Chapter 8

Univariate Regression

Software and Documentation by:

Carla Mae Bann

This chapter presents ViSta-Regress, the ViSta procedure for performing univariate multiple and simple regression. ViSta-Regress can perform ordinary least squares, monotonic, and robust univariate multiple regression. The visualization and report methods of ViSta-Regress include regression diagnostics and plots for examining the structure of the data and for checking model assumptions.

8.1 Introduction

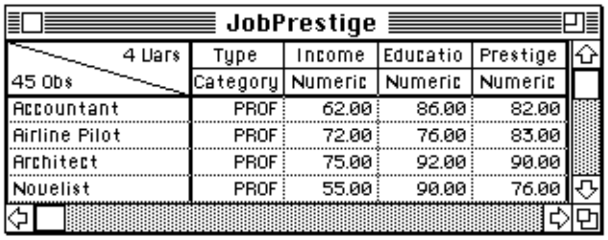
Multiple regression is a widely used data analysis technique. It allows researchers to predict one response variable using two or more predictor variables. The most commonly used type of multiple regression is ordinary least squares regression (OLS). The OLS regression model is represented as $Y = \beta X + \text{error}$ where Y are the actual values of the response variable, β is a vector of regression coefficients, and X is a matrix of predictor variables.

The ordinary least squares regression model makes the following assumptions: (1) the errors are normally distributed with mean 0 and constant variance; (2) errors are independent; and (3) the independent variables are measured without error. OLS regression is the most efficient method when the assumptions of the model are satisfied. It produces coefficients which have the minimum variance of all linear unbiased estimators. However, when the data fail to meet the assumptions of the OLS model or when there are problems such as outliers or colinearity present in the data, the OLS regression coefficients will be inaccurate. These coefficients may not provide good estimates of the population parameters.

Regression diagnostics may be used to determine whether data violate the OLS model assumptions. If the data fail to meet the assumptions of OLS regression, alternative regression methods may be used to analyze the data. One alternative regression method which is useful when the data contain outliers is robust regression. Robust regression methods produce regression coefficients which are not influenced by outliers. Monotonic regression is another alternative method which is useful when the relationship between the response variable and the predictor variables is nonlinear.

The ViSta-Regress program makes identifying outliers and checking the OLS regression model assumptions easy for all levels of users. Also, it provides easy access to monotonic and robust regression methods. This program allows users to compare the results from an ordinary least squares regression, robust regression, and monotonic regression, thereby enabling researchers to determine which technique is most appropriate for their data.

An example will be used to demonstrate the various features and uses of the ViSta-Regress program. The data for this example consist of a small dataset of 45 occupations reported by Duncan (1961). The first four observations of this dataset are displayed in Figure 1. For each of the occupations, information is reported on the income level, education, and prestige of individuals in the occupation. Income is operationalized as the percentage of males in the occupation earning more than $3500 in 1950. The Education variable measures percentage of males in the occupation who were high school graduates in 1950 and Prestige is expressed as the percentage of raters rating the occupation as excellent or good in prestige. Income and education are the predictor variables and prestige is the response variable.



4 Vars 45 Obs	Type	Income	Educatio	Prestige
	Category	Numeric	Numeric	Numeric
Accountant	PROF	62.00	86.00	82.00
Airline Pilot	PROF	72.00	76.00	83.00
Architect	PROF	75.00	92.00	90.00
Novelist	PROF	55.00	90.00	76.00

Figure 1: Jobs Datasheet

The analysis of Duncan's jobs data is described fully in Young & Bann (in press). Briefly, the data were first analyzed using ordinary least squares regression. After examining the diagnostic plots, it appeared that the data contained outliers. Robust regression suggested that four of the observations were outliers. Next, a monotonic regression was performed and it appeared that there was no systematic curvilinearity in the data. Finally, based on the results of these analyses, the four outliers (Minister, Reporter, Railroad Conductor, and Railroad Engineer) were removed from the dataset and an OLS regression was performed on the remaining observations.

The plot shown in Figure 2 is the result of this analysis. This plot is a scatterplot of the actual values versus the predicted values of the response variable. The 45-degree angle line through the plot may be used to determine how well the regression model predicts the observations. Observations, such as Welfare Worker, which fall on the line have the same predicted and actual values, so the model is able to predict those observations perfectly. However, the model does not provide a good fit to points which are far from the line, such as Insurance Agent and Machinist. Those observations have predicted values which differ from their actual values.

