# Chapter 102
## The REG Procedure

## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overview: REG Procedure</td>
<td>8430</td>
</tr>
<tr>
<td>Getting Started: REG Procedure</td>
<td>8432</td>
</tr>
<tr>
<td>Simple Linear Regression</td>
<td>8432</td>
</tr>
<tr>
<td>Polynomial Regression</td>
<td>8436</td>
</tr>
<tr>
<td>Using PROC REG Interactively</td>
<td>8447</td>
</tr>
<tr>
<td>Syntax: REG Procedure</td>
<td>8448</td>
</tr>
<tr>
<td>PROC REG Statement</td>
<td>8450</td>
</tr>
<tr>
<td>ADD Statement</td>
<td>8464</td>
</tr>
<tr>
<td>BY Statement</td>
<td>8464</td>
</tr>
<tr>
<td>CODE Statement</td>
<td>8464</td>
</tr>
<tr>
<td>DELETE Statement</td>
<td>8465</td>
</tr>
<tr>
<td>FREQ Statement</td>
<td>8465</td>
</tr>
<tr>
<td>ID Statement</td>
<td>8466</td>
</tr>
<tr>
<td>MODEL Statement</td>
<td>8466</td>
</tr>
<tr>
<td>MTEST Statement</td>
<td>8477</td>
</tr>
<tr>
<td>OUTPUT Statement</td>
<td>8479</td>
</tr>
<tr>
<td>PRINT Statement</td>
<td>8480</td>
</tr>
<tr>
<td>REFIT Statement</td>
<td>8481</td>
</tr>
<tr>
<td>RESTRICT Statement</td>
<td>8482</td>
</tr>
<tr>
<td>REWEIGHT Statement</td>
<td>8483</td>
</tr>
<tr>
<td>STORE Statement</td>
<td>8486</td>
</tr>
<tr>
<td>TEST Statement</td>
<td>8487</td>
</tr>
<tr>
<td>VAR Statement</td>
<td>8488</td>
</tr>
<tr>
<td>WEIGHT Statement</td>
<td>8488</td>
</tr>
<tr>
<td>Details: REG Procedure</td>
<td>8488</td>
</tr>
<tr>
<td>Missing Values</td>
<td>8488</td>
</tr>
<tr>
<td>Input Data Sets</td>
<td>8489</td>
</tr>
<tr>
<td>Output Data Sets</td>
<td>8492</td>
</tr>
<tr>
<td>Interactive Analysis</td>
<td>8497</td>
</tr>
<tr>
<td>Model-Selection Methods</td>
<td>8501</td>
</tr>
<tr>
<td>Criteria Used in Model-Selection Methods</td>
<td>8504</td>
</tr>
<tr>
<td>Limitations in Model-Selection Methods</td>
<td>8504</td>
</tr>
<tr>
<td>Parameter Estimates and Associated Statistics</td>
<td>8505</td>
</tr>
<tr>
<td>Predicted and Residual Values</td>
<td>8507</td>
</tr>
<tr>
<td>Models of Less Than Full Rank</td>
<td>8511</td>
</tr>
<tr>
<td>Collinearity Diagnostics</td>
<td>8513</td>
</tr>
</tbody>
</table>
Overview: REG Procedure

The REG procedure is one of many regression procedures in the SAS System. It is a general-purpose procedure for regression, while other SAS regression procedures provide more specialized applications.

Other SAS/STAT procedures that perform at least one type of regression analysis are the CATMOD, GENMOD, GLM, LOGISTIC, MIXED, NLIN, ORTHOREG, PROBIT, RSREG, and TRANSREG procedures. SAS/ETS procedures are specialized for applications in time series or simultaneous systems. These other SAS/STAT regression procedures are summarized in Chapter 4, “Introduction to Regression Procedures,” which also contains an overview of regression techniques and defines many of the statistics computed by PROC REG and other regression procedures.
PROC REG provides the following capabilities:

- multiple MODEL statements
- nine model-selection methods
- interactive changes both in the model and the data used to fit the model
- linear equality restrictions on parameters
- tests of linear hypotheses and multivariate hypotheses
- collinearity diagnostics
- predicted values, residuals, studentized residuals, confidence limits, and influence statistics
- correlation or crossproduct input
- requested statistics available for output through output data sets
- ODS Graphics. For more information, see the section “ODS Graphics” on page 8557.

Nine model-selection methods are available in PROC REG. In the simplest method, PROC REG fits the complete model that you specify. The other eight methods involve various ways of including or excluding variables from the model. You specify these methods with the SELECTION= option in the MODEL statement.

The methods are identified in the following list and are explained in detail in the section “Model-Selection Methods” on page 8501.

- NONE no model selection. This is the default. The complete model specified in the MODEL statement is fit to the data.
- FORWARD forward selection. This method starts with no variables in the model and adds variables.
- BACKWARD backward elimination. This method starts with all variables in the model and deletes variables.
- STEPWISE stepwise regression. This is similar to the FORWARD method except that variables already in the model do not necessarily stay there.
- MAXR forward selection to fit the best one-variable model, the best two-variable model, and so on. Variables are switched so that R square is maximized.
- MINR similar to the MAXR method, except that variables are switched so that the increase in R square from adding a variable to the model is minimized.
- RSQUARE finds a specified number of models with the highest R square in a range of model sizes.
- ADJRSQ finds a specified number of models with the highest adjusted R square in a range of model sizes.
- CP finds a specified number of models with the lowest $C_p$ in a range of model sizes.
Simple Linear Regression

Suppose that a response variable $Y$ can be predicted by a linear function of a regressor variable $X$. You can estimate $\beta_0$, the intercept, and $\beta_1$, the slope, in

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

for the observations $i = 1, 2, \ldots, n$. Fitting this model with the REG procedure requires only the following MODEL statement, where $y$ is the outcome variable and $x$ is the regressor variable.

```sas
proc reg;
   model y=x;
run;
```

For example, you might use regression analysis to find out how well you can predict a child’s weight if you know that child’s height. The Class data set used in this example is available in the Sashelp library.

The equation of interest is

$$\text{Weight} = \beta_0 + \beta_1 \text{Height} + \epsilon$$

The variable Weight is the response or dependent variable in this equation, and $\beta_0$ and $\beta_1$ are the unknown parameters to be estimated. The variable Height is the regressor or independent variable, and $\epsilon$ is the unknown error. The following commands invoke the REG procedure and fit this model to the data.

```sas
ods graphics on;
proc reg data=sashelp.class;
   model Weight = Height;
run;
```

Figure 102.1 includes some information concerning model fit.

The $F$ statistic for the overall model is highly significant ($F = 57.076$, $p < 0.0001$), indicating that the model explains a significant portion of the variation in the data.

The degrees of freedom can be used in checking accuracy of the data and model. The model degrees of freedom are one less than the number of parameters to be estimated. This model estimates two parameters, $\beta_0$ and $\beta_1$; thus, the degrees of freedom should be $2 - 1 = 1$. The corrected total degrees of freedom are always one less than the total number of observations in the data set, in this case $19 - 1 = 18$. 
Several simple statistics follow the ANOVA table. The Root MSE is an estimate of the standard deviation of the error term. The coefficient of variation, or Coeff Var, is a unitless expression of the variation in the data. The R-square and Adj R-square are two statistics used in assessing the fit of the model; values close to 1 indicate a better fit. The R-square of 0.77 indicates that Height accounts for 77% of the variation in Weight.

**Figure 102.1** ANOVA Table

The REG Procedure  
Model: MODEL1  
Dependent Variable: Weight

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>7193.24912</td>
<td>7193.24912</td>
<td>57.08</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>17</td>
<td>2142.48772</td>
<td>126.02869</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>18</td>
<td>9335.73684</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 11.22625  R-Square 0.7705  
Dependent Mean 100.02632  Adj R-Sq 0.7570  
Coeff Var 11.22330

The “Parameter Estimates” table in Figure 102.2 contains the estimates of $\beta_0$ and $\beta_1$. The table also contains the $t$ statistics and the corresponding $p$-values for testing whether each parameter is significantly different from zero. The $p$-values ($t = -4.43$, $p = 0.0004$ and $t = 7.55$, $p < 0.0001$) indicate that the intercept and Height parameter estimates, respectively, are highly significant.

From the parameter estimates, the fitted model is

$$\text{Weight} = -143.0 + 3.9 \times \text{Height}$$
If ODS Graphics is enabled, then PROC REG produces a variety of plots. Figure 102.3 shows a plot of the residuals versus the regressor and Figure 102.4 shows a panel of diagnostic plots.

**Figure 102.3** Residuals vs. Regressor
A trend in the residuals would indicate nonconstant variance in the data. The plot of residuals by predicted values in the upper-left corner of the diagnostics panel in Figure 102.4 might indicate a slight trend in the residuals; they appear to increase slightly as the predicted values increase. A fan-shaped trend might indicate the need for a variance-stabilizing transformation. A curved trend (such as a semicircle) might indicate the need for a quadratic term in the model. Since these residuals have no apparent trend, the analysis is considered to be acceptable.
Polynomial Regression

Consider a response variable $Y$ that can be predicted by a polynomial function of a regressor variable $X$. You can estimate $\beta_0$, the intercept; $\beta_1$, the slope due to $X$; and $\beta_2$, the slope due to $X^2$, in

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \epsilon_i$$

for the observations $i = 1, 2, \ldots, n$.

Consider the following example on population growth trends. The population of the United States from 1790 to 2000 is fit to linear and quadratic functions of time. Note that the quadratic term, $\text{YearSq}$, is created in the DATA step; this is done since polynomial effects such as $\text{Year}\times\text{Year}$ cannot be specified in the MODEL statement in PROC REG. The data are as follows:

```plaintext
data USPopulation;
  input Population @@;
  retain Year 1780;
  Year = Year+10;
  YearSq = Year*Year;
  Population = Population/1000;
datalines;
 3929 5308 7239 9638 12866 17069 23191 31443 39818 50155
 62947 75994 91972 105710 122775 131669 151325 179323 203211
 226542 248710 281422
;

proc reg data=USPopulation plots=ResidualByPredicted;
  var YearSq;
  model Population=Year / r clm cli;
run;
```

The DATA option ensures that the procedure uses the intended data set. Any variable that you might add to the model but that is not included in the first MODEL statement must appear in the VAR statement.

The “Analysis of Variance” and “Parameter Estimates” tables are displayed in Figure 102.5.
Polynomial Regression

**Figure 102.5** ANOVA Table and Parameter Estimates

The REG Procedure
Model: MODEL1
Dependent Variable: Population

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>146869</td>
<td>146869</td>
<td>228.92</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>12832</td>
<td>641.58160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>21</td>
<td>159700</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 25.32946  R-Square 0.9197
Dependent Mean 94.64800  Adj R-Sq 0.9156
Coeff Var 26.76175

| Variable  | DF  | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|-----------|-----|--------------------|----------------|---------|------|------|
| Intercept | 1   | -2345.85498        | 161.39279      | -14.54  | <.0001 |
| Year      | 1   | 1.28786            | 0.08512        | 15.13   | <.0001 |

The Model $F$ statistic is significant ($F = 228.92$, $p < 0.0001$), indicating that the model accounts for a significant portion of variation in the data. The R-square indicates that the model accounts for 92% of the variation in population growth. The fitted equation for this model is

$$\text{Population} = -2345.85 + 1.29 \times \text{Year}$$

In the MODEL statement, three options are specified: R requests a residual analysis to be performed, CLI requests 95% confidence limits for an individual value, and CLM requests these limits for the expected value of the dependent variable. You can request specific 100$(1 - \alpha)$% limits with the ALPHA= option in the PROC REG or MODEL statement.

**Figure 102.6** shows the “Output Statistics” table. The residual, its standard error, the studentized residual, and Cook’s $D$ are displayed for each observation. The studentized residual is the residual divided by its standard error. Cook’s $D$ is a measure of the change in the predicted values upon deletion of that observation from the data set; hence, it measures the influence of the observation on the estimated regression coefficients. The studentized residuals and Cook’s $D$ statistics are displayed in **Figure 102.7**.

Studentized residuals follow a beta distribution and can be used to identify outlying or extreme observations:

- Having a few large absolute studentized residuals is no cause for alarm.
- Many observations having absolute studentized residuals greater than two might indicate an inadequate model.
- Absolute studentized residuals greater than three are rarer; they should be regarded as extreme, and the data should be reviewed carefully.
### The REG Procedure

#### Model: MODEL1

**Dependent Variable:** Population

#### Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Mean Error</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
<th>Residual Std Error</th>
<th>Student Residual</th>
<th>Cook's D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.93</td>
<td>-40.5778</td>
<td>10.4424</td>
<td>-62.3602</td>
<td>-18.7953</td>
<td>10.4424</td>
<td>44.5068</td>
<td>23.077</td>
</tr>
<tr>
<td>2</td>
<td>5.31</td>
<td>-27.6991</td>
<td>9.7238</td>
<td>-47.9826</td>
<td>-7.4156</td>
<td>5.0071</td>
<td>28.8968</td>
<td>23.389</td>
</tr>
<tr>
<td>4</td>
<td>9.64</td>
<td>-1.9418</td>
<td>8.3617</td>
<td>-19.3841</td>
<td>15.5004</td>
<td>5.5071</td>
<td>53.6991</td>
<td>23.909</td>
</tr>
<tr>
<td>5</td>
<td>12.87</td>
<td>10.9368</td>
<td>7.7314</td>
<td>-5.1906</td>
<td>-44.3060</td>
<td>66.1797</td>
<td>7.4156</td>
<td>24.121</td>
</tr>
<tr>
<td>6</td>
<td>17.07</td>
<td>23.8155</td>
<td>7.1470</td>
<td>-27.6991</td>
<td>8.9070</td>
<td>28.8968</td>
<td>78.7148</td>
<td>23.666</td>
</tr>
<tr>
<td>7</td>
<td>23.19</td>
<td>36.6941</td>
<td>6.6208</td>
<td>-14.8205</td>
<td>38.7239</td>
<td>41.2719</td>
<td>23.077</td>
<td>23.077</td>
</tr>
<tr>
<td>8</td>
<td>31.44</td>
<td>49.5727</td>
<td>6.1675</td>
<td>-1.9418</td>
<td>50.5048</td>
<td>53.6991</td>
<td>23.909</td>
<td>24.121</td>
</tr>
<tr>
<td>9</td>
<td>39.82</td>
<td>62.4514</td>
<td>5.8044</td>
<td>10.9368</td>
<td>63.7547</td>
<td>7.4156</td>
<td>53.6991</td>
<td>23.909</td>
</tr>
<tr>
<td>10</td>
<td>50.16</td>
<td>75.3300</td>
<td>5.5491</td>
<td>-5.1906</td>
<td>86.9053</td>
<td>8.2455</td>
<td>78.7148</td>
<td>23.909</td>
</tr>
<tr>
<td>11</td>
<td>62.95</td>
<td>88.2087</td>
<td>5.4170</td>
<td>-5.1906</td>
<td>96.9090</td>
<td>9.7238</td>
<td>24.121</td>
<td>24.121</td>
</tr>
<tr>
<td>12</td>
<td>75.99</td>
<td>101.0873</td>
<td>5.4170</td>
<td>-5.1906</td>
<td>112.3870</td>
<td>10.9368</td>
<td>78.7148</td>
<td>23.909</td>
</tr>
<tr>
<td>13</td>
<td>91.97</td>
<td>113.9660</td>
<td>5.5491</td>
<td>-5.1906</td>
<td>123.5413</td>
<td>12.6858</td>
<td>24.121</td>
<td>24.121</td>
</tr>
<tr>
<td>15</td>
<td>122.78</td>
<td>139.7233</td>
<td>6.1675</td>
<td>-5.1906</td>
<td>151.5885</td>
<td>15.9368</td>
<td>25.1750</td>
<td>24.714</td>
</tr>
<tr>
<td>16</td>
<td>131.67</td>
<td>152.6019</td>
<td>6.6208</td>
<td>-5.1906</td>
<td>166.1426</td>
<td>17.6858</td>
<td>25.1750</td>
<td>24.714</td>
</tr>
<tr>
<td>17</td>
<td>151.33</td>
<td>165.4805</td>
<td>7.1470</td>
<td>-5.1906</td>
<td>180.3890</td>
<td>19.3523</td>
<td>25.1750</td>
<td>24.714</td>
</tr>
<tr>
<td>19</td>
<td>203.21</td>
<td>191.2378</td>
<td>8.3617</td>
<td>-5.1906</td>
<td>208.6801</td>
<td>23.1163</td>
<td>25.1750</td>
<td>24.714</td>
</tr>
<tr>
<td>20</td>
<td>226.54</td>
<td>204.1165</td>
<td>9.0283</td>
<td>-5.1906</td>
<td>222.9439</td>
<td>24.8237</td>
<td>25.1750</td>
<td>24.714</td>
</tr>
</tbody>
</table>
Figure 102.7 Studentized Residuals

Figure 102.8 shows the residual statistics table. A fairly close agreement between the PRESS statistic (see Table 102.7) and the sum of squared residuals indicates that the MSE is a reasonable measure of the predictive accuracy of the fitted model (Neter, Wasserman, and Kutner 1990).

**Figure 102.8 Residual Statistics**

<table>
<thead>
<tr>
<th>Sum of Residuals</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Squared Residuals</td>
<td>12832</td>
</tr>
<tr>
<td>Predicted Residual SS (PRESS)</td>
<td>16662</td>
</tr>
</tbody>
</table>

Graphical representations are very helpful in interpreting the information in the “Output Statistics” table. When ODS Graphics is enabled, the REG procedure produces a default set of diagnostic plots that are appropriate for the requested analysis.
Figure 102.9 displays a panel of diagnostics plots. These diagnostics indicate an inadequate model:

- The plots of residual and studentized residual versus predicted value show a clear quadratic pattern.

- The plot of studentized residual versus leverage seems to indicate that there are two outlying data points. However, the plot of Cook’s $D$ distance versus observation number reveals that these two points are just the data points for the endpoint years 1790 and 2000. These points show up as apparent outliers because the departure of the linear model from the underlying quadratic behavior in the data shows up most strongly at these endpoints.

- The normal quantile plot of the residuals and the residual histogram are not consistent with the assumption of Gaussian errors. This occurs as the residuals themselves still contain the quadratic behavior that is not captured by the linear model.

- The plot of the dependent variable versus the predicted value exhibits a quadratic form around the 45-degree line that represents a perfect fit.

- The “Residual-Fit” (or RF) plot consisting of side-by-side quantile plots of the centered fit and the residuals shows that the spread in the residuals is no greater than the spread in the centered fit. For inappropriate models, the spread of the residuals in such a plot is often greater than the spread of the centered fit. In this case, the RF plot shows that the linear model does indeed capture the increasing trend in the data, and hence accounts for much of the variation in the response.
Figure 102.9 Diagnostics Panel
Figure 102.10 shows a plot of residuals versus Year. Again you can see the quadratic pattern that strongly indicates that a quadratic term should be added to the model.

Figure 102.10 Residual Plot

![Residual Plot](image)

Figure 102.11 shows the FitPlot consisting of a scatter plot of the data overlaid with the regression line, and 95% confidence and prediction limits. Note that this plot also indicates that the model fails to capture the quadratic nature of the data. This plot is produced for models containing a single regressor. You can use the ALPHA= option in the model statement to change the significance level of the confidence band and prediction limits.

![FitPlot](image)
These default plots provide strong evidence that the $\text{YearSq}$ needs to be added to the model. You can use the interactive feature of PROC REG to do this by specifying the following statements:

```plaintext
   add YearSq;
   print;
   run;
```

The ADD statement requests that $\text{YearSq}$ be added to the model, and the PRINT command causes the model to be refit and displays the ANOVA and parameter estimates for the new model. The print statement also produces updated ODS graphical displays.

Figure 102.12 displays the ANOVA table and parameter estimates for the new model.
The overall $F$ statistic is still significant ($F = 8864.19, p < 0.0001$). The R-square has increased from 0.9197 to 0.9989, indicating that the model now accounts for 99.9% of the variation in Population. All effects are significant with $p < 0.0001$ for each effect in the model.

The fitted equation is now

$$\text{Population} = 21631 - 24.046 \times \text{Year} + 0.0067 \times \text{YearSq}$$

Figure 102.13 show the panel of diagnostics for this quadratic polynomial model. These diagnostics indicate that this model is considerably more successful than the corresponding linear model:

- The plots of residuals and studentized residuals versus predicted values exhibit no obvious patterns.
- The points on the plot of the dependent variable versus the predicted values lie along a 45-degree line, indicating that the model successfully predicts the behavior of the dependent variable.
- The plot of studentized residual versus leverage shows that the years 1790 and 2000 are leverage points with 2000 showing up as an outlier. This is confirmed by the plot of Cook’s $D$ distance versus observation number. This suggests that while the quadratic model fits the current data well, the model might not be quite so successful over a wider range of data. You might want to investigate whether the population trend over the last couple of decades is growing slightly faster than quadratically.
When a model contains more than one regressor, PROC REG does not produce a fit plot. However, when all the regressors in the model are functions of a single variable, it is appropriate to plot predictions and residuals as a function of that variable. You request such plots by using the PLOTS=PREDICTIONS option in the PROC REG statement, as the following code illustrates:

```plaintext
proc reg data=USPopulation plots=predictions(X=Year);
   model Population=Year Yearsq;
quit;
ods graphics off;
```
Figure 102.14 shows the data, predictions, and residuals by Year. These plots confirm that the quadratic polynomial model successfully model the growth in U.S. population between the years 1780 and 2000.

To complete an analysis of these data, you might want to examine influence statistics and, since the data are essentially time series data, examine the Durbin-Watson statistic.
Using PROC REG Interactively

The REG procedure can be used interactively. After you specify a model with a MODEL statement and run PROC REG with a RUN statement, a variety of statements can be executed without reinvoking PROC REG.

The section “Interactive Analysis” on page 8497 describes which statements can be used interactively. These interactive statements can be executed singly or in groups by following the single statement or group of statements with a RUN statement. Note that the MODEL statement can be repeated. This is an important difference from the GLM procedure, which supports only one MODEL statement.

If you use PROC REG interactively, you can end the REG procedure with a DATA step, another PROC step, an ENDSAS statement, or a QUIT statement. The syntax of the QUIT statement is

```plaintext
quit;
```

When you are using PROC REG interactively, additional RUN statements do not end PROC REG but tell the procedure to execute additional statements.

When a BY statement is used with PROC REG, interactive processing is not possible; that is, once the first RUN statement is encountered, processing proceeds for each BY group in the data set, and no further statements are accepted by the procedure.

When you use PROC REG interactively, you can fit a model, perform diagnostics, and then refit the model and perform diagnostics on the refitted model. Most of the interactive statements implicitly refit the model; for example, if you use the ADD statement to add a variable to the model, the regression equation is automatically recomputed. The two exceptions to this automatic recomputing are the PAINT and REWEIGHT statements. These two statements do not cause the model to be refitted. To refit the model, you can follow these statements either with a REFIT statement, which causes the model to be explicitly recomputed, or with another interactive statement that causes the model to be implicitly recomputed.
Syntax: REG Procedure

The following statements are available in the REG procedure:

```
PROC REG <options>;<label:>
   MODEL dependents = <regressors> </options>;
   BY variables;
   FREQ variable;
   ID variables;
   VAR variables;
   WEIGHT variable;
   ADD variables;
   CODE <options>;
   DELETE variables;<label:>
   MTEST <equation, . . . , equation> </options>;
   OUTPUT <OUT=SAS-data-set> <keyword=names> <. . . keyword=names>;
   PAINT <condition | ALLOBS > </options> | <STATUS | UNDO>;
   PLOT <yvariable* xvariable> <symbol> <. . . yvariable* xvariable> <symbol> </options>;
   PRINT <options> <ANOVA> <MODELDATA>;
   REFIT ;
   RESTRICT equation, . . . , equation;
   REWEIGHT <condition | ALLOBS > </options> | <STATUS | UNDO>;
   STORE <options>;
   TEST equation, <, . . . , equation> </option>;
```

Although there are numerous statements and options available in PROC REG, many analyses use only a few of them. Often you can find the features you need by looking at an example or by scanning this section.

In the preceding list, brackets denote optional specifications, and vertical bars denote a choice of one of the specifications separated by the vertical bars. In all cases, `label` is optional.

The PROC REG statement is required. To fit a model to the data, you must specify the MODEL statement. If you want to use only the options available in the PROC REG statement, you do not need a MODEL statement, but you must use a VAR statement. (See the example in the section “OUTSSCP= Data Sets” on page 8497.) Several MODEL statements can be used. In addition, several MTEST, OUTPUT, PAINT, PLOT, PRINT, RESTRICT, and TEST statements can follow each MODEL statement.

The ADD, DELETE, and REWEIGHT statements are used interactively to change the regression model and the data used in fitting the model. The ADD, DELETE, MTEST, OUTPUT, PLOT, PRINT, RESTRICT, and TEST statements implicitly refit the model; changes made to the model are reflected in the results from these statements. The REFIT statement is used to refit the model explicitly and is most helpful when it follows PAINT and REWEIGHT statements, which do not refit the model.
The BY, FREQ, ID, VAR, and WEIGHT statements are optionally specified once for the entire PROC step, and they must appear before the first RUN statement.

When a TYPE=CORR, TYPE=COV, or TYPE=SSCP data set is used as an input data set to PROC REG, statements and options that require the original data are not available. Specifically, the OUTPUT, PAINT, PLOT, and REWEIGHT statements and the MODEL and PRINT statement options P, R, CLM, CLI, DW, DWPROB, INFLUENCE, PARTIAL, and PARTIALDATA are disabled.

You can specify the following statements with the REG procedure in addition to the PROC REG statement:

- **ADD** adds independent variables to the regression model.
- **BY** specifies variables to define subgroups for the analysis.
- **CODE** requests that the procedure write SAS DATA step code to a file or catalog entry for computing predicted values according to the fitted model.
- **DELETE** deletes independent variables from the regression model.
- **FREQ** specifies a frequency variable.
- **ID** names a variable to identify observations in the tables.
- **MODEL** specifies the dependent and independent variables in the regression model, requests a model selection method, displays predicted values, and provides details on the estimates (according to which options are selected).
- **MTEST** performs multivariate tests across multiple dependent variables.
- **OUTPUT** creates an output data set and names the variables to contain predicted values, residuals, and other diagnostic statistics.
- **PAINT** paints points in scatter plots.
- **PLOT** generates scatter plots.
- **PRINT** displays information about the model and can reset options.
- **REFIT** refits the model.
- **RESTRICT** places linear equality restrictions on the parameter estimates.
- **REWEIGHT** excludes specific observations from analysis or changes the weights of observations used.
- **STORE** requests that the procedure requests that the procedure save the estimated parameters of the fitted model.
- **TEST** performs an $F$ test on linear functions of the parameters.
- **VAR** lists variables for which crossproducts are to be computed, variables that can be interactively added to the model, or variables to be used in scatter plots.
- **WEIGHT** declares a variable to weight observations.

The CODE and STORE statements are also used by many other procedures. A summary description of functionality and syntax for these statements is also shown after the PROC REG statement in alphabetical order, but you can find full documentation about them in the section “STORE Statement” on page 520 in Chapter 19, “Shared Concepts and Topics.”
PROC REG Statement

PROC REG <options> ;

The PROC REG statement invokes the REG procedure. The PROC REG statement is required. If you want to fit a model to the data, you must also use a MODEL statement. If you want to use only the PROC REG options, you do not need a MODEL statement, but you must use a VAR statement. If you do not use a MODEL statement, then the COVOUT and OUTEST= options are not available.

Table 102.1 summarizes the options available in the PROC REG statement. Note that any option specified in the PROC REG statement applies to all MODEL statements.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>DATA=</td>
<td>Names a data set to use for the regression</td>
</tr>
<tr>
<td>OUTEST=</td>
<td>Outputs a data set that contains parameter estimates and other model fit summary statistics</td>
</tr>
<tr>
<td>OUTSSCP=</td>
<td>Outputs a data set that contains sums of squares and crossproducts</td>
</tr>
<tr>
<td>COVOUT</td>
<td>Outputs the covariance matrix for parameter estimates to the OUTEST= data set</td>
</tr>
<tr>
<td>EDF</td>
<td>Outputs the number of regressors, the error degrees of freedom, and the model R square to the OUTEST= data set</td>
</tr>
<tr>
<td>OUTSEB</td>
<td>Outputs standard errors of the parameter estimates to the OUTEST= data set</td>
</tr>
<tr>
<td>OUTSTB</td>
<td>Outputs standardized parameter estimates to the OUTEST= data set. Use only with the RIDGE= or PCOMIT= option.</td>
</tr>
<tr>
<td>OUTVIF</td>
<td>Outputs the variance inflation factors to the OUTEST= data set. Use only with the RIDGE= or PCOMIT= option.</td>
</tr>
<tr>
<td>PCOMIT=</td>
<td>Performs incomplete principal component analysis and outputs estimates to the OUTEST= data set</td>
</tr>
<tr>
<td>PRESS</td>
<td>Outputs the PRESS statistic to the OUTEST= data set</td>
</tr>
<tr>
<td>RIDGE=</td>
<td>Performs ridge regression analysis and outputs estimates to the OUTEST= data set</td>
</tr>
<tr>
<td>RSQUARE</td>
<td>Same effect as the EDF option</td>
</tr>
<tr>
<td>TABLEOUT</td>
<td>Outputs standard errors, confidence limits, and associated test statistics of the parameter estimates to the OUTEST= data set</td>
</tr>
</tbody>
</table>

ODS Graphics Options

PLOTS= Produces ODS graphical displays

Display Options

CORR Displays correlation matrix for variables listed in MODEL and VAR statements
SIMPLE Displays simple statistics for each variable listed in MODEL and VAR statements
USSCP Displays uncorrected sums of squares and crossproducts matrix
ALL Displays all statistics (CORR, SIMPLE, and USSCP)
Following are explanations of the options that you can specify in the PROC REG statement (in alphabetical order).

Note that any option specified in the PROC REG statement applies to all MODEL statements.

**ALL**
requests the display of many tables. Using the ALL option in the PROC REG statement is equivalent to specifying ALL in every MODEL statement. The ALL option also implies the CORR, SIMPLE, and USSCP options.

**ALPHA=number**
sets the significance level used for the construction of confidence intervals. The value must be between 0 and 1; the default value of 0.05 results in 95% intervals. This option affects the PROC REG option TABLEOUT; the MODEL options CLB, CLI, and CLM; the OUTPUT statement keywords LCL, LCLM, UCL, and UCLM; the PLOT statement keywords LCL., LCLM., UCL., and UCLM.; and the PLOT statement options CONF and PRED.

**CORR**
displays the correlation matrix for all variables listed in the MODEL or VAR statement.

**COVOUT**
outputs the covariance matrices for the parameter estimates to the OUTTEST= data set. This option is valid only if the OUTTEST= option is also specified. See the section “OUTTEST= Data Set” on page 8492.

**DATA=SAS-data-set**
names the SAS data set to be used by PROC REG. The data set can be an ordinary SAS data set or a TYPE=CORR, TYPE=COV, or TYPE=SSCP data set. If one of these special TYPE= data sets is used, the OUTPUT, PAINT, PLOT, and REWEIGHT statements, ODS Graphics, and some options in the MODEL and PRINT statements are not available. See Appendix A, “Special SAS Data Sets,” for more information about TYPE= data sets. If the DATA= option is not specified, PROC REG uses the most recently created SAS data set.

**EDF**
outputs the number of regressors in the model excluding and including the intercept, the error degrees of freedom, and the model R square to the OUTTEST= data set.

**NOPRINT**
suppresses the normal display of results. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 20, “Using the Output Delivery System,” for more information.
OUTEST=SAS-data-set requests that parameter estimates and optional model fit summary statistics be output to this data set. See the section “OUTEST= Data Set” on page 8492 for details. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see SAS Language Reference: Concepts.

OUTSEB outputs the standard errors of the parameter estimates to the OUTEST= data set. The value SEB for the variable _TYPE_ identifies the standard errors. If the RIDGE= or PCOMIT= option is specified, additional observations are included and identified by the values RIDGESEB and IPCSEB, respectively, for the variable _TYPE_. The standard errors for ridge regression estimates and IPC estimates are limited in their usefulness because these estimates are biased. This option is available for all model selection methods except RSQUARE, ADJRSQ, and CP.

OUTSSCP=SAS-data-set requests that the sums of squares and crossproducts matrix be output to this TYPE=SSCP data set. See the section “OUTSSCP= Data Sets” on page 8497 for details. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see SAS Language Reference: Concepts.

OUTSTB outputs the standardized parameter estimates as well as the usual estimates to the OUTEST= data set when the RIDGE= or PCOMIT= option is specified. The values RIDGESTB and ICPSTB for the variable _TYPE_ identify ridge regression estimates and IPC estimates, respectively.

OUTVIF outputs the variance inflation factors (VIF) to the OUTEST= data set when the RIDGE= or PCOMIT= option is specified. The factors are the diagonal elements of the inverse of the correlation matrix of regressors as adjusted by ridge regression or IPC analysis. These observations are identified in the output data set by the values RIDGEVIF and IPCVIF for the variable _TYPE_.

PCOMIT=list requests an incomplete principal component (IPC) analysis for each value m in the list. The procedure computes parameter estimates by using all but the last m principal components. Each value of m produces a set of IPC estimates, which are output to the OUTEST= data set. The values of m are saved by the variable _PCOMIT_, and the value of the variable _TYPE_ is set to IPC to identify the estimates. Only nonnegative integers can be specified with the PCOMIT= option.

If you specify the PCOMIT= option, RESTRICT statements are ignored.

PLOTS < (global-plot-options) > <= plot-request<(options)> >=

controls the plots produced through ODS Graphics. When you specify only one plot request, you can omit the parentheses around the plot request. Here are some examples:

```
plots = none
plots = diagnostics(unpack)
plots = (all fit(stats=none))
plots(label) = (rstudentbyleverage cooks)
plots(only) = (diagnostics(stats=all) fit(nocli stats=(aic sbc)))
```
ODS Graphics must be enabled before plots can be requested. For example:

```plaintext
ods graphics on;
proc reg;
    model y = x1-x10;
run;
proc reg plots=diagnostics(stats=(default aic sbc));
    model y = x1-x10;
run;
ods graphics off;
```

For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 623 in Chapter 21, “Statistical Graphics Using ODS.”

If ODS Graphics is enabled but you do not specify the PLOTS= option, then PROC REG produces a default set of plots. Table 102.2 lists the default set of plots produced.

<table>
<thead>
<tr>
<th>Plot</th>
<th>Conditional On</th>
</tr>
</thead>
<tbody>
<tr>
<td>DiagnosticsPanel</td>
<td>Unconditional</td>
</tr>
<tr>
<td>ResidualPlot</td>
<td>Unconditional</td>
</tr>
<tr>
<td>FitPlot</td>
<td>Model with one regressor (excluding intercept)</td>
</tr>
<tr>
<td>PartialPlot</td>
<td>PARTIAL option specified in MODEL statement</td>
</tr>
<tr>
<td>RidgePanel</td>
<td>RIDGE= option specified in PROC REG or MODEL statement</td>
</tr>
</tbody>
</table>

For models with multiple dependent variables, separate plots are produced for each dependent variable. For jobs with more than one MODEL statement, plots are produced for each model statement.

The global-options apply to all plots generated by the REG procedure, unless it is altered by a specific-plot-option. The following global-plot-options are available:

- **LABEL**
  specifies that the LABEL option be applied to each plot that supports a LABEL option. See the descriptions of the specific plots for details.

- **MAXPOINTS=NONE | max < heat-max>**
  suppresses most plots that require processing more than max points. When the number of points exceeds max but does not exceed heat-max divided by the number of independent variables, heat maps are displayed instead of scatter plots for the fit and residual plots. All other plots are suppressed when the number of points exceeds max. The default is MAXPOINTS=5000 150000. These cutoffs are ignored if you specify MAXPOINTS=NONE.

- **MODELLABEL**
  requests that the model label be displayed in the upper-left corner of all plots. This option is useful when you use more than one MODEL statement.
ONLY
suppress the default plots. Only plots specifically requested are displayed.

STATS=ALL | DEFAULT | NONE | (plot-statistics)
requests statistics that are included on the fit plot and diagnostics panel. Table 102.3 lists the statistics that you can request. STATS=ALL requests all these statistics; STATS=NONE suppresses them.

Table 102.3 Statistics Available on Plots

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJRSQ</td>
<td>x</td>
<td>adjusted R-square</td>
</tr>
<tr>
<td>AIC</td>
<td></td>
<td>Akaike’s information criterion</td>
</tr>
<tr>
<td>BIC</td>
<td></td>
<td>Sawa’s Bayesian information criterion</td>
</tr>
<tr>
<td>CP</td>
<td></td>
<td>Mallows’ $C_p$ statistic</td>
</tr>
<tr>
<td>COEFFVAR</td>
<td></td>
<td>coefficient of variation</td>
</tr>
<tr>
<td>DEPMEAN</td>
<td></td>
<td>mean of dependent</td>
</tr>
<tr>
<td>DEFAULT</td>
<td></td>
<td>all default statistics</td>
</tr>
<tr>
<td>EDF</td>
<td>x</td>
<td>error degrees of freedom</td>
</tr>
<tr>
<td>GMSEP</td>
<td></td>
<td>estimated MSE of prediction, assuming multivariate normality</td>
</tr>
<tr>
<td>JP</td>
<td></td>
<td>final prediction error</td>
</tr>
<tr>
<td>MSE</td>
<td>x</td>
<td>mean squared error</td>
</tr>
<tr>
<td>NOBS</td>
<td>x</td>
<td>number of observations used</td>
</tr>
<tr>
<td>NPARM</td>
<td>x</td>
<td>number of parameters in the model (including the intercept)</td>
</tr>
<tr>
<td>PC</td>
<td></td>
<td>Amemiya’s prediction criterion</td>
</tr>
<tr>
<td>RSQUARE</td>
<td>x</td>
<td>R-square</td>
</tr>
<tr>
<td>SBC</td>
<td></td>
<td>SBC statistic</td>
</tr>
<tr>
<td>SP</td>
<td></td>
<td>SP statistic</td>
</tr>
<tr>
<td>SSE</td>
<td></td>
<td>error sum of squares</td>
</tr>
</tbody>
</table>

You request statistics in addition to the default set by including the keyword DEFAULT in the plot-statistics list.

UNPACK
suppresses paneling.

USEALL
specifies that predicted values at data points with missing dependent variable(s) be included on appropriate plots. By default, only points used in constructing the SSCP matrix appear on plots.
The following specific plots are available:

**ADJRSQ < (adjrsq-options) >**

   displays the adjusted R-square values for the models examined when you request variable selection with the SELECTION= option in the MODEL statement.

   The following *adjrsq-options* are available for models where you request the RSQUARE, ADJRSQ, or CP selection method:

   **LABEL**
   
   requests that the model number corresponding to the one displayed in the “Subset Selection Summary” table be used to label the model with the largest adjusted R-square statistic at each value of the number of parameters.

   **LABELVARS**
   
   requests that the list (excluding the intercept) of the regressors in the relevant model be used to label the model with the largest adjusted R-square statistic at each value of the number of parameters.

**AIC < (aic-options) >**

   displays Akaike’s information criterion (AIC) for the models examined when you request variable selection with the SELECTION= option in the MODEL statement.

   The following *aic-options* are available for models where you request the RSQUARE, ADJRSQ, or CP selection method:

   **LABEL**
   
   requests that the model number corresponding to the one displayed in the “Subset Selection Summary” table be used to label the model with the smallest AIC statistic at each value of the number of parameters.

   **LABELVARS**
   
   requests that the list (excluding the intercept) of the regressors in the relevant model be used to label the model with the smallest AIC statistic at each value of the number of parameters.

**ALL**

   produces all appropriate plots.

**BIC < (bic-options) >**

   displays Sawa’s Bayesian information criterion (BIC) for the models examined when you request variable selection with the SELECTION= option in the MODEL statement.

   The following *bic-options* are available for models where you request the RSQUARE, ADJRSQ, or CP selection method:

   **LABEL**
   
   requests that the model number corresponding to the one displayed in the “Subset Selection Summary” table be used to label the model with the smallest BIC statistic at each value of the number of parameters.

   **LABELVARS**
   
   requests that the list (excluding the intercept) of the regressors in the relevant model be used to label the model with the smallest BIC statistic at each value of the number of parameters.
COOKSD < (LABEL) >
plots Cook’s D statistic by observation number. Observations whose Cook’s D statistic lies above the horizontal reference line at value $4/n$, where $n$ is the number of observations used, are deemed to be influential (Rawlings, Pantula, and Dickey 1998). If you specify the LABEL option, then points deemed as influential are labeled. If you do not specify an ID variable, the observation number within the current BY group is used as the label. If you specify one or more ID variables in one or more ID statements, then the first ID variable you specify is used for the labeling.

CP < (cp-options) >
displays Mallows’ $C_p$ statistic for the models examined when you request variable selection with the SELECTION= option in the MODEL statement. For models where you request the RSQUARE, ADJRSQ, or CP selection, reference lines corresponding to the equations $C_p = p$ and $C_p = 2p - p_{full}$, where $p_{full}$ is the number of parameters in the full model (excluding the intercept) and $p$ is the number of parameters in the subset model (including the intercept), are displayed on the plot of $C_p$ versus $p$. For the purpose of parameter estimation, Hocking (1976) suggests selecting a model where $C_p \leq 2p - p_{full}$. For the purpose of prediction, Hocking suggests the criterion $C_p \leq p$. Mallows (1973) suggests that all subset models with $C_p$ small and near $p$ be considered for further study.

The following cp-options are available for models where you request the RSQUARE, ADJRSQ, or CP selection method:

LABEL
requests that the model number corresponding to the one displayed in the “Subset Selection Summary” table be used to label the model with the smallest $C_p$ statistic at each value of the number of parameters.

LABELVARS
requests that the list (excluding the intercept) of the regressors in the relevant model be used to label the model with the smallest $C_p$ statistic at each value of the number of parameters.

CRITERIA | CRITERIONPANEL < (criteria-options) >
produces a panel of fit criteria for the models examined when you request variable selection with the SELECTION= option in the MODEL statement. The fit criteria displayed are R-square, adjusted R-square, Mallows’ $C_p$, Akaike’s information criterion (AIC), Sawa’s Bayesian information criterion (BIC), and Schwarz’s Bayesian information criterion (SBC). For SELECTION=RSQUARE, SELECTION=ADJRSQ, or SELECTION=CP, scatter plots of these statistics versus the number of parameters (including the intercept) are displayed. For other selection methods, line plots of these statistics as function of the selection step number are displayed.

The following criteria-options are available:

LABEL
requests that the model number corresponding to the one displayed in the “Subset Selection Summary” table be used to label the best model at each value of the number of parameters. This option applies only to the RSQUARE, ADJRSQ, and CP selection methods.
**LABELVARS**
requests that the list (excluding the intercept) of the regressors in the relevant model be used
to label the best model at each value of the number of parameters. Since these labels are
typically long, LABELVARS is supported only when the panel is unpacked. This option
applies only to the RSQUARE, ADJRSQ, and CP selection methods.

**UNPACK**
suppresses paneling. Separate plots are produced for each of the six fit statistics. For
models where you request the RSQUARE, ADJRSQ, or CP selection, two reference lines
corresponding to the equations $C_p = p$ and $C_p = 2p - p_{full}$, where $p_{full}$ is the number
of parameters in the full model (excluding the intercept) and $p$ is the number of parameters
in the subset model (including the intercept), are displayed on the plot of $C_p$ versus $p$. For
the purpose of parameter estimation, Hocking (1976) suggests selecting a model where
$C_p \leq 2p - p_{full}$. For the purpose of prediction, Hocking suggests the criterion $C_p \leq p$.
Mallows (1973) suggests that all subset models with $C_p$ small and near $p$ be considered for
further study.

**DFBETAS < (DFBETAS-options)>**
produces panels of DFBETAS by observation number for the regressors in the model. Note
that each panel contains at most six plots, and multiple panels are used in the case where
there are more than six regressors (including the intercept) in the model. Observations whose
DFBETAS’ statistics for a regressor are greater in magnitude than $2/ \sqrt{n}$, where $n$ is the number
of observations used, are deemed to be influential for that regressor (Rawlings, Pantula, and
Dickey 1998).

The following **DFBETAS-options** are available:

- **COMMONAXES**
specifies that the same DFBETAS axis be used in all panels when multiple panels are needed.
  By default, the DFBETAS axis is chosen independently for each panel. If you also specify
  the UNPACK option, then the same DFBETAS axis is used for each regressor.

- **LABEL**
specifies that observations whose magnitude are greater than $2/ \sqrt{n}$ be labeled. If you do
  not specify an ID variable, the observation number within the current BY group is used as
  the label. If you specify one or more ID variables on one or more ID statements, then the
  first ID variable you specify is used for the labeling.

- **UNPACK**
suppresses paneling. The DFBETAS statistics for each regressor are displayed on separate
  plots.

**DFFITS < (LABEL)>**
plots the DFFITS statistic by observation number. Observations whose DFFITS’ statistic is
greater in magnitude than $2 \sqrt{p/n}$, where $n$ is the number of observations used and $p$ is the
number of regressors, are deemed to be influential (Rawlings, Pantula, and Dickey 1998). If you
specify the LABEL option, then these influential observations are labeled. If you do not specify
an ID variable, the observation number within the current BY group is used as the label. If you
specify one or more ID variables in one or more ID statements, then the first ID variable you
specify is used for the labeling.
DIAGNOSTICS < (diagnostics-options) >
produces a summary panel of fit diagnostics:

- residuals versus the predicted values
- studentized residuals versus the predicted values
- studentized residuals versus the leverage
- normal quantile plot of the residuals
- dependent variable values versus the predicted values
- Cook’s D versus observation number
- histogram of the residuals
- “Residual-Fit” (or RF) plot consisting of side-by-side quantile plots of the centered fit and the residuals
- box plot of the residuals if you specify the STATS=NONE suboption

You can specify the following diagnostics-options:

STATS=stats-options

determines which model fit statistics are included in the panel. See the global STATS= suboption for details. The PLOTS= suboption of the DIAGNOSTICSPANEL option overrides the global PLOTS= suboption.

UNPACK
produces the eight plots in the panel as individual plots. Note that you can also request individual plots in the panel by name without having to unpack the panel.

FITPLOT | FIT < (fit-options) >
produces a scatter plot of the data overlaid with the regression line, confidence band, and prediction band for models that depend on at most one regressor excluding the intercept. When the number of points exceeds the MAXPOINTS=max value, a heat map is displayed instead of a scatter plot. By default, heat maps are not displayed if the number of observations times the number of independent variables is greater than 150,000. See the MAXPOINTS= option.

You can specify the following fit-options:

NOCLI
suppresses the prediction limits.

NOCLM
suppresses the confidence limits.

NOLIMITS
suppresses the confidence and prediction limits.

STATS=stats-options

determines which model fit statistics are included in the panel. See the global STATS= suboption for details. The PLOTS= suboption of the FITPLOT option overrides the global PLOTS= suboption.
OBSERVEDBYPREDICTED < (LABEL) >
plots dependent variable values by the predicted values. If you specify the LABEL option, then points deemed as outliers or influential (see the RSTUDENTBYLEVERAGE option for details) are labeled.

NONE
suppresses all plots.

PARTIAL < (UNPACK) >
produces panels of partial regression plots for each regressor with at most six regressors per panel. If you specify the UNPACK option, then all partial plot panels are unpacked.

