SAS/STAT® 9.3 User’s Guide
The REG Procedure
(Chapter)
Chapter 76
The REG Procedure

Contents

Overview: REG Procedure .................................................. 6306
Getting Started: REG Procedure ........................................... 6308
  Simple Linear Regression ................................................. 6308
  Polynomial Regression .................................................... 6312
  Using PROC REG Interactively ......................................... 6322
Syntax: REG Procedure ..................................................... 6323
  PROC REG Statement ...................................................... 6325
  ADD Statement ............................................................. 6338
  BY Statement .............................................................. 6339
  DELETE Statement ......................................................... 6339
  FREQ Statement ........................................................... 6339
  ID Statement ............................................................... 6340
  MODEL Statement .......................................................... 6340
  MTEST Statement .......................................................... 6351
  OUTPUT Statement ......................................................... 6353
  PAINT Statement .......................................................... 6355
  PLOT Statement ............................................................ 6358
  PRINT Statement ........................................................... 6370
  REFIT Statement ........................................................... 6371
  RESTRICT Statement ....................................................... 6371
  REWEIGHT Statement ..................................................... 6373
  TEST Statement ............................................................ 6376
  VAR Statement ............................................................. 6377
  WEIGHT Statement ........................................................ 6377
Details: REG Procedure ..................................................... 6378
  Missing Values ............................................................... 6378
  Input Data Sets ............................................................. 6378
  Output Data Sets ........................................................... 6382
  Interactive Analysis ....................................................... 6389
  Model-Selection Methods ................................................ 6393
  Criteria Used in Model-Selection Methods ........................... 6396
  Limitations in Model-Selection Methods .............................. 6397
  Parameter Estimates and Associated Statistics ..................... 6397
  Predicted and Residual Values ......................................... 6400
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 6438.

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 6393.

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 $R^2$ is maximized.
MINR similar to the MAXR method, except that variables are switched so that the increase in $R^2$ from adding a variable to the model is minimized.
RSQUARE finds a specified number of models with the highest $R^2$ in a range of model sizes.
ADJRSQ finds a specified number of models with the highest adjusted $R^2$ 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.

```plaintext
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 following data are from a study of nineteen children. Height and weight are measured for each child.

```plaintext
title 'Simple Linear Regression';
data Class;
  input Name $ Height Weight Age @@;
datalines;
Alfred 69.0 112.5 14  Alice 56.5 84.0 13  Barbara 65.3 98.0 13
Carol  62.8 102.5 14  Henry 63.5 102.5 14  James 57.3 83.0 12
Jane    59.8  84.5 12  Janet 62.5 112.5 15  Jeffrey 62.5 84.0 13
John    59.0  99.5 12  Joyce 51.3  50.5 11  Judy   64.3  90.0 14
Louise  56.3  77.0 12  Mary  66.5 112.0 15  Philip 72.0 150.0 16
Robert  64.8 128.0 12  Ronald 67.0 133.0 15  Thomas 57.5  85.0 11
William 66.5 112.0 15  ;
```

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.
ods graphics on;
proc reg;
   model Weight = Height;
run;
ods graphics off;

Figure 76.1 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 76.1 ANOVA Table

<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 76.2 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
Weight = $-143.0 + 3.9 \times \text{Height}$

**Figure 76.2** Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variable</strong></td>
</tr>
<tr>
<td>Intercept</td>
</tr>
<tr>
<td>Height</td>
</tr>
</tbody>
</table>

If ODS Graphics is enabled, then PROC REG produces a variety of plots. **Figure 76.3** shows a plot of the residuals versus the regressor and **Figure 76.4** shows a panel of diagnostic plots.

**Figure 76.3** 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 76.4 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.
Chapter 76: The REG Procedure

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, YearSq, is created in the DATA step; this is done since polynomial effects such as Year*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
;
ods graphics on;
proc reg data=USPopulation plots=ResidualByPredicted;
  var Year Sq;
  model Population=Year / r clm cli;
run;
```

The “Analysis of Variance” and “Parameter Estimates” tables are displayed in Figure 76.5.
The REG Procedure
Model: MODEL1
Dependent Variable: Population

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>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

Parameter Estimates

| 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 76.6 shows the “Output Statistics” table. The residual, its standard error, and the studentized residuals are displayed for each observation. The studentized residual is the residual divided by its standard error. The magnitude of each studentized residual is shown in a print plot. Studentized residuals follow a $t$ distribution and can be used to identify outlying or extreme observations. Asterisks (*) extending beyond the dashed lines indicate that the residual is more than three standard errors from zero. Many observations having absolute studentized residuals greater than two might indicate an inadequate model. 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.
### Output Statistics

The REG Procedure
Model: MODEL1
Dependent Variable: Population

#### Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Variable</th>
<th>Value</th>
<th>Mean</th>
<th>Predict</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.9290</td>
<td>-40.5778</td>
<td>10.4424</td>
<td>-62.3602</td>
<td>-18.7953</td>
<td>-97.7280</td>
<td>44.5068</td>
</tr>
<tr>
<td>2</td>
<td>5.3080</td>
<td>-27.6991</td>
<td>9.7238</td>
<td>-47.9826</td>
<td>-7.4156</td>
<td>-84.2950</td>
<td>28.8968</td>
</tr>
<tr>
<td>3</td>
<td>7.2390</td>
<td>-14.8205</td>
<td>9.0283</td>
<td>-33.6533</td>
<td>4.0123</td>
<td>-70.9128</td>
<td>42.7179</td>
</tr>
<tr>
<td>4</td>
<td>9.6380</td>
<td>-1.9418</td>
<td>8.3617</td>
<td>-19.3841</td>
<td>15.5004</td>
<td>-57.5827</td>
<td>53.6991</td>
</tr>
<tr>
<td>5</td>
<td>12.8660</td>
<td>10.9368</td>
<td>7.7314</td>
<td>-5.1906</td>
<td>27.0643</td>
<td>-44.3060</td>
<td>66.1797</td>
</tr>
<tr>
<td>6</td>
<td>17.0690</td>
<td>23.1555</td>
<td>7.1470</td>
<td>8.9070</td>
<td>38.7239</td>
<td>-31.0839</td>
<td>78.7148</td>
</tr>
<tr>
<td>7</td>
<td>23.1910</td>
<td>36.9411</td>
<td>6.6208</td>
<td>22.8834</td>
<td>50.5048</td>
<td>-17.9174</td>
<td>91.3056</td>
</tr>
<tr>
<td>8</td>
<td>31.4430</td>
<td>49.5727</td>
<td>6.1675</td>
<td>36.7075</td>
<td>62.4380</td>
<td>-4.8073</td>
<td>103.9528</td>
</tr>
<tr>
<td>9</td>
<td>39.8180</td>
<td>62.4514</td>
<td>5.8044</td>
<td>50.3436</td>
<td>74.5592</td>
<td>8.2455</td>
<td>116.6573</td>
</tr>
<tr>
<td>10</td>
<td>50.1550</td>
<td>75.3300</td>
<td>5.5491</td>
<td>63.7547</td>
<td>86.9053</td>
<td>21.2406</td>
<td>129.4195</td>
</tr>
<tr>
<td>11</td>
<td>62.9470</td>
<td>88.2087</td>
<td>5.4170</td>
<td>76.9090</td>
<td>99.5084</td>
<td>34.1776</td>
<td>142.2395</td>
</tr>
<tr>
<td>12</td>
<td>75.9940</td>
<td>101.0873</td>
<td>5.4170</td>
<td>89.7876</td>
<td>112.3870</td>
<td>47.0562</td>
<td>155.1184</td>
</tr>
<tr>
<td>13</td>
<td>91.9720</td>
<td>113.9660</td>
<td>5.4941</td>
<td>102.3907</td>
<td>125.5413</td>
<td>59.8765</td>
<td>168.0553</td>
</tr>
<tr>
<td>14</td>
<td>105.7100</td>
<td>126.8446</td>
<td>5.8044</td>
<td>114.7368</td>
<td>138.9524</td>
<td>72.6387</td>
<td>181.0505</td>
</tr>
<tr>
<td>15</td>
<td>122.7750</td>
<td>139.7233</td>
<td>6.1675</td>
<td>126.8580</td>
<td>152.5885</td>
<td>85.3432</td>
<td>194.1033</td>
</tr>
<tr>
<td>16</td>
<td>131.6690</td>
<td>152.6019</td>
<td>6.6208</td>
<td>138.7912</td>
<td>166.4126</td>
<td>97.9904</td>
<td>207.2134</td>
</tr>
<tr>
<td>17</td>
<td>151.3250</td>
<td>165.4805</td>
<td>7.1470</td>
<td>150.5721</td>
<td>180.3890</td>
<td>110.5812</td>
<td>220.3799</td>
</tr>
<tr>
<td>18</td>
<td>179.3230</td>
<td>188.3592</td>
<td>7.7314</td>
<td>162.2317</td>
<td>194.4866</td>
<td>123.1163</td>
<td>233.6020</td>
</tr>
<tr>
<td>19</td>
<td>203.2110</td>
<td>211.2378</td>
<td>8.3617</td>
<td>173.7956</td>
<td>208.6801</td>
<td>135.5969</td>
<td>246.8787</td>
</tr>
<tr>
<td>20</td>
<td>226.5420</td>
<td>240.1165</td>
<td>9.0283</td>
<td>185.2837</td>
<td>222.9466</td>
<td>148.0241</td>
<td>260.2088</td>
</tr>
<tr>
<td>21</td>
<td>248.7100</td>
<td>269.9951</td>
<td>9.7238</td>
<td>196.7116</td>
<td>237.2786</td>
<td>160.3992</td>
<td>273.5910</td>
</tr>
<tr>
<td>22</td>
<td>281.4220</td>
<td>329.7873</td>
<td>10.4424</td>
<td>208.0913</td>
<td>251.6562</td>
<td>172.7235</td>
<td>287.0240</td>
</tr>
</tbody>
</table>