8.2 Using the Regression Analysis Procedure

ViSta-Regress can perform three kinds of multiple regression: Ordinary least squares (OLS) univariate multiple regression; robust univariate multiple regression; and monotonic univariate multiple regression. It also contains several regression diagnostics and plots. The following sections describe how to use the ViSta-Regress program.

8.2.1 Open Data

The jobs data shown in Figure 1 can be loaded into ViSta using the **Open Data** or **Load Data** menu items.

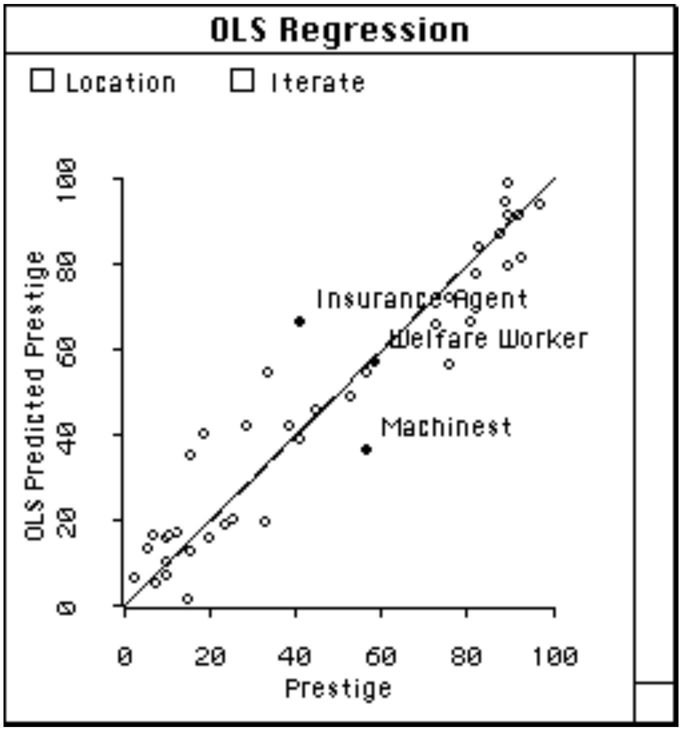
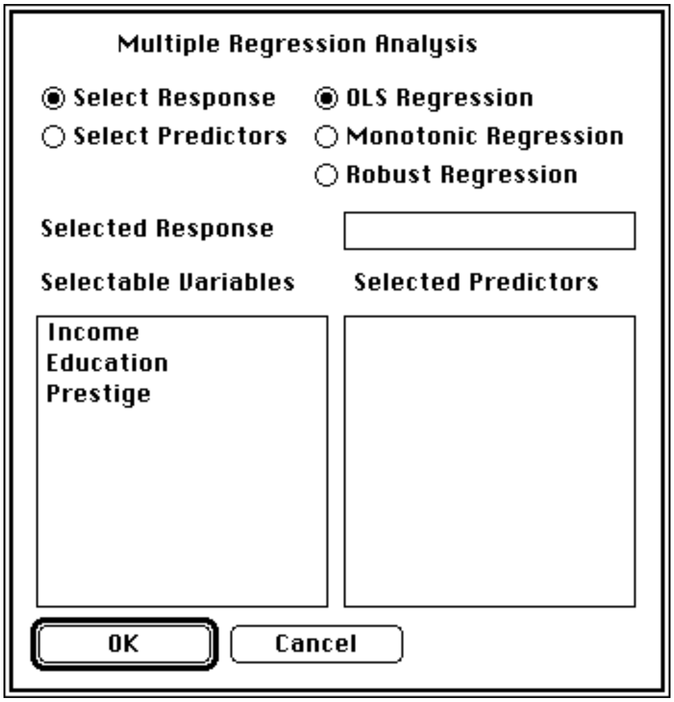


Figure 2: OLS Regression Plot of Jobs Data

8.2.2 Analysis Options

You can perform an ordinary least squares regression analysis by selecting the **Regression Analysis** item from the **Analyze** menu, by using the Regress button on the toolbar, or by typing the `(regression-analysis)` function in the listener window. As a result, the options dialog box shown in Figure 3 will be displayed. The response variable may be selected by clicking on the button which reads **Select Response Variable** and then clicking on the appropriate variable in the box labeled **Variables**. This variable name will then be entered into the **Response Variable** box. Then, to select the predictor variables, click on the button labeled **Select Predictor Variables** and then click on the appropriate variables in the **Variables** box. Once the variables are selected you perform OLS regression analysis by clicking on the **OK** button.