PREDICTIONS (X=numeric-variable < prediction-options >)
produces a panel of two plots whose horizontal axis is the variable you specify in the required X= suboption. The upper plot in the panel is a scatter plot of the residuals. The lower plot shows the data overlaid with the regression line, confidence band, and prediction band. This plot is appropriate for models where all regressors are known to be functions of the single variable that you specify in the X= suboption.

You can specify the following prediction-options:

NOCLI
suppresses the prediction limits.

NOCLM
suppresses the confidence limits

NOLIMITS
suppresses the confidence and prediction limits

SMOOTH
requests a nonparametric smoothing of the residuals as a function of the variable you specify in the X= suboption. This nonparametric fit is a loess fit that uses local linear polynomials, linear interpolation, and a smoothing parameter that is selected to yield a local minimum of the corrected Akaike’s information criterion (AICC). See Chapter 75, “The LOESS Procedure,” for details. The SMOOTH option is not supported when a FREQ statement is used.

UNPACK
suppresses paneling.

QQPLOT | QQ
produces a normal quantile plot of the residuals.

RESIDUALBOXPLOT | BOXPLOT < (LABEL) >
produces a box plot consisting of the residuals. If you specify label option, points deemed far-outliers are labeled. If you do not specify an ID variable, the observation number within the current BY group is used as the label. If you specify one or more ID variables in one or more ID statements, then the first ID variable you specify is used for the labeling.
RESIDUALBYPREDICTED < (LABEL) > plots residuals by predicted values. If you specify the LABEL option, then points deemed as outliers or influential (see the RSTUDENTBYLEVERAGE option for details) are labeled.

RESIDUALCHART < (residual-chart-options) > RC < (residual-chart-options) > produces the residual chart and enables you to specify residual-chart-options. This chart displays studentized residuals and Cook’s D in side-by-side bar charts. This chart is also displayed when you specify the R option in the MODEL statement.

Unlike most graphs, the height of this chart can vary as a function of the number of observations that appear in the chart. You can specify the following residual-chart-options to control the height and other aspects of the chart:

COMPUTEHEIGHT=a b < max > CH=a b < max > specifies the constants for computing the height of the chart. For n observations, intercept a, slope b, and maximum height max, the height is \( \text{min}(a + b(n + 1), \text{max}) \). By default, COMPUTEHEIGHT=150 15 1650. Thus, the default height in pixels is \( \text{min}(150 + 15(n + 1), 1650) \). The default unit is pixels, and you can use the UNIT= residual-chart-option to change the unit to inches or centimeters.

MAX=max species the maximum number of points to display in each chart. When the number of points exceeds max, charts of up to max observations are displayed until all observations are displayed.

SETHEIGHT=height SH=height specifies the height of the chart. By default, the height is based on the COMPUTEHEIGHT= option. The default unit is pixels, and you can use the UNIT= residual-chart-option to change the unit to inches or centimeters.

UNIT=PX | IN | CM specifies the unit (pixels, inches, or centimeters) for the SETHEIGHT= and COMPUTEHEIGHT= residual-chart-options. Inches equals pixels divided by 96, and centimeters equals inches times 2.54. By default, UNIT=PX.

UNPACK suppresses paneling. The studentized residuals and Cook’s D are displayed in separate charts. When you specify the UNPACK residual-chart-option, residuals, standard errors, and other values that go into the computations are added to each chart.

RESIDUALS < (residual-options) > produces panels of the residuals versus the regressors in the model. Each panel contains at most six plots, and multiple panels are used when the model contains more than six regressors (including the intercept). When the number of points exceeds the MAXPOINTS=max value, a heat map is displayed instead of a scatter plot. By default, heat maps are not displayed if the number of observations times the number of independent variables is greater than 150,000. See the MAXPOINTS= option. You can specify the following residual-options:
SMOOTH requests a nonparametric smoothing of the residuals for each regressor. Each nonparametric fit is a loess fit that uses local linear polynomials, linear interpolation, and a smoothing parameter that is selected to yield a local minimum of the corrected Akaike’s information criterion (AICC). See Chapter 75, “The LOESS Procedure,” for details. The SMOOTH option is not supported when a FREQ statement is used.

UNPACK suppresses paneling.

RESIDUALHISTOGRAM produces a histogram of the residuals.

RFPLOT | RF produces a “Residual-Fit” (or RF) plot consisting of side-by-side quantile plots of the centered fit and the residuals. This plot “shows how much variation in the data is explained by the fit and how much remains in the residuals” (Cleveland 1993).

RIDGE | RIDGE PANEL | RIDGE PLOT < (ridge-options) > creates panels of VIF values and standardized ridge estimates by ridge values for each coefficient. The VIF values for each coefficient are connected by lines and are displayed in the upper plot in each panel. The points corresponding to the standardized estimates of each coefficient are connected by lines and are displayed in the lower plot in each panel. By default, at most 10 coefficients are represented in a panel and multiple panels are produced for models with more than 10 regressors. For ridge estimates to be computed and plotted, the OUTEST= option must be specified in the PROC REG statement, and the RIDGE= list must be specified in either the PROC REG or the MODEL statement. (See Example 102.5.)

The following ridge-options are available:

COMMONAXES specifies that the same VIF axis and the same standardized estimate axis are used in all panels when multiple panels are needed. By default, these axes are chosen independently for the regressors shown in each panel.

RIDGEAXIS=LINEAR | LOG specifies the axis type used to display the ridge parameters. The default is RIDGEAXIS=LINEAR. Note that the point with the ridge parameter equal to zero is not displayed if you specify RIDGEAXIS=LOG.

UNPACK suppresses paneling. The traces of the VIF statistics and standardized estimates are shown in separate plots.

VARSPERPLOT=ALL

VARSPERPLOT=number specifies the maximum number of regressors displayed in each panel or in each plot if you additionally specify the UNPACK option. If you specify VARSPERPLOT=ALL, then the VIF values and ridge traces for all regressors are displayed in a single panel.
VIFAXIS=LINEAR | LOG
specifies the axis type used to display the VIF statistics. The default is VIFAXIS=LINEAR.

RSQUARE < (rsquare-options)>
displays the R-square values for the models examined when you request variable selection with the SELECTION= option in the MODEL statement.

The following rsquare-options are available for models where you request the RSQUARE, ADJRSQ, or CP selection method:

LABEL
requests that the model number corresponding to the one displayed in the “Subset Selection Summary” table be used to label the model with the largest R-square statistic at each value of the number of parameters.

LABELVARS
requests that the list (excluding the intercept) of the regressors in the relevant model be used to label the model with the largest R-square statistic at each value of the number of parameters.

RSTUDENTBYLEVERAGE < (LABEL)>
plots studentized residuals by leverage. Observations whose studentized residuals lie outside the band between the reference lines RSTUDENT = ±2 are deemed outliers. Observations whose leverage values are greater than the vertical reference LEVERAGE = 2p/n, where p is the number of parameters including the intercept and n is the number of observations used, are deemed influential (Rawlings, Pantula, and Dickey 1998). If you specify the LABEL option, then points deemed as outliers or influential are labeled. If you do not specify an ID variable, the observation number within the current BY group is used as the label. If you specify one or more ID variables in one or more ID statements, then the first ID variable you specify is used for the labeling.

RSTUDENTBYPREDICTED < (LABEL)>
plots studentized residuals by predicted values. If you specify the LABEL option, then points deemed as outliers or influential (see the RSTUDENTBYLEVERAGE option for details) are labeled.

SBC < (sbc-options)>
displays Schwarz’s Bayesian information criterion (SBC) for the models examined when you request variable selection with the SELECTION= option in the MODEL statement.

The following sbc-options are available for models where you request the RSQUARE, ADJRSQ, or CP selection method:

LABEL
requests that the model number corresponding to the one displayed in the “Subset Selection Summary” table be used to label the model with the smallest SBC statistic at each value of the number of parameters.
**LABELVARS**
requests that the list (excluding the intercept) of the regressors in the relevant model be used to label the model with the smallest SBC statistic at each value of the number of parameters.

**PRESS**
outputs the PRESS statistic to the OUTEST= data set. The values of this statistic are saved in the variable _PRESS_. This option is available for all model selection methods except RSQUARE, ADJRSQ, and CP.

**RIDGE=list**
requests a ridge regression analysis and specifies the values of the ridge constant $k$ (see the section “Computations for Ridge Regression and IPC Analysis” on page 8536). Each value of $k$ produces a set of ridge regression estimates that are placed in the OUTTEST= data set. The values of $k$ are saved by the variable _RIDGE_, and the value of the variable _TYPE_ is set to RIDGE to identify the estimates. Only nonnegative numbers can be specified with the RIDGE= option. Example 102.5 illustrates this option.

If ODS Graphics is enabled (see the section “ODS Graphics” on page 8557), then ridge regression plots are automatically produced. These plots consist of panels containing ridge traces for the regressors, with at most eight ridge traces per panel.

If you specify the RIDGE= option, RESTRICT statements are ignored.

**RSQUARE**
has the same effect as the EDF option.

**SIMPLE**
displays the sum, mean, variance, standard deviation, and uncorrected sum of squares for each variable used in PROC REG.

**SINGULAR=n**
tunes the mechanism used to check for singularities. The default value is machine dependent but is approximately 1E–9 on most machines. This option is rarely needed.

Singularity checking is described in the section “Computational Methods” on page 8537.

**TABLEOUT**
outputs the standard errors and $100(1 - \alpha)\%$ confidence limits for the parameter estimates, the $t$ statistics for testing if the estimates are zero, and the associated $p$-values to the OUTTEST= data set. The _TYPE_ variable values STDERR, LnB, UnB, T, and PVALUE, where $n = 100(1 - \alpha)$, identify these rows in the OUTTEST= data set. The $\alpha$ level can be set with the ALPHA= option in the PROC REG or MODEL statement. The OUTTEST= option must be specified in the PROC REG statement for this option to take effect.

**USSCP**
displays the uncorrected sums-of-squares and crossproducts matrix for all variables used in the procedure.
ADD Statement

```
ADD variables;
```

The ADD statement adds independent variables to the regression model. Only variables used in the VAR statement or used in MODEL statements before the first RUN statement can be added to the model. You can use the ADD statement interactively to add variables to the model or to include a variable that was previously deleted with a DELETE statement. Each use of the ADD statement modifies the MODEL label.

See the section “Interactive Analysis” on page 8497 for an example.

BY Statement

```
BY variables;
```

You can specify a BY statement in PROC REG to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement in the REG procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

When a BY statement is used with PROC REG, interactive processing is not possible; that is, once the first RUN statement is encountered, processing proceeds for each BY group in the data set, and no further statements are accepted by the procedure. A BY statement that appears after the first RUN statement is ignored.

For more information about BY-group processing, see the discussion in SAS Language Reference: Concepts. For more information about the DATASETS procedure, see the discussion in the Base SAS Procedures Guide.

CODE Statement

```
CODE < options >;
```

The CODE statement writes SAS DATA step code for computing predicted values of the fitted model either to a file or to a catalog entry. This code can then be included in a DATA step to score new data.

Table 102.4 summarizes the options available in the CODE statement.
### DELETE Statement

**DELETE** `variables` ;

The DELETE statement deletes independent variables from the regression model. The DELETE statement performs the opposite function of the ADD statement and is used in a similar manner. Each use of the DELETE statement modifies the MODEL label.

For an example of how the ADD statement is used (and how the DELETE statement can be used), see the section “Interactive Analysis” on page 8497.

### FREQ Statement

**FREQ** `variable` ;

When a FREQ statement appears, each observation in the input data set is assumed to represent `n` observations, where `n` is the value of the FREQ variable. The analysis produced when you use a FREQ statement is the same as an analysis produced by using a data set that contains `n` observations in place of each observation in the input data set. When the procedure determines degrees of freedom for significance tests, the total number of observations is considered to be equal to the sum of the values of the FREQ variable.

If the value of the FREQ variable is missing or is less than 1, the observation is not used in the analysis. If the value is not an integer, only the integer portion is used.

The FREQ statement must appear before the first RUN statement, or it is ignored.

---

**Table 102.4** CODE Statement Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CATALOG=</td>
<td>Names the catalog entry where the generated code is saved</td>
</tr>
<tr>
<td>DUMMIES</td>
<td>Retains the dummy variables in the data set</td>
</tr>
<tr>
<td>ERROR</td>
<td>Computes the error function</td>
</tr>
<tr>
<td>FILE=</td>
<td>Names the file where the generated code is saved</td>
</tr>
<tr>
<td>FORMAT=</td>
<td>Specifies the numeric format for the regression coefficients</td>
</tr>
<tr>
<td>GROUP=</td>
<td>Specifies the group identifier for array names and statement labels</td>
</tr>
<tr>
<td>IMPUTE</td>
<td>Imputes predicted values for observations with missing or invalid covariates</td>
</tr>
<tr>
<td>LINESIZE=</td>
<td>Specifies the line size of the generated code</td>
</tr>
<tr>
<td>LOOKUP=</td>
<td>Specifies the algorithm for looking up CLASS levels</td>
</tr>
<tr>
<td>RESIDUAL</td>
<td>Computes residuals</td>
</tr>
</tbody>
</table>

For details about the syntax of the CODE statement, see the section “CODE Statement” on page 400 in Chapter 19, “Shared Concepts and Topics.”
ID Statement

ID variables;

When one of the MODEL statement options CLI, CLM, P, R, and INFLUENCE is requested, the variables listed in the ID statement are displayed beside each observation. These variables can be used to identify each observation. If the ID statement is omitted, the observation number is used to identify the observations.

Although there are no restrictions on the length of ID variables, PROC REG might truncate ID values to 16 characters for display purposes.

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 numeric variables in the data set being analyzed. For example, if you want to specify a quadratic term for variable \( X_1 \) in the model, you cannot use \( X_1 \times X_1 \) in the MODEL statement but must create a new variable (for example, \( X1SQUARE=X1\times X1 \)) in a DATA step and use this new variable in the MODEL statement. The label in the MODEL statement is optional.

Table 102.5 summarizes the options available in the MODEL statement. Equations for the statistics available are given in the section “Model Fit and Diagnostic Statistics” on page 8514.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model Selection and Details of Selection</td>
<td></td>
</tr>
<tr>
<td>SELECTION=</td>
<td>Specifies model selection method</td>
</tr>
<tr>
<td>BEST=</td>
<td>Specifies maximum number of subset models displayed or output to the OUTEST= data set</td>
</tr>
<tr>
<td>DETAILS</td>
<td>Produces summary statistics at each step</td>
</tr>
<tr>
<td>DETAILS=</td>
<td>Specifies the display details for FORWARD, BACKWARD, and STEPWISE methods</td>
</tr>
<tr>
<td>GROUPNAMES=</td>
<td>Provides names for groups of variables</td>
</tr>
<tr>
<td>INCLUDE=</td>
<td>Includes first ( n ) variables in the model</td>
</tr>
<tr>
<td>MAXSTEP=</td>
<td>Specifies maximum number of steps that might be performed</td>
</tr>
<tr>
<td>NOINT</td>
<td>Fits a model without the intercept term</td>
</tr>
<tr>
<td>PCOMIT=</td>
<td>Performs incomplete principal component analysis and outputs estimates to the OUTEST= data set</td>
</tr>
<tr>
<td>RIDGE=</td>
<td>Performs ridge regression analysis and outputs estimates to the OUTEST= data set</td>
</tr>
<tr>
<td>SLE=</td>
<td>Sets criterion for entry into model</td>
</tr>
<tr>
<td>SLS=</td>
<td>Sets criterion for staying in model</td>
</tr>
<tr>
<td>START=</td>
<td>Specifies number of variables in model to begin the comparing and switching process</td>
</tr>
</tbody>
</table>
### Table 102.5  continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>STOP=</td>
<td>Stops selection criterion</td>
</tr>
</tbody>
</table>

**Statistics**
- **ADJRSQ**: Computes adjusted R square
- **AIC**: Computes Akaike’s information criterion
- **B**: Computes parameter estimates for each model
- **BIC**: Computes Sawa’s Bayesian information criterion
- **CP**: Computes Mallows’ $C_p$ statistic
- **GMSEP**: Computes estimated MSE of prediction assuming multivariate normality
- **JP**: Computes $J_p$, the final prediction error
- **MSE**: Computes MSE for each model
- **PC**: Computes Amemiya’s prediction criterion
- **RMSE**: Displays root MSE for each model
- **SBC**: Computes the SBC statistic
- **SP**: Computes $S_p$ statistic for each model
- **SSE**: Computes error sum of squares for each model

**Data Set Options**
- **EDF**: Outputs the number of regressors, the error degrees of freedom, and the model R square to the OUTEST= data set
- **OUTSEB**: Outputs standard errors of the parameter estimates to the OUTEST= data set
- **OUTSTB**: Outputs standardized parameter estimates to the OUTEST= data set. Use only with the RIDGE= or PCOMIT= option.
- **OUTVIF**: Outputs the variance inflation factors to the OUTEST= data set. Use only with the RIDGE= or PCOMIT= option.
- **PRESS**: Outputs the PRESS statistic to the OUTEST= data set
- **RSQUARE**: Has same effect as the EDF option

**Regression Calculations**
- **I**: Displays inverse of sums of squares and crossproducts
- **XPX**: Displays sums-of-squares and crossproducts matrix

**Details on Estimates**
- **ACOV**: Displays heteroscedasticity- consistent covariance matrix of estimates and heteroscedasticity-consistent standard errors
- **ACOVMETHOD=**: Specifies method for computing the asymptotic heteroscedasticity-consistent covariance matrix
- **COLLIN**: Produces collinearity analysis
- **COLLINPOINT**: Produces collinearity analysis with intercept adjusted out
- **CORRB**: Displays correlation matrix of estimates
- **COVB**: Displays covariance matrix of estimates
- **HCC**: Displays heteroscedasticity-consistent standard errors
- **HCCMETHOD=**: Specifies method for computing the asymptotic heteroscedasticity-consistent covariance matrix
- **LACKFIT**: Performs lack-of-fit test
<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>PARTIALR2</td>
<td>Displays squared semipartial correlation coefficients computed using Type I sums of squares</td>
</tr>
<tr>
<td>PCORR1</td>
<td>Displays squared partial correlation coefficients computed using Type I sums of squares</td>
</tr>
<tr>
<td>PCORR2</td>
<td>Displays squared partial correlation coefficients computed using Type II sums of squares</td>
</tr>
<tr>
<td>SCORR1</td>
<td>Displays squared semipartial correlation coefficients computed using Type I sums of squares</td>
</tr>
<tr>
<td>SCORR2</td>
<td>Displays squared semipartial correlation coefficients computed using Type II sums of squares</td>
</tr>
<tr>
<td>SEQB</td>
<td>Displays a sequence of parameter estimates during selection process</td>
</tr>
<tr>
<td>SPEC</td>
<td>Tests that first and second moments of model are correctly specified</td>
</tr>
<tr>
<td>SRT</td>
<td>Displays the studentized residuals by using a series of asterisks</td>
</tr>
<tr>
<td>SS1</td>
<td>Displays the sequential sums of squares</td>
</tr>
<tr>
<td>SS2</td>
<td>Displays the partial sums of squares</td>
</tr>
<tr>
<td>STB</td>
<td>Displays standardized parameter estimates</td>
</tr>
<tr>
<td>TOL</td>
<td>Displays tolerance values for parameter estimates</td>
</tr>
<tr>
<td>WHITE</td>
<td>Displays heteroscedasticity-consistent standard errors</td>
</tr>
<tr>
<td>VIF</td>
<td>Computes variance-inflation factors</td>
</tr>
</tbody>
</table>

Predicted and Residual Values

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLB</td>
<td>Computes $100(1 - \alpha)$% confidence limits for the parameter estimates</td>
</tr>
<tr>
<td>CLI</td>
<td>Computes $100(1 - \alpha)$% confidence limits for an individual predicted value</td>
</tr>
<tr>
<td>CLM</td>
<td>Computes $100(1 - \alpha)$% confidence limits for the expected value of the dependent variable</td>
</tr>
<tr>
<td>DW</td>
<td>Computes a Durbin-Watson statistic</td>
</tr>
<tr>
<td>DWPROB</td>
<td>Computes a Durbin-Watson statistic and $p$-value</td>
</tr>
<tr>
<td>INFLUENCE</td>
<td>Computes influence statistics</td>
</tr>
<tr>
<td>P</td>
<td>Computes predicted values</td>
</tr>
<tr>
<td>PARTIAL</td>
<td>Displays partial regression plots for each regressor</td>
</tr>
<tr>
<td>PARTIALDATA</td>
<td>Displays partial regression data</td>
</tr>
<tr>
<td>R</td>
<td>Produces analysis of residuals</td>
</tr>
</tbody>
</table>

Display Options and Other Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>Requests the following options: ACOV, CLB, CLI, CLM, CORRB, COVB, HCC, I, P, PCORR1, PCORR2, R, SCORR1, SCORR2, SEQB, SPEC, SS1, SS2, STB, TOL, VIF, XPX</td>
</tr>
<tr>
<td>ALPHA=</td>
<td>Sets significance value for confidence and prediction intervals and tests</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>Suppresses display of results</td>
</tr>
<tr>
<td>SIGMA=</td>
<td>Specifies the true standard deviation of error term for computing CP and BIC</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>Sets criterion for checking for singularity</td>
</tr>
</tbody>
</table>
You can specify the following options in the MODEL statement after a slash (/).

ACOV
displays the estimated asymptotic covariance matrix of the estimates under the hypothesis of heteroscedasticity and heteroscedasticity-consistent standard errors of parameter estimates. See the HCCMETHOD= option and the HCC option and the section “Testing for Heteroscedasticity” on page 8530 for more information.

ACOVMETHOD=0,1,2, or 3
See the HCCMETHOD= option.

ADJRSQ
computes R square adjusted for degrees of freedom for each model selected (Darlington 1968; Judge et al. 1980).

AIC
outputs Akaike’s information criterion for each model selected (Akaike 1969; Judge et al. 1980) to the OUTEST= data set. If SELECTION=ADJRSQ, SELECTION=RSQUARE, or SELECTION=CP is specified, then the AIC statistic is also added to the SubsetSelSummary table.

ALL
requests all these options: ACOV, CLB, CLI, CLM, CORRB, COVB, HCC, I, P, PCORR1, PCORR2, R, SCORR1, SCORR2, SEQB, SPEC, SS1, SS2, STB, TOL, VIF, and XPX.

ALPHA=number
sets the significance level used for the construction of confidence intervals for the current MODEL statement. The value must be between 0 and 1; the default value of 0.05 results in 95% intervals. This option affects the MODEL options CLB, CLI, and CLM; the OUTPUT statement keywords LCL, LCLM, UCL, and UCLM; the PLOT statement keywords LCL., LCLM., UCL., and UCLM.; and the PLOT statement options CONF and PRED. If you specify this option in the MODEL statement, it takes precedence over the ALPHA= option in the PROC REG statement.

B
is used with the RSQUARE, ADJRSQ, and CP model-selection methods to compute estimated regression coefficients for each model selected.

BEST=n
is used with the RSQUARE, ADJRSQ, and CP model-selection methods. If SELECTION=CP or SELECTION=ADJRSQ is specified, the BEST= option specifies the maximum number of subset models to be displayed or output to the OUTEST= data set. For SELECTION=RSQUARE, the BEST= option requests the maximum number of subset models for each size.

If the BEST= option is used without the B option (displaying estimated regression coefficients), the variables in each MODEL are listed in order of inclusion instead of the order in which they appear in the MODEL statement.

If the BEST= option is omitted and the number of regressors is less than 11, all possible subsets are evaluated. If the BEST= option is omitted and the number of regressors is greater than 10, the number of subsets selected is, at most, equal to the number of regressors. A small value of the BEST= option greatly reduces the CPU time required for large problems.
BIC
outputs Sawa’s Bayesian information criterion for each model selected (Sawa 1978; Judge et al. 1980) to the OUTEST= data set. If SELECTION=ADJRSQ, SELECTION=RSQUARE, or SELECTION=CP is specified, then the BIC statistic is also added to the SubsetSelSummary table.

CLB
requests the 100(1 – \( \alpha \))% upper and lower confidence limits for the parameter estimates. By default, the 95% limits are computed; the ALPHA= option in the PROC REG or MODEL statement can be used to change the \( \alpha \) level. If any of the MODEL statement options ACOV, HCC, or WHITE are in effect, then the CLB option also produces heteroscedasticity-consistent 100(1 – \( \alpha \))% upper and lower confidence limits for the parameter estimates.

CLI
requests the 100(1 – \( \alpha \))% upper and lower confidence limits for an individual predicted value. By default, the 95% limits are computed; the ALPHA= option in the PROC REG or MODEL statement can be used to change the \( \alpha \) level. The confidence limits reflect variation in the error, as well as variation in the parameter estimates. See the section “Predicted and Residual Values” on page 8507 and Chapter 4, “Introduction to Regression Procedures,” for more information.

CLM
displays the 100(1 – \( \alpha \))% upper and lower confidence limits for the expected value of the dependent variable (mean) for each observation. By default, the 95% limits are computed; the ALPHA= in the PROC REG or MODEL statement can be used to change the \( \alpha \) level. 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. See the section “Predicted and Residual Values” on page 8507 and Chapter 4, “Introduction to Regression Procedures,” for more information.

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 the section “Collinearity Diagnostics” on page 8513.

COLLINOINT
requests the same analysis as the COLLIN option with the intercept variable adjusted out rather than included in the diagnostics. See the section “Collinearity Diagnostics” on page 8513.

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

COVB
displays the estimated covariance matrix of the estimates. This matrix is \((X'X)^{-1}s^2\), where \(s^2\) is the estimated mean squared error.

CP
outputs Mallows’ \(C_p\) statistic for each model selected (Mallows 1973; Hocking 1976) to the OUTEST= data set. See the section “Criteria Used in Model-Selection Methods” on page 8504 for a discussion of the use of \(C_p\). If SELECTION=ADJRSQ, SELECTION=RSQUARE, or SELECTION=CP is specified, then the \(C_p\) statistic is also added to the SubsetSelSummary table.
DETAILS

**DETAILS=** name

specifies the level of detail produced when the BACKWARD, FORWARD, or STEPWISE method is used, where name can be ALL, STEPS, or SUMMARY. The DETAILS or DETAILS=ALL option produces entry and removal statistics for each variable in the model building process, ANOVA and parameter estimates at each step, and a selection summary table. The option DETAILS=STEPS provides the step information and summary table. The option DETAILS=SUMMARY produces only the summary table. The default if the DETAILS option is omitted is DETAILS=STEPS.

**DW**

calculates a Durbin-Watson statistic to test whether or not the errors have first-order autocorrelation. (This test is appropriate only for time series data.) Note that your data should be sorted by the date/time ID variable before you use this option. The sample autocorrelation of the residuals is also produced. See the section “Autocorrelation in Time Series Data” on page 8535.

**DWPROB**

calculates a Durbin-Watson statistic and a $p$-value to test whether or not the errors have first-order autocorrelation. Note that it is not necessary to specify the DW option if the DWPROB option is specified. (This test is appropriate only for time series data.) Note that your data should be sorted by the date/time ID variable before you use this option. The sample autocorrelation of the residuals is also produced. See the section “Autocorrelation in Time Series Data” on page 8535.

**EDF**

outputs the number of regressors in the model excluding and including the intercept, the error degrees of freedom, and the model R square to the OUTEST= data set.

**GMSEP**

outputs the estimated mean square error of prediction assuming that both independent and dependent variables are multivariate normal (Stein 1960; Darlington 1968) to the OUTEST= data set. (Note that Hocking’s formula (1976, eq. 4.20) contains a misprint: “$n - 1$” should read “$n - 2$.”) If SELECTION=ADJRSQ, SELECTION=RSQUARE, or SELECTION=CP is specified, then the GMSEP statistic is also added to the SubsetSelSummary table.

**GROUPNAMES=** name1 name2 . . .

provides names for variable groups. This option is available only in the BACKWARD, FORWARD, and STEPWISE methods. The group name can be up to 32 characters. Subsets of independent variables listed in the MODEL statement can be designated as variable groups. This is done by enclosing the appropriate variables in braces. Variables in the same group are entered into or removed from the regression model at the same time. However, if the tolerance of any variable (see the TOL option on page 8477) in a group is less than the setting of the SINGULAR= option, then the variable is not entered into the model with the rest of its group. If the GROUPNAMES= option is not used, then the names GROUP1, GROUP2, . . . , GROUPn are assigned to groups encountered in the MODEL statement. Variables not enclosed by braces are used as groups of a single variable.
For example:

```plaintext
model y=(x1 x2) x3 / selection=stepwise
  groupnames='x1 x2' 'x3';
```

Another example:

```plaintext
model y=(ht wgt age) bodyfat / selection=forward
  groupnames='htwgtage' 'bodyfat';
```

HCC requests heteroscedasticity-consistent standard errors of the parameter estimates. You can use the HCCMETHOD= option to specify the method used to compute the heteroscedasticity-consistent covariance matrix.

HCCMETHOD=0,1,2, or 3 specifies the method used to obtain a heteroscedasticity-consistent covariance matrix for use with the ACOV, HCC, or WHITE option in the MODEL statement and for heteroscedasticity-consistent tests with the TEST statement. The default is HCCMETHOD=0. See the section “Testing for Heteroscedasticity” on page 8530 for details.

I displays the \((X'X)^{-1}\) matrix. The inverse of the crossproducts matrix is bordered by the parameter estimates and SSE matrices.

INCLUDE=n forces the first \(n\) independent variables listed in the MODEL statement to be included in all models. The selection methods are performed on the other variables in the MODEL statement. The INCLUDE= option is not available with SELECTION=NONE.

INFLUENCE requests a detailed analysis of the influence of each observation on the estimates and the predicted values. See the section “Influence Statistics” on page 8516 for details.

JP outputs \(J_p\), the estimated mean square error of prediction for each model selected assuming that the values of the regressors are fixed and that the model is correct to the OUTEST= data set. The \(J_p\) statistic is also called the final prediction error (FPE) by Akaike (Nicholson 1948; Lord 1950; Mallows 1967; Darlington 1968; Rothman 1968; Akaike 1969; Hocking 1976; Judge et al. 1980) If SELECTION=ADJRSQ, SELECTION=RSQUARE, or SELECTION=CP is specified, then the \(J_p\) statistic is also added to the SubsetSelSummary table.

LACKFIT performs a lack-of-fit test. See the section “Testing for Lack of Fit” on page 8531 for more information. See Draper and Smith (1981) for a discussion of lack-of-fit tests.
MSE

computes the mean square error for each model selected (Darlington 1968).

\textbf{MAXSTEP} = n

specifies the maximum number of steps that are done when \texttt{SELECTION=FORWARD}, \texttt{SELECTION=BACKWARD}, or \texttt{SELECTION=STEPWISE} is used. The default value is the number of independent variables in the model for the FORWARD and BACKWARD methods and three times this number for the stepwise method.

\textbf{NOINT}

suppresses the intercept term that is otherwise included in the model.

\textbf{NOPRINT}

suppresses the normal display of regression results. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 20, “Using the Output Delivery System,” for more information.

\textbf{OUTSEB}

outputs the standard errors of the parameter estimates to the OUTEST= data set. The value SEB for the variable \_TYPE\_ identifies the standard errors. If the RIDGE= or PCOMIT= option is specified, additional observations are included and identified by the values RIDGESEB and IPCSEB, respectively, for the variable \_TYPE\_. The standard errors for ridge regression estimates and incomplete principal components (IPC) estimates are limited in their usefulness because these estimates are biased. This option is available for all model-selection methods except RSQUARE, ADJRSQ, and CP.

\textbf{OUTSTB}

outputs the standardized parameter estimates as well as the usual estimates to the OUTEST= data set when the RIDGE= or PCOMIT= option is specified. The values RIDGESTB and IPCSTB for the variable \_TYPE\_ identify ridge regression estimates and IPC estimates, respectively.

\textbf{OUTVIF}

outputs the variance inflation factors (VIF) to the OUTEST= data set when the RIDGE= or PCOMIT= option is specified. The factors are the diagonal elements of the inverse of the correlation matrix of regressors as adjusted by ridge regression or IPC analysis. These observations are identified in the output data set by the values RIDGEVIF and IPCVIF for the variable \_TYPE\_.

\textbf{P}

calculates predicted values from the input data and the estimated model. The display includes the observation number, the ID variable (if one is specified), the actual and predicted values, and the residual. If the CLI, CLM, or R option is specified, the P option is unnecessary. See the section “Predicted and Residual Values” on page 8507 for more information.

\textbf{PARTIAL}

requests partial regression leverage plots for each regressor. You can use the \texttt{PARTIALDATA} option to obtain a tabular display of the partial regression leverage data. If ODS Graphics is enabled (see the section “ODS Graphics” on page 8557), then these partial plots are produced in panels with up to six plots per panel. See the section “Influence Statistics” on page 8516 for more information.
PARTIALDATA
requests partial regression leverage data for each regressor. You can request partial regression leverage plots of these data with the PARTIAL option. See the section “Influence Statistics” on page 8516 for more information.

PARTIALR2 < ( < TESTS > < SEQTESTS > ) >
See the SCORR1 option.

PC
outputs Amemiya’s prediction criterion for each model selected (Amemiya 1976; Judge et al. 1980) to the OUTEST= data set. If SELECTION=ADJRSQ, SELECTION=RSQUARE, or SELECTION=CP is specified, then the PC statistic is also added to the SubsetSelSummary table.

PCOMIT=list
requests an IPC analysis for each value \( m \) in the list. The procedure computes parameter estimates by using all but the last \( m \) principal components. Each value of \( m \) produces a set of IPC estimates, which is output to the OUTEST= data set. The values of \( m \) are saved by the variable _PCOMIT_, and the value of the variable _TYPE_ is set to IPC to identify the estimates. Only nonnegative integers can be specified with the PCOMIT= option.

If you specify the PCOMIT= option, RESTRICT statements are ignored. The PCOMIT= option is ignored if you use the SELECTION= option in the MODEL statement.

PCORR1
displays the squared partial correlation coefficients computed using Type I sum of squares (SS). This is calculated as \( SS/(SS+SSE) \), where SSE is the error sum of squares.

PCORR2
displays the squared partial correlation coefficients computed using Type II sums of squares. These are calculated the same way as with the PCORR1 option, except that Type II SS are used instead of Type I SS.

PRESS
outputs the PRESS statistic to the OUTEST= data set. The values of this statistic are saved in the variable _PRESS_. This option is available for all model-selection methods except RSQUARE, ADJRSQ, and CP.

R
requests an analysis of the residuals. The results include everything requested by the P option plus the standard errors of the mean predicted and residual values, the studentized residual, and Cook’s \( D \) statistic to measure the influence of each observation on the parameter estimates. See the section “Predicted and Residual Values” on page 8507 for more information.

RIDGE=list
requests a ridge regression analysis and specifies the values of the ridge constant \( k \) (see the section “Computations for Ridge Regression and IPC Analysis” on page 8536). Each value of \( k \) produces a set of ridge regression estimates that are placed in the OUTEST= data set. The values of \( k \) are saved by the variable _RIDGE_, and the value of the variable _TYPE_ is set to RIDGE to identify the estimates. Only nonnegative numbers can be specified with the RIDGE= option. Example 102.5 illustrates this option.
If you specify the RIDGE= option, RESTRICT statements are ignored. The RIDGE= option is ignored if you use the SELECTION= option in the MODEL statement.

**RMSE**

displays the root mean square error for each model selected.

**RSQUARE**

has the same effect as the EDF option.

**SBC**

outputs the SBC statistic for each model selected (Schwarz 1978; Judge et al. 1980) to the OUTEST= data set. If SELECTION=ADJRSQ, SELECTION=RSQUARE, or SELECTION=CP is specified, then the SBC statistic is also added to the SubsetSelSummary table.

**SCORR1 < ( < TESTS > < SEQTESTS > ) >**

displays the squared semipartial correlation coefficients computed using Type I sums of squares. This is calculated as SS/SST, where SST is the corrected total SS. If the NOINT option is used, the uncorrected total SS is used in the denominator. The optional arguments TESTS and SEQTESTS request are sequentially added to a model. The *F*-test values are computed as the Type I sum of squares for the variable in question divided by a mean square error. If you specify the TESTS option, the denominator MSE is the residual mean square for the full model specified in the MODEL statement. If you specify the SEQTESTS option, the denominator MSE is the residual mean square for the model containing all the independent variables that have been added to the model up to and including the variable in question. The TESTS and SEQTESTS options are not supported if you specify model selection methods or the RIDGE or PCOMIT options. Note that the PARTIALR2 option is a synonym for the SCORR1 option.

**SCORR2 < ( TESTS ) >**

displays the squared semipartial correlation coefficients computed using Type II sums of squares. These are calculated the same way as with the SCORR1 option, except that Type II SS are used instead of Type I SS. The optional TEST argument requests *F* tests and *p*-values as variables are sequentially added to a model. The *F*-test values are computed as the Type II sum of squares for the variable in question divided by the residual mean square for the full model specified in the MODEL statement. The TESTS option is not supported if you specify model selection methods or the RIDGE or PCOMIT options.

**SELECTION=name**

specifies the method used to select the model, where *name* can be FORWARD (or F), BACKWARD (or B), STEPWISE, MAXR, MINR, RSQUARE, ADJRSQ, CP, or NONE (use the full model). The default method is NONE. See the section “Model-Selection Methods” on page 8501 for a description of each method.

**SEQB**

produces a sequence of parameter estimates as each variable is entered into the model. This is displayed as a matrix where each row is a set of parameter estimates.

**SIGMA=n**

specifies the true standard deviation of the error term to be used in computing the CP and BIC statistics. If the SIGMA= option is not specified, an estimate from the full model is used. This option is available in the RSQUARE, ADJRSQ, and CP model-selection methods only.
SINGULAR=n
tunes the mechanism used to check for singularities. If you specify this option in the MODEL statement, it takes precedence over the SINGULAR= option in the PROC REG statement. The default value is machine dependent but is approximately 1E–9 on most machines. This option is rarely needed. Singularity checking is described in the section “Computational Methods” on page 8537.

SLE=value
SLENTRY=value
specifies the significance level for entry into the model used in the FORWARD and STEPWISE methods. The defaults are 0.50 for FORWARD and 0.15 for STEPWISE.

SLSTAY=value
SLS=value
specifies the significance level for staying in the model for the BACKWARD and STEPWISE methods. The defaults are 0.10 for BACKWARD and 0.15 for STEPWISE.

SP
outputs the $S_P$ statistic for each model selected (Hocking 1976) to the OUTEST= data set. If SELECTION=ADJRSQ, SELECTION=Rsquare, or SELECTION=CP is specified, then the SP statistic is also added to the SubsetSelSummary table.

SPE
performs a test that the first and second moments of the model are correctly specified. See the section “Testing for Heteroscedasticity” on page 8530 for more information.

SRT

displays a column in the “Output Statistics” table that shows the magnitude of studentized residuals by using a series of asterisks. This table is produced by the R option in the MODEL statement. When ODS Graphics is not enabled, the column is displayed in the table and the graph is not produced. When ODS Graphics is enabled, a graphical chart of the studentized residuals and Cook’s $D$ is produced, and the column is omitted from the table unless you specify the SRT option.

SS1
displays the sequential sums of squares (Type I SS) along with the parameter estimates for each term in the model. See Chapter 15, “The Four Types of Estimable Functions," for more information about the different types of sums of squares.

SS2
displays the partial sums of squares (Type II SS) along with the parameter estimates for each term in the model. See the SS1 option also.

SSE
computes the error sum of squares for each model selected.

START=s
specifies the number of regressors to be added or reported in a model.

When SELECTION= MAXR, MINR, or STEPWISE, START=s requests that the first s regressors be added to the model before beginning the comparing-and-switching process. Thus, START=s behaves like INCLUDE=s in that it forces selection to begin with the first s regressors included in the model,
but it differs in that START=s allows these variables to be subsequently removed from the model in the course of variable selection. For these methods, the default is START=0.

When SELECTION= RSQUARE, ADJRSQ, or CP, START=s specifies the smallest number of regressors to be reported in a subset model. For these methods, the default is START=1.

The START= option cannot be used with model-selection methods other than the six described here.

**STB** produces 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.

**STOP=s** causes PROC REG to stop when it has found the “best” s-variable model, where s is the STOP value. For the RSQUARE, ADJRSQ, and CP methods, STOP=s specifies the largest number of regressors to be reported in a subset model. For the MAXR and MINR methods, STOP=s specifies the largest number of regressors to be included in the model.

The default setting for the STOP= option is the number of variables in the MODEL statement. This option can be used only with the MAXR, MINR, RSQUARE, ADJRSQ, and CP methods.

**TOL** produces tolerance values for the estimates. Tolerance for a variable is defined as \(1 - R^2\), where \(R^2\) is obtained from the regression of the variable on all other regressors in the model. See the section “Collinearity Diagnostics” on page 8513 for more details.

**VIF** produces variance inflation factors with the parameter estimates. Variance inflation is the reciprocal of tolerance. See the section “Collinearity Diagnostics” on page 8513 for more detail.

**WHITE**

See the HCC option.

**XPX**

displays the \(X'X\) crossproducts matrix for the model. The crossproducts matrix is bordered by the \(X'Y\) and \(Y'Y\) matrices.

---

**MTEST Statement**

```
< label: > MTEST < equation < , . . . , equation > > < / options > ;
```

where each equation is a linear function composed of coefficients and variable names. The label is optional.

The MTEST statement is used to test hypotheses in multivariate regression models where there are several dependent variables fit to the same regressors. If no equations or options are specified, the MTEST statement tests the hypothesis that all estimated parameters except the intercept are zero.

The hypotheses that can be tested with the MTEST statement are of the form

\[(L\beta - c)M = 0\]
where $\mathbf{L}$ is a linear function on the regressor side, $\mathbf{\beta}$ is a matrix of parameters, $\mathbf{e}$ is a column vector of constants, $\mathbf{j}$ is a row vector of ones, and $\mathbf{M}$ is a linear function on the dependent side. The special case where the constants are zero is

$$\mathbf{L}\mathbf{\beta}\mathbf{M} = 0$$

See the section “Multivariate Tests” on page 8532 for more details.

Each linear function extends across either the regressor variables or the dependent variables. If the equation is across the dependent variables, then the constant term, if specified, must be zero. The equations for the regressor variables form the $\mathbf{L}$ matrix and $\mathbf{c}$ vector in the preceding formula; the equations for dependent variables form the $\mathbf{M}$ matrix. If no equations for the dependent variables are given, PROC REG uses an identity matrix for $\mathbf{M}$, testing the same hypothesis across all dependent variables. If no equations for the regressor variables are given, PROC REG forms a linear function corresponding to a test that all the nonintercept parameters are zero.

As an example, consider the following statements:

```plaintext
model y1 y2 y3=x1 x2 x3;
mtest x1,x2;
mtest y1-y2, y2-y3, x1;
mtest y1-y2;
```

The first MTEST statement tests the hypothesis that the $X_1$ and $X_2$ parameters are zero for $Y_1$, $Y_2$, and $Y_3$. In addition, the second MTEST statement tests the hypothesis that the $X_1$ parameter is the same for all three dependent variables. For the same model, the third MTEST statement tests the hypothesis that all parameters except the intercept are the same for dependent variables $Y_1$ and $Y_2$.

You can specify the following options in the MTEST statement:

**CANPRINT**

displays the canonical correlations for the hypothesis combinations and the dependent variable combinations. If you specify

```plaintext
mtest / canprint;
```

the canonical correlations between the regressors and the dependent variables are displayed.

**DETAILS**

displays the $\mathbf{M}$ matrix and various intermediate calculations.

**MSTAT=FAPPROX | EXACT**

specifies the method of evaluating the multivariate test statistics. The default is MSTAT=FAPPROX, which specifies that the multivariate tests are evaluated by using the usual approximations based on the $F$ distribution, as discussed in the “Multivariate Tests” section in Chapter 4, “Introduction to Regression Procedures.” Alternatively, you can specify MSTAT=EXACT to compute exact $p$-values for three of the four tests (Wilks’ lambda, the Hotelling-Lawley trace, and Roy’s greatest root) and an improved $F$ approximation for the fourth (Pillai’s trace). While MSTAT=EXACT provides better control of the significance probability for the tests, especially for Roy’s greatest root, computations for the exact $p$-values can be appreciably more demanding, and are in fact infeasible for large problems (many dependent variables). Thus, although MSTAT=EXACT is more accurate for most data, it is not the default method.
PRINT displays the H and E matrices.

**OUTPUT Statement**

```
OUTPUT < OUT=SAS-data-set> < keyword=names > < . . . keyword=names > ;
```

The OUTPUT statement creates a new SAS data set that saves diagnostic measures calculated after fitting the model. The OUTPUT statement refers to the most recent MODEL statement. At least one `keyword=names` specification is required.

All the variables in the original data set are included in the new data set, along with variables created in the OUTPUT statement. These new variables contain the values of a variety of statistics and diagnostic measures that are calculated for each observation in the data set. If you want to create a SAS data set in a permanent library, you must specify a two-level name. For more information about permanent libraries and SAS data sets, see SAS Language Reference: Concepts.

The OUTPUT statement cannot be used when a TYPE=CORR, TYPE=COV, or TYPE=SSCP data set is used as the input data set for PROC REG. See the section “Input Data Sets” on page 8489 for more details.

The statistics created in the OUTPUT statement are described in this section. More details are given in the section “Predicted and Residual Values” on page 8507 and the section “Influence Statistics” on page 8516. Also see Chapter 4, “Introduction to Regression Procedures,” for definitions of the statistics available from the REG procedure.

You can specify the following options in the OUTPUT statement:

- **OUT=SAS data set**
  gives the name of the new data set. By default, the procedure uses the DATA\* convention to name the new data set.

- **keyword=names**
  specifies the statistics to include in the output data set and names the new variables that contain the statistics. Specify a `keyword` for each desired statistic (see the following list of `keywords`), an equal sign, and the variable or variables to contain the statistic.

  In the output data set, the first variable listed after a `keyword` in the OUTPUT statement contains that statistic for the first dependent variable listed in the MODEL statement; the second variable contains the statistic for the second dependent variable in the MODEL statement, and so on. The list of variables following the equal sign can be shorter than the list of dependent variables in the MODEL statement. In this case, the procedure creates the new names in order of the dependent variables in the MODEL statement.