#### Std Error

<table>
<thead>
<tr>
<th>Obs</th>
<th>Residual</th>
<th>Student Residual</th>
<th>Cook's D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>23.077</td>
<td>1.929</td>
<td>0.381</td>
</tr>
<tr>
<td>2</td>
<td>23.389</td>
<td>1.411</td>
<td>0.172</td>
</tr>
<tr>
<td>3</td>
<td>23.666</td>
<td>0.932</td>
<td>0.063</td>
</tr>
<tr>
<td>4</td>
<td>23.909</td>
<td>0.484</td>
<td>0.014</td>
</tr>
<tr>
<td>5</td>
<td>24.121</td>
<td>0.088</td>
<td>0.000</td>
</tr>
<tr>
<td>6</td>
<td>24.300</td>
<td>-0.237</td>
<td>0.003</td>
</tr>
<tr>
<td>7</td>
<td>24.449</td>
<td>-0.552</td>
<td>0.011</td>
</tr>
<tr>
<td>8</td>
<td>24.567</td>
<td>-0.738</td>
<td>0.017</td>
</tr>
<tr>
<td>9</td>
<td>24.655</td>
<td>-0.916</td>
<td>0.023</td>
</tr>
<tr>
<td>10</td>
<td>24.714</td>
<td>-1.019</td>
<td>0.026</td>
</tr>
<tr>
<td>11</td>
<td>24.743</td>
<td>-1.021</td>
<td>0.025</td>
</tr>
<tr>
<td>12</td>
<td>24.743</td>
<td>-1.041</td>
<td>0.025</td>
</tr>
<tr>
<td>13</td>
<td>24.743</td>
<td>-0.890</td>
<td>0.020</td>
</tr>
<tr>
<td>14</td>
<td>24.655</td>
<td>-0.857</td>
<td>0.020</td>
</tr>
<tr>
<td>15</td>
<td>24.567</td>
<td>-0.690</td>
<td>0.015</td>
</tr>
<tr>
<td>16</td>
<td>24.449</td>
<td>-0.856</td>
<td>0.027</td>
</tr>
<tr>
<td>17</td>
<td>24.300</td>
<td>-0.583</td>
<td>0.015</td>
</tr>
<tr>
<td>18</td>
<td>24.121</td>
<td>0.040</td>
<td>0.000</td>
</tr>
<tr>
<td>19</td>
<td>23.909</td>
<td>0.501</td>
<td>0.015</td>
</tr>
<tr>
<td>20</td>
<td>23.666</td>
<td>0.948</td>
<td>0.065</td>
</tr>
<tr>
<td>21</td>
<td>23.389</td>
<td>1.356</td>
<td>0.159</td>
</tr>
<tr>
<td>22</td>
<td>23.077</td>
<td>2.234</td>
<td>0.511</td>
</tr>
</tbody>
</table>
Figure 76.7 shows the residual statistics table. A fairly close agreement between the PRESS statistic (see Table 76.8) 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 76.7** Residual Statistics

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Residuals</td>
<td>0</td>
</tr>
<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 76.8 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 76.8 Diagnostics Panel

Figure 76.9 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 76.9 Residual Plot

Figure 76.10 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.
These default plots provide strong evidence that the 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 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 76.11 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 76.12 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:

```sas
proc reg data=USPopulation plots=predictions(X=Year);
    model Population=Year Yearsq;
quit;
```
ods graphics off;

Figure 76.13 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.

**Figure 76.13** Predictions and Residuals by Year

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 6389 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

```
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.
The following statements are available in PROC REG:

```
PROC REG <options>;
  <label:>MODEL dependents=< regressors > < / options >;
  BY variables;
  FREQ variable;
  ID variables;
  VAR variables;
  WEIGHT variable;
  ADD variables;
  DELETE variables;
  <label:>MTEST <equation, . . . ,equation > < / options >;
  OUTPUT <OUT=SAS-data-set>< keyword=names> < . . . keyword=names >;
  PAINT <condition | ALLOBS > < / options > | < STATUS / UNDO >;
  RESTRICT equation, . . . ,equation ;
  REWEIGHT <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 >;
  <label:>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 6388.) 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.
- **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.
- **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.
PROC REG Statement

PROC REG <options> ;

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 76.1 lists the options you can use with 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><strong>Data Set Options</strong></td>
<td></td>
</tr>
<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 $R^2$ 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>
<tr>
<td><strong>ODS Graphics Options</strong></td>
<td></td>
</tr>
<tr>
<td>PLOTS=</td>
<td>produces ODS graphical displays</td>
</tr>
<tr>
<td><strong>Traditional Graphics Options</strong></td>
<td></td>
</tr>
<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><strong>Display Options</strong></td>
<td></td>
</tr>
<tr>
<td>CORR</td>
<td>displays correlation matrix for variables listed in MODEL and VAR statements</td>
</tr>
</tbody>
</table>
Table 76.1  continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIMPLE</td>
<td>displays simple statistics for each variable listed in MODEL and VAR statements</td>
</tr>
<tr>
<td>USSCP</td>
<td>displays uncorrected sums of squares and crossproducts matrix</td>
</tr>
<tr>
<td>ALL</td>
<td>displays all statistics (CORR, SIMPLE, and USSCP)</td>
</tr>
<tr>
<td>NOPRINT</td>
<td>suppresses output</td>
</tr>
<tr>
<td>LINEPRINTER</td>
<td>creates printer plots</td>
</tr>
</tbody>
</table>

Other Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALPHA=number</td>
<td>sets significance value for confidence and prediction intervals and tests</td>
</tr>
<tr>
<td>SINGULAR=</td>
<td>sets criterion for checking for singularity</td>
</tr>
</tbody>
</table>

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.

**ANNOTATE=SAS-data-set**
**ANNO=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.

**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 OUTEST= data set. This option is valid only if the OUTEST= option is also specified. See the section “OUTEST= Data Set” on page 6382.

**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 \( R^2 \) to the OUTEST= data set.

**GOUT=**

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.

**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=**

requests that parameter estimates and optional model fit summary statistics be output to this data set. See the section “OUTEST= Data Set” on page 6382 for details. If you want to create a permanent SAS data set, you must specify a two-level name (refer to the section “SAS Files” in *SAS Language Reference: Concepts* for more information about permanent SAS data sets).

**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=**

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 6388 for details. If you want to create a permanent SAS data set, you must specify a two-level name (refer to the section “SAS Files” in *SAS Language Reference: Concepts* for more information about permanent SAS data sets).

**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.
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 cooksd)
plots(only) = (diagnostics(stats=all) fit(nocli stats=(aic sbc))

ODS Graphics must be enabled before requesting plots. For example:

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 609 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 76.2 lists the default set of plots produced.
Table 76.2  Default Graphs 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 | number**

specifies that plots with elements that require processing more than number points be suppressed. The default is MAXPOINTS=5000. This cutoff is 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 76.3 lists the statistics that you can request. STATS=ALL requests all these statistics; STATS=NONE suppresses them.

Table 76.3  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>
</tbody>
</table>
Table 76.3 continued

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Default</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<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 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 Mallow’s $C_p$ 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, Mallow’s $C_p$, 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 selecting 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 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 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:
Chapter 76: The REG Procedure

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.

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 smooth 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 selected that yields a local minimum of the corrected Akaike information criterion (AICC). See Chapter 52, “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.

RESIDUALS < residual-options) >
produces panels of the residuals versus 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.

The following residual-options are available:

SMOOTH
requests a nonparametric smooth of the residuals for each regressor. Each nonparametric fit is a loess fit that uses local linear polynomials, linear interpolation, and a smoothing parameter selected that yields a local minimum of the corrected Akaike information criterion (AICC). See Chapter 52, “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 | RIDGEPROGRAM | RIDGEPLAINT < (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 76.5.)

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.

VARSERPERPLOT=ALL
VARSERPERPLOT=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 VARSERPERPLOT=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 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 OUTTEST= 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 6432). 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 76.5 illustrates this option.

If ODS Graphics is enabled (see the section “ODS Graphics” on page 6438), 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-7 on most machines. This option is rarely needed.

Singularity checking is described in the section “Computational Methods” on page 6433.

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 OUTEST= data set. The _TYPE_ variable values STDERR, LnB, UnB, T, and PVALUE, where \( n = 100(1 - \alpha) \), identify these rows in the OUTEST= data set. The \( \alpha \) level can be set with the ALPHA= option in the PROC REG or MODEL statement. The OUTEST= 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 6389 for an example.
BY Statement

BY variables;

You can specify a BY statement with PROC REG to obtain separate analyses on 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 for 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.

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 6389.

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.

### 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

\(< \textit{label:}> \textbf{MODEL} \; \textit{dependents}=\textit{< regressors>}; \) \(< \;/ \textit{options}>; \)

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 \( X1 \) in the model, you cannot use \( X1*X1 \) in the MODEL statement but must create a new variable (for example, \( X1SQUARE=X1*X1 \)) in a DATA step and use this new variable in the MODEL statement. The label in the MODEL statement is optional.

Table 76.4 lists the options available in the MODEL statement. Equations for the statistics available are given in the section “Model Fit and Diagnostic Statistics” on page 6407.

<table>
<thead>
<tr>
<th>Table 76.4 MODEL Statement Options</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Option</strong></td>
</tr>
<tr>
<td>-----------</td>
</tr>
<tr>
<td><strong>Model Selection and Details of Selection</strong></td>
</tr>
<tr>
<td>SELECTION=</td>
</tr>
<tr>
<td>BEST=</td>
</tr>
<tr>
<td>DETAILS</td>
</tr>
<tr>
<td>DETAILS=</td>
</tr>
</tbody>
</table>
Table 76.4  continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<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</td>
</tr>
<tr>
<td></td>
<td>estimates to the OUTEST= data set</td>
</tr>
<tr>
<td>RIDGE=</td>
<td>performs ridge regression analysis and outputs estimates to the</td>
</tr>
<tr>
<td></td>
<td>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</td>
</tr>
<tr>
<td></td>
<td>and switching process</td>
</tr>
<tr>
<td>STOP=</td>
<td>stops selection criterion</td>
</tr>
</tbody>
</table>

**Statistics**

<table>
<thead>
<tr>
<th>Statistic</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADJRSQ</td>
<td>computes adjusted $R^2$</td>
</tr>
<tr>
<td>AIC</td>
<td>computes Akaike’s information criterion</td>
</tr>
<tr>
<td>B</td>
<td>computes parameter estimates for each model</td>
</tr>
<tr>
<td>BIC</td>
<td>computes Sawa’s Bayesian information criterion</td>
</tr>
<tr>
<td>CP</td>
<td>computes Mallows’ $C_p$ statistic</td>
</tr>
<tr>
<td>GMSEP</td>
<td>computes estimated MSE of prediction assuming multivariate</td>
</tr>
<tr>
<td></td>
<td>normality</td>
</tr>
<tr>
<td>JP</td>
<td>computes $J_p$, the final prediction error</td>
</tr>
<tr>
<td>MSE</td>
<td>computes MSE for each model</td>
</tr>
<tr>
<td>PC</td>
<td>computes Amemiya’s prediction criterion</td>
</tr>
<tr>
<td>RMSE</td>
<td>displays root MSE for each model</td>
</tr>
<tr>
<td>SBC</td>
<td>computes the SBC statistic</td>
</tr>
<tr>
<td>SP</td>
<td>computes $S_p$ statistic for each model</td>
</tr>
<tr>
<td>SSE</td>
<td>computes error sum of squares for each model</td>
</tr>
</tbody>
</table>

