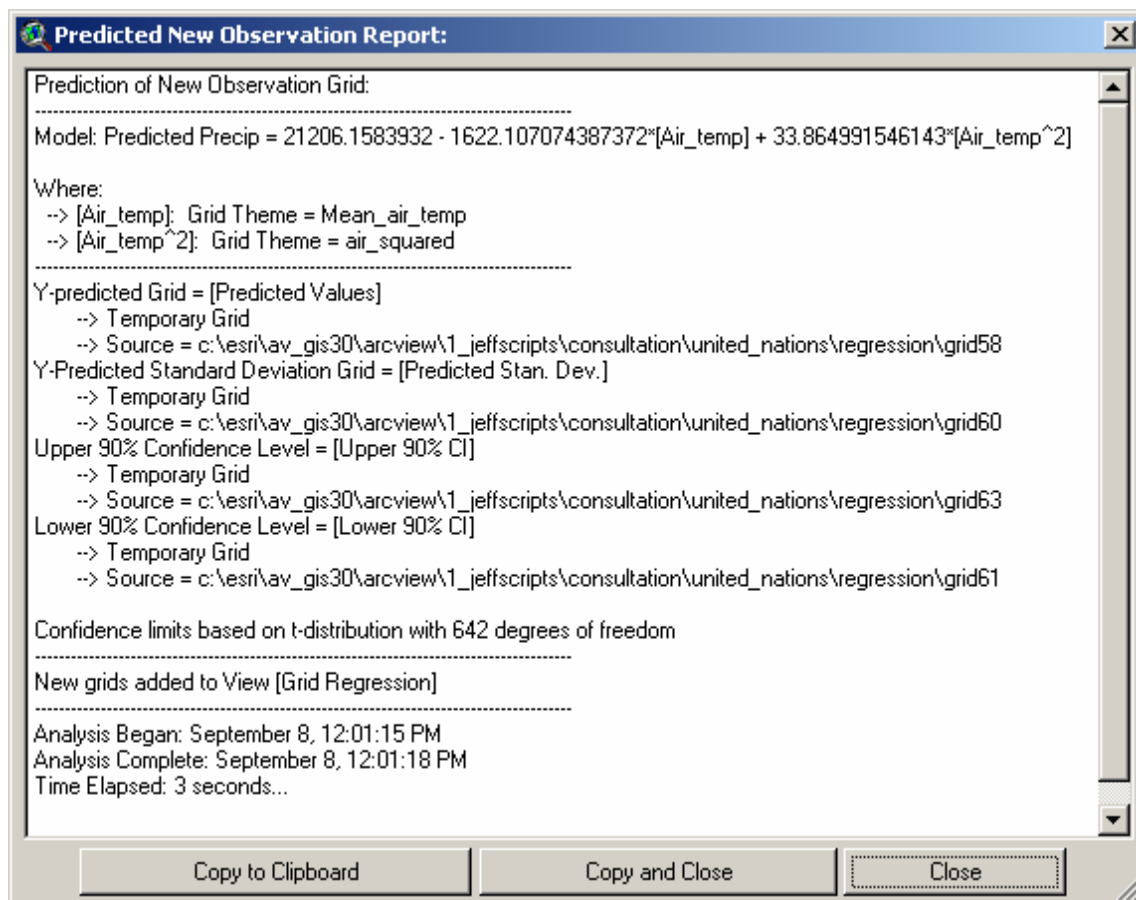
Click 'OK' and specify which grids correspond with each predictor variable. Note that even if your model contains multiple transformations of a single variable, your view must contain separate grids for each transformation.



Simply double-click on the predictor variable name and select the appropriate grid theme from the list:

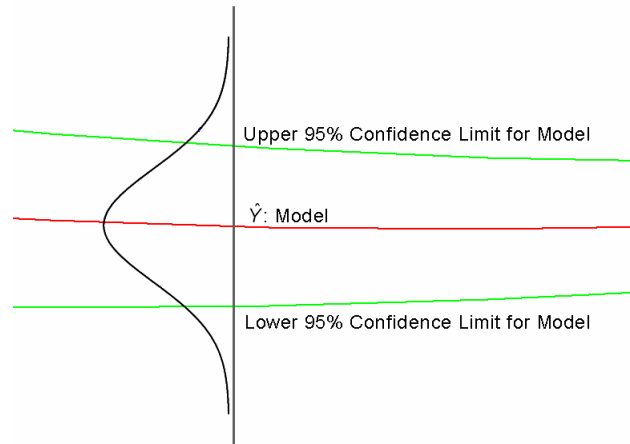


After you have specified all your predictor variable grids, click 'OK' on the main dialog to calculate the grid of predicted new observations:



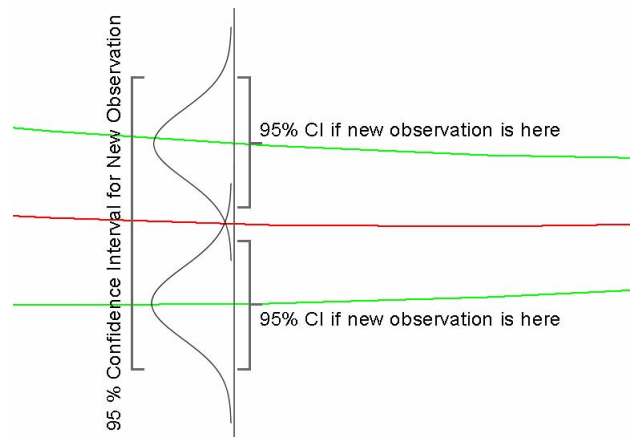
Why are the confidence intervals so large?

You may notice that the confidence interval around your predicted new observations are surprisingly large, and certainly larger than the confidence bands around the original model. This is because the predicted value is a random variable and therefore has an additional degree of uncertainty added on top of the original model uncertainty. For example, the uncertainty of the original model can easily be observed by looking at the confidence bands around the regression line:



The true model could lie anywhere, and the 95% confidence level can loosely be interpreted to mean that there is approximately a 95% chance that the true model lies between the bands.

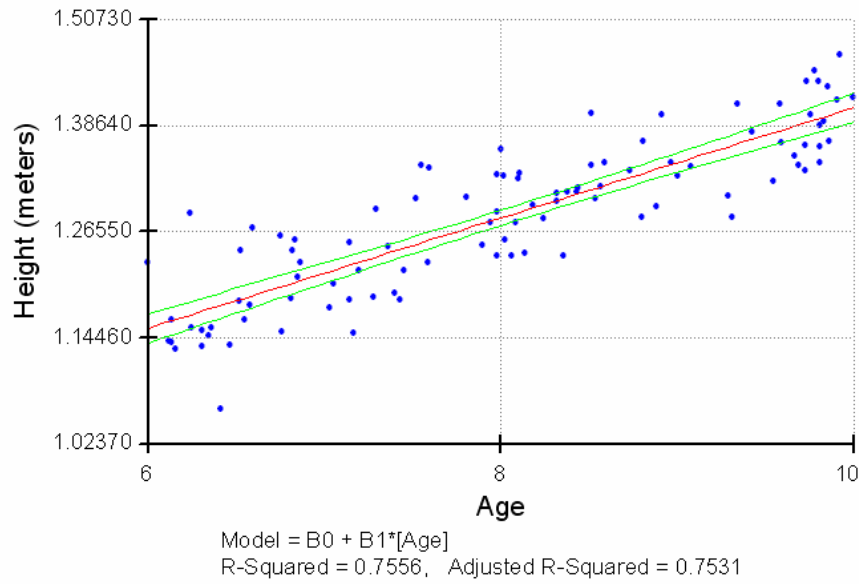
When we predict a new observation using some confidence level, then the new observation will always be located on the regression line. However, because of the uncertainty about the regression line, we must consider that the new observation itself could lie anywhere between the confidence bands, and therefore we must combine the uncertainty about the new predicted observation with the uncertainty about the model, which automatically expands the original confidence interval:



The confidence intervals are relatively large because the new predicted observation could lie anywhere between the combined confidence intervals around the original model confidence interval.

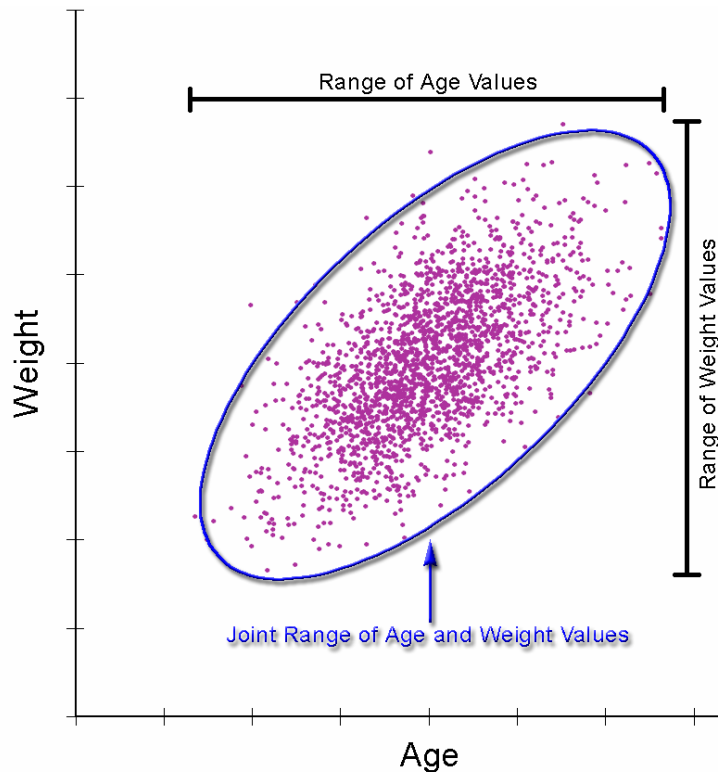
Beware of predicting outside the range of predictor values:

You should always be wary of predicting new observations using independent variable values that are outside the range of your original dataset. Just because you observed a statistically significant and predictable relationship between your predictor and response variables does not mean that the relationship continues over the entire possible range of predictor variables. As a simple example, consider the relationship between age and height for children ages 6-10:



There is a clear and well-defined linear relationship here, and it appears that we could use this model to make reasonably accurate predictions of a child's height based on their age. The model is a good one for children between 6 and 10. We run into problems, however, if we try to predict outside this age range. Because children typically stop growing in late adolescence, this model is not appropriate for older ages. Using this model, we would predict that a 30-year old would be 2.7 meters (8 ft. 10 in.) and a 60-year old would be 4.7 meters (15 ft. 5 in.). These predictions are clearly nonsense.

The problem with predicting outside the range of the data is easy to visualize using a simple linear regression example such as that above. Unfortunately this problem can be difficult to identify and avoid when using more complex models. If we wanted to predict a child's height based on both their age and weight, we need to be aware that the joint range of age and weight may be difficult to define. For example, because children tend to weigh more as they grow older, the joint distribution of age and weight may look something like this:



The joint range of Age+Weight combinations is not as clearly defined as the simple range of Age or Weight individually, but we can approximate it by drawing an ellipse around the general cloud of points. We can then consider the region inside the ellipse to have been sampled and therefore more appropriate for predicting new values. The region outside the ellipse has not been sampled and therefore predictions should be made with caution.

A key point to note here is that we may pick a combination of age and weight that is within the respective ranges of sampled values, but this combination may not be within the joint range of sampled values. For example, we never sampled a child at the minimum of the age range and the maximum of the weight range, so it may be inappropriate to predict a child's height based on such a combination of values.

When working with multiple predictor variables, it can be difficult to know whether a particular combination of values lies within the joint range of sampled predictor values. This extension does not offer any tools to do so, either, so you simply need to be aware of the issue and consider whether your sample data are appropriate for your prediction.

A Warning About Regression with Spatial Data:

Although regression is a useful and powerful tool, it should be noted that some aspects of it often violate basic regression assumptions. This problem is especially true with grid regression. The end results of these violations would likely be that your estimated parameters (i.e. your slope, y-intercept and R-square values) are probably a bit off, and in particular your true R-square value is likely to be less than the calculated value. This may not be a problem in many cases because regression is still a good method for identifying relationships between our independent variable and our predictor variables, and therefore helps us to predict what our independent variable will likely be doing in different areas based on our predictor variables. We do, however, have to be careful to report that there is some uncertainty about our model because of these violations, and be cautious when our R-squared value is near the limits of what we consider to be significant.

In particular, the violations are:

1. *With grid regression, we did not measure at every point:* The fact that we are using grids usually implies that we know more about our independent variables than we actually do. We are regressing data using sample points that completely cover the entire area, and it is rare that we have actually measured all our variables at every one of these sample points. In fact, grids are generally created by some interpolation method in which values are only measured at a few points, and the rest of the region is estimated (or interpolated) based on the values at these sample points. Different grids may even be generated from different sets of sample points, at different resolutions, or by different interpolation methods. Therefore we are often not as certain of the true variable values at each sample point as we would be if we actually measured at that point.
2. *Lack of Independence:* Most statistical techniques assume that each sample point is independent of the others, such that the values you measure at that point are completely unrelated to those points around it. This is not the case with most spatial phenomena, however, and the problem is even more pronounced with grid data. In fact, the interpolation methods often used to generate grids rely on the fact that locations near a point are likely to be more similar to that point than locations farther away, and the interpolation process uses that relationship to estimate what the values should be in the locations that weren't measured. The concept that points close to each other are often more similar than points that are far away is referred to as "spatial autocorrelation", and the degree to which a dataset is spatially autocorrelated can actually be useful information in its own right.

One method to avoid spatial lack of independence is to build a semi-variogram and identify the spatial separation distance at which spatial autocorrelation drops to an acceptable level. This extension currently does not offer that option, although I would like to include it in a future revision.

Additional Reading:

For those who would like to learn about regression in depth, there are many texts available that cover it thoroughly. Two such texts that the author recommends are:

- Applied Linear Statistical Models, 4th ed. by Michael H. Kutner, Christopher J. Nachtschiem, William Wasserman and John Neter (1408 pages, published by McGraw-Hill/Irwin, 1996)
- Applied Regression Analysis, 3rd ed. By Norman R. Draper and Harry Smith (706 pages, published by Wiley-Interscience, 1998)

There are also some interesting new developments in the field of spatial regression and the author recommends the following texts for discussion:

- Geographically Weighted Regression: The analysis of spatially varying relationships. A. Stewart Fotheringham, Chris Brunsdon and Martin Charlton. (269 pages, published by John Wiley & Sons Ltd., 2002).
- Quantitative Geography: Perspectives on spatial data analysis. A. Stewart Fotheringham, Chris Brunsdon and Martin Charlton. (270 pages, published by Sage Publications, 2000).