  For example, the following SAS statements create an output data set named b:

  ```
  proc reg data=a;
    model y z=x1 x2;
    output out=b
      p=yhat zhat
      r=yresid zresid;
  run;
  ```
In addition to the variables in the input data set, `b` contains the following variables:

- `yhat`, with values that are predicted values of the dependent variable `y`
- `zhat`, with values that are predicted values of the dependent variable `z`
- `yresid`, with values that are the residual values of `y`
- `zresid`, with values that are the residual values of `z`

You can specify the following *keywords* in the *OUTPUT statement.* See the section “Model Fit and Diagnostic Statistics” on page 8514 for computational formulas.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>COOKD=names</code></td>
<td>Cook’s <em>D</em> influence statistic</td>
</tr>
<tr>
<td><code>COVRATIO=names</code></td>
<td>standard influence of observation on covariance of betas, as discussed in the section “Influence Statistics” on page 8516</td>
</tr>
<tr>
<td><code>DFFITS=names</code></td>
<td>standard influence of observation on predicted value</td>
</tr>
<tr>
<td><code>H=names</code></td>
<td>leverage, ( x_i (X'X)^{-1} x_i' )</td>
</tr>
<tr>
<td><code>LCL=names</code></td>
<td>lower bound of a ( 100(1 - \alpha) % ) confidence interval for an individual prediction. This includes the variance of the error, as well as the variance of the parameter estimates.</td>
</tr>
<tr>
<td><code>LCLM=names</code></td>
<td>lower bound of a ( 100(1 - \alpha) % ) confidence interval for the expected value (mean) of the dependent variable</td>
</tr>
<tr>
<td>`PREDICTED</td>
<td>P=names`</td>
</tr>
<tr>
<td><code>PRESS=names</code></td>
<td>( i )th residual divided by ( 1 - h ), where ( h ) is the leverage, and where the model has been refit without the ( i )th observation</td>
</tr>
<tr>
<td>`RESIDUAL</td>
<td>R=names`</td>
</tr>
<tr>
<td><code>RSTUDENT=names</code></td>
<td>a studentized residual with the current observation deleted</td>
</tr>
<tr>
<td><code>STDI=names</code></td>
<td>standard error of the individual predicted value</td>
</tr>
<tr>
<td><code>STDP=names</code></td>
<td>standard error of the mean predicted value</td>
</tr>
<tr>
<td><code>STDR=names</code></td>
<td>standard error of the residual</td>
</tr>
<tr>
<td><code>STUDENT=names</code></td>
<td>studentized residuals, which are the residuals divided by their standard errors</td>
</tr>
<tr>
<td><code>UCL=names</code></td>
<td>upper bound of a ( 100(1 - \alpha) % ) confidence interval for an individual prediction</td>
</tr>
<tr>
<td><code>UCLM=names</code></td>
<td>upper bound of a ( 100(1 - \alpha) % ) confidence interval for the expected value (mean) of the dependent variable</td>
</tr>
</tbody>
</table>

**PRINT Statement**

```
PRINT < options > < ANOVA > < MODELDATA > ;
```

The PRINT statement enables you to interactively display the results of MODEL statement options, produce an ANOVA table, display the data for variables used in the current model, or redisplay the options specified
in a `MODEL` or a previous `PRINT` statement. In addition, like most other interactive statements in PROC REG, the `PRINT` statement implicitly refits the model; thus, effects of `REWEIGHT` statements are seen in the resulting tables. If ODS Graphics is enabled (see the section “ODS Graphics” on page 8557), the `PRINT` statement also requests the use of the ODS graphical displays associated with the current model.

The following specifications can appear in the `PRINT` statement:

```plaintext
options
  interactively displays the results of `MODEL` statement options, where `options` is one or more of the following: ACOV, ALL, CLI, CLM, COLLIN, COLLINOINT, CORRB, COVB, DW, I, INFLUENCE, P, PARTIAL, PCORR1, PCORR2, R, SCORR1, SCORR2, SEQB, SPEC, SS1, SS2, STB, TOL, VIF, or XPX. See the section “`MODEL` Statement” on page 8466 for a description of these `options`.

ANOVA
  produces the ANOVA table associated with the current model. This is either the model specified in the last `MODEL` statement or the model that incorporates changes made by ADD, DELETE, or `REWEIGHT` statements after the last `MODEL` statement.

MODELDATA
  displays the data for variables used in the current model.
```

Use the statement

```plaintext
print;
```

to reprint options in the most recently specified `PRINT` or `MODEL` statement.

Options that require original data values, such as R or INFLUENCE, cannot be used when a TYPE=CORR, TYPE=COV, or TYPE=SSCP data set is used as the input data set to PROC REG. See the section “Input Data Sets” on page 8489 for more detail.

---

**REFIT Statement**

```plaintext
REFIT;
```

The `REFIT` statement causes the current model and corresponding statistics to be recomputed immediately. No output is generated by this statement. The `REFIT` statement is needed after one or more `REWEIGHT` statements to cause them to take effect before subsequent `PAINT` or `REWEIGHT` statements. This is sometimes necessary when you are using statistical conditions in `REWEIGHT` statements. For example, consider the following statements:

```plaintext
paint student.>2;
plot student.*p.;
reweight student.>2;
refit;
paint student.>2;
plot student.*p.;
```

The second `PAINT` statement paints any additional observations that meet the condition after deleting observations and refitting the model. The `REFIT` statement is used because the `REWEIGHT` statement does not cause the model to be recomputed. In this particular example, the same effect could be achieved by replacing the `REFIT` statement with a `PLOT` statement.
Most interactive statements can be used to implicitly refit the model; any plots or statistics produced by these statements reflect changes made to the model and changes made to the data used to compute the model. The two exceptions are the PAINT and REWEIGHT statements, which do not cause the model to be recomputed.

### RESTRICT Statement

**RESTRICT**  

RESTRICT  

```plaintext
equation <, . . . , equation>
```

A RESTRICT statement is used to place restrictions on the parameter estimates in the MODEL preceding it. More than one RESTRICT statement can follow each MODEL statement. Each RESTRICT statement replaces any previous RESTRICT statement. To lift all restrictions on a model, submit a new MODEL statement. If there are several restrictions, separate them with commas. The statement

```plaintext
restrict equation1=equation2=equation3;
```

is equivalent to imposing the two restrictions

```plaintext
restrict equation1=equation2;
restrict equation2=equation3;
```

Each restriction is written as a linear equation and can be written as

```plaintext
equation
```

or

```plaintext
equation = equation
```

The form of each `equation` is

```plaintext
c_1 \times variable_1 \pm c_2 \times variable_2 \pm \cdots \pm c_n \times variable_n
```

where the `c_j`'s are constants and the `variable_j`'s are any regressor variables.

When no equal sign appears, the linear combination is set equal to zero. Each variable name mentioned must be a variable in the MODEL statement to which the RESTRICT statement refers. The keyword INTERCEPT can also be used as a variable name, and it refers to the intercept parameter in the regression model.

Note that the parameters associated with the variables are restricted, not the variables themselves. Restrictions should be consistent and not redundant.

Examples of valid RESTRICT statements include the following:

```plaintext
restrict x1;
restrict a+b=1;
restrict a=b=c;
restrict a=b, b=c;
restrict 2*f=g+h, intercept+f=0;
restrict f=g=h=intercept;
```

The third and fourth statements in this list produce identical restrictions. You cannot specify
restrict f-g=0,
f-intercept=0,
g-intercept=1;

because the three restrictions are not consistent. If these restrictions are included in a RESTRICT statement, one of the restrict parameters is set to zero and has zero degrees of freedom, indicating that PROC REG is unable to apply a restriction.

The restrictions usually operate even if the model is not of full rank. Check to ensure that DF = –1 for each restriction. In addition, the model DF should decrease by 1 for each restriction.

The parameter estimates are those that minimize the quadratic criterion (SSE) subject to the restrictions. If a restriction cannot be applied, its parameter value and degrees of freedom are listed as zero.

The method used for restricting the parameter estimates is to introduce a Lagrangian parameter for each restriction (Pringle and Rayner 1971). The estimates of these parameters are displayed with test statistics. Note that the \( t \) statistic reported for the Lagrangian parameters does not follow a Student’s \( t \) distribution, but its square follows a beta distribution (LaMotte 1994). The \( p \)-value for these parameters is computed using the beta distribution.

The Lagrangian parameter \( \gamma \) measures the sensitivity of the SSE to the restriction constant. If the restriction constant is changed by a small amount \( \epsilon \), the SSE is changed by \( 2\gamma \epsilon \). The \( t \) ratio tests the significance of the restrictions. If \( \gamma \) is zero, the restricted estimates are the same as the unrestricted estimates, and a change in the restriction constant in either direction increases the SSE.

RESTRICT statements are ignored if the PCOMIT= or RIDGE= option is specified in the PROC REG statement.

---

**REWEIGHT Statement**

```
REWEIGHT < condition | ALLOBS > < / options > ;
REWEIGHT < STATUS | UNDO > ;
```

The REWEIGHT statement interactively changes the weights of observations that are used in computing the regression equation. The REWEIGHT statement can change observation weights, or set them to zero, which causes selected observations to be excluded from the analysis. When a REWEIGHT statement sets observation weights to zero, the observations are not deleted from the data set. More than one REWEIGHT statement can be used. The requests from all REWEIGHT statements are applied to the subsequent statements. Each use of the REWEIGHT statement modifies the MODEL label.

The model and corresponding statistics are not recomputed after a REWEIGHT statement. For example, consider the following statements:

```
reweight r.>0;
reweight r.>0;
```

The second REWEIGHT statement does not exclude any additional observations since the model is not recomputed after the first REWEIGHT statement. Either use a REFIT statement to explicitly refit the model, or implicitly refit the model by following the REWEIGHT statement with any other interactive statement except a PAINT statement or another REWEIGHT statement.
The REWEIGHT statement cannot be used if a TYPE=CORR, TYPE=COV, or TYPE=SSCP data set is used as an input data set to PROC REG. Note that the syntax used in the REWEIGHT statement is the same as the syntax in the PAINT statement.

The syntax of the REWEIGHT statement is described in the following sections.

For detailed examples of using this statement, see the section “Reweighting Observations in an Analysis” on page 8524.

**Specifying Condition**

*Condition* is used to find observations to be reweighted. The syntax of *condition* is

\[
\text{variable compare value}
\]

or

\[
\text{variable compare value logical variable compare value}
\]

where

- **variable** is one of the following:
  - a variable name in the input data set
  - OBS., which is the observation number
  - *keyword*, where *keyword* is a keyword for a statistic requested in the OUTPUT statement. The *keyword* specification is applied to all dependent variables in the model.

- **compare** is an operator that compares **variable** to **value**. *Compare* can be any one of the following: <, <=, >, >=, =, ≠. The operators LT, LE, GT, GE, EQ, and NE, respectively, can be used instead of the preceding symbols. See the “Expressions” chapter in *SAS Language Reference: Concepts* for more information about comparison operators.

- **value** gives an unformatted value of **variable**. Observations are selected to be reweighted if they satisfy the condition created by **variable compare value**. **Value** can be a number or a character string. If **value** is a character string, it must be eight characters or less and must be enclosed in quotes. In addition, **value** is case-sensitive. In other words, the following two statements are not the same:

\[
\text{reweight name='steve';}
\]

\[
\text{reweight name='Steve';}
\]

- **logical** is one of two logical operators. Either AND or OR can be used. To specify AND, use AND or the symbol &. To specify OR, use OR or the symbol |.
Here are some examples of the variable compare value form:

```plaintext
reweight obs. le 10;
reweight temp=55;
reweight type='new';
```

Here are some examples of the variable compare value logical variable compare value form:

```plaintext
reweight obs.<=10 and residual.<2;
reweight student.<-2 or student.>2;
reweight name='Mary' | name='Susan';
```

Using ALLOBS

Instead of specifying condition, you can use the ALLOBS option to select all observations. This is most useful when you want to restore the original weights of all observations. For example,

```plaintext
reweight allob / reset;
```

resets weights for all observations and uses all observations in the subsequent analysis. Note that

```plaintext
reweight allob;
```

specifies that all observations be excluded from analysis. Consequently, using ALLOBS is useful only if you also use one of the options discussed in the following section.

Options in the REWEIGHT Statement

The following options can be used when either a condition, ALLOBS, or nothing is specified before the slash. If only an option is listed, the option applies to the observations selected in the previous REWEIGHT statement, not to the observations selected by reapplying the condition from the previous REWEIGHT statement. For example, consider the following statements:

```plaintext
reweight r.>0 / weight=0.1;
refit;
reweight;
```

The second REWEIGHT statement excludes from the analysis only those observations selected in the first REWEIGHT statement. No additional observations are excluded even if there are new observations that meet the condition in the first REWEIGHT statement.

**NOTE:** Options are not available when either the UNDO or STATUS option is used.

**NOLIST**

suppresses the display of the selected observation numbers. If you omit the NOLIST option, a list of observations selected is written to the log.

**RESET**

resets the observation weights to their original values as defined by the WEIGHT statement or to WEIGHT=1 if no WEIGHT statement is specified. For example,

```plaintext
reweight / reset;
```
Chapter 102: The REG Procedure

resets observation weights to the original weights in the data set. If previous REWEIGHT statements have been submitted, this REWEIGHT statement applies only to the observations selected by the previous REWEIGHT statement. Note that, although the RESET option does reset observation weights to their original values, it does not cause the model and corresponding statistics to be recomputed.

**WEIGHT=**value
changes observation weights to the specified nonnegative real number. If you omit the WEIGHT= option, the observation weights are set to zero, and observations are excluded from the analysis. For example:

```
reweight name='Alan';
...other interactive statements
reweight / weight=0.5;
```

The first REWEIGHT statement changes weights to zero for all observations with `name='Alan'`, effectively deleting these observations. The subsequent analysis does not include these observations. The second REWEIGHT statement applies only to those observations selected by the previous REWEIGHT statement, and it changes the weights to 0.5 for all the observations with `NAME='Alan'`. Thus, the next analysis includes all original observations; however, those observations with `NAME='Alan'` have their weights set to 0.5.

**STATUS and UNDO**

If you omit `condition` and the ALLOBS options, you can specify one of the following options.

**STATUS**
writes to the log the observation’s number and the weight of all reweighted observations. If an observation’s weight has been set to zero, it is reported as deleted. However, the observation is not deleted from the data set, only from the analysis.

**UNDO**
undoes the changes made by the most recent REWEIGHT statement. Weights might be, but are not necessarily, reset. For example, consider the following statements:

```
reweight student.>2 / weight=0.1;
reweight;
reweight undo;
```

The first REWEIGHT statement sets the weights of observations that satisfy the condition to 0.1. The second REWEIGHT statement sets the weights of the same observations to zero. The third REWEIGHT statement undoes the second, changing the weights back to 0.1.

**STORE Statement**

```
STORE < OUT=>item-store-name </ LABEL='label'> ;
```
The STORE statement requests that the procedure save the estimated parameters of the fitted model. The resulting item store is a binary file format that cannot be modified. The contents of the item store can be processed with the PLM procedure.

For details about the syntax of the STORE statement, see the section “STORE Statement” on page 520 in Chapter 19, “Shared Concepts and Topics.”

**NOTE:** The information stored by the STORE statement in PROC REG is a subset of what is usually stored by other procedures that implement this statement. In particular, PROC REG stores only the estimated parameters of the model, so that you can later use the CODE statement in PROC PLM to write SAS DATA step code for prediction to a file or catalog entry. With only this subset of information, many other postprocessing features of PROC PLM are not available for item stores that are created by PROC REG.

---

**TEST Statement**

\[
<\text{label}:> \text{TEST equation, } , \ldots, \text{equation} </\text{option}> ;
\]

The TEST statement tests hypotheses about the parameters estimated in the preceding MODEL statement. It has the same syntax as the RESTRICT statement except that it supports an **option**. Each **equation** specifies a linear hypothesis to be tested. The rows of the hypothesis are separated by commas.

Variable names must correspond to regressors, and each variable name represents the coefficient of the corresponding variable in the model. An optional label is useful to identify each test with a name. The keyword INTERCEPT can be used instead of a variable name to refer to the model’s intercept.

The REG procedure performs an \( F \) test for the joint hypotheses specified in a single TEST statement. More than one TEST statement can accompany a MODEL statement. The numerator is the usual quadratic form of the estimates; the denominator is the mean squared error. If hypotheses can be represented by

\[
L\beta = c
\]

then the numerator of the \( F \) test is

\[
Q = (Lb - c)'(L(X'X)^{-1}L')^{-1}(Lb - c)
\]

divided by degrees of freedom, where \( b \) is the estimate of \( \beta \). For example:

```plaintext
model y=a1 a2 b1 b2;
aplus: test a1+a2=1;
b1: test b1=0, b2=0;
b2: test b1, b2;
```

The last two statements are equivalent; since no constant is specified, zero is assumed.

Note that, when the ACOV, HCC, or WHITE option is specified in the MODEL statement, tests are recomputed using the heteroscedasticity-consistent covariance matrix specified with the HCCMETHOD= option in the MODEL statement (see the section “Testing for Heteroscedasticity” on page 8530).
One option can be specified in the TEST statement after a slash (/):

PRINT
  displays intermediate calculations. This includes $L(X'X)^{-1}L'$ bordered by $Lb - c$, and $(L(X'X)^{-1}L')^{-1}$ bordered by $(L(X'X)^{-1}L')^{-1}(Lb - c)$.

VAR Statement

VAR variables ;

The VAR statement is used to include numeric variables in the crossproducts matrix that are not specified in the first MODEL statement.

Variables not listed in MODEL statements before the first RUN statement must be listed in the VAR statement if you want the ability to add them interactively to the model with an ADD statement, to include them in a new MODEL statement, or to plot them in a scatter plot with the PLOT statement.

In addition, if you want to use options in the PROC REG statement and do not want to fit a model to the data (with a MODEL statement), you must use a VAR statement.

WEIGHT Statement

WEIGHT variable ;

A WEIGHT statement names a variable in the input data set with values that are relative weights for a weighted least squares fit. If the weight value is proportional to the reciprocal of the variance for each observation, then the weighted estimates are the best linear unbiased estimates (BLUE).

Values of the weight variable must be nonnegative. If an observation’s weight is zero, the observation is deleted from the analysis. If a weight is negative or missing, it is set to zero, and the observation is excluded from the analysis. A more complete description of the WEIGHT statement can be found in Chapter 50, “The GLM Procedure.”

Observation weights can be changed interactively with the REWEIGHT statement.

Details: REG Procedure

Missing Values

PROC REG constructs only one crossproducts matrix for the variables in all regressions. If any variable needed for any regression is missing, the observation is excluded from all estimates. If you include variables with missing values in the VAR statement, the corresponding observations are excluded from all analyses, even if you never include the variables in a model. PROC REG assumes that you might want to include these variables after the first RUN statement and deletes observations with missing values.
Input Data Sets

PROC REG does not compute new regressors. For example, if you want a quadratic term in your model, you should create a new variable when you prepare the input data. For example, the statement

\[
\text{model } y=x_1 x_1\times x_1;
\]

is not valid. Note that this MODEL statement is valid in the GLM procedure.

The input data set for most applications of PROC REG contains standard rectangular data, but special TYPE=CORR, TYPE=COV, and TYPE=SSCP data sets can also be used. TYPE=CORR and TYPE=COV data sets created by the CORR procedure contain means and standard deviations. In addition, TYPE=CORR data sets contain correlations and TYPE=COV data sets contain covariances. TYPE=SSCP data sets created in previous runs of PROC REG that used the OUTSSCP= option contain the sums of squares and crossproducts of the variables.


These summary files save CPU time. It takes \( nk^2 \) operations (where \( n \) = number of observations and \( k \) = number of variables) to calculate crossproducts; the regressions are of the order \( k^3 \). When \( n \) is in the thousands and \( k \) is less than 10, you can save 99% of the CPU time by reusing the SSCP matrix rather than recomputing it.

When you want to use a special SAS data set as input, PROC REG must determine the TYPE for the data set. PROC CORR and PROC REG automatically set the type for their output data sets. However, if you create the data set by some other means (such as a DATA step), you must specify its type with the TYPE= data set option. If the TYPE for the data set is not specified when the data set is created, you can specify TYPE= as a data set option in the DATA= option in the PROC REG statement. For example:

\[
\text{proc reg data=a(type=corr);}
\]

When a TYPE=CORR, TYPE=COV, or TYPE=SSCP data set is used with PROC REG, statements and options that require the original data values have no effect. The OUTPUT, PAINT, PLOT, and REWEIGHT statements and the MODEL and PRINT statement options P, R, CLM, CLI, DW, INFLUENCE, and PARTIAL are disabled since the original observations needed to calculate predicted and residual values are not present.

Example Using TYPE=CORR Data Set

The following statements use PROC CORR to produce an input data set for PROC REG. The fitness data for this analysis can be found in Example 102.2.

\[
\text{proc corr data=fitness outp=r noprint;}
\]
\[
\text{var Oxygen RunTime Age Weight RunPulse MaxPulse RestPulse;}
\]
\[
\text{run;}
\]
\[
\text{proc print data=r;}
\]
\[
\text{run;}
\]
\[
\text{proc reg data=r;}
\]
\[
\text{model Oxygen=RunTime Age Weight;}
\]
\[
\text{run;}
\]

Since the OUTP= data set from PROC CORR is automatically set to TYPE=CORR, the TYPE= data set option is not required in this example. The data set containing the correlation matrix is displayed by the
PRINT procedure as shown in Figure 102.15. Figure 102.16 shows results from the regression that uses the TYPE=CORR data as an input data set.

**Figure 102.15** TYPE=CORR Data Set Created by PROC CORR

<table>
<thead>
<tr>
<th>Obs</th>
<th>TYPE</th>
<th>NAME</th>
<th>Oxygen</th>
<th>RunTime</th>
<th>Age</th>
<th>Weight</th>
<th>RunPulse</th>
<th>MaxPulse</th>
<th>RestPulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MEAN</td>
<td></td>
<td>47.3758</td>
<td>10.5861</td>
<td>47.6774</td>
<td>77.4445</td>
<td>169.645</td>
<td>173.774</td>
<td>53.4516</td>
</tr>
<tr>
<td>2</td>
<td>STD</td>
<td></td>
<td>5.3272</td>
<td>1.3874</td>
<td>5.2114</td>
<td>8.3286</td>
<td>10.252</td>
<td>9.164</td>
<td>7.6194</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td></td>
<td>31.0000</td>
<td>31.0000</td>
<td>31.0000</td>
<td>31.0000</td>
<td>31.000</td>
<td>31.000</td>
<td>31.0000</td>
</tr>
<tr>
<td>4</td>
<td>CORR</td>
<td>Oxygen</td>
<td>1.0000</td>
<td></td>
<td>-0.8622</td>
<td>-0.3046</td>
<td>-0.1628</td>
<td>-0.398</td>
<td>-0.237</td>
</tr>
<tr>
<td>5</td>
<td>CORR</td>
<td>RunTime</td>
<td>-0.8622</td>
<td>1.0000</td>
<td>0.1887</td>
<td>0.1435</td>
<td>0.314</td>
<td>0.226</td>
<td>0.4504</td>
</tr>
<tr>
<td>6</td>
<td>CORR</td>
<td>Age</td>
<td>-0.3046</td>
<td>0.1887</td>
<td>1.0000</td>
<td>-0.2335</td>
<td>-0.338</td>
<td>-0.433</td>
<td>-0.1641</td>
</tr>
<tr>
<td>7</td>
<td>CORR</td>
<td>Weight</td>
<td>-0.1628</td>
<td>0.1435</td>
<td>-0.2335</td>
<td>1.0000</td>
<td>0.182</td>
<td>0.249</td>
<td>0.0440</td>
</tr>
<tr>
<td>8</td>
<td>CORR</td>
<td>RunPulse</td>
<td>-0.3980</td>
<td>0.3136</td>
<td>-0.3379</td>
<td>0.1815</td>
<td>1.000</td>
<td>0.930</td>
<td>0.3525</td>
</tr>
<tr>
<td>9</td>
<td>CORR</td>
<td>MaxPulse</td>
<td>-0.2367</td>
<td>0.2261</td>
<td>-0.4329</td>
<td>0.2494</td>
<td>0.930</td>
<td>1.000</td>
<td>0.3051</td>
</tr>
<tr>
<td>10</td>
<td>CORR</td>
<td>RestPulse</td>
<td>-0.3994</td>
<td>0.4504</td>
<td>-0.1641</td>
<td>0.0440</td>
<td>0.352</td>
<td>0.305</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

**Figure 102.16** Regression on TYPE=CORR Data Set

The following example uses the saved crossproducts matrix:

```plaintext
proc reg data=fitness outsscp=sscp noprint;
   model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse;
run;
proc print data=sscp;
run;
proc reg data=sscp;
   model Oxygen=RunTime Age Weight;
run;
```
First, all variables are used to fit the data and create the SSCP data set. Figure 102.17 shows the PROC PRINT display of the SSCP data set. The SSCP data set is then used as the input data set for PROC REG, and a reduced model is fit to the data.

**Figure 102.17** TYPE=SSCP Data Set Created by PROC REG

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>TYPE</em></th>
<th><em>NAME</em></th>
<th>Intercept</th>
<th>RunTime</th>
<th>Age</th>
<th>Weight</th>
<th>RunPulse</th>
<th>MaxPulse</th>
<th>RestPulse</th>
<th>Oxygen</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SSCP</td>
<td>Intercept</td>
<td>31.00</td>
<td>328.17</td>
<td>1478.00</td>
<td>2400.78</td>
<td>5259.00</td>
<td>5387.00</td>
<td>1657.00</td>
<td>1468.65</td>
</tr>
<tr>
<td>2</td>
<td>SSCP</td>
<td>RunTime</td>
<td>328.17</td>
<td>3531.80</td>
<td>15687.24</td>
<td>25464.71</td>
<td>55806.29</td>
<td>57113.72</td>
<td>17684.05</td>
<td>15356.14</td>
</tr>
<tr>
<td>3</td>
<td>SSCP</td>
<td>Age</td>
<td>1478.00</td>
<td>71282.00</td>
<td>114158.90</td>
<td>250194.00</td>
<td>256218.00</td>
<td>78806.00</td>
<td>69767.75</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>SSCP</td>
<td>Weight</td>
<td>2400.78</td>
<td>114158.90</td>
<td>188008.20</td>
<td>407745.67</td>
<td>417764.62</td>
<td>128409.28</td>
<td>113522.26</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>SSCP</td>
<td>RunPulse</td>
<td>5259.00</td>
<td>55806.29</td>
<td>250194.00</td>
<td>407745.67</td>
<td>895317.00</td>
<td>916499.00</td>
<td>281928.00</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>SSCP</td>
<td>MaxPulse</td>
<td>5387.00</td>
<td>57113.72</td>
<td>256218.00</td>
<td>417764.62</td>
<td>938641.00</td>
<td>288583.00</td>
<td>248497.31</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>SSCP</td>
<td>RestPulse</td>
<td>1657.00</td>
<td>78806.00</td>
<td>128409.28</td>
<td>281928.00</td>
<td>288583.00</td>
<td>90311.00</td>
<td>78015.41</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>SSCP</td>
<td>Oxygen</td>
<td>1468.65</td>
<td>69767.75</td>
<td>113522.26</td>
<td>248497.31</td>
<td>254866.75</td>
<td>78015.41</td>
<td>70429.86</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td></td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
</tr>
</tbody>
</table>

Figure 102.18 also shows the PROC REG results for the reduced model. (For the PROC REG results for the full model, see Figure 102.30.)

**Figure 102.18** Regression on TYPE=SSCP Data Set

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>656.27095</td>
<td>218.75698</td>
<td>30.27</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>195.11060</td>
<td>7.22632</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.68818   R-Square 0.7708
Dependent Mean 47.37581   Adj R-Sq 0.7454
Coeff Var 5.67416

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>93.12615</td>
<td>7.55916</td>
<td>12.32</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RunTime</td>
<td>1</td>
<td>-3.14039</td>
<td>0.36738</td>
<td>-8.55</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>-0.17388</td>
<td>0.09955</td>
<td>-1.75</td>
<td>0.0921</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>1</td>
<td>-0.05444</td>
<td>0.06181</td>
<td>-0.88</td>
<td>0.3862</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the preceding example, the TYPE= data set option is not required since PROC REG sets the OUTSSCP= data set to TYPE=SSCP.
Output Data Sets

OUTEST= Data Set

The OUTEST= specification produces a TYPE=EST output SAS data set containing estimates and optional statistics from the regression models. For each BY group on each dependent variable occurring in each MODEL statement, PROC REG outputs an observation to the OUTEST= data set. The variables output to the data set are as follows:

- the BY variables, if any
- _MODEL_, a character variable containing the label of the corresponding MODEL statement, or MODELn if no label is specified, where n is 1 for the first MODEL statement, 2 for the second model statement, and so on
- _TYPE_, a character variable with the value 'PARMS' for every observation
- _DEPVAR_, the name of the dependent variable
- _RMSE_, the root mean squared error or the estimate of the standard deviation of the error term
- Intercept, the estimated intercept, unless the NOINT option is specified
- all the variables listed in any MODEL or VAR statement. Values of these variables are the estimated regression coefficients for the model. A variable that does not appear in the model corresponding to a given observation has a missing value in that observation. The dependent variable in each model is given a value of –1.

If you specify the COVOUT option, the covariance matrix of the estimates is output after the estimates; the _TYPE_ variable is set to the value 'COV' and the names of the rows are identified by the character variable, _NAME_.

If you specify the TABLEOUT option, the following statistics listed by _TYPE_ are added after the estimates:

- STDERR, the standard error of the estimate
- T, the t statistic for testing if the estimate is zero
- PVALUE, the associated p-value
- LNb, the 100(1 − α) lower confidence limit for the estimate, where n is the nearest integer to 100(1 − α) and α defaults to 0.05 or is set by using the ALPHA= option in the PROC REG or MODEL statement
- UNb, the 100(1 − α) upper confidence limit for the estimate
Specifying the option ADJRSQ, AIC, BIC, CP, EDF, GMSEP, JP, MSE, PC, RSQUARE, SBC, SP, or SSE in the PROC REG or MODEL statement automatically outputs these statistics and the model $R^2$ for each model selected, regardless of the model selection method. Additional variables, in order of occurrence, are as follows:

- `_IN_`, the number of regressors in the model not including the intercept
- `_P_`, the number of parameters in the model including the intercept, if any
- `_EDF_`, the error degrees of freedom
- `_SSE_`, the error sum of squares, if the SSE option is specified
- `_MSE_`, the mean squared error, if the MSE option is specified
- `_RSQ_`, the $R^2$ statistic
- `_ADJRSQ_`, the adjusted $R^2$ statistic, if the ADJRSQ option is specified
- `_CP_`, the $C_p$ statistic, if the CP option is specified
- `_SP_`, the $S_p$ statistic, if the SP option is specified
- `_JP_`, the $J_p$ statistic, if the JP option is specified
- `_PC_`, the PC statistic, if the PC option is specified
- `_GMSEP_`, the GMSEP statistic, if the GMSEP option is specified
- `_AIC_`, the AIC statistic, if the AIC option is specified
- `_BIC_`, the BIC statistic, if the BIC option is specified
- `_SBC_`, the SBC statistic, if the SBC option is specified

The following statements produce and display the OUTEST= data set. This example uses the population data given in the section “Polynomial Regression” on page 8436. Figure 102.19 through Figure 102.21 show the regression equations and the resulting OUTEST= data set.

```plaintext
proc reg data=USPopulation outest=est;
   m1: model Population=Year;
   m2: model Population=Year YearSq;
run;
proc print data=est;
run;
```
**Figure 102.19** Regression Output for Model M1

The REG Procedure  
Model: m1  
Dependent Variable: Population

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>146869</td>
<td>146869</td>
<td>228.92</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>20</td>
<td>12832</td>
<td>641.58160</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>21</td>
<td>159700</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 25.32946  
R-Square 0.9197  
Dependent Mean 94.64800  
Adj R-Sq 0.9156  
Coeff Var 26.76175

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|
| Intercept| 1  | -2345.85498        | 161.39279      | -14.54  | <.0001|
| Year     | 1  | 1.28786            | 0.08512        | 15.13   | <.0001|

**Figure 102.20** Regression Output for Model M2

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>159529</td>
<td>79765</td>
<td>8864.19</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>170.97193</td>
<td>8.99852</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>21</td>
<td>159700</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.99975  
R-Square 0.9989  
Dependent Mean 94.64800  
Adj R-Sq 0.9988  
Coeff Var 3.16938

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|
| Intercept| 1  | 21631              | 639.50181      | 33.82   | <.0001|
| Year     | 1  | -24.04581          | 0.67547        | -35.60  | <.0001|
| YearSq   | 1  | 0.00668            | 0.00017820     | 37.51   | <.0001|

**Figure 102.21** OUTEST= Data Set

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>MODEL</em></th>
<th><em>TYPE</em></th>
<th><em>DEPVAR</em></th>
<th><em>RMSE</em></th>
<th>Intercept</th>
<th>Year</th>
<th>Population</th>
<th>YearSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m1</td>
<td>PARMS</td>
<td>Population</td>
<td>25.3295</td>
<td>-2345.85</td>
<td>1.2879</td>
<td>-1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>m2</td>
<td>PARMS</td>
<td>Population</td>
<td>2.9998</td>
<td>21630.89</td>
<td>-24.0458</td>
<td>-1 .006684346</td>
<td></td>
</tr>
</tbody>
</table>
The following modification of the previous example uses the TABLEOUT and ALPHA= options to obtain additional information in the OUTEST= data set:

```plaintext
proc reg data=USPopulation outest=est tableout alpha=0.1;
  m1: model Population=Year/noprint;
  m2: model Population=Year YearSq/noprint;
run;
proc print data=est;
run;
```

Notice that the TABLEOUT option causes standard errors, \( t \) statistics, \( p \)-values, and confidence limits for the estimates to be added to the OUTEST= data set. Also note that the ALPHA= option is used to set the confidence level at 90%. The OUTEST= data set is shown in Figure 102.22.

![Figure 102.22 The OUTEST= Data Set When TABLEOUT Is Specified](image)

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>MODEL</em></th>
<th><em>TYPE</em></th>
<th><em>DEPVAR</em></th>
<th><em>RMSE</em></th>
<th>Intercept</th>
<th>Year</th>
<th>Population</th>
<th>YearSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m1</td>
<td>PARMS</td>
<td>Population</td>
<td>25.3295</td>
<td>-2345.85</td>
<td>1.2879</td>
<td>-1</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>m1</td>
<td>STDERR</td>
<td>Population</td>
<td>25.3295</td>
<td>161.39</td>
<td>0.0851</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>m1</td>
<td>T</td>
<td>Population</td>
<td>25.3295</td>
<td>-14.54</td>
<td>15.1300</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>m1</td>
<td>PVALUE</td>
<td>Population</td>
<td>25.3295</td>
<td>0.00</td>
<td>0.0000</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>5</td>
<td>m1</td>
<td>L90B</td>
<td>Population</td>
<td>25.3295</td>
<td>-2624.21</td>
<td>1.1411</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>6</td>
<td>m1</td>
<td>U90B</td>
<td>Population</td>
<td>25.3295</td>
<td>-2067.50</td>
<td>1.4347</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>7</td>
<td>m2</td>
<td>PARMS</td>
<td>Population</td>
<td>2.9998</td>
<td>21630.89</td>
<td>-24.0458</td>
<td>-1</td>
<td>0.0067</td>
</tr>
<tr>
<td>8</td>
<td>m2</td>
<td>STDERR</td>
<td>Population</td>
<td>2.9998</td>
<td>639.50</td>
<td>0.6755</td>
<td>.</td>
<td>0.0002</td>
</tr>
<tr>
<td>9</td>
<td>m2</td>
<td>T</td>
<td>Population</td>
<td>2.9998</td>
<td>33.82</td>
<td>-35.5988</td>
<td>.</td>
<td>37.5096</td>
</tr>
<tr>
<td>10</td>
<td>m2</td>
<td>PVALUE</td>
<td>Population</td>
<td>2.9998</td>
<td>0.00</td>
<td>0.0000</td>
<td>.</td>
<td>0.0000</td>
</tr>
<tr>
<td>11</td>
<td>m2</td>
<td>L90B</td>
<td>Population</td>
<td>2.9998</td>
<td>20525.11</td>
<td>-25.2138</td>
<td>.</td>
<td>0.0064</td>
</tr>
<tr>
<td>12</td>
<td>m2</td>
<td>U90B</td>
<td>Population</td>
<td>2.9998</td>
<td>22736.68</td>
<td>-22.8778</td>
<td>.</td>
<td>0.0070</td>
</tr>
</tbody>
</table>

A slightly different OUTEST= data set is created when you use the RSQUARE selection method. The following statements request only the “best” model for each subset size but ask for a variety of model selection statistics, as well as the estimated regression coefficients. An OUTEST= data set is created and displayed. See Figure 102.23 and Figure 102.24 for the results.

```plaintext
proc reg data=fitness outest=est;
  model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse /
        selection=rsquare mse jp gmsep cp aic bic sbc b best=1;
run;
proc print data=est;
run;
```
Chapter 102: The REG Procedure

Figure 102.23 PROC REG Output for Physical Fitness Data: Best Models

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

R-Square Selection Method

<table>
<thead>
<tr>
<th>Number in</th>
<th>R-Square</th>
<th>C(p)</th>
<th>AIC</th>
<th>BIC</th>
<th>Estimated MSE of Prediction</th>
<th>J(p)</th>
<th>MSE</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.7434</td>
<td>13.6988</td>
<td>64.5341</td>
<td>65.4673</td>
<td>8.0546</td>
<td>8.0199</td>
<td>7.53384</td>
<td>67.40210</td>
</tr>
<tr>
<td>2</td>
<td>0.7642</td>
<td>12.3894</td>
<td>63.9050</td>
<td>64.8212</td>
<td>7.9478</td>
<td>7.8621</td>
<td>7.16842</td>
<td>68.20695</td>
</tr>
<tr>
<td>3</td>
<td>0.8111</td>
<td>6.9596</td>
<td>59.0373</td>
<td>61.3127</td>
<td>6.8583</td>
<td>6.7253</td>
<td>5.95669</td>
<td>64.77326</td>
</tr>
<tr>
<td>4</td>
<td>0.8368</td>
<td>4.8800</td>
<td>56.4995</td>
<td>60.3996</td>
<td>6.3984</td>
<td>6.2053</td>
<td>5.34346</td>
<td>63.66941</td>
</tr>
<tr>
<td>5</td>
<td>0.8480</td>
<td>5.1036</td>
<td>56.2996</td>
<td>61.5667</td>
<td>6.4565</td>
<td>6.1782</td>
<td>5.17634</td>
<td>64.90250</td>
</tr>
<tr>
<td>6</td>
<td>0.8487</td>
<td>7.0000</td>
<td>58.1616</td>
<td>64.0748</td>
<td>6.9870</td>
<td>6.5804</td>
<td>5.36825</td>
<td>68.19952</td>
</tr>
</tbody>
</table>

Parameter Estimates

<table>
<thead>
<tr>
<th>Number in</th>
<th>R-Square</th>
<th>Intercept</th>
<th>Age</th>
<th>Weight</th>
<th>RunTime</th>
<th>RunPulse</th>
<th>RestPulse</th>
<th>MaxPulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0.7434</td>
<td>82.42177</td>
<td>.</td>
<td>.</td>
<td>-3.31056</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>0.7642</td>
<td>88.46229</td>
<td>-0.15037</td>
<td>.</td>
<td>-3.20395</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>0.8111</td>
<td>111.71806</td>
<td>-0.25640</td>
<td>.</td>
<td>-2.82538</td>
<td>-0.13091</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>0.8368</td>
<td>98.14789</td>
<td>-0.19773</td>
<td>.</td>
<td>-2.76758</td>
<td>-0.34811</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>5</td>
<td>0.8480</td>
<td>102.20428</td>
<td>-0.21962</td>
<td>-0.07230</td>
<td>-2.68252</td>
<td>-0.37340</td>
<td>.</td>
<td>0.27051</td>
</tr>
<tr>
<td>6</td>
<td>0.8487</td>
<td>102.93448</td>
<td>-0.22697</td>
<td>-0.07417</td>
<td>-2.62865</td>
<td>-0.36963</td>
<td>-0.02153</td>
<td>0.30491</td>
</tr>
</tbody>
</table>

Figure 102.24 PROC PRINT Output for Physical Fitness Data: OUTEST= Data Set

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>MODEL</em></th>
<th><em>TYPE</em></th>
<th><em>DEPVAR</em></th>
<th><em>RMSE</em></th>
<th>Intercept</th>
<th>Age</th>
<th>Weight</th>
<th>RunTime</th>
<th>RunPulse</th>
<th>RestPulse</th>
<th>MaxPulse</th>
<th><em>IN</em></th>
<th><em>P</em></th>
<th>EDF</th>
<th>MSE</th>
<th>RSQ</th>
<th>CP</th>
<th>JP</th>
<th>GMSEP</th>
<th>AIC</th>
<th>BIC</th>
<th>SBC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.74478</td>
<td>82.422</td>
<td>.</td>
<td>.</td>
<td>-3.31056</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>29</td>
<td>7.53384</td>
<td>0.74338</td>
<td>13.6988</td>
<td>8.0199</td>
<td>8.0546</td>
<td>64.5341</td>
</tr>
<tr>
<td>2</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.67739</td>
<td>88.46229</td>
<td>-0.15037</td>
<td>.</td>
<td>-3.20395</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1</td>
<td>2</td>
<td>3</td>
<td>28</td>
<td>7.16842</td>
<td>0.76425</td>
<td>12.3894</td>
<td>7.86214</td>
<td>7.9478</td>
<td>63.9050</td>
</tr>
<tr>
<td>3</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.44063</td>
<td>111.71806</td>
<td>-0.25640</td>
<td>.</td>
<td>-2.82538</td>
<td>-0.13091</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1</td>
<td>3</td>
<td>4</td>
<td>27</td>
<td>5.95669</td>
<td>0.81109</td>
<td>6.9596</td>
<td>6.72530</td>
<td>6.5833</td>
<td>59.0373</td>
</tr>
<tr>
<td>4</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.31159</td>
<td>98.14789</td>
<td>-0.19773</td>
<td>.</td>
<td>-2.76758</td>
<td>-0.34811</td>
<td>.</td>
<td>.</td>
<td>.</td>
<td>-1</td>
<td>4</td>
<td>5</td>
<td>26</td>
<td>5.34346</td>
<td>0.83682</td>
<td>4.8800</td>
<td>6.20531</td>
<td>6.39837</td>
<td>56.4995</td>
</tr>
<tr>
<td>5</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.31695</td>
<td>102.93448</td>
<td>-0.22697</td>
<td>-0.07417</td>
<td>-2.62865</td>
<td>-0.36963</td>
<td>-0.02153</td>
<td>0.30491</td>
<td>.</td>
<td>-1</td>
<td>5</td>
<td>6</td>
<td>25</td>
<td>5.17634</td>
<td>0.84800</td>
<td>5.1063</td>
<td>6.17821</td>
<td>6.45651</td>
<td>56.2986</td>
</tr>
<tr>
<td>6</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.31695</td>
<td>102.93448</td>
<td>-0.22697</td>
<td>-0.07417</td>
<td>-2.62865</td>
<td>-0.36963</td>
<td>-0.02153</td>
<td>0.30491</td>
<td>.</td>
<td>-1</td>
<td>6</td>
<td>7</td>
<td>24</td>
<td>5.36825</td>
<td>0.84867</td>
<td>7.0000</td>
<td>6.58043</td>
<td>6.98700</td>
<td>58.1616</td>
</tr>
</tbody>
</table>

Obs | MaxPulse | Oxygen | _IN_ | _P_ | EDF | MSE | RSQ | CP | JP | GMSEP | AIC | BIC | SBC  |
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.</td>
<td>-1</td>
<td>1</td>
<td>2</td>
<td>29</td>
<td>7.53384</td>
<td>0.74338</td>
<td>13.6988</td>
<td>8.01990</td>
<td>8.0546</td>
<td>64.5341</td>
<td>65.4673</td>
<td>67.40210</td>
</tr>
<tr>
<td>2</td>
<td>.</td>
<td>-1</td>
<td>2</td>
<td>3</td>
<td>28</td>
<td>7.16842</td>
<td>0.76425</td>
<td>12.3894</td>
<td>7.86214</td>
<td>7.94778</td>
<td>63.9050</td>
<td>64.8212</td>
<td>68.20695</td>
</tr>
<tr>
<td>3</td>
<td>.</td>
<td>-1</td>
<td>3</td>
<td>4</td>
<td>27</td>
<td>5.95669</td>
<td>0.81109</td>
<td>6.9596</td>
<td>6.72530</td>
<td>6.8533</td>
<td>59.0373</td>
<td>61.3127</td>
<td>64.7733</td>
</tr>
<tr>
<td>4</td>
<td>0.27051</td>
<td>-1</td>
<td>4</td>
<td>5</td>
<td>26</td>
<td>5.34346</td>
<td>0.83682</td>
<td>4.8800</td>
<td>6.20531</td>
<td>6.39837</td>
<td>56.4995</td>
<td>60.3996</td>
<td>63.6694</td>
</tr>
<tr>
<td>5</td>
<td>0.30491</td>
<td>-1</td>
<td>5</td>
<td>6</td>
<td>25</td>
<td>5.17634</td>
<td>0.84800</td>
<td>5.1063</td>
<td>6.17821</td>
<td>6.45651</td>
<td>56.2986</td>
<td>61.5667</td>
<td>64.90250</td>
</tr>
<tr>
<td>6</td>
<td>0.30322</td>
<td>-1</td>
<td>6</td>
<td>7</td>
<td>24</td>
<td>5.36825</td>
<td>0.84867</td>
<td>7.0000</td>
<td>6.58043</td>
<td>6.98700</td>
<td>58.1616</td>
<td>64.0748</td>
<td>68.1995</td>
</tr>
</tbody>
</table>
OUTSSCP= Data Sets

The OUTSSCP= option produces a TYPE=SSCP output SAS data set containing sums of squares and crossproducts. A special row (observation) and column (variable) of the matrix called Intercept contain the number of observations and sums. Observations are identified by the character variable _NAME_. The data set contains all variables used in MODEL statements. You can specify additional variables that you want included in the crossproducts matrix with a VAR statement.

The SSCP data set is used when a large number of observations are explored in many different runs. The SSCP data set can be saved and used for subsequent runs, which are much less expensive since PROC REG never reads the original data again. If you run PROC REG once to create only a SSCP data set, you should list all the variables that you might need in a VAR statement or include all the variables that you might need in a MODEL statement.

The following statements use the fitness data from Example 102.2 to produce an output data set with the OUTSSCP= option. The resulting output is shown in Figure 102.25.

```sas
proc reg data=fitness outsscp=sscp;
  var Oxygen RunTime Age Weight RestPulse RunPulse MaxPulse;
run;
proc print data=sscp;
run;
```

Since a model is not fit to the data and since the only request is to create the SSCP data set, a MODEL statement is not required in this example. However, since the MODEL statement is not used, the VAR statement is required.

Figure 102.25 SSCP Data Set Created with OUTSSCP= Option: REG Procedure

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>TYPE</em></th>
<th><em>NAME</em></th>
<th>Intercept</th>
<th>Oxygen</th>
<th>RunTime</th>
<th>Age</th>
<th>Weight</th>
<th>RestPulse</th>
<th>RunPulse</th>
<th>MaxPulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SSCP</td>
<td>Intercept</td>
<td>31.00</td>
<td>1468.65</td>
<td>328.17</td>
<td>1478.00</td>
<td>2400.78</td>
<td>1657.00</td>
<td>5259.00</td>
<td>5387.00</td>
</tr>
<tr>
<td>2</td>
<td>SSCP</td>
<td>Oxygen</td>
<td>1468.65</td>
<td>70429.86</td>
<td>15356.14</td>
<td>69767.75</td>
<td>113522.26</td>
<td>78015.41</td>
<td>248497.31</td>
<td>254866.75</td>
</tr>
<tr>
<td>3</td>
<td>SSCP</td>
<td>RunTime</td>
<td>328.17</td>
<td>15356.14</td>
<td>3531.80</td>
<td>15687.24</td>
<td>25464.71</td>
<td>17684.05</td>
<td>55806.29</td>
<td>57113.72</td>
</tr>
<tr>
<td>4</td>
<td>SSCP</td>
<td>Age</td>
<td>1478.00</td>
<td>69767.75</td>
<td>15687.24</td>
<td>71282.00</td>
<td>114158.90</td>
<td>78806.00</td>
<td>250194.00</td>
<td>256218.00</td>
</tr>
<tr>
<td>5</td>
<td>SSCP</td>
<td>Weight</td>
<td>2400.78</td>
<td>113522.26</td>
<td>25464.71</td>
<td>114158.90</td>
<td>188008.20</td>
<td>128409.28</td>
<td>407745.67</td>
<td>417764.62</td>
</tr>
<tr>
<td>6</td>
<td>SSCP</td>
<td>RestPulse</td>
<td>1657.00</td>
<td>78015.41</td>
<td>17684.05</td>
<td>78806.00</td>
<td>128409.28</td>
<td>90311.00</td>
<td>281928.00</td>
<td>288583.00</td>
</tr>
<tr>
<td>7</td>
<td>SSCP</td>
<td>RunPulse</td>
<td>5259.00</td>
<td>248497.31</td>
<td>55806.29</td>
<td>250194.00</td>
<td>407745.67</td>
<td>281928.00</td>
<td>895317.00</td>
<td>916499.00</td>
</tr>
<tr>
<td>8</td>
<td>SSCP</td>
<td>MaxPulse</td>
<td>5387.00</td>
<td>254866.75</td>
<td>57113.72</td>
<td>256218.00</td>
<td>417764.62</td>
<td>288583.00</td>
<td>916499.00</td>
<td>938641.00</td>
</tr>
<tr>
<td>9</td>
<td>N</td>
<td></td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
</tr>
</tbody>
</table>

Interactive Analysis

PROC REG enables you to change interactively both the model and the data used to compute the model, and to produce and highlight scatter plots. See the section “Using PROC REG Interactively” on page 8447 for an overview of interactive analysis that uses PROC REG. The following statements can be used interactively (without reinvoking PROC REG): ADD, DELETE, MODEL, MTEST, OUTPUT, PAINT, PLOT, PRINT, REFIT, RESTRICT, REWEIGHT, and TEST. All interactive features are disabled if there is a BY statement.
The ADD, DELETE, and REWEIGHT statements can be used to modify the current MODEL. Every use of an ADD, DELETE, or REWEIGHT statement causes the model label to be modified by attaching an additional number to it. This number is the cumulative total of the number of ADD, DELETE, or REWEIGHT statements following the current MODEL statement.