**Data Set Options**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDF</td>
<td>outputs the number of regressors, the error degrees of freedom,</td>
</tr>
<tr>
<td></td>
<td>and the model $R^2$ to the OUTEST= data set</td>
</tr>
<tr>
<td>OUTSEB</td>
<td>outputs standard errors of the parameter estimates to the OUT-</td>
</tr>
<tr>
<td></td>
<td>EST= data set</td>
</tr>
<tr>
<td>OUTSTB</td>
<td>outputs standardized parameter estimates to the OUTTEST= data set.</td>
</tr>
<tr>
<td></td>
<td>Use only with the RIDGE= or PCOMIT= option.</td>
</tr>
<tr>
<td>OUTVIF</td>
<td>outputs the variance inflation factors to the OUTTEST= data set.</td>
</tr>
<tr>
<td></td>
<td>Use only with the RIDGE= or PCOMIT= option.</td>
</tr>
<tr>
<td>PRESS</td>
<td>outputs the PRESS statistic to the OUTTEST= data set</td>
</tr>
<tr>
<td>RSQUARE</td>
<td>has same effect as the EDF option</td>
</tr>
</tbody>
</table>

**Regression Calculations**

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>displays inverse of sums of squares and crossproducts</td>
</tr>
<tr>
<td>XPX</td>
<td>displays sums-of-squares and crossproducts matrix</td>
</tr>
</tbody>
</table>
### Option Details on Estimates

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACOV</td>
<td>displays heteroscedasticity-consistent covariance matrix of estimates and</td>
</tr>
<tr>
<td></td>
<td>heteroscedasticity-consistent standard errors</td>
</tr>
<tr>
<td>ACOVMETHOD=</td>
<td>specifies method for computing the asymptotic heteroscedasticity-consistent</td>
</tr>
<tr>
<td></td>
<td>covariance matrix</td>
</tr>
<tr>
<td>COLLIN</td>
<td>produces collinearity analysis</td>
</tr>
<tr>
<td>COLLINOINT</td>
<td>produces collinearity analysis with intercept adjusted out</td>
</tr>
<tr>
<td>CORRB</td>
<td>displays correlation matrix of estimates</td>
</tr>
<tr>
<td>COVB</td>
<td>displays covariance matrix of estimates</td>
</tr>
<tr>
<td>HCC</td>
<td>displays heteroscedasticity-consistent standard errors</td>
</tr>
<tr>
<td>HCCMETHOD=</td>
<td>specifies method for computing the asymptotic heteroscedasticity-consistent</td>
</tr>
<tr>
<td></td>
<td>covariance matrix</td>
</tr>
<tr>
<td>LACKFIT</td>
<td>performs lack-of-fit test</td>
</tr>
<tr>
<td>PARTIALR2</td>
<td>displays squared semipartial correlation coefficients computed using Type I</td>
</tr>
<tr>
<td></td>
<td>sums of squares</td>
</tr>
<tr>
<td>PCORR1</td>
<td>displays squared partial correlation coefficients computed using Type I</td>
</tr>
<tr>
<td></td>
<td>sums of squares</td>
</tr>
<tr>
<td>PCORR2</td>
<td>displays squared partial correlation coefficients computed using Type II</td>
</tr>
<tr>
<td></td>
<td>sums of squares</td>
</tr>
<tr>
<td>SCORR1</td>
<td>displays squared semipartial correlation coefficients computed using Type I</td>
</tr>
<tr>
<td></td>
<td>sums of squares</td>
</tr>
<tr>
<td>SCORR2</td>
<td>displays squared semipartial correlation coefficients computed using Type II</td>
</tr>
<tr>
<td></td>
<td>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>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</td>
</tr>
<tr>
<td></td>
<td>value</td>
</tr>
<tr>
<td>CLM</td>
<td>computes $100(1 - \alpha)%$ confidence limits for the expected value of the</td>
</tr>
<tr>
<td></td>
<td>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>
</tbody>
</table>
### Table 76.4 continued

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<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**

- **ALL** 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, XFX
- **ALPHA=** sets significance value for confidence and prediction intervals and tests
- **NOPRINT** suppresses display of results
- **SIGMA=** specifies the true standard deviation of error term for computing CP and BIC
- **SINGULAR=** sets criterion for checking for singularity

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 6425 for more information.

**ACOVMETHOD=0,1,2, or 3**

See the `HCCMETHOD=` option.

**ADJRSQ**

computes $R^2$ 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 XFX.

**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 6400 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 6400 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 6405.

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 6405.

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 6396 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 6431.

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 6431.

EDF
outputs the number of regressors in the model excluding and including the intercept, the error degrees of freedom, and the model $R^2$ 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 SE-
LECTION=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 6351) 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:

```
model y={x1 x2} x3 / selection=stepwise
    groupnames='x1 x2' 'x3';
```

Another example:

```
model y={ht wgt age} bodyfat / selection=forward
    groupnames='ht wgt age' '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 co-
variance 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 Het-
eroscedasticity” on page 6425 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 mod-
els. 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 6409 for details.
outputs $J_\rho$, 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 $J_\rho$ 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 $J_\rho$ 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 6426 for more information. Refer to Draper and Smith (1981) for a discussion of lack-of-fit tests.

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

MAXSTEP=$n$
specifies the maximum number of steps that are done when SELECTION=FORWARD, SELECTION=BACKWARD, or 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.

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

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.

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.

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.

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_.

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 6400 for more information.

**PARTIAL**
requests partial regression leverage plots for each regressor. You can use the 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 6438), then these partial plots are produced in panels with up to six plots per panel. See the section “Influence Statistics” on page 6409 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 6409 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=** \(m\)
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
A statistic to measure the influence of each observation on the parameter estimates. See the section “Predicted and Residual Values” on page 6400 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 6432). 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 76.5 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.
Chapter 76: The REG Procedure

**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 6393 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−7 on most machines. This option is rarely needed. Singularity checking is described in the section “Computational Methods” on page 6433.

**SLENTRY=** *value*
**SLE=** *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.

**SPEC**
performs a test that the first and second moments of the model are correctly specified. See the section “Testing for Heteroscedasticity” on page 6425 for more information.

**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

is used to begin the comparing-and-switching process in the MAXR, MINR, and STEPWISE methods for a model containing the first s independent variables in the MODEL statement, where s is the START value. For these methods, the default is START=0.

For the RSQUARE, ADJRSQ, and CP methods, 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 6405 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 6405 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 \(L\) 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

\[L\beta M = 0\]

See the section “Multivariate Tests” on page 6427 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 \(L\) matrix and \(c\) vector in the preceding formula; the equations for dependent variables form the \(M\) matrix. If no equations for the dependent variables are given, PROC REG uses an identity matrix for \(M\), 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 \(X1\) and \(X2\) parameters are zero for \(Y1\), \(Y2\), and \(Y3\). In addition, the second MTEST statement tests the hypothesis that the \(X1\) 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 \(Y1\) and \(Y2\).

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 \(M\) matrix and various intermediate calculations.

**MSTAT=FAPPROX**
**MSTAT=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 permanent SAS data set, you must specify a two-level name (for example, `libref.data-set-name`).

For more information about permanent SAS data sets, refer to the section “SAS Files” in *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 6378 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 6400 and the section “Influence Statistics” on page 6409. 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 DATAn 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:

```sas
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 6407 for computational formulas.

**Table 76.5  Keywords for OUTPUT Statement**

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>COOKD=names</td>
<td>Cook’s D influence statistic</td>
</tr>
<tr>
<td>COVRATIO=names</td>
<td>standard influence of observation on covariance of betas, as discussed in the section “Influence Statistics” on page 6409</td>
</tr>
<tr>
<td>DFFITS=names</td>
<td>standard influence of observation on predicted value</td>
</tr>
<tr>
<td>H=names</td>
<td>leverage, $x_i (X'X)^{-1} x_i'$</td>
</tr>
<tr>
<td>LCL=names</td>
<td>lower bound of a 100(1 − α)% 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>LCLM=names</td>
<td>lower bound of a 100(1 − α)% confidence interval for the expected value (mean) of the dependent variable</td>
</tr>
<tr>
<td>PREDICTED</td>
<td>P=names</td>
</tr>
<tr>
<td>PRESS=names</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>RSTUDENT=names</td>
<td>a studentized residual with the current observation deleted</td>
</tr>
</tbody>
</table>
### Table 76.5  continued

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>STDI=names</td>
<td>standard error of the individual predicted value</td>
</tr>
<tr>
<td>STDP=names</td>
<td>standard error of the mean predicted value</td>
</tr>
<tr>
<td>STDR=names</td>
<td>standard error of the residual</td>
</tr>
<tr>
<td>STUDENT=names</td>
<td>studentized residuals, which are the residuals divided by their standard errors</td>
</tr>
<tr>
<td>UCL=names</td>
<td>upper bound of a 100(1 − α)% confidence interval for an individual prediction</td>
</tr>
<tr>
<td>UCLM=names</td>
<td>upper bound of a 100(1 − α)% confidence interval for the expected value (mean) of the dependent variable</td>
</tr>
</tbody>
</table>

### PAINT Statement

**PAINT**<br/>

PAINT <condition | ALLOBS > </options> ;<br/>

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. Refer to 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

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

and

```
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:

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

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

```
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,

```
paint allops / 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

```
paint r.>0 / symbol='a';
reweight r.>0;
refit;
paint / symbol='b';
```
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,

