Text Version

Expand Print

Building a Multiple Regression Model

In this demonstration, we build a multiple linear regression model by using PROC GLMSELECT.

Notice that this program starts by specifying a title for the results. The TITLE statement is followed by two SAS procedures. PROC GLMSELECT builds the model and PROC UNIVARIATE performs further analysis that enables us to determine whether the assumptions of the analysis were met.

Let's take a closer look at the PROC GLMSELECT step. The PROC GLMSELECT statement specifies the input data set school.

The MODEL statement specifies the response (**Reading3**) and the predictors (**Words1**, **Letters1**, and **Phonics1**). Remember that PROC GLMSELECT uses the stepwise method, by default, to select variables for a model. However, we want the model to contain the three predictors that are specified. The SELECTION=NONE option turns off the automatic selection of variables.

Finally, in the OUTPUT statement, the OUT= option saves a temporary data set named out. This data set contains the residuals (indicated by the statistic keyword R), which are stored in a variable named **Residuals**. After we build the model, we'll use this output data set as the input data set for the PROC UNIVARIATE analysis.

Now we submit this code.

```
title 'School Data: Regression and Diagnostics';
proc glmselect data=mydata.school;
          model reading3 = words1 letters1 phonics1 / selection=none;
          output out=out r=residuals;
run;
```

Let's look at the results.

In the first set of tables, notice that the number of observations used is less than the number of observations read. This difference indicates that SAS detected some missing values in the data set, and did not use the incomplete cases to build the model. The Dimensions table shows the number of effects and the number of parameters considered. The Least Squares Summary table displays the value of SBC (Schwarz's Bayesian criterion) at each step when an effect enters the model. Notice that the optimal value of the criterion occurs when the last variable, **Phonics1**, is added.

The Analysis of Variance table shows three sources of variation: the model for the fitted regression, the error for the residual error, and the corrected total for the total variation after correcting for the mean. In this table, note the following:

- The model has three degrees of freedom. You can think of degrees of freedom as the number of independent pieces of information, or the number of values in the final calculation of a statistic that are free to vary. Remember that this is the number of parameters minus one. In this example, there are four parameters: the intercept and three slopes (one for each of the independent variables).
- The sum of squares values for the term are as follows: the model sum of squares is 168543, the error sum of squares is 73453, and the corrected total sum of squares is 241996.
- The mean square is the sum of squares divided by the degrees of freedom. The mean square model is 56181 and is calculated by dividing the model sum of squares by the model degrees of freedom, which provides the average sum of squares for the model. The mean square error is 489.68756, which is an estimate of the population variance. SAS calculates this by dividing the error sum of squares by the error degrees of freedom, which provides the average sum of squares for the error.
- Next is the F value for testing the hypothesis that all parameters are zero except for the intercept. This is calculated by dividing the mean square for model by the mean square for error. Here, the F statistic is 114.73.
- The *p*-value is the probability of getting an F statistic greater than or equal to that observed if the hypothesis is true. Here, the *p*-value is less than 0.0001, which is smaller than any reasonable alpha level. Therefore, we reject the null hypothesis and conclude that at least one slope is not equal to zero. In other words, we can say that this model is better than the baseline model for predicting **Reading3**.

The next table displays the following fit statistics for the selected model:

- The root mean squared error is an estimate of the standard deviation of the error term. It is calculated as the square root of the mean square error.

- The dependent mean is the sample mean of the dependent variable. In this case, it is the mean of the **Reading3** score, which is 49.24026.
- The R square statistic, which has a value between 0 and 1, indicates the proportion of variance in the response that is accounted for by the model. It is also called the coefficient of determination. Here the value of R square is 0.6965, which means that approximately 69.65% of the variability in the response (**Reading3**) is explained by the model. As the value of R square approaches 1, the proportion of variability in the data that is explained by the independent variable increases. However, keep in mind that R square never decreases as you add more predictor variables to the model. Thus, R square can be misleading when used to compare the models, if the models have a different number of parameters.
- The adjusted R square is the R square adjusted for degrees of freedom. In other words, it takes into account the number of terms in the model. Therefore, the adjusted R square is a more appropriate statistic to use when comparing models. The adjusted R square for this model is 0.6904.
- Other fit statistics are displayed that can be used to compare models. These include the AIC (Akaike's information criterion), AICC (corrected Akaike's information criterion), and SBC.