A more detailed explanation of changing the data used to compute the model is given in the section “Reweighting Observations in an Analysis” on page 8524.

The following statements illustrate the usefulness of the interactive features. First, the full regression model is fit to the Sashelp.Class data, and Figure 102.26 is produced.

```sas
ods graphics on;
proc reg data=sashelp.Class plots(modelLabel only)=ResidualByPredicted;
   model Weight=Age Height;
run;
```

![Figure 102.26 Interactive Analysis: Full Model](image)

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>7215.63710</td>
<td>3607.81855</td>
<td>27.23</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>2120.09974</td>
<td>132.50623</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>18</td>
<td>9335.73684</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 11.51114
R-Square 0.7729
Dependent Mean 100.02632
Adj R-Sq 0.7445
Coeff Var 11.50811

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|
| Intercept| 1  | -141.22376         | 33.38309       | -4.23   | 0.0006|
| Age      | 1  | 1.27839            | 3.11010        | 0.41    | 0.6865|
| Height   | 1  | 3.59703            | 0.90546        | 3.97    | 0.0011|

Next, the regression model is reduced by the following statements, and Figure 102.27 is produced.

```sas
delete age;
print;
run;
```
Note that the MODEL label has been changed from MODEL1 to MODEL1.1, since the original MODEL has been changed by the delete statement.

When ODS Graphics is enabled, updated plots are produced whenever a PRINT statement is used. The option

```plaintext
plots(modelLabel only)=ResidualByPredicted
```

in the PROC REG statement specifies that the only plot produced is a scatter plot of residuals by predicted values. The MODELLABEL option specifies that the current model label is added to the plot.

The following statements generate a scatter plot of the residuals against the predicted values from the full model. Figure 102.28 is produced, and the scatter plot shows a possible outlier.

```plaintext
add age;  
print;  
run;
```
The following statements delete the observation with the largest residual, refit the regression model, and produce a scatter plot of residuals against predicted values for the refitted model. Figure 102.29 shows the new scatter plot.

```
   reweight r.>20;
   print;
   run;
```
The nine methods of model selection implemented in PROC REG are specified with the SELECTION= option in the MODEL statement. Each method is discussed in this section.

**Full Model Fitted (NONE)**

This method is the default and provides no model selection capability. The complete model specified in the MODEL statement is used to fit the model. For many regression analyses, this might be the only method you need.

**Forward Selection (FORWARD)**

The forward-selection technique begins with no variables in the model. For each of the independent variables, the FORWARD method calculates $F$ statistics that reflect the variable’s contribution to the model if it is included. The $p$-values for these $F$ statistics are compared to the SLENTRY= value that is specified in the MODEL statement (or to 0.50 if the SLENTRY= option is omitted). If no $F$ statistic has a significance level
greater than the SLENTRY= value, the FORWARD selection stops. Otherwise, the FORWARD method adds the variable that has the largest $F$ statistic to the model. The FORWARD method then calculates $F$ statistics again for the variables still remaining outside the model, and the evaluation process is repeated. Thus, variables are added one by one to the model until no remaining variable produces a significant $F$ statistic. Once a variable is in the model, it stays.

**Backward Elimination (BACKWARD)**

The backward elimination technique begins by calculating $F$ statistics for a model which includes all of the independent variables. Then the variables are deleted from the model one by one until all the variables remaining in the model produce $F$ statistics significant at the SLSTAY= level specified in the MODEL statement (or at the 0.10 level if the SLSTAY= option is omitted). At each step, the variable showing the smallest contribution to the model is deleted.

**Stepwise (STEPWISE)**

The stepwise method is a modification of the forward-selection technique and differs in that variables already in the model do not necessarily stay there. As in the forward-selection method, variables are added one by one to the model, and the $F$ statistic for a variable to be added must be significant at the SLENTRY= level. After a variable is added, however, the stepwise method looks at all the variables already included in the model and deletes any variable that does not produce an $F$ statistic significant at the SLSTAY= level. Only after this check is made and the necessary deletions are accomplished can another variable be added to the model. The stepwise process ends when none of the variables outside the model has an $F$ statistic significant at the SLENTRY= level and every variable in the model is significant at the SLSTAY= level, or when the variable to be added to the model is the one just deleted from it.

**Maximum R$^2$ Improvement (MAXR)**

The maximum R square improvement technique does not settle on a single model. Instead, it tries to find the “best” one-variable model, the “best” two-variable model, and so forth, although it is not guaranteed to find the model with the largest R square for each size.

The MAXR method begins by finding the one-variable model producing the highest R square. Then another variable, the one that yields the greatest increase in R square, is added. Once the two-variable model is obtained, each of the variables in the model is compared to each variable not in the model. For each comparison, the MAXR method determines if removing one variable and replacing it with the other variable increases R square. After comparing all possible switches, the MAXR method makes the switch that produces the largest increase in R square. Comparisons begin again, and the process continues until the MAXR method finds that no switch could increase R square. Thus, the two-variable model achieved is considered the “best” two-variable model the technique can find. Another variable is then added to the model, and the comparing-and-switching process is repeated to find the “best” three-variable model, and so forth.

The difference between the STEPWISE method and the MAXR method is that all switches are evaluated before any switch is made in the MAXR method. In the STEPWISE method, the “worst” variable might be removed without considering what adding the “best” remaining variable might accomplish. The MAXR method might require much more computer time than the STEPWISE method.
Minimum $R^2$ (MINR) Improvement

The MINR method closely resembles the MAXR method, but the switch chosen is the one that produces the smallest increase in $R$ square. For a given number of variables in the model, the MAXR and MINR methods usually produce the same “best” model, but the MINR method considers more models of each size.

$R^2$ Selection (RSQUARE)

The RSQUARE method finds subsets of independent variables that best predict a dependent variable by linear regression in the given sample. You can specify the largest and smallest number of independent variables to appear in a subset and the number of subsets of each size to be selected. The RSQUARE method can efficiently perform all possible subset regressions and display the models in decreasing order of $R$ square magnitude within each subset size. Other statistics are available for comparing subsets of different sizes. These statistics, as well as estimated regression coefficients, can be displayed or output to a SAS data set.

The subset models selected by the RSQUARE method are optimal in terms of $R$ square for the given sample, but they are not necessarily optimal for the population from which the sample is drawn or for any other sample for which you might want to make predictions. If a subset model is selected on the basis of a large $R$ square value or any other criterion commonly used for model selection, then all regression statistics computed for that model under the assumption that the model is given a priori, including all statistics computed by PROC REG, are biased.

While the RSQUARE method is a useful tool for exploratory model building, no statistical method can be relied on to identify the “true” model. Effective model building requires substantive theory to suggest relevant predictors and plausible functional forms for the model.

The RSQUARE method differs from the other selection methods in that RSQUARE always identifies the model with the largest $R$ square for each number of variables considered. The other selection methods are not guaranteed to find the model with the largest $R$ square. The RSQUARE method requires much more computer time than the other selection methods, so a different selection method such as the STEPWISE method is a good choice when there are many independent variables to consider.

Adjusted $R^2$ Selection (ADJRSQ)

This method is similar to the RSQUARE method, except that the adjusted $R$ square statistic is used as the criterion for selecting models, and the method finds the models with the highest adjusted $R$ square within the range of sizes.

Mallows’ $C_p$ Selection (CP)

This method is similar to the ADJRSQ method, except that Mallows’ $C_p$ statistic is used as the criterion for model selection. Models are listed in ascending order of $C_p$.

Additional Information about Model-Selection Methods

Reviews of model-selection methods by Hocking (1976) and Judge et al. (1980) describe these and other variable-selection methods.
Criteria Used in Model-Selection Methods

When many significance tests are performed, each at a level of, for example, 5%, the overall probability of rejecting at least one true null hypothesis is much larger than 5%. If you want to guard against including any variables that do not contribute to the predictive power of the model in the population, you should specify a very small SLE= significance level for the FORWARD and STEPWISE methods and a very small SLS= significance level for the BACKWARD and STEPWISE methods.

In most applications, many of the variables considered have some predictive power, however small. If you want to choose the model that provides the best prediction computed using the sample estimates, you need only to guard against estimating more parameters than can be reliably estimated with the given sample size, so you should use a moderate significance level, perhaps in the range of 10% to 25%.

In addition to R square, the $C_p$ statistic is displayed for each model generated in the model-selection methods. The $C_p$ statistic is proposed by Mallows (1973) as a criterion for selecting a model. It is a measure of total squared error defined as

$$C_p = \frac{SSE_p}{s^2} - (N - 2p)$$

where $s^2$ is the MSE for the full model, and $SSE_p$ is the sum-of-squares error for a model with $p$ parameters including the intercept, if any. If $C_p$ is plotted against $p$, Mallows recommends the model where $C_p$ first approaches $p$. When the right model is chosen, the parameter estimates are unbiased, and this is reflected in $C_p$ near $p$. For further discussion, see Daniel and Wood (1980).

The adjusted R square statistic is an alternative to R square that is adjusted for the number of parameters in the model. The adjusted R square statistic is calculated as

$$ADJRSQ = 1 - \frac{(n - i)(1 - R^2)}{n - p}$$

where $n$ is the number of observations used in fitting the model, and $i$ is an indicator variable that is 1 if the model includes an intercept, and 0 otherwise.

Limitations in Model-Selection Methods

The use of model-selection methods can be time-consuming in some cases because there is no built-in limit on the number of independent variables, and the calculations for a large number of independent variables can be lengthy. The recommended limit on the number of independent variables for the MINR method is $20 + i$, where $i$ is the value of the INCLUDE= option.

For the RSQUARE, ADJRSQ, or CP method, with a large value of the BEST= option, adding one more variable to the list from which regressors are selected might significantly increase the CPU time. Also, the time required for the analysis is highly dependent on the data and on the values of the BEST=, START=, and STOP= options.
Parameter Estimates and Associated Statistics

The following example uses the fitness data from Example 102.2. Figure 102.31 shows the parameter estimates and the tables from the SS1, SS2, STB, CLB, COVB, and CORRB options:

```plaintext
proc reg data=fitness;
   model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse / ss1 ss2 stb clb covb corrb;
run;
```

The procedure first displays an analysis of variance table (Figure 102.30). The $F$ statistic for the overall model is significant, indicating that the model explains a significant portion of the variation in the data.

![Figure 102.30 ANOVA Table](image)

The procedure next displays parameter estimates and some associated statistics (Figure 102.31). First, the estimates are shown, followed by their standard errors. The next two columns of the table contain the $t$ statistics and the corresponding probabilities for testing the null hypothesis that the parameter is not significantly different from zero. These probabilities are usually referred to as $p$-values. For example, the Intercept term in the model is estimated to be 102.9 and is significantly different from zero. The next two columns of the table are the result of requesting the SS1 and SS2 options, and they show sequential and partial sums of squares (SS) associated with each variable. The standardized estimates (produced by the STB option) are the parameter estimates that result when all variables are standardized to a mean of 0 and a variance of 1. These estimates are computed by multiplying the original estimates by the standard deviation of the regressor (independent) variable and then dividing by the standard deviation of the dependent variable. The CLB option adds the upper and lower 95% confidence limits for the parameter estimates; the $\alpha$ level can be changed by specifying the ALPHA= option in the PROC REG or MODEL statement.

![Parameter Estimates and Associated Statistics](image)
### Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Parameter &gt;</th>
<th>Type I SS</th>
<th>Type II SS</th>
<th>Standardized Estimate</th>
<th>95% Confidence Limits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>102.93448</td>
<td>12.40326</td>
<td>8.30</td>
<td>&lt; .0001</td>
<td>69578</td>
<td>369.72831</td>
<td>0</td>
<td>77.33541 128.53355</td>
</tr>
<tr>
<td>RunTime</td>
<td>1</td>
<td>-2.62865</td>
<td>0.38456</td>
<td>-6.84</td>
<td>&lt; .0001</td>
<td>632.90010</td>
<td>520.82210</td>
<td>-0.68460</td>
<td>-3.42235 -1.83496</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>-0.22697</td>
<td>0.09984</td>
<td>-2.27</td>
<td>0.0322</td>
<td>17.76563</td>
<td>27.74577</td>
<td>-0.71133</td>
<td>-0.36460 -0.05806</td>
</tr>
<tr>
<td>Weight</td>
<td>1</td>
<td>-0.07418</td>
<td>0.05459</td>
<td>-1.36</td>
<td>0.1869</td>
<td>5.60522</td>
<td>9.91059</td>
<td>-0.11597</td>
<td>-0.22204 -0.00926</td>
</tr>
<tr>
<td>RunPulse</td>
<td>1</td>
<td>-0.36963</td>
<td>0.11985</td>
<td>-3.08</td>
<td>0.0051</td>
<td>38.87574</td>
<td>51.05806</td>
<td>-0.71133</td>
<td>-0.36460 -0.05806</td>
</tr>
<tr>
<td>MaxPulse</td>
<td>1</td>
<td>0.30322</td>
<td>0.13650</td>
<td>2.22</td>
<td>0.0360</td>
<td>26.82640</td>
<td>26.49142</td>
<td>-0.11597</td>
<td>0.02150 0.58493</td>
</tr>
<tr>
<td>RestPulse</td>
<td>1</td>
<td>-0.02153</td>
<td>0.06605</td>
<td>-0.33</td>
<td>0.7473</td>
<td>0.57051</td>
<td>0.57051</td>
<td>-0.30800</td>
<td>-0.15786 0.11480</td>
</tr>
</tbody>
</table>

The final two tables are produced as a result of requesting the COVB and CORRB options (Figure 102.32). These tables show the estimated covariance matrix of the parameter estimates, and the estimated correlation matrix of the estimates.

### Covariance of Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intercept</th>
<th>RunTime</th>
<th>Age</th>
<th>Weight</th>
<th>RunPulse</th>
<th>MaxPulse</th>
<th>RestPulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>153.84081152</td>
<td>0.7678373769</td>
<td>-0.902049478</td>
<td>-0.178237818</td>
<td>0.280796516</td>
<td>-0.832761667</td>
<td>-0.147954715</td>
</tr>
<tr>
<td>RunTime</td>
<td>0.7678373769</td>
<td>0.1478880839</td>
<td>-0.014191688</td>
<td>-0.004417672</td>
<td>-0.009047784</td>
<td>0.0046249498</td>
<td>-0.010915224</td>
</tr>
<tr>
<td>Age</td>
<td>-0.902049478</td>
<td>-0.014191688</td>
<td>0.009967521</td>
<td>0.0010219105</td>
<td>-0.001203914</td>
<td>0.0035823843</td>
<td>0.0014897532</td>
</tr>
<tr>
<td>Weight</td>
<td>-0.178237818</td>
<td>-0.004417672</td>
<td>0.0010219105</td>
<td>0.0029804131</td>
<td>-0.001372241</td>
<td>0.0003799295</td>
<td></td>
</tr>
<tr>
<td>RunPulse</td>
<td>0.280796516</td>
<td>-0.009047784</td>
<td>-0.001203914</td>
<td>0.0009644683</td>
<td>0.0143647273</td>
<td>-0.014952457</td>
<td>-0.000764507</td>
</tr>
<tr>
<td>MaxPulse</td>
<td>-0.832761667</td>
<td>0.0046249498</td>
<td>0.0035823843</td>
<td>-0.001372241</td>
<td>0.0186309364</td>
<td>0.0003425724</td>
<td></td>
</tr>
<tr>
<td>RestPulse</td>
<td>-0.147954715</td>
<td>-0.010915224</td>
<td>0.0014897532</td>
<td>0.0003799295</td>
<td>-0.000764507</td>
<td>0.0003425724</td>
<td>0.0043631674</td>
</tr>
</tbody>
</table>

### Correlation of Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intercept</th>
<th>RunTime</th>
<th>Age</th>
<th>Weight</th>
<th>RunPulse</th>
<th>MaxPulse</th>
<th>RestPulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1.0000</td>
<td>0.1610</td>
<td>-0.7285</td>
<td>-0.2632</td>
<td>0.1889</td>
<td>-0.4919</td>
<td>-0.1806</td>
</tr>
<tr>
<td>RunTime</td>
<td>0.1610</td>
<td>1.0000</td>
<td>-0.3696</td>
<td>-0.2104</td>
<td>0.1875</td>
<td>-0.1006</td>
<td>0.2629</td>
</tr>
<tr>
<td>Age</td>
<td>-0.7285</td>
<td>-0.3696</td>
<td>1.0000</td>
<td>0.1875</td>
<td>1.0000</td>
<td>0.1474</td>
<td>-0.1842</td>
</tr>
<tr>
<td>Weight</td>
<td>-0.2632</td>
<td>-0.2104</td>
<td>0.1875</td>
<td>1.0000</td>
<td>0.1474</td>
<td>1.0000</td>
<td>-0.9140</td>
</tr>
<tr>
<td>RunPulse</td>
<td>0.1889</td>
<td>-0.1963</td>
<td>-0.1006</td>
<td>0.1474</td>
<td>1.0000</td>
<td>-0.9140</td>
<td>-0.0966</td>
</tr>
<tr>
<td>MaxPulse</td>
<td>-0.4919</td>
<td>0.0881</td>
<td>0.2629</td>
<td>-0.1842</td>
<td>-0.9140</td>
<td>1.0000</td>
<td>0.0380</td>
</tr>
<tr>
<td>RestPulse</td>
<td>-0.1806</td>
<td>-0.4297</td>
<td>0.2259</td>
<td>0.1054</td>
<td>-0.0966</td>
<td>0.0380</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

For further discussion of the parameters and statistics, see the section “Displayed Output” on page 8538, and Chapter 4, “Introduction to Regression Procedures.”
Predicted and Residual Values

The display of the predicted values and residuals is controlled by the P, R, CLM, and CLI options in the MODEL statement. The P option causes PROC REG to display the observation number, the ID value (if an ID statement is used), the actual value, the predicted value, and the residual. The R, CLI, and CLM options also produce the items under the P option. Thus, P is unnecessary if you use one of the other options.

The R option requests more detail, especially about the residuals. The standard errors of the mean predicted value and the residual are displayed. The studentized residual, which is the residual divided by its standard error, is both displayed and plotted. A measure of influence, Cook’s D, is displayed and plotted. Cook’s D measures the change to the estimates that results from deleting each observation (Cook 1977, 1979). This statistic is very similar to DFFITS.

The CLM option requests that PROC REG display the 100(1 − α)% lower and upper confidence limits for the mean predicted values. This accounts for the variation due to estimating the parameters only. If you want a 100(1 − α)% confidence interval for observed values, then you can use the CLI option, which adds in the variability of the error term. The α level can be specified with the ALPHA= option in the PROC REG or MODEL statement.

You can use these statistics in PLOT and PAINT statements. This is useful in performing a variety of regression diagnostics. For definitions of the statistics produced by these options, see Chapter 4, “Introduction to Regression Procedures.”

The following statements use the U.S. population data found in the section “Polynomial Regression” on page 8436. The results are shown in Figure 102.33 and Figure 102.34.

```sas
ods graphics on;

data USPop2;
  input Year @@;
  YearSq=Year*Year;
datalines;
2010 2020 2030
;

data USPop2;
  set USPopulation USPop2;
run;

proc reg data=USPop2;
  id Year;
  model Population=Year YearSq / r cli clm;
run;
```
## Figure 102.33  Regression Using the R, CLI, and CLM Options

**The REG Procedure**

**Model: MODEL1**

**Dependent Variable: Population**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>159529</td>
<td>79765</td>
<td>8864.19</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>170.97193</td>
<td>8.99852</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>21</td>
<td>159700</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Root MSE**: 2.99975
- **R-Square**: 0.9989
- **Dependent Mean**: 94.64800
- **Adj R-Sq**: 0.9988
- **Coeff Var**: 3.16938

### Parameter Estimates

| Variable | DF | Parameter | Standard Error | t Value | Pr > |t| |
|----------|----|-----------|----------------|---------|------|---|
| Intercept| 1  | 21631     | 639.50181      | 33.82   | <.0001 |
| Year     | 1  | -24.04581 | 0.67547        | -35.60  | <.0001 |
| YearSq   | 1  | 0.00668   | 0.00017820     | 37.51   | <.0001 |

---

The REG Procedure

Model: MODEL1

Dependent Variable: Population

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>159529</td>
<td>79765</td>
<td>8864.19</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>19</td>
<td>170.97193</td>
<td>8.99852</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>21</td>
<td>159700</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **Root MSE**: 2.99975
- **R-Square**: 0.9989
- **Dependent Mean**: 94.64800
- **Adj R-Sq**: 0.9988
- **Coeff Var**: 3.16938

### Parameter Estimates

| Variable | DF | Parameter | Standard Error | t Value | Pr > |t| |
|----------|----|-----------|----------------|---------|------|---|
| Intercept| 1  | 21631     | 639.50181      | 33.82   | <.0001 |
| Year     | 1  | -24.04581 | 0.67547        | -35.60  | <.0001 |
| YearSq   | 1  | 0.00668   | 0.00017820     | 37.51   | <.0001 |
### Figure 10.2.34 Regression Using the R, CLI, and CLM Options

**The REG Procedure**  
**Model: MODEL1**  
**Dependent Variable: Population**

<table>
<thead>
<tr>
<th>Obs</th>
<th>Year</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Std Error Mean</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
<th>Residual</th>
<th>Std Error Residual</th>
<th>Student Residual</th>
<th>Cook’s D</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1800</td>
<td>5.31</td>
<td>5.7226</td>
<td>1.4560</td>
<td>2.6751</td>
<td>8.7701</td>
<td>-1.2565</td>
<td>12.7017</td>
<td>-0.4146</td>
<td>2.623</td>
</tr>
<tr>
<td>3</td>
<td>1810</td>
<td>7.24</td>
<td>6.5694</td>
<td>1.2118</td>
<td>4.0331</td>
<td>9.1057</td>
<td>-0.2021</td>
<td>13.3409</td>
<td>0.6696</td>
<td>2.744</td>
</tr>
<tr>
<td>4</td>
<td>1820</td>
<td>9.64</td>
<td>8.7531</td>
<td>1.0305</td>
<td>6.5963</td>
<td>10.9100</td>
<td>2.1144</td>
<td>15.3918</td>
<td>0.8849</td>
<td>2.817</td>
</tr>
<tr>
<td>5</td>
<td>1830</td>
<td>12.87</td>
<td>12.2737</td>
<td>0.9163</td>
<td>10.3558</td>
<td>14.1916</td>
<td>5.7087</td>
<td>18.8386</td>
<td>0.5923</td>
<td>2.856</td>
</tr>
<tr>
<td>6</td>
<td>1840</td>
<td>17.07</td>
<td>17.1311</td>
<td>0.8650</td>
<td>15.3207</td>
<td>18.9415</td>
<td>10.5968</td>
<td>23.6655</td>
<td>-0.0621</td>
<td>2.872</td>
</tr>
<tr>
<td>7</td>
<td>1850</td>
<td>23.19</td>
<td>23.3254</td>
<td>0.8613</td>
<td>21.5227</td>
<td>25.1281</td>
<td>16.7932</td>
<td>29.8576</td>
<td>-0.1344</td>
<td>2.873</td>
</tr>
<tr>
<td>8</td>
<td>1860</td>
<td>31.44</td>
<td>30.8566</td>
<td>0.8846</td>
<td>29.0051</td>
<td>32.7080</td>
<td>24.3107</td>
<td>37.4024</td>
<td>0.5864</td>
<td>2.866</td>
</tr>
<tr>
<td>9</td>
<td>1870</td>
<td>39.82</td>
<td>39.7246</td>
<td>0.9163</td>
<td>37.8067</td>
<td>41.6425</td>
<td>33.1597</td>
<td>46.2896</td>
<td>0.0934</td>
<td>2.856</td>
</tr>
<tr>
<td>10</td>
<td>1880</td>
<td>50.16</td>
<td>49.9295</td>
<td>0.9436</td>
<td>47.9545</td>
<td>51.9046</td>
<td>43.3476</td>
<td>56.5114</td>
<td>0.2255</td>
<td>2.847</td>
</tr>
<tr>
<td>11</td>
<td>1890</td>
<td>62.95</td>
<td>61.4713</td>
<td>0.9590</td>
<td>59.4641</td>
<td>63.4785</td>
<td>54.8797</td>
<td>68.0629</td>
<td>1.4757</td>
<td>2.842</td>
</tr>
<tr>
<td>12</td>
<td>1900</td>
<td>75.99</td>
<td>74.3499</td>
<td>0.9590</td>
<td>72.3427</td>
<td>76.3571</td>
<td>67.7831</td>
<td>80.9415</td>
<td>1.6441</td>
<td>2.842</td>
</tr>
<tr>
<td>13</td>
<td>1910</td>
<td>91.97</td>
<td>88.6565</td>
<td>0.9436</td>
<td>86.5904</td>
<td>90.5405</td>
<td>81.9836</td>
<td>95.1473</td>
<td>3.4065</td>
<td>2.847</td>
</tr>
<tr>
<td>14</td>
<td>1920</td>
<td>105.71</td>
<td>104.1178</td>
<td>0.9163</td>
<td>102.2000</td>
<td>106.0357</td>
<td>97.5529</td>
<td>110.6828</td>
<td>1.5922</td>
<td>2.856</td>
</tr>
<tr>
<td>15</td>
<td>1930</td>
<td>122.78</td>
<td>121.0071</td>
<td>0.8846</td>
<td>119.1556</td>
<td>122.8585</td>
<td>114.4612</td>
<td>127.5529</td>
<td>1.7679</td>
<td>2.866</td>
</tr>
<tr>
<td>16</td>
<td>1940</td>
<td>131.67</td>
<td>139.2332</td>
<td>0.8613</td>
<td>137.4305</td>
<td>141.0359</td>
<td>132.7010</td>
<td>145.7654</td>
<td>-7.5642</td>
<td>2.873</td>
</tr>
<tr>
<td>17</td>
<td>1950</td>
<td>151.33</td>
<td>158.7962</td>
<td>0.8650</td>
<td>156.9858</td>
<td>160.6066</td>
<td>152.2618</td>
<td>165.3306</td>
<td>-7.4712</td>
<td>2.872</td>
</tr>
<tr>
<td>18</td>
<td>1960</td>
<td>179.32</td>
<td>179.6961</td>
<td>0.9163</td>
<td>177.7782</td>
<td>181.6139</td>
<td>173.1311</td>
<td>186.2610</td>
<td>-0.3731</td>
<td>2.856</td>
</tr>
<tr>
<td>19</td>
<td>1970</td>
<td>203.21</td>
<td>201.9328</td>
<td>1.0305</td>
<td>199.7759</td>
<td>204.0896</td>
<td>195.2941</td>
<td>208.5715</td>
<td>1.2782</td>
<td>2.817</td>
</tr>
<tr>
<td>20</td>
<td>1980</td>
<td>226.54</td>
<td>225.5064</td>
<td>1.2118</td>
<td>222.9701</td>
<td>228.0427</td>
<td>218.7349</td>
<td>232.2779</td>
<td>1.0356</td>
<td>2.744</td>
</tr>
<tr>
<td>22</td>
<td>2000</td>
<td>281.42</td>
<td>276.6642</td>
<td>1.7565</td>
<td>272.9877</td>
<td>280.3407</td>
<td>269.3884</td>
<td>283.9400</td>
<td>4.7578</td>
<td>2.432</td>
</tr>
<tr>
<td>23</td>
<td>2010</td>
<td>304.2484</td>
<td>2.1073</td>
<td>299.8377</td>
<td>308.6591</td>
<td>296.5754</td>
<td>311.9214</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>24</td>
<td>2020</td>
<td>333.1695</td>
<td>2.5040</td>
<td>327.9285</td>
<td>338.4104</td>
<td>324.9910</td>
<td>341.3479</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
</tbody>
</table>
Chapter 102: The REG Procedure

Figure 102.35  Studentized Residuals and Cook’s $D$

After producing the usual analysis of variance and parameter estimates tables (Figure 102.33), the procedure displays the results of requesting the options for predicted and residual values (Figure 102.34). For each observation, the requested information is shown. Note that the ID variable is used to identify each observation. Also note that, for observations with missing dependent variables, the predicted value, standard error of the predicted value, and confidence intervals for the predicted value are still available.

The studentized residuals and Cook’s $D$ statistics in Figure 102.34 and Figure 102.35 are displayed as a result of specifying the R option. The large absolute studentized residuals for 1940 and 1950 (best seen in Figure 102.35) indicate that the overall model is inadequate for explaining the population in these two years. You can use ODS Graphics to obtain plots of studentized residuals by predicted values or leverage; see Example 102.1 for a similar example.
Models of Less Than Full Rank

If the model is not full rank, there are an infinite number of least squares solutions for the estimates. PROC REG chooses a nonzero solution for all variables that are linearly independent of previous variables and a zero solution for other variables. This solution corresponds to using a generalized inverse in the normal equations, and the expected values of the estimates are the Hermite normal form of $X$ multiplied by the true parameters:

$$E(b) = (X'X)^{-1}(X'X)\beta$$

Degrees of freedom for the zeroed estimates are reported as zero. The hypotheses that are not testable have $t$ tests reported as missing. The message that the model is not full rank includes a display of the relations that exist in the matrix.

The following statements use the fitness data from Example 102.2. The variable $Dif=RunPulse-RestPulse$ is created. When this variable is included in the model along with $RunPulse$ and $RestPulse$, there is a linear dependency (or exact collinearity) between the independent variables. Figure 102.36 shows how this problem is diagnosed.

```plaintext
data fit2;
  set fitness;
  Dif=RunPulse-RestPulse;
run;
proc reg data=fit2;
  model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse Dif;
run;
```
**Figure 102.36** Model That Is Not Full Rank: REG Procedure

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>722.54361</td>
<td>120.42393</td>
<td>22.43</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>128.83794</td>
<td>5.36825</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.31695
R-Square 0.8487
Dependent Mean 47.37581
Adj R-Sq 0.8108
Coeff Var 4.89057

| Parameter Estimates | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|--------------------|----|--------------------|----------------|---------|------|-----|
| Intercept          | 1  | 102.93448          | 12.40326       | 8.30    | <.0001 |
| RunTime            | 1  | -2.62865           | 0.38456        | -6.84   | <.0001 |
| Age                | 1  | -0.22697           | 0.09984        | -2.27   | 0.0322 |
| Weight             | 1  | -0.07418           | 0.05459        | -1.36   | 0.1869 |
| RunPulse B         | B  | -0.36963           | 0.11985        | -3.08   | 0.0051 |
| MaxPulse           | 1  | 0.30322            | 0.13650        | 2.22    | 0.0360 |
| RestPulse B        | B  | -0.02153           | 0.06605        | -0.33   | 0.7473 |
| Dif                | 0  | 0                  | .              | .       | .     |

PROC REG produces a message informing you that the model is less than full rank. Parameters with DF=0 are not estimated, and parameters with DF=B are biased. In addition, the form of the linear dependency among the regressors is displayed.
Collinearity Diagnostics

When a regressor is nearly a linear combination of other regressors in the model, the affected estimates are unstable and have high standard errors. This problem is called *collinearity* or *multicollinearity*. It is a good idea to find out which variables are nearly collinear with which other variables. The approach in PROC REG follows that of Belsley, Kuh, and Welsch (1980). PROC REG provides several methods for detecting collinearity with the COLLIN, COLLINOINT, TOL, and VIF options.

The COLLIN option in the MODEL statement requests that a collinearity analysis be performed. First, $X'X$ is scaled to have 1s on the diagonal. If you specify the COLLINOINT option, the intercept variable is adjusted out first. Then the eigenvalues and eigenvectors are extracted. The analysis in PROC REG is reported with eigenvalues of $X'X$ rather than singular values of $X$. The eigenvalues of $X'X$ are the squares of the singular values of $X$.

The condition indices are the square roots of the ratio of the largest eigenvalue to each individual eigenvalue. The largest condition index is the condition number of the scaled $X$ matrix. Belsley, Kuh, and Welsch (1980) suggest that, when this number is around 10, weak dependencies might be starting to affect the regression estimates. When this number is larger than 100, the estimates might have a fair amount of numerical error (although the statistical standard error almost always is much greater than the numerical error).

For each variable, PROC REG produces the proportion of the variance of the estimate accounted for by each principal component. A collinearity problem occurs when a component associated with a high condition index contributes strongly (variance proportion greater than about 0.5) to the variance of two or more variables.

The VIF option in the MODEL statement provides the variance inflation factors (VIF). These factors measure the inflation in the variances of the parameter estimates due to collinearities that exist among the regressor (independent) variables. There are no formal criteria for deciding if a VIF is large enough to affect the predicted values.

The TOL option requests the tolerance values for the parameter estimates. The tolerance is defined as $1 / \text{VIF}$.

For a complete discussion of the preceding methods, see Belsley, Kuh, and Welsch (1980). For a more detailed explanation of using the methods with PROC REG, see Freund and Littell (1986).

This example uses the COLLIN option on the fitness data found in Example 102.2. The following statements produce Figure 102.37.

```plaintext
proc reg data=fitness;
    model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse / tol vif collin;
run;
```
Figure 102.37  Regression Using the TOL, VIF, and COLLIN Options

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>722.54361</td>
<td>120.42393</td>
<td>22.43</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>128.83794</td>
<td>5.36825</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 2.31695  R-Square 0.8487
Dependent Mean 47.37581  Adj R-Sq 0.8108
Coeff Var 4.89057

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| Tolerance | Variance Inflation |
|----------|----|--------------------|----------------|---------|------|-----------|-------------------|
| Intercept| 1  | 102.93448          | 12.40326       | 8.30    | <.0001| .          | 0                 |
| RunTime  | 1  | -2.62865           | 0.38456        | -6.84   | <.0001| 0.62859   | 1.59087           |
| Age      | 1  | -0.22697           | 0.09984        | -2.27   | 0.0322| 0.66101   | 1.51284           |
| Weight   | 1  | -0.07418           | 0.09984        | -0.74   | 0.4601| 0.86555   | 1.15533           |
| RunPulse | 1  | -0.36963           | 0.11985        | -3.08   | 0.0051| 0.11852   | 8.43727           |
| MaxPulse | 1  | 0.30322            | 0.13650        | 2.22    | 0.0360| 0.11437   | 8.74385           |
| RestPulse| 1  | -0.02153           | 0.06605        | -0.33   | 0.7473| 0.70642   | 1.41559           |

Collinearity Diagnostics

<table>
<thead>
<tr>
<th>Number</th>
<th>Eigenvalue</th>
<th>Condition Index</th>
<th>Intercept</th>
<th>RunTime</th>
<th>Age</th>
<th>Weight</th>
<th>RunPulse</th>
<th>MaxPulse</th>
<th>RestPulse</th>
<th>Proportion of Variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6.94991</td>
<td>1.00000</td>
<td>0.00002326</td>
<td>0.00021086</td>
<td>0.00015451</td>
<td>0.00019651</td>
<td>0.00000634</td>
<td>0.00027850</td>
<td>1.00000</td>
<td>0.00002326</td>
</tr>
<tr>
<td>2</td>
<td>0.01868</td>
<td>19.29087</td>
<td>0.00218</td>
<td>0.02522</td>
<td>0.14632</td>
<td>0.01042</td>
<td>0.00000244</td>
<td>0.0000743</td>
<td>0.39064</td>
<td>0.00000218</td>
</tr>
<tr>
<td>3</td>
<td>0.01503</td>
<td>21.50072</td>
<td>0.00638</td>
<td>0.60897</td>
<td>0.03186</td>
<td>0.18313</td>
<td>0.00419</td>
<td>0.0123</td>
<td>0.19030</td>
<td>0.0000638</td>
</tr>
<tr>
<td>4</td>
<td>0.00911</td>
<td>27.62115</td>
<td>0.00133</td>
<td>0.12501</td>
<td>0.11284</td>
<td>0.44442</td>
<td>0.01506</td>
<td>0.00833</td>
<td>0.36475</td>
<td>0.00007966</td>
</tr>
<tr>
<td>5</td>
<td>0.00607</td>
<td>33.82918</td>
<td>0.09746</td>
<td>0.49660</td>
<td>0.10330</td>
<td>0.06948</td>
<td>0.00561</td>
<td>0.02026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>0.00102</td>
<td>82.63757</td>
<td>0.18981</td>
<td>0.01455</td>
<td>0.06210</td>
<td>0.02283</td>
<td>0.91277</td>
<td>0.98357</td>
<td>0.00568</td>
<td></td>
</tr>
</tbody>
</table>

Model Fit and Diagnostic Statistics

This section gathers the formulas for the statistics available in the MODEL, PLOT, and OUTPUT statements. The model to be fit is \( Y = X\beta + \epsilon \), and the parameter estimate is denoted by \( b = (X'X)^{-1}X'Y \). The subscript \( i \) denotes values for the \( i \)th observation, the parenthetical subscript \( (i) \) means that the statistic is computed by using all observations except the \( i \)th observation, and the subscript \( jj \) indicates the \( j \)th diagonal matrix entry. The ALPHA= option in the PROC REG or MODEL statement is used to set the \( \alpha \) value for the \( t \) statistics.

Table 102.7 contains the summary statistics for assessing the fit of the model.
Table 102.7  Formulas and Definitions for Model Fit Summary Statistics

<table>
<thead>
<tr>
<th>Model Option or Statistic</th>
<th>Definition or Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n$</td>
<td>the number of observations</td>
</tr>
<tr>
<td>$p$</td>
<td>the number of parameters including the intercept</td>
</tr>
<tr>
<td>$i$</td>
<td>1 if there is an intercept, 0 otherwise</td>
</tr>
<tr>
<td>$\hat{\sigma}^2$</td>
<td>the estimate of pure error variance from the SIGMA= option or from fitting the full model</td>
</tr>
<tr>
<td>$\text{SST}_0$</td>
<td>the uncorrected total sum of squares for the dependent variable</td>
</tr>
<tr>
<td>$\text{SST}_1$</td>
<td>the total sum of squares corrected for the mean for the dependent variable</td>
</tr>
<tr>
<td>$\text{SSE}$</td>
<td>the error sum of squares</td>
</tr>
<tr>
<td>$\text{MSE}$</td>
<td>$\frac{\text{SSE}}{n-p}$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>$1 - \frac{\text{SSE}}{\text{SST}_i}$</td>
</tr>
<tr>
<td>$\text{ADJRSQ}$</td>
<td>$1 - \frac{(n-i)(1-R^2)}{n-p}$</td>
</tr>
<tr>
<td>$\text{AIC}$</td>
<td>$n \ln \left( \frac{\text{SSE}}{n} \right) + 2p$</td>
</tr>
<tr>
<td>$\text{BIC}$</td>
<td>$n \ln \left( \frac{\text{SSE}}{n} \right) + 2(p+2)q - 2q^2$ where $q = \frac{n\hat{\sigma}^2}{\text{SSE}}$</td>
</tr>
<tr>
<td>$\text{CP (C_p)}$</td>
<td>$\frac{\text{SSE}}{\hat{\sigma}^2} + 2p - n$</td>
</tr>
<tr>
<td>$\text{GMSEP}$</td>
<td>$\frac{\text{MSE}(n+1)(n-2)}{n(n-p-1)} = \frac{1}{n} S_p (n+1)(n-2)$</td>
</tr>
<tr>
<td>$\text{JP (J_p)}$</td>
<td>$\frac{n}{n+p} \cdot \text{MSE}$</td>
</tr>
<tr>
<td>$\text{PC}$</td>
<td>$\frac{n}{n-p} \cdot (1 - R^2) = J_p \left( \frac{n}{\text{SST}_i} \right)$</td>
</tr>
<tr>
<td>$\text{PRESS}$</td>
<td>the sum of squares of pred$_r_i$ (see Table 102.8)</td>
</tr>
<tr>
<td>$\text{RMSE}$</td>
<td>$\sqrt{\text{MSE}}$</td>
</tr>
<tr>
<td>$\text{SBC}$</td>
<td>$n \ln \left( \frac{\text{SSE}}{n} \right) + p \ln(n)$</td>
</tr>
<tr>
<td>$\text{SP (S_p)}$</td>
<td>$\frac{\text{MSE}}{n-p-1}$</td>
</tr>
</tbody>
</table>

Table 102.8 contains the diagnostic statistics and their formulas; these formulas and further information can be found in Chapter 4, "Introduction to Regression Procedures," and in the section "Influence Statistics" on page 8516. Each statistic is computed for each observation.
Table 102.8  Formulas and Definitions for Diagnostic Statistics

<table>
<thead>
<tr>
<th>MODEL Option or Statistic</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED (Ŷᵢ)</td>
<td>Xᵢb</td>
</tr>
<tr>
<td>RES (rᵢ)</td>
<td>Yᵢ − Ŷᵢ</td>
</tr>
<tr>
<td>H (hᵢ)</td>
<td>xᵢ(X’X)⁻¹xᵢ’</td>
</tr>
<tr>
<td>STDP</td>
<td>√hᵢσ²</td>
</tr>
<tr>
<td>STDI</td>
<td>√(1 + hᵢ)σ²</td>
</tr>
<tr>
<td>STDR</td>
<td>√(1 − hᵢ)σ²</td>
</tr>
<tr>
<td>LCL</td>
<td>Ŷᵢ − tα/2 STDI</td>
</tr>
<tr>
<td>LCLM</td>
<td>Ŷᵢ + tα/2 STDP</td>
</tr>
<tr>
<td>UCL</td>
<td>Ŷᵢ + tα/2 STDI</td>
</tr>
<tr>
<td>UCLM</td>
<td>Ŷᵢ + tα/2 STDP</td>
</tr>
<tr>
<td>STUDENT</td>
<td>rᵢ/STDRᵢ</td>
</tr>
<tr>
<td>RSTUDENT</td>
<td>rᵢ/(δ(i)√(1 − hᵢ))</td>
</tr>
<tr>
<td>COOKD</td>
<td>1 / p * (STUDENT² STDP²) / STDR²</td>
</tr>
<tr>
<td>COVRATIO</td>
<td>det(δ²(X’X)⁻¹) / det(δ²(X’X)⁻¹)</td>
</tr>
<tr>
<td>DFFITS</td>
<td>(Ŷᵢ − Ŷ(i)) / (δ(i)√hᵢ)</td>
</tr>
<tr>
<td>DFBETASᵢ</td>
<td>bⱼ − b(i)ⱼ / δ(i)√(X’X)jj</td>
</tr>
<tr>
<td>PRESS(predᵢ)</td>
<td>rᵢ(1 − hᵢ)</td>
</tr>
</tbody>
</table>

Influence Statistics

This section discusses the INFLUENCE option, which produces several influence statistics, and the PARTIAL option, which produces partial regression leverage plots.

The INFLUENCE Option

The INFLUENCE option (in the MODEL statement) requests the statistics proposed by Belsley, Kuh, and Welsch (1980) to measure the influence of each observation on the estimates. Influential observations are those that, according to various criteria, appear to have a large influence on the parameter estimates.

Let b(i) be the parameter estimates after deleting the ith observation; let s(i)² be the variance estimate after deleting the ith observation; let X(i) be the X matrix without the ith observation; let Ŷ(i) be the ith value
predicted without using the \( i \)th observation; let \( r_i = y_i - \hat{y}_i \) be the \( i \)th residual; and let \( h_i \) be the \( i \)th diagonal of the projection matrix for the predictor space, also called the hat matrix:

\[
h_i = x_i (X'X)^{-1} x_i'
\]

Belsley, Kuh, and Welsch (1980) propose a cutoff of \( 2p/n \), where \( n \) is the number of observations used to fit the model and \( p \) is the number of parameters in the model. Observations with \( h_i \) values above this cutoff should be investigated.

For each observation, PROC REG first displays the residual, the studentized residual (RSTUDENT), and the \( h_i \). The studentized residual RSTUDENT differs slightly from STUDENT since the error variance is estimated by \( s(i)^2 \) without the \( i \)th observation, not by \( s^2 \). For example,

\[
\text{RSTUDENT} = \frac{r_i}{s(i) \sqrt{(1 - h_i)}}
\]

Observations with RSTUDENT larger than 2 in absolute value might need some attention.

The COVRATIO statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the \( i \)th observation:

\[
\text{COVRATIO} = \frac{\det(s^2(i)(X'_i X(i))^{-1})}{\det(s^2(X'X)^{-1})}
\]

Belsley, Kuh, and Welsch (1980) suggest that observations with

\[|\text{COVRATIO} - 1| \geq \frac{3p}{n}\]

where \( p \) is the number of parameters in the model and \( n \) is the number of observations used to fit the model, are worth investigation.

The DFFITS statistic is a scaled measure of the change in the predicted value for the \( i \)th observation and is calculated by deleting the \( i \)th observation. A large value indicates that the observation is very influential in its neighborhood of the \( X \) space.

\[
\text{DFFITS} = \frac{\hat{y}_i - \hat{y}_{(i)}}{s(i) \sqrt{h_{(i)}}}
\]

Large values of DFFITS indicate influential observations. A general cutoff to consider is 2; a size-adjusted cutoff recommended by Belsley, Kuh, and Welsch (1980) is \( 2\sqrt{p/n} \), where \( n \) and \( p \) are as defined previously.

The DFFITS statistic is very similar to Cook’s \( D \), defined in the section “Predicted and Residual Values” on page 8507.
The DFBETAS statistics are the scaled measures of the change in each parameter estimate and are calculated by deleting the $i$th observation:

$$\text{DFBETAS}_j = \frac{b_j - b_{(i)j}}{s(i) \sqrt{(X'X)_{jj}}}$$

where $(X'X)_{jj}$ is the $(j, j)$ element of $(X'X)^{-1}$.

In general, large values of DFBETAS indicate observations that are influential in estimating a given parameter. Belsley, Kuh, and Welsch (1980) recommend 2 as a general cutoff value to indicate influential observations and $2/s\sqrt{n}$ as a size-adjusted cutoff.