```
paint / symbol='#';
```
changes the painting symbol for the observations selected by the most recent PAINT statement to ’#’. As another example,
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';
\Codecomment{...other interactive statements}
paint obs. =1 / symbol='b';
\Codecomment{...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**

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 6438) 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 76.6). 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 6432 for more information)

- keywords for model fit summary statistics available in the OUTEST= data set with _TYPE_ = PARMS (see Table 76.6). 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 6368, 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 6368 for more information.

Traditional Graphics

The display of traditional graphics is described in the following paragraphs, the options are summarized in Table 76.6 and described in the section “Dictionary of PLOT Statement Options” on page 6363.

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=, NO-COLLECT, 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 76.7 lists options to control the display of the equation.

Four statistics are displayed by default in the right margin: the number of observations, $R^2$, the adjusted $R^2$, 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 76.7.

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 6363 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 6363 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 6363 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 76.6 Keywords for PLOT Statement *xvariables***

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diagnostic Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>COOKD.</td>
<td>Cook’s <em>D</em> 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 − α)% confidence interval for individual prediction</td>
</tr>
<tr>
<td>LCLM.</td>
<td>lower bound of 100(1 − α)% confidence interval for the mean of the dependent variable</td>
</tr>
<tr>
<td>PREDICTED.</td>
<td>predicted values</td>
</tr>
<tr>
<td>PRED.</td>
<td></td>
</tr>
<tr>
<td>PRESS.</td>
<td>residuals from refitting the model with current observation deleted</td>
</tr>
<tr>
<td>RESIDUAL.</td>
<td>residuals</td>
</tr>
<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 − α)% confidence interval for individual prediction</td>
</tr>
<tr>
<td>UCLM.</td>
<td>upper bound of 100(1 − α)% confidence interval for the mean of the dependent variables</td>
</tr>
<tr>
<td><strong>Other Keywords Used with Diagnostic Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>NPP.</td>
<td>normal probability-probability plot</td>
</tr>
<tr>
<td>NQQ.</td>
<td>normal quantile-quantile plot</td>
</tr>
<tr>
<td>OBS.</td>
<td>observation number (cannot plot against OUTEST= statistics)</td>
</tr>
<tr>
<td><strong>Model Fit Summary Statistics</strong></td>
<td></td>
</tr>
<tr>
<td>ADJRSQ.</td>
<td>adjusted R-square</td>
</tr>
<tr>
<td>AIC.</td>
<td>Akaike’s information criterion</td>
</tr>
<tr>
<td>BIC.</td>
<td>Sawa’s Bayesian information criterion</td>
</tr>
<tr>
<td>CP.</td>
<td>Mallows’ <em>Cp</em> statistic</td>
</tr>
<tr>
<td>EDF.</td>
<td>error degrees of freedom</td>
</tr>
<tr>
<td>GMSEP.</td>
<td>estimated MSE of prediction, assuming multivariate normality</td>
</tr>
<tr>
<td>IN.</td>
<td>number of regressors in the model not including the intercept</td>
</tr>
<tr>
<td>JP.</td>
<td>final prediction error</td>
</tr>
</tbody>
</table>
Table 76.6  continued

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSE.</td>
<td>mean squared error</td>
</tr>
<tr>
<td>NP.</td>
<td>number of parameters in the model (including the intercept)</td>
</tr>
<tr>
<td>PC.</td>
<td>Amemiya’s prediction criterion</td>
</tr>
<tr>
<td>RMSE.</td>
<td>root MSE</td>
</tr>
<tr>
<td>RSQ.</td>
<td>R-square</td>
</tr>
<tr>
<td>SBC.</td>
<td>SBC statistic</td>
</tr>
<tr>
<td>SP.</td>
<td>SP statistic</td>
</tr>
<tr>
<td>SSE.</td>
<td>error sum of squares</td>
</tr>
</tbody>
</table>

Summary of PLOT Statement Graphics Options

The following table lists the PLOT statement options by function. 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 6363.

Table 76.7  Traditional Graphics Options

<table>
<thead>
<tr>
<th>Option</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>General Graphics Options</strong></td>
<td></td>
</tr>
<tr>
<td>ANNOTATE=</td>
<td>specifies the annotate data set</td>
</tr>
<tr>
<td>SAS-data-set</td>
<td></td>
</tr>
<tr>
<td>CHOCKING=color</td>
<td>requests a reference line for ( C_p ) model selection criteria</td>
</tr>
<tr>
<td>CMALLOWS=color</td>
<td>requests a reference line for the ( C_p ) model selection criterion</td>
</tr>
<tr>
<td>CONF</td>
<td>requests plots of ( 100(1-\alpha)% ) confidence intervals for the mean</td>
</tr>
<tr>
<td>DESCRIPTION=</td>
<td>specifies a description for graphics catalog member</td>
</tr>
<tr>
<td>'string'</td>
<td></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>
<tr>
<td><strong>Axis and Legend Options</strong></td>
<td></td>
</tr>
<tr>
<td>LEGEND=LEGENDrn</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>
<tr>
<td><strong>Reference Line Options</strong></td>
<td></td>
</tr>
<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>
<tr>
<td><strong>Color Options</strong></td>
<td></td>
</tr>
</tbody>
</table>
Table 76.7  continued

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

**Options for Displaying the Fitted Model Equation**

- MODELFONT=\textcolor{\textcolor{font}}{} specifies font of model equation and model label
- MODELHT=\textcolor{\textcolor{value}}{} specifies text height of model equation and model label
- MODELLAB=\textcolor{\textcolor{label}}{} specifies model label
- NOMODEL suppresses display of the fitted model and the label

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

- AIC displays Akaike’s information criterion
- BIC displays Sawa’s Bayesian information criterion
- CP displays Mallows’ $C_p$ statistic
- EDF displays the error degrees of freedom
- GMSEP displays the estimated MSE of prediction assuming multivariate normality
- IN displays the number of regressors in the model not including the intercept
- JP displays the $J_p$ statistic
- MSE displays the mean squared error
- NOSTAT suppresses display of the default statistics: the number of observations, R-square, adjusted R-square, and root mean square error
- NP displays the number of parameters in the model including the intercept, if any
- PC displays the PC statistic
- SBC displays the SBC statistic
- SP displays the $S_p$ statistic
- SSE displays the error sum of squares
- STATFONT=\textcolor{\textcolor{font}}{} specifies font of text displayed in the margin
- STATHT=\textcolor{\textcolor{value}}{} specifies height of text displayed in the margin

**Dictionary of PLOT Statement Options**

The following entries describe the PLOT statement \textit{options} in detail. Note that these \textit{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 $C_p = 2p - p_{full}$ 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=LEGENDn**

- specifies the LEGENDn statement to be used. The LEGENDn statement is a global graphics statement; see *SAS/GRAPH: Reference* for more information.
LHREF=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.

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 6360 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.

STATFONT=font specifies the font used for displaying the statistics that appear in the plot margin. See SAS/GRAPH: Reference for tables of software fonts.

STATHT=height specifies the text height of the statistics that appear in the plot margin.

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.

VAXIS=values VA=values specifies tick mark values for the vertical axis.

VREF=values specifies where reference lines perpendicular to the vertical axis are to appear.
Line Printer Plots

Line printer plots are requested with the LINEPRINTER option in the 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', . . . , '9' are used for variable values 1, 2, . . . , 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 LINEPRINTER option is used, you can specify the following options in the 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;
```

### 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 6438), 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:

**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 6340 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

```
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 6378 for more detail.
REFIT Statement

**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;
paint student.*p.;
  reweight student.>2;
  refit;
paint student.>2;
paint 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** 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

```
  equation
```

or
equation = equation

The form of each equation is

\[ c_1 \times \text{variable}_1 \pm c_2 \times \text{variable}_2 \pm \cdots \pm c_n \times \text{variable}_n \]

where the \( c_j \)'s are constants and the \( \text{variable}_j \)'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:

```
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**

```plaintext
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:

```plaintext
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 6419.

### Specifying Condition

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

```plaintext
variable compare value
```

or

```plaintext
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. Refer to 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:

```
reweight name='steve';
```

```
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:

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

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

```
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,

```
reweight allobs / reset;
```

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

```
reweight allobs;
```

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:

```
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,

```
reweight / reset;
```

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';
\Codecomment{...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:

```plaintext
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.

TEST Statement

```plaintext
< label: > TEST equation,<…,equation> </ 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:
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 6425).

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**

\[
\text{VAR } \text{variables;} \\
\]

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**

\[
\text{WEIGHT } \text{variable;} \\
\]

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 41, “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

```
model y=x1 x1*x1;
```

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:

```
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 76.2.

```plaintext
proc corr data=fitness outp=r noprint;
  var Oxygen RunTime Age Weight RunPulse MaxPulse RestPulse;
proc print data=r;
proc reg data=r;
  model Oxygen=RunTime Age Weight;
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 76.14. Figure 76.15 shows results from the regression that uses the TYPE=CORR data as an input data set.

**Figure 76.14** TYPE=CORR Data Set Created by PROC CORR

<table>
<thead>
<tr>
<th></th>
<th>R</th>
<th>R</th>
<th>R</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MEAN</td>
<td>STD</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>47.3758</td>
<td>5.3272</td>
<td>31.0000</td>
</tr>
<tr>
<td>2</td>
<td>10.5861</td>
<td>1.3874</td>
<td>31.0000</td>
</tr>
<tr>
<td>3</td>
<td>47.6774</td>
<td>5.2114</td>
<td>31.0000</td>
</tr>
<tr>
<td>4</td>
<td>77.4445</td>
<td>8.3286</td>
<td>31.0000</td>
</tr>
<tr>
<td>5</td>
<td>169.645</td>
<td>10.252</td>
<td>31.0000</td>
</tr>
<tr>
<td>6</td>
<td>173.774</td>
<td>9.1640</td>
<td>31.0000</td>
</tr>
<tr>
<td>7</td>
<td>53.4516</td>
<td>7.6194</td>
<td>31.0000</td>
</tr>
<tr>
<td>8</td>
<td>1 MEAN</td>
<td>47.3758</td>
<td>10.5861</td>
</tr>
<tr>
<td>9</td>
<td>2 STD</td>
<td>5.3272</td>
<td>1.3874</td>
</tr>
<tr>
<td>10</td>
<td>3 N</td>
<td>31.0000</td>
<td>31.0000</td>
</tr>
<tr>
<td>11</td>
<td>4 CORR Oxygen</td>
<td>1.0000</td>
<td>-0.8622</td>
</tr>
<tr>
<td>12</td>
<td>5 CORR RunTime</td>
<td>-0.8622</td>
<td>1.0000</td>
</tr>
<tr>
<td>13</td>
<td>6 CORR Age</td>
<td>-0.3046</td>
<td>0.1887</td>
</tr>
<tr>
<td>14</td>
<td>7 CORR Weight</td>
<td>-0.1628</td>
<td>0.1435</td>
</tr>
<tr>
<td>15</td>
<td>8 CORR RunPulse</td>
<td>-0.3980</td>
<td>0.3136</td>
</tr>
<tr>
<td>16</td>
<td>9 CORR MaxPulse</td>
<td>-0.2367</td>
<td>0.2261</td>
</tr>
<tr>
<td>17</td>
<td>10 CORR RestPulse</td>
<td>-0.3994</td>
<td>0.4504</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></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>47.3758</td>
<td>10.5861</td>
<td>47.6774</td>
<td>77.4445</td>
<td>169.645</td>
<td>173.774</td>
</tr>
<tr>
<td>2</td>
<td>STD</td>
<td>5.3272</td>
<td>1.3874</td>
<td>5.2114</td>
<td>8.3286</td>
<td>10.252</td>
<td>9.1640</td>
</tr>
<tr>
<td>3</td>
<td>N</td>
<td>31.0000</td>
<td>31.0000</td>
<td>31.0000</td>
<td>31.0000</td>
<td>31.0000</td>
<td>31.0000</td>
</tr>
<tr>
<td>4</td>
<td>CORR Oxygen</td>
<td>1.0000</td>
<td>-0.8622</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>CORR RunTime</td>
<td>-0.8622</td>
<td>1.0000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>CORR Age</td>
<td>-0.3046</td>
<td>0.1887</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CORR Weight</td>
<td>-0.1628</td>
<td>0.1435</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>CORR RunPulse</td>
<td>-0.3980</td>
<td>0.3136</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>CORR MaxPulse</td>
<td>-0.2367</td>
<td>0.2261</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>CORR RestPulse</td>
<td>-0.3994</td>
<td>0.4504</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 76.14 TYPE=CORR Data Set Created by PROC CORR
Figure 76.15 Regression on TYPE=CORR Data Set

