SAS Work Shop
PROC REG
Handout #3

Statistical Programs
College of Agriculture
HTTP://WWW.UIDAHO.EDU/AG/STATPROG

Regression Diagnostics

Residual Analysis: One of the most important aspects of the regression technique is the residual analysis. This involves numeric and graphical inspection of the model residuals defined as the observed values minus the predicted values. The ability of PROC REG to do such analyses is unequalled in other SAS procedures and is the main reason for developing regression models using PROC REG rather than PROC GLM. Residual analysis in PROC REG can be approached in three basic ways outlined below.

MODEL Statement Options: As mentioned earlier, some MODEL statement options are related to diagnostics, and in particular, residual analysis. The most obvious of these is the R option. This requests the following: observed values, predicted values, residuals, standard errors, studentized residuals, and Cook's D statistic. The P option will print only the observed value, predicted value and the residual. Standardized estimates can be obtained with the STB option. The first order correlation of the residuals can also be examined with the DW (Durban-Watson) option.

PLOT and PAINT Statements: One of the most effective and simple methods of residual analysis is plotting of residuals vs predicted values and regressors. Problems such as heteroskedasticity, and systematic lack of fit tend to stand out in such plots. Recent versions of PROC REG (versions 6 and higher) have made producing these plots simple using a PLOT statement. The syntax for PLOT is similar to that of the PROC PLOT procedure: `PLOT <y var>* <x var>=<symbol#>.` The differences in the PROC REG PLOT statement are the variable names and the options available. The variables used can come from the data set being analyzed, or from the results of the current regression analysis. If the later is to be plotted, say the residuals, the

specific variable name to use is: residual.
(Note this has a period at the end). Other potential variables available are rstudent., and predicted. . In fact, any variables available in the OUTPUT statement can be used.

Example:

```
PROC REG DATA=PHOTO;
     MODEL PHOTO=IRRAD;
     PLOT PHOTO*IRRAD='O' PREDICTED.*IRRAD='P'/OVERLAY;
     PLOT RSTUDENT.*PREDICTED.='+' RSTUDENT.*IRRAD='+';
```

In this example the photosynthetic rate is regressed against irradiation. The first PLOT statement requests the observed and predicted values be plotted against irradiation as O's and P's, respectively. The OVERLAY option forces both plots to be on the same graph. The second PLOT statement looks at studentized residuals vs predicted values and irradiation. Both are plotted as a + symbol and are on separate graphs. Other options for PLOT allow for multiple plots per printed page, default symbols and clearing of graphs.

In some cases, it is useful to identify points on a given graph according to some criteria. For this, PROC REG uses the PAINT statement. The PAINT statement is issued

preceding a PLOT statement and defines the criteria and plotting symbol to be used. The syntax is:
PAINT var name <condition> / options . This can be very useful when trying to understand the structure of the data or locating troublesome data points.

Example:

```
     PROC REG DATA=PHOTO;
          MODEL PHOTO=IRRAD;
          PAINT CO2 > 600/SYMBOL='#';
          PLOT PHOTO*IRRAD='O' PREDICTED.*IRRAD='P'/OVERLAY;
          PLOT RSTUDENT.*PREDICTED.='+' RSTUDENT.*IRRAD='+';
```

This example produces the same plots as before, but now changes the plotting symbol to # for those observations with CO_2 levels greater than 600. Any variable in the data set or from the analysis can be used for the criteria in the PAINT statement. Multiple PAINT statements are allowed and are cumulative.

OUTPUT Statement: The third method for addressing residual analysis is the OUTPUT statement. This allows the results obtained from the MODEL statement options to be put into a data set for further analysis -- it may be used to test the distributional assumption of the residuals, for example. It also permits the values to be exported to software other than SAS for plotting or analysis. The syntax for OUTPUT is:
***OUTPUT OUT=(data set name) option=var$_1$ option=var$_2$ ... option=var$_n$*.**
The data set name specifies where the output will go and the options are what statistics are requested.
Example:

```
PROC REG DATA=PHOTO;
     MODEL PHOTO = CO2;
     OUTPUT OUT=PRED P=YHAT RSTUDENT=RESID L95M=LOW U95M=HIGH;

PROC UNIVARIATE PLOT NORMAL;
```

```
    VAR RESID;
```

The OUTPUT statement used in this example creates a data set named PRED which contains several new variables. These are (in order requested): YHAT=predicted values, RESID=studentized residuals, LOW=lower 95% CI on mean values, and HIGH=upper 95% CI on mean values. SAS gives the 95% confidence levels for the last two by default.

These are not the only variables in PRED, however. The data set created by OUTPUT will have the new requested variables **and all** the original variables. Thus, there is no need to merge together this new data set and the old one! The second half of the example runs a univariate summary procedure on the residuals of the analysis and specifically calls the PLOT and NORMAL options which produce a stem and leaf diagram and test for normality of the residuals. This step allows the user to examine the variance, skewness, and other summary information on the residuals.

Influence and Collinearity:

Other diagnostic features of PROC REG examine the influence of data points and multicollinearity among regressors. These are invoked as MODEL statement options and produce a variety of printed output.

Influence: The INFLUENCE option of PROC REG produces several measures of influence for each observation. These include residual, studentized residual, h_i (leverage), and the statistics DFFITS and DFBETA.

Collinearity: Collinearity implies a lack of independence between regressors and can lead to biased estimates with inflated errors. The PROC REG options for examining collinearity are COLLIN, VIF and TOL. The main option here is COLLIN which outputs condition numbers and variance proportions associated with regressors. The options

VIF and TOL give the statistics Variance Inflation Factor and Tolerance, respectively, which are the inverse of one another.

Example:

```
          PROC REG DATA=PHOTO;
               MODEL PHOTO = IRRAD CO2 RESIST/INFLUENCE COLLIN;
```