The following statements use a subset of the data in the population example in the section “Polynomial Regression” on page 8436. The INFLUENCE option produces the tables shown in Figure 102.38 and Figure 102.39.

```
proc reg data=USPopulation;
  where Year <= 1970;
  model Population=Year YearSq / influence;
run;
```

**Figure 102.38** Regression Using the INFLUENCE Option

<table>
<thead>
<tr>
<th></th>
<th>Obs</th>
<th>Residual</th>
<th>RStudent</th>
<th>Hat Diag</th>
<th>H</th>
<th>Cov Ratio</th>
<th>DFFITS</th>
<th>Intercept</th>
<th>Year</th>
<th>YearSq</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>-1.1094</td>
<td>-0.4972</td>
<td>0.3865</td>
<td>1.8834</td>
<td>0.3946</td>
<td>-0.2842</td>
<td>0.2810</td>
<td>-0.2779</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>0.2691</td>
<td>0.1082</td>
<td>0.2501</td>
<td>1.6147</td>
<td>0.0625</td>
<td>0.0376</td>
<td>0.0370</td>
<td>0.0365</td>
<td></td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0.9305</td>
<td>0.3561</td>
<td>0.1652</td>
<td>1.4176</td>
<td>0.1584</td>
<td>0.0666</td>
<td>0.0651</td>
<td>0.0636</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.7908</td>
<td>0.2941</td>
<td>0.1184</td>
<td>1.3531</td>
<td>0.1078</td>
<td>0.0182</td>
<td>-0.0172</td>
<td>0.0161</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.2110</td>
<td>0.0774</td>
<td>0.0983</td>
<td>1.3444</td>
<td>0.0256</td>
<td>-0.0030</td>
<td>0.0033</td>
<td>-0.0035</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>-0.6629</td>
<td>-0.2431</td>
<td>0.0951</td>
<td>1.3255</td>
<td>0.0788</td>
<td>0.0296</td>
<td>-0.0302</td>
<td>0.0307</td>
<td></td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>-0.8869</td>
<td>-0.3268</td>
<td>0.1009</td>
<td>1.3214</td>
<td>-0.1095</td>
<td>0.0609</td>
<td>-0.0616</td>
<td>0.0621</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>-0.2501</td>
<td>-0.0923</td>
<td>0.1095</td>
<td>1.3605</td>
<td>0.0324</td>
<td>0.0216</td>
<td>-0.0217</td>
<td>0.0218</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>-0.7593</td>
<td>-0.2820</td>
<td>0.1164</td>
<td>1.3519</td>
<td>-0.1023</td>
<td>0.0743</td>
<td>-0.0745</td>
<td>0.0747</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>-0.5757</td>
<td>-0.2139</td>
<td>0.1190</td>
<td>1.3650</td>
<td>-0.0786</td>
<td>0.0586</td>
<td>-0.0587</td>
<td>0.0587</td>
<td></td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>0.7938</td>
<td>0.2949</td>
<td>0.1164</td>
<td>1.3499</td>
<td>0.0170</td>
<td>-0.0784</td>
<td>0.0783</td>
<td>-0.0781</td>
<td></td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>1.1492</td>
<td>0.4265</td>
<td>0.1095</td>
<td>1.3144</td>
<td>0.1496</td>
<td>-0.1018</td>
<td>0.1014</td>
<td>-0.1009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>3.1664</td>
<td>1.2189</td>
<td>0.1009</td>
<td>1.0168</td>
<td>0.4084</td>
<td>-0.2357</td>
<td>0.2338</td>
<td>-0.2318</td>
<td></td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>1.6746</td>
<td>0.6207</td>
<td>0.0951</td>
<td>1.2430</td>
<td>0.2013</td>
<td>-0.0811</td>
<td>0.0798</td>
<td>-0.0784</td>
<td></td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>2.2406</td>
<td>0.8407</td>
<td>0.0983</td>
<td>1.1724</td>
<td>0.2776</td>
<td>-0.0427</td>
<td>0.0404</td>
<td>-0.0380</td>
<td></td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>-6.6335</td>
<td>-3.1845</td>
<td>0.1184</td>
<td>0.2924</td>
<td>-1.1673</td>
<td>-0.1531</td>
<td>0.1636</td>
<td>-0.1747</td>
<td></td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>-6.0147</td>
<td>-2.8433</td>
<td>0.1652</td>
<td>0.3989</td>
<td>-1.2649</td>
<td>-0.4843</td>
<td>0.4958</td>
<td>-0.5076</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18</td>
<td>1.6770</td>
<td>0.6847</td>
<td>0.2501</td>
<td>1.4757</td>
<td>0.3954</td>
<td>0.2240</td>
<td>-0.2274</td>
<td>0.2308</td>
<td></td>
</tr>
<tr>
<td></td>
<td>19</td>
<td>3.9895</td>
<td>1.9947</td>
<td>0.3865</td>
<td>0.9766</td>
<td>1.5831</td>
<td>1.0902</td>
<td>-1.1025</td>
<td>1.1151</td>
<td></td>
</tr>
</tbody>
</table>
In Figure 102.38, observations 16, 17, and 19 exceed or are near the cutoff value of 2 for RSTUDENT. None of the observations exceeds the general cutoff of 2 for DFFITS or the DFBETAS, but observations 16, 17, and 19 exceed at least one of the size-adjusted cutoffs for these statistics. Observations 1 and 19 exceed the cutoff for the hat diagonals, and observations 1, 2, 16, 17, and 18 exceed the cutoffs for COVRATIO. Taken together, these statistics indicate that you should look first at observations 16, 17, and 19 and then perhaps investigate the other observations that exceeded a cutoff.

When ODS Graphics is enabled, you can request influence diagnostic plots by using the PLOTS= option in the PROC REG statement as shown in the following statements:

```sas
ods graphics on;
proc reg data=USPopulation
   plots(label)=(CooksD RStudentByLeverage DFFITS DFBETAS);
   where Year <= 1970;
   id Year;
   model Population=Year YearSq;
run;
```

The LABEL suboption specified in the PLOTS(LABEL)= option requests that observations that exceed the relevant cutoffs for the statistics being plotted are labeled. Since `Year` has been named in an ID statement, the value of `Year` is used for the labels. The requested plots are shown in Figure 102.40.
Chapter 102: The REG Procedure

Figure 102.40 Influence Diagnostics

Cook's D for Population

Outlier and Leverage Diagnostics for Population

Legend:
- Outlier
- Leverage
- Outlier and Leverage
Figure 102.40 continued
When ODS Graphics is enabled, you can request a chart of the studentized residuals and Cook’s D by using the PLOTS=RESIDUALCHART option as shown in the following statements:

```plaintext
proc reg data=USPopulation plots(only)=residualchart;
  where Year <= 1970;
  id Year;
  model Population=Year YearSq;
run;
```

The chart is shown in Figure 102.41.

**Figure 102.41** Studentized Residuals and Cook’s D Chart

---

**The PARTIAL and PARTIALDATA Options**

The PARTIAL option in the MODEL statement produces partial regression leverage plots. If ODS Graphics is not enabled, this option requires the use of the LINEPRINTER option in the PROC REG statement. One plot is created for each regressor in the current full model. For example, plots are produced for regressors included by using ADD statements; plots are not produced for interim models in the various model-selection methods but only for the full model. If you use a model-selection method and the final model contains only a subset of the original regressors, the PARTIAL option still produces plots for all regressors in the full model. If ODS Graphics is enabled, these plots are produced as high-resolution graphics, in panels with a maximum of six partial regression leverage plots per panel. Multiple panels are displayed for models with more than six regressors.

For a given regressor, the partial regression leverage plot is the plot of the dependent variable and the regressor after they have been made orthogonal to the other regressors in the model. These can be obtained by plotting the residuals for the dependent variable against the residuals for the selected regressor, where the residuals for the dependent variable are calculated with the selected regressor omitted, and the residuals for the selected...
regressor are calculated from a model where the selected regressor is regressed on the remaining regressors. A line fit to the points has a slope equal to the parameter estimate in the full model.

When ODS Graphics is not enabled, points in the plot are marked by the number of replicates appearing at one position. The symbol ‘*’ is used if there are 10 or more replicates. If an ID statement is specified, the leftmost nonblank character in the value of the ID variable is used as the plotting symbol.

The PARTIALDATA option in the MODEL statement produces a table that contains the partial regression data that are displayed in the partial regression leverage plots. You can request partial regression data even if you do not requests plots with the PARTIAL option.

The following statements use the fitness data in Example 102.2 with the PARTIAL option and ODS Graphics to produce the partial regression leverage plots. The plots are shown in Figure 102.42.

```plaintext
ods graphics on;
proc reg data=fitness;
   model Oxygen=RunTime Weight Age / partial;
run;
ods graphics off;
```

**Figure 102.42** Partial Regression Leverage Plots

![Partial Regression Leverage Plots](image-url)
Reweighting Observations in an Analysis

Reweighting observations is an interactive feature of PROC REG that enables you to change the weights of observations used in computing the regression equation. Observations can also be deleted from the analysis (not from the data set) by changing their weights to zero. In the following statements, the Sashelp.Class data are used to illustrate some of the features of the REWEIGHT statement. First, the full model is fit, and the residuals are displayed in Figure 102.43.

```
proc reg data=sashelp.Class;
  model Weight=Age Height / p;
  id Name;
run;
```

**Figure 102.43** Full Model for Sashelp.Class Data, Residuals Shown

<table>
<thead>
<tr>
<th>Obs</th>
<th>Name</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alfred</td>
<td>112.5</td>
<td>124.8686</td>
<td>-12.3686</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>84.0</td>
<td>78.6273</td>
<td>5.3727</td>
</tr>
<tr>
<td>3</td>
<td>Barbara</td>
<td>98.0</td>
<td>110.2812</td>
<td>-12.2812</td>
</tr>
<tr>
<td>4</td>
<td>Carol</td>
<td>102.5</td>
<td>102.5670</td>
<td>-0.0670</td>
</tr>
<tr>
<td>5</td>
<td>Henry</td>
<td>102.5</td>
<td>105.0849</td>
<td>-2.5849</td>
</tr>
<tr>
<td>6</td>
<td>James</td>
<td>83.0</td>
<td>80.2266</td>
<td>2.7734</td>
</tr>
<tr>
<td>7</td>
<td>Jane</td>
<td>84.5</td>
<td>89.2191</td>
<td>-4.7191</td>
</tr>
<tr>
<td>8</td>
<td>Janet</td>
<td>112.5</td>
<td>102.7663</td>
<td>9.7337</td>
</tr>
<tr>
<td>9</td>
<td>Jeffrey</td>
<td>84.0</td>
<td>100.2095</td>
<td>-16.2095</td>
</tr>
<tr>
<td>10</td>
<td>John</td>
<td>99.5</td>
<td>86.3415</td>
<td>13.1585</td>
</tr>
<tr>
<td>11</td>
<td>Joyce</td>
<td>50.5</td>
<td>57.3660</td>
<td>-6.8660</td>
</tr>
<tr>
<td>12</td>
<td>Judy</td>
<td>90.0</td>
<td>107.9625</td>
<td>-17.9625</td>
</tr>
<tr>
<td>13</td>
<td>Louise</td>
<td>77.0</td>
<td>76.6295</td>
<td>0.3705</td>
</tr>
<tr>
<td>14</td>
<td>Mary</td>
<td>112.0</td>
<td>117.1544</td>
<td>-5.1544</td>
</tr>
<tr>
<td>15</td>
<td>Philip</td>
<td>150.0</td>
<td>138.2164</td>
<td>11.7836</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>128.0</td>
<td>107.2043</td>
<td>20.7957</td>
</tr>
<tr>
<td>17</td>
<td>Ronald</td>
<td>133.0</td>
<td>118.9529</td>
<td>14.0471</td>
</tr>
<tr>
<td>18</td>
<td>Thomas</td>
<td>85.0</td>
<td>79.6676</td>
<td>5.3324</td>
</tr>
<tr>
<td>19</td>
<td>William</td>
<td>112.0</td>
<td>117.1544</td>
<td>-5.1544</td>
</tr>
</tbody>
</table>

**Output Statistics**

- Sum of Residuals: 0
- Sum of Squared Residuals: 2120.09974
- Predicted Residual SS (PRESS): 3272.72186
Upon examining the data and residuals, you realize that observation 17 (Ronald) was mistakenly included in the analysis. Also, you would like to examine the effect of reweighting to 0.5 those observations with residuals that have absolute values greater than or equal to 17. The following statements show how you request this reweighting:

```
reweight obs.=17;
reweight r. le -17 or r. ge 17 / weight=0.5;
print p;
run;
```

At this point, a message appears (in the log) that tells you which observations have been reweighted and what the new weights are. Figure 102.44 is produced.

**Figure 102.44** Model with Reweighted Observations

The REG Procedure
Model: MODEL1.2
Dependent Variable: Weight

<table>
<thead>
<tr>
<th>Output Statistics</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>Obs</td>
<td>Name</td>
<td>Weight</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>Alfred</td>
<td>1.0</td>
<td>112.5</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>1.0</td>
<td>84.0</td>
</tr>
<tr>
<td>3</td>
<td>Barbara</td>
<td>1.0</td>
<td>98.0</td>
</tr>
<tr>
<td>4</td>
<td>Carol</td>
<td>1.0</td>
<td>102.5</td>
</tr>
<tr>
<td>5</td>
<td>Henry</td>
<td>1.0</td>
<td>102.5</td>
</tr>
<tr>
<td>6</td>
<td>James</td>
<td>1.0</td>
<td>83.0</td>
</tr>
<tr>
<td>7</td>
<td>Jane</td>
<td>1.0</td>
<td>84.5</td>
</tr>
<tr>
<td>8</td>
<td>Janet</td>
<td>1.0</td>
<td>112.5</td>
</tr>
<tr>
<td>9</td>
<td>Jeffrey</td>
<td>1.0</td>
<td>84.0</td>
</tr>
<tr>
<td>10</td>
<td>John</td>
<td>1.0</td>
<td>99.5</td>
</tr>
<tr>
<td>11</td>
<td>Joyce</td>
<td>1.0</td>
<td>50.5</td>
</tr>
<tr>
<td>12</td>
<td>Judy</td>
<td>0.5</td>
<td>90.0</td>
</tr>
<tr>
<td>13</td>
<td>Louise</td>
<td>1.0</td>
<td>77.0</td>
</tr>
<tr>
<td>14</td>
<td>Mary</td>
<td>1.0</td>
<td>112.0</td>
</tr>
<tr>
<td>15</td>
<td>Philip</td>
<td>1.0</td>
<td>150.0</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>0.5</td>
<td>128.0</td>
</tr>
<tr>
<td>17</td>
<td>Ronald</td>
<td>0.0</td>
<td>133.0</td>
</tr>
<tr>
<td>18</td>
<td>Thomas</td>
<td>1.0</td>
<td>85.0</td>
</tr>
<tr>
<td>19</td>
<td>William</td>
<td>1.0</td>
<td>112.0</td>
</tr>
</tbody>
</table>

| Sum of Residuals | 0 |
| Sum of Squared Residuals | 1500.61194 |
| Predicted Residual SS (PRESS) | 2287.57621 |

The first REWEIGHT statement excludes observation 17, and the second REWEIGHT statement reweights observations 12 and 16 to 0.5. An important feature to note from this example is that the model is not refit until after the PRINT statement. REWEIGHT statements do not cause the model to be refit. This is so that multiple REWEIGHT statements can be applied to a subsequent model.
In this example, since the intent is to reweight observations with large residuals, the observation that was mistakenly included in the analysis should be deleted; then the model should be fit for those remaining observations, and the observations with large residuals should be reweighted. To accomplish this, use the REFIT statement. Note that the model label has been changed from MODEL1 to MODEL1.2 since two REWEIGHT statements have been used. The following statements produce Figure 102.45:

```
reweight allobs / weight=1.0;
reweight obs.=17;
refit;
reweight r. le -17 or r. ge 17 / weight=.5;
print;
run;
```

**Figure 102.45** Observations Excluded from Analysis, Model Refitted, and Observations Reweighted

<table>
<thead>
<tr>
<th>Obs</th>
<th>Name</th>
<th>Weight</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alfred</td>
<td>1.0</td>
<td>112.5</td>
<td>120.9716</td>
<td>-8.4716</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>1.0</td>
<td>84.0</td>
<td>79.5342</td>
<td>4.4658</td>
</tr>
<tr>
<td>3</td>
<td>Barbara</td>
<td>1.0</td>
<td>98.0</td>
<td>107.0746</td>
<td>-9.0746</td>
</tr>
<tr>
<td>4</td>
<td>Carol</td>
<td>1.0</td>
<td>102.5</td>
<td>101.5681</td>
<td>0.9319</td>
</tr>
<tr>
<td>5</td>
<td>Henry</td>
<td>1.0</td>
<td>102.5</td>
<td>103.7588</td>
<td>-1.2588</td>
</tr>
<tr>
<td>6</td>
<td>James</td>
<td>1.0</td>
<td>83.0</td>
<td>79.7204</td>
<td>3.2796</td>
</tr>
<tr>
<td>7</td>
<td>Jane</td>
<td>1.0</td>
<td>84.5</td>
<td>85.5443</td>
<td>-0.0443</td>
</tr>
<tr>
<td>8</td>
<td>Janet</td>
<td>1.0</td>
<td>112.5</td>
<td>102.9467</td>
<td>9.5533</td>
</tr>
<tr>
<td>9</td>
<td>Jeffrey</td>
<td>1.0</td>
<td>84.0</td>
<td>98.3117</td>
<td>-14.3117</td>
</tr>
<tr>
<td>10</td>
<td>John</td>
<td>1.0</td>
<td>99.5</td>
<td>85.0407</td>
<td>14.4593</td>
</tr>
<tr>
<td>11</td>
<td>Joyce</td>
<td>1.0</td>
<td>50.5</td>
<td>58.6253</td>
<td>-8.1253</td>
</tr>
<tr>
<td>12</td>
<td>Judy</td>
<td>1.0</td>
<td>90.0</td>
<td>106.2625</td>
<td>-16.2625</td>
</tr>
<tr>
<td>13</td>
<td>Louise</td>
<td>1.0</td>
<td>77.0</td>
<td>76.5908</td>
<td>0.4092</td>
</tr>
<tr>
<td>14</td>
<td>Mary</td>
<td>1.0</td>
<td>112.0</td>
<td>115.4651</td>
<td>-3.4651</td>
</tr>
<tr>
<td>15</td>
<td>Philip</td>
<td>1.0</td>
<td>150.0</td>
<td>134.9953</td>
<td>15.0047</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>0.5</td>
<td>128.0</td>
<td>103.1923</td>
<td>24.8077</td>
</tr>
<tr>
<td>17</td>
<td>Ronald</td>
<td>0.0</td>
<td>133.0</td>
<td>117.0299</td>
<td>15.9701</td>
</tr>
<tr>
<td>18</td>
<td>Thomas</td>
<td>1.0</td>
<td>85.0</td>
<td>78.0288</td>
<td>6.9712</td>
</tr>
<tr>
<td>19</td>
<td>William</td>
<td>1.0</td>
<td>112.0</td>
<td>115.4651</td>
<td>-3.4651</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Output Statistics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Residuals</td>
<td>0</td>
</tr>
<tr>
<td>Sum of Squared Residuals</td>
<td>1637.81879</td>
</tr>
<tr>
<td>Predicted Residual SS (PRESS)</td>
<td>2473.87984</td>
</tr>
</tbody>
</table>

Notice that this results in a slightly different model than the previous set of statements: only observation 16 is reweighted to 0.5. Also note that the model label is now MODEL1.5 since five REWEIGHT statements have been used for this model.
Another important feature of the REWEIGHT statement is the ability to nullify the effect of a previous or all REWEIGHT statements. First, assume that you have several REWEIGHT statements in effect and you want to restore the original weights of all the observations. The following REWEIGHT statement accomplishes this and produces Figure 102.46:

```plaintext
reweight allobs / reset;
print;
run;
```

**Figure 102.46** Restoring Weights of All Observations

The resulting model is identical to the original model specified at the beginning of this section. Notice that the model label is now MODEL1.6. Note that the Weight column does not appear, since all observations have been reweighted to have weight=1.

Now suppose you want only to undo the changes made by the most recent REWEIGHT statement. Use REWEIGHT UNDO for this. The following statements produce Figure 102.47:
```plaintext
test
```
Reweighting Observations in an Analysis

```itespace
reweight r. le -12 or r. ge 12 / weight=.75;
reweight r. le -17 or r. ge 17 / weight=.5;
reweight / reset;
print;
run;
```

**Figure 102.48** REWEIGHT Statement with RESET option

**The REG Procedure**
**Model: MODEL1.12**
**Dependent Variable: Weight**

<table>
<thead>
<tr>
<th>Obs</th>
<th>Name</th>
<th>Weight</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alfred</td>
<td>0.75</td>
<td>0.75</td>
<td>112.5</td>
<td>126.0076</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>1.00</td>
<td>1.00</td>
<td>84.0</td>
<td>77.8727</td>
</tr>
<tr>
<td>3</td>
<td>Barbara</td>
<td>0.75</td>
<td>0.75</td>
<td>98.0</td>
<td>111.2805</td>
</tr>
<tr>
<td>4</td>
<td>Carol</td>
<td>1.00</td>
<td>1.00</td>
<td>102.5</td>
<td>102.4703</td>
</tr>
<tr>
<td>5</td>
<td>Henry</td>
<td>1.00</td>
<td>1.00</td>
<td>102.5</td>
<td>105.1278</td>
</tr>
<tr>
<td>6</td>
<td>James</td>
<td>1.00</td>
<td>1.00</td>
<td>83.0</td>
<td>80.2290</td>
</tr>
<tr>
<td>7</td>
<td>Jane</td>
<td>1.00</td>
<td>1.00</td>
<td>84.5</td>
<td>89.7199</td>
</tr>
<tr>
<td>8</td>
<td>Janet</td>
<td>1.00</td>
<td>1.00</td>
<td>112.5</td>
<td>102.0122</td>
</tr>
<tr>
<td>9</td>
<td>Jeffrey</td>
<td>0.75</td>
<td>0.75</td>
<td>84.0</td>
<td>100.6507</td>
</tr>
<tr>
<td>10</td>
<td>John</td>
<td>0.75</td>
<td>0.75</td>
<td>99.5</td>
<td>86.6828</td>
</tr>
<tr>
<td>11</td>
<td>Joyce</td>
<td>1.00</td>
<td>1.00</td>
<td>50.5</td>
<td>56.7703</td>
</tr>
<tr>
<td>12</td>
<td>Judy</td>
<td>1.00</td>
<td>1.00</td>
<td>90.0</td>
<td>108.1649</td>
</tr>
<tr>
<td>13</td>
<td>Louise</td>
<td>1.00</td>
<td>1.00</td>
<td>77.0</td>
<td>76.4327</td>
</tr>
<tr>
<td>14</td>
<td>Mary</td>
<td>1.00</td>
<td>1.00</td>
<td>112.0</td>
<td>117.1975</td>
</tr>
<tr>
<td>15</td>
<td>Philip</td>
<td>1.00</td>
<td>1.00</td>
<td>150.0</td>
<td>138.7581</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>1.00</td>
<td>1.00</td>
<td>128.0</td>
<td>108.7016</td>
</tr>
<tr>
<td>17</td>
<td>Ronald</td>
<td>0.75</td>
<td>0.75</td>
<td>133.0</td>
<td>119.0957</td>
</tr>
<tr>
<td>18</td>
<td>Thomas</td>
<td>1.00</td>
<td>1.00</td>
<td>85.0</td>
<td>80.3076</td>
</tr>
<tr>
<td>19</td>
<td>William</td>
<td>1.00</td>
<td>1.00</td>
<td>112.0</td>
<td>117.1975</td>
</tr>
</tbody>
</table>

Sum of Residuals: 0
Sum of Squared Residuals: 1879.08980
Predicted Residual SS (PRESS): 2959.57279

Note that observations that meet the condition of the second REWEIGHT statement (residuals with an absolute value greater than or equal to 17) now have weights reset to their original value of 1. Observations 1, 3, 9, 10, and 17 have weights of 0.75, but observations 12 and 16 (which meet the condition of the second REWEIGHT statement) have their weights reset to 1.

Notice how the last three examples show three ways to change weights back to a previous value. In the first example, ALLOBS and the RESET option are used to change weights for all observations back to their original values. In the second example, the UNDO option is used to negate the effect of a previous REWEIGHT statement, thus changing weights for observations selected in the previous REWEIGHT statement to the weights specified in still another REWEIGHT statement. In the third example, the RESET option is used to change weights for observations selected in a previous REWEIGHT statement back to their
original values. Finally, note that the label MODEL1.12 indicates that 12 REWEIGHT statements have been applied to the original model.

Testing for Heteroscedasticity

The regression model is specified as $y_i = x_i \beta + \epsilon_i$, where the $\epsilon_i$'s are identically and independently distributed: $E(\epsilon) = 0$ and $E(\epsilon'\epsilon) = \sigma^2 I$. If the $\epsilon_i$'s are not independent or their variances are not constant, the parameter estimates are unbiased, but the estimate of the covariance matrix is inconsistent.

In the case of heteroscedasticity, if the regression data are from a simple random sample, then White (1980), showed that matrix

$$ HC_0 = (X'X)^{-1} (X' \text{diag}(e_i^2)X)(X'X)^{-1} $$

where

$$ e_i = y_i - x_i b $$

is an asymptotically consistent estimate of the covariance matrix. MacKinnon and White (1985) introduced three alternative heteroscedasticity-consistent covariance matrix estimators that are all asymptotically equivalent to the estimator $HC_0$ but that typically have better small sample behavior. These estimators labeled $HC_1$, $HC_2$, and $HC_3$ are defined as follows:

$$ HC_1 = \frac{n}{n - p} HC_0 $$

where $n$ is the number of observations and $p$ is the number of regressors including the intercept.

$$ HC_2 = (X'X)^{-1} X' \text{diag}(\frac{e_i^2}{1 - h_{ii}})X(X'X)^{-1} $$

where

$$ h_{ii} = x_i (X'X)^{-1} x'_i $$

is the leverage of the $i$th observation.

$$ HC_3 = (X'X)^{-1} X' \text{diag}(\frac{e_i^2}{(1 - h_{ii})^2})X(X'X)^{-1} $$

Long and Ervin (2000) studied the performance of these estimators and recommend using the $HC_3$ estimator if the sample size is less than 250.
You can use the HCCMETHOD=0,1,2, or 3 in the MODEL statement to select a heteroscedasticity-consistent covariance matrix estimator, with HC0 being the default. The ACOV option in the MODEL statement displays the heteroscedasticity-consistent covariance matrix estimator in effect and adds heteroscedasticity-consistent standard errors, also known as White standard errors, to the parameter estimates table. If you specify the HCC or WHITE option in the MODEL statement, but do not also specify the ACOV option, then the heteroscedasticity-consistent standard errors are added to the parameter estimates table but the heteroscedasticity-consistent covariance matrix is not displayed.

The SPEC option performs a model specification test. The null hypothesis for this test maintains that the errors are homoscedastic and independent of the regressors and that several technical assumptions about the model specification are valid. For details, see theorem 2 and assumptions 1–7 of White (1980). When the model is correctly specified and the errors are independent of the regressors, the rejection of this null hypothesis is evidence of heteroscedasticity. In implementing this test, an estimator of the average covariance matrix (White 1980, p. 822) is constructed and inverted. The nonsingularity of this matrix is one of the assumptions in the null hypothesis about the model specification. When PROC REG determines this matrix to be numerically singular, a generalized inverse is used and a note to this effect is written to the log. In such cases, care should be taken in interpreting the results of this test.

When you specify the SPEC, ACOV, HCC, or WHITE option in the MODEL statement, tests listed in the TEST statement are performed with both the usual covariance matrix and the heteroscedasticity-consistent covariance matrix requested with the HCCMETHOD= option. Tests performed with the consistent covariance matrix are asymptotic. For more information, see White (1980).

Both the ACOV and SPEC options can be specified in a MODEL or PRINT statement.

Testing for Lack of Fit

The test for lack of fit compares the variation around the model with “pure” variation within replicated observations. This measures the adequacy of the specified model. In particular, if there are \( n_i \) replicated observations \( Y_{i1}, \ldots, Y_{in_i} \) of the response all at the same values \( x_i \) of the regressors, then you can predict the true response at \( x_i \) either by using the predicted value \( \hat{Y}_i \) based on the model or by using the mean \( \bar{Y}_i \) of the replicated values. The test for lack of fit decomposes the residual error into a component due to the variation of the replications around their mean value (the “pure” error) and a component due to the variation of the mean values around the model prediction (the “bias” error):

\[
\sum_{i} \sum_{j=1}^{n_i} (Y_{ij} - \hat{Y}_i)^2 = \sum_{i} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i} n_i (\bar{Y}_i - \hat{Y}_i)^2
\]

If the model is adequate, then both components estimate the nominal level of error; however, if the bias component of error is much larger than the pure error, then this constitutes evidence that there is significant lack of fit.

If some observations in your design are replicated, you can test for lack of fit by specifying the LACKFIT option in the MODEL statement (see Example 102.6). Note that, since all other tests use total error rather than pure error, you might want to hand-calculate the tests with respect to pure error if the lack of fit is significant. On the other hand, significant lack of fit indicates that the specified model is inadequate, so if this is a problem you can also try to refine the model.
Chapter 102: The REG Procedure

Multivariate Tests

The MTEST statement described in the section “MTEST Statement” on page 8477 can test hypotheses involving several dependent variables in the form

\[(L\beta - cj)M = 0\]

where \(L\) is a linear function on the regressor side, \(\beta\) is a matrix of parameters, \(c\) is a column vector of constants, \(j\) is a row vector of ones, and \(M\) is a linear function on the dependent side. The special case where the constants are zero is

\[L\beta M = 0\]

To test this hypothesis, PROC REG constructs two matrices called \(H\) and \(E\) that correspond to the numerator and denominator of a univariate \(F\) test:

\[H = M'(LB - cj)'(L(X'X)^{-1}L')^{-1}(LB - cj)M\]
\[E = M'(Y'Y - B'(X'X)B)M\]

These matrices are displayed for each MTEST statement if the PRINT option is specified.

Four test statistics based on the eigenvalues of \(E^{-1}H\) or \((E + H)^{-1}H\) are formed. These are Wilks’ lambda, Pillai’s trace, the Hotelling-Lawley trace, and Roy’s greatest root. These test statistics are discussed in Chapter 4, “Introduction to Regression Procedures.”

The following code creates MANOVA data from Morrison (1976):

```sas
* Manova Data from Morrison (1976, 190);
data a;
  input sex $ drug $ @;
do rep=1 to 4;
  input y1 y2 @;
  sexcode=(sex='m')-(sex='f');
drug1=(drug='a')-(drug='c');
drug2=(drug='b')-(drug='c');
  sexdrug1=sexcode*drug1;
  sexdrug2=sexcode*drug2;
  output;
end;
datalines;

m a 5 6 5 4 9 9 7 6
m b 7 6 7 7 9 12 6 8
m c 21 15 14 11 17 12 12 10
```
The following statements perform a multivariate analysis of variance and produce Figure 102.49 through Figure 102.52:

```
proc reg;
  model y1 y2=sexcode drug1 drug2 sexdrug1 sexdrug2;
  y1y2drug: mtest y1=y2, drug1,drug2;
  drugshow: mtest drug1, drug2 / print canprint;
run;
```

**Figure 102.49**  Multivariate Analysis of Variance: REG Procedure

```
The REG Procedure
Model: MODEL1
Dependent Variable: y1

              Source DF    Sum of Squares     Mean Square    F Value    Pr > F
Model          5       316.00000       63.200000       12.04     <.0001
Error         18       94.50000        5.250000
Corrected Total 23       410.50000

              Dependent Mean     Root MSE    R-Square    Adj R-Sq    Coeff Var
9.75000          2.29129     0.76980     0.70580     23.50039
```

**Figure 102.50**  Multivariate Analysis of Variance: REG Procedure

```
The REG Procedure
Model: MODEL1
Dependent Variable: y1

              Source DF    Sum of Squares     Mean Square    F Value    Pr > F
Model          5       69.33333       13.86667        2.19      0.1008
Error         18       114.00000       6.33333
Corrected Total 23       183.33333

              Dependent Mean     Root MSE    R-Square    Adj R-Sq    Coeff Var
8.66667           2.51661     0.37820     0.20550     29.03782
```
Figure 102.50 continued

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|--------|-------|-----|
| Intercept| 1  | 8.66667            | 0.51370        | 16.87  | <.0001|
| sexcode  | 1  | 0.16667            | 0.51370        | 0.32   | 0.7493|
| drug1    | 1  | -1.41667           | 0.72648        | -1.95  | 0.0669|
| drug2    | 1  | -0.16667           | 0.72648        | -0.23  | 0.8211|
| sexdrug1 | 1  | -1.16667           | 0.72648        | -1.61  | 0.1257|
| sexdrug2 | 1  | -0.41667           | 0.72648        | -0.57  | 0.5734|

Figure 102.51 Multivariate Analysis of Variance: First Test

The REG Procedure
Model: MODEL1
Multivariate Test: y1y2drug

Multivariate Statistics and Exact F Statistics
S=1    M=0    N=8

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Value</th>
<th>F Value</th>
<th>Num DF</th>
<th>Den DF</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wilks’ Lambda</td>
<td>0.2805</td>
<td>23.08</td>
<td>2</td>
<td>18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pillai’s Trace</td>
<td>0.7194</td>
<td>23.08</td>
<td>2</td>
<td>18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>2.5645</td>
<td>23.08</td>
<td>2</td>
<td>18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Roy’s Greatest Root</td>
<td>2.5645</td>
<td>23.08</td>
<td>2</td>
<td>18</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

The four multivariate test statistics are all highly significant, giving strong evidence that the coefficients of drug1 and drug2 are not the same across dependent variables y1 and y2.

Figure 102.52 Multivariate Analysis of Variance: Second Test

The REG Procedure
Model: MODEL1
Multivariate Test: drugshow

<table>
<thead>
<tr>
<th>Error Matrix (E)</th>
</tr>
</thead>
<tbody>
<tr>
<td>94.5  76.5</td>
</tr>
<tr>
<td>76.5  114</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Hypothesis Matrix (H)</th>
</tr>
</thead>
<tbody>
<tr>
<td>301  97.5</td>
</tr>
<tr>
<td>97.5  36.33333333333333</td>
</tr>
</tbody>
</table>
Autocorrelation in Time Series Data

When regression is performed on time series data, the errors might not be independent. Often errors are autocorrelated; that is, each error is correlated with the error immediately before it. Autocorrelation is also a symptom of systematic lack of fit. The DW option provides the Durbin-Watson $d$ statistic to test that the autocorrelation is zero:

$$d = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

The value of $d$ is close to 2 if the errors are uncorrelated. The distribution of $d$ is reported by Durbin and Watson (1951). Tables of the distribution are found in most econometrics textbooks, such as Johnston (1972) and Pindyck and Rubinfeld (1981).

The sample autocorrelation estimate is displayed after the Durbin-Watson statistic. The sample is computed as

$$r = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}$$
This autocorrelation of the residuals might not be a very good estimate of the autocorrelation of the true errors, especially if there are few observations and the independent variables have certain patterns. If there are missing observations in the regression, these measures are computed as though the missing observations did not exist.

Positive autocorrelation of the errors generally tends to make the estimate of the error variance too small, so confidence intervals are too narrow and true null hypotheses are rejected with a higher probability than the stated significance level. Negative autocorrelation of the errors generally tends to make the estimate of the error variance too large, so confidence intervals are too wide and the power of significance tests is reduced. With either positive or negative autocorrelation, least squares parameter estimates are usually not as efficient as generalized least squares parameter estimates. For more details, see Judge et al. (1985, Chapter 8) and the SAS/ETS User’s Guide.

The following SAS statements request the DWPROB option for the U.S. population data (see Figure 102.53). If you use the DW option instead of the DWPROB option, then p-values are not produced.

```sas
proc reg data=USPopulation;
   model Population=Year YearSq / dwProb;
run;
```

**Figure 102.53** Regression Using DW Option

<table>
<thead>
<tr>
<th>The REG Procedure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model: MODEL1</td>
</tr>
<tr>
<td>Dependent Variable: Population</td>
</tr>
<tr>
<td>----------------------</td>
</tr>
<tr>
<td>Durbin-Watson D 1.191</td>
</tr>
<tr>
<td>Pr &lt; DW 0.0050</td>
</tr>
<tr>
<td>Pr &gt; DW 0.9950</td>
</tr>
<tr>
<td>Number of Observations 22</td>
</tr>
<tr>
<td>1st Order Autocorrelation 0.323</td>
</tr>
</tbody>
</table>

**Computations for Ridge Regression and IPC Analysis**

In ridge regression analysis, the crossproduct matrix for the independent variables is centered (the NOINT option is ignored if it is specified) and scaled to one on the diagonal elements. The ridge constant \( k \) (specified with the RIDGE= option) is then added to each diagonal element of the crossproduct matrix. The ridge regression estimates are the least squares estimates obtained by using the new crossproduct matrix.

Let \( X \) be an \( n \times p \) matrix of the independent variables after centering the data, and let \( Y \) be an \( n \times 1 \) vector corresponding to the dependent variable. Let \( D \) be a \( p \times p \) diagonal matrix with diagonal elements as in \( X'X \). The ridge regression estimate corresponding to the ridge constant \( k \) can be computed as

\[
D^{-\frac{1}{2}}(Z'Z + kI_p)^{-1}Z'Y
\]

where \( Z = XD^{-\frac{1}{2}} \) and \( I_p \) is a \( p \times p \) identity matrix.

For IPC analysis, the smallest \( m \) eigenvalues of \( Z'Z \) (where \( m \) is specified with the PCOMIT= option) are omitted to form the estimates.

For information about ridge regression and IPC standardized parameter estimates, parameter estimate standard errors, and variance inflation factors, see Rawlings, Pantula, and Dickey (1998); Neter, Wasserman, and
Kutner (1990); Marquardt and Snee (1975). Unlike Rawlings, Pantula, and Dickey (1998), the REG procedure uses the mean squared errors of the submodels instead of the full model MSE to compute the standard errors of the parameter estimates.

Construction of Q-Q and P-P Plots

If a normal probability-probability or quantile-quantile plot for the variable \( x \) is requested, the \( n \) nonmissing values of \( x \) are first ordered from smallest to largest:

\[
x(1) \leq x(2) \leq \cdots \leq x(n)
\]

If a Q-Q plot is requested (with a PLOT statement of the form PLOT yvariable*NQQ.), the \( i \)th-ordered value \( x(i) \) is represented by a point with y-coordinate \( x(i) \) and x-coordinate \( \Phi^{-1} \left( \frac{i-0.375}{n+0.25} \right) \), where \( \Phi(\cdot) \) is the standard normal distribution.

If a P-P plot is requested (with a PLOT statement of the form PLOT yvariable*NPP.), the \( i \)th-ordered value \( x(i) \) is represented by a point with y-coordinate \( \frac{i}{n} \) and x-coordinate \( \Phi \left( \frac{x(i)-\mu}{\sigma} \right) \), where \( \mu \) is the mean of the nonmissing \( x \)-values and \( \sigma \) is the standard deviation. If an \( x \)-value has multiplicity \( k \) (that is, \( x(i) = \cdots = x(i+k-1) \)), then only the point \( \left( \Phi \left( \frac{x(i)-\mu}{\sigma} \right), \frac{i+k-1}{n} \right) \) is displayed.

Computational Methods

The REG procedure first composes a crossproducts matrix. The matrix can be calculated from input data, reformed from an input correlation matrix, or read in from an SSCP data set. For each model, the procedure selects the appropriate crossproducts from the main matrix. The normal equations formed from the crossproducts are solved by using a sweep algorithm (Goodnight 1979). The method is accurate for data that are reasonably scaled and not too collinear.

The mechanism that PROC REG uses to check for singularity involves the diagonal (pivot) elements of \( X'X \) as it is being swept. If a pivot is less than SINGULAR*CSS, then a singularity is declared and the pivot is not swept (where CSS is the corrected sum of squares for the regressor and SINGULAR is machine dependent but is approximately 1E–9 on most machines or reset in the PROC REG statement).

The sweep algorithm is also used in many places in the model-selection methods. The RSQUARE method uses the leaps-and-bounds algorithm by Furnival and Wilson (1974).

Computer Resources in Regression Analysis

The REG procedure is efficient for ordinary regression; however, requests for optional features can greatly increase the amount of time required.

The major computational expense in the regression analysis is the collection of the crossproducts matrix. For \( p \) variables and \( n \) observations, the time required is proportional to \( np^2 \). For each model run, PROC REG needs time roughly proportional to \( k^3 \), where \( k \) is the number of regressors in the model. Include an additional \( nk^2 \) for the R, CLM, or CLI option and another \( nk^2 \) for the INFLUENCE option.
Most of the memory that PROC REG needs to solve large problems is used for crossproducts matrices. PROC REG requires \(4p^2\) bytes for the main crossproducts matrix plus \(4k^2\) bytes for the largest model. If several output data sets are requested, memory is also needed for buffers.

See the section “Input Data Sets” on page 8489 for information about how to use TYPE=SSCP data sets to reduce computing time.

**Displayed Output**

Many of the more specialized tables are described in detail in previous sections. Most of the formulas for the statistics are in Chapter 4, “Introduction to Regression Procedures,” while other formulas can be found in the section “Model Fit and Diagnostic Statistics” on page 8514 and the section “Influence Statistics” on page 8516.

The analysis-of-variance table includes the following:

- the Source of the variation, Model for the fitted regression, Error for the residual error, and C Total for the total variation after correcting for the mean. The Uncorrected Total Variation is produced when the NOINT option is used.
- the degrees of freedom (DF) associated with the source
- the Sum of Squares for the term
- the Mean Square, the sum of squares divided by the degrees of freedom
- the F Value for testing the hypothesis that all parameters are zero except for the intercept. This is formed by dividing the mean square for Model by the mean square for Error.
- the Prob>F, the probability of getting a greater F statistic than that observed if the hypothesis is true. This is the significance probability.

Other statistics displayed include the following:

- Root MSE is an estimate of the standard deviation of the error term. It is calculated as the square root of the mean square error.
- Dep Mean is the sample mean of the dependent variable.
- C.V. is the coefficient of variation, computed as 100 times Root MSE divided by Dep Mean. This expresses the variation in unitless values.
- R-square is a measure between 0 and 1 that indicates the portion of the (corrected) total variation that is attributed to the fit rather than left to residual error. It is calculated as SS(Model) divided by SS(Total). It is also called the coefficient of determination. It is the square of the multiple correlation—in other words, the square of the correlation between the dependent variable and the predicted values.
• Adj R-square, the adjusted R square, is a version of R square that has been adjusted for degrees of freedom. It is calculated as

$$\hat{R}^2 = 1 - \frac{(n - i)(1 - R^2)}{n - p}$$

where \(i\) is equal to 1 if there is an intercept and 0 otherwise, \(n\) is the number of observations used to fit the model, and \(p\) is the number of parameters in the model.

The parameter estimates and associated statistics are then displayed, and they include the following:

• the Variable used as the regressor, including the name Intercept to represent the estimate of the intercept parameter

• the degrees of freedom (DF) for the variable. There is one degree of freedom unless the model is not full rank.

• the Parameter Estimate

• the Standard Error, the estimate of the standard deviation of the parameter estimate

• T for H0: Parameter=0, the \(t\) test that the parameter is zero. This is computed as the Parameter Estimate divided by the Standard Error.

• the Prob > |T|, the probability that a \(t\) statistic would obtain a greater absolute value than that observed given that the true parameter is zero. This is the two-tailed significance probability.

If model-selection methods other than NONE, RSQUARE, ADJRSQ, and CP are used, the analysis-of-variance table and the parameter estimates with associated statistics are produced at each step. Also displayed are the following:

• C(p), which is Mallows’ \(C_p\) statistic

• bounds on the condition number of the correlation matrix for the variables in the model (Berk 1977)

After statistics for the final model are produced, the following is displayed when the method chosen is FORWARD, BACKWARD, or STEPWISE:

• a Summary table listing Step number, Variable Entered or Removed, Partial and Model R-square, and C(p) and F statistics

The RSQUARE method displays its results beginning with the model containing the fewest independent variables and producing the largest R square. Results for other models with the same number of variables are then shown in order of decreasing R square, and so on, for models with larger numbers of variables. The ADJRSQ and CP methods group models of all sizes together and display results beginning with the model having the optimal value of adjusted R square and \(C_p\), respectively.

For each model considered, the RSQUARE, ADJRSQ, and CP methods display the following:
Chapter 102: The REG Procedure

- Number in Model or IN, the number of independent variables used in each model
- R-square or RSQ, the squared multiple correlation coefficient

If the B option is specified, the RSQUARE, ADJRSQ, and CP methods produce the following:

- Parameter Estimates, the estimated regression coefficients

If the B option is not specified, the RSQUARE, ADJRSQ, and CP methods display the following:

- Variables in Model, the names of the independent variables included in the model

---

Plot Options Superseded by ODS Graphics

You can select one of the following three types of graphics in PROC REG: ODS, traditional, and line printer. ODS Graphics is the preferred method of creating graphs, superseding the other two. This section describes the options that are available on the PROC REG, PAINT, and PLOT statements for traditional and line printer graphics.

When ODS Graphics is enabled, you can use the PLOTS= option in the PROC REG statement to create plots by using ODS Graphics. For more information about ODS Graphics options see the PLOTS= option in the section “Syntax: REG Procedure” on page 8448.

If ODS Graphics is not enabled and you specify the LINEPRINTER option, line printer plots are produced; otherwise traditional graphics are produced.

Table 102.9 summarizes the options available in the PROC REG statement for line printer and traditional graphics.

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ANNOTATE=</td>
<td>Specifies an annotation data set</td>
</tr>
<tr>
<td>GOUT=</td>
<td>Specifies the graphics catalog in which graphics output is saved</td>
</tr>
<tr>
<td>LINEPRINTER</td>
<td>Creates printer plots</td>
</tr>
</tbody>
</table>

The following options are used to produce line printer and traditional graphics:

ANNOTATE=SAS-data-set
ANO= SAS-data-set

specifies an input data set containing annotate variables, as described in SAS/GRAPH: Reference. You can use this data set to add features to the traditional graphics that you request with the PLOT statement. Features provided in this data set are applied to all plots produced in the current run of PROC REG.
To add features to individual plots, use the ANNOTATE= option in the PLOT statement. This option cannot be used if the LINEPRINTER option is specified.

**GOUT=** graphics-catalog

specifies the graphics catalog in which traditional graphics output is saved. The default *graphics-catalog* is WORK.GSEG. The GOUT= option cannot be used if the LINEPRINTER option is specified.

**LINEPRINTER | LP**

creates printer plots. If you do not specify this option, requested plots are created on a high-resolution graphics device. See the PLOTS= option for information about using ODS graphics to create modern statistical graphics.

**PAINT Statement**

```
PAINT < condition | ALLOBS > </ options > ;
PAINT < STATUS | UNDO > ;
```

The PAINT statement is used with line printer plots. See the PLOTS= option for information about using ODS graphics to create modern statistical graphics.

The PAINT statement selects observations to be painted or highlighted in a scatter plot on line printer output; the PAINT statement is ignored if the LINEPRINTER option is not specified in the PROC REG statement.

All observations that satisfy *condition* are painted using some specific symbol. The PAINT statement does not generate a scatter plot and must be followed by a PLOT statement, which does generate a scatter plot. Several PAINT statements can be used before a PLOT statement, and all prior PAINT statement requests are applied to all later PLOT statements.

The PAINT statement lists the observation numbers of the observations selected, the total number of observations selected, and the plotting symbol used to paint the points.

