Linear Regression Analysis using PROC GLM

Regression analysis is a statistical method of obtaining an equation that represents a linear relationship between two variables (simple linear regression), or between a single dependent and several independent variables (multiple linear regression). Either the GLM procedure or the REG procedure can be used to perform simple and multiple linear regression. PROC GLM can also handle a wide variety of other linear models. The statements required to perform a regression analyses in either procedure are similar.

The general form of the PROC GLM statement is

PROC GLM options ;

Options which you may need to run a regression analysis include the DATA= option. Other options apply to other types of linear models.

The simplest form of the PROC GLM statement is

PROC GLM ;

A MODEL statement is also required in order to perform a regression analysis. The form of the MODEL statement is

MODEL dependent = independents / options;

The dependent variable is specified on the left of the equal sign, and the independent variable(s) on the right of the equal sign, separated by blanks. Allowable options include INTERCEPT (or just INT), which tells PROC GLM to print hypothesis tests associated with the intercept as an effect in the model. By default, PROC GLM includes the intercept in the model, but does not print associated tests of hypotheses. The option NOINT tells PROC GLM to omit the intercept term from the model. The CLI option tells PROC GLM to print confidence limits for individual predicted values for each observation. The ALPHA= option specifies the alpha level for confidence intervals. By default the alpha level is 0.05. The option P will cause the observed, predicted, and residual values to be printed for each observation that does not contain missing values. An example of the MODEL statement is

MODEL weight = height age / CLI P ;

The OUTPUT statement can be used to create a SAS data set that contains all the input data, as well as predicted values, confidence limits, residuals, and regression diagnostics. The form of the OUTPUT statement is

OUTPUT OUT=datasetname keyword=name ;

The keywords are used to specify which values to store in the output data set. Useful keyword options include PREDICTED (or P) for predicted values, L95 and U95 for lower and upper confidence limits for an individual prediction, and RESIDUAL (or R) for residual values. An example of an OUTPUT statement is

OUTPUT OUT=stats P=pred R=res L95=lower U95=upper;

The OUTPUT statement is useful when creating a data set that will be used later by another SAS procedure (such as PROC PLOT).

A BY statement can be used with PROC GLM to obtain separate plots on observations in groups defined by the BY variables. When a BY statement appears, PROC GLM expects the data to be sorted in the order of the BY variables.

Example: Simple Linear Regression
Develop a simple linear regression equation to predict the age (days) of a rat from its body weight (grams). The data is given below in the following SAS program:

```
DATA rats;
INPUT weight age ;
DATALINES;
76 28
89 33
154 35
189 37
180 38
180 47
```

```
241 66
320 67
271 70
370 82
;
PROC GLM  DATA=rats;
        MODEL   age = weight  / p;
        OUTPUT  OUT=stats  P=pred  R=res  L95=lower  U95=upper;
RUN;

PROC PLOT DATA=stats;
        PLOT   age *weight  pred*weight='*'  / OVERLAY  ;
RUN;
```

At the top of the OUTPUT window we see information on the number of observations in the data set. Next we see an overall analysis of variance table for the specified model. This table shows the breakdown of the total sum of squares for the dependent variable (age) into orthogonal components: the portion of variation accounted for by the model (the model sum of squares), and the portion that the model does not account (the error sum of squares). These are found under the column "Sum of Squares"

In this section, we also see a column labeled "F Value" and another labeled "Pr > F". The F value for the model is calculated by dividing the mean square for the model by the error mean square. For the example above, the model F value is 61.56, and the probability of getting a greater F value is 0.0001. This indicates that the model explains a significant portion of the variation in the ages.

Beneath the analysis of variance table are some computed statistics. The R-Square value is computed as the ratio of the sum of squares for the model divided by the sum of squares for the corrected total. R^2 measures how much variation in the dependent variable (age) can be explained by the model. The value of R^2 can range from 0 to 1. In general the larger the value of R^2 the better the fit of the model. In our printout we see a value of 0.884995, indicating that the model can account for over 88% of the value of age, just by knowing the weight.