For those who would like to learn more about spatial autocorrelation, some of the classic references are:

- Spatial Statistics, by Brian D. Ripley (252 pages, published by Wiley Series in Probability and Mathematical Statistics, 1981)
 - Spatial Autocorrelation, by A.D. Cliff and J.K. Ord (178 pages, published by Pion Limited, 1973)
 - Spatial Processes: Models and Applications, by A.D. Cliff and J.K. Ord (266 pages, published by Pion Limited, 1981)
-


Manually Transforming Variables:



The model-building function in this extension allows you to apply several transformations to your predictor variables automatically (see p. 16). However, it provides no functions to automatically transform your response variable. If you need your response variable transformed, you will need to do this manually prior to running the extension. For example, SPSS (1999) offers the following pre-defined curve-fitting model designed to fit S-shaped curves:

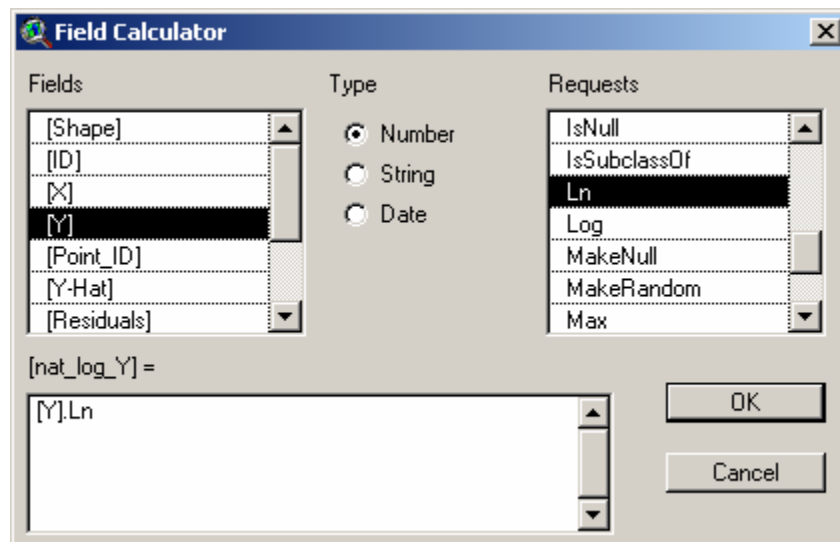
$$\ln Y = \beta_0 + \beta_1 \frac{1}{X} + \varepsilon$$

You cannot apply this model automatically using this extension because there are no means to transform \hat{Y} . However, you can manually transform \hat{Y} beforehand and then run the extension using the transformed variable.

Transforming Variables in Themes and Tables

This function requires that you create a new field in your table, so you must open your theme attribute table using the  button if you are using a shapefile or theme.

- 1) Set your table to *Editable* by clicking the “Table” menu, then “Start Editing”.
- 2) Add a new field to your attribute table by clicking the “Edit” menu, then “Add Field”. Make sure your new field is numeric and that it has a sufficient number of decimal places.
- 3) Clear any current selection in your table by clicking the Clear Selection button .
- 4) Select your new field by clicking on the field name at the top of the table. It should have an inset appearance.
- 5) Click the Calculate button  to open the Field Calculator. If your Response field was named [Y] and your newly-created field was named [nat_log_y], and you wished to perform a natural log transformation, then you would fill out the Field Calculator dialog as follows:



- 6) The list of “Requests” contains several other transformations you can apply to your field.

HINTS: If you wish to raise your response values to a power, then the calculation string would be:

$$[Y]^{(aPower)}$$

If you wish to transform by e^Y , then use the following calculation string:

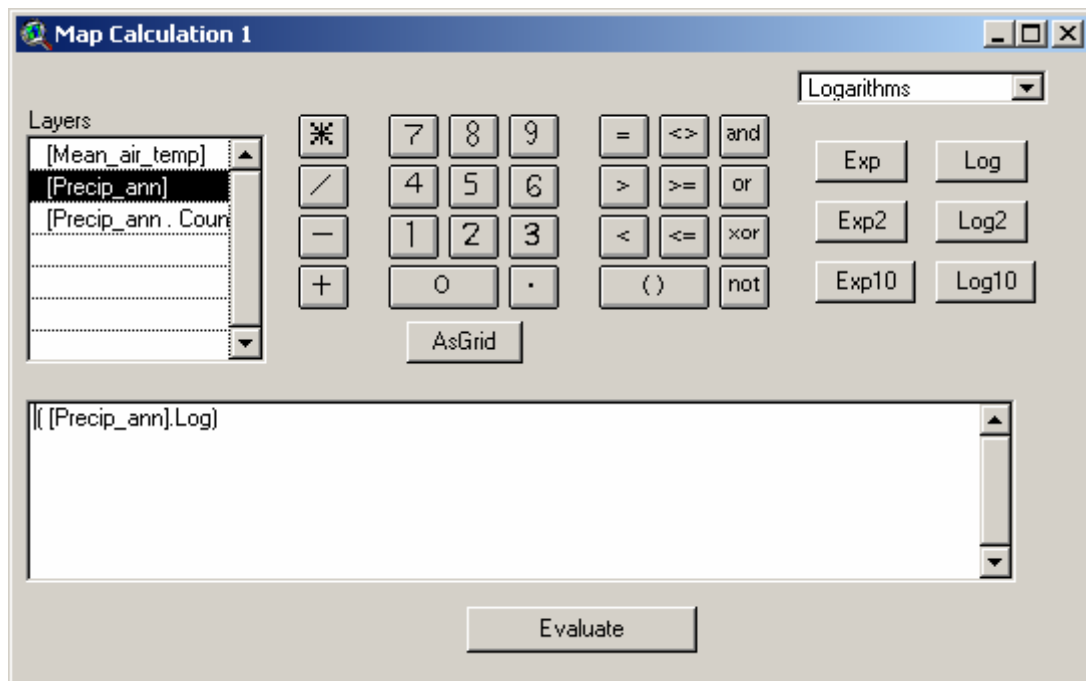
$$\text{Number.GetEuler}^{([Y])}$$

- 7) **IMPORTANT:** Make sure your values are appropriate for the transformation! For example, if you apply a natural log transformation to negative numbers, you will get null values which cannot be used in regression and which will trigger a “Singular Matrix” error (see *Troubleshooting* on p. 90).
- 8) Save edits by clicking the “Table” menu, then “Stop Editing”.

Transforming Grids:

This transformation will create an entirely new grid, and the new grid should be used in the regression function in place of the original response grid.

- 1) Make sure your response grid is in your view.
- 2) Click the “Map Calculator...” menu item in the “Analysis” menu.
- 3) If your response grid is named “Precip_ann” and you wish to perform a natural log transformation, fill out the Map Calculator dialog as follows:



- 4) There are several other transformation options available by clicking on the drop-down box in the upper right corner of the window.

HINTS: To raise your grid values by a power, use the following calculation string:

$$[\text{Your_grid}]^{(aPower)}$$

If you wish to transform by e^Y , then use the following calculation string:

$$[\text{Your_grid}].\text{exp}$$


- 5) After the map calculator finishes, a new grid will be added to your view. **IMPORTANT:** This is a temporary grid and will only be saved permanently if you save your project or specifically save your dataset with the “Save Dataset” menu item (in the “Theme” menu).

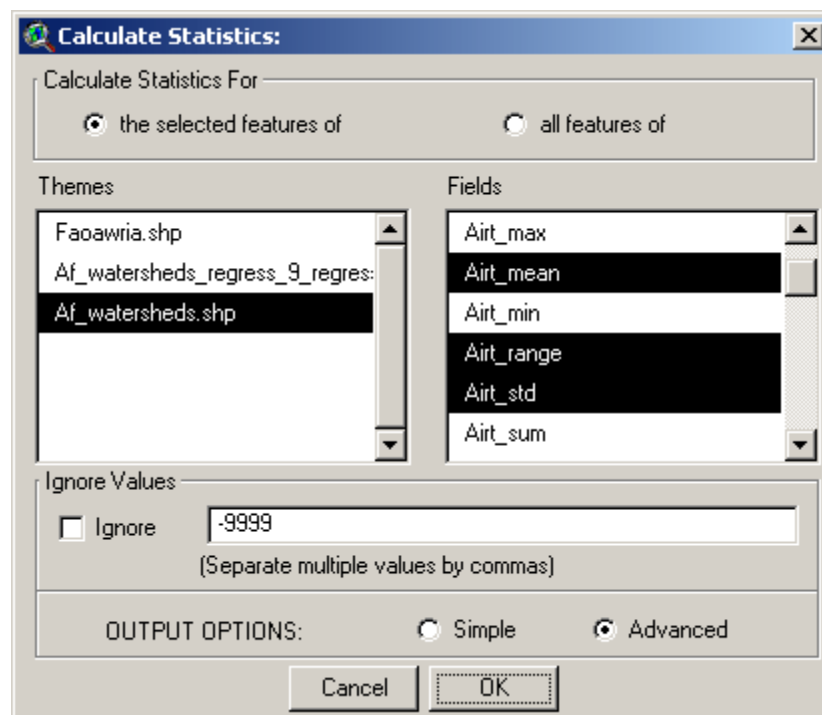
If you want to inspire confidence, give plenty of statistics. It does not matter that they should be accurate, or even intelligible, as long as there is enough of them.
Lewis Carroll

Field Summary Statistics:

This tool provides functions similar to those available in the basic ArcView “Statistics...” options under the standard “Field” menu item in the Table menu, with the exception that there are both more options and a higher level of precision used for any calculations. The tool may be used to generate statistics on either a theme in a view or a field in a table.

Summary Statistics on a Theme:

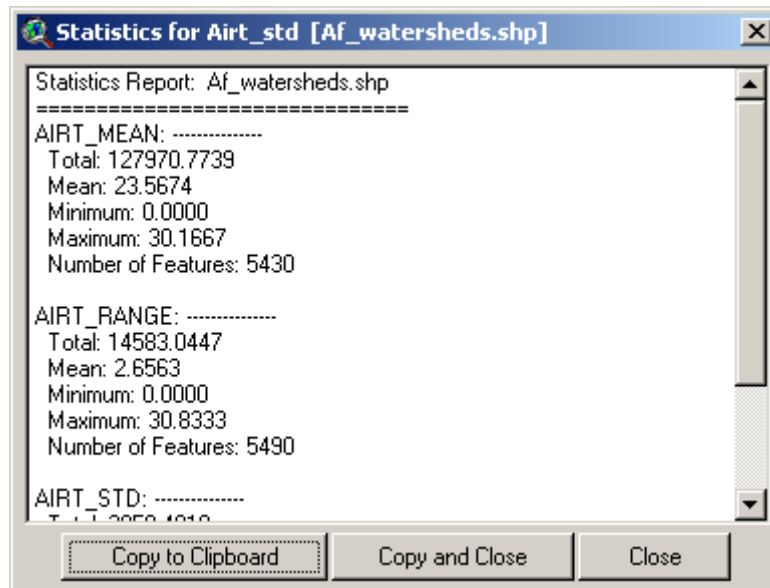
The  button will only be enabled if the user has at least one feature theme in the current View. When the user clicks the button, they will be prompted to identify the theme and/or fields for calculating the statistics.



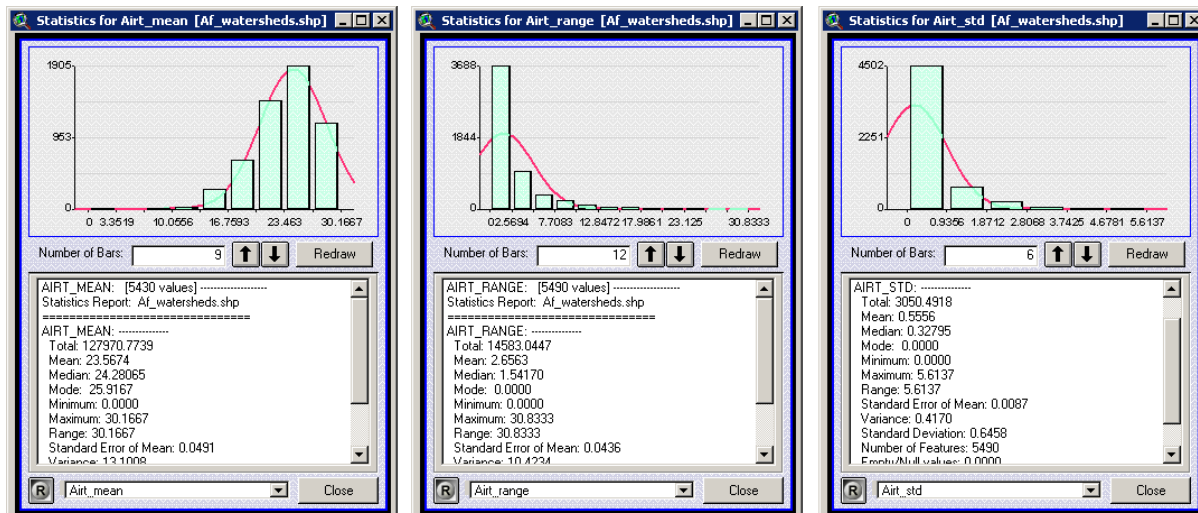
Also, the user may choose to calculate statistics on either all the features or only the selected features. If no features are selected, this tool will use all the features regardless of which option is chosen. The user also may choose to calculate statistics on multiple fields at one time.

The user may specify certain values they do not wish to include in the analysis. For example, it is common practice to designate some number to mean “No Data”, or to identify values not involved in the analysis. Researchers often use -9999 or -99999 for this purpose, especially with datasets where such a value would be impossible (e.g., elevation, population, area, etc.) The user may designate as many of these values as desired by entering them into the “Ignore Values” section, and checking the “Ignore” box.

The user may choose between either Simple or Advanced output. Simple output includes the *Sum*, *Number of Features*, *Mean*, *Minimum* and *Maximum*, and is reported in a text box:




Advanced output includes the *Sum, Mean, Median, Mode(s), Minimum, Maximum, Range, Standard Error of Mean, Variance, Standard Deviation, Number of Features, and Number of Null Values*, and is reported in a histogram:



Although only one histogram window will be open, the user may choose which set of statistics to view by choosing the field from the drop-down box at the bottom of the window. Also, the user may change the number of histogram bars to display by clicking the up/down arrows and selecting “Redraw”. The red line behind the histogram bars demonstrates how the bars should be arranged if the data were normally distributed. In the above examples, the mean air temperature values follow the normal distribution better than the range and standard deviation of air temperature values. The “R” button at the window’s bottom left is the “Refresh” button, and can be used if the image becomes corrupted. Clicking this button will redraw the image .