On a plot, paint symbols take precedence over all other symbols. If any position contains more than one painted point, the paint symbol for the observation plotted last is used.

The PAINT statement cannot be used when a TYPE=CORR, TYPE=COV, or TYPE=SSCP data set is used as the input data set for PROC REG. Also, the PAINT statement cannot be used for models with more than one dependent variable. Note that the syntax for the PAINT statement is the same as the syntax for the REWEIGHT statement.

**Specifying Condition**

*Condition* is used to select observations to be painted. The syntax of *condition* is

```
variable compare value
```

or

```
variable compare value logical variable compare value
```

where
variable is one of the following:

- a variable name in the input data set
- OBS., which is the observation number
- keyword, where keyword is a keyword for a statistic requested in the OUTPUT statement

compare is an operator that compares variable to value. Compare can be any one of the following: <, <=, >, >=, =, ^=. The operators LT, LE, GT, GE, EQ, and NE, respectively, can be used instead of the preceding symbols. See the “Expressions” section in SAS Language Reference: Concepts for more information about comparison operators.

value gives an unformatted value of variable. Observations are selected to be painted if they satisfy the condition created by variable compare value. Value can be a number or a character string. If value is a character string, it must be eight characters or less and must be enclosed in quotes. In addition, value is case-sensitive. In other words, the statements

```plaintext
paint name='henry';
```

and

```plaintext
paint name='Henry';
```

are not the same.

logical is one of two logical operators. Either AND or OR can be used. To specify AND, use AND or the symbol &. To specify OR, use OR or the symbol |.

Here are some examples of the variable compare value form:

```plaintext
paint name='Henry';
paint residual.>=20;
paint obs.=99;
```

Here are some examples of the variable compare value logical variable compare value form:

```plaintext
paint name='Henry'|name='Mary';
paint residual.>=20 or residual.<=10;
paint obs.>=11 and residual.<=20;
```

Using ALLOBS

Instead of specifying condition, the ALLOBS option can be used to select all observations. This is most useful when you want to unpaint all observations. For example,

```plaintext
paint allobs / reset;
```

resets the symbols for all observations.

Options in the PAINT Statement

The following options can be used when either a condition is specified, the ALLOBS option is specified, or nothing is specified before the slash. If only an option is listed, the option applies to the observations selected in the previous PAINT statement, not to the observations selected by reapplying the condition from the previous PAINT statement. For example, in the statements
the second PAINT statement paints only those observations selected in the first PAINT statement. No additional observations are painted even if, after refitting the model, there are new observations that meet the condition in the first PAINT statement.

NOTE: Options are not available when either the UNDO or STATUS option is used.

You can specify the following options after a slash (/).

**NOLIST**
suppresses the display of the selected observation numbers. If the NOLIST option is not specified, a list of observations selected is written to the log. The list includes the observation numbers and painting symbol used to paint the points. The total number of observations selected to be painted is also shown.

**RESET**
changes the painting symbol to the current default symbol, effectively unpainting the observations selected. If you set the default symbol by using the SYMBOL= option in the PLOT statement, the RESET option in the PAINT statement changes the painting symbol to the symbol you specified. Otherwise, the default symbol of ‘1’ is used.

**SYMBOL=’character’**
specifies a painting symbol. If the SYMBOL= option is omitted, the painting symbol is either the one used in the most recent PAINT statement or, if there are no previous PAINT statements, the symbol ‘@’. For example,

```plaintext
paint / symbol='#';
```
changes the painting symbol for the observations selected by the most recent PAINT statement to ‘#’. As another example,

```plaintext
paint temp lt 22 / symbol='c';
```
changes the painting symbol to ‘c’ for all observations with TEMP<22. In general, the numbers 1, 2, ..., 9 and the asterisk are not recommended as painting symbols. These symbols are used as default symbols in the PLOT statement, where they represent the number of replicates at a point. If SYMBOL=’ ’ is used, no painting is done in the current plot. If SYMBOL=’ ’ is used, observations are painted with a blank and are no longer seen on the plot.

**STATUS and UNDO**
Instead of specifying condition or the ALLOBS option, you can use the STATUS or UNDO option as follows:
STATUS

lists (in the log) the observation number and plotting symbol of all currently painted observations.

UNDO

undoes changes made by the most recent PAINT statement. Observations might be, but are not necessarily, unpainted. For example:

```plaintext
   paint obs. <=10 / symbol='a';
   ...other interactive statements
   paint obs.=1 / symbol='b';
   ...other interactive statements
   paint undo;
```

The last PAINT statement changes the plotting symbol used for observation 1 back to ‘a’. If the statement

```plaintext
   paint / reset;
```

is used instead, observation 1 is unpainted.

**PLOT Statement**

```plaintext
   PLOT <yvariable* xvariable> <symbol> <...yvariable* xvariable> <symbol> </options> ;
```

The PLOT statement is used with line printer and traditional graphics. See the PLOTS= option for information about using ODS graphics to create modern statistical graphics.

The PLOT statement in PROC REG displays scatter plots with *yvariable* on the vertical axis and *xvariable* on the horizontal axis. Line printer plots are generated if the LINEPRINTER option is specified in the PROC REG statement; otherwise, the traditional graphics are created. Points in line printer plots can be marked with *symbols*, while global graphics statements such as GOPTIONS and SYMBOL are used to enhance the traditional graphics. Note that the plots you request by using the PLOT statement are independent of the ODS graphical displays (see the section “ODS Graphics” on page 8557) that are available in PROC REG.

As with most other interactive statements, the PLOT statement implicitly refits the model. For example, if a PLOT statement is preceded by a REWEIGHT statement, the model is recomputed, and the plot reflects the new model.

If there are multiple MODEL statements preceding a PLOT statement, then the PLOT statement refers to the latest MODEL statement.

The PLOT statement cannot be used when a TYPE=CORR, TYPE=COV, or TYPE=SSCP data set is used as input to PROC REG.

You can specify several PLOT statements for each MODEL statement, and you can specify more than one plot in each PLOT statement.

**Specifying Yvariables, Xvariables, and Symbol**

More than one *yvariable*<*xvariable* pair can be specified to request multiple plots. The *yvariables* and *xvariables* can be as follows:

- any variables specified in the VAR or MODEL statement before the first RUN statement
• *keyword*., where *keyword* is a regression diagnostic statistic available in the OUTPUT statement (see Table 102.10). For example,

```
plot predicted.*residual.;
```

generates one plot of the predicted values by the residuals for each dependent variable in the MODEL statement. These statistics can also be plotted against any of the variables in the VAR or MODEL statements.

• the keyword OBS. (the observation number), which can be plotted against any of the preceding variables

• the keyword NPP. or NQQ., which can be used with any of the preceding variables to construct normal P-P or Q-Q plots, respectively (see the section “Construction of Q-Q and P-P Plots” on page 8537 for more information)

• *keywords* for model fit summary statistics available in the OUTEST= data set with _TYPE_ = PARMS (see Table 102.10). A SELECTION= method (other than NONE) must be requested in the MODEL statement for these variables to be plotted. If one member of a *yvariable*:*xvariable* pair is from the OUTEST= data set, the other member must also be from the OUTEST= data set.

The OUTPUT statement and the OUTEST= option are not required when their *keywords* are specified in the PLOT statement.

The *yvariable* and *xvariable* specifications can be replaced by a set of variables and statistics enclosed in parentheses. When this occurs, all possible combinations of *yvariable* and *xvariable* are generated. For example, the following two statements are equivalent:

```
plot (y1 y2)*(x1 x2);
plot y1*x1 y1*x2 y2*x1 y2*x2;
```

The statement

```
plot;
```

is equivalent to respecifying the most recent PLOT statement without any options. However, the line printer options COLLECT, HPLOTS=, SYMBOL=, and VPLOTS=, described in the section “Line Printer Plots” on page 8553, apply across PLOT statements and remain in effect if they have been previously specified.

Options used for the traditional graphics are described in the following section; see “Line Printer Plots” on page 8553 for more information.

**Traditional Graphics**

The display of traditional graphics is described in the following paragraphs, the options are summarized in Table 102.10 and described in the section “Dictionary of PLOT Statement Options” on page 8549.

Several line printer statements and options are not supported for the traditional graphics. In particular the PAINT statement is disabled, as are the PLOT statement options CLEAR, COLLECT, HPLOTS=, NOCOLLECT, SYMBOL=, and VPLOTS=. To display more than one plot per page or to collect plots from multiple PLOT statements, use the PROC GREPLAY statement (see SAS/GRAPH: Reference). Also note that traditional graphics options are not recognized for line printer plots.
The fitted model equation and a label are displayed in the top margin of the plot; this display can be suppressed with the NOMODEL option. If the label is requested but cannot fit on one line, it is not displayed. The equation and label are displayed on one line when possible; if more lines are required, the label is displayed in the first line with the model equation in successive lines. If displaying the entire equation causes the plot to be unacceptably small, the equation is truncated. Table 102.11 lists options to control the display of the equation.

Four statistics are displayed by default in the right margin: the number of observations, R square, the adjusted R square, and the root mean square error. The display of these statistics can be suppressed with the NOSTAT option. You can specify other options to request the display of various statistics in the right margin; see Table 102.11.

A default reference line at zero is displayed if residuals are plotted. If the dependent variable is plotted against the independent variable in a simple linear regression model, the fitted regression line is displayed by default. Default reference lines can be suppressed with the NOLINE option; the lines are not displayed if the OVERLAY option is specified.

Specialized plots are requested with special options. For each coefficient, the RIDGEPLOT option plots the ridge estimates against the ridge values \( k \); see the description of the RIDGEPLOT option in the section “Dictionary of PLOT Statement Options” on page 8549 for more details. The CONF option plots 100(1 – \( \alpha \))% confidence intervals for the mean while the PRED option plots 100(1 – \( \alpha \))% prediction intervals; see the description of these options in the section “Dictionary of PLOT Statement Options” on page 8549 for more details.

If a SELECTION= method is requested, the fitted model equation and the statistics displayed in the margin correspond to the selected model. For the ADJRSQ and CP methods, the selected model is treated as a submodel of the full model. If a CP.*NP. plot is requested, the CHOCKING= and CMALLOWS= options display model selection reference lines; see the descriptions of these options in the section “Dictionary of PLOT Statement Options” on page 8549 for more details.

**PLOT Statement variable Keywords**  The following table lists the keywords available as PLOT statement xvariables and yvariables. All keywords have a trailing dot; for example, “COOKD.” requests Cook’s D statistic. Neither the OUTPUT statement nor the OUTEST= option needs to be specified.

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic Statistics</td>
<td></td>
</tr>
<tr>
<td>COOKD.</td>
<td>Cook’s D influence statistics</td>
</tr>
<tr>
<td>COVRATIO.</td>
<td>standard influence of observation on covariance of betas</td>
</tr>
<tr>
<td>DFFITS.</td>
<td>standard influence of observation on predicted value</td>
</tr>
<tr>
<td>H.</td>
<td>leverage</td>
</tr>
<tr>
<td>LCL.</td>
<td>lower bound of 100(1 – ( \alpha ))% confidence interval for individual prediction</td>
</tr>
<tr>
<td>LCLM.</td>
<td>lower bound of 100(1 – ( \alpha ))% confidence interval for the mean of the dependent variable</td>
</tr>
<tr>
<td>PREDICTED.</td>
<td>predicted values</td>
</tr>
<tr>
<td>| PRED. | P.</td>
<td></td>
</tr>
<tr>
<td>PRESS.</td>
<td>residuals from refitting the model with current observation deleted</td>
</tr>
<tr>
<td>RESIDUAL. | R.</td>
<td>residuals</td>
</tr>
</tbody>
</table>

Table 102.10  Keywords for PLOT Statement xvariables
Table 102.10  continued

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RSTUDENT.</td>
<td>studentized residuals with the current observation deleted</td>
</tr>
<tr>
<td>STDI.</td>
<td>standard error of the individual predicted value</td>
</tr>
<tr>
<td>STDP.</td>
<td>standard error of the mean predicted value</td>
</tr>
<tr>
<td>STDR.</td>
<td>standard error of the residual</td>
</tr>
<tr>
<td>STUDENT.</td>
<td>residuals divided by their standard errors</td>
</tr>
<tr>
<td>UCL.</td>
<td>upper bound of 100(1 − (\alpha))% confidence interval for individual</td>
</tr>
<tr>
<td>UCLM.</td>
<td>prediction</td>
</tr>
</tbody>
</table>

Other Keywords Used with Diagnostic Statistics

NPP.        | normal probability-probability plot                                        |
NQQ.        | normal quantile-quantile plot                                              |
OBS.        | observation number (cannot plot against OUTEST= statistics)                |

Model Fit Summary Statistics

ADJRSQ.     | adjusted R-square                                                          |
AIC.        | Akaike’s information criterion                                             |
BIC.        | Sawa’s Bayesian information criterion                                      |
CP.         | Mallows’ \(C_p\) statistic                                                |
EDF.        | error degrees of freedom                                                   |
GMSEP.      | estimated MSE of prediction, assuming multivariate normality              |
IN.         | number of regressors in the model not including the intercept              |
JP.         | final prediction error                                                     |
MSE.        | mean squared error                                                          |
NP.         | number of parameters in the model (including the intercept)                |
PC.         | Amemiya’s prediction criterion                                             |
RMSE.       | root MSE                                                                    |
RSQ.        | R-square                                                                    |
SBC.        | SBC statistic                                                               |
SP.         | SP statistic                                                                |
SSE.        | error sum of squares                                                        |

Summary of PLOT Statement Graphics Options  Table 102.11 summarizes the options available in the PLOT statement. These options are available unless the LINEPRINTER option is specified in the PROC REG statement. For complete descriptions, see the section “Dictionary of PLOT Statement Options” on page 8549.

Table 102.11  Traditional Graphics Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
</table>
| General Graphics Options
| ANNOTATE= SAS-data-set          | Specifies the annotate data set                                            |
| CHOCKING=color            | Requests a reference line for \(C_p\) model selection criteria             |
| CMALLOWS=color            | Requests a reference line for the \(C_p\) model selection criterion         |
Table 102.11  continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONF</td>
<td>Requests plots of $100(1 - \alpha)%$ confidence intervals for the mean</td>
</tr>
<tr>
<td>DESCRIPTION= 'string'</td>
<td>Specifies a description for graphics catalog member</td>
</tr>
<tr>
<td>NAME='string'</td>
<td>Names the plot in the graphics catalog</td>
</tr>
<tr>
<td>OVERLAY</td>
<td>Overlays plots from the same model</td>
</tr>
<tr>
<td>PRED</td>
<td>Requests plots of $100(1 - \alpha)%$ prediction intervals for individual responses</td>
</tr>
<tr>
<td>RIDGEPLOT</td>
<td>Requests the ridge trace for ridge regression</td>
</tr>
</tbody>
</table>

**Axis and Legend Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LEGEND=LEGENDn</td>
<td>Specifies LEGEND statement to be used</td>
</tr>
<tr>
<td>NOLEGEND</td>
<td>Suppresses display of the legend</td>
</tr>
<tr>
<td>HAXIS=values</td>
<td>Specifies tick mark values for horizontal axis</td>
</tr>
<tr>
<td>VAXIS=values</td>
<td>Specifies tick mark values for vertical axis</td>
</tr>
</tbody>
</table>

**Reference Line Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>HREF=values</td>
<td>Specifies reference lines perpendicular to horizontal axis</td>
</tr>
<tr>
<td>LHREF=linetype</td>
<td>Specifies line style for HREF= lines</td>
</tr>
<tr>
<td>LLINE=linetype</td>
<td>Specifies line style for lines displayed by default</td>
</tr>
<tr>
<td>LVREF=linetype</td>
<td>Specifies line style for VREF= lines</td>
</tr>
<tr>
<td>NOLINE</td>
<td>Suppresses display of any default reference line</td>
</tr>
<tr>
<td>VREF=values</td>
<td>Specifies reference lines perpendicular to vertical axis</td>
</tr>
</tbody>
</table>

**Color Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAXIS=color</td>
<td>Specifies color for axis line and tick marks</td>
</tr>
<tr>
<td>CFRAME=color</td>
<td>Specifies color for frame</td>
</tr>
<tr>
<td>CHREF=color</td>
<td>Specifies color for HREF= lines</td>
</tr>
<tr>
<td>CLINE=color</td>
<td>Specifies color for lines displayed by default</td>
</tr>
<tr>
<td>CTEXT=color</td>
<td>Specifies color for text</td>
</tr>
<tr>
<td>CVREF=color</td>
<td>Specifies color for VREF= lines</td>
</tr>
</tbody>
</table>

**Options for Displaying the Fitted Model Equation**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MODELFONT=font</td>
<td>Specifies font of model equation and model label</td>
</tr>
<tr>
<td>MODELHT=value</td>
<td>Specifies text height of model equation and model label</td>
</tr>
<tr>
<td>MODELLAB='label'</td>
<td>Specifies model label</td>
</tr>
<tr>
<td>NOMODEL</td>
<td>Suppresses display of the fitted model and the label</td>
</tr>
</tbody>
</table>

**Options for Displaying Statistics in the Plot Margin**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC</td>
<td>Displays Akaike’s information criterion</td>
</tr>
<tr>
<td>BIC</td>
<td>Displays Sawa’s Bayesian information criterion</td>
</tr>
<tr>
<td>CP</td>
<td>Displays Mallows’ $C_p$ statistic</td>
</tr>
<tr>
<td>EDF</td>
<td>Displays the error degrees of freedom</td>
</tr>
<tr>
<td>GMSEP</td>
<td>Displays the estimated MSE of prediction assuming multivariate normality</td>
</tr>
<tr>
<td>IN</td>
<td>Displays the number of regressors in the model not including the intercept</td>
</tr>
<tr>
<td>JP</td>
<td>Displays the $J_p$ statistic</td>
</tr>
<tr>
<td>MSE</td>
<td>Displays the mean squared error</td>
</tr>
</tbody>
</table>
Table 102.11  continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOSTAT</td>
<td>Suppresses display of the default statistics: the number of observations,</td>
</tr>
<tr>
<td></td>
<td>R-square, adjusted R-square, and root mean square error</td>
</tr>
<tr>
<td>NP</td>
<td>Displays the number of parameters in the model including the intercept, if</td>
</tr>
<tr>
<td></td>
<td>any</td>
</tr>
<tr>
<td>PC</td>
<td>Displays the PC statistic</td>
</tr>
<tr>
<td>SBC</td>
<td>Displays the SBC statistic</td>
</tr>
<tr>
<td>SP</td>
<td>Displays the $S_p$ statistic</td>
</tr>
<tr>
<td>SSE</td>
<td>Displays the error sum of squares</td>
</tr>
<tr>
<td>STATFONT=</td>
<td>Specifies font of text displayed in the margin</td>
</tr>
<tr>
<td>STATHT=</td>
<td>Specifies height of text displayed in the margin</td>
</tr>
</tbody>
</table>

Dictionary of PLOT Statement Options  The following entries describe the PLOT statement options in detail. Note that these options are available unless you specify the LINEPRINTER option in the PROC REG statement.

AIC
displays Akaike’s information criterion in the plot margin.

ANNOTATE=SAS-data-set
ANNO=SAS-data-set
specifies an input data set that contains appropriate variables for annotation. This applies only to displays created with the current PLOT statement. See SAS/GRAPH: Reference for more information.

BIC
displays Sawa’s Bayesian information criterion in the plot margin.

CAXIS=color
CAXES=color
CA=color
specifies the color for the axes, frame, and tick marks.

CFRAME=color
CFR=color
specifies the color for filling the area enclosed by the axes and the frame.

CHOCKING=color
requests reference lines corresponding to the equations $C_p = p$ and $C_p = 2p - p_{full}$, where $p_{full}$ is the number of parameters in the full model (excluding the intercept) and $p$ is the number of parameters in the subset model (including the intercept). The color must be specified; the $C_p = p$ line is solid and the $C_p = 2p - p_{full}$ line is dashed. Only PLOT statements of the form PLOT CP.*NP. produce these lines.

For the purpose of parameter estimation, Hocking (1976) suggests selecting a model where $C_p \leq 2p - p_{full}$. For the purpose of prediction, Hocking suggests the criterion $C_p \leq p$. You can request
the single reference line $C_p = p$ with the CMALLOWS= option. If, for example, you specify both
CHOCKING=RED and CMALLOWS=BLUE, then the $C_p = 2p - p_{\text{full}}$ line is red and the $C_p = p$
line is blue.

CHREF=color

CH=color

specifies the color for lines requested with the HREF= option.

CLINE=color

CL=color

specifies the color for lines displayed by default. See the NOLINE option for details.

CMALLOWS=color

requests a $C_p = p$ reference line, where $p$ is the number of parameters (including the intercept) in the
subset model. The color must be specified; the line is solid. Only PLOT statements of the form PLOT
CP.*NP. produce this line.

Mallows (1973) suggests that all subset models with $C_p$ small and near $p$ be considered for further
study. See the CHOCKING= option for related model-selection criteria.

CONF

is a keyword used as a shorthand option to request plots that include $(100 - \alpha)\%$ confidence intervals
for the mean response. The ALPHA= option in the PROC REG or MODEL statement selects the
significance level $\alpha$, which is 0.05 by default. The CONF option is valid for simple regression models
only, and is ignored for plots where confidence intervals are inappropriate. The CONF option replaces
the CONF95 option; however, the CONF95 option is still supported when the ALPHA= option is not
specified. The OVERLAY option is ignored when the CONF option is specified.

CP

displays Mallows’ $C_p$ statistic in the plot margin.

CTEXT=color

CT=color

specifies the color for text including tick mark labels, axis labels, the fitted model label and equation,
the statistics displayed in the margin, and legends.

CVREF=color

CV=color

specifies the color for lines requested with the VREF= option.

DESCRIPTION='string'

DESC='string'

specifies a descriptive string, up to 40 characters, that appears in the description field of the PROC
GREPLAY master menu.

EDF

displays the error degrees of freedom in the plot margin.
GMSEP displays the estimated mean square error of prediction in the plot margin. Note that the estimate is calculated under the assumption that both independent and dependent variables have a multivariate normal distribution.

HAXIS=values
HA=values
specifies tick mark values for the horizontal axis.

HREF=values
specifies where reference lines perpendicular to the horizontal axis are to appear.

IN displays the number of regressors in the model (not including the intercept) in the plot margin.

JP displays the $J_P$ statistic in the plot margin.

LEGEND=LEGEND$n$
specifies the LEGEND$n$ statement to be used. The LEGEND$n$ statement is a global graphics statement; see SAS/GRAPH: Reference for more information.

LHREF=linetype
LH=linetype
specifies the line style for lines requested with the HREF= option. The default linetype is 2. Note that LHREF=1 requests a solid line. See SAS/GRAPH: Reference for a table of available line types.

LLINE=linetype
LL=linetype
specifies the line style for reference lines displayed by default; see the NOLINE option for details. The default linetype is 2. Note that LLINE=1 requests a solid line.

LVREF=linetype
LV=linetype
specifies the line style for lines requested with the VREF= option. The default linetype is 2. Note that LVREF=1 requests a solid line.

MODELFONT=font
specifies the font used for displaying the fitted model label and the fitted model equation. See SAS/GRAPH: Reference for tables of software fonts.

MODELHT=height
specifies the text height for the fitted model label and the fitted model equation.

MODELLAB='label'
specifies the label to be displayed with the fitted model equation. By default, no label is displayed. If the label does not fit on one line, it is not displayed. See the section “Traditional Graphics” on page 8545 for more information.
MSE
displays the mean squared error in the plot margin.

NAME='string'  
specifies a descriptive string, up to eight characters, that appears in the name field of the PROC GREPLAY master menu. The default string is REG.

NOLEGEND  
suppresses the display of the legend.

NOLINE  
suppresses the display of default reference lines. A default reference line at zero is displayed if residuals are plotted. If the dependent variable is plotted against the independent variable in a simple regression model, then the fitted regression line is displayed by default. Default reference lines are not displayed if the OVERLAY option is specified.

NOMODEL  
suppresses the display of the fitted model equation.

NOSTAT  
suppresses the display of statistics in the plot margin. By default, the number of observations, R-square, adjusted R-square, and root MSE are displayed.

NP  
displays the number of regressors in the model including the intercept, if any, in the plot margin.

OVERLAY  
overlays all plots specified in the PLOT statement from the same model on one set of axes. The variables for the first plot label the axes. The procedure automatically scales the axes to fit all of the variables unless the HAXIS= or VAXIS= option is used. Default reference lines are not displayed. A default legend is produced; the LEGEND= option can be used to customize the legend.

PC  
displays the PC statistic in the plot margin.

PRED  
is a keyword used as a shorthand option to request plots that include \((100 - \alpha)\%\) prediction intervals for individual responses. The ALPHA= option in the PROC REG or MODEL statement selects the significance level \(\alpha\), which is 0.05 by default. The PRED option is valid for simple regression models only, and is ignored for plots where prediction intervals are inappropriate. The PRED option replaces the PRED95 option; however, the PRED95 option is still supported when the ALPHA= option is not specified. The OVERLAY option is ignored when the PRED option is specified.

RIDGEPLOT  
creates overlaid plots of ridge estimates against ridge values for each coefficient. The points corresponding to the estimates of each coefficient in the plot are connected by lines. For ridge estimates to be computed and plotted, the OUTEST= option must be specified in the PROC REG statement, and the RIDGE=list must be specified in either the PROC REG or MODEL statement.
SBC
displays the SBC statistic in the plot margin.

SP
displays the $S_p$ statistic in the plot margin.

SSE
displays the error sum of squares in the plot margin.

\texttt{STATFONT=font}
specifies the font used for displaying the statistics that appear in the plot margin. See \textit{SAS/GRAPH: Reference} for tables of software fonts.

\texttt{STATHT=height}
specifies the text height of the statistics that appear in the plot margin.

\texttt{USEALL}
specifies that predicted values at data points with missing dependent variable(s) be included on appropriate plots. By default, only points used in constructing the SSCP matrix appear on plots.

\texttt{VAXIS=values}
\texttt{VA=values}
specifies tick mark values for the vertical axis.

\texttt{VREF=values}
specifies where reference lines perpendicular to the vertical axis are to appear.

\textbf{Line Printer Plots}

Line printer plots are requested with the LINEPRINTER option in the \texttt{PROC REG} statement. Points in line printer plots can be marked with symbols, which can be specified as a single character enclosed in quotes or the name of any variable in the input data set.

If a character variable is used for the symbol, the first (leftmost) nonblank character in the formatted value of the variable is used as the plotting symbol. If a character in quotes is specified, that character becomes the plotting symbol. If a character is used as the plotting symbol, and if there are different plotting symbols needed at the same point, the symbol ‘?’ is used at that point.

If an unformatted numeric variable is used for the symbol, the symbols ‘1’, ‘2’, \ldots, ‘9’ are used for variable values 1, 2, \ldots, 9. For noninteger values, only the integer portion is used as the plotting symbol. For values of 10 or greater, the symbol ‘*’ is used. For negative values, a ‘?’ is used. If a numeric variable is used, and if there is more than one plotting symbol needed at the same point, the sum of the variable values is used at that point. If the sum exceeds 9, the symbol ‘*’ is used.

If a symbol is not specified, the number of replicates at the point is displayed. The symbol ‘*’ is used if there are 10 or more replicates.

If the \texttt{LINEPRINTER} option is used, you can specify the following options in the \texttt{PLOT} statement after a slash (/):
CLEAR

clears any collected scatter plots before plotting begins but does not turn off the COLLECT option. Use this option when you want to begin a new collection with the plots in the current PLOT statement. For more information about collecting plots, see the COLLECT and NOCOLLECT options in this section.

COLLECT

specifies that plots begin to be collected from one PLOT statement to the next and that subsequent plots show an overlay of all collected plots. This option enables you to overlay plots before and after changes to the model or to the data used to fit the model. Plots collected before changes are unaffected by the changes and can be overlaid on later plots. You can request more than one plot with this option, and you do not need to request the same number of plots in subsequent PLOT statements. If you specify an unequal number of plots, plots in corresponding positions are overlaid. For example, the statements

```plaintext
plot residual.*predicted. y*x / collect;
run;
```

produce two plots. If these statements are then followed by

```plaintext
plot residual.*x;
run;
```

two plots are again produced. The first plot shows residual against X values overlaid on residual against predicted values. The second plot is the same as that produced by the first PLOT statement.

Axes are scaled for the first plot or plots collected. The axes are not rescaled as more plots are collected. Once specified, the COLLECT option remains in effect until the NOCOLLECT option is specified.

HPLOTS=number

sets the number of scatter plots that can be displayed across the page. The procedure begins with one plot per page. The value of the HPLOTS= option remains in effect until you change it in a later PLOT statement. See the VPLOTS= option for an example.

NOCOLLECT

specifies that the collection of scatter plots ends after adding the plots in the current PLOT statement. PROC REG starts with the NOCOLLECT option in effect. After you specify the NOCOLLECT option, any following PLOT statement produces a new plot that contains only the plots requested by that PLOT statement.

For more information, see the COLLECT option.

OVERLAY

enables requested scatter plots to be superimposed. The axes are scaled so that points on all plots are shown. If the HPLOTS= or VPLOTS= option is set to more than one, the overlaid plot occupies the first position on the page. The OVERLAY option is similar to the COLLECT option in that both options produce superimposed plots. However, OVERLAY superimposes only the plots in the associated PLOT statement; COLLECT superimposes plots across PLOT statements. The OVERLAY option can be used when the COLLECT option is in effect.
**SYMBOL='character'**

changes the default plotting symbol used for all scatter plots produced in the current and in subsequent PLOT statements. Both SYMBOL='' and SYMBOL=' ' are allowed.

If the SYMBOL= option has not been specified, the default symbol is ‘1’ for positions with one observation, ‘2’ for positions with two observations, and so on. For positions with more than 9 observations, ‘*’ is used. The SYMBOL= option (or a plotting symbol) is needed to avoid any confusion caused by this default convention. Specifying a particular symbol is especially important when either the OVERLAY or COLLECT option is being used.

If you specify the SYMBOL= option and use a number for *character*, that number is used for all points in the plot. For example, the statement

```
plot y*x / symbol='1';
```

produces a plot with the symbol ‘1’ used for all points.

If you specify a plotting symbol and the SYMBOL= option, the plotting symbol overrides the SYMBOL= option. For example, in the statements

```
plot y*x y*v='.' / symbol='*';
```

the symbol used for the plot of Y against X is ‘*’, and a ‘.’ is used for the plot of Y against V.

If a paint symbol is defined with a PAINT statement, the paint symbol takes precedence over both the SYMBOL= option and the default plotting symbol for the PLOT statement.

**VPLOTS=number**

sets the number of scatter plots that can be displayed down the page. The procedure begins with one plot per page. The value of the VPLOTS= option remains in effect until you change it in a later PLOT statement.

For example, to specify a total of six plots per page, with two rows of three plots, use the HPLOTS= and VPLOTS= options as follows:

```
plot y1*x1 y1*x2 y1*x3 y2*x1 y2*x2 y2*x3 / hplots=3 vplots=2;
run;
```

---

**ODS Table Names**

PROC REG assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information about ODS, see Chapter 20, “Using the Output Delivery System.”
<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACovEst</td>
<td>Consistent covariance of estimates matrix</td>
<td>MODEL</td>
<td>ALL, ACOV</td>
</tr>
<tr>
<td>ACovTestANOVA</td>
<td>Test ANOVA using ACOV estimates</td>
<td>TEST</td>
<td>ACOV (MODEL statement)</td>
</tr>
<tr>
<td>ANOVA</td>
<td>Model ANOVA table</td>
<td>MODEL</td>
<td>Default</td>
</tr>
<tr>
<td>CanCorr</td>
<td>Canonical correlations for hypothesis combinations</td>
<td>MTEST</td>
<td>CANPRINT</td>
</tr>
<tr>
<td>CollinDiag</td>
<td>Collinearity Diagnostics table</td>
<td>MODEL</td>
<td>COLLIN</td>
</tr>
<tr>
<td>CollinDiagNoInt</td>
<td>Collinearity Diagnostics for no intercept model</td>
<td>MODEL</td>
<td>COLLINOINT</td>
</tr>
<tr>
<td>ConditionBounds</td>
<td>Bounds on condition number</td>
<td>MODEL</td>
<td>(SELECTION=BACKWARD</td>
</tr>
<tr>
<td>Corr</td>
<td>Correlation matrix for analysis variables</td>
<td>PROC</td>
<td>ALL, CORR</td>
</tr>
<tr>
<td>CorrB</td>
<td>Correlation of estimates</td>
<td>MODEL</td>
<td>CORRB</td>
</tr>
<tr>
<td>CovB</td>
<td>Covariance of estimates</td>
<td>MODEL</td>
<td>COVB</td>
</tr>
<tr>
<td>CrossProducts</td>
<td>Bordered model $X'X$ matrix</td>
<td>MODEL</td>
<td>ALL, XPX</td>
</tr>
<tr>
<td>DWStatistic</td>
<td>Durbin-Watson statistic</td>
<td>MODEL</td>
<td>ALL, DW</td>
</tr>
<tr>
<td>DependenceEquations</td>
<td>Linear dependence equations</td>
<td>MODEL</td>
<td>Default if needed</td>
</tr>
<tr>
<td>Eigenvalues</td>
<td>MTest eigenvalues</td>
<td>MTEST</td>
<td>CANPRINT</td>
</tr>
<tr>
<td>Eigenvectors</td>
<td>MTest eigenvectors</td>
<td>MTEST</td>
<td>CANPRINT</td>
</tr>
<tr>
<td>EntryStatistics</td>
<td>Entry statistics for selection methods</td>
<td>MODEL</td>
<td>(SELECTION=BACKWARD</td>
</tr>
<tr>
<td>ErrorPlusHypothesis</td>
<td>MTest error plus hypothesis matrix $H+E$</td>
<td>MTEST</td>
<td>PRINT</td>
</tr>
<tr>
<td>ErrorSSCP</td>
<td>MTest error matrix $E$</td>
<td>MTEST</td>
<td>PRINT</td>
</tr>
<tr>
<td>FitStatistics</td>
<td>Model fit statistics</td>
<td>MODEL</td>
<td>Default</td>
</tr>
<tr>
<td>HypothesisSSCP</td>
<td>MTest hypothesis matrix</td>
<td>MTEST</td>
<td>PRINT</td>
</tr>
<tr>
<td>InvMTestCov</td>
<td>$\text{Inv}(L \text{ Ginv}(X'X) L')$ and $\text{Inv}(Lb-c)$</td>
<td>MTEST</td>
<td>DETAILS</td>
</tr>
<tr>
<td>InvTestCov</td>
<td>$\text{Inv}(L \text{ Ginv}(X'X) L')$ and $\text{Inv}(Lb-c)$</td>
<td>TEST</td>
<td>PRINT</td>
</tr>
<tr>
<td>InvXPX</td>
<td>Bordered $X'X$ inverse matrix</td>
<td>MODEL</td>
<td>I</td>
</tr>
<tr>
<td>MTestCov</td>
<td>$L \text{ Ginv}(X'X) L'$ and $Lb-c$</td>
<td>MTEST</td>
<td>DETAILS</td>
</tr>
<tr>
<td>MTransform</td>
<td>MTest matrix $M$, across dependents</td>
<td>MTEST</td>
<td>DETAILS</td>
</tr>
<tr>
<td>MultStat</td>
<td>Multivariate test statistics</td>
<td>MTEST</td>
<td>Default</td>
</tr>
<tr>
<td>NObs</td>
<td>Number of observations</td>
<td>MTEST</td>
<td>Default</td>
</tr>
</tbody>
</table>
### Table 102.12 continued

<table>
<thead>
<tr>
<th>ODS Table Name</th>
<th>Description</th>
<th>Statement</th>
<th>Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>OutputStatistics</td>
<td>Output statistics table</td>
<td>MODEL</td>
<td>ALL, CLI, CLM,</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>INFLUENCE, P, R</td>
</tr>
<tr>
<td>PartialData</td>
<td>Partial regression leverage data</td>
<td>MODEL</td>
<td>PARTIALDATA</td>
</tr>
<tr>
<td>ParameterEstimates</td>
<td>Model parameter estimates</td>
<td>MODEL</td>
<td>Default if SELECTION= is not specified</td>
</tr>
<tr>
<td>RemovalStatistics</td>
<td>Removal statistics for selection methods</td>
<td>MODEL</td>
<td>(SELECTION=BACKWARD</td>
</tr>
<tr>
<td>ResidualStatistics</td>
<td>Residual statistics and PRESS statistic</td>
<td>MODEL</td>
<td>ALL, CLI, CLM</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>INFLUENCE, P, R</td>
</tr>
<tr>
<td>SelParmEst</td>
<td>Parameter estimates for selection methods</td>
<td>MODEL</td>
<td>SELECTION=BACKWARD</td>
</tr>
<tr>
<td>SelectionSummary</td>
<td>Selection summary for FORWARD, BACKWARD, and STEPWISE methods</td>
<td>MODEL</td>
<td>SELECTION=BACKWARD</td>
</tr>
<tr>
<td>SeqParmEst</td>
<td>Sequential parameter estimates</td>
<td>MODEL</td>
<td>SEQB</td>
</tr>
<tr>
<td>SimpleStatistics</td>
<td>Simple statistics for analysis variables</td>
<td>PROC</td>
<td>ALL, SIMPLE</td>
</tr>
<tr>
<td>SpecTest</td>
<td>White’s heteroscedasticity test</td>
<td>MODEL</td>
<td>ALL, SPEC</td>
</tr>
<tr>
<td>SubsetSelSummary</td>
<td>Selection summary for RSQUARE, ADJRSQ, and CP methods</td>
<td>MODEL</td>
<td>SELECTION=RSQUARE</td>
</tr>
<tr>
<td>TestANOVA</td>
<td>Test ANOVA table</td>
<td>TEST</td>
<td>Default</td>
</tr>
<tr>
<td>TestCov</td>
<td>( L' \text{Ginv}(X'X) L' ) and ( L ' - c ) )</td>
<td>TEST</td>
<td>PRINT</td>
</tr>
<tr>
<td>USSCP</td>
<td>Uncorrected SSCP matrix for analysis variables</td>
<td>PROC</td>
<td>ALL, USSCP</td>
</tr>
</tbody>
</table>

**ODS Graphics**

Statistical procedures use ODS Graphics to create graphs as part of their output. ODS Graphics is described in detail in Chapter 21, “Statistical Graphics Using ODS.”

Before you create graphs, ODS Graphics must be enabled (for example, by specifying the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 623 in Chapter 21, “Statistical Graphics Using ODS.”
The overall appearance of graphs is controlled by ODS styles. Styles and other aspects of using ODS Graphics are discussed in the section “A Primer on ODS Statistical Graphics” on page 622 in Chapter 21, “Statistical Graphics Using ODS.”

The following sections describe the ODS graphical displays produced by PROC REG.

**Diagnostics Panel**

The “Diagnostics Panel” provides a display that you can use to get an overall assessment of your model. See Figure 102.9 for an example.

The panel contains the following plots:

- residuals versus the predicted values
- externally studentized residuals (RSTUDENT) versus the predicted values
- externally studentized residuals versus the leverage
- normal quantile-quantile plot (Q-Q plot) of the residuals
- dependent variable values versus the predicted values
- Cook’s $D$ versus observation number
- histogram of the residuals
- “Residual-Fit” (or RF) plot consisting of side-by-side quantile plots of the centered fit and the residuals
- box plot of the residuals if you specify the STATS=NONE suboption

Patterns in the plots of residuals or studentized residuals versus the predicted values, or spread of the residuals being greater than the spread of the centered fit in the RF plot, are indications of an inadequate model. Patterns in the spread about the 45-degree reference line in the plot of the dependent variable values versus the predicted values are also indications of an inadequate model.

The Q-Q plot, residual histogram, and box plot of the residuals are useful for diagnosing violations of the normality and homoscedasticity assumptions. If the data in a Q-Q plot come from a normal distribution, the points will cluster tightly around the reference line. A normal density is overlaid on the residual histogram to help in detecting departures from normality.

Following Rawlings, Pantula, and Dickey (1998), reference lines are shown on the relevant plots to identify observations deemed outliers or influential. Observations whose externally studentized residual magnitudes exceed 2 are deemed outliers. Observations whose leverage value exceeds $2p/n$ or whose Cook’s $D$ value exceeds $4/n$ are deemed influential ($p$ is the number of regressors including the intercept, and $n$ is the number of observations used in the analysis). If you specify the LABEL suboption of the PLOTS=DIAGNOSTICS option, then the points deemed outliers or influential are labeled on the appropriate plots.

Fit statistics are shown in the lower right of the plot and can be customized or suppressed by using the STATS= suboption of the PLOTS=DIAGNOSTICS option.
Residuals by Regressor Plots

Panels of plots of the residuals versus each of the regressors in the model are produced by default. Patterns in these plots are indications of an inadequate model. To help in detecting patterns, you can use the SMOOTH= suboption of the PLOTS=RESIDUALS option to add loess fits to these residual plots. See Figure 102.1.7 for an example.

Fit and Prediction Plots

A fit plot consisting of a scatter plot of the data overlaid with the regression line, as well as confidence and prediction limits, is produced for models depending on a single regressor. Fit statistics are shown to the right of the plot and can be customized or suppressed by using the STATS= suboption of the PLOTS=FIT option.

When a model contains more than one regressor, a fit plot is not appropriate. However, if all the regressors in the model are transformations of a single variable in the input data set, then you can request a scatter plot of the dependent variable overlaid with a fit line and confidence and prediction limits versus this variable. You can also plot residuals versus this variable. You request these plots, shown in a panel, with the PLOTS=PREDICTION option. See Figure 102.14 for an example.

Influence Plots

In addition to the “Cook’s D Plot” and the “RStudent By Leverage Plot,” you can request plots of the DFBETAS and DFFITS statistics versus observation number by using the PLOTS=DFBETAS and PLOTS=DFFITS options. You can also obtain partial regression leverage plots by using the PLOTS=PARTIAL option. See the section “Influence Statistics” on page 8516 for examples of these plots and details about their interpretation.

Ridge and VIF Plots

When you use ridge regression, you can request plots of the variance inflation factor (VIF) values and standardized ridge estimates by ridge values for each coefficient with the PLOTS=RIDGE option. See Example 102.5 for examples.

Variable Selection Plots

When you request variable selection by using the SELECTION= option in the MODEL statement, you can request plots of fit criteria for the models examined by using the PLOTS=CRITERIA option. The fit criteria are displayed versus the step number for the FORWARD, BACKWARD, and STEPWISE selection methods and the step at which the optimal value of each criterion is obtained is indicated using a “Star” marker. For the all-subset-based selection methods (SELECTION=RSQUARE|ADJRSQ|ICP), the fit criteria are displayed versus the number of observations in the model.

The criteria are shown in a panel, but you can use the UNPACK suboption of the PLOTS=CRITERIA option to obtain separate plots for each criterion. You can also use the LABEL suboption of the PLOTS=CRITERIA option to request that optimal models be labeled on the plots. Example 102.2 provides several examples.

Heat Maps

PROC REG can produce either fit and residual scatter plots for smaller data sets or heat maps for larger data sets. The global plot option MAXPOINTS=max heat-max controls which of these are produced. When the number of points exceeds the value of max but does not exceed the value of heat-max divided by the number of independent variables, heat maps are displayed instead of scatter plots for the fit and residual plots. All
other graphs are suppressed when the number of points exceeds `max`. The default is `MAXPOINTS=5000 150000`. These cutoffs are ignored if you specify `MAXPOINTS=NONE`. The following statements create both scatter plots and heat maps with artificial data:

```sas
data x;
  do i = 1 to 25000;
    x = 2 * normal(104);
    y = x + sin(x * 2) + 3 * normal(104);
    output;
  end;
run;

ods graphics on;
proc reg data=x plots(maxpoints=30000);
  model y = x;
quit;

proc reg data=x;
  model y = x;
quit;
```

Scatter plots are displayed in **Output 102.54**, and heat maps are displayed in **Output 102.55**.

**Figure 102.54**  Scatter Plot

![Residuals for y](image-url)
Figure 102.54  continued

Fit Plot for $y$

Observations 25000
Parameters 2
Error DF 24998
MSE 9.6745
R-Square 0.294
Adj R-Square 0.294

Figure 102.55  Heat Map

Residuals for $y$

N Obs = 25000
The heat maps show more clearly that the sine function is not fit well by the linear fit function.

**ODS Graph Names**

PROC REG assigns a name to each graph it creates using ODS. You can use these names to reference the graphs when using ODS. The names are listed in Table 102.13.

<table>
<thead>
<tr>
<th>ODS Graph Name</th>
<th>Plot Description</th>
<th>PLOTS Option</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdjrsqPlot</td>
<td>Adjusted R-square statistic for models examined doing variable selection</td>
<td>ADJRSQ</td>
</tr>
<tr>
<td>AICPlot</td>
<td>AIC statistic for models examined doing variable selection</td>
<td>AIC</td>
</tr>
<tr>
<td>BICPlot</td>
<td>BIC statistic for models examined doing variable selection</td>
<td>BIC</td>
</tr>
<tr>
<td>CooksDChart</td>
<td>Cook’s $D$ chart</td>
<td></td>
</tr>
<tr>
<td>CooksDPlot</td>
<td>Cook’s $D$ statistic versus observation number</td>
<td>PLOTS=RESIDUALCHART(UNPACK)</td>
</tr>
<tr>
<td>CPPlot</td>
<td>$C_p$ statistic for models examined doing variable selection</td>
<td>CP</td>
</tr>
<tr>
<td>DFFITSPlot</td>
<td>DFFITS statistics versus observation number</td>
<td>DFFITS</td>
</tr>
<tr>
<td>ODS Graph Name</td>
<td>Plot Description</td>
<td>PLOTS Option</td>
</tr>
<tr>
<td>----------------------</td>
<td>----------------------------------------------------------------------------------</td>
<td>-------------------------------</td>
</tr>
<tr>
<td>DFBETASPanel</td>
<td>Panel of DFBETAS statistics versus observation number</td>
<td>DFBETAS</td>
</tr>
<tr>
<td>DFBETASPlot</td>
<td>DFBETAS statistics versus observation number</td>
<td>DFBETAS(UNPACK)</td>
</tr>
<tr>
<td>DiagnosticsPanel</td>
<td>Panel of fit diagnostics</td>
<td>DIAGNOSTICS</td>
</tr>
<tr>
<td>FitPlot</td>
<td>Regression line, confidence limits, and prediction limits overlaid on a scatter plot of the data</td>
<td>FIT, MAXPOINTS=max not exceeded</td>
</tr>
<tr>
<td>FitPlot</td>
<td>Regression line overlaid on a heat map of the data</td>
<td>FIT, MAXPOINTS=max exceeded</td>
</tr>
<tr>
<td>ObservedByPredicted</td>
<td>Dependent variable versus predicted values</td>
<td>OBSERVEDBYPREDICTED</td>
</tr>
<tr>
<td>PartialPlot</td>
<td>Partial regression plot</td>
<td>PARTIAL</td>
</tr>
<tr>
<td>PredictionPanel</td>
<td>Panel of residuals and fit versus specified variable</td>
<td>PREDICTIONS</td>
</tr>
<tr>
<td>PredictionPlot</td>
<td>Regression line, confidence limits, and prediction limits versus specified variable</td>
<td>PREDICTIONS(UNPACK)</td>
</tr>
<tr>
<td>PredictionResidualPlot</td>
<td>Residuals versus specified variable</td>
<td>PREDICTIONS(UNPACK)</td>
</tr>
<tr>
<td>QQPlot</td>
<td>Normal quantile plot of residuals</td>
<td>QQ</td>
</tr>
<tr>
<td>ResidualBoxPlot</td>
<td>Box plot of residuals</td>
<td>BOXPLOT</td>
</tr>
<tr>
<td>ResidualByPredicted</td>
<td>Residuals versus predicted values</td>
<td>RESIDUALBYPREDICTED</td>
</tr>
<tr>
<td>ResidualHistogram</td>
<td>Histogram of fit residuals</td>
<td>RESIDUALHISTOGRAM</td>
</tr>
<tr>
<td>ResidualPlot</td>
<td>Scatter plot of residuals versus regressor</td>
<td>RESIDUALS, MAXPOINTS=max not exceeded</td>
</tr>
<tr>
<td>ResidualPlot</td>
<td>Heat map of residuals versus regressor</td>
<td>RESIDUALS, MAXPOINTS=max exceeded</td>
</tr>
<tr>
<td>RFPlot</td>
<td>Side-by-side plots of quantiles of centered fit and residuals</td>
<td>RF</td>
</tr>
<tr>
<td>RidgePanel</td>
<td>Plot of VIF and ridge traces</td>
<td>RIDGE</td>
</tr>
<tr>
<td>RidgePlot</td>
<td>Plot of ridge traces</td>
<td>RIDGE(UNPACK)</td>
</tr>
<tr>
<td>RSquarePlot</td>
<td>R-square statistic for models examined doing variable selection</td>
<td>RSQUARE</td>
</tr>
<tr>
<td>RStudentByLeverage</td>
<td>Studentized residuals versus leverage</td>
<td>RSTUDENTBYLEVERAGE</td>
</tr>
<tr>
<td>RStudentByPredicted</td>
<td>Studentized residuals versus predicted values</td>
<td>RSTUDENTBYPREDICTED</td>
</tr>
<tr>
<td>SBCPlot</td>
<td>SBC statistic for models examined doing variable selection</td>
<td>SBC</td>
</tr>
<tr>
<td>SelectionCriterionPanel</td>
<td>Panel of fit statistics for models examined doing variable selection</td>
<td>CRITERIA</td>
</tr>
<tr>
<td>StudResCooksDChart</td>
<td>Studentized residual and Cook's D chart</td>
<td>PLOTS=RESIDUALCHART</td>
</tr>
<tr>
<td>StudentResChart</td>
<td>Studentized residual chart</td>
<td>PLOTS=RESIDUALCHART(UNPACK)</td>
</tr>
<tr>
<td>VIFPlot</td>
<td>Plot of VIF traces</td>
<td>RIDGE(UNPACK)</td>
</tr>
</tbody>
</table>
Examples: REG Procedure

Example 102.1: Modeling Salaries of Major League Baseball Players

This example features the use of ODS Graphics in the process of building models by using the REG procedure and highlights the use of fit and influence diagnostics.