There are three possible ways for the user to select robust or monotonic regression. First of all, you can choose the desired method in the dialog box. Second, you can ask for a visualization and then click on the Iterate button on the Regression plot. Finally, you can type the regression-analysis function into the listener window and use the `:method` keyword argument, as explained below.



Multiple Regression Analysis

● Select Response
○ Select Predictors

● OLS Regression
○ Monotonic Regression
○ Robust Regression

Selected Response

Selectable Variables
Income
Education
Prestige

Selected Predictors

Figure 3: Options Dialog Box

As mentioned earlier, the user may perform a regression analysis by typing `(regression-analysis)` into the listener window. The `(regression-analysis)` method with its keywords and their default values is:

```
(regression-analysis
   :data current-data
   :title “Multiple Regression Analysis”
   :name (strcat “REG-”(send current-data :name))
   :method “OLS”
   :iterations 20
   :max-rsq 1
   :min-rsq-improve .001
   :response nil
   :predictors nil)
```

To perform a regression-analysis using the default values, simply type (regression-analysis). To change the default values, any or all of these keywords may be used.

`:data`	The data keyword is used to change the name of the dataset to be analyzed. The dataset name is entered without quotes.
`:title`	The `:title` keyword is used to specify a title for the ViSta report. The title is entered with quotes.
`:name`	The argument for the `:name` keyword is the name of the model object which is created.
`:method`	The `:method` keyword is used to specify the type of regression analysis to be performed. The possible arguments are “OLS”, “Robust”, and “Monotonic.”
`:iterations`	The argument to the `:iteration` keyword is the number of iterations to be performed for either monotonic or robust regression.
`:max-rsq`	The `:max-rsq` keyword specifies the maximum model R^2 which is to be obtained with the monotonic regression.
`:min-rsq-improve`	The argument to `:min-rsq-improve` is the minimum amount that the model R^2 should increase from iteration to the next of the monotonic regression.
`:response`	The argument to the `:response` keyword is a list containing the name of the response variable in quotes.
`:predictors`	The argument to the `:predictors` keyword is a list containing predictor variable names in quotes.

For example, to perform a robust regression analysis on Duncan’s jobs data using Income and Education as the predictor variables and Prestige as the response variable and changing the number of iterations to 10, type:

```
(regression-analysis
     :method “Robust”
     :iterations 10
     :response (list “Prestige”)
     :predictors (list “Income” “Education”))
```

8.2.3 Report

A report of the analysis may be obtained by selecting the **Report Model** item from the **Model** menu, by typing `(report-model)`, or by sending the model object the `:report` message. The Report dialog box shown in Figure 4 will then appear. A standard ViSta regression report can be obtained by simply clicking on **OK**. This report contains the parameter estimates with two-tailed t tests, fit statistics, and a test of the model. Diagnostics such as the autocorrelation, variance inflation factors, and DFFITS are also included in this report.

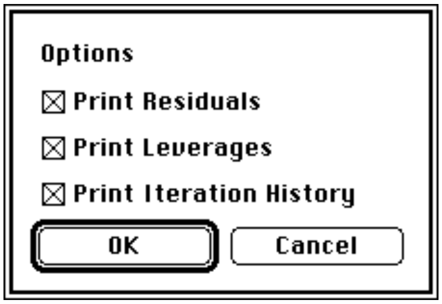


Figure 4: Report Dialog Box

Other information may be obtained by selecting any or all of the three options in the dialog box. The user may select **Print Leverages** to obtain a list of the leverages and Cook's distances for each observation. When **Print Residuals** is selected a list of the raw, studentized, and externally studentized residuals for each observation is displayed. **Print Iteration History** displays a list of the model R^2 and change in R^2 for each iteration for the robust or monotonic methods.