Summary Statistics on a Field in a Table:

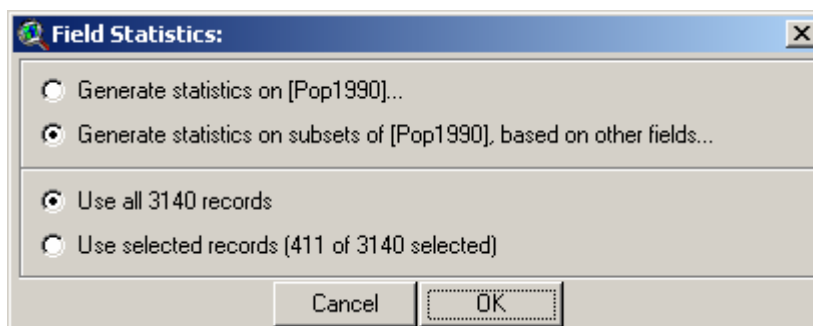
The  button in the Table button bar will be enabled only if a numeric field has been selected. This tool will allow the user to generate a large number of statistics on the values within a given field. The user may choose from: mean; standard error of the mean; confidence intervals; minimum; 1st quartile; median; 3rd quartile; maximum; range; variance; standard deviation; average absolute deviation; skewness

(normal and Fisher's G1); kurtosis (normal and Fisher's G2); number of records; number of null values; mode; and lastly, total sum for any attribute field(s) within a set of selected records.

This tool also allows users to break up the dataset into subsets based on one or more additional fields and generate multiple statistics for each subset of data. For example, if a person had a table of county-level statistics for all the counties in the United States, this tool would let them calculate a single set of statistics for all counties combined, or separate sets of statistics for each state or region.

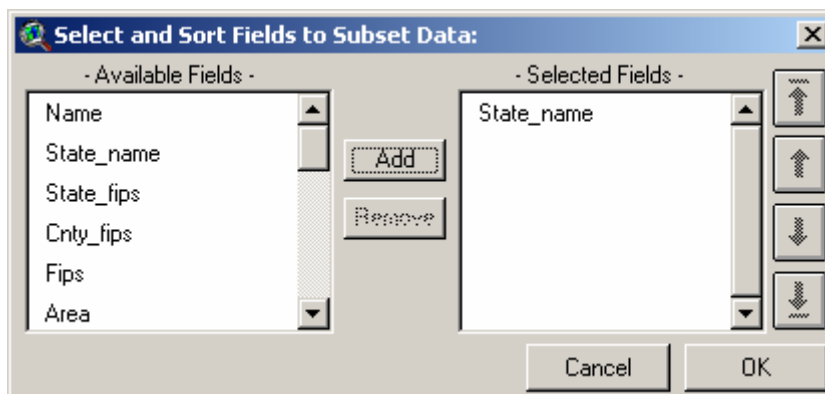
Users can use a single or multiple fields as classification fields to divide the data into subsets. If the user chooses multiple fields, then this extension will develop a separate set of statistics for each unique combination of classification values.

Begin by selecting the field containing your data, clicking the  to open the "Field Statistics" dialog, and setting your preferences:



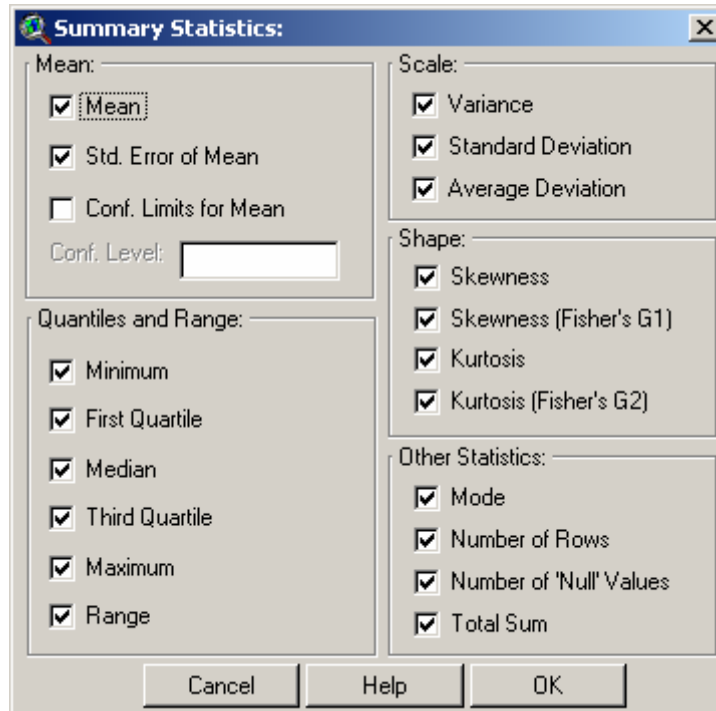
Generating Statistics on Multiple Subsets of Data:

Note that you have options to generate statistics on all data in the field or on subsets of that data. If you choose to generate statistics on subsets of the data, you will next be prompted to specify the fields containing your classification values. For example, if you wanted to analyze county statistics by state, then you would need to specify the field containing the state names.

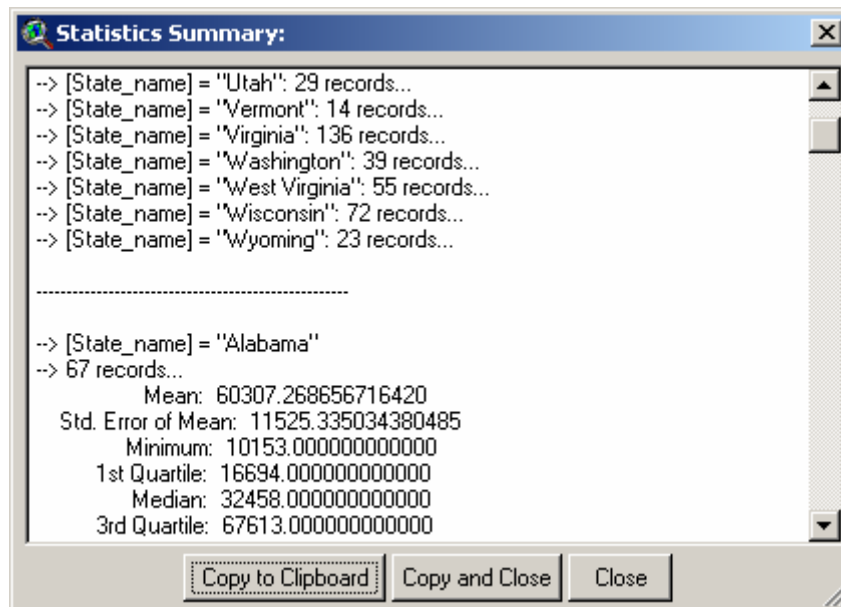


Note that you may choose multiple fields and change their order. If you choose multiple fields, then this extension will generate statistics for each unique combination of field values. The field order will not change the statistics produced, but will change the order they are presented in the final report.

Click 'OK' and specify the statistics you would like to generate. These statistics are described in detail below:

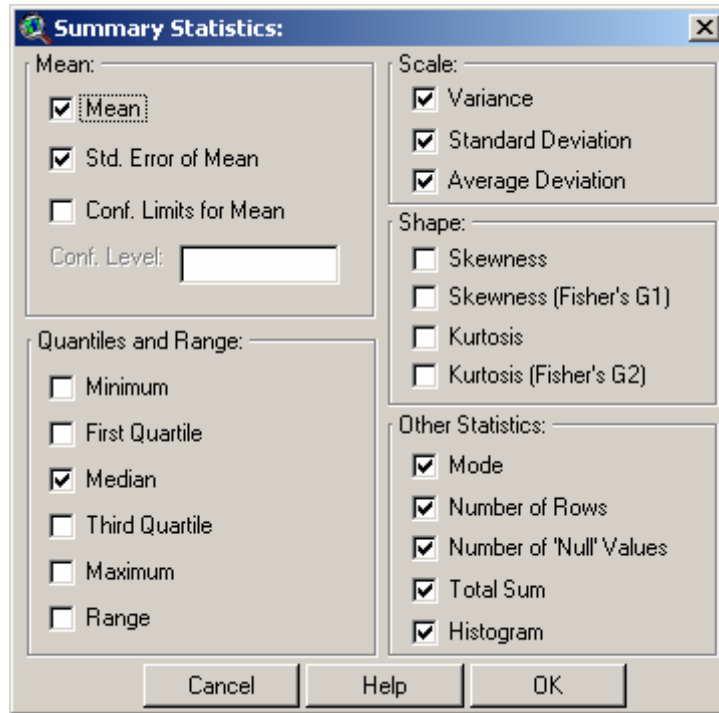


Click 'OK' and the extension will generate a report:

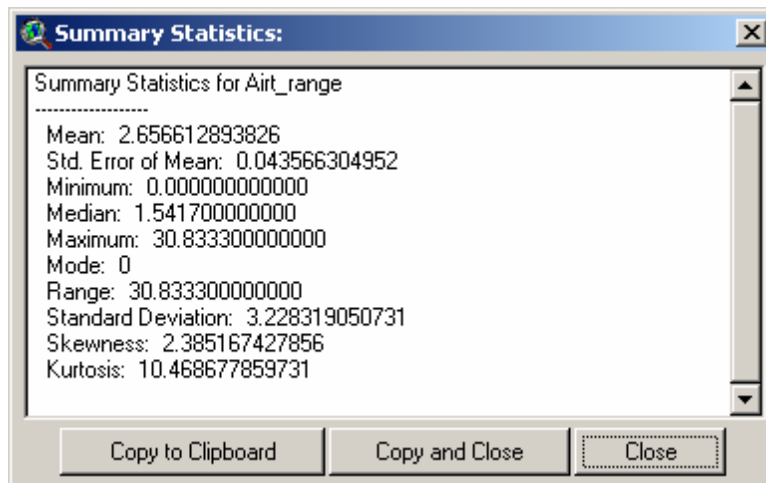


Generating Statistics on a Single Dataset:

If you choose to calculate statistics on all data in that field, you will next be prompted to specify your statistics in the Summary Statistics dialog. This version is slightly different in that here you have the option to create a histogram if you wish.



Choose the desired statistics and then click “OK.” If you selected the Histogram option, the output will appear in a histogram as illustrated above. If the Histogram option was not selected, the output will appear in a report window:



This tool may also be accessed with Avenue code, which enables more advanced users to pass these statistics to variables, and then use the calculated values in other places. Please review *Calculating Summary Statistics with Avenue* on p. 67 for details on accessing the Avenue script directly.

- 1) Mean: Calculated as: $\frac{\sum x}{n}$
- 2) Std Error of the Mean: Calculated as: $s\{\bar{Y}\} = \frac{s}{\sqrt{n}}$, where s = sample standard deviation

- 3) Confidence Interval: The confidence limits for population mean μ with a confidence coefficient $(1 - \alpha)$, given a sample population mean \bar{X} , are calculated as:

$$\bar{X} \pm t_{(1-\alpha/2; n-1)} s\{\bar{Y}\},$$

where $s\{\bar{Y}\} = \frac{s}{\sqrt{n}}$, s = Sample Standard Deviation

and $t_{(1-\alpha/2; n-1)}$ = value from the t distribution at $p = (1 - \alpha/2)$ and $n - 1$ degrees of freedom.

- 4) Minimum: The lowest value in the data set.
- 5) Quartiles and Median: Those values at which at most $(P)\%$ of the data lie below the value and at most $(1 - P)\%$ of the data lie above the value. There are different ways to calculate quartile values, which produce similar but different results. Some methods draw the quartile values from the data set itself, so that the value called the “quartile” will always be found in the data. This script uses a different method which occasionally calculates a quartile value which represents the midpoint between two values from the data set, applying the following algorithm:

Assuming the data have been sorted from lowest to highest:

$$\text{Quartile 1 Index} = (N + 1) \times 0.25 = Q(1)$$

$$\text{Quartile 2 Index} = (N + 1) \times 0.50 = Q(2)$$

$$\text{Quartile 3 Index} = (N + 1) \times 0.75 = Q(3)$$

If $Q(N)$ is an integer, then:

$$\text{Quartile} = Q(N)^{\text{th}} \text{ Value}$$

If $Q(N)$ is not an integer, then:

$$R = \text{that value at which } R < N < R + 1, \text{ and}$$

$$\text{Quartile} = \frac{(Q(R)^{\text{th}} \text{ Value}) + (Q(R + 1)^{\text{th}} \text{ Value})}{2}$$

- 6) Maximum: The highest value in the data set.
- 7) Range: Maximum - Minimum

8) Variance: Calculated as:
$$\text{Variance} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$$

9) Standard Deviation: Calculated as:
$$\text{Std. Deviation} = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$$

10) Average Deviation: Calculated as:
$$\text{Avg. Deviation} = \frac{\sum_{i=1}^n |y_i - \bar{y}|}{n}$$

- 11) Skewness: Measures the degree of asymmetry of the sample data around the mean.

$$\text{Skewness} = \frac{m_3}{m_2^{3/2}},$$

$$\text{where: } m_2 = 2^{\text{nd}} \text{ moment} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

$$\text{and: } m_3 = 3^{\text{rd}} \text{ moment} = \frac{\sum_{i=1}^n (y_i - \bar{y})^3}{n}$$

- 12) Skewness (Fisher's G1): There are alternative methods to calculate *skewness* measures of the data. S-PLUS uses the *Fisher's G1* variation, calculated as:

$$\text{Fisher's G1} = \frac{b_1 \sqrt{n(n-1)}}{n-2}$$

$$\text{where: } b_1 = \frac{m_3}{m_2^{3/2}} \quad (\text{standard measure of skewness}),$$

$$\text{and: } m_2 = 2^{\text{nd}} \text{ moment} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

$$\text{and: } m_3 = 3^{\text{rd}} \text{ moment} = \frac{\sum_{i=1}^n (y_i - \bar{y})^3}{n}$$

- 13) Kurtosis: Measures the “peakedness” or “pointedness” in a distribution, and calculated as:

$$\text{Kurtosis} = \frac{m_4}{m_2^2},$$

$$\text{where: } m_2 = 2^{\text{nd}} \text{ moment} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

$$\text{and: } m_4 = 4^{\text{th}} \text{ moment} = \frac{\sum_{i=1}^n (y_i - \bar{y})^4}{n}$$

- 14) Kurtosis (Fisher's G2): As with *Skewness*, there are alternative ways to calculate kurtosis. S-PLUS uses the *Fisher's G2* variation, calculated as:

$$\text{Fisher's G2} = \frac{(n+1)(n-1)}{(n-2)(n-3)} \left[b_2 - \frac{3(n-1)}{n+1} \right]$$