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>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

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|-------------------|----------------|---------|------|---|
| Intercept| 1  | 93.12615          | 7.55916        | 12.32   | <.0001 |
| RunTime  | 1  | -3.14039          | 0.36738        | -8.55   | <.0001 |
| Age      | 1  | -0.17388          | 0.09955        | -1.75   | 0.0921 |
| Weight   | 1  | -0.05444          | 0.06181        | -0.88   | 0.3862 |

The following example uses the saved crossproducts matrix:

```
proc reg data=fitness outsscp=sscp noprint;
model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse;
proc print data=sscp;
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 76.16 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 76.16** 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>
</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>
</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>
</tr>
<tr>
<td>3</td>
<td>SSCP</td>
<td>Age</td>
<td>1478.00</td>
<td>15687.24</td>
<td>71282.00</td>
<td>114158.90</td>
</tr>
<tr>
<td>4</td>
<td>SSCP</td>
<td>Weight</td>
<td>2400.78</td>
<td>25464.71</td>
<td>114158.90</td>
<td>188008.20</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>
</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>
</tr>
<tr>
<td>7</td>
<td>SSCP</td>
<td>RestPulse</td>
<td>1657.00</td>
<td>17684.05</td>
<td>78806.00</td>
<td>128409.28</td>
</tr>
<tr>
<td>8</td>
<td>SSCP</td>
<td>Oxygen</td>
<td>1468.65</td>
<td>15356.14</td>
<td>69767.75</td>
<td>113522.26</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>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</th>
<th>RunPulse</th>
<th>MaxPulse</th>
<th>RestPulse</th>
<th>Oxygen</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5259.00</td>
<td>5387.00</td>
<td>1657.00</td>
<td>1468.65</td>
</tr>
<tr>
<td>2</td>
<td>55806.29</td>
<td>57113.72</td>
<td>17684.05</td>
<td>15356.14</td>
</tr>
<tr>
<td>3</td>
<td>250194.00</td>
<td>256218.00</td>
<td>78806.00</td>
<td>69767.75</td>
</tr>
<tr>
<td>4</td>
<td>407745.67</td>
<td>417764.62</td>
<td>128409.28</td>
<td>113522.26</td>
</tr>
<tr>
<td>5</td>
<td>895317.00</td>
<td>916499.00</td>
<td>281928.00</td>
<td>248497.31</td>
</tr>
<tr>
<td>6</td>
<td>916499.00</td>
<td>938641.00</td>
<td>288583.00</td>
<td>254866.75</td>
</tr>
<tr>
<td>7</td>
<td>281928.00</td>
<td>288583.00</td>
<td>90311.00</td>
<td>78015.41</td>
</tr>
<tr>
<td>8</td>
<td>248497.31</td>
<td>254866.75</td>
<td>78015.41</td>
<td>70429.86</td>
</tr>
<tr>
<td>9</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
<td>31.00</td>
</tr>
</tbody>
</table>

**Figure 76.17** also shows the PROC REG results for the reduced model. (For the PROC REG results for the full model, see Figure 76.29.)
Figure 76.17  Regression on TYPE=SSCP Data Set

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>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

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|------|---|
| Intercept| 1  | 93.12615           | 7.55916        | 12.32   | <.0001 |
| RunTime  | 1  | -3.14039           | 0.36738        | -8.55   | <.0001 |
| Age      | 1  | -0.17388           | 0.09955        | -1.75   | 0.0921 |
| Weight   | 1  | -0.05444           | 0.06181        | -0.88   | 0.3862 |

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−$\alpha$) lower confidence limit for the estimate, where $n$ is the nearest integer to 100(1−$\alpha$) and $\alpha$ defaults to 0.05 or is set by using the ALPHA= option in the PROC REG or MODEL statement
• **UₙB**, the 100(1−$\alpha$) 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 $R^2$, 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
The following statements produce and display the OUTEST= data set. This example uses the population data given in the section “Polynomial Regression” on page 6312. Figure 76.18 through Figure 76.20 show the regression equations and the resulting OUTEST= data set.

```
proc reg data=USPopulation outest=est;
  m1: model Population=Year;
  m2: model Population=Year YearSq;
proc print data=est;
run;
```

**Figure 76.18** Regression Output for Model M1

<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

**Parameter Estimates**

| 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 _JP_ statistic, if the JP option is specified
- The _PC_ statistic, if the PC option is specified
- The _GMSEP_ statistic, if the GMSEP option is specified
- The _AIC_ statistic, if the AIC option is specified
- The _BIC_ statistic, if the BIC option is specified
- The _SBC_ statistic, if the SBC option is specified
**Figure 76.19**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 |          |        |

**Parameter Estimates**

| 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  | 0.006684346         | 0.00017820     | 37.51   | <.0001|

**Figure 76.20**OUTEST= Data Set

<table>
<thead>
<tr>
<th>Obs</th>
<th>MODEL</th>
<th>TYPE</th>
<th>DEPVAR</th>
<th>RMSE</th>
<th>Intercept</th>
<th>Year</th>
<th>Population</th>
<th>YearSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>m1</td>
<td>FARMS</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>FARMS</td>
<td>Population</td>
<td>2.9998</td>
<td>21630.89</td>
<td>-24.0458</td>
<td>-1</td>
<td>.006684346</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:

```r
proc reg data=USPopulation outest=est tableout alpha=0.1;
  m1: model Population=Year/noprint;
  m2: model Population=Year YearSq/noprint;
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 76.21**.
Chapter 76: The REG Procedure

Figure 76.21 The OUTEST= Data Set When TABLEOUT Is Specified

<table>
<thead>
<tr>
<th>Obs</th>
<th>MODEL</th>
<th>TYPE</th>
<th>DEPVAR</th>
<th>RMSE</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>37.5096</td>
<td>.</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 76.22 and Figure 76.23 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;  
proc print data=est;  
run;
```
**Figure 76.22** PROC REG Output for Physical Fitness Data: Best Models

<table>
<thead>
<tr>
<th>Model</th>
<th>R-Square</th>
<th>Selection Method</th>
<th>AIC</th>
<th>BIC</th>
<th>Estimated MSE</th>
<th>J(p)</th>
<th>MSE</th>
</tr>
</thead>
<tbody>
<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>
</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>
</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>
</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>
</tr>
<tr>
<td>5</td>
<td>0.8480</td>
<td>5.1063</td>
<td>56.2986</td>
<td>61.5667</td>
<td>6.4565</td>
<td>6.1782</td>
<td>5.17634</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>
</tr>
</tbody>
</table>

Parameter Estimates:

<table>
<thead>
<tr>
<th>Model</th>
<th>R-Square</th>
<th>Intercept</th>
<th>Age</th>
<th>Weight</th>
<th>RunTime</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7434</td>
<td>82.42177</td>
<td>.</td>
<td>.</td>
<td>-3.31056</td>
</tr>
<tr>
<td>2</td>
<td>0.7642</td>
<td>88.46229</td>
<td>-0.15037</td>
<td>.</td>
<td>-3.20395</td>
</tr>
<tr>
<td>3</td>
<td>0.8111</td>
<td>111.71806</td>
<td>-0.25640</td>
<td>.</td>
<td>-2.82538</td>
</tr>
<tr>
<td>4</td>
<td>0.8368</td>
<td>98.14789</td>
<td>-0.19773</td>
<td>.</td>
<td>-2.76758</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>
</tr>
<tr>
<td>6</td>
<td>0.8487</td>
<td>102.93448</td>
<td>-0.22697</td>
<td>-0.07418</td>
<td>-2.62865</td>
</tr>
</tbody>
</table>

Parameter Estimates:

<table>
<thead>
<tr>
<th>Model</th>
<th>R-Square</th>
<th>RunPulse</th>
<th>RestPulse</th>
<th>MaxPulse</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.7434</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>2</td>
<td>0.7642</td>
<td>.</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>3</td>
<td>0.8111</td>
<td>-0.13091</td>
<td>.</td>
<td>.</td>
</tr>
<tr>
<td>4</td>
<td>0.8368</td>
<td>-0.34811</td>
<td>.</td>
<td>0.27051</td>
</tr>
<tr>
<td>5</td>
<td>0.8480</td>
<td>-0.37340</td>
<td>.</td>
<td>0.30491</td>
</tr>
<tr>
<td>6</td>
<td>0.8487</td>
<td>-0.36963</td>
<td>-0.02153</td>
<td>0.30322</td>
</tr>
</tbody>
</table>
**Figure 76.23** 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>
</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>
</tr>
<tr>
<td>2</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.67739</td>
<td>88.462</td>
<td>-0.15037</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.44063</td>
<td>111.718</td>
<td>-0.25640</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.31159</td>
<td>98.148</td>
<td>-0.19773</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.27516</td>
<td>102.204</td>
<td>-0.21962</td>
<td>-0.072302</td>
</tr>
<tr>
<td>6</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>Oxygen</td>
<td>2.31695</td>
<td>102.934</td>
<td>-0.22697</td>
<td>-0.074177</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 76.2 to produce an output data set with the OUTSSCP= option. The resulting output is shown in Figure 76.24.

