Step 5 – Record Your Results

Since multiple regression can be quite involved, it is best make notes of why you did what you did at different steps of the analysis. Jot down what decisions you made and what you have found. Explain what you did, why you did it, what conclusions you reached, which outliers you deleted, areas for further investigation, and so on. Be sure to examine the following sections closely and in the indicated order:

1. Analysis of Variance Section. Check for the overall significance of the model.
2. Regression Equation and Coefficient Sections. Significant individual variables are noted here.

Regression analysis is a complicated statistical tool that frequently demands revisions of the model. Your notes of the analysis process as well as of the interpretation will be worth their weight in gold when you come back to an analysis a few days later!

Multiple Regression Technical Details

This section presents the technical details of least squares regression analysis using a mixture of summation and matrix notation. Because this module also calculates weighted multiple regression, the formulas will include the weights, w_j. When weights are not used, the w_j are set to one.

Define the following vectors and matrices:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_j \\ \vdots \\ y_N \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 \; x_{11} \cdots x_{1p} \\ \vdots \\ 1 \; x_{1j} \cdots x_{pj} \\ \vdots \\ 1 \; x_{1N} \cdots x_{pN} \end{bmatrix}, \quad \mathbf{e} = \begin{bmatrix} e_1 \\ \vdots \\ e_j \\ \vdots \\ e_N \end{bmatrix}, \quad \mathbf{1} = \begin{bmatrix} 1 \\ \vdots \\ 1 \\ \vdots \\ 1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix}$$

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & \ddots & 0 & 0 & \vdots \\ 0 & 0 & w_j & 0 & 0 \\ \vdots & 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 0 & w_N \end{bmatrix}$$

Least Squares

Using this notation, the least squares estimates are found using the equation.

$$\mathbf{b} = (\mathbf{X'WX})^{-1}\mathbf{X'WY}$$

Note that when the weights are not used, this reduces to

$$\mathbf{b} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$$

The predicted values of the dependent variable are given by

$$\hat{\mathbf{Y}} = \mathbf{b'X}$$

The residuals are calculated using

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$$

Estimated Variances

An estimate of the variance of the residuals is computed using

$$s^2 = \frac{\mathbf{e'We}}{N - p - 1}$$

An estimate of the variance of the regression coefficients is calculated using

$$\mathrm{V}\begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix} = s^2\left(\mathbf{X'WX}\right)^{-1}$$

An estimate of the variance of the predicted mean of Y at a specific value of X, say X_0, is given by

$$s^2_{Y_m|X_0} = s^2\left(1, X_0\right)\left(\mathbf{X'WX}\right)^{-1}\begin{pmatrix} 1 \\ X_0 \end{pmatrix}$$

An estimate of the variance of the predicted value of Y for an individual for a specific value of X, say X_0, is given by

$$s^2_{Y_I|X_0} = s^2 + s^2_{Y_m|X_0}$$

Hypothesis Tests of the Intercept and Slopes

Using these variance estimates and assuming the residuals are normally distributed, hypothesis tests may be constructed using the Student's t distribution with N - p - 1 degrees of freedom using

$$t_{b_i} = \frac{b_i - \mathrm{B}_i}{s_{b_i}}$$

Usually, the hypothesized value of B_i is zero, but this does not have to be the case.

Confidence Intervals of the Intercept and Slope

A $100(1-\alpha)\%$ confidence interval for the true regression coefficient, β_i, is given by

$$b_{\mathrm{i}} \pm \left(t_{1-\alpha/2,N-p-1}\right)s_{b_i}$$

Confidence Interval of Y for Given X

A $100(1-\alpha)\%$ confidence interval for the mean of Y at a specific value of X, say X_0, is given by

$$b'X_0 \pm \left(t_{1-\alpha/2,N-p-1}\right)s_{Y_m|X_0}$$

A $100(1-\alpha)\%$ prediction interval for the value of Y for an individual at a specific value of X, say X_0, is given by

$$b'X_0 \pm \left(t_{1-\alpha/2,N-p-1}\right)s_{Y_I|X_0}$$

R^2 (Percent of Variation Explained)

Several measures of the goodness-of-fit of the regression model to the data have been proposed, but by far the most popular is R^2. R^2 is the square of the correlation coefficient between Y and $\hat{Y}$. It is the proportion of the