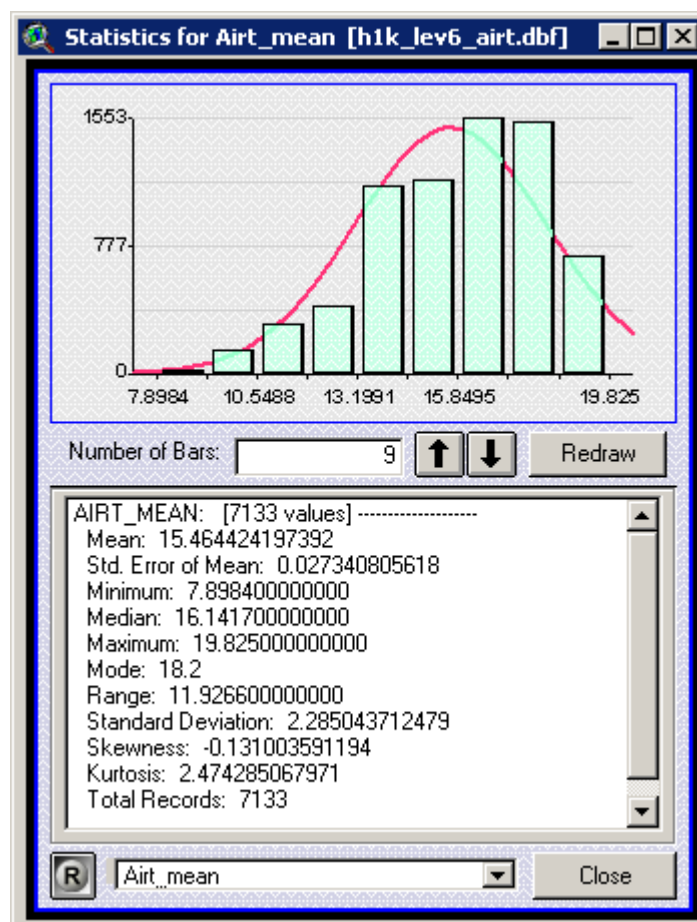
$$\text{where: } b_2 = \frac{m_4}{m_2^2} \quad (\text{standard measure of kurtosis}),$$

$$\text{and: } m_2 = 2^{\text{nd}} \text{ moment} = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n}$$

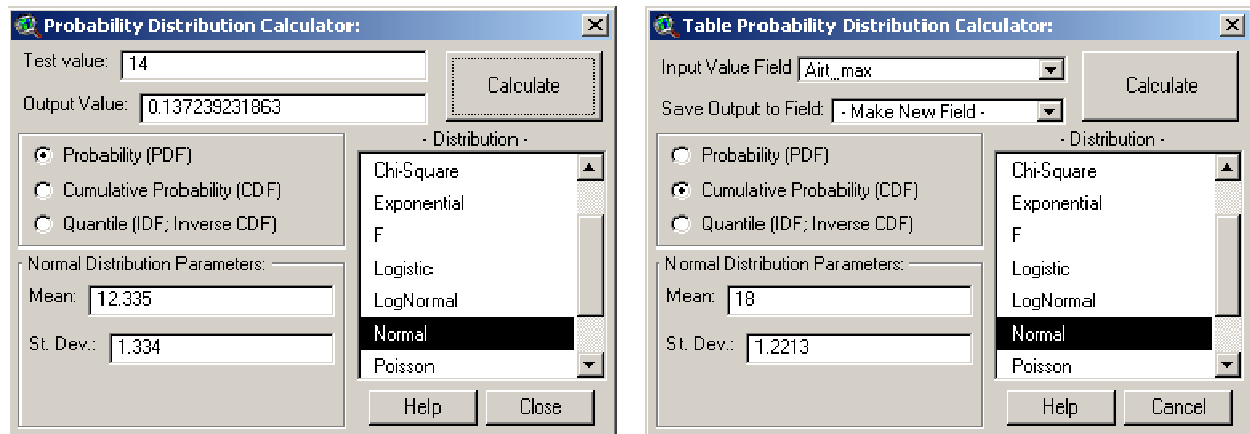
$$\text{and: } m_4 = 4^{\text{th}} \text{ moment} = \frac{\sum_{i=1}^n (y_i - \bar{y})^4}{n}$$

- 15) Mode: That value that occurs most often. There could be multiple modes or no modes in the data. If no value in the dataset is found more than once, this option will report that no modes were found. If no value occurs more than once, no mode is returned.
- 16) Number of Rows: The total number of rows of data examined during the analysis.
- 17) Number of 'Null' Values: The total number of cases of missing data. These are represented as "null" numbers in the table, which are different than zeros.
- 18) Total Sum: Sum of all non-null values, calculated as: $\sum x$
- 19) Histogram: This is a graphic illustrating the shape of the data and is useful for visually determining if the data are normally distributed. You may change the number of vertical bars by clicking the up/down arrows and then the "Redraw" button. The red line behind the bars shows how the data would appear if they were perfectly normally distributed. The drop-down box at the bottom of the illustration (containing the words "Airt_mean" in this example) shows the field that the statistics were calculated from. If you generated this histogram from a theme in a view, you may have selected multiple fields to calculate statistics from. This option allows you to choose which set of statistics to view.


The button with the "R" in the lower left-hand corner is a Refresh button. If the histogram window gets corrupted somehow, click the "R" button to regenerate it.




Probability Distribution Calculators:



This extension includes two versions of a Probability Distribution Calculator, each of which calculate distribution data based on a variety of distributions and parameters. The **Probability Distribution**

Calculator is started from within a View, and is opened by clicking on the  button in the View toolbar. You simply enter the input and parameter values, specify whether you are calculating Probability, Cumulative Probability or Quantile values, and click “Calculate”, and the result appears in the “Output Value” window. This calculator stays open until you close it and you can leave it open as you do other things in ArcView.

The “Table Probability Distribution Calculator” is designed to work on all selected records in a table, applying the distribution parameters to each value and saving the results to a field in that table. This

calculator is opened from within a Table by clicking on the  button in the Table toolbar. Select the field containing the “Input” values, then decide whether to create a new field or use an existing field to save the “Output” values, then click “Calculate” to generate distribution values for all selected records. The window stays open until you click “Calculate” or “Cancel”.

The Distribution functions included with this extension may be grouped in 3 categories. In general, the *Probability Density Functions* return the probability that the Test Value = X given that particular distribution. The *Cumulative Distribution Functions* return the probability that the Test Value \leq X, given that particular distribution. The *Quantile Functions* (sometimes referred to as *Inverse Density Functions* or *Percent Point Functions*) return the Value X at which $P(X) = [\text{specified probability}]$, given that particular distribution.

Functions and Probability Distributions			
Distribution	Probability Density Function	Cumulative Distribution Function	Quantile Function
Beta	PDF_Beta	CDF_Beta	IDF_Beta
Binomial	PDF_Binomial	CDF_Binomial	IDF_Binomial
Cauchy	PDF_Cauchy	CDF_Cauchy	IDF_Cauchy
Chi-Square	PDF_ChiSquare	CDF_ChiSquare	IDF_ChiSquare

Exponential	PDF_Exp	CDF_Exp	IDF_Exp
F	PDF_F	CDF_F	IDF_F
Logistic	PDF_Logistic	CDF_Logistic	IDF_Logistic
LogNormal	PDF_LogNormal	CDF_LogNormal	IDF_LogNormal
Normal	PDF_Normal_Simpsons	CDF_Normal	IDF_Normal
Poisson	PDF_Poisson	CDF_Poisson	IDF_Poisson
Student's T	PDF_StudentsT	CDF_StudentsT	IDF_StudentsT
Weibull	PDF_Weibull	CDF_Weibull	IDF_Weibull

Equations for each function are included in the *Distribution Functions, Parameters and Usages* (p. 69), but some of them do not have closed formulas which can be calculated and therefore must be computed numerically. Those interested should refer to the references to find source code and computational methods of calculating these functions. We recommend Croarkin & Tobias (date unknown) and McLaughlin (2001) for illustrations of the various distributions, and Press et al. (1988-1997) and Burkardt (2001) for computational methods. All of these sources are available on-line.

The descriptions in *Functions, Parameters and Usages* (p. 69) include four methods of utilizing each function. The first method describes how to use the *Probability Distribution Calculators* to calculate values. There are three additional methods available for programmers who may want to access the functions through Avenue code. Simply copy the line of code exactly as written, substituting your parameter variable names in the proper places.

Avenue Functions:

- 1) The first *Avenue* option sends your parameters to a central script called "Regression_jen.DistFunc", which checks for possible errors in the parameters (e.g. using a negative value for Degrees of Freedom). If the script finds errors, it will halt operation and alert you to the problem. If it doesn't find errors, it forwards your parameters to the appropriate script and returns the result. Users may want to review the script "Regression_jen.ProbDlogCalculate" for an example of this option. **IMPORTANT:** Users should be aware this script only checks whether the input values follow the rules described in *Functions, Parameters and Usages* on p. 69. It doesn't check for programming errors, such as sending a non-numeric value to the script.
- 2) The second *Avenue* option is similar to the first. It sends your parameters to a central script to check for errors (in this case, "Regression_jen.TableDistFunc"), but it doesn't halt operation if it finds an error. Rather, it returns an error message (in *String* format) detailing the problem. We recommend this option for cases in which the user wants to conduct calculations on a series of values (i.e. records in a table), but doesn't want the function to stop if it finds an illegal value (e.g., possibly a record with no data). This option would allow the user to insert an "if-then" statement in their code to check if the result is a *String* or a *Number*. Numerical responses would indicate successful calculations while *String* responses could be appended to a running report of unsuccessful calculations. Users may want to review the script "Regression_jen.ProbTabDlogCalculate" for an example of this option.
- 3) If you'd like to skip the error-checking routines, use the third *Avenue* option to send your parameters to the relevant script directly.

Calculating Summary Statistics with Avenue

The Summary Statistics tool collects a series of True/False and Numerical parameters from the user and sends them to a script called "Regression_jen.prob_Stat_CalcFieldStats", which does the necessary calculations and returns a list of results. The tool then prints those results up in a Report window for the user.

Avenue programmers can bypass the dialog and send values to the script directly if they wish, and then they will have the desired statistics directly available to them in a list. For example, many statistical calculations require such things as means, standard deviations, variances, quartiles, etc. The user may want to generate these values early in a script and then use them in later calculations. The "Regression_jen.prob_Stat_CalcFieldStats" script makes it simple to generate such values from data in a table.

This option is a little simpler than the standard Avenue method for generating statistics, which is to create a new file on the hard drive and then use the "Summarize" request to save statistics to that file. It also offers a larger variety of statistical output, including such things as confidence intervals, standard error of the mean, average deviation, and kurtosis/skewness values. This option is also a little slower on large datasets, however, and it doesn't divide up the dataset into subsets like "Summarize" does.

The function can be used with just a few lines of code:

```
ListOfResults = av.Run("Regression_jen.prob_Stat_CalcFieldStats", {ListOfInputParameters,
    theVTab, theField})
```

The object "theVTab" is a VTab object containing your data, and "theField" is a Field object in the VTab, reflecting the field you want to calculate statistics on.

The "ListOfInputParameters" must contain 22 values, most of which are Boolean (true/false) reflecting whether you want that particular statistic calculated. Note that the last value should be set to "False".

```
ListOfInputParameters = {CalcMean, CalcSEMean, CalcConInt, Con_Level,
    CalcMinimum, Calc1stQuart, CalcMedian, Calc3rdQuart, CalcMaximum,
    CalcVariance, CalcStandDev, CalcAvgDev, CalcSkewness, CalcSkewFish,
    CalcKurtosis, CalcKurtFish, CalcCount, CalcNumNull, CalcSum, CalcRange,
    CalcMode, False}
```

Where:

- CalcMean: **Boolean**, *True* if you want to calculate the mean.
- CalcSEMean: **Boolean**, *True* if you want to calculate the standard error of the mean.
- CalcConInt: **Boolean**, *True* if you want to calculate confidence intervals of the mean.
- Con_Level: **Number**, $0 \leq p \leq 1$, where $p = \text{probability} = (1 - \alpha)$
- CalcMinimum: **Boolean**, *True* if you want to calculate the minimum value.
- Calc1stQuart: **Boolean**, *True* if you want to calculate the 1st quartile.
- CalcMedian: **Boolean**, *True* if you want to calculate the median.
- Calc3rdQuart: **Boolean**, *True* if you want to calculate the 3rd quartile.
- CalcMaximum: **Boolean**, *True* if you want to calculate the maximum value.
- CalcVariance: **Boolean**, *True* if you want to calculate the variance.
- CalcStandDev: **Boolean**, *True* if you want to calculate the standard deviation.
- CalcAvgDev: **Boolean**, *True* if you want to calculate the absolute average deviation.
- CalcSkewness: **Boolean**, *True* if you want to calculate the standard skewness.

CalcSkewFish: **Boolean**, *True* if you want to calculate the Fisher's G1 skewness.

CalcKurtosis: **Boolean**, *True* if you want to calculate the standard kurtosis.

CalcKurtFish: **Boolean**, *True* if you want to calculate the Fisher's G2 kurtosis.

CalcCount: **Boolean**, *True* if you want to calculate the total number of rows of data.

CalcNumNull: **Boolean**, *True* if you want to calculate the number of null values.

CalcSum: **Boolean**, *True* if you want to calculate the sum.

CalcRange: **Boolean**, *True* if you want to calculate the Range.

CalcMode: **Boolean**, *True* if you want to calculate the Mode. .

ForHistogram: **False**, intended only for internal use.

When the script finishes, it will return a list of 19 values to you representing the various statistics you requested. If you did not request a particular statistic, then it will not be calculated and the return list will contain a "nil" object in it's place. Note that if you requested a confidence interval, the upper and lower levels are returned as a separate list (3rd object in the Return List).

```
Return list: {Mean, Standard Error of Mean, {Lower Confidence Level,
Upper Confidence Level}, Minimum, 1st Quartile, Median, 3rd
Quartile, Maximum, Variance, Standard Deviation, Skewness,
Fisher's GI Skewness, Kurtosis, Fisher's G2 Kurtosis,
Record Count, Number of Null Values, Sum, Range, Mode}
```

For example: If you had a table of population demographic data containing a field of Annual Income values, and you were interested in the mean annual income plus a 95% confidence interval around that mean, then you would set up the code as follows:

```
theDemographyVTab = YourTable.GetVTab
theField = theDemographyVTab.FindField("Income")
theInputParameters = {True, False, True, 0.95, False, False, False, False,
False, False, False, False, False, False, False, False, False, False, False}
theReturnList = av.Run("Regression_jen.prob_Stat_CalcFieldStats", {theInputParameters,
theDemographyVTab , theField})
theMeanIncome = theReturnList.Get(0)
theLowerConfidenceLimit = theReturnList.Get(2).Get(0)
theUpperConfidenceLimit = theReturnList.Get(2).Get(1)
```

All the objects in "theReturnList" will be "nil" objects except for the ones at indices 0 and 2. The Mean will be at index 0, the Lower 95% Confidence Limit will be the first item in index 2, and the Upper 95% Confidence Limit will be the second item in index 2.