```sas
proc reg data=fitness outsscp=sscp;
var Oxygen RunTime Age Weight RestPulse RunPulse MaxPulse;
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 76.24** SSCP Data Set Created with OUTSSCP= Option: REG Procedure

```
Obs  TYPE_ _NAME_ Intercept Oxygen RunTime Age
1    SSCP  Intercept   31.00    1468.65  328.17  1478.00
2    SSCP  Oxygen       1468.65  70429.86 15356.14  69767.75
3    SSCP  RunTime      328.17   15356.14  3531.80  15687.24
4    SSCP  Age           1478.00  69767.75  15687.24    71282.00
5    SSCP  Weight       2400.78  113522.26 25464.71  114158.90
6    SSCP  RestPulse    1657.00   78015.41  17684.05    78806.00
7    SSCP  RunPulse     5259.00  248497.31 55806.29  250194.00
8    SSCP  MaxPulse     5387.00  254866.75 57113.72  256218.00
9    N     N            31.00     31.00    31.00     31.00
```

```
Obs  Weight RestPulse  RunPulse MaxPulse
1    2400.78    1657.00     5259.00     5387.00
2    113522.26  78015.41     248497.31    254866.75
3    25464.71  17684.05      55806.29     57113.72
4    114158.90  78806.00     250194.00    256218.00
5    188008.20 128409.28     407745.67    417764.62
6    128409.28  90311.00     281928.00    288583.00
7    407745.67  281928.00    895317.00    916499.00
8    417764.62  288583.00    916499.00    938641.00
9    31.00       31.00       31.00       31.00
```

**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 6322 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 6419.

The following statements illustrate the usefulness of the interactive features. First, the full regression model is fit to the Class data (see the section “Getting Started: REG Procedure” on page 6308), and Figure 76.25 is produced.
The REG Procedure  
Model: MODEL1  
Dependent Variable: Weight  

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>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

Parameter Estimates  

| 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 76.26 is produced.
Figure 76.26  Interactive Analysis: Reduced Model

<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>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>R-Square</th>
<th>Adj R-Sq</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Root MSE</td>
<td>11.22625</td>
<td>0.7705</td>
<td>0.7570</td>
<td></td>
</tr>
<tr>
<td>Dependent Mean</td>
<td>100.02632</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coeff Var</td>
<td>11.22330</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Variable  | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|-----------|----|--------------------|----------------|---------|-------|
| Intercept | 1  | -143.02692         | 32.27459       | -4.43   | 0.0004|
| Height    | 1  | 3.89903            | 0.51609        | 7.55    | <.0001|

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

```
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 76.27 is produced, and the scatter plot shows a possible outlier.

```
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 76.28 shows the new scatter plot.

```sas
reweight r.>20;
print;
run;

ods graphics off;
```
Model-Selection Methods

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
Chapter 76: The REG Procedure

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 \( R^2 \) 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 \( R^2 \) for each size.

The MAXR method begins by finding the one-variable model producing the highest \( R^2 \). Then another variable, the one that yields the greatest increase in \( R^2 \), 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 \( R^2 \). After comparing all possible switches, the MAXR method makes the switch that produces the largest increase in \( R^2 \). Comparisons begin again, and the process continues until the MAXR method finds that no switch could increase \( R^2 \). 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 $R^2$. 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 $R^2$ 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 $R^2$ 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 $R^2$ 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 $R^2$ for each number of variables considered. The other selection methods are not guaranteed to find the model with the largest $R^2$. 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 $R^2$ statistic is used as the criterion for selecting models, and the method finds the models with the highest adjusted $R^2$ 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

If the RSQUARE or STEPWISE procedure (as documented in SAS User’s Guide: Statistics, Version 5 Edition) is requested, PROC REG with the appropriate model-selection method is actually used.

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 $R^2$, 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, refer to Daniel and Wood (1980).

The adjusted $R^2$ statistic is an alternative to $R^2$ that is adjusted for the number of parameters in the model. The adjusted $R^2$ statistic is calculated as

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

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 76.2. Figure 76.30 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 76.29). The $F$ statistic for the overall model is significant, indicating that the model explains a significant portion of the variation in the data.

**Figure 76.29** ANOVA Table

<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 |

The procedure next displays parameter estimates and some associated statistics (Figure 76.30). First, the estimates are shown, followed by their standard errors. The next two columns of the table contain the $t$
Chapter 76: The REG Procedure

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.

**Figure 76.30** SS1, SS2, STB, CLB, COVB, and CORRB Options: Parameter Estimates

| Variable  | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| | Type I SS |
|-----------|----|--------------------|----------------|---------|------|---------------------|------------|
| Intercept | 1  | 102.93448          | 12.40326       | 8.30    | <.0001| 69578                |
| RunTime   | 1  | -2.62865           | 0.38456        | -6.84   | <.0001| 632.90010            |
| Age       | 1  | -0.22697           | 0.09984        | -2.27   | 0.0322| 17.76563             |
| Weight    | 1  | -0.07418           | 0.05459        | -1.36   | 0.1869| 5.60522              |
| RunPulse  | 1  | -0.36963           | 0.11985        | -3.08   | 0.0051| 38.87574             |
| MaxPulse  | 1  | 0.30322            | 0.13650        | 2.22    | 0.0360| 26.82640             |
| RestPulse | 1  | -0.02153           | 0.06605        | -0.33   | 0.7473| 0.57051              |

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</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>369.72831</td>
<td>0</td>
<td>77.33541</td>
</tr>
<tr>
<td>RunTime</td>
<td>1</td>
<td>250.82210</td>
<td>-0.68460</td>
<td>-3.42235</td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>27.74577</td>
<td>-0.22204</td>
<td>-0.43303</td>
</tr>
<tr>
<td>Weight</td>
<td>1</td>
<td>9.91059</td>
<td>-0.11597</td>
<td>-0.31685</td>
</tr>
<tr>
<td>RunPulse</td>
<td>1</td>
<td>51.05806</td>
<td>-0.71133</td>
<td>-0.61699</td>
</tr>
<tr>
<td>MaxPulse</td>
<td>1</td>
<td>26.49142</td>
<td>0.52161</td>
<td>0.02150</td>
</tr>
<tr>
<td>RestPulse</td>
<td>1</td>
<td>0.57051</td>
<td>-0.03080</td>
<td>-0.15786</td>
</tr>
</tbody>
</table>

The final two tables are produced as a result of requesting the COVB and CORRB options (Figure 76.31). These tables show the estimated covariance matrix of the parameter estimates, and the estimated correlation matrix of the estimates.
Figure 76.31 SS1, SS2, STB, CLB, COVB, and CORRB Options: Covariances and Correlations

<table>
<thead>
<tr>
<th>Variable</th>
<th>Intercept Covariance</th>
<th>RunTime Covariance</th>
<th>Age Covariance</th>
<th>Weight Covariance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.280796516</td>
<td>-0.832761667</td>
<td>-0.147954715</td>
<td>-0.178237818</td>
</tr>
<tr>
<td>RunTime</td>
<td>-0.009047784</td>
<td>0.0046249498</td>
<td>-0.010915224</td>
<td>-0.001372241</td>
</tr>
<tr>
<td>Age</td>
<td>0.0009644683</td>
<td>-0.001372241</td>
<td>0.0003799295</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>0.0143647273</td>
<td>-0.014952457</td>
<td>0.0003425724</td>
<td></td>
</tr>
<tr>
<td>RunPulse</td>
<td>-0.000764507</td>
<td>0.0003425724</td>
<td>0.0043631674</td>
<td></td>
</tr>
</tbody>
</table>

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

For further discussion of the parameters and statistics, see the section “Displayed Output” on page 6433, 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. 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 $100(1 - \alpha)\%$ 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 $100(1 - \alpha)\%$ confidence interval for observed values, then you can use the CLI option, which adds in the variability of the error term. The $\alpha$ 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 6312. The results are shown in Figure 76.32 and Figure 76.33.

```plaintext
data USPop2;
    input Year @@;
    YearSq=Year*Year;
datalines;
2010 2020 2030
;
%data USPop2;
    set USPopulation USPop2;

proc reg data=USPop2;
    id Year;
    model Population=Year YearSq / r cli clm;
run;
```
Figure 76.32  Regression Using the R, CLI, and CLM Options