8.2.4 Statistical Visualization

You can obtain a visualization of the analysis results by selecting **Visualize Model** item from the **Model** menu or by typing (visualize-model) in the listener window. The visualization consists of eight plots, six of which may be seen simultaneously in a six cell spreadplot. Different plots are available depending on the type of regression method used. The following sections describe the plots that are available for each method.

8.2.4.1 OLS Regression Plots

The **OLS Regression Plot** is a plot of the predicted values of the response variable versus the actual values of the response variable. A line at the 45-degree angle is drawn through the plot. Points falling on this line have the same predicted and actual value. This plot also has a button labeled **Iterate** which may be used to select either robust or monotonic regression.

The **Added Variable Plot** shows the relationship between the response and a single predictor, controlling for the effects of all of the other predictors. If the plot is linear, then there is a linear relationship between the response and the plotted predictor, an assumption underlying the OLS regression analysis. The plot is also useful for determining the contribution of a particular predictor variable. If the plot is linear the variable contributes to the relationship, but if the plot shows no pattern the variable does not contribute. The **Y-axis** button allows the user to select a different predictor variable.

The two **Influence** plots are plots of influence statistics versus the predicted values of the response variable. These plots are designed to reveal outliers. They will appear as points that are separated from the main point-cloud. The **Y-axis** button allows the information shown on the Y-axis to be changed. The user has an option of leverages or Cook's distances.

The two **Residual** plots are plots of residuals versus the predicted values of the response variable. The **Y-axis** button may be used to change the type of residuals displayed in the plot. The user may select raw residuals, Bayes residuals, studentized residuals, or externally studentized residuals.

The **Observation** window contains a list of the observations. Clicking on one of the observations in the window will highlight the points in the other graphs that correspond to that observation.

Each of the graphs contain a button labeled **Location.** This button is used to change the location of the plots. When the user clicks on this button, a dialog box appears (Figure 5). Each of the options corresponds to the section of the screen where the plot will be displayed. For example, if Upper Left is selected, the plot will be moved to the upper left corner of the screen.

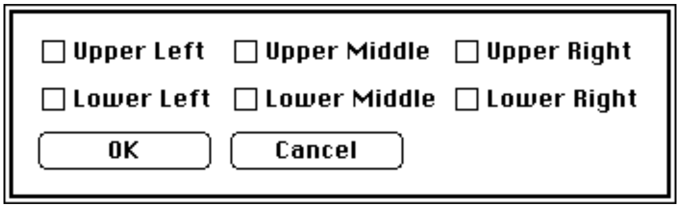
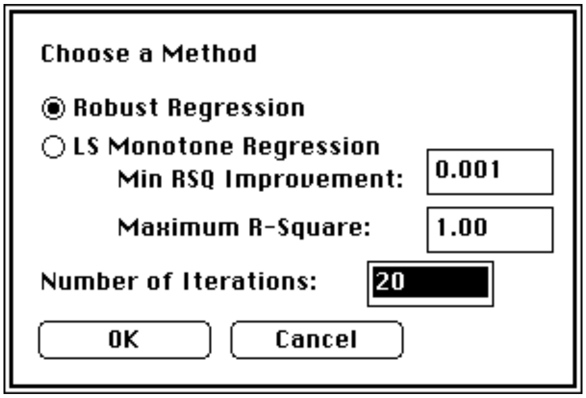


Figure 5: Location Dialog Box

Each window contains a close box in the upper lefthand corner. After clicking on this box, a dialog box will appear which offers the user the choice of closing that specific plot or all of the plots. If the user closes all of the plots, he/she may view the plots again by selecting **Visualize Model** under the **Model** menu.

If the user wishes to perform a robust regression or monotonic regression, he/she can click on the **Iterate** button located in the Linear Regression plot. The dialog box shown in Figure 6 will appear. The user is given the option of either a robust or monotonic regression. The user may specify the number of iterations to be performed. Also, if the user chooses Monotonic regression, he/she may choose to specify the maximum R^2 or minimum R^2 improvement. The program will then calculate the new regression and the plots in the spreadplot will change to display the new information. The dialog boxes associated with the **Y-axis** buttons on the residuals and influence plots will also be changed so there is an option of viewing either the information for the ordinary least squares regression or the new regression method.