In general, all the possible statistics can be obtained with the following lines of code. Simply copy and paste the appropriate lines into your script:

```
theMean = theReturnList.Get(0)
theSEMean = theReturnList.Get(1)
if (Calculating_Confidence_Intervals)
    LowerCI = theReturnList.Get(2).Get(0)
    UpperCI = theReturnList.Get(2).Get(1)
end
theMinimum = theReturnList.Get(3)
theQ1 = theReturnList.Get(4)
theMedian = theReturnList.Get(5)
theQ3 = theReturnList.Get(6)
theMaximum = theReturnList.Get(7)
theVar = theReturnList.Get(8)
theStdDev = theReturnList.Get(9)
theAvgDev = theReturnList.Get(10)
theSkew = theReturnList.Get(11)
theFisherSkew = theReturnList.Get(12)
theKurt = theReturnList.Get(13)
```

```

theFisherKurt = theReturnList.Get(14)
theCount = theReturnList.Get(15)
theNumberNull = theReturnList.Get(16)
theSum = theReturnList.Get(17)
theRange = theReturnList.Get(18)
theMode = theReturnList.Get(19)

```

Distribution Functions, Parameters and Usages

Probability Density Functions:

1. **PDF_Beta:** This function returns the probability that the Test Value = X, assuming a Beta distribution and the specified Shape parameters. This is the standardized Beta function, where Location = 0 and Scale (upper bound) = 1. According to McLaughlin (2001), parameters *Shape1* and *Shape2* can be any positive value, but they rarely exceed 10. The function becomes nearly flat if the values get much larger than this.

a) Parameters:

- i) Test Value: Number
- ii) Shape1: Number > 0
- iii) Shape2: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Beta* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_Beta", {Test Value, Shape1, Shape2}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_Beta", {Test Value, Shape1, Shape2}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_Beta", {Test Value, Shape1, Shape2})

- c) Function:
- $$\text{Beta PDF} = \frac{\Gamma(S_1 + S_2)}{\Gamma(S_1)\Gamma(S_2)} y^{S_1-1} (1-y)^{S_2-1}$$
- where: y = Test Value, S_1 = Shape 1, S_2 = Shape 2
- and: $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$

2. **PDF_Binomial:** The Binomial distribution is used when there are exactly two mutually exclusive outcomes of a trial. This function returns the probability of getting X successes out of N trials, given a probability of success = P.

a) Parameters:

- i) # Successes: Integer ≥ 0
- ii) # Trials: Integer ≥ 2 , # Successes
- iii) Probability of Success: Number ($0 \leq p \leq 1$)

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Binomial* distribution.

- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_Binomial", {#Success, #Trials, Probability of Success}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_Binomial", {#Success, #Trials, Probability of Success}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_Binomial", {#Success, #Trials, Probability of Success})

c) Function: Binomial PDF = $\left(\frac{B!}{y!(B-y)!} \right) A^y (1-A)^{B-y}$

where: y = #Successes, A = Probability of Success, B = #Trials

- 3. PDF_Cauchy:** This function returns the probability that the Test Value = X , assuming a *Cauchy* distribution with the specified mean and standard deviation. The *Standardized Cauchy distribution* is that with *Location* = 0 and *Scale* = 1.

a) Parameters:

- i) Test Value: Number
- ii) Location: Number
- iii) Scale: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Cauchy* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_Cauchy", {Test Value, Location, Scale}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_Cauchy", {Test Value, Location, Scale}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_Cauchy", {Test Value, Location, Scale})

c) Function: Cauchy PDF = $\frac{1}{\pi B \left[1 + \left(\frac{y-A}{B} \right)^2 \right]}$

where: y = Test Value, A = Location, B = Scale

- 4. PDF_ChiSquare:** This function returns the probability that the Test Value = X , assuming a *Chi-Square* distribution with the specified Degrees of Freedom. The *Chi-Square* distribution results when v (where v = Degrees of Freedom) independent variables with standard normal distributions are squared and summed (Croarkin & Tobias, Date unknown).

a) Parameters:

- i) Test Value: Number ≥ 0
- ii) Degrees of Freedom: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Chi-Square* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_ChiSquare", {Test Value, DF}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_ChiSquare", {Test Value, DF}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_ChiSquare", {Test Value, DF})

c) Function:
$$\text{Chi-Square PDF} = \frac{e^{-\frac{y}{2}} x^{\frac{v}{2}-1}}{2^{\frac{v}{2}} \Gamma\left(\frac{v}{2}\right)}$$

where: y = Test Value, S_1 = Shape 1, S_2 = Shape 2

and: $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$

5. PDF_Exp: This function returns the probability that the Test Value = X , assuming an *Exponential* distribution with the specified mean. This script uses the 1-parameter version of the equation (i.e. assuming *Location* = 0). The *Standard Exponential Distribution* is that which has *Mean* = 1.

a) Parameters:

- i) Test Value: Number ≥ 0
- ii) Mean: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Exponential* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_Exp", {Test Value, Mean}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_Exp", {Test Value, Mean}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_Exp", {Test Value, Mean})

c) Function:
$$\text{Exponential PDF} = \frac{1}{\beta} e^{-\frac{x}{\beta}}$$

where: x = Test Value, β = Mean (or Scale Parameter)

6. PDF_F: This function returns the probability that the Test Value = X , assuming an *F* distribution with the specified Degrees of Freedom. The *F* distribution is the ratio of two *Chi-Square* distributions with ratios v_1 and v_2 respectively.

a) Parameters:

- i) Test Value: Number ≥ 1
- ii) 1st Degrees of Freedom: Number > 1

iii) 2nd Degrees of Freedom: Number > 1

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *F* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_F", {Test Value, DF1, DF2}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_F", {Test Value, DF1, DF2}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_F", {Test Value, DF1, DF2})

$$F \text{ PDF} = \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right) \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} x^{\frac{v_1}{2} - 1}}{\Gamma\left(\frac{v_1}{2}\right) \Gamma\left(\frac{v_2}{2}\right) \left(1 + \frac{v_1 x}{v_2}\right)^{\frac{v_1 + v_2}{2}}}$$

c) Function:

where: x = Test Value, v_1 = DF1, v_2 = DF2

$$\text{and: } \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

7. PDF_Logistic: This function returns the probability that the Test Value = X , assuming a *Logistic* distribution with the specified mean and scale.

a) Parameters:

- i) Test Value: Number
- ii) Mean: Number
- iii) Scale: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Logistic* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_Logistic", {Test Value, Mean, Scale}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_Logistic", {Test Value, Mean, Scale}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_Logistic", {Test Value, Mean, Scale})

$$c) \text{ Function: Logistic PDF} = \frac{1}{B} \frac{\exp\left(\frac{y - A}{B}\right)}{\left[1 + \exp\left(\frac{y - A}{B}\right)\right]^2}$$

where: y = Test Value, A = Mean, B = Scale

8. PDF_LogNormal: This function returns the probability that the Test Value = X , assuming a *LogNormal* distribution with the specified mean and scale. A *LogNormal* distribution occurs when

natural logarithms of variable X are normally distributed. The *Standard LogNormal Distribution* is that with Mean = 0 and Scale = 1. The *2-Parameter LogNormal Distribution* is that with Mean = 0.

a) Parameters:

- i) Test Value: Number ≥ 0
- ii) Mean: Number > 0
- iii) Scale: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *LogNormal* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_LogNormal", {Test Value, Mean, Scale}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_LogNormal", {Test Value, Mean, Scale}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_LogNormal", {Test Value, Mean, Scale})

c) Function:
$$\text{LogNormal PDF} = \frac{1}{B\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(y) - A}{B}\right)^2\right)$$

where: y = Test Value, A = Mean, B = Scale

9. PDF_Normal: This function returns the probability that the Test Value = X , assuming a *Normal* distribution with the specified mean and standard deviation. The *Standard Normal Distribution* is that with Mean = 0 and Standard Deviation = 1.

a) Parameters:

- i) Test Value: Number
- ii) Mean: Number
- iii) Standard Deviation: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Normal* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_Normal", {Test Value, Mean, St. Dev.}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_Normal", {Test Value, Mean, St. Dev.}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_Normal", {Test Value, Mean, St. Dev.})

c) Function:
$$\text{Normal PDF} = \frac{1}{B\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{y - A}{B}\right)^2\right)$$

where: y = Test Value, A = Mean, B = Scale

10. PDF_Poisson: This function returns the probability that the Specified Number of Events = X , assuming a *Poisson* distribution with the specified mean.

a) Parameters:

- i) # Events: Integer ≥ 0
- ii) Mean: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Poisson* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_Poisson", {# Events, Mean}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_Poisson", {# Events, Mean}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_Poisson", {# Events, Mean})

c) Function: Poisson PDF = $\frac{\exp^{-A} A^y}{y!}$

where: y = Test value, A = Expectation (mean)

11. PDF_StudentsT: This function returns the probability that the Test Value = X , assuming a *Students T* distribution with the specified Degrees of Freedom. A *Student's T* distribution with 1df is a *Cauchy* Distribution, and it approaches a *Normal* distribution when $DF > 30$. Various sources recommend using the *Normal* distribution if $DF > 100$.

a) Parameters:

- i) Test Value: Number
- ii) Degrees of Freedom: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Student's T* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_StudentsT", {Test Value, DF}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_StudentsT", {Test Value, DF}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_StudentsT", {Test Value, DF})

c) Function: Student's T PDF = $\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v} \Gamma\left(\frac{v}{2}\right)} \left[1 + \frac{y^2}{v}\right]^{-\frac{v+1}{2}}$

where: y = Test Value, v = Degrees of Freedom

and: $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$

12. PDF_Weibull: This function returns the probability that the Test Value = X , assuming a *Weibull* distribution with the specified Location, Scale and Shape parameters. The *Standardized Weibull Distribution* is that with *Location* = 0 and *Scale* = 1. The *2-Parameter Weibull Distribution* is that with *Location* = 0.

a) Parameters:

- i) Test Value: Number > Location
- ii) Location: Number
- iii) Scale: Number > 0
- iv) Shape: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Probability (PDF)" and *Weibull* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"PDF_Weibull", {Test Value, Location, Scale, Number}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"PDF_Weibull", {Test Value, Location, Scale, Number}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_PDF_Weibull", {Test Value, Location, Scale, Number})

c) Function: Weibull PDF =
$$\left(\frac{C}{B}\right)\left(\frac{y-A}{B}\right)^{C-1} \exp\left(-\left(\frac{y-A}{B}\right)^C\right)$$

where: y = Test Value, A = Location, B = Scale, C = Shape

Cumulative Distribution Functions:

1. CDF_Beta: This function returns the probability that the Test Value $\leq X$, assuming a Beta distribution with the specified Shape parameters. This is the *Standardized Beta function*, where Location = 0 and Scale (upper bound) = 1. According to McLaughlin (2001), parameters *Shape1* and *Shape2* can be any positive value, but they rarely exceed 10. The function becomes nearly flat if the values get much larger than this.

a) Parameters:

- i) Test Value: Number
- ii) Shape1: Number > 0
- iii) Shape2: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Beta* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_Beta", {Test Value, Shape1, Shape2}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_Beta", {Test Value, Shape1, Shape2}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_Beta", {Test Value, Shape1, Shape2})

Beta CDF = $I(y, S_1, S_2)$ (From Press et al, 1997)

c) Function: where: y = Test Value, S_1 = Shape 1, S_2 = Shape 2

$$\text{and: } I(x, a, b) = \frac{B_x(a, b)}{B(a, b)} = \frac{1}{B(a, b)} \int_0^x t^{a-1} (1-t)^{b-1} dt$$

$$\text{and: } B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$$

2. **CDF_Binomial:** The *Binomial* distribution is used when there are exactly two mutually exclusive outcomes of a trial. This function returns the probability of getting $\leq X$ successes out of N trials, given a probability of success = P .

a) Parameters:

- i) # Successes: Integer ≥ 0
- ii) # Trials: Integer ≥ 2 , # Successes
- iii) Probability of Success: Number ($0 \leq p \leq 1$)

b) Usages:

- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Binomial* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_Binomial", {"#Success, #Trials, Probability of Success}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_Binomial", {"#Success, #Trials, Probability of Success}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_Binomial", {"#Success, #Trials, Probability of Success})

c) Function: Binomial CDF = $\sum_{i=1}^y \left(\frac{B!}{i!(B-i)!} \right) A^i (1-A)^{B-i}$

where: y = #Successes, A = Probability of Success, B = #Trials

3. **CDF_Cauchy:** This function returns the probability that the Test Value $\leq X$, assuming a *Cauchy* distribution with the specified Location and Scale parameters. The *Standardized Cauchy distribution* has *Location* = 0 and *Scale* = 1.

a) Parameters:

- i) Test Value: Number
- ii) Location: Number
- iii) Scale: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Cauchy* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_Cauchy", {Test Value, Location, Scale}})

- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_Cauchy", {Test Value, Location, Scale}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_Cauchy", {Test Value, Location, Scale})

c) Function: Cauchy CDF = $\frac{1}{2} + \frac{1}{\pi} \tan^{-1} \left(\frac{y - A}{B} \right)$

where: y = Test Value, A = Location, B = Scale

4. **CDF_ChiSquare:** This function returns the probability that the Test Value $\leq X$, assuming a *Chi-Square* distribution with the specified Degrees of Freedom. The *Chi-Square* distribution results when v (where v = Degrees of Freedom) independent variables with standard normal distributions are squared and summed (Croarkin & Tobias, Date unknown).

a) Parameters:

- i) Test Value: Number ≥ 0
- ii) Degrees of Freedom: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Chi-Square* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_ChiSquare", {Test Value, DF}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_ChiSquare", {Test Value, DF}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_ChiSquare", {Test Value, DF})

$$\text{Chi-Square CDF} = \frac{\gamma\left(\frac{v}{2}, \frac{x}{2}\right)}{\Gamma\left(\frac{v}{2}\right)}$$

c) Function: where: y = Test Value, S_1 = Shape 1, S_2 = Shape 2

and: $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$

and: $\gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt$

5. **CDF_Exp:** This function returns the probability that the Test Value $\leq X$, assuming an *Exponential* distribution with the specified mean. This script uses the 1-parameter version of the equation (i.e. assuming *Location* = 0). The *Standard Exponential Distribution* is that which has *Mean* = 1.

a) Parameters:

- i) Test Value: Number ≥ 0
- ii) Mean: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Exponential* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_Exp", {Test Value, Mean}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_Exp", {Test Value, Mean}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_Exp", {Test Value, Mean})

c) Function: Exponential CDF = $1 - e^{-\frac{x}{\beta}}$

where: x = Test value, β = Mean (or Scale Parameter)

6. CDF_F: This function returns the probability that the Test Value $\leq X$, assuming an *F* distribution with the specified Degrees of Freedom. The *F* distribution is the ratio of two *Chi-Square* distributions with ratios v_1 and v_2 respectively.

a) Parameters:

- i) Test Value: Number ≥ 1
- ii) 1st Degrees of Freedom: Number > 1
- iii) 2nd Degrees of Freedom: Number > 1

b) Usages:

- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *F* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_F", {Test Value, DF1, DF2}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_F", {Test Value, DF1, DF2}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_F", {Test Value, DF1, DF2})

$$F \text{ CDF} = 1 - I\left(k, \frac{v_1}{2}, \frac{v_2}{2}\right)$$

$$\text{where: } k = \left(\frac{v_2}{v_2 + v_1 y}\right)$$

c) Function: and: y = Test Value, S_1 = Shape 1, S_2 = Shape 2

$$\text{and: } I(x, a, b) \equiv \frac{B_x(a, b)}{B(a, b)} = \frac{1}{B(a, b)} \int_0^x t^{a-1} (1-t)^{b-1} dt$$

$$\text{and: } B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$$

(From Croarkin & Tobias, Date Unknown; Press et al, 1997)

7. CDF_Logistic: This function returns the probability that the Test Value $\leq X$, assuming a *Logistic* distribution with the specified mean and scale.

a) Parameters:

- i) Test Value: Number
 - ii) Mean: Number
 - iii) Scale: Number > 0
- b) Usages:
- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Logistic* distribution.
 - ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_Logistic", {Test Value, Mean, Scale}})
 - iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_Logistic", {Test Value, Mean, Scale}})
 - iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_Logistic", {Test Value, Mean, Scale})

c) Function: Logistic CDF =
$$\frac{1}{1 + \exp\left(\frac{A - y}{B}\right)}$$

where: y = Test Value, A = Mean, B = Scale

8. CDF_LogNormal: This function returns the probability that the Test Value $\leq X$, assuming a *LogNormal* distribution with the specified mean and scale. A *LogNormal* distribution occurs when natural logarithms of variable X are normally distributed. The *Standard LogNormal Distribution* is that with Mean = 0 and Scale = 1. The *2-Parameter LogNormal Distribution* is that with Mean = 0.