Also printed are the coefficient of variation (C.V.), the root mean square error (Root MSE), which is the square root of the error mean square

and is also known as the standard deviation of the dependent variable (age).The mean of the dependent variable is also printed.

In the next section of output we see the sum of squares attributable to each variable in the model. The TYPE I SS measures the increment in the sum of squares for the model as each variable is added to the model. The TYPE III SS measures the sum of squares due to adding that variable last in the model. In this example, since only one independent variable was included in the model, these sums of squares are equal. The "F Value" and "Pr > F" for the type III sum of squares are equivalent to the results of a t-test for testing that the regression coefficient equals zero. In our example, we see that this probability is 0.0001, indicating that the coefficient of the variable weight is significantly different from zero.

The next section is a report on the parameter estimates. The intercept is estimated to be 10.866, and the results of the t-test (testing the null hypothesis that the parameter equals zero) yield a p-value of 0.0819, indicating that the intercept is "somewhat" useful in the model. The coefficient of the weight term is 0.190404, with a p-value of 0.001, indicating that the coefficient of weight does not equal zero. Hence the independent variable weight does contribute significantly to the model. Standard errors of each parameter estimate are also printed. The next part of the printed output shows the observed and predicted values of age, along with the residual vales. The First Order Autocorrelation and Durbin-Watson D statistics measure the presence of a first -order atucorrelation are also computed.

A plot of predicted ages and actual ages vs. weight appear on the next section of output. The plot of residuals that appears next can be used to detect patterns in the residuals, and may be used to determine if a nonlinear model may be more appropriate. In our plot, the residuals appear to be randomly scattered about the line r=0. There appears to be no pattern in the residuals plot. A linear model seems appropriate.

Example: Multiple Linear Regression

In a study of grade school children, ages, heights, weights and scores on a physical fitness exam were obtained from a random sample of 20 children. The data is given below. Find a multiple linear regression equation relating the scores to the ages, heights, and weights of the children.

Our SAS program might look like:

```
DATA  phys;
INPUT  score age height weight;
DATALINES;
58  7  47.5  53
54  7   45  50
55  9  52.5  85
74  7  48  52
86  9  55  76
98  8  51  64
96  9  53  75
70  7  46  75
40  7  48  68
67  9  50.5  74
41  6  45  40
41  7  48.5  66
47  8  50.5  65
45  8  49.0  70
92  9 51.5 70
50  7  46.5  60
98  9  53.5  77
42  8  45  65
64  8  52.5  65
70  8  51.5  67
;
PROC GLM DATA = phys;
        MODEL   score = age   height   weight    / p;
RUN;
```

When we look at the analysis of variance table in the output, we see that the model does seem to explain a significant portion of the variation in the physical fitness scores, as indicated by a p-value of 0.0143 under "Pr > F" . Although the model is significant in explaining scores, the model is not a very good fit to the data. The r-square value of 0.4738 indicates that less than half of the variation in scores is accounted for by the model.

In the next section of output we see the sum of squares attributable to each variable in the model. The TYPE I SS measures the increment in the

sum of squares for the model as each variable is added to the model. The TYPE III SS measures the sum of squares due to adding that variable last in the model. In this example, we have three independent variables included in the model. Hence these sums of squares are different. The "F Value" and "Pr > F" for the type III sum of squares are equivalent to the results of a t-test for testing that the regression coefficient equals zero. In our example, it appears that none of the variables is significant in explaining the test scores. However, looking at the TYPE I SS we see that age is significant when it is the first variable added to the model. These results may seem contradictory. But you have to remember how TYPE I and TYPE III SS are computed. A reasonable explanation for this seeming discrepancy is that age is likely to be significantly correlated with both height and weight. Thus when age is the last variable entered into the model, most of the information contained in the variable age has already been "used" by the variables height and weight in explaining test scores. When this type of situation arises you may want to repeat the analysis using some type of selection procedure.

The section for parameter estimates also indicates that none of the variables are significant in explaining test scores.

Example: Stepwise Selection Multiple Regression
Run the same model using the stepwise selection method.