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

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>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 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 76.33** 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>Variable</th>
<th>Predicted Value</th>
<th>Std Error</th>
<th>95% CL Mean</th>
<th>95% CL Predict</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1790</td>
<td>3.9290</td>
<td>6.2127</td>
<td>1.7565</td>
<td>2.5362</td>
<td>-1.0631</td>
</tr>
<tr>
<td>2</td>
<td>1800</td>
<td>5.3080</td>
<td>5.7226</td>
<td>1.4560</td>
<td>2.6751</td>
<td>8.7701</td>
</tr>
<tr>
<td>3</td>
<td>1810</td>
<td>7.3380</td>
<td>6.5694</td>
<td>1.2118</td>
<td>4.0331</td>
<td>9.1057</td>
</tr>
<tr>
<td>4</td>
<td>1820</td>
<td>9.6380</td>
<td>8.7531</td>
<td>1.0305</td>
<td>6.5963</td>
<td>10.9100</td>
</tr>
<tr>
<td>5</td>
<td>1830</td>
<td>12.8660</td>
<td>12.2737</td>
<td>0.9163</td>
<td>10.3558</td>
<td>14.1916</td>
</tr>
<tr>
<td>6</td>
<td>1840</td>
<td>17.0690</td>
<td>17.1311</td>
<td>0.8650</td>
<td>15.3207</td>
<td>18.9415</td>
</tr>
<tr>
<td>7</td>
<td>1850</td>
<td>23.1910</td>
<td>23.3254</td>
<td>0.8613</td>
<td>21.5227</td>
<td>25.1281</td>
</tr>
<tr>
<td>8</td>
<td>1860</td>
<td>31.4430</td>
<td>30.8566</td>
<td>0.8846</td>
<td>29.0051</td>
<td>32.7080</td>
</tr>
<tr>
<td>9</td>
<td>1870</td>
<td>39.8180</td>
<td>39.7246</td>
<td>0.9163</td>
<td>37.8067</td>
<td>41.6425</td>
</tr>
<tr>
<td>10</td>
<td>1880</td>
<td>50.1550</td>
<td>49.9295</td>
<td>0.9436</td>
<td>47.9545</td>
<td>51.9046</td>
</tr>
<tr>
<td>11</td>
<td>1890</td>
<td>62.9470</td>
<td>61.4713</td>
<td>0.9590</td>
<td>59.4641</td>
<td>63.4785</td>
</tr>
<tr>
<td>12</td>
<td>1900</td>
<td>75.9940</td>
<td>74.3499</td>
<td>0.9590</td>
<td>72.3427</td>
<td>76.3571</td>
</tr>
<tr>
<td>13</td>
<td>1910</td>
<td>91.9720</td>
<td>88.5655</td>
<td>0.9436</td>
<td>86.5904</td>
<td>90.5405</td>
</tr>
<tr>
<td>14</td>
<td>1920</td>
<td>105.7100</td>
<td>104.1178</td>
<td>0.9163</td>
<td>102.2000</td>
<td>106.0357</td>
</tr>
<tr>
<td>15</td>
<td>1930</td>
<td>122.7750</td>
<td>121.0071</td>
<td>0.8846</td>
<td>119.1556</td>
<td>122.8585</td>
</tr>
<tr>
<td>16</td>
<td>1940</td>
<td>131.6690</td>
<td>139.2332</td>
<td>0.8613</td>
<td>137.4305</td>
<td>141.0359</td>
</tr>
<tr>
<td>17</td>
<td>1950</td>
<td>151.3250</td>
<td>158.7962</td>
<td>0.8650</td>
<td>156.9858</td>
<td>160.6066</td>
</tr>
<tr>
<td>18</td>
<td>1960</td>
<td>179.3250</td>
<td>179.6961</td>
<td>0.9163</td>
<td>177.7782</td>
<td>181.6139</td>
</tr>
<tr>
<td>19</td>
<td>1970</td>
<td>203.2110</td>
<td>201.9328</td>
<td>1.0305</td>
<td>199.7759</td>
<td>204.0896</td>
</tr>
<tr>
<td>20</td>
<td>1980</td>
<td>226.5420</td>
<td>225.5064</td>
<td>1.2118</td>
<td>222.9701</td>
<td>228.0427</td>
</tr>
<tr>
<td>21</td>
<td>1990</td>
<td>248.7100</td>
<td>250.4168</td>
<td>1.4560</td>
<td>247.3693</td>
<td>253.4644</td>
</tr>
<tr>
<td>22</td>
<td>2000</td>
<td>281.4220</td>
<td>276.6642</td>
<td>1.7565</td>
<td>272.9877</td>
<td>280.3407</td>
</tr>
<tr>
<td>23</td>
<td>2010</td>
<td>.</td>
<td>.</td>
<td>2.1073</td>
<td>299.8377</td>
<td>308.6591</td>
</tr>
<tr>
<td>24</td>
<td>2020</td>
<td>.</td>
<td>.</td>
<td>3.042484</td>
<td>327.9285</td>
<td>338.4104</td>
</tr>
<tr>
<td>25</td>
<td>2030</td>
<td>.</td>
<td>.</td>
<td>3.634274</td>
<td>357.2665</td>
<td>369.5883</td>
</tr>
</tbody>
</table>

Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Year</th>
<th>Residual</th>
<th>Std Error</th>
<th>Student Residual</th>
<th>Cook's D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1790</td>
<td>-2.2837</td>
<td>2.432</td>
<td>-0.939</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1800</td>
<td>-0.4146</td>
<td>2.623</td>
<td>-0.158</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1810</td>
<td>0.6696</td>
<td>2.744</td>
<td>0.244</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1820</td>
<td>0.8849</td>
<td>2.817</td>
<td>0.314</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1830</td>
<td>0.5923</td>
<td>2.856</td>
<td>0.207</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>1840</td>
<td>-0.0621</td>
<td>2.872</td>
<td>-0.0216</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>1850</td>
<td>-0.1344</td>
<td>2.873</td>
<td>-0.0468</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1860</td>
<td>0.5864</td>
<td>2.866</td>
<td>0.205</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>1870</td>
<td>0.0934</td>
<td>2.856</td>
<td>0.0327</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1880</td>
<td>0.2255</td>
<td>2.847</td>
<td>0.0792</td>
<td></td>
</tr>
</tbody>
</table>
The REG Procedure
Model: MODEL1
Dependent Variable: Population

Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Year</th>
<th>Residual</th>
<th>Std Error Residual</th>
<th>Student Residual</th>
<th>Cook's D</th>
<th>2-1 0 1 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>1890</td>
<td>1.4757</td>
<td>2.842</td>
<td>0.519</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>1900</td>
<td>1.6441</td>
<td>2.842</td>
<td>0.578</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1910</td>
<td>3.4065</td>
<td>2.847</td>
<td>1.196</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>1920</td>
<td>1.5922</td>
<td>2.856</td>
<td>0.557</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>1930</td>
<td>1.7679</td>
<td>2.866</td>
<td>0.617</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>1940</td>
<td>-7.5642</td>
<td>2.873</td>
<td>-2.632</td>
<td></td>
<td>*****</td>
</tr>
<tr>
<td>17</td>
<td>1950</td>
<td>-7.4712</td>
<td>2.872</td>
<td>-2.601</td>
<td></td>
<td>*****</td>
</tr>
<tr>
<td>18</td>
<td>1960</td>
<td>-0.3731</td>
<td>2.856</td>
<td>-0.131</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>1970</td>
<td>1.7282</td>
<td>2.817</td>
<td>0.454</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1980</td>
<td>1.0356</td>
<td>2.744</td>
<td>0.377</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1990</td>
<td>-1.7068</td>
<td>2.623</td>
<td>-0.651</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>2000</td>
<td>4.7578</td>
<td>2.432</td>
<td>1.957</td>
<td></td>
<td>***</td>
</tr>
<tr>
<td>23</td>
<td>2010</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>2020</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>2030</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

After producing the usual analysis of variance and parameter estimates tables (Figure 76.32), the procedure displays the results of requesting the options for predicted and residual values (Figure 76.33). 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 columnar print plot of studentized residuals and Cook’s $D$ statistics are displayed as a result of requesting the R option. In the plot of studentized residuals, the large number of observations with absolute values greater than two indicates an inadequate model. You can use ODS Graphics to obtain plots of studentized residuals by predicted values or leverage; see Example 76.1 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 76.2. 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 76.34 shows how this problem is diagnosed.

```latex
\begin{verbatim}
data fit2;
  set fitness; Dif=RunPulse-RestPulse;
proc reg data=fit2;
  model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse Dif;
run;
\end{verbatim}
```

**Figure 76.34** Model That Is Not Full Rank: REG Procedure

<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

| Variable   | DF | 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  | -0.36963 | 0.11985        | -3.08   | 0.0051|
| MaxPulse   | 1  | 0.30322  | 0.13650        | 2.22    | 0.0360|
| RestPulse  | 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 $1/VIF$.

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

This example uses the COLLIN option on the fitness data found in Example 76.2. The following statements produce Figure 76.35.

```plaintext
proc reg data=fitness;
   model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse
   / tol vif collin;
run;
```
### Chapter 76: The REG Procedure

**Figure 76.35** 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  |
Coeff Var  | 4.89057  |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| Tolerance |
|----------|----|--------------------|----------------|---------|------|-----------|
| Intercept| 1  | 102.93448          | 12.40326       | 8.30    | <.0001| .          |
| RunTime  | 1  | -2.62865           | 0.38456        | -6.84   | <.0001| 0.62859   |
| Age      | 1  | -0.22697           | 0.09984        | -2.27   | 0.0322| 0.66101   |
| Weight   | 1  | -0.07418           | 0.05459        | -1.36   | 0.1869| 0.86555   |
| RunPulse | 1  | -0.36963           | 0.11985        | -3.08   | 0.0051| 0.11852   |
| MaxPulse | 1  | 0.30322            | 0.13650        | 2.22    | 0.0360| 0.11437   |
| RestPulse| 1  | -0.02153           | 0.06605        | -0.33   | 0.7473| 0.70642   |

Parameter Estimates

<table>
<thead>
<tr>
<th>Variable</th>
<th>DF</th>
<th>Variance</th>
<th>Inflation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>1</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>RunTime</td>
<td>1</td>
<td>1.59087</td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>1</td>
<td>1.51284</td>
<td></td>
</tr>
<tr>
<td>Weight</td>
<td>1</td>
<td>1.15553</td>
<td></td>
</tr>
<tr>
<td>RunPulse</td>
<td>1</td>
<td>8.43727</td>
<td></td>
</tr>
<tr>
<td>MaxPulse</td>
<td>1</td>
<td>8.74385</td>
<td></td>
</tr>
<tr>
<td>RestPulse</td>
<td>1</td>
<td>1.41559</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 \( \hat{b} = (X'X)^{-1}X'Y \). 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 76.8 contains the summary statistics for assessing the fit of the model.

### Table 76.8 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>
</tbody>
</table>
Table 76.8 continued

<table>
<thead>
<tr>
<th>MODEL Option or Statistic</th>
<th>Definition or Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>SST&lt;sub&gt;1&lt;/sub&gt;</td>
<td>the total sum of squares corrected for the mean for the dependent variable</td>
</tr>
<tr>
<td>SSE</td>
<td>the error sum of squares</td>
</tr>
<tr>
<td>MSE</td>
<td>$\frac{SSE}{n-p}$</td>
</tr>
<tr>
<td>$R^2$</td>
<td>$1 - \frac{SSE}{SST_i}$</td>
</tr>
<tr>
<td>ADJRSQ</td>
<td>$1 - \frac{(n-i)(1-R^2)}{n-p}$</td>
</tr>
<tr>
<td>AIC</td>
<td>$n \ln \left( \frac{SSE}{n} \right) + 2p$</td>
</tr>
<tr>
<td>BIC</td>
<td>$n \ln \left( \frac{SSE}{n} \right) + 2(p+2)q - 2q^2$ where $q = \frac{n\hat{\sigma}^2}{SSE}$</td>
</tr>
<tr>
<td>CP ($C_p$)</td>
<td>$\frac{SSE}{\hat{\sigma}^2} + 2p - n$</td>
</tr>
<tr>
<td>GMSEP</td>
<td>$\frac{MSE(n+1)(n-2)}{n(n-p-1)} = \frac{1}{n} S_p(n+1)(n-2)$</td>
</tr>
<tr>
<td>JP ($J_p$)</td>
<td>$\frac{n+p}{n} \cdot MSE$</td>
</tr>
<tr>
<td>PRESS</td>
<td>$\frac{n+p}{n-p} (1 - R^2) = J_p \left( \frac{n}{SST_i} \right)$</td>
</tr>
<tr>
<td>RMSE</td>
<td>$\sqrt{MSE}$</td>
</tr>
<tr>
<td>SBC</td>
<td>$n \ln \left( \frac{SSE}{n} \right) + p \ln(n)$</td>
</tr>
<tr>
<td>SP ($S_p$)</td>
<td>$\frac{MSE}{n-p-1}$</td>
</tr>
</tbody>
</table>

Table 76.9 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 6409. Each statistic is computed for each observation.