The Sashelp.Baseball data set contains salary and performance information for Major League Baseball players who played at least one game in both the 1986 and 1987 seasons, excluding pitchers. The salaries (Sports Illustrated, April 20, 1987) are for the 1987 season and the performance measures are from 1986 (Collier Books, The 1987 Baseball Encyclopedia Update). The following step displays in Output 102.1.1 the variables in the data set:

```plaintext
proc contents varnum data=sashelp.baseball;
  ods select position;
run;
```

Output 102.1.1  Sashelp.Baseball Data Set

The CONTENTS Procedure

Variables in Creation Order

<table>
<thead>
<tr>
<th>#</th>
<th>Variable</th>
<th>Type</th>
<th>Len</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Name</td>
<td>Char</td>
<td>18</td>
<td>Player's Name</td>
</tr>
<tr>
<td>2</td>
<td>Team</td>
<td>Char</td>
<td>14</td>
<td>Team at the End of 1986</td>
</tr>
<tr>
<td>3</td>
<td>nAtBat</td>
<td>Num</td>
<td>8</td>
<td>Times at Bat in 1986</td>
</tr>
<tr>
<td>4</td>
<td>nHits</td>
<td>Num</td>
<td>8</td>
<td>Hits in 1986</td>
</tr>
<tr>
<td>5</td>
<td>nHome</td>
<td>Num</td>
<td>8</td>
<td>Home Runs in 1986</td>
</tr>
<tr>
<td>6</td>
<td>nRuns</td>
<td>Num</td>
<td>8</td>
<td>Runs in 1986</td>
</tr>
<tr>
<td>7</td>
<td>nRBI</td>
<td>Num</td>
<td>8</td>
<td>RBIs in 1986</td>
</tr>
<tr>
<td>8</td>
<td>nBB</td>
<td>Num</td>
<td>8</td>
<td>Walks in 1986</td>
</tr>
<tr>
<td>9</td>
<td>YrMajor</td>
<td>Num</td>
<td>8</td>
<td>Years in the Major Leagues</td>
</tr>
<tr>
<td>10</td>
<td>CrAtBat</td>
<td>Num</td>
<td>8</td>
<td>Career Times at Bat</td>
</tr>
<tr>
<td>11</td>
<td>CrHits</td>
<td>Num</td>
<td>8</td>
<td>Career Hits</td>
</tr>
<tr>
<td>12</td>
<td>CrHome</td>
<td>Num</td>
<td>8</td>
<td>Career Home Runs</td>
</tr>
<tr>
<td>13</td>
<td>CrRuns</td>
<td>Num</td>
<td>8</td>
<td>Career Runs</td>
</tr>
<tr>
<td>14</td>
<td>CrRbi</td>
<td>Num</td>
<td>8</td>
<td>Career RBIs</td>
</tr>
<tr>
<td>15</td>
<td>CrBB</td>
<td>Num</td>
<td>8</td>
<td>Career Walks</td>
</tr>
<tr>
<td>16</td>
<td>League</td>
<td>Char</td>
<td>8</td>
<td>League at the End of 1986</td>
</tr>
<tr>
<td>17</td>
<td>Division</td>
<td>Char</td>
<td>8</td>
<td>Division at the End of 1986</td>
</tr>
<tr>
<td>18</td>
<td>Position</td>
<td>Char</td>
<td>8</td>
<td>Position(s) in 1986</td>
</tr>
<tr>
<td>19</td>
<td>nOuts</td>
<td>Num</td>
<td>8</td>
<td>Put Outs in 1986</td>
</tr>
<tr>
<td>20</td>
<td>nAssts</td>
<td>Num</td>
<td>8</td>
<td>Assists in 1986</td>
</tr>
<tr>
<td>21</td>
<td>nError</td>
<td>Num</td>
<td>8</td>
<td>Errors in 1986</td>
</tr>
<tr>
<td>22</td>
<td>Salary</td>
<td>Num</td>
<td>8</td>
<td>1987 Salary in $ Thousands</td>
</tr>
<tr>
<td>23</td>
<td>Div</td>
<td>Char</td>
<td>16</td>
<td>League and Division</td>
</tr>
<tr>
<td>24</td>
<td>logSalary</td>
<td>Num</td>
<td>8</td>
<td>Log Salary</td>
</tr>
</tbody>
</table>
Suppose you want to investigate whether you can model the players’ salaries for the 1987 season based on batting statistics for the previous season and lifetime batting performance. Since the variation in salaries is much greater for higher salaries, it is appropriate to apply a log transformation for this analysis. The following statements begin the analysis:

```plaintext
ods graphics on;
proc reg data=sashelp.baseball;
id name team league;
model logSalary = nhits nruns nrbi nbb yrmajor crhits;
run;
```

Output 102.1.2 shows the default output produced by PROC REG. The number of observations table shows that 59 observations are excluded because they have missing values for at least one of the variables used in the analysis. The analysis of variance and parameter estimates tables provide details about the fitted model.

Output 102.1.2  Default Output from PROC REG

The REG Procedure
Model: MODEL1
Dependent Variable: logSalary Log Salary

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>121.53052</td>
<td>20.25509</td>
<td>60.56</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>256</td>
<td>85.62322</td>
<td>0.33447</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>262</td>
<td>207.15373</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.57833  R-Square 0.5867
Dependent Mean 5.92722  Adj R-Sq 0.5770
Coeff Var 9.75719

| Variable         | Label            | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|------------------|------------------|----|--------------------|----------------|---------|------|---|
| Intercept        | Intercept        | 1  | 4.14614            | 0.13612        | 30.46   | <.0001|
| nHits            | Hits in 1986     | 1  | 0.00663            | 0.00210        | 3.15    | 0.0018|
| nRuns            | Runs in 1986     | 1  | 0.00019890         | 0.00398        | 0.05    | 0.9602|
| nRBI             | RBIs in 1986     | 1  | 0.00125            | 0.00235        | 0.53    | 0.5947|
| nBB              | Walks in 1986    | 1  | 0.00672            | 0.00239        | 2.81    | 0.0054|
| YrMajor          | Years in the Major Leagues | 1  | 0.07108            | 0.01925        | 3.69    | 0.0003|
| CrHits           | Career Hits      | 1  | 0.00023910         | 0.00014571     | 1.64    | 0.1020|

Before you accept a regression model, it is important to examine influence and fit diagnostics to see whether the model might be unduly influenced by a few observations and whether the data support the assumptions that underlie the linear regression. To facilitate such investigations, you can obtain diagnostic plots by enabling ODS Graphics.
Output 102.1.3 shows a panel of diagnostic plots. The plot of externally studentized residuals (RStudent) by leverage values reveals that there is one observation with very high leverage that might be overly influencing the fit produced. The plot of Cook’s $D$ by observation also indicates two highly influential observations. To investigate further, you can use the PLOTS= option in the PROC REG statement as follows to produce labeled versions of these plots:
Example 102.1: Modeling Salaries of Major League Baseball Players

```sas
proc reg data=sashelp.baseball
   plots(only label)=(RStudentByLeverage CooksD);
   id name team league;
   model logSalary = nhits nruns nrbi nbb yrmajor crhits;
run;
```

Output 102.1.4 and Output 102.1.5 reveal that Pete Rose is the highly influential observation. You might obtain a better fit to the remaining data if you omit his statistics when building the model.

Output 102.1.4 Outlier and Leverage Diagnostics
The following statements use a WHERE statement to omit Pete Rose’s statistics when building the model. An alternative way to do this within PROC REG is to use a REWEIGHT statement. See “Reweighting Observations in an Analysis” on page 8524 for details about reweighting.

```
proc reg data=sashelp.baseball
   plots=(RStudentByLeverage(label) residuals(smooth));
   where name^="Rose, Pete";
   id name team league;
   model logSalary = nhits nruns nrbi nbb yrmajor crhits;
run;
```

Output 102.1.6 shows the new fit diagnostics panel. You can see that there are still several influential and outlying observations. One possible reason for observing outliers is that the linear model specified is not appropriate to capture the variation in this data. You can often see evidence of an inappropriate model by observing patterns in plots of residuals.
Output 102.1.6  Fit Diagnostics

Output 102.1.7 shows plots of the residuals by the regressors in the model. When you specify the RESIDUALS(SMOOTH) suboption of the PLOTS option in the PROC REG statement, a loess fit is overlaid on each of these plots. You can see the same clear pattern in the residual plots for YrMajor and CrHits. Players near the start of their careers and players near the end of their careers get paid less than the model predicts.
You can address this lack of fit by using polynomials of degree 2 for these two variables as shown in the following statements:

```plaintext
data baseball;
  set sashelp.baseball(where=(name^="Rose, Pete");
  YrMajor2 = yrmajor*yrmajor;
  CrHits2 = crhits * crhits;
run;

proc reg data=baseball
  plots=(diagnostics(stats=none) RStudentByLeverage(label) CooksD(label) Residuals(smooth) DFFITS(label) DFBETAS ObservedByPredicted(label));
  id name team league;
  model logSalary = nhits nruns nrbi nbb yrmajor crhits
                   yrmajor2 crhits2;
run;
```
Example 102.1: Modeling Salaries of Major League Baseball Players

Output 102.1.8 shows the analysis of variance and parameter estimates for this model. Note that the R-square value of 0.787 for this model is considerably larger than the R-square value of 0.587 for the initial model shown in Output 102.1.2.

Output 102.1.8  Output from PROC REG

The REG Procedure
Model: MODEL1
Dependent Variable: logSalary Log Salary

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>8</td>
<td>162.73473</td>
<td>20.34184</td>
<td>117.13</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>253</td>
<td>43.93712</td>
<td>0.17366</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>261</td>
<td>206.67186</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.41673 R-Square 0.7874
Dependent Mean 5.92458 Adj R-Sq 0.7807
Coeff Var 7.03393

Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Label</th>
<th>DF</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>t Value</th>
<th>Pr &gt;</th>
<th>t</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>Intercept</td>
<td>1</td>
<td>3.78922</td>
<td>0.11581</td>
<td>32.72</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nHits</td>
<td>Hits in 1986</td>
<td>1</td>
<td>-0.00012753</td>
<td>0.00159</td>
<td>-0.08</td>
<td>0.9363</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nRBI</td>
<td>Runs in 1986</td>
<td>1</td>
<td>0.00215</td>
<td>0.00288</td>
<td>0.75</td>
<td>0.4549</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nBB</td>
<td>RBIs in 1986</td>
<td>1</td>
<td>0.00431</td>
<td>0.00172</td>
<td>2.51</td>
<td>0.0127</td>
<td></td>
<td></td>
</tr>
<tr>
<td>nRBI2</td>
<td>Walks in 1986</td>
<td>1</td>
<td>0.00501</td>
<td>0.00173</td>
<td>2.90</td>
<td>0.0040</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YrMajor</td>
<td>Years in the Major Leagues</td>
<td>1</td>
<td>0.23908</td>
<td>0.03443</td>
<td>6.94</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CrHits</td>
<td>Career Hits</td>
<td>1</td>
<td>0.00170</td>
<td>0.00027562</td>
<td>6.18</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>YrMajor2</td>
<td></td>
<td>1</td>
<td>-0.01440</td>
<td>0.00165</td>
<td>-8.73</td>
<td>&lt;.0001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CrHits2</td>
<td></td>
<td>1</td>
<td>-3.31739E-7</td>
<td>1.001272E-7</td>
<td>-3.31</td>
<td>0.0011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The plots of residuals by regressors in Output 102.1.9 and Output 102.1.10 show that the strong pattern in the plots for CrMajors and CrHits has been reduced, although there is still some indication of a pattern remaining in these residuals. This suggests that a quadratic function might be insufficient to capture dependence of salary on these regressors.
Output 102.1.9 Residuals by Regressors

Output 102.1.10 Residuals by Regressors
Output 102.1.11 show the diagnostics plots; three of the plots, with points of interest labeled, are shown individually in Output 102.1.12, Output 102.1.13, and Output 102.1.14. The STATS=NONE suboption specified in the PLOTS=Diagnostics option replaces the inset of statistics with a box plot of the residuals in the fit diagnostics panel. The observed by predicted value plot reveals a reasonably successful model for explaining the variation in salary for most of the players. However, the model tends to overpredict the salaries of several players near the lower end of the salary range. This bias can also be seen in the distribution of the residuals that you can see in the histogram, Q-Q plot, and box plot in Output 102.1.11.
Output 102.1.12 Outlier and Leverage Diagnostics

Output 102.1.13 Observed by Predicted Values
Example 102.1: Modeling Salaries of Major League Baseball Players

The RStudent by leverage plot in Output 102.1.12 and the Cook’s $D$ plot in Output 102.1.14 show that there are still a number of influential observations. By specifying the DFFITS and DFBETAS suboptions of the PLOTS= option, you obtain the additional influence diagnostics plots shown in Output 102.1.15 and Output 102.1.16. For more information about the interpretation of DFFITS and DFBETAS statistics, see the section “Influence Statistics” on page 8516.

You can continue this analysis by investigating how the influential observations that are identified in the various influence plots affect the fit. You can also use PROC ROBUSTREG to obtain a fit that is resistant to the presence of high-leverage points and outliers.

You can create marginal fit diagnostic plots outside PROC REG by using the %Marginal autocall macro. Marginal model plots (Cook and Weisberg 1997) display the marginal relationship between the response and each predictor. Each plot contains a scatter plot of the two variables overlaid with two curves: a smooth fit function for the variables in the plot (labeled “Data”) and the overall model predicted values (labeled “Model”). One additional plot displays predicted values on the horizontal axis rather than an independent variable. (See Output 102.1.17.) When the two curves are similar in each of the graphs, there is evidence that the model fits well. When they differ in at least one of the graphs, there is evidence that the model does not fit well.
Chapter 102: The REG Procedure

Output 102.1.15  DFFITS

Influence Diagnostics for logSalary

Output 102.1.16  DFBETAS

Influence Diagnostics for logSalary

<table>
<thead>
<tr>
<th></th>
<th>Intercept</th>
<th>nHits</th>
<th>nRuns</th>
<th>nRBI</th>
<th>nBB</th>
<th>YrMajor</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The following statements refit the last model, create a data set that contains predicted values, and use the %Marginal autocall macro to display the marginal model plots:

```sas
%let ind = nhits nruns nrbi nbb yrmajor crhits yrmajor2 crhits2;
proc reg data=baseball plots=none;
   model logSalary = &ind;
   output out=pred p=p;
run;
%marginal(dependent=logSalary, predicted=p, independents=&ind)
```

The independent variables are specified in a macro variable for convenience to avoid listing them twice; this is not required. Output 102.1.17 displays the results and shows that the model fits less well for the players who spent the most time in the major leagues.

The macro requires you to specify the dependent variable, predicted values variable, and independent variables. The macro input data set is the output data set from PROC REG. Additional options are available. For more information about the %Marginal macro, see the section “Example 22.7: Marginal Model Plots” on page 846 in Chapter 22, “ODS Graphics Template Modification.”

**Output 102.1.17** Marginal Model Plots

![Marginal Models for logSalary](image)
Example 102.2: Aerobic Fitness Prediction

Aerobic fitness (measured by the ability to consume oxygen) is fit to some simple exercise tests. The goal is to develop an equation to predict fitness based on the exercise tests rather than on expensive and cumbersome oxygen consumption measurements. Three model-selection methods are used: forward selection, backward selection, and MAXR selection. Here are the data:

`*-------------------Data on Physical Fitness-------------------*`
`| These measurements were made on men involved in a physical fitness course at N.C.State Univ. The variables are Age (years), Weight (kg), Oxygen intake rate (ml per kg body weight per minute), time to run 1.5 miles (minutes), heart rate while resting, heart rate while running (same time Oxygen rate measured), and maximum heart rate recorded while running. |
| ***Certain values of MaxPulse were changed for this analysis.*** |
`*--------------------------------------------------------------*;`

data fitness;
  input Age Weight Oxygen RunTime RestPulse RunPulse MaxPulse @@;
datalines;
  44  89.47  44.609  11.37  62  178  182  40  75.07  45.313  10.07  62  185  185
  44  85.84  45.313  9.22  55  178  180  47  77.45  44.811  11.63  58  176  176
  40  75.98  45.681  11.95  70  176  180  43  81.19  49.091  10.85  64  162  170
  44  81.42  39.442  13.08  63  174  176  38  81.87  60.055  8.17  40  166  186
  44  73.03  50.541  10.13  45  168  168  45  87.66  37.388  14.03  56  186  192
  45  66.45  44.754  11.12  51  176  176  47  79.15  47.273  10.60  47  162  164
  54  83.12  51.855  10.33  50  166  170  49  81.42  49.156  8.95  44  180  185
  51  69.63  40.836  10.95  57  168  172  51  77.91  46.672  10.00  48  162  168
  48  91.63  46.774  10.25  48  162  164  49  73.37  50.388  10.08  67  168  168
  57  73.37  39.407  12.63  58  174  176  54  79.38  46.080  11.17  62  156  165
  52  76.32  45.441  9.63  48  164  166  50  70.87  54.625  8.92  48  146  155
  51  87.25  45.118  11.08  48  172  172  54  91.63  39.203  12.88  44  168  172
  51  73.71  45.790  10.47  59  186  188  57  59.08  50.545  9.93  49  148  155
  49  76.32  48.673  9.40  56  186  188  48  61.24  47.920  11.50  52  170  176
  52  82.78  47.467  10.50  53  170  172 ;`

The following statements demonstrate the FORWARD, BACKWARD, and MAXR model selection methods:

```
proc reg data=fitness;
  model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse / selection=forward;
  model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse / selection=backward;
  model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse / selection=maxr;
run;
```

Output 102.2.1 shows the sequence of models produced by the FORWARD model-selection method.
Output 102.2.1  Forward Selection Method: PROC REG

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Forward Selection: Step 1
Variable RunTime Entered: R-Square = 0.7434 and C(p) = 13.6988

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>632.90010</td>
<td>632.90010</td>
<td>84.01</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>29</td>
<td>218.48144</td>
<td>7.53384</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bounds on condition number: 1, 1

Forward Selection: Step 2
Variable Age Entered: R-Square = 0.7642 and C(p) = 12.3894

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>650.66573</td>
<td>325.33287</td>
<td>45.38</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>200.71581</td>
<td>7.16842</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>88.46229</td>
<td>5.37264</td>
<td>1943.41071</td>
<td>271.11</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.15037</td>
<td>0.09551</td>
<td>17.76563</td>
<td>2.48</td>
<td>0.1267</td>
</tr>
<tr>
<td>RunTime</td>
<td>-3.20395</td>
<td>0.35877</td>
<td>571.67751</td>
<td>79.75</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Bounds on condition number: 1.0369, 4.1478

Forward Selection: Step 3
Variable RunPulse Entered: R-Square = 0.8111 and C(p) = 6.9596

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>690.55086</td>
<td>230.18362</td>
<td>38.64</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>160.83069</td>
<td>5.95669</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Output 102.2.1 continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>111.71806</td>
<td>10.23509</td>
<td>709.69014</td>
<td>119.14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.25640</td>
<td>0.09623</td>
<td>42.28867</td>
<td>7.10</td>
<td>0.0129</td>
</tr>
<tr>
<td>RunTime</td>
<td>-2.82538</td>
<td>0.35828</td>
<td>370.43529</td>
<td>62.19</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>RunPulse</td>
<td>-0.13091</td>
<td>0.05059</td>
<td>39.88512</td>
<td>6.70</td>
<td>0.0154</td>
</tr>
</tbody>
</table>

**Bounds on condition number:** 1.3548, 11.597

**Forward Selection: Step 4**

**Variable MaxPulse Entered: R-Square = 0.8368 and C(p) = 4.8800**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>712.45153</td>
<td>178.11288</td>
<td>33.33</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>26</td>
<td>138.93002</td>
<td>5.34346</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>98.14789</td>
<td>11.78569</td>
<td>370.57373</td>
<td>69.35</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.19773</td>
<td>0.09564</td>
<td>22.84231</td>
<td>4.27</td>
<td>0.0488</td>
</tr>
<tr>
<td>RunTime</td>
<td>-2.76758</td>
<td>0.34054</td>
<td>352.93570</td>
<td>66.05</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>RunPulse</td>
<td>-0.34811</td>
<td>0.11750</td>
<td>46.90089</td>
<td>8.78</td>
<td>0.0064</td>
</tr>
<tr>
<td>MaxPulse</td>
<td>0.27051</td>
<td>0.13362</td>
<td>21.90067</td>
<td>4.10</td>
<td>0.0533</td>
</tr>
</tbody>
</table>

**Bounds on condition number:** 8.4182, 76.851

**Forward Selection: Step 5**

**Variable Weight Entered: R-Square = 0.8480 and C(p) = 5.1063**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>721.97309</td>
<td>144.39462</td>
<td>27.90</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25</td>
<td>129.40845</td>
<td>5.17634</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>102.20428</td>
<td>11.97929</td>
<td>376.78935</td>
<td>72.79</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.21962</td>
<td>0.09550</td>
<td>27.37429</td>
<td>5.29</td>
<td>0.0301</td>
</tr>
<tr>
<td>Weight</td>
<td>-0.07230</td>
<td>0.05331</td>
<td>9.52157</td>
<td>1.84</td>
<td>0.1871</td>
</tr>
<tr>
<td>RunTime</td>
<td>-2.68252</td>
<td>0.34099</td>
<td>320.35968</td>
<td>61.89</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>RunPulse</td>
<td>-0.37340</td>
<td>0.11714</td>
<td>52.59624</td>
<td>10.16</td>
<td>0.0038</td>
</tr>
<tr>
<td>MaxPulse</td>
<td>0.30491</td>
<td>0.13394</td>
<td>26.82640</td>
<td>5.18</td>
<td>0.0316</td>
</tr>
</tbody>
</table>

**Bounds on condition number:** 8.7312, 104.83
The final variable available to add to the model, RestPulse, is not added since it does not meet the 50% (the default value of the SLE option is 0.5 for FORWARD selection) significance-level criterion for entry into the model.

The BACKWARD model-selection method begins with the full model. Output 102.2.2 shows the steps of the BACKWARD method. RestPulse is the first variable deleted, followed by Weight. No other variables are deleted from the model since the variables remaining (Age, RunTime, RunPulse, and MaxPulse) are all significant at the 10% (the default value of the SLS option is 0.1 for the BACKWARD elimination method) significance level.

**Output 102.2.2** Backward Selection Method: PROC REG

**Backward Elimination: Step 0**

All Variables Entered: R-Square = 0.8487 and C(p) = 7.0000

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>722.54361</td>
<td>120.42393</td>
<td>22.43</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>128.83794</td>
<td>5.36825</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bounds on condition number: 8.7438, 137.13

**Backward Elimination: Step 1**

Variable RestPulse Removed: R-Square = 0.8480 and C(p) = 5.1063

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>102.93448</td>
<td>12.40326</td>
<td>369.72831</td>
<td>68.87</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.22697</td>
<td>0.09984</td>
<td>27.74577</td>
<td>5.17</td>
<td>0.0322</td>
</tr>
<tr>
<td>Weight</td>
<td>-0.07418</td>
<td>0.05459</td>
<td>9.91059</td>
<td>1.85</td>
<td>0.1869</td>
</tr>
<tr>
<td>RunTime</td>
<td>-2.62865</td>
<td>0.38456</td>
<td>250.82210</td>
<td>46.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>RunPulse</td>
<td>-0.36963</td>
<td>0.11985</td>
<td>51.05806</td>
<td>9.51</td>
<td>0.0051</td>
</tr>
<tr>
<td>RestPulse</td>
<td>-0.02153</td>
<td>0.06605</td>
<td>0.57051</td>
<td>0.11</td>
<td>0.7473</td>
</tr>
<tr>
<td>MaxPulse</td>
<td>0.30322</td>
<td>0.13650</td>
<td>26.49142</td>
<td>4.93</td>
<td>0.0360</td>
</tr>
</tbody>
</table>

Bounds on condition number: 8.7438, 137.13
The MAXR method tries to find the “best” one-variable model, the “best” two-variable model, and so on. Output 102.2.3 shows that the one-variable model contains RunTime; the two-variable model contains RunTime and Age; the three-variable model contains RunTime, Age, and RunPulse; the four-variable model contains Age, RunTime, RunPulse, and MaxPulse; the five-variable model contains Age, Weight, RunTime, RunPulse, and MaxPulse; and finally, the six-variable model contains all the variables in the MODEL statement.
### Example 102.2: Aerobic Fitness Prediction

#### Variable Selection

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>82.42177</td>
<td>3.85530</td>
<td>3443.36654</td>
<td>457.05</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>RunTime</td>
<td>-3.31056</td>
<td>0.36119</td>
<td>632.90010</td>
<td>84.01</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

**Bounds on condition number:** 1, 1

The above model is the best 1-variable model found.

### Maximum R-Square Improvement: Step 2

Variable Age Entered: R-Square = 0.7642 and C(p) = 12.3894

#### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>650.66573</td>
<td>325.33287</td>
<td>45.38</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>28</td>
<td>200.71581</td>
<td>7.16842</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Bounds on condition number:** 1.0369, 4.1478

The above model is the best 2-variable model found.

### Maximum R-Square Improvement: Step 3

Variable RunPulse Entered: R-Square = 0.8111 and C(p) = 6.9596

#### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>690.55086</td>
<td>230.18362</td>
<td>38.64</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>27</td>
<td>160.83069</td>
<td>5.95669</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>111.71806</td>
<td>10.23509</td>
<td>709.69014</td>
<td>119.14</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.25640</td>
<td>0.09623</td>
<td>42.28867</td>
<td>7.10</td>
<td>0.0129</td>
</tr>
<tr>
<td>RunTime</td>
<td>-2.82538</td>
<td>0.35828</td>
<td>370.43529</td>
<td>62.19</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>RunPulse</td>
<td>-0.13091</td>
<td>0.05059</td>
<td>39.88512</td>
<td>6.70</td>
<td>0.0154</td>
</tr>
</tbody>
</table>
Output 102.2.3 continued

Bounds on condition number: 1.3548, 11.597

The above model is the best 3-variable model found.

Maximum R-Square Improvement: Step 4

Variable MaxPulse Entered: R-Square = 0.8368 and C(p) = 4.8800

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>4</td>
<td>712.45153</td>
<td>178.11288</td>
<td>33.33</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>26</td>
<td>138.93002</td>
<td>5.34346</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bounds on condition number: 8.4182, 76.851

The above model is the best 4-variable model found.

Maximum R-Square Improvement: Step 5

Variable Weight Entered: R-Square = 0.8480 and C(p) = 5.1063

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>721.97309</td>
<td>144.39462</td>
<td>27.90</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>25</td>
<td>129.40845</td>
<td>5.17634</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bounds on condition number: 8.4182, 76.851

The above model is the best 4-variable model found.
Output 102.2.3 continued

Bounds on condition number: 8.7312, 104.83

The above model is the best 5-variable model found.

Maximum R-Square Improvement: Step 6

Variable RestPulse Entered: R-Square = 0.8487 and C(p) = 7.0000

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>6</td>
<td>722.54361</td>
<td>120.42393</td>
<td>22.43</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>24</td>
<td>128.83794</td>
<td>5.36825</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variable</th>
<th>Parameter Estimate</th>
<th>Standard Error</th>
<th>Type II SS</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>102.93448</td>
<td>12.40326</td>
<td>369.72831</td>
<td>68.87</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Age</td>
<td>-0.22697</td>
<td>0.09984</td>
<td>27.74577</td>
<td>5.17</td>
<td>0.0322</td>
</tr>
<tr>
<td>Weight</td>
<td>-0.07418</td>
<td>0.05459</td>
<td>9.91059</td>
<td>1.85</td>
<td>0.1869</td>
</tr>
<tr>
<td>RunTime</td>
<td>-2.62865</td>
<td>0.38456</td>
<td>250.82210</td>
<td>46.72</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>RunPulse</td>
<td>-0.36963</td>
<td>0.11985</td>
<td>51.05806</td>
<td>9.51</td>
<td>0.0051</td>
</tr>
<tr>
<td>RestPulse</td>
<td>-0.02153</td>
<td>0.06605</td>
<td>0.57051</td>
<td>0.11</td>
<td>0.7473</td>
</tr>
<tr>
<td>MaxPulse</td>
<td>0.30322</td>
<td>0.13650</td>
<td>26.49142</td>
<td>4.93</td>
<td>0.0360</td>
</tr>
</tbody>
</table>

Bounds on condition number: 8.7438, 137.13

Note that for all three of these methods, RestPulse contributes least to the model. In the case of forward selection, it is not added to the model. In the case of backward selection, it is the first variable to be removed from the model. In the case of MAXR selection, RestPulse is included only for the full model.

For the STEPWISE, BACKWARD, and FORWARD selection methods, you can control the amount of detail displayed by using the DETAILS option, and you can use ODS Graphics to produce plots that show how selection criteria progress as the selection proceeds. For example, the following statements display only the selection summary table for the FORWARD selection method (Output 102.2.4) and produce the plots shown in Output 102.2.5 and Output 102.2.6.

```plaintext
ods graphics on;
proc reg data=fitness plots=(criteria sbc);
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
      / selection=forward details=summary;
run;
```
Output 102.2.4 Forward Selection Summary

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable Entered</th>
<th>Number Vars In</th>
<th>Partial R-Square</th>
<th>Model R-Square</th>
<th>C(p)</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>RunTime</td>
<td>1</td>
<td>0.7434</td>
<td>0.7434</td>
<td>13.6988</td>
<td>84.01</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>2</td>
<td>Age</td>
<td>2</td>
<td>0.0209</td>
<td>0.7642</td>
<td>12.3894</td>
<td>2.48</td>
<td>0.1267</td>
</tr>
<tr>
<td>3</td>
<td>RunPulse</td>
<td>3</td>
<td>0.0468</td>
<td>0.8111</td>
<td>6.9596</td>
<td>6.70</td>
<td>0.0154</td>
</tr>
<tr>
<td>4</td>
<td>MaxPulse</td>
<td>4</td>
<td>0.0257</td>
<td>0.8368</td>
<td>4.8800</td>
<td>4.10</td>
<td>0.0533</td>
</tr>
<tr>
<td>5</td>
<td>Weight</td>
<td>5</td>
<td>0.0112</td>
<td>0.8480</td>
<td>5.1063</td>
<td>1.84</td>
<td>0.1871</td>
</tr>
</tbody>
</table>

Output 102.2.5 show how six fit criteria progress as the forward selection proceeds. The step at which each criterion achieves its best value is indicated. For example, the BIC criterion achieves its minimum value for the model at step 4. Note that this does not mean that the model at step 4 achieves the smallest BIC criterion among all possible models that use a subset of the regressors; the model at step 4 yields the smallest BIC statistic among the models at each step of the forward selection. Output 102.2.6 show the progression of the SBC statistic in its own plot. If you want to see six of the selection criteria in individual plots, you can specify the UNPACK suboption of the PLOTS=CRITERIA option in the PROC REG statement.

Output 102.2.5 Fit Criteria
Next, the RSQUARE model-selection method is used to request R square and $C_p$ statistics for all possible combinations of the six independent variables. The following statements produce Output 102.2.7:

```plaintext
proc reg data=fitness plots=(criteria(label) cp);
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse /
       selection=rsquare cp;
   title 'Physical fitness data: all models';
run;
```
## Output 102.2.7
All Models by the RSQUARE Method: PROC REG

### Physical fitness data: all models

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

R-Square Selection Method

<table>
<thead>
<tr>
<th>Model Index</th>
<th>Number in Model</th>
<th>R-Square</th>
<th>C(p)</th>
<th>Variables in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0.7434</td>
<td>13.6988</td>
<td>RunTime</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.1595</td>
<td>106.3021</td>
<td>RestPulse</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.1584</td>
<td>106.4769</td>
<td>RunPulse</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.0928</td>
<td>116.8818</td>
<td>Age</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0.0560</td>
<td>122.7072</td>
<td>MaxPulse</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>0.0265</td>
<td>127.3948</td>
<td>Weight</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.7642</td>
<td>12.3894</td>
<td>Age RunTime</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0.7614</td>
<td>12.8372</td>
<td>RunTime RunPulse</td>
</tr>
<tr>
<td>9</td>
<td>2</td>
<td>0.7452</td>
<td>15.4069</td>
<td>RunTime MaxPulse</td>
</tr>
<tr>
<td>10</td>
<td>2</td>
<td>0.7449</td>
<td>15.4523</td>
<td>Weight RunTime</td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td>0.7435</td>
<td>15.6746</td>
<td>RunTime RestPulse</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>0.3760</td>
<td>73.9645</td>
<td>Age RestPulse</td>
</tr>
<tr>
<td>13</td>
<td>2</td>
<td>0.3003</td>
<td>85.9742</td>
<td>Age RestPulse</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>0.2894</td>
<td>87.6951</td>
<td>RunPulse MaxPulse</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>0.2600</td>
<td>92.3638</td>
<td>Age MaxPulse</td>
</tr>
<tr>
<td>16</td>
<td>2</td>
<td>0.2350</td>
<td>96.3209</td>
<td>RunPulse RestPulse</td>
</tr>
<tr>
<td>17</td>
<td>2</td>
<td>0.1806</td>
<td>104.9523</td>
<td>Weight RestPulse</td>
</tr>
<tr>
<td>18</td>
<td>2</td>
<td>0.1740</td>
<td>105.9939</td>
<td>RestPulse MaxPulse</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>0.1669</td>
<td>107.1332</td>
<td>Weight RunPulse</td>
</tr>
<tr>
<td>20</td>
<td>2</td>
<td>0.1506</td>
<td>109.7057</td>
<td>Age Weight</td>
</tr>
<tr>
<td>21</td>
<td>2</td>
<td>0.0675</td>
<td>122.8881</td>
<td>Weight MaxPulse</td>
</tr>
<tr>
<td>22</td>
<td>3</td>
<td>0.8111</td>
<td>6.9596</td>
<td>Age RunTime RunPulse</td>
</tr>
<tr>
<td>23</td>
<td>3</td>
<td>0.8100</td>
<td>7.1350</td>
<td>RunTime RunPulse MaxPulse</td>
</tr>
<tr>
<td>24</td>
<td>3</td>
<td>0.7817</td>
<td>11.6167</td>
<td>Age RunTime MaxPulse</td>
</tr>
<tr>
<td>25</td>
<td>3</td>
<td>0.7708</td>
<td>13.3453</td>
<td>Age Weight RunTime</td>
</tr>
<tr>
<td>26</td>
<td>3</td>
<td>0.7673</td>
<td>13.8974</td>
<td>Age RunTime RestPulse</td>
</tr>
<tr>
<td>27</td>
<td>3</td>
<td>0.7619</td>
<td>14.7619</td>
<td>RunTime RunPulse RestPulse</td>
</tr>
<tr>
<td>28</td>
<td>3</td>
<td>0.7618</td>
<td>14.7729</td>
<td>Weight RunTime RunPulse</td>
</tr>
<tr>
<td>29</td>
<td>3</td>
<td>0.7462</td>
<td>17.2588</td>
<td>Weight RunTime MaxPulse</td>
</tr>
<tr>
<td>30</td>
<td>3</td>
<td>0.7452</td>
<td>17.4060</td>
<td>RunTime RestPulse MaxPulse</td>
</tr>
<tr>
<td>31</td>
<td>3</td>
<td>0.7451</td>
<td>17.4243</td>
<td>Weight RunTime RestPulse</td>
</tr>
<tr>
<td>32</td>
<td>3</td>
<td>0.4666</td>
<td>61.5873</td>
<td>Age RunPulse RestPulse</td>
</tr>
<tr>
<td>33</td>
<td>3</td>
<td>0.4223</td>
<td>68.6250</td>
<td>Age RunPulse MaxPulse</td>
</tr>
<tr>
<td>34</td>
<td>3</td>
<td>0.4091</td>
<td>70.7102</td>
<td>Age Weight RunPulse</td>
</tr>
<tr>
<td>35</td>
<td>3</td>
<td>0.3900</td>
<td>73.7424</td>
<td>Age RestPulse MaxPulse</td>
</tr>
<tr>
<td>36</td>
<td>3</td>
<td>0.3568</td>
<td>79.0013</td>
<td>Age Weight RestPulse</td>
</tr>
<tr>
<td>37</td>
<td>3</td>
<td>0.3538</td>
<td>79.4891</td>
<td>RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>38</td>
<td>3</td>
<td>0.3208</td>
<td>84.7216</td>
<td>Weight RunPulse MaxPulse</td>
</tr>
<tr>
<td>39</td>
<td>3</td>
<td>0.2902</td>
<td>89.5693</td>
<td>Age Weight MaxPulse</td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>0.2447</td>
<td>96.7952</td>
<td>Weight RunPulse RestPulse</td>
</tr>
<tr>
<td>41</td>
<td>3</td>
<td>0.1882</td>
<td>105.7430</td>
<td>Weight RestPulse MaxPulse</td>
</tr>
</tbody>
</table>
Output 102.2.7 continued

Physical fitness data: all models

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

R-Square Selection Method

<table>
<thead>
<tr>
<th>Model Index</th>
<th>Number in Model</th>
<th>R-Square</th>
<th>C(p)</th>
<th>Variables in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>42</td>
<td>4</td>
<td>0.8368</td>
<td>4.8800</td>
<td>Age RunTime RunPulse MaxPulse</td>
</tr>
<tr>
<td>43</td>
<td>4</td>
<td>0.8165</td>
<td>8.1035</td>
<td>Age Weight RunTime RunPulse</td>
</tr>
<tr>
<td>44</td>
<td>4</td>
<td>0.8158</td>
<td>8.2056</td>
<td>Weight RunTime RunPulse MaxPulse</td>
</tr>
<tr>
<td>45</td>
<td>4</td>
<td>0.8117</td>
<td>8.8683</td>
<td>Age RunTime RunPulse RestPulse</td>
</tr>
<tr>
<td>46</td>
<td>4</td>
<td>0.8104</td>
<td>9.0697</td>
<td>RunTime RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>47</td>
<td>4</td>
<td>0.7862</td>
<td>12.9039</td>
<td>Age Weight RunTime MaxPulse</td>
</tr>
<tr>
<td>48</td>
<td>4</td>
<td>0.7834</td>
<td>13.3468</td>
<td>Age RunTime RestPulse MaxPulse</td>
</tr>
<tr>
<td>49</td>
<td>4</td>
<td>0.7750</td>
<td>14.6788</td>
<td>Age Weight RunTime RestPulse</td>
</tr>
<tr>
<td>50</td>
<td>4</td>
<td>0.7623</td>
<td>16.7058</td>
<td>Weight RunTime RunPulse RestPulse</td>
</tr>
<tr>
<td>51</td>
<td>4</td>
<td>0.7462</td>
<td>19.2550</td>
<td>Weight RunTime RestPulse MaxPulse</td>
</tr>
<tr>
<td>52</td>
<td>4</td>
<td>0.5034</td>
<td>57.7590</td>
<td>Age Weight RunPulse RestPulse</td>
</tr>
<tr>
<td>53</td>
<td>4</td>
<td>0.5025</td>
<td>57.9092</td>
<td>Age RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>54</td>
<td>4</td>
<td>0.4717</td>
<td>62.7830</td>
<td>Age Weight RunPulse MaxPulse</td>
</tr>
<tr>
<td>55</td>
<td>4</td>
<td>0.4256</td>
<td>70.0963</td>
<td>Age Weight RestPulse MaxPulse</td>
</tr>
<tr>
<td>56</td>
<td>4</td>
<td>0.3858</td>
<td>76.4100</td>
<td>Weight RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>57</td>
<td>5</td>
<td>0.8480</td>
<td>5.1063</td>
<td>Age Weight RunTime RunPulse MaxPulse</td>
</tr>
<tr>
<td>58</td>
<td>5</td>
<td>0.8370</td>
<td>6.8461</td>
<td>Age RunTime RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>59</td>
<td>5</td>
<td>0.8176</td>
<td>9.9348</td>
<td>Age Weight RunTime RunPulse RestPulse</td>
</tr>
<tr>
<td>60</td>
<td>5</td>
<td>0.8161</td>
<td>10.1685</td>
<td>Weight RunTime RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>61</td>
<td>5</td>
<td>0.7887</td>
<td>14.5111</td>
<td>Age Weight RunTime RestPulse MaxPulse</td>
</tr>
<tr>
<td>62</td>
<td>5</td>
<td>0.5541</td>
<td>51.7233</td>
<td>Age Weight RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>63</td>
<td>6</td>
<td>0.8487</td>
<td>7.0000</td>
<td>Age Weight RunTime RunPulse RestPulse MaxPulse</td>
</tr>
</tbody>
</table>

The models in Output 102.2.7 are arranged first by the number of variables in the model and then by the magnitude of R square for the model.
Output 102.2.8 shows the panel of fit criteria for the RSQUARE selection method. The best models (based on the R-square statistic) for each subset size are indicated on the plots. The LABEL suboption specifies that these models are labeled by the model number that appears in the summary table shown in Output 102.2.7.

**Output 102.2.8** Fit Criteria

Output 102.2.9 shows the plot of the \( C_p \) criterion by number of regressors in the model. Useful reference lines suggested by Mallows (1973) and Hocking (1976) are included on the plot. However, because all possible subset models are included on this plot, the better models are all compressed near the bottom of the plot.
The following statements use the BEST=20 option in the model statement and SELECTION=CP to restrict attention to the models that yield the 20 smallest values of the $C_p$ statistic:

```sas
proc reg data=fitness plots(only)=cp(label);
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
       / selection=cp best=20;
run;
ods graphics off;
```
Output 102.2.10 shows the summary table listing the regressors in the 20 models that yield the smallest $C_p$ values, and Output 102.2.11 presents the results graphically. Reference lines $C_p = 2p - p_{full}$ and $C_p = p$ are shown on this plot. See the PLOTS=CP option on page 8456 for interpretations of these lines. For the Fitness data, these lines indicate that a six-variable model is a reasonable choice for doing parameter estimation, while a five-variable model might be suitable for doing prediction.

**Output 102.2.10**  
$C_p$ Selection Summary: PROC REG

**The REG Procedure**  
**Model:** MODEL1  
**Dependent Variable:** Oxygen

<table>
<thead>
<tr>
<th>Model Index</th>
<th>Number in Model</th>
<th>C(p)</th>
<th>R-Square</th>
<th>Variables in Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>4.8800</td>
<td>0.8368</td>
<td>Age RunTime RunPulse MaxPulse</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>5.1063</td>
<td>0.8480</td>
<td>Age Weight RunTime RunPulse MaxPulse</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>6.8461</td>
<td>0.8370</td>
<td>Age RunTime RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>6.9596</td>
<td>0.8111</td>
<td>Age RunTime RunPulse</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>7.0000</td>
<td>0.8487</td>
<td>Age Weight RunTime RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>7.1350</td>
<td>0.8100</td>
<td>RunTime RunPulse MaxPulse</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>8.1035</td>
<td>0.8165</td>
<td>Age Weight RunTime RunPulse</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>8.2056</td>
<td>0.8158</td>
<td>Weight RunTime RunPulse MaxPulse</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>8.8683</td>
<td>0.8117</td>
<td>Age RunTime RunPulse RestPulse</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>9.0697</td>
<td>0.8104</td>
<td>RunTime RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>11</td>
<td>5</td>
<td>9.9348</td>
<td>0.8176</td>
<td>Age Weight RunTime RunPulse RestPulse</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>10.1685</td>
<td>0.8161</td>
<td>Weight RunTime RunPulse RestPulse MaxPulse</td>
</tr>
<tr>
<td>13</td>
<td>3</td>
<td>11.6167</td>
<td>0.7817</td>
<td>Age RunTime MaxPulse</td>
</tr>
<tr>
<td>14</td>
<td>2</td>
<td>12.3894</td>
<td>0.7642</td>
<td>Age RunTime</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>12.8372</td>
<td>0.7614</td>
<td>RunTime RunPulse</td>
</tr>
<tr>
<td>16</td>
<td>4</td>
<td>12.9039</td>
<td>0.7862</td>
<td>Age Weight RunTime MaxPulse</td>
</tr>
<tr>
<td>17</td>
<td>3</td>
<td>13.3453</td>
<td>0.7708</td>
<td>Age Weight RunTime</td>
</tr>
<tr>
<td>18</td>
<td>4</td>
<td>13.3468</td>
<td>0.7834</td>
<td>Age RunTime RestPulse MaxPulse</td>
</tr>
<tr>
<td>19</td>
<td>1</td>
<td>13.6988</td>
<td>0.7434</td>
<td>RunTime</td>
</tr>
<tr>
<td>20</td>
<td>3</td>
<td>13.8974</td>
<td>0.7673</td>
<td>Age RunTime RestPulse</td>
</tr>
</tbody>
</table>
Before making a final decision about which model to use, you would want to perform collinearity diagnostics. Note that, since many different models have been fit and the choice of a final model is based on \( R \) square, the statistics are biased and the \( p \)-values for the parameter estimates are not valid.
Example 102.3: Predicting Weight by Height and Age

In this example, the weights of schoolchildren are modeled as a function of their heights and ages. The example shows the use of a BY statement with PROC REG, multiple MODEL statements, and the OUTEST= and OUTSSCP= options, which create data sets. Here are the data:

*-------------Data on Age, Weight, and Height of Children----------*
| Age (months), height (inches), and weight (pounds) were       |
| recorded for a group of school children.                      |
| From Lewis and Taylor (1967).                                 |
*-----------------------------------------------------------------

    data htwt;
      input sex $ age :3.1 height weight @@;
      datalines;
    f 143 56.3  85.0 f 155 62.3 105.0 f 153 63.3 108.0 f 161 59.0 92.0
    f 191 62.5 112.5 f 171 62.5 112.0 f 185 59.0 104.0 f 142 56.5 69.0
    f 160 62.0  94.5 f 140 53.8  68.5 f 139 61.5 104.0 f 178 61.5 103.5
    f 157 64.5 123.5 f 149 58.3  93.0 f 143 51.3  50.5 f 145 58.8  89.0
    f 191 65.3 107.0 f 150 59.5  78.5 f 147 61.3 115.0 f 180 63.3 114.0
    ... more lines ...
    m 164 66.5 112.0 m 189 65.0 114.0 m 164 61.5 140.0 m 167 62.0 107.5
    m 151 59.3  87.0
    ;

Modeling is performed separately for boys and girls. Since the BY statement is used, interactive processing is not possible in this example; no statements can appear after the first RUN statement.