Choose a Method

● Robust Regression

○ LS Monotone Regression

Min RSQ Improvement: 0.001

Maximum R-Square: 1.00

Number of Iterations: 20

Figure 6: Iterate Dialog Box

8.2.4.2 Robust Regression Plots

If the user has selected robust regression, the **OLS Regression Plot** will no longer be shown and a **Robust Regression Plot** will be displayed instead. To view the **OLS Regression Plot** again, the user may click on **OLS Regression** under the **Spreadplot** menu. Also, in place of the **Added Variable Plot**, a plot of **Robust Regression Weights** will be shown.

The **Robust Regression Plot** is a plot of the predicted values found using the robust regression versus the actual values of the response variable. The 45-degree angle line through the plot is useful for determining how close the actual and predicted values are.

The **Robust Weights Plot** displays the regression weights calculated on each iteration. This enables the user to track how the weight assigned to an observation has changed from one iteration to the next. When the user clicks on a point at the end of one of the weight lines, the name of the observation corresponding to that weight will be highlighted in the Observation window.

8.2.4.3 Monotonic Regression Plots

If monotonic regression was selected, the **OLS Regression Plot** will be replaced with a **Transformation Plot**, the second influence plot will be replaced with an **R-Squared/Beta Plot**, and the **Added Variable Plot** will be replaced with the **Predictor Variables** window.

The **Transformation Plot** is a plot of the predicted values from the monotonic regression versus the actual values of the response variable. There are also two lines drawn on the plot. The first line is a 45-degree angle line which allows the user to determine if the observation has the same predicted and actual values. The second line is a line of the optimally scaled response variable and the actual response variable. Comparing these two lines is useful for determining if the data contain any curvilinearity.

The **Predictor Variables** window contains a list of the predictor variables. If the user clicks on one of the variable names in this window, the point in the R^2/Beta plot which corresponds to that variable will be highlighted.

The **R2/Beta Plot** displays the values for the model R^2 and the standardized regression coefficients for each iteration. This plot allows the user to determine how these values have changed from one iteration to the next. When the user clicks on a point at the end of one of the Beta lines, the name of the variable corresponding to that regression coefficient will be highlighted in the Predictor Variables window.

8.3 Algorithm

8.3.1 Ordinary Least Squares Regression Algorithm

Ordinary least squares multiple regression consists of fitting this model:

$$Y = \beta_0 + \sum_i \beta_i X_i + error \qquad \textbf{(EQ 1)}$$

where β_i are the parameter estimates and X_i are the predictor variables. The parameter estimates (β_i) are chosen as those values that minimize the sum of squared residuals, $\sum(\hat{Y} - Y)^2$ where $\hat{Y} = \beta_o + \sum_i \beta_i X_i$ is the predicted value of the response variable and Y is the actual value of the response variable.

8.3.2 Robust Regression Algorithm

The purpose of robust regression is to produce regression coefficients which are not sensitive to outliers and nonnormal error distributions. Robust regression methods usually accomplish this by performing weighted least squares regression and downweighting observations which deviate from the majority of the data.

The robust regression section of the ViSta-Regress program uses the biweight estimator developed by John Tukey (Mosteller & Tukey, 1977). This estimator is one type of M-estimator. M-estimators produce regression coefficients which minimize a function of the residuals (ρ). The ρ function of the biweight estimator is:

$$\rho(u_i) = \left(\frac{c^2}{3}\right)\left(1 - \left[1 - \left(\frac{u_i}{c}\right)^2\right]^3\right) \text{ if } |u_i| \le c \qquad \textbf{(EQ 2)}$$

$$\rho(u_i) = \frac{c^2}{3} \text{ if } |u_i| > c \qquad \textbf{(EQ 3)}$$

where u_i is the residual of the ith observation and c is a tuning constant.