Having established that the model is significant, we need to look further to decide which variables have slopes that are significantly different from zero. To do this, we examine the next table, which contains the parameter estimates and the corresponding t-test values. The *t*-values and *p*-values in the table test the null hypothesis that the slope for each independent variable is equal to zero. Presuming an alpha equal to 0.05, you reject the null hypothesis in each case. That is, all of these predictor variables are significant in predicting **Reading3**.

Based on the Parameter Estimates table, the estimated regression equation can be written as **Reading3** = -10.79366 + 0.93737 * **Words1** + 0.70787 * **Letters1** + 0.84782 * **Phonics1**.

You should be careful when you interpret the tests of hypothesis for the parameter estimates. They test the significance of each variable when it is added to a model that contains all of the other independent variables. As a result, if the independent variables in the model are correlated with one another, the significance of both variables can be hidden in these tests. Therefore, you should not remove more than one variable at a time from the model, based on these tests. Note that the significance level of the test does not depend on the order in which you list the independent variables in the model. It does depend on the variables included in the model.

Now that we have a modeling equation, are we finished? No, we should now verify that the assumptions of the analysis were met. Remember that the PROC GLMSELECT step contains an OUTPUT statement that saved the residuals in a data set named out, to be used for further analysis. In this PROC UNIVARIATE step, the out data set is specified as the input data set. PROC UNIVARIATE will produce histograms and normal quantile plots, as well as applying formal tests of normality, so that we can assess the assumption of normality of the residuals. The VAR statement specifies the numeric variable to analyze—in this case, **Residuals**.

The HISTOGRAM statement creates a histogram of the residuals. This statement also specifies two options. The NORMAL option requests an overlaid distribution plot as well as a series of goodness-of-fit tests based on the empirical distribution function. The KERNEL option requests an overlaid kernel distribution plot.

The QQPLOT statement creates a normal quantile plot for residuals. In this statement, the NORMAL option followed by the MU=EST and SIGMA=EST suboptions in parentheses requests a reference line on the QQ-plot. EST means "estimate." The estimates of the mean and standard deviation are the sample mean and sample standard deviation.

We submit the code.

```
proc univariate data=out;
   var residuals;
   histogram residuals / normal kernel;
   qqplot residuals / normal(mu=est sigma=est);
run;
title;
```

Now let's see if the results validate the normality assumptions.

First, we examine the graphs at the bottom of the results. The histogram of the residuals is shown with a normal density curve and a kernel density curve overlaid. The normal density, represented by the line with the lower peak, is constructed assuming that the data are from a normal distribution that has the same mean and variance as the sample data. The kernel density is represented by the line with the higher peak. It makes minimal assumptions about the functional form of the data and enables the data to describe the shape of the curve. The histogram of the residuals shows a fairly normal distribution.

Similarly the QQ plot shows that the residuals closely follow the reference line. However, note that there is a slight indication of an S-shaped curve. There also appears to be one large error.

At the top of the PROC UNIVARIATE results, the first two tables display descriptive statistics like the mean, median, and skewness. The skewness statistic is 0.51 and the mean is larger than the median, both of which indicate a possible skewed-to-the-right distribution of the residuals.

In the Goodness-of-Fit Tests for Normal Distribution table, all of the normality tests reject the null hypothesis that the residuals are normally distributed. Keep in mind that these results should always be evaluated in conjunction with the histogram and QQ plot (normal probability plot) of the residuals. All of the normality tests depend on the sample size. As the sample size becomes larger, increasingly smaller departures from normality can be detected. A small sample size likely yields a less powerful test and you might want to use a higher alpha value. For additional details about the Tests for Normal Distribution table, click the Information button.

Close