Regression Analysis

Required Statements:

MODEL: Like Analysis of Variance, the MODEL statement for regression in GLM has dependent and independent variables and would have the form:
MODEL *Dependent var. = Independent var.* In ordinary linear regression however, all the variables are quantitative. The asterisk notation can also be used here which represents the interaction or multiplicative effects of more than one variable. An example might be **MODEL y = X_1 X_1*X_2.** The vertical bar notation is also applicable, but is not normally used. Nested terms are not used in regression models.

Additional Statements and Options:

MODEL Statement Options: In regression it is usually of interest to examine predicted values and residuals. These can be printed using the options P and either R or STUDENT. The last two refer to regular residuals and studentized residuals, respectively. Also available are confidence intervals on the predicted curve (CLM) and individual points (CLI) both of which are calculated as 95% confidence limits. Some control over the model to be fitted is given by options such as INT and NOINT which tell SAS to fit a model with and without an intercept. SAS will fit an intercept by default.

OUTPUT Statement: The statistics that can be specified here are similar to those of the MODEL statement - predicted values and diagnostics. With the OUTPUT statement however, the requested items are put into a new SAS data set and are available to other SAS procedures such as UNIVARIATE and PLOT. This is very important in terms of regression diagnostics.

Both the OUTPUT and MODEL statement options provide access to useful diagnostics, but the user should be aware that PROC REG is much more adept at doing regression analysis and has many more options available. It should also be noted that many of the features listed are only available in **versions 6** or higher.

Example 5 - Simple Linear Regression SLR

```
PROC GLM;
     MODEL YIELD = pH / P;
     OUTPUT OUT=FIT P=PRED STUDENT=RESID;
```

NOTES: Notice no CLASS statement was used. The model implied here is of the form : YIELD = $\$_0 + \$_1$* pH. The MODEL statement also requests that predicted values and studentized residuals be printed. The OUTPUT statement creates the data set FIT which contains the variables PRED and RESID. This is equivalent to the MODEL statement options. It is important to note that data set FIT also contains <u>all</u> of the original variables.

Example 6 - Multiple Linear Regression (Polynomial Regression)

```
PROC GLM;
     MODEL YIELD = pH pH*pH / SS1;
     OUTPUT OUT=QUAD P=PRED STUDENT=RESID;
```

NOTES: This example is a special case of multiple regression where the independent terms are polynomial. The model is YIELD = $\$_0 + \$_1$* pH + $\$_2$* pH^2. It is important to notice how easily these terms were specified using the asterisk notation. No DATA step was required to create these new variables. The options work the same as earlier examples. A new option of SS1 has been added in this case though, which requests that sequential sums of squares be computed. This is useful in polynomial regression analysis.

Although PROC GLM will handle regression models, PROC REG should normally be used since it provides a complete REGRESSION analysis including measures of collinearity and influence which are unavailable in PROC GLM.