The robust regression weights are calculated using this formula:

$$w_i = \frac{\psi(u_i)}{u_i} \qquad \textbf{(EQ 4)}$$

where ψ is an influence function which determines the amount of influence an observation has on the calculation of the regression coefficients. It is calculated as the first derivative of ρ. The formula for the ψ function of the biweight estimator is:

$$\psi(u_i) = u_i\left[1 - \left(\frac{u_i}{c}\right)^2\right]^2 \text{ if } |u_i| \le c \qquad \textbf{(EQ 5)}$$

$$\psi(u_i) = 0 \text{ if } |u_i| > c \qquad \textbf{(EQ 6)}$$

This formula shows that when a residual is greater than the tuning constant (c), that observation is given a weight of 0 and it is essentially dropped from the analysis. The larger the tuning constant chosen, the fewer observations that will be dropped from the analysis and the smaller the tuning constant, the more observations that will be dropped.

The tuning constant also determines the robustness of the estimator. If a tuning constant of 4.685 is used, the biweight estimator is 95% efficient when sampling from a normal population (Hamilton, 1992). The ViSta-Regress program uses 4.685 as the tuning constant.

The formulas shown above reveal a problem when calculating a robust regression. The regression coefficients are needed to calculate the weights and the weights are needed to find the regression coefficients. To solve this problem, the iteratively reweighted least squares algorithm is used (Holland & Welsh, 1977). The steps of this algorithm are:

(1) Obtain initial estimate of regression coefficients using OLS regression

(2) Use the coefficients to calculate the weights

(3) Use the weights to perform a weighted least squares regression and obtain new estimates of the regression coefficients

(4) Repeat steps 2 and 3 until a convergence criterion has been met.

The convergence criterion used in ViSta-Regress is the maximum relative change in the regression coefficients. The default is that the loop will terminate when the maximum relative change is less than .001.

8.3.3 Monotonic Regression Algorithm

Monotonic regression allows a researcher to calculate a multiple regression using a response variable which is nonlinearly related to the predictor variables. The algorithm for monotonic regression uses an Alternating Least Squares Optimal Scaling method (ALSOS) developed by Young et al. (1976). The procedure is an iterative process which alternates between two steps, a parameter estimation step and an optimal scaling step. In the parameter estimation step, the data are assumed to be constant and parameter estimates are calculated which maximize the fit between the model and the data. In the optimal scaling step, the parameter estimates are held constant while scores are assigned to the data. The ViSta-Regress monotonic regression procedure may be used only with data at the ordinal level of measurement.

The optimal scaling procedure monotonically transforms the raw data observations in order to obtain the optimally scaled data. The observations must be assigned scores which are monotonic with the raw data. In ViSta-Regress only the response variable is optimally scaled and it is assumed to be at the ordinal level. Therefore, the order of the optimally scaled response variable must be the same as the order of the original response variable.

In order to ensure that the optimally scaling procedure retains the order of the variable, the following procedure is used. First, the response variable observations are ordered from smallest to largest. The first observation (the smallest value) is assigned the value of its predicted value. Then the predicted value for each observation is compared to the score assigned to the preceding observation. If it is larger than the preceding observation, then it is set equal to its predicted value. However, if it is smaller than the preceding observation, then the mean of the predicted values for both observations is calculated. These two observations form a block and are both assigned the value of the mean. Then the mean is compared to the score assigned to the observation preceding both of these observations. If the mean is larger than the preceding observation, then nothing is done. However, if the mean is smaller than the value for the preceding observation, that observation must be added to the block and the three observations are set equal to the mean of their predicted values. This continues until all of the observations have been scored.

The optimal scaling procedure may also be expressed by the following matrices:

$$(U^TU)^{-1}U^TY \qquad \textbf{(EQ 7)}$$

(Young et al., 1976). The indicator matrix, U, is an (n x n_b) matrix with a row for each of the n observations and a column for each of the n_b blocks needed to maintain the order restriction. This matrix contains 0's and 1's which indicate the blocks of observations which must be merged to maintain order. U^TU is a diagonal (n_b x n_b) matrix with a row and column for each block and diagonal values equal to the number of observations in each block. U^TY is an n_b column vector which contains the sum of the y's. The product $(U^TU)^{-1}U^T$ forms an n_b column vector with the mean of the appropriate y's as its elements.

These are the steps of the monotonic regression algorithm:
(1) Initial estimates: Use OLS regression coefficients as starting parameter values.
(2) Check convergence criterion. If it has been met, end the loop.
(3) Optimal Scaling of Response Variable: Solve for U in
(4) OLS regression on optimally scaled data, getting new parameter estimates.
(5) Return to step 2.