PROC REG Statement

PROC REG *options;*

These options can be specified in the PROC REG statement:

DATA=*SASdataset*
: names the SAS data set to be used by PROC REG. If DATA= is not specified, REG uses the most recently created SAS data set.

OUTEST=*SASdataset*
: requests that parameter estimates be output to this data set. See **Output Data Sets** later in this chapter for details. If you want to create a permanent SAS data set, you must specify a two-level name (see "SAS Files" in *SAS User's Guide: Basics* for more information on permanent SAS data sets).

OUTSSCP=*SASdataset*
: requests that the sums of squares and crossproducts matrix be output to this TYPE=SSCP data set. See **Output Data Sets** for details. If you want to create a permanent SAS data set, you must specify a two-level name (see "SAS Files" in *SAS User's Guide: Basics* for more information on permanent SAS data sets).

NOPRINT
: suppresses the normal printed output. Using this option on the PROC REG statement is equivalent to specifying NOPRINT on each MODEL statement.

SIMPLE
: prints the "simple" descriptive statistics for each variable used in REG.

MODEL Statement

label: MODEL *dependents=regressors / options;*

After the keyword MODEL, the dependent (response) variables are specified, followed by an equal sign and the regressor variables. Variables specified in the MODEL statement must be variables in the data set being analyzed. The label is optional.

General options

NOPRINT
: suppresses the normal printout of regression results.

NOINT
: suppresses the intercept term that is normally included in the model automatically.

ALL
: requests all the features of these options: XPX, SS1, SS2, STB, TOL, COVB, CORRB, SEQB, P, R, CLI, CLM, SPEC, ACOV, PCORR1, PCORR2, SCORR1, SCORR2.

Options to request regression calculations

XPX
: prints the **X'X** crossproducts matrix for the model. The crossproducts matrix is bordered with the **X'Y** and **Y'Y** matrices.

I
: prints the $(\mathbf{X'X})^{-1}$ matrix. The inverse of the crossproducts matrix is bordered with the parameter estimates and SSE matrices.

Options for details on the estimates

SS1
: prints the sequential sums of squares (Type I SS) along with the parameter estimates for each term in the model.

SS2
: prints the partial sums of squares (Type II SS) along with the parameter estimates for each term in the model.

STB
: prints standardized regression coefficients. A standardized regression coefficient is computed by dividing a parameter estimate by the ratio of the sample standard deviation of the dependent variable to the sample standard deviation of the regressor.

TOL
: prints tolerance values for the estimates. Tolerance is defined as $1-R^2$ for a variable with respect to all other regressor variables in the model.

VIF
: prints variance inflation factors with the parameter estimates. Variance inflation is the reciprocal of tolerance.

COVB
: prints the estimated covariance matrix of the estimates. This matrix is $(\mathbf{X'X})^{-1}S^2$ where S^2 is the mean squared error.

CORRB
: prints the correlation matrix of the estimates. This is the $(\mathbf{X'X})^{-1}$ matrix scaled to unit diagonals.

SEQB
: prints a sequence of parameter estimates as each variable is entered into the model. This is printed as a lower triangular matrix where each row is a set of parameter estimates.

COLLIN
: requests a detailed analysis of collinearity among the regressors. This includes eigenvalues, condition indices, and decomposition of the variances of the estimates with respect to each eigenvalue. (See **Collinearity Diagnostics** below.)

COLLINOINT
: requests the same analysis as the COLLIN option with the intercept variable adjusted out rather than included in the diagnostics. (Also see **Parameter Estimates and Associated Statistics** later in this chapter.)

MODEL statement (cont'd)

Options for predicted values and residuals

P calculates predicted values from the input data and the estimated model. The printout includes the observation number, the first ID variable if specified, the actual and predicted values, and the residual.

R requests that the residual be analyzed. The printed output includes everything requested by the P option plus the standard errors of the predicted and residual values, the studentized residual, and Cook's D statistic to measure the influence of each observation on the parameter estimates.

CLM prints the 95% upper- and lower-confidence limits for the expected value of the dependent variable (mean) for each observation. This is not a prediction interval (see the CLI option) because it takes into account only the variation in the parameter estimates, not the variation in the error term.

CLI requests the 95% upper- and lower-confidence limits for an individual predicted value. The confidence limits reflect variation in the error, as well as variation in the parameter estimates.

DW calculates a Durbin-Watson statistic to test whether or not the errors have first-order autocorrelation. (This test is only appropriate for time-series data.) The sample autocorrelation of the residuals is also printed. (See **Autocorrelation in Time-Series Data**.)

INFLUENCE requests a detailed analysis of the influence of each observation on the estimates and the predicted values. (See **Influence Diagnostics**.)

PARTIAL requests partial regression leverage plots for each regressor. (See the sections **Influence Diagnostics** and **Predicted Values and Residuals**.)

TEST Statement

label: TEST *equation1*,
equation2,
.
.
.
equationk;

```
MODEL Y = A1 A2 B1 B2;
APLUS: TEST A1+A2=1;
B1:    TEST B1=0, B2=0;
```

label: TEST *equation1,..., equationk / options;*

OUTPUT Statement

The OUTPUT statement specifies an output data set containing statistics calculated for each observation. For each statistic, specify the keyword, an equal sign, and a variable name for the statistic in the output data set. If the MODEL has several dependent variables, then a list of output variable names can be specified after each keyword to correspond to the list of dependent variables.

OUTPUT OUT=*SASdataset*
PREDICTED=*names* or P=*names*
RESIDUAL=*names* or R=*names*
L95M=*names*
U95M=*names*
L95=*names*
U95=*names*
STDP=*names*
STDR=*names*
STDI=*names*
STUDENT=*names*
COOKD=*names*
H=*names*
PRESS=*names*
RSTUDENT=*names*
DFFITS=*names*
COVRATIO=*names*;

with

PREDICTED= P= predicted values.

RESIDUAL= R= residuals, calculated as ACTUAL minus PREDICTED.

L95M= lower bound of a 95% confidence interval for the expected value (mean) of the dependent variable.

U95M= upper bound of a 95% confidence interval for the expected value (mean) of the dependent variable.

L95= lower bound of a 95% confidence interval for an individual prediction. This includes the variance of the error, as well as the variance of the parameter estimates.

U95= upper bound of a 95% confidence interval for an individual prediction.

STDP= standard error of the mean predicted value.

STDR= standard error of the residual.

STDI= standard error of the individual predicted value.

STUDENT= studentized residuals, the residual divided by its standard error.

COOKD= Cook's D influence statistic.

H= leverage, $x_i(\mathbf{X}'\mathbf{X})^{-1}x_i'$.

PRESS= residual for estimates dropping this observation, which is the residual divided by (1-*h*) where *h* is leverage above.

RSTUDENT= a studentized residual with the current observation deleted.

DFFITS= standard influence of observation on predicted value.

COVRATIO= standard influence of observation on covariance of betas, as discussed with INFLUENCE option.