The following statements produce Output 102.3.1 through Output 102.3.4:

    proc reg outest=est1 outsscp=sscp1 rsquare;
      by sex;
      eq1: model weight=height;
      eq2: model weight=height age;
      run;

    proc print data=sscp1;
      title2 'SSCP type data set';
      run;

    proc print data=est1;
      title2 'EST type data set';
      run;
### Output 102.3.1  Height and Weight Data: Submodel for Female Children

**The REG Procedure**  
**Model: eq1**  
**Dependent Variable: weight**

```plaintext
sex=f
```

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>21507</td>
<td>21507</td>
<td>141.09</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>109</td>
<td>16615</td>
<td>152.42739</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>110</td>
<td>38121</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 12.34615  
R-Square 0.5642  
Dependent Mean 98.87838  
Adj R-Sq 0.5602  
Coeff Var 12.48620

| Parameter Estimates | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|--------------------|--------------------|----------------|---------|------|---|
| Intercept          | -153.12891         | 21.24814       | -7.21   | <.0001 |
| height             | 4.16361            | 0.35052        | 11.88   | <.0001 |

### Output 102.3.2  Height and Weight Data: Full Model for Female Children

**The REG Procedure**  
**Model: eq2**  
**Dependent Variable: weight**

```plaintext
sex=f
```

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>22432</td>
<td>11216</td>
<td>77.21</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>108</td>
<td>15689</td>
<td>145.26700</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>110</td>
<td>38121</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 12.05268  
R-Square 0.5884  
Dependent Mean 98.87838  
Adj R-Sq 0.5808  
Coeff Var 12.18939

| Parameter Estimates | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|--------------------|--------------------|----------------|---------|------|---|
| Intercept          | -150.59698         | 20.76730       | -7.25   | <.0001 |
| height             | 3.60378            | 0.40777        | 8.84    | <.0001 |
| age                | 1.90703            | 0.75543        | 2.52    | 0.0130 |
### Output 102.3.3 Height and Weight Data: Submodel for Male Children

**The REG Procedure**  
**Model:** eq1  
**Dependent Variable:** weight

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>1</td>
<td>31126</td>
<td>31126</td>
<td>206.24</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>124</td>
<td>18714</td>
<td>150.92222</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>125</td>
<td>49840</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Root MSE</th>
<th>R-Square</th>
<th>Dependent Mean</th>
<th>Adj R-Sq</th>
<th>Coeff Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.28504</td>
<td>0.6245</td>
<td>103.44841</td>
<td>0.6215</td>
<td>11.87552</td>
</tr>
</tbody>
</table>

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|-------------------|----------------|---------|------|
| Intercept | 1  | -125.69807        | 15.99362       | -7.86   | <.0001|
| height   | 1  | 3.68977           | 0.25693        | 14.36   | <.0001|

### Output 102.3.4 Height and Weight Data: Full Model for Male Children

**The REG Procedure**  
**Model:** eq2  
**Dependent Variable:** weight

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>32975</td>
<td>16487</td>
<td>120.24</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>123</td>
<td>16866</td>
<td>137.11922</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>125</td>
<td>49840</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Root MSE</th>
<th>R-Square</th>
<th>Dependent Mean</th>
<th>Adj R-Sq</th>
<th>Coeff Var</th>
</tr>
</thead>
<tbody>
<tr>
<td>11.70979</td>
<td>0.6616</td>
<td>103.44841</td>
<td>0.6561</td>
<td>11.31945</td>
</tr>
</tbody>
</table>

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|-------------------|----------------|---------|------|
| Intercept | 1  | -113.71346       | 15.59021       | -7.29   | <.0001|
| height   | 1  | 2.68075          | 0.36809        | 7.28    | <.0001|
| age      | 1  | 3.08167          | 0.83927        | 3.67    | 0.0004|
For both female and male children, the overall $F$ statistics for both models are significant, indicating that the model explains a significant portion of the variation in the data. For females, the full model is

$$\text{weight} = -150.57 + 3.60 \times \text{height} + 1.91 \times \text{age}$$

and for males, the full model is

$$\text{weight} = -113.71 + 2.68 \times \text{height} + 3.08 \times \text{age}$$

The OUTSSCP= data set is shown in Output 102.3.5. Note how the BY groups are separated. Observations with _TYPE_='N' contain the number of observations in the associated BY group. Observations with _TYPE_='SSCP' contain the rows of the uncorrected sums of squares and crossproducts matrix. The observations with _NAME_='Intercept' contain crossproducts for the intercept.

Output 102.3.5 SSCP Matrix

<table>
<thead>
<tr>
<th>Obs</th>
<th>sex</th>
<th><em>TYPE</em></th>
<th><em>NAME</em></th>
<th>Intercept</th>
<th>height</th>
<th>weight</th>
<th>age</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 f</td>
<td>SSCP</td>
<td>Intercept</td>
<td>111.0</td>
<td>6718.40</td>
<td>10975.50</td>
<td>1824.90</td>
<td></td>
</tr>
<tr>
<td>2 f</td>
<td>SSCP</td>
<td>height</td>
<td>6718.4</td>
<td>407879.32</td>
<td>669469.85</td>
<td>110818.32</td>
<td></td>
</tr>
<tr>
<td>3 f</td>
<td>SSCP</td>
<td>weight</td>
<td>10975.5</td>
<td>669469.85</td>
<td>1123360.75</td>
<td>182444.95</td>
<td></td>
</tr>
<tr>
<td>4 f</td>
<td>SSCP</td>
<td>age</td>
<td>1824.9</td>
<td>110818.32</td>
<td>182444.95</td>
<td>30363.81</td>
<td></td>
</tr>
<tr>
<td>5 f</td>
<td>N</td>
<td></td>
<td>111.0</td>
<td>111.00</td>
<td>111.00</td>
<td>111.00</td>
<td></td>
</tr>
<tr>
<td>6 m</td>
<td>SSCP</td>
<td>Intercept</td>
<td>126.0</td>
<td>7825.00</td>
<td>13034.50</td>
<td>2072.10</td>
<td></td>
</tr>
<tr>
<td>7 m</td>
<td>SSCP</td>
<td>height</td>
<td>7825.0</td>
<td>488243.60</td>
<td>817919.60</td>
<td>129432.57</td>
<td></td>
</tr>
<tr>
<td>8 m</td>
<td>SSCP</td>
<td>weight</td>
<td>13034.5</td>
<td>817919.60</td>
<td>1398238.75</td>
<td>217717.45</td>
<td></td>
</tr>
<tr>
<td>9 m</td>
<td>SSCP</td>
<td>age</td>
<td>2072.1</td>
<td>129432.57</td>
<td>217717.45</td>
<td>34515.95</td>
<td></td>
</tr>
<tr>
<td>10 m</td>
<td>N</td>
<td></td>
<td>126.0</td>
<td>126.00</td>
<td>126.00</td>
<td>126.00</td>
<td></td>
</tr>
</tbody>
</table>

The OUTEST= data set is displayed in Output 102.3.6; again, the BY groups are separated. The _MODEL_ column contains the labels for models from the MODEL statements. If no labels are specified, the defaults MODEL1 and MODEL2 would appear as values for _MODEL_. Note that _TYPE_='PARMS' for all observations, indicating that all observations contain parameter estimates. The _DEPVAR_ column displays the dependent variable, and the _RMSE_ column gives the root mean square error for the associated model. The Intercept column gives the estimate for the intercept for the associated model, and variables with the same name as variables in the original data set (height, age) give parameter estimates for those variables. The dependent variable, weight, is shown with a value of –1. The _IN_ column contains the number of regressors in the model not including the intercept; _P_ contains the number of parameters in the model; _EDF_ contains the error degrees of freedom; and _RSQ_ contains the R square statistic. Finally, note that the _IN_, _P_, _EDF_, and _RSQ_ columns appear in the OUTEST= data set since the RSQUARE option is specified in the PROC REG statement.
Output 102.3.6 OUTEST Data Set

<table>
<thead>
<tr>
<th>Obs</th>
<th>sex</th>
<th><em>MODEL</em></th>
<th><em>TYPE</em></th>
<th><em>DEPVAR</em></th>
<th><em>RMSE</em></th>
<th>Intercept</th>
<th>height</th>
<th>weight</th>
<th>age</th>
<th><em>IN</em></th>
<th><em>P</em></th>
<th><em>EDF</em></th>
<th><em>RSQ</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>f</td>
<td>eq1</td>
<td>PARMS</td>
<td>weight</td>
<td>12.3461</td>
<td>-153.129</td>
<td>4.16361</td>
<td>1</td>
<td>2</td>
<td>109</td>
<td>0.56416</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>eq2</td>
<td>PARMS</td>
<td>weight</td>
<td>12.0527</td>
<td>-150.597</td>
<td>3.60378</td>
<td>-1</td>
<td>1.90703</td>
<td>2</td>
<td>3</td>
<td>108</td>
<td>0.58845</td>
</tr>
<tr>
<td>3</td>
<td>m</td>
<td>eq1</td>
<td>PARMS</td>
<td>weight</td>
<td>12.2850</td>
<td>-125.698</td>
<td>3.68977</td>
<td>-1</td>
<td>1</td>
<td>124</td>
<td>0.62451</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>m</td>
<td>eq2</td>
<td>PARMS</td>
<td>weight</td>
<td>11.7098</td>
<td>-113.713</td>
<td>2.68075</td>
<td>-1</td>
<td>3.08167</td>
<td>2</td>
<td>3</td>
<td>123</td>
<td>0.66161</td>
</tr>
</tbody>
</table>

Example 102.4: Regression with Quantitative and Qualitative Variables

At times it is desirable to have independent variables in the model that are qualitative rather than quantitative. This is easily handled in a regression framework. Regression uses qualitative variables to distinguish between populations. There are two main advantages of fitting both populations in one model. You gain the ability to test for different slopes or intercepts in the populations, and more degrees of freedom are available for the analysis.

Regression with qualitative variables is different from analysis of variance and analysis of covariance. Analysis of variance uses qualitative independent variables only. Analysis of covariance uses quantitative variables in addition to the qualitative variables in order to account for correlation in the data and reduce MSE; however, the quantitative variables are not of primary interest and merely improve the precision of the analysis.

Consider the case where $Y_i$ is the dependent variable, $X_{1i}$ is a quantitative variable, $X_{2i}$ is a qualitative variable taking on values 0 or 1, and $X_{1i}X_{2i}$ is the interaction. The variable $X_{2i}$ is called a dummy, binary, or indicator variable. With values 0 or 1, it distinguishes between two populations. The model is of the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}X_{2i} + \epsilon_i$$

for the observations $i = 1, 2, \ldots, n$. The parameters to be estimated are $\beta_0, \beta_1, \beta_2$, and $\beta_3$. The number of dummy variables used is one less than the number of qualitative levels. This yields a nonsingular $X'X$ matrix. See Chapter 10 of Neter, Wasserman, and Kutner (1990) for more details.

An example from Neter, Wasserman, and Kutner (1990) follows. An economist is investigating the relationship between the size of an insurance firm and the speed at which it implements new insurance innovations. He believes that the type of firm might affect this relationship and suspects that there might be some interaction between the size and type of firm. The dummy variable in the model enables the two firms to have different intercepts. The interaction term enables the firms to have different slopes as well.

In this study, $Y_i$ is the number of months from the time the first firm implemented the innovation to the time it was implemented by the $i$th firm. The variable $X_{1i}$ is the size of the firm, measured in total assets of the firm. The variable $X_{2i}$ denotes the firm type; it is 0 if the firm is a mutual fund company and 1 if the firm is a stock company. The dummy variable enables each firm type to have a different intercept and slope.

The previous model can be broken down into a model for each firm type by plugging in the values for $X_{2i}$. If $X_{2i} = 0$, the model is

$$Y_i = \beta_0 + \beta_1 X_{1i} + \epsilon_i$$
This is the model for a mutual company. If $X_2 = 1$, the model for a stock firm is

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1 + \epsilon_i$$

This model has intercept $\beta_0 + \beta_2$ and slope $\beta_1 + \beta_3$.

The data follow. Note that the interaction term is created in the DATA step since polynomial effects such as size\(\times\)type are not allowed in the MODEL statement in the REG procedure.

```plaintext
title 'Regression with Quantitative and Qualitative Variables';
data insurance;
    input time size type @@;
    sizetype=size*type;
datalines;
17 151 0 26 92 0 21 175 0 30 31 0 22 104 0 0 277 0 12 210 0
19 120 0 4 290 0 16 238 0 28 164 1 15 272 1 11 295 1 38 68 1
31 85 1 21 224 1 20 166 1 13 305 1 30 124 1 14 246 1
;
```

The following statements begin the analysis and produce the ANOVA table in Output 102.4.1:

```plaintext
proc reg data=insurance;
    model time = size type sizetype;
run;
```

### Output 102.4.1 ANOVA Table and Parameter Estimates

**Regression with Quantitative and Qualitative Variables**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3</td>
<td>1504.41904</td>
<td>501.47301</td>
<td>45.49</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>16</td>
<td>176.38096</td>
<td>11.02381</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>1680.80000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Root MSE</th>
<th>R-Square</th>
<th>0.8951</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dependent Mean</td>
<td>19.40000</td>
<td>Adj R-Sq</td>
</tr>
<tr>
<td>Coeff Var</td>
<td>17.11450</td>
<td></td>
</tr>
</tbody>
</table>

| Variable   | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|------------|----|--------------------|----------------|---------|-------|---|
| Intercept  | 1  | 33.83837           | 2.44065        | 13.86   | <.0001|
| size       | 1  | -0.10153           | 0.01305        | -7.78   | <.0001|
| type       | 1  | 8.13125            | 3.65405        | 2.23    | 0.0408|
| sizetype   | 1  | -0.00041714       | 0.01833        | -0.02   | 0.9821|
```

---

The overall $F$ statistic is significant ($F = 45.490, p < 0.0001$). The interaction term is not significant ($t = -0.023, p = 0.9821$). Hence, this term should be removed and the model refitted, as shown in the following statements:

```plaintext
delete sizetype;
print;
run;
```

The `DELETE` statement removes the interaction term (`sizetype`) from the model. The new ANOVA and parameter estimates tables are shown in Output 102.4.2.

**Output 102.4.2** ANOVA Table and Parameter Estimates

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>2</td>
<td>1504.41333</td>
<td>752.20667</td>
<td>72.50</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>17</td>
<td>176.38667</td>
<td>10.37569</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>19</td>
<td>1680.80000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Root MSE: 3.22113
- R-Square: 0.8951
- Dependent Mean: 19.40000
- Adj R-Sq: 0.8827
- Coeff Var: 16.60377

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>size</td>
</tr>
<tr>
<td>type</td>
</tr>
</tbody>
</table>

The overall $F$ statistic is still significant ($F = 72.497, p < 0.0001$). The intercept and the coefficients associated with `size` and `type` are significantly different from zero ($t = 18.675, p < 0.0001; t = -11.443, p < 0.0001; t = 5.521, p < 0.0001$ respectively). Notice that the $R$ square did not change with the omission of the interaction term.

The fitted model is

\[
time = 33.87 - 0.102 \times \text{size} + 8.055 \times \text{type}
\]

The fitted model for a mutual fund company ($X_{2i} = 0$) is

\[
time = 33.87 - 0.102 \times \text{size}
\]

and the fitted model for a stock company ($X_{2i} = 1$) is

\[
time = (33.87 + 8.055) - 0.102 \times \text{size}
\]
So the two models have different intercepts but the same slope.

The following statements first use an OUTPUT statement to save the residuals and predicted values from the new model in the OUT= data set. Next PROC SGPLOT is used to produce Output 102.4.3, which plots residuals versus predicted values. The firm type is used as the plot symbol; this can be useful in determining if the firm types have different residual patterns.

```plaintext
output out=out r=r p=p;
run;

proc sgplot data=out;
  scatter x=p y=r / markerchar=type group=type;
run;
```

**Output 102.4.3** Plot of Residual vs. Predicted Values

The residuals show no major trend. Neither firm type by itself shows a trend either. This indicates that the model is satisfactory.

The following statements produce the plot of the predicted values versus size that appears in Output 102.4.4, where the firm type is again used as the plotting symbol:
proc sgplot data=out;
    scatter x=size y=p / markerchar=type group=type;
run;

**Output 102.4.4** Plot of Predicted vs. Size

The different intercepts are very evident in this plot.
**Example 102.5: Ridge Regression for Acetylene Data**

This example uses the acetylene data in Marquardt and Snee (1975) to illustrate the RIDGEPLOT and OUTVIF options. Here are the data:

```plaintext
data acetyl;
  input x1-x4 @@;
  x1x2 = x1 * x2;
  x1x1 = x1 * x1;
  label x1 = 'reactor temperature(celsius)'
     x2 = 'h2 to n-heptone ratio'
     x3 = 'contact time/sec'
     x4 = 'conversion percentage'
     x1x2 = 'temperature-ratio interaction'
     x1x1 = 'squared temperature';
datalines;
  1300 7.5 .012 49 1300 9 .012 50.2 1300 11 .0115 50.5
  1300 13.5 .013 48.5 1300 17 .0135 47.5 1300 23 .012 44.5
  1200 5.3 .04 28 1200 7.5 .038 31.5 1200 11 .032 34.5
  1200 13.5 .026 35 1200 17 .034 38 1200 23 .041 38.5
  1100 5.3 .084 15 1100 7.5 .098 17 1100 11 .092 20.5
  1100 17 .086 29.5;
ods graphics on;
proc reg data=acetyl outvif
   outest=b ridge=0 to 0.02 by .002;
   model x4=x1 x2 x3 x1x2 x1x1;
run;
proc print data=b;
run;
```

When ODS Graphics is enabled and you request ridge regression by using the RIDGE= option in the PROC REG statement, PROC REG produces a panel showing variance inflation factors (VIF) in the upper plot in the panel and ridge traces in the lower plot. This panel is shown in Output 102.5.1.
The OUTVIF option outputs the variance inflation factors to the OUTEST= data set that is shown in Output 102.5.2.
### Output 102.5.2 OUTEST Data Set Showing VIF Values

<table>
<thead>
<tr>
<th>Obs</th>
<th><em>MODEL</em></th>
<th><em>TYPE</em></th>
<th><em>DEPVAR</em></th>
<th><em>RIDGE</em></th>
<th><em>PCOMIT</em></th>
<th><em>RMSE</em></th>
<th>Intercept</th>
<th>x1</th>
<th>x2</th>
<th>x3</th>
<th>x1x2</th>
<th>x1x1</th>
<th>x4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>x4</td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.15596</td>
<td>390.538</td>
<td>-0.78</td>
<td>10.174</td>
<td>-121.626</td>
<td>-0.008</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.000</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>7682.37</td>
<td>320.022</td>
<td>53.525</td>
<td>344.545</td>
<td>6643.32</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.000</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1.15596</td>
<td>390.538</td>
<td>-0.78</td>
<td>10.174</td>
<td>-121.626</td>
<td>-0.008</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.002</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>11.18</td>
<td>58.731</td>
<td>10.744</td>
<td>63.208</td>
<td>11.22</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.002</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.69721</td>
<td>-103.388</td>
<td>0.05</td>
<td>4.404</td>
<td>-9.065</td>
<td>-0.003</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.004</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>4.36</td>
<td>23.939</td>
<td>9.996</td>
<td>25.744</td>
<td>5.15</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.004</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.22340</td>
<td>-93.797</td>
<td>0.06</td>
<td>2.839</td>
<td>-21.338</td>
<td>-0.002</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.006</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>2.93</td>
<td>13.011</td>
<td>9.383</td>
<td>13.976</td>
<td>3.81</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.006</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.47752</td>
<td>-87.687</td>
<td>0.06</td>
<td>2.110</td>
<td>-28.447</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.008</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>2.36</td>
<td>8.224</td>
<td>8.838</td>
<td>8.821</td>
<td>3.23</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.008</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.62677</td>
<td>-83.593</td>
<td>0.06</td>
<td>1.689</td>
<td>-33.377</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.010</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>2.04</td>
<td>5.709</td>
<td>8.343</td>
<td>6.112</td>
<td>2.89</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.010</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.72505</td>
<td>-80.603</td>
<td>0.06</td>
<td>1.414</td>
<td>-37.177</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.012</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>1.84</td>
<td>4.226</td>
<td>7.891</td>
<td>4.514</td>
<td>2.65</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.012</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.79477</td>
<td>-78.276</td>
<td>0.06</td>
<td>1.221</td>
<td>-40.297</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.014</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>1.69</td>
<td>3.279</td>
<td>7.476</td>
<td>3.493</td>
<td>2.46</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.014</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.84693</td>
<td>-76.381</td>
<td>0.06</td>
<td>1.078</td>
<td>-42.965</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.016</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>1.57</td>
<td>2.637</td>
<td>7.094</td>
<td>2.801</td>
<td>2.31</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.016</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.88750</td>
<td>-74.785</td>
<td>0.06</td>
<td>0.968</td>
<td>-45.309</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.018</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>1.47</td>
<td>2.182</td>
<td>6.741</td>
<td>2.310</td>
<td>2.18</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.018</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.92004</td>
<td>-73.407</td>
<td>0.06</td>
<td>0.880</td>
<td>-47.407</td>
<td>-0.000</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.020</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.</td>
<td>.</td>
<td>1.39</td>
<td>1.847</td>
<td>6.415</td>
<td>1.949</td>
<td>2.06</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.020</td>
<td></td>
<td></td>
<td>.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.94679</td>
<td>-72.193</td>
<td>0.06</td>
<td>0.809</td>
<td>-49.310</td>
<td>-0.000</td>
<td>0.00</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
If you want to obtain separate plots containing the ridge traces and VIF traces, you can specify the UNPACK suboption in the PLOTS=RIDGE option. You can also request that one or both of the VIF axis and ridge parameter axis be displayed on a logarithmic scale. You can see in Output 102.5.1 that the VIF traces for several of the parameters are nearly indistinguishable when displayed on a linear scale. The following code illustrates how you obtain separate VIF and ridge traces with the VIF values displayed on a logarithmic scale. Note that you can obtain plots of VIF values even though you do not specify the OUTVIF option in the PROC REG statement.

```
proc reg data=acetyl plots(only)=ridge(unpack VIFaxis=log)
   outest=b ridge=0 to 0.02 by .002;
   model x4=x1 x2 x3 x1x2 x1x1;
   run;
ods graphics off;
```

The requested plots are shown in Output 102.5.3 and Output 102.5.4.

Output 102.5.3  VIF Traces
Example 102.6: Chemical Reaction Response

This example shows how you can use lack-of-fit tests with the REG procedure. See the section “Testing for Lack of Fit” on page 8531 for details about lack-of-fit tests.

In a study of the percentage of raw material that responds in a reaction, researchers identified the following five factors:

- the feed rate of the chemicals (FeedRate), ranging from 10 to 15 liters per minute
- the percentage of the catalyst (Catalyst), ranging from 1% to 2%
- the agitation rate of the reactor (AgitRate), ranging from 100 to 120 revolutions per minute
- the temperature (Temperature), ranging from 140 to 180 degrees Celsius
- the concentration (Concentration), ranging from 3% to 6%
The following data set contains the results of an experiment designed to estimate main effects for all factors:

```plaintext
data reaction;
  input FeedRate Catalyst AgitRate Temperature Concentration ReactionPercentage;
datalines;
  10.0 1.0 100 140 6.0 37.5
  10.0 1.0 120 180 3.0 28.5
  10.0 2.0 100 180 3.0 40.4
  10.0 2.0 120 140 6.0 48.2
  15.0 1.0 100 180 6.0 50.7
  15.0 1.0 120 140 3.0 28.9
  15.0 2.0 100 140 3.0 43.5
  15.0 2.0 120 180 6.0 64.5
  12.5 1.5 110 160 4.5 39.0
  12.5 1.5 110 160 4.5 40.3
  12.5 1.5 110 160 4.5 38.7
  12.5 1.5 110 160 4.5 39.7
;```

The first eight runs of this experiment enable orthogonal estimation of the main effects for all factors. The last four comprise four replicates of the centerpoint.

The following statements fit a linear model. Because this experiment includes replications, you can test for lack of fit by using the LACKFIT option in the MODEL statement.

```plaintext
proc reg data=reaction;
  model ReactionPercentage=FeedRate Catalyst AgitRate Temperature Concentration / lackfit;
run;
```

Output 102.6.1 shows that the lack of fit for the linear model is significant, indicating that a more complex model is required. Models that include interactions should be investigated. In this case, this will require additional experimentation to obtain appropriate data for estimating the effects.

**Output 102.6.1** Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
<th>F Value</th>
<th>Pr &gt; F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>5</td>
<td>990.27000</td>
<td>198.05400</td>
<td>33.29</td>
<td>0.0003</td>
</tr>
<tr>
<td>Error</td>
<td>6</td>
<td>35.69917</td>
<td>5.94986</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lack of Fit</td>
<td>3</td>
<td>34.15167</td>
<td>11.38389</td>
<td>22.07</td>
<td>0.0151</td>
</tr>
<tr>
<td>Pure Error</td>
<td>3</td>
<td>1.54750</td>
<td>0.51583</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>11</td>
<td>1025.96917</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Output 102.6.1** Analysis of Variance

- **Root MSE**: 2.43923
- **R-Square**: 0.9652
- **Dependent Mean**: 41.65833
- **Adj R-Sq**: 0.9362
- **Coeff Var**: 5.85533
References


Subject Index

adjusted R² selection (REG), 8503
alpha level
   REG procedure, 8451
annotate
   global data set (REG), 8540
   local data set (REG), 8549
autocorrelation
   REG procedure, 8535
backward elimination
   REG procedure, 8431, 8502
collinearity
   REG procedure, 8513
correlation
   matrix (REG), 8451
covariance matrix
   REG procedure, 8451
COVRATIO statistic, 8517
crossproducts matrix
   REG procedure, 8537
delete variables (REG), 8465
deleting observations
   REG procedure, 8524
DFBETAS statistic (REG), 8518
DFFITS statistic
   REG procedure, 8517
diagnostic statistics
   REG procedure, 8514, 8515
fit diagnostics
   examples (REG), 8564
forward selection
   REG procedure, 8431, 8501
graphics
   keywords (REG), 8546
   options (REG), 8547
   traditional plots (REG), 8545
hat matrix, 8517
heteroscedasticity
   testing (REG), 8530
hypothesis tests
   multivariate (REG), 8532
   REG procedure, 8477, 8487
incomplete principal components
   REG procedure, 8452, 8474
   influence diagnostics
      examples (REG), 8564
   influence statistics
      REG procedure, 8516
   IPC analysis
      REG procedure, 8452, 8474, 8536
lack of fit
   examples (REG), 8607
lack-of-fit
testing (REG), 8531
line printer plots
   REG procedure, 8553
LMSELECT procedure
   ODS Graphics, 8452
Mallows’ Cₚ selection
   REG procedure, 8503
marginal model plots
   REG procedure, 8577
model
   fit summary (REG), 8514
model building
   examples (REG), 8564
model selection
   examples (REG), 8578
   REG procedure, 8431, 8501, 8503, 8504
multicollinearity
   REG procedure, 8513
multivariate tests
   REG procedure, 8532
non-full-rank models
   REG procedure, 8511
ODS graph names
   REG procedure, 8562
ODS GRAPHICS
   examples (REG), 8564
ODS Graphics
   LMSELECT procedure, 8452
P-P plots
   REG procedure, 8537
parameter estimates
   example (REG), 8505
   REG procedure, 8539
partial regression leverage plots
REG procedure, 8522
plots
  keywords (REG), 8546
  line printer (REG), 8553
  options (REG), 8547, 8549
  traditional (REG), 8545
polynomial regression
  REG procedure, 8436
predicted values
  REG procedure, 8504, 8507
prediction
  example (REG), 8578
Q-Q plots
  REG procedure, 8537
qualitative variables
  REG procedure, 8598
R$^2$ improvement
  REG procedure, 8502, 8503
R$^2$ selection
  REG procedure, 8503
refitting models
  REG procedure, 8526
REG procedure
  adding variables, 8464
  adjusted R$^2$ selection, 8503
  alpha level, 8451
  annotations, 8540, 8549
  ANOVA table, 8538
  autocorrelation, 8535
  backward elimination, 8431, 8502
  collinearity, 8513
  computational methods, 8537
  correlation matrix, 8451
  covariance matrix, 8451
  crossproducts matrix, 8537
  delete variables, 8465
  deleting observations, 8524
  diagnostic statistics, 8514, 8515
  dictionary of options, 8549
  fit diagnostics, 8564
  forward selection, 8431, 8501
  graphics keywords and options, 8546, 8547
  graphics plots, traditional, 8545
  heteroscedasticity, testing, 8530
  hypothesis tests, 8477, 8487
  incomplete principal components, 8452, 8474
  influence diagnostics, 8564
  influence statistics, 8516
  input data sets, 8489
  interactive analysis, 8447, 8497
  introductory example, 8432
  IPC analysis, 8452, 8474, 8536
lack of fit, 8607
lack-of-fit, testing, 8531
line printer plots, 8553
Mallows’ C$^p$ selection, 8503
marginal model plots, 8577
missing values, 8488
model building, 8564
model fit summary statistics, 8514
model selection, 8431, 8501, 8503, 8504, 8578
multicollinearity, 8513
multivariate tests, 8532
new regressors, 8489
non-full-rank models, 8511
ODS graph names, 8562
ODS GRAPHICS, 8564
ODS table names, 8555
output data sets, 8492, 8497
P-P plots, 8537
parameter estimates, 8505, 8539
partial regression leverage plots, 8522
plot keywords and options, 8546, 8547, 8549
plots, traditional, 8545
polynomial regression, 8436
predicted values, 8504, 8507, 8578
Q-Q plots, 8537
regression
  analysis (REG), 8430
residuals
  REG procedure, 8507
restoring weights
  REG procedure, 8527
reweighting observations
  REG procedure, 8524
ridge regression, 8463, 8474, 8536, 8552, 8603
singularities, 8537
stepwise selection, 8431, 8502
summary statistics, 8514
sweep algorithm, 8537
time series data, 8535
variance inflation factors (VIF), 8452
studentized residual, 8517
summary statistics
  REG procedure, 8514
sweep algorithm
  REG procedure, 8537
time series data
  REG procedure, 8535
variance inflation factors (VIF)
  REG procedure, 8452
VIF, see variance inflation factors (VIF)
## Syntax Index

<table>
<thead>
<tr>
<th>Option</th>
<th>Statement/Procedure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOV option</td>
<td>MODEL statement (REG)</td>
<td>8469</td>
</tr>
<tr>
<td>ACOVMETHOD= option</td>
<td>MODEL statement (REG)</td>
<td>8469</td>
</tr>
<tr>
<td>ADD statement</td>
<td>REG procedure</td>
<td>8464</td>
</tr>
<tr>
<td>ADJRSQ option</td>
<td>MODEL statement (REG)</td>
<td>8469</td>
</tr>
<tr>
<td>AIC option</td>
<td>MODEL statement (REG)</td>
<td>8469</td>
</tr>
<tr>
<td>ALL option</td>
<td>MODEL statement (REG)</td>
<td>8469</td>
</tr>
<tr>
<td>ALLOBS option</td>
<td>PAINT statement (REG)</td>
<td>8542</td>
</tr>
<tr>
<td>ANOVA option</td>
<td>PRINT statement (REG)</td>
<td>8481</td>
</tr>
<tr>
<td>B option</td>
<td>MODEL statement (REG)</td>
<td>8469</td>
</tr>
<tr>
<td>BEST= option</td>
<td>MODEL statement (REG)</td>
<td>8469</td>
</tr>
<tr>
<td>BIC option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>BY statement</td>
<td>REG procedure</td>
<td>8464</td>
</tr>
<tr>
<td>CANPRINT option</td>
<td>MTEST statement (REG)</td>
<td>8478</td>
</tr>
<tr>
<td>CAXIS= option</td>
<td>PLOT statement (REG)</td>
<td>8549</td>
</tr>
<tr>
<td>CFRA ME= option</td>
<td>PLOT statement (REG)</td>
<td>8549</td>
</tr>
<tr>
<td>CHOCKING= option</td>
<td>PLOT statement (REG)</td>
<td>8549</td>
</tr>
<tr>
<td>CHREF= option</td>
<td>PLOT statement (REG)</td>
<td>8550</td>
</tr>
<tr>
<td>CLB option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>CLEAR option</td>
<td>PLOT statement (REG)</td>
<td>8554</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Option</th>
<th>Statement/Procedure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLI option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>CLINE= option</td>
<td>PLOT statement (REG)</td>
<td>8550</td>
</tr>
<tr>
<td>CLM option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>CMALLOWS= option</td>
<td>PLOT statement (REG)</td>
<td>8550</td>
</tr>
<tr>
<td>CODE statement</td>
<td>REG procedure</td>
<td>8464</td>
</tr>
<tr>
<td>COLLECT option</td>
<td>PLOT statement (REG)</td>
<td>8554</td>
</tr>
<tr>
<td>COLLIN option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>COLLINOINT option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>CONF option</td>
<td>PLOT statement (REG)</td>
<td>8550</td>
</tr>
<tr>
<td>CORR option</td>
<td>PROC REG statement</td>
<td>8451</td>
</tr>
<tr>
<td>CORRB option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>COVB option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>COVOUT option</td>
<td>PROC REG statement</td>
<td>8451</td>
</tr>
<tr>
<td>CP option</td>
<td>MODEL statement (REG)</td>
<td>8470</td>
</tr>
<tr>
<td>CTEXT= option</td>
<td>PLOT statement (REG)</td>
<td>8550</td>
</tr>
<tr>
<td>CVREF= option</td>
<td>PLOT statement (REG)</td>
<td>8550</td>
</tr>
<tr>
<td>DATA= option</td>
<td>PROC REG statement</td>
<td>8451</td>
</tr>
<tr>
<td>DELETE statement</td>
<td>REG procedure</td>
<td>8465</td>
</tr>
<tr>
<td>DESCRIPTION= option</td>
<td>PLOT statement (REG)</td>
<td>8550</td>
</tr>
<tr>
<td>DETAILS option</td>
<td>MODEL statement (REG)</td>
<td>8471</td>
</tr>
<tr>
<td>DW option</td>
<td>MTEST statement (REG)</td>
<td>8478</td>
</tr>
<tr>
<td>DWPROB option</td>
<td>MODEL statement (REG)</td>
<td>8471</td>
</tr>
<tr>
<td>EDF option</td>
<td>MODEL statement (REG)</td>
<td>8471</td>
</tr>
</tbody>
</table>
PLOT statement (REG), 8550
PROC REG statement, 8451

FREQ statement
REG procedure, 8465

GMSEP option
MODEL statement (REG), 8471
PLOT statement (REG), 8551
GOUT= option
PROC REG statement, 8541
GROUPNAMES= option
MODEL statement (REG), 8471

HAXIS= option
PLOT statement (REG), 8551
HCC option
MODEL statement (REG), 8472
HCCMETHOD= option
MODEL statement (REG), 8472
HPLOTS= option
PLOT statement (REG), 8554
HREF= option
PLOT statement (REG), 8551

I option
MODEL statement (REG), 8472
ID statement
REG procedure, 8466
IN option
PLOT statement (REG), 8551
INCLUDE= option
MODEL statement (REG), 8472
INFLUENCE option
MODEL statement (REG), 8472

JP option
MODEL statement (REG), 8472
PLOT statement (REG), 8551

keyword= option
OUTPUT statement (REG), 8479

LACKFIT option
MODEL statement (REG), 8472
LEGEND= option
PLOT statement (REG), 8551
LHREF= option
PLOT statement (REG), 8551
LINEPRINTER option
PROC REG statement, 8541
LLINE= option
PLOT statement (REG), 8551
LVREF= option
PLOT statement (REG), 8551

MAXSTEP option
MODEL statement (REG), 8473
MODEL statement
REG procedure, 8466
MODELDATA option
PRINT statement (REG), 8481
MODELFONT option
PLOT statement (REG), 8551
MODELHT option
PLOT statement (REG), 8551
MODELLAB option
PLOT statement (REG), 8551
MSE option
MODEL statement (REG), 8473
PLOT statement (REG), 8552
MSTAT= option
MTEST statement (REG), 8478
MTEST statement
REG procedure, 8477

NAME= option
PLOT statement (REG), 8552
NOCOLLECT option
PLOT statement (REG), 8554
NOINT option
MODEL statement (REG), 8473
NOLEGEND option
PLOT statement (REG), 8552
NOLINE option
PLOT statement (REG), 8552
NOLIST option
PAINT statement (REG), 8543
REWEIGHT statement (REG), 8485
NOMODEL option
PLOT statement (REG), 8552
NOPRINT option
MODEL statement (REG), 8473
PROC REG statement, 8451
NOSTAT option
PLOT statement (REG), 8552
NP option
PLOT statement (REG), 8552

OUT= option
OUTPUT statement (REG), 8479
OUTEST= option
PROC REG statement, 8452
OUTPUT statement
REG procedure, 8479
OUTSEB option
MODEL statement (REG), 8473
PROC REG statement, 8452
OUTSSCP= option
PROC REG statement, 8452
OUTSTB option
MODEL statement (REG), 8473
PROC REG statement, 8452
OUTVIF option
MODEL statement (REG), 8473
PROC REG statement, 8452
OVERLAY option
PLOT statement (REG), 8552, 8554
P option
MODEL statement (REG), 8473
PAINT statement
REG procedure, 8541
PARTIAL option
MODEL statement (REG), 8473
PARTIALDATA option
MODEL statement (REG), 8474
PARTIALR2 option
MODEL statement (REG), 8474
PC option
MODEL statement (REG), 8474
PLOT statement (REG), 8552
PCOMIT= option
MODEL statement (REG), 8474
PROC REG statement, 8452
PCORR1 option
MODEL statement (REG), 8474
PCORR2 option
MODEL statement (REG), 8474
PLOT option
PROC REG statement, 8452
PLOT statement
REG procedure, 8544
PLOTS option
PROC REG statement, 8452
PRED option
PLOT statement (REG), 8552
PRESS option
MODEL statement (REG), 8474
PROC REG statement, 8463
PRINT option
MTEST statement (REG), 8479
TEST statement (REG), 8488
PRINT statement, REG procedure, 8480
PROC REG statement, see REG procedure
R option
MODEL statement (REG), 8474
REFIT statement, REG procedure, 8481
REG procedure
syntax, 8448
REG procedure, ADD statement, 8464
REG procedure, BY statement, 8464
REG procedure, CODE statement, 8464
REG procedure, DELETE statement, 8465
REG procedure, FREQ statement, 8465
REG procedure, ID statement, 8466
REG procedure, MODEL statement, 8466
ACOV option, 8469
ACOVMETHOD= option, 8469
ADJRSQ option, 8469
AIC option, 8469
ALL option, 8469
ALPHA= option, 8469
B option, 8469
BEST= option, 8469
BIC option, 8470
CLB option, 8470
CLI option, 8470
CLM option, 8470
COLLIN option, 8470
COLLINOINT option, 8470
CORRB option, 8470
COVB option, 8470
CP option, 8470
DETAILS option, 8471
DW option, 8471
DWPROB option, 8471
EDF option, 8471
GMSEP option, 8471
GROUPNAMES= option, 8471
HCC option, 8472
HCCMETHOD= option, 8472
I option, 8472
INCLUDE= option, 8472
INFLUENCE option, 8472
JP option, 8472
LACKFIT option, 8472
MAXSTEP option, 8473
MSE option, 8473
NOINT option, 8473
NPRINT option, 8473
OUTSEB option, 8473
OUTSTB option, 8473
OUTVIF option, 8473
P option, 8473
PARTIAL option, 8473
PARTIALDATA option, 8474
PARTIALR2 option, 8474
PC option, 8474
PCOMIT= option, 8474
PCORR1 option, 8474
PCORR2 option, 8474
PRESS option, 8474
R option, 8474
RIDGE= option, 8474
RMSE option, 8475
RSQUARE option, 8475
SBC option, 8475
SCORR1 option, 8475
SCORR2 option, 8475
SELECTION= option, 8431, 8475
SEQB option, 8475
SIGMA= option, 8475
SINGULAR= option, 8476
SLENGTH= option, 8476
SP option, 8476
SPEC option, 8476
SRT option, 8476
SS1 option, 8476
SS2 option, 8476
SSE option, 8476
START= option, 8476
STB option, 8477
STOP= option, 8477
TOL option, 8477
VIF option, 8477
WHITE option, 8477
XPX option, 8477
REG procedure, MTEST statement, 8477
CANPRINT option, 8478
DETAILS option, 8478
MSTAT= option, 8478
PRINT option, 8479
REG procedure, OUTPUT statement, 8479
keyword= option, 8479
OUT= option, 8479
REG procedure, PAINT statement, 8541
ALLOBS option, 8542
NOLIST option, 8543
RESET option, 8543
STATUS option, 8544
SYMBOL= option, 8543
UNDO option, 8544
REG procedure, PLOT statement, 8544
AIC option, 8549
ANNOTATE= option, 8549
BIC option, 8549
CAXIS= option, 8549
CFRAME= option, 8549
CHOCKING= option, 8549
CHREF= option, 8550
CLEAR option, 8550
CLINE= option, 8550
CMALLOWS= option, 8550
COLLECT option, 8554
CONF option, 8550
CP option, 8550
CTEXT= option, 8550
CVREF= option, 8550
DESCRIPTION= option, 8550
EDF option, 8550
GMSEP option, 8551
HAXIS= option, 8551
HPOLTS= option, 8554
HREF= option, 8551
IN option, 8551
JP option, 8551
LEGEND= option, 8551
LHREF= option, 8551
LLINE= option, 8551
LVREF= option, 8551
MODELFONT option, 8551
MODELHT option, 8551
MODELLAB option, 8551
MSE option, 8552
NAME= option, 8552
NOCOLLECT option, 8554
NOLENGTHEN option, 8552
NOLINE option, 8552
NOMODEL option, 8552
NOSTAT option, 8552
NP option, 8552
OVERLAY option, 8552, 8554
PC option, 8552
PRED option, 8552
RIDGEPLOT option, 8552
SBC option, 8553
SP option, 8553
SSE option, 8553
STATFONT option, 8553
STATHT option, 8553
summary of options, 8546, 8547
SYMBOL= option, 8555
USEALL option, 8555
VAXIS= option, 8553
VPLTS= option, 8555
VREF= option, 8553
REG procedure, PRINT statement, 8480
ANOVA option, 8481
MODELDATA option, 8481
REG procedure, PROC REG statement, 8450
ALL option, 8451
ALPHA= option, 8451
ANNOTATE= option, 8540
CORR option, 8451
COVOUT option, 8451
DATA= option, 8451
EDF option, 8451
GOUT= option, 8451
LINEPRINTER option, 8451
NOPRINT option, 8451
OUTEST= option, 8452
OUTSEB option, 8452
OUTSSCP= option, 8452
OUTSTB option, 8452
OUTVIF option, 8452
PCOMIT= option, 8452
PLOT option, 8452
PLOTS option, 8452
PRESS option, 8463
RIDGE= option, 8463
RSQUARE option, 8463
SIMPLE option, 8463
SINGULAR= option, 8463
TABLEOUT option, 8463
USSCP option, 8463
REG procedure, REFIT statement, 8481
REG procedure, RESTRICT statement, 8482
REG procedure, REWEIGHT statement, 8483
ALLOBS option, 8485
NOLIST option, 8485
RESET option, 8485
STATUS option, 8486
UNDO option, 8486
WEIGHT= option, 8486
REG procedure, STORE statement, 8486
REG procedure, TEST statement, 8487
PRINT option, 8488
REG procedure, VAR statement, 8488
REG procedure, WEIGHT statement, 8488
RESET option
PAINT statement (REG), 8543
REWEIGHT statement (REG), 8485
RESTRICT statement
REG procedure, 8482
REWEIGHT statement, REG procedure, 8483
RIDGE= option
MODEL statement (REG), 8474
PROC REG statement, 8463
RIDGEPLOT option
PLOT statement (REG), 8552
RMSE option
MODEL statement (REG), 8475
RSQUARE option
MODEL statement (REG), 8475
PROC REG statement, 8463
SBC option
MODEL statement (REG), 8475
PLOT statement (REG), 8553
SCORR1 option
MODEL statement (REG), 8475
SCORR2 option
MODEL statement (REG), 8475
SELECTION= option
MODEL statement (REG), 8475
REG procedure, MODEL statement, 8431
SEQB option
MODEL statement (REG), 8475
SIGMA= option
MODEL statement (REG), 8475
SIMPLE option
PROC REG statement, 8463
SINGULAR= option
MODEL statement (REG), 8476
PROC REG statement, 8463
SLENTRY= option
MODEL statement (REG), 8476
SLSTAY= option
MODEL statement (REG), 8476
SP option
MODEL statement (REG), 8476
PLOT statement (REG), 8553
SPEC option
MODEL statement (REG), 8476
SRT option
MODEL statement (REG), 8476
SS1 option
MODEL statement (REG), 8476
SS2 option
MODEL statement (REG), 8476
SSE option
MODEL statement (REG), 8476
PLOT statement (REG), 8553
START= option
MODEL statement (REG), 8476
STATFONT option
PLOT statement (REG), 8553
STATHT option
PLOT statement (REG), 8553
STATUS option
PAINT statement (REG), 8544
REWEIGHT statement (REG), 8486
STB option
MODEL statement (REG), 8477
STOP= option
MODEL statement (REG), 8477
STORE statement
REG procedure, 8477
SYMBOL= option
PAINT statement (REG), 8543
PLOT statement (REG), 8555
TABLEOUT option
PROC REG statement, 8463
TEST statement
REG procedure, 8487
TOL option
MODEL statement (REG), 8477
UNDO option
PAINT statement (REG), 8544
REWEIGHT statement (REG), 8486
USEALL option
   PLOT statement (REG), 8553
USSCP option
   PROC REG statement, 8463

VAR statement
   REG procedure, 8488
VAXIS= option
   PLOT statement (REG), 8553
VIF option
   MODEL statement (REG), 8477
VPLOTS= option
   PLOT statement (REG), 8555
VREF= option
   PLOT statement (REG), 8553

WEIGHT statement
   REG procedure, 8488
WEIGHT= option
   REWEIGHT statement (REG), 8486
WHITE option
   MODEL statement (REG), 8477

XPX option
   MODEL statement (REG), 8477