Note that PROC GLM will not perform model selection methods. We must use SAS's regression procedure (PROC REG) to do this. The SAS program is

```
DATA  phys;
INPUT  score age height weight;
DATALINES;
58  7  47.5  53
54  7   45  50
55  9  52.5  85
74  7  48  52
86  9  55  76
98  8  51  64
96  9  53  75
70  7  46  75
40  7  48  68
```

```
67  9  50.5  74
41  6  45  40
41  7  48.5  66
47  8  50.5  65
45  8  49.0  70
92  9 51.5 70
50  7  46.5  60
98  9  53.5  77
42  8  45  65
64  8  52.5  65
70  8  51.5  67
;
PROC REG DATA = phys;
        MODEL   score = age   height   weight    / p
                SELECTION = STEPWISE
                SLENTRY =0.3
                SLSTAY = 0.3;
RUN;
```

The STEPWISE selection method begins by fitting an intercept term to the data. If you do not want an intercept term in your model, use the option NOINT in the MODEL statement. After fitting the intercept, SAS performs an "Analysis of Variables Not in the Model". For STEPWISE, the first variable selected for entry into the model will be the variable with the lowest p-value (assuming that this p-value is less than the value of SLENTRY). The option SLENTRY=.3 specifies the significance level for entry into the model. The default value is 0.50 for FORWARD selection and 0.15 for STEPWISE selection. The option SLSTAY=.3 specifies the significance level for remaining in the model. The default value is 0.10 for BACKWARD selection and 0.15 for STEPWISE selection.

On the SAS output (Step 1), we see that the variable height has entered the model with a p-value of 0.0019.

After entering a new variable into the model, SAS performs a BACKWARD elimination step to see if any of the variables in the current model can be removed. SAS will next determine which (if any) of the remaining explanatory variables should be entered into the model. None of the remaining variables meets the criterion for entry into the model. The

STEPWISE selection procedure terminates. A summary of the variables selected for entry and the variables removed from the model at each step of the process is given in the SAS output.

We have mentioned three selection methods available in PROC REG. These three methods are FORWARD for forward selection, BACKWARD for backward selection, and STEPWISE for stepwise selection. These methods are specified using the SELECTION= option in the MODEL statement. Intercept parameters are always forced to stay in the model unless the NOINT option is specified in the MODEL statement.

When SELECTION=FORWARD, PROC REG first estimates parameters for variables forced into the model, i.e. the intercept. (There is a way to force some of the explanatory variables into the model. Use the INCLUDE= variable(s) option in the MODEL statement) Next SAS computes the adjusted Chi-square statistics for all variables not in the model and examines the largest of these statistics. If it meets the criterion for entry (SLENTRY) into the model, the variable with this largest adjusted Chi-square is entered into the model. Once a variable is entered into the model it is never removed from the model. The process is repeated until none of the remaining variables meets the specified entry level.

When SELECTION=BACKWARD, parameters for the complete model as specified in the MODEL statement are estimated. The univariate tests based on the Maximum Likelihood Estimates are examined. The least significant variable that does not meet the criterion for staying (SLSTAY) in the model is removed. Once a variable is removed from the model it remains excluded. The process is repeated until no other variable meets the specified level for removal.

SELECTION=STEPWISE is similar to SELECTION=FORWARD except that variables already in the model do not necessarily remain in the model. Variables are entered into the model and removed from the model in such a way that each forward selection step is followed by one or more backward elimination steps. The stepwise selection process terminates if no further variable can be added to the model, or if the variable just entered into the model is the only variable removed in the subsequent backward elimination.

Other selection methods are available, for example RSQUARE. The RSQUARE selection procedure performs an "all subsets regression" and prints the models in decreasing order of R^2 magnitude within each subset size. The subset models selected by RSQUARE are optimal in terms of R^2 for the given sample, but are not necessarily optimal for the population from which the sample was obtained, or any other sample for which you may wish to make predictions.

While model selection techniques are useful for exploratory model building, no statistical method can be relied upon to identify the "true" model. Effective model building requires substantive theory to suggest relevant predictors and plausible functional forms for the model.