- a) Parameters:
- i) Test Value: Number ≥ 0
 - ii) Mean: Number > 0
 - iii) Scale: Number > 0
- b) Usages:
- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *LogNormal* distribution.
 - ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_LogNormal, {Test Value, Mean, Scale}})
 - iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_LogNormal, {Test Value, Mean, Scale}})
 - iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_LogNormal", {Test Value, Mean, Scale})

c) Function: LogNormal CDF =
$$\Phi\left(\frac{\ln(y) - A}{B}\right)$$

where: y = Test Value, A = Mean, B = Scale

and: $\Phi(x)$ = Cumulative Distribution Function of the Normal Distribution

- 9. CDF_Normal_Simpsons:** This function returns the probability that the Test Value $\leq X$, assuming a *Normal* distribution with the specified mean and standard deviation. Because the formula for this function does not exist in a closed form, it must be computed numerically. This script uses the *Simpson's* approximation method (Stewart 1998, p. 421-424) to calculate a highly accurate estimate of the Normal cumulative distribution function (accuracy to > 12 decimal places). The *Standard Normal Distribution* is that with Mean = 0 and Standard Deviation = 1.

a) Parameters:

- i) Test Value: Number
- ii) Mean: Number
- iii) Standard Deviation: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Normal* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_Normal_Simpsons, {Test Value, Mean, St. Dev.}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_Normal_Simpsons, {Test Value, Mean, St. Dev.}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_Normal_Simpsons", {Test Value, Mean, St. Dev.})

c) Function:
$$\text{Normal CDF} = \Phi\left(\frac{y - A}{B}\right)$$

where: y = Test Value, A = Mean, B = Scale

and: $\Phi(x)$ = Cumulative Distribution Function of the Normal Distribution

- 10. CDF_Poisson:** This function returns the probability that the specified Number of Events will be $\leq X$, assuming a *Poisson* distribution with the specified mean.

a) Parameters:

- i) # Events: Integer ≥ 0
- ii) Mean: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Poisson* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_Poisson, {# Events, Mean}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_Poisson, {# Events, Mean}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_Poisson", {# Events, Mean})

$$\text{Poisson CDF} = \frac{\gamma(y, A)}{\Gamma(y)}$$

c) Function: where: y = Test value, A = Expectation (mean)

$$\text{and: } \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

$$\text{and: } \gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt$$

11. CDF_StudentsT: This function returns the probability that the Test Value $\leq X$, assuming a *Students T* distribution with the specified Degrees of Freedom. A *Student's T* distribution with 1df is a *Cauchy* Distribution, and it approaches a *Normal* distribution when $DF > 30$. Various sources recommend using the *Normal* distribution if $DF > 100$.

a) Parameters:

i) Test Value: Number

ii) Degrees of Freedom: Number > 0

b) Usages:

i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Student's T* distribution.

ii) (Avenue): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_StudentsT, {Test Value, DF}})

iii) (Avenue): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_StudentsT, {Test Value, DF}})

iv) (Avenue): theProb = av.Run("Regression_jen.prob_Stat_CDF_StudentsT", {Test Value, DF})

c) Function: The CDF_StudentsT T Function is dependent on whether the test value is positive or negative:

$$\text{Student's T CDF} = \begin{cases} \frac{1}{2} I\left(\frac{v}{v+y^2}, \frac{v}{2}, \frac{1}{2}\right), & t \equiv y \leq 0 \\ 1 - \frac{1}{2} I\left(\frac{v}{v+y^2}, \frac{v}{2}, \frac{1}{2}\right), & t \equiv y > 0 \end{cases}$$

where: y = Test Value, v = Degrees of Freedom

$$\text{and: } I(x, a, b) \equiv \frac{B_x(a, b)}{B(a, b)} \equiv \frac{1}{B(a, b)} \int_0^x t^{a-1} (1-t)^{b-1} dt$$

$$\text{and: } B(a, b) = \int_0^1 t^{a-1} (1-t)^{b-1} dt$$

12. CDF_Weibull: This function returns the probability that the Test Value $\leq X$, assuming a *Weibull* distribution with the specified Location, Scale and Shape parameters. The *Standardized Weibull Distribution* is that with *Location* = 0 and *Scale* = 1. The *2-Parameter Weibull Distribution* is that with *Location* = 0.

a) Parameters:

i) Test Value: Number $>$ Location

- ii) Location: Number
 - iii) Scale: Number > 0
 - iv) Shape: Number > 0
- b) Usages:
- i) From "Probability Distribution Calculator", select "Cumulative Probability (CDF)" and *Weibull* distribution.
 - ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"CDF_Weibull, {Test Value, Location, Scale, Number}})
 - iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"CDF_Weibull, {Test Value, Location, Scale, Number}})
 - iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_CDF_Weibull", {Test Value, Location, Scale, Number})
- c) Function: Weibull CDF = $1 - \exp\left(-\left(\frac{y-A}{B}\right)^C\right)$
 where: y = Test Value, A = Location, B = Scale, C = Shape

Quantiles (also referred to as Inverse Density Functions or Percent Point Functions).

1. IDF_Beta: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the Beta distribution with the two specified Shape parameters. Because the formula for this function does not exist in a closed form, it must be computed numerically. This script arrives at the X -value through an iterative process, repeatedly testing X -values with the *CDF_Beta* function until it arrives at $P\text{-value}$ that is within 1×10^{-12} units from the specified $P\text{-value}$ (this usually takes between 30-60 iterations).

- a) Parameters:
- i) P-value: Number ($0 \leq p \leq 1$)
 - ii) Shape1: Number > 0
 - iii) Shape2: Number > 0
- b) Usages:
- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Beta* distribution.
 - ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_Beta, {P-value, Shape1, Shape2}})
 - iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_Beta, {P-value, Shape1, Shape2}})
 - iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_Beta", {P-value, Shape1, Shape2})

c) Function:
$$\text{Beta IDF} = \int_0^y \frac{\Gamma(S_1 + S_2)}{\Gamma(S_1)\Gamma(S_2)} y^{S_1-1} (1-y)^{S_2-1} dy$$

where: y = Test Value, S_1 = Shape 1, S_2 = Shape 2

and: $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$

2. IDF_Binomial: This function takes the specified probability and returns the value X such that the Probability of getting $(X - 1)$ successes \leq the Specified Probability. This function takes an iterative approach to finding the correct X value, repeatedly trying different values of X until it reaches the correct one. This iterative process rarely takes more than 25 repetitions.

a) Parameters:

- i) P-value = Number ($0 \geq p \geq 1$)
- ii) # Trials = Integer ≥ 2
- iii) Probability of Success = Number ($0 \geq p \geq 1$)

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Binomial* distribution.
- ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_Binomial", {P-value, NumTrials, Probability of Success}})
- iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_Binomial", {P-value, NumTrials, Probability of Success}})
- iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_Binomial", {P-value, NumTrials, Probability of Success})

Binomial IDF: Iterative Process, repeatedly testing values of y , such that:

c) Function:
$$p = \sum_{i=1}^y \left(\frac{B!}{i!(B-i)!} \right) A^i (1-A)^{B-i}$$

where: y = #Successes, A = Probability of Success, B = #Trials

Until: $P(y - 1) \leq \text{User-Specified probability}$

3. IDF_Cauchy: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the Cauchy distribution with the specified location and scale parameters. The *Standardized Cauchy distribution* has *Location* = 0 and *Scale* = 1.

a) Parameters:

- i) P-value: Number ($0 \geq p \geq 1$)
- ii) Location: Number
- iii) Scale: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Cauchy* distribution.

- ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_Cauchy", {P-value, location, Scale}})
- iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_Cauchy", {P-value, location, Scale}})
- iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_Cauchy", {P-value, Location, Scale})

c) Function: Cauchy IDF = $A - \frac{B}{\tan(\pi p)}$

where: A = Location, B = Scale, p = Probability

4. IDF_ChiSquare: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the Chi-Square distribution with the specified Degrees of Freedom. Because the formula for this function does not exist in a closed form, it must be computed numerically. This script arrives at the X -value through an iterative process, repeatedly testing X -values with the *CDF_ChiSquare* function until it arrives at $P\text{-value}$ that is within 1×10^{-12} units from the specified $P\text{-value}$ (this usually takes between 30-60 iterations). The *Chi-Square* distribution results when ν (where ν = Degrees of Freedom) independent variables with standard normal distributions are squared and summed (Croarkin & Tobias, Date unknown).

a) Parameters:

- i) P-value: Number ($0 \leq p \leq 1$)
- ii) Degrees of Freedom: Number

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Chi-Square* distribution.
- ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_ChiSquare", {P-Value, F}})
- iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_ChiSquare", {P-Value, F}})
- iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_ChiSquare", {P-Value, DF})

$$\text{Chi-Square IDF} = \int_0^y \frac{e^{-\frac{y}{2}} x^{\frac{\nu}{2}-1}}{2^{\frac{\nu}{2}} \Gamma\left(\frac{\nu}{2}\right)} dy$$

c) Function:

where: y = Test Value, S_1 = Shape 1, S_2 = Shape 2

$$\text{and: } \Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$$

5. IDF_Exp: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the Exponential distribution with the specified mean. This script uses the 1-parameter version of the equation (i.e. assuming *Location* = 0). The *Standard Exponential Distribution* is that which has *Mean* = 1.

a) Parameters:

i) P-value: Number ($0 \leq p \leq 1$)

ii) Mean: Number > 0

b) Usages:

i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Exponential* distribution.

ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_Exp, {P-value, Mean}})

iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_Exp, {P-value, Mean}})

iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_Exp", {P-value, Mean})

c) Function: $\text{Exponential IDF} = -\beta \ln(1 - p)$
where: β = Mean (or Scale Parameter)
and: p = Specified Probability

6. IDF_F: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the F distribution with the specified Degrees of Freedom. Because the formula for this function does not exist in a closed form, it must be computed numerically. This script arrives at the X -value through an iterative process, repeatedly testing X -values with the CDF_F function until it arrives at $P\text{-value}$ that is within 1×10^{-12} units from the specified $P\text{-value}$ (this usually takes between 30-60 iterations). The F distribution is the ratio of two *Chi-Square* distributions with ratios v_1 and v_2 respectively.

a) Parameters:

i) Test Value: Number ≥ 1

ii) 1st Degrees of Freedom: Number > 1

iii) 2nd Degrees of Freedom: Number > 1

b) Usages:

i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and F distribution.

ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_F, {P-value, DF1, DF2}})

iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_F, {P-value, DF1, DF2}})

iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_F", {P-value, DF1, DF2})

c) Function:
$$F \text{ IDF} = \int_0^x \frac{\Gamma\left(\frac{v_1 + v_2}{2}\right) \left(\frac{v_1}{v_2}\right)^{\frac{v_1}{2}} x^{\frac{v_1}{2}-1}}{\Gamma\left(\frac{v_1}{2}\right) \Gamma\left(\frac{v_2}{2}\right) \left(1 + \frac{v_1 x}{v_2}\right)^{\frac{v_1 + v_2}{2}}} dx$$

where: x = Test Value, v_1 = DF1, v_2 = DF2

and: $\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt$

7. IDF_Logistic: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the *Logistic* distribution with the specified mean and scale parameters.

a) Parameters:

- i) P-value: Number ($0 \leq p \leq 1$)
- ii) Mean: Number
- iii) Scale: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Logistic* distribution.
- ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_Logistic", {P-value, Mean, Scale}})
- iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_Logistic", {P-value, Mean, Scale}})
- iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_Logistic", {P-value, Mean, Scale})

c) Function:
$$\text{Logistic IDF} = A + B \ln\left(\frac{p}{1-p}\right)$$

where: p = Probability, A = Mean, B = Scale

8. IDF_LogNormal: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the *LogNormal* distribution with the specified mean and scale parameters. Because the formula for this function does not exist in a closed form, it must be computed numerically. This script arrives at the X -value through an iterative process, repeatedly testing X -values with the *CDF_LogNormal* function until it arrives at $P\text{-value}$ that is within 1×10^{-12} units from the specified $P\text{-value}$ (this usually takes between 30-60 iterations). A *LogNormal* distribution occurs when natural logarithms of variable X are normally distributed. The *Standard LogNormal Distribution* is that with Mean = 0 and Scale = 1. The *2-Parameter LogNormal Distribution* is that with Mean = 0.

a) Parameters:

- i) P-value: Number ($0 \leq p \leq 1$)
- ii) Mean: Number > 0
- iii) Scale: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *LogNormal* distribution.
- ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_LogNormal, {P-value, Mean, Scale}})
- iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_LogNormal, {P-value, Mean, Scale}})
- iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_LogNormal", {P-value, Mean, Scale})

c) Function:
$$\text{LogNormal IDF} = \int_0^y \left(\frac{1}{B\sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{\ln(y) - A}{B} \right)^2 \right) \right) dy$$

where: y = Test Value, A = Mean, B = Scale

- 9. IDF_Normal:** This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the *Normal* distribution with the specified mean and standard deviation. Because the formula for this function does not exist in a closed form, it must be computed numerically. This script arrives at the X -value through an iterative process, repeatedly testing X -values with the *CDF_Normal_Simpsons* function until it arrives at $P\text{-value}$ that is within 1×10^{-12} units from the specified $P\text{-value}$ (this usually takes between 30-60 iterations). Furthermore, there is no closed formula for calculating the Normal cumulative distribution function, so this script uses the *Simpson's* approximation method (Stewart 1998, p. 421-424) to calculate a highly accurate estimate (accuracy to > 12 decimal places). The *Standard Normal Distribution* is that with Mean = 0 and Standard Deviation = 1.

a) Parameters:

- i) P-value: Number ($0 \leq p \leq 1$)
- ii) Mean: Number
- iii) Standard Deviation: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Normal* distribution.
- ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_Normal, {P-value, Mean, St. Dev.}})
- iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_Normal, {P-value, Mean, St. Dev.}})
- iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_Normal", {P-value, Mean, St. Dev.})

c) Function:
$$\text{Normal IDF} = \int_0^y \left(\frac{1}{B\sqrt{2\pi}} \exp \left(-\frac{1}{2} \left(\frac{y - A}{B} \right)^2 \right) \right) dy$$

where: y = Test Value, A = Mean, B = Scale

- 10. IDF_Poisson:** This function takes the specified probability and returns the value X such that the Probability of getting $(X - 1)$ events \leq the Specified Probability. This function takes an iterative

approach to finding the correct X value, repeatedly trying different values of X until it reaches the correct one.

a) Parameters:

- i) P-value: Number ($0 \leq p \leq 1$)
- ii) Mean: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Poisson* distribution.
- ii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_Poisson", {P-value, Mean}})
- iii) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_Poisson", {P-value, Mean}})
- iv) (*Avenue*): theProb = av.Run("Regression_jen.prob_Stat_IDF_Poisson", {P-value, Mean})

Poisson IDF: Iterative Process, repeatedly testing values of y , such that:

$$p = \frac{\gamma(y, A)}{\Gamma(y)}$$

c) Function: where: y = Test value, A = Expectation (mean)

$$\text{and: } \Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

$$\text{and: } \gamma(x, y) = \int_0^y t^{x-1} e^{-t} dt$$

Until: $P(y - 1) \leq \text{User-Specified probability}$

11. IDF_StudentsT: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the *Student's T* distribution with the specified Degrees of Freedom. Because the formula for this function does not exist in a closed form, it must be computed numerically. This script arrives at the X -value through an iterative process, repeatedly testing X -values with the *CDF_StudentsT* function until it arrives at $P\text{-value}$ that is within 1×10^{-12} units from the specified $P\text{-value}$ (this usually takes between 30-60 iterations). A *Student's T* distribution with 1df is a *Cauchy* Distribution, and it approaches a *Normal* distribution when $DF > 30$. Various sources recommend using the *Normal* distribution if $DF > 100$.

a) Parameters:

- i) P-value: Number ($0 \leq p \leq 1$)
- ii) Degrees of Freedom: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Student's T* distribution.
- ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_StudentsT", {P-value, DF}})
- iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_StudentsT", {P-value, DF}})
- iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_StudentsT", {P-value, DF})

c) Function:
$$\text{Student's T IDF} = \int_{-\infty}^y \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{\pi v} \Gamma\left(\frac{v}{2}\right)} \left[1 + \frac{y^2}{v}\right]^{-\frac{v+1}{2}} dy$$

where: y = Test Value, v = Degrees of Freedom
and: $\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$

12. IDF_Weibull: This function takes the specified probability and returns the value X , such that $P(X) = P\text{-value}$, given the *Weibull* distribution with the specified Location, Scale and Shape parameters. The *Standardized Weibull Distribution* is that with *Location* = 0 and *Scale* = 1. The *2-Parameter Weibull Distribution* is that with *Location* = 0.

a) Parameters:

- i) Test Value: Number > Location
- ii) Location: Number
- iii) Scale: Number > 0
- iv) Shape: Number > 0

b) Usages:

- i) From "Probability Distribution Calculator", select "Quantile (IDF; Inverse CDF)" and *Weibull* distribution.
- ii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_DistFunc", {"IDF_Weibull", {P-value, Location, Scale, Number}})
- iii) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_TableDistFunc", {"IDF_Weibull", {P-value, Location, Scale, Number}})
- iv) (*Avenue*): theX = av.Run("Regression_jen.prob_Stat_IDF_Weibull", {P-value, Location, Scale, Number})

c) Function: $\text{Weibull IDF} = A + B \sqrt[C]{-\ln p}$

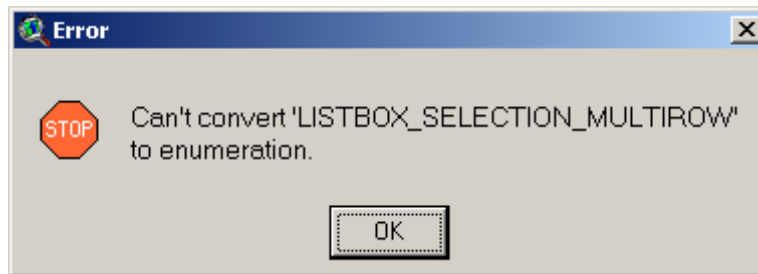
where: p = Probability, A = Location, B = Scale, C = Shape

Troubleshooting:

If you encounter some strange crash, please click the menu item "Check Regression Scripts" in either the View, Table or Project Help menu. Click this as soon as you are able following the crash. With any luck, that function will generate a report with enough information for the author to find and fix the problem.

Otherwise, the problem may be found and explained below:

Problem: Extension Fails to Load, with the following error message:

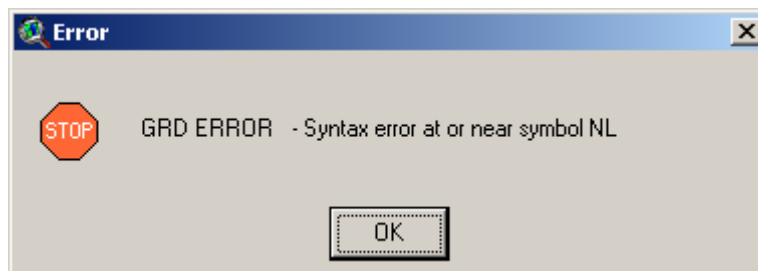


Solution: This problem is caused by an outdated version of the Dialog Designer. For some reason, some versions of ArcView 3 were shipped with an older version of Dialog Designer which didn't support this "LISTBOX_SELECTION_MULTIROW" option (which basically means that a listbox on a dialog is set so that you can select multiple items from the list).

ESRI has a newer version of the Dialog Designer available on their website for free download. Please link to:

http://support.esri.com/index.cfm?fa=downloads_patchesServicePacks.viewPatch&PID=25&MetaID=483

Problem: Extension crashes in mid-operation, producing an obscure message stating that there is a syntax error at or near symbol NL:



This is sometimes followed by the infamous Segmentation Violation message:

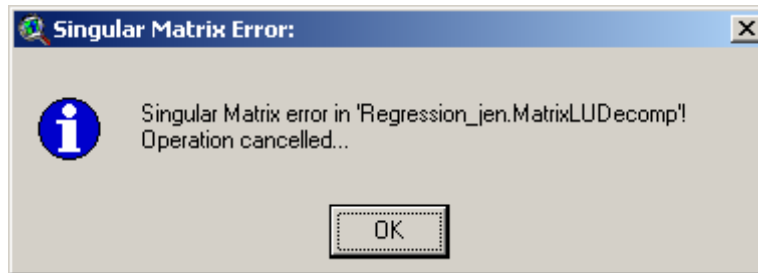


Sometimes ArcView crashes completely and vanishes without showing these messages; while other times it vanishes after showing these messages. Sometimes it keeps working in an unstable state.

Solution: There is no simple solution to this problem. It is due to a bug in Spatial Analyst which causes SA to crash after approximately 32,500 grid operations or if SA tries to hold > 50 grids in memory at one time. You can force the error to occur by writing a short script that checks the cell value at a particular point, then loops over 32,500 iterations. You can also trigger it by running the Zonal Statistics function on a point theme containing over 32,500 points or trying to do any grid operation that accesses > 50 grids.

I am unaware of a simple way to work around this problem. If possible, use smaller point data sets or try to use fewer grids in your analysis. Alternatively, ArcGIS 9 is expected to fix this problem.

Problem: Extension stops in mid-operation, producing an error message stating that there is a singular matrix error:

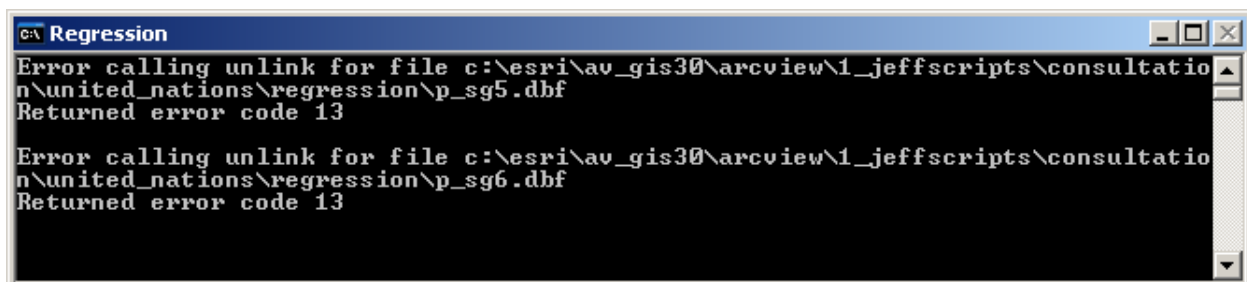


A couple of issues could cause this:

1. If there is something unusual about your data that makes it impossible to invert the matrix of predictor values $[(X'X)^{-1}]$, then the matrix is considered “singular” and the regression analysis cannot continue. This issue might arise if all your response values corresponded with the exact same predictor value so that the scatterplot looks like a column of points. In this case, there are really an infinite number of equally valid regression lines, all going through the same midpoint and all having different slopes.
2. This error can also be triggered if you are using invalid data. For example, if you have negative values in your predictor data and you attempt a log transformation (which is impossible; you can't take logs of negative numbers without resorting to complex imaginary numbers, and ArcView doesn't do imaginary numbers), then the script will react as if you sent it a singular matrix.

Solution: There is no solution if you are genuinely using a singular matrix. Some data are simply not appropriate for regression analysis.

Problem: Extension produces an error message in a DOS window stating that there is an “Error calling unlink” for some particular file. The extension continues operating and the analysis completes successfully. The DOS window stays open until ArcView is shut down.



This error is triggered when ArcView attempts to delete a file and is unable to do so. For some reason the ArcView programmers decided to alert you to this problem by saying that there was an “error calling unlink” for that file rather than simply saying it could not be deleted, which I assume means that ArcView believes that some other program is using that file.

The files this extension attempts to delete are temporary statistics tables used to report grid statistics. This extension generates the files, examines the data and adds them to the report, then deletes the files. There is usually no problem, but occasionally one of the files can't be deleted.

Solution: This isn't really a problem and there is no solution that I have been able to figure out. I don't know why ArcView is occasionally unable to delete the statistics file. The extension will continue to operate normally and you can minimize the DOS window if you do not want to see it. The worst thing that happens is that you get a small unwanted file sitting on your hard drive, adding to the general clutter of files. You can easily delete the file manually using Windows Explorer

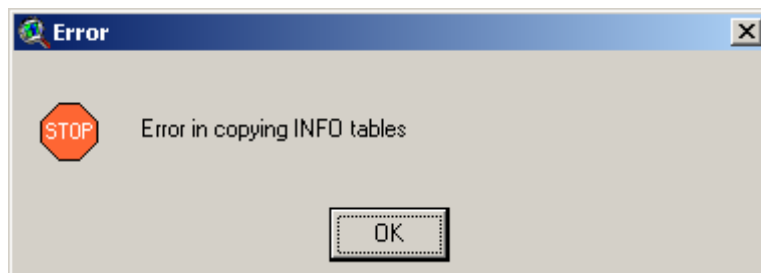
Problem: Unable to find grid in a directory, even though you know it is there.

Solution: This is probably due to a space or invalid character in the pathname. For some reason, Spatial Analyst doesn't recognize a grid if it lies in a folder with a space or period in it. For example, if you store your grids in the standard default Windows directory "My Documents," you will probably not see the grid listed in the "Add Theme" dialog at all. The "Add Theme" dialog will show you all the shapefiles and images, but no grids. The only solution is to move your grid to a different file location where it does not lie in a path with invalid characters.

Problem: You load your grids but you are unable to do any calculations on them. They don't act like grids.

Solution: You may have loaded them as images. Grids can be loaded as either images or grids, and you can't do any of the grid functions on them if you have loaded them as images. Try adding them to your view again, but make sure that you have "Grid Data Source" selected instead of "Image Data Source".

Problem: Extension crashes in mid-calculation, with the message:



Solution: This error may be caused by either a corrupt INFO directory or if your working directory pathname is too long. I am unaware of the exact pathname size that triggers the error, but I think it is around 80 characters. If you have over 80 characters in your pathname and you see this error, then you can probably avoid it by changing your work directory to someplace closer to the root.

Revisions:

Version 3.0 (August 13, 2005)

- Adds functions to build custom models and conduct multiple regression analyses.

Version 3.1 (September 8, 2005)

- Corrects a bug that produced an error message stating that “Assertion ‘Positive buffer size’ failed.”
- Corrects a bug triggered by applying an inverse transformation to a predictor grid.
- Modifies the Field Statistics tools so that you can generate statistics on multiple subsets of your data, based on one or more category fields.
- Makes several minor formatting changes to the regression report.
- Adds tools to the regression and scatterplot GUI allowing you to describe your model and to predict new observations using that model.
- Adds tools to the scatterplot GUI allowing you to modify and enhance the graphic attributes of the plot.

Version 3.1a (September 22, 2005)

- Again corrects a bug that produced an error message stating that “Assertion ‘Positive buffer size’ failed. This error appears to only affect ArcView 3.2a and earlier installations.

Version 3.1b (October 12, 2005)

- Corrects a bug which may say either “Variable theLinkText has not been initialized” or “Variable theConfBandTheme has not been initialized”, and which was related to generating a scatterplot without generating confidence bands.
- Corrects an issue related to Grid-based regression scatterplots, in which the scatterplot point X-coordinates were close, but not exactly equal to, the predictor values in the point attribute table.
- Corrects another issue related to Grid-based regression scatterplots in which the analysis was constrained within a polygon, in which the X- and Y-scales were incorrect.

Version 3.1c (February 18, 2006)

- Corrects a bug which appears to occur only on Asian or Chinese installations of ArcView. The problem was that I used some unusual characters in the Avenue code, which caused ArcView to crash when the extension was loaded on a Chinese computer.

Version 3.1e (August 29, 2006)

- Corrects a bug which appears when you conduct grid regression and do not choose to calculate confidence bands or intervals, producing a message stating that ArcView cannot convert a string to a number.
- Added functions to check the extension scripts. These functions are added as menu items in the View, Table and Project “Help” menus.

References:

- Abramowitz, Milton; Stegun, Irene A. 1972. Handbook of Mathematical Functions.
- Burkardt, John. 2001. PROB - Probability Density Functions. Sample Fortran code for calculating density functions. Available on-line at <http://www.psc.edu/~burkardt/src/prob/prob.html>.
- Croarkin, Carol; Tobias, Paul. (Date Unknown). Engineering and Statistics Handbook, available on-line at: (<http://www.itl.nist.gov/div898/handbook/>). National Institute of Standards and Technology. Visited August 13, 2005.
- Cliff, A.D. and J.K. Ord. 1973. Spatial Autocorrelation. Pion Limited. 178 pp.
- Cliff, A.D. and J.K. Ord. 1981. Spatial Processes: Models and Applications. 266 pp.
- de Graaf, G., F.J.B. Martin, J. Aguilar-Manjarrez & J. Jenness. 2003. Geographic Information Systems in fisheries management and planning. Technical manual. FAO Fisheries Technical Paper No. 449. Food and Agriculture Organization of the United Nations. Rome. 162p.
- Denis, Daniel J. 2000. The origins of correlation and regression: Francis Galton or Auguste Bravais and the error theorists? Paper presented at the 61st Annual Convention of the Canadian Psychological Association. Ottawa, Canada. Available at <http://www.york.ac.uk/depts/maths/histstat/bravais.htm>. Visited September 11, 2005.
- Draper, Norman R. and Smith, Harry. 1998. Applied Regression Analysis. 3rd ed. New York: John Wiley & Sons, Inc.; 706 pages. (Wiley Series in Probability and Statistics)
- Fotheringham, A. Stewart, Chris Brunsdon and Martin Charlton. 2002. Geographically Weighted Regression: The analysis of spatially varying relationships. John Wiley & Sons Ltd. 269 pp.
- Fotheringham, A. Stewart, Chris Brunsdon and Martin Charlton. 2000. Quantitative Geography: Perspectives on spatial data analysis. Sage Publications. 270 pp.
- Jeffrey, Alan. 2000. Handbook of Mathematical Formulas and Integrals (2nd Ed). Academic Press
- McLaughlin, Michael P. 2001. Regress+, Appendix A. A Compendium of Common Probability Distributions (version 2.3). Available on-line at: http://www.causaScientia.org/math_stat/Dists/Compendium.pdf. Visited August 13, 2005.
- Neter, John; Wasserman, William; Nachtschiem, Christopher J.; and Kutner, Michael H. 1996. Applied linear statistical models: regression, analysis of variance and experimental design. 4th ed. Burr Ridge, Illinois: McGraw-Hill/Irwin; 1408 pages.
- Ott, R. Lyman. 1993. An Introduction to Statistical Methods and Data Analysis (4th Ed). Duxbury Press.
- Press, William H.; Teukolsky, Saul A.; Vetterling William T.; and Flannery, Brian P. 1997. Numerical Recipes in C; the Art of Scientific Computing (2nd Ed). Cambridge University Press. (ISBN 0-521-43108-5). (<http://lib-www.lanl.gov/numerical/bookcpdf.html> - See Chapter 6, sections 1-4)
- Ripley, B.D. 1981. Spatial Statistics. John Wiley & Sons. Wiley Series in Probability and Mathematical Statistics. 252 pp.
- Salsburg, David A. 2001. The lady tasting tea – how statistics revolutionized science in the twentieth century. Owl Books; Henry Holt and Company, LLC. New York. 340 pp.
- S-Plus Help Files. 2001. S-Plus 6 for Windows. Product Information available at: <http://www.insightful.com/support/splus60win/default.asp>
- S-PLUS. 2005. S-PLUS 7 Guide to Statistics, Volume 1, Insightful Corporation, Seattle, WA.
- SPSS Help Files. 1999. SPSS for Windows Release 9. Product Information available at: <http://www.spss.com/>
- Stewart, James. 1998. Calculus, Concepts and Contexts. Brooks/Cole Publishing Company. p. 421-424
- Stanton, Jeffrey M. 2001. Galton, Pearson, and the peas: a brief history of linear regression for statistics instructors. Journal of Statistics Education 9(3).

Index:

A

Abramowitz, M.....94
accuracy.....80, 87
Aguilar-Manjarrez, J.....1, 2, 94
analysis boundary.....31, 32
ANOVA.....1, 16, 22, 30
ArcView GIS.....1, 2, 5, 6, 19, 30, 39, 57, 65, 90, 91, 92, 99
assumptions... ..8, 33, 53, 98
Avenue.....61, 66, 67, 69, 70, 71, 72, 73, 74, 75, 76, 77, 78,
79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89
average deviation.....62, 67

B

best fit.....6, 7, 8, 16, 21, 26
Box, George E. P.....6
Bravais, Auguste.....6, 94
Burkardt, J.....66, 94

C

calculate.....1, 23, 34, 43, 45, 47, 49, 55, 57, 59, 60, 62, 63,
64, 65, 66, 67, 68, 80, 87
calculator.....1, 5, 55, 65, 69, 70, 71, 72, 73, 74, 75, 76, 77,
78, 79, 80, 81, 82, 83, 84, 85, 86, 87, 88, 89
causality.....6
Charlton, M.....54, 94
classification.....59
Cliff, A.D.....54, 94
coefficient of multiple determination (R^2)8
confidence bands.....8, 18, 22, 23, 29, 30, 38, 49, 50
confidence interval... 1, 8, 16, 21, 29, 43, 44, 45, 49, 50, 58,
62, 67, 68
confidence limit.....21, 23, 43, 45, 47, 62, 68
crash.....30, 91
critical value.....1
Croarkin, C.....66, 70, 77, 84, 94

D

data snooping.....26
de Graaf, G.....1, 94
decimal precision.....36, 37
Denis, Daniel J.....6, 94
Density Functions
Cumulative Density Functions.....65, 75, 80, 87
Inverse Density Functions.....65, 82
Probability Density Functions.....65, 69, 94
describe model.....24, 35, 93
descriptive statistics.....16, 17, 29

Distribution

Beta.....1, 65, 69, 75, 82
Binomial.....1, 65, 69, 70, 76, 83, 98
Cauchy.....1, 65, 70, 74, 76, 77, 81, 83, 84, 88
Chi-Square.....1, 65, 70, 71, 77, 78, 84, 85
Exponential.....1, 6, 9, 10, 17, 66, 71, 77, 78, 84, 85
F.....1, 16, 44, 45, 47, 55, 58, 66, 67, 71, 72, 78, 84, 85,
94
Logistic.....1, 66, 72, 78, 79, 86, 98
LogNormal.....1, 66, 72, 73, 79, 86, 87
Normal... 1, 58, 59, 66, 70, 73, 74, 77, 80, 81, 84, 87, 88
Poisson.....1, 66, 74, 80, 87, 88
Student's T.....66, 74, 81, 88
Weibull.....1, 66, 75, 81, 82, 89
Distribution Functions.....66, 69
Dooley, J.....2
Draper, N.R.....6, 9, 22, 28, 53, 94

E

error.. 7, 8, 17, 30, 43, 44, 56, 58, 61, 66, 68, 90, 91, 92, 93,
94

F

FAO.....1, 2, 94
field calculator.....55
Fisher, Ronald Aylmer... ..1, 6, 59, 63, 68
Fisher.....1, 6, 59, 63, 68
Flagstaff.....1, 2, 99
Flannery, B.P.....94
font attributes... ..35, 36, 37, 38, 39
Fotheringham, A.....54, 94

G

Galton, Francis... ..6, 94
Gauss, Carl Friedrich.....6
geographically weighted regression.. ..54, 94
GRD ERROR.....30
grid mask.....31, 32
grid theme.....47, 48

H

histogram.....2, 18, 58, 60, 61, 64

I

independence.. 1, 6, 8, 16, 17, 21, 22, 23, 29, 30, 50, 53, 70,
77, 84, 98
Inland Water Resources and Aquaculture Service.....1, 2
interpolation.....53

J

Jeffrey, A.....94
Jenness, J.S..... 1, 2, 94, 99

K

Kendall, L.....2
kurtosis.. 1, 59, 63, 67, 68
Kutner, M.H.....28, 53, 94

L

lack of independence..53
least squares..... 1, 6, 7, 8
Legendre, Adrien-Marie...6
linking tables.....90
LISTBOX_SELECTION_MULTIOROW.....90

M

manage data sources.....41
map calculator.....56
Marttin, F.J.B..... 1, 94
maximum..... 1, 29, 30, 52, 57, 58, 62, 67, 68
McLaughlin, M.P.....66, 69, 75, 94
mean.. .1, 6, 8, 18, 21, 22, 23, 27, 29, 30, 34, 43, 44, 50, 57,
58, 61, 62, 64, 67, 68, 70, 71, 72, 73, 74, 77, 78, 79, 80,
84, 85, 86, 87, 88
median.. 1, 29, 58, 62, 67, 68
minimum..... 1, 29, 52, 57, 58, 62, 67, 68
mode.. 1, 29, 58, 59, 64, 68
model builder.....17
models
2nd Order..... 10, 17, 21
3rd Order.....9, 10, 17
complex..... 8, 9, 10, 27, 51
defining.....17
exponential..... 1, 6, 9, 10, 17, 66, 71, 77, 78, 84, 85
inverse.....1, 9, 10, 15, 17, 43, 65, 82, 83, 84, 85, 86, 87,
88, 89, 93
linear.....9, 25
natural log.....9, 17, 55, 56, 73, 79, 86
MSE.....43
multiple linear regression.....1, 6

N

Nachtschiem, C.J..... 28, 53, 94
Nefisco Foundation..... 1
Neter, J...6, 9, 28, 43, 53, 94

O

Ord, J.K..... 54, 94
Ott, R.L..... 94

P

parameter.. 1, 8, 9, 16, 21, 29, 30, 43, 44, 47, 65, 66, 71, 73,
75, 77, 79, 81, 84, 86, 89
Pearson, Karl... 6, 94
planned modifications..... 98
point theme..... 91
polygon theme..... 1, 32
polygon..... 1, 32
polynomial
cubic..... 10
quadratic..... 10
precision..... 57
predicted values.....16, 18, 22, 23, 29, 30, 47
predicting new observations... .5, 21, 22, 24, 35, 43, 44, 45,
47, 49, 50, 51, 52, 93
Press, W.H..... 66, 94
probability.. 1, 2, 5, 6, 22, 28, 30, 54, 65, 66, 67, 69, 70, 71,
72, 73, 74, 75, 76, 77, 78, 79, 80, 81, 82, 83, 84, 85, 86,
87, 88, 89, 94
proportion..... 23, 26, 30
p-value..... 22, 25, 27, 84

Q

quartile.....1, 6, 58, 62, 67, 68

R

R²..... 1, 8, 16, 22, 25, 26, 27, 30, 33, 53
range..... 1, 6, 8, 17, 18, 25, 29, 50, 51, 52, 58, 62, 68
refresh..... 5, 38, 58, 64
residuals.. 30
restricting number of points..... 30
Ripley, B.D... 54, 94
Riva, C..... 2
Rocky Mountain Research Station..... 2

S

Salsburg, D..... 6, 94
sample size..... 8, 43

scatterplot....1, 5, 8, 9, 16, 18, 19, 23, 25, 27, 30, 31, 33, 34,
35, 36, 37, 38, 39, 40, 41, 43, 44, 91, 93
Segmentation Violation..90
selection, clear..55
sequential sums of squares.....16, 23
simple linear regression..6, 17, 21, 51
singular matrix.....17, 56, 91
skewness.....1, 58, 62, 63, 67, 68
slope.....1, 6, 21, 53
Smith, H.....6, 9, 22, 28, 53, 94
Spatial Analyst.....2, 91, 92
spatial autocorrelation.....8, 53, 54, 94
S-Plus.....94
SPSS.....55, 94
s-shaped curve..14, 15
standard deviation....1, 6, 21, 29, 30, 43, 44, 45, 47, 58, 62,
67, 68, 70, 73, 80, 87
standard error of the mean..1, 58, 67
standardized residuals.....16, 18, 23, 30
Stanton, Jeffrey M.....6, 94
Stegun, I.A.....94
Stewart, C.B.....54, 80, 87, 94
Stewart, J.....54, 80, 87, 94
summary statistics.....1, 18, 57, 58, 60, 61, 67
syntax error.....90

T

Teukolsky, S.A.....94
text labels.....35
Tobias, P.....66, 70, 77, 84, 94
transforming variables.....55, 98
transformations.....10, 17, 46, 48, 55, 56, 91, 93, 98
transforming grids.....56
troubleshooting.....17, 30, 56, 90

U

unlink.....91

V

variables
dependent....1, 6, 8, 16, 17, 21, 22, 23, 29, 30, 33, 34, 81
independent.....1, 6, 8, 16, 17, 21, 22, 23, 29, 30, 50, 53,
70, 77, 84
predictor..1, 6, 8, 9, 10, 17, 18, 24, 29, 30, 35, 43, 44, 45,
46, 47, 48, 49, 50, 52, 53, 55, 91, 93, 98
response..1, 6, 8, 10, 11, 12, 13, 14, 15, 18, 29, 50, 55,
56, 91, 98
variance.....1, 6, 8, 22, 23, 28, 29, 30, 33, 58, 62, 67, 68, 94
Vetterling, W.T.....94

W

Walker, S.....2
Wasserman, W... ..28, 53, 94

X

X-Axis.....5, 18, 30, 36

Y

Y -Axis.....5, 37

Z

zonal statistics.....91

Planned Modifications:

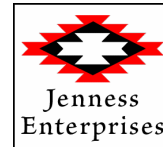
1. SEMI-VARIOGRAMS: Write functions to generate semi-variograms for themes and grids. These semi-variograms are extremely useful for determining how spatially autocorrelated the data are, and what sample point separation distance is best to avoid that autocorrelation and therefore meet the assumption of sample independence.
 2. LOGISTIC REGRESSION: Include functions to perform binomial logistic regression, which is a very useful method to analyze binomial phenomena (such as whether a point on the landscape is or is not useful habitat).
 3. SPATIALLY-WEIGHTED REGRESSION: I would like to learn more about spatially weighted regression and make it available within this extension.
 4. TRANSFORM RESPONSE VARIABLE: The extension currently offers several transformations of predictor variables, but sometimes the best model is one which transforms both the predictors and the response. You can always transform the response variable manually prior to running the analysis (see *Manually Transforming Variables* on p. 55), but it would be more convenient to do it within the analysis itself.
 5. Add AIC (Akaike Information Criterion) values to reports.
-

Enjoy! Please contact the author if you have problems or find bugs.

Jeff Jenness

Jenness Enterprises
3020 N. Schevene Blvd.
Flagstaff, AZ 86004
USA

jeffj@jennessent.com
<http://www.jennessent.com>
(928) 607-4638



Updates to this extension and an on-line version of this manual are available at

<http://www.jennessent.com/arcview/regression.htm>

Please visit *Jenness Enterprises* [ArcView Extensions](http://www.jennessent.com/arcview/extensions) site for more ArcView Extensions and other software by the author. We also offer customized ArcView-based [GIS consultation](http://www.jennessent.com/gisconsultation) services to help you meet your specific data analysis and application development needs.