Table 76.9 Formulas and Definitions for Diagnostic Statistics

<table>
<thead>
<tr>
<th>MODEL Option or Statistic</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>PRED ($\hat{Y}_i$)</td>
<td>$X_i b$</td>
</tr>
<tr>
<td>RES ($r_i$)</td>
<td>$Y_i - \hat{Y}_i$</td>
</tr>
<tr>
<td>H ($h_i$)</td>
<td>$x_i (X'X)^{-1} x_i'$</td>
</tr>
<tr>
<td>STDP</td>
<td>$\sqrt{h_i \hat{\sigma}^2}$</td>
</tr>
</tbody>
</table>
### Table 76.9  continued

<table>
<thead>
<tr>
<th>MODEL Option or Statistic</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>STDI</td>
<td>$\sqrt{(1 + h_i)\hat{\sigma}^2}$</td>
</tr>
<tr>
<td>STDR</td>
<td>$\sqrt{(1 - h_i)\hat{\sigma}^2}$</td>
</tr>
<tr>
<td>LCL</td>
<td>$\hat{y}<em>i - t</em>{\alpha / 2} \text{STDI}$</td>
</tr>
<tr>
<td>LCLM</td>
<td>$\hat{y}<em>i - t</em>{\alpha / 2} \text{STDP}$</td>
</tr>
<tr>
<td>UCL</td>
<td>$\hat{y}<em>i + t</em>{\alpha / 2} \text{STDI}$</td>
</tr>
<tr>
<td>UCLM</td>
<td>$\hat{y}<em>i + t</em>{\alpha / 2} \text{STDP}$</td>
</tr>
<tr>
<td>STUDENT</td>
<td>$\frac{r_i}{\text{STDR}_i}$</td>
</tr>
<tr>
<td>RSTUDENT</td>
<td>$\frac{\delta_{(i)}}{\sqrt{1 - h_i}}$</td>
</tr>
<tr>
<td>COOKD</td>
<td>$\frac{1}{p} \text{STUDENT}^2 \frac{\text{STDP}^2}{\text{STDR}^2}$</td>
</tr>
<tr>
<td>COVRATIO</td>
<td>$\frac{\text{det}(\hat{\sigma}^2 (X'X)^{-1})}{\text{det}(\hat{\sigma}^2)^2}$</td>
</tr>
<tr>
<td>DFFITS</td>
<td>$\frac{(\hat{\sigma}<em>{(i)} \sqrt{h_i})}{b_j - b</em>{(i)}j}$</td>
</tr>
<tr>
<td>DFBETAS</td>
<td>$\frac{\delta_{(i)} \sqrt{(X'X)_{jj}}}{r_i}$</td>
</tr>
<tr>
<td>PRESS(predr)</td>
<td>$\frac{1}{1 - h_i}$</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 $\mathbf{b}(i)$ be the parameter estimates after deleting the $i$th observation; let $s(i)^2$ be the variance estimate after deleting the $i$th observation; let $\mathbf{X}(i)$ be the $\mathbf{X}$ matrix without the $i$th observation; let $\hat{y}(i)$ be the $i$th 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 \( \frac{2p}{n} \), 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 \( s^2(i) \) 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:

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

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 6400.
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 $(j, j)$th 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 $2\sqrt{n}$ as a size-adjusted cutoff.

The following statements use the population example in the section “Polynomial Regression” on page 6312. See Figure 76.32 for the fitted regression equation. The INFLUENCE option produces the tables shown in Figure 76.36 and Figure 76.37.

```sas
proc reg data=USPopulation;
   model Population=Year YearSq / influence;
run;
```
### Figure 76.36 Regression Using the INFLUENCE Option

The REG Procedure  
Model: MODEL1  
Dependent Variable: Population

#### Output Statistics

<table>
<thead>
<tr>
<th>Obs</th>
<th>Residual</th>
<th>RStudent</th>
<th>Hat Diag</th>
<th>Cov Ratio</th>
<th>DFFITS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.2837</td>
<td>-0.9361</td>
<td>0.3429</td>
<td>1.5519</td>
<td>-0.6762</td>
</tr>
<tr>
<td>2</td>
<td>-0.4146</td>
<td>-0.1540</td>
<td>0.2356</td>
<td>1.5325</td>
<td>-0.0855</td>
</tr>
<tr>
<td>3</td>
<td>0.6696</td>
<td>0.2379</td>
<td>0.1632</td>
<td>1.3923</td>
<td>0.1050</td>
</tr>
<tr>
<td>4</td>
<td>0.8849</td>
<td>0.3065</td>
<td>0.1180</td>
<td>1.3128</td>
<td>0.1121</td>
</tr>
<tr>
<td>5</td>
<td>0.5923</td>
<td>0.2021</td>
<td>0.0933</td>
<td>1.2883</td>
<td>0.0648</td>
</tr>
<tr>
<td>6</td>
<td>-0.0621</td>
<td>-0.0210</td>
<td>0.0831</td>
<td>1.2827</td>
<td>-0.0063</td>
</tr>
<tr>
<td>7</td>
<td>-0.1344</td>
<td>-0.0455</td>
<td>0.0824</td>
<td>1.2813</td>
<td>-0.0136</td>
</tr>
<tr>
<td>8</td>
<td>0.5864</td>
<td>0.1994</td>
<td>0.0870</td>
<td>1.2796</td>
<td>0.0615</td>
</tr>
<tr>
<td>9</td>
<td>0.0934</td>
<td>0.0318</td>
<td>0.0933</td>
<td>1.2969</td>
<td>0.0102</td>
</tr>
<tr>
<td>10</td>
<td>0.2255</td>
<td>0.0771</td>
<td>0.0990</td>
<td>1.3040</td>
<td>0.0255</td>
</tr>
<tr>
<td>11</td>
<td>1.4757</td>
<td>0.5090</td>
<td>0.1022</td>
<td>1.2550</td>
<td>0.1717</td>
</tr>
<tr>
<td>12</td>
<td>1.6441</td>
<td>0.5680</td>
<td>0.1022</td>
<td>1.2420</td>
<td>0.1916</td>
</tr>
<tr>
<td>13</td>
<td>3.4065</td>
<td>1.2109</td>
<td>0.0990</td>
<td>1.0320</td>
<td>0.4013</td>
</tr>
<tr>
<td>14</td>
<td>1.5922</td>
<td>0.5470</td>
<td>0.0933</td>
<td>1.2345</td>
<td>0.1755</td>
</tr>
<tr>
<td>15</td>
<td>1.7679</td>
<td>0.6064</td>
<td>0.0870</td>
<td>1.2123</td>
<td>0.1871</td>
</tr>
<tr>
<td>16</td>
<td>-7.5642</td>
<td>-3.2147</td>
<td>0.0824</td>
<td>0.3286</td>
<td>-0.9636</td>
</tr>
<tr>
<td>17</td>
<td>-7.4712</td>
<td>-3.1550</td>
<td>0.0831</td>
<td>0.3425</td>
<td>-0.9501</td>
</tr>
<tr>
<td>18</td>
<td>-0.3731</td>
<td>-0.1272</td>
<td>0.0933</td>
<td>1.2936</td>
<td>-0.0408</td>
</tr>
<tr>
<td>19</td>
<td>1.2782</td>
<td>0.4440</td>
<td>0.1180</td>
<td>1.2906</td>
<td>0.1624</td>
</tr>
<tr>
<td>20</td>
<td>1.0356</td>
<td>0.3687</td>
<td>0.1632</td>
<td>1.3741</td>
<td>0.1628</td>
</tr>
<tr>
<td>21</td>
<td>-1.7068</td>
<td>-0.6406</td>
<td>0.2356</td>
<td>1.4380</td>
<td>-0.3557</td>
</tr>
<tr>
<td>22</td>
<td>4.7578</td>
<td>2.1312</td>
<td>0.3429</td>
<td>0.9113</td>
<td>1.5395</td>
</tr>
</tbody>
</table>

Output Statistics

\[---------DFBETAS---------\]

<table>
<thead>
<tr>
<th>Obs</th>
<th>Intercept</th>
<th>Year</th>
<th>YearSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-0.4924</td>
<td>0.4862</td>
<td>-0.4802</td>
</tr>
<tr>
<td>2</td>
<td>-0.0540</td>
<td>0.0531</td>
<td>-0.0523</td>
</tr>
<tr>
<td>3</td>
<td>0.0517</td>
<td>-0.0505</td>
<td>0.0494</td>
</tr>
<tr>
<td>4</td>
<td>0.0335</td>
<td>-0.0322</td>
<td>0.0310</td>
</tr>
<tr>
<td>5</td>
<td>0.0040</td>
<td>-0.0032</td>
<td>0.0025</td>
</tr>
<tr>
<td>6</td>
<td>0.0012</td>
<td>-0.0012</td>
<td>0.0013</td>
</tr>
<tr>
<td>7</td>
<td>0.0054</td>
<td>-0.0055</td>
<td>0.0056</td>
</tr>
<tr>
<td>8</td>
<td>-0.0339</td>
<td>0.0343</td>
<td>-0.0347</td>
</tr>
<tr>
<td>9</td>
<td>-0.0067</td>
<td>0.0067</td>
<td>-0.0068</td>
</tr>
<tr>
<td>10</td>
<td>-0.0182</td>
<td>0.0183</td>
<td>-0.0183</td>
</tr>
<tr>
<td>11</td>
<td>-0.1272</td>
<td>0.1275</td>
<td>-0.1276</td>
</tr>
<tr>
<td>12</td>
<td>-0.1426</td>
<td>0.1426</td>
<td>-0.1424</td>
</tr>
<tr>
<td>13</td>
<td>-0.2895</td>
<td>0.2889</td>
<td>-0.2880</td>
</tr>
<tr>
<td>14</td>
<td>-0.1173</td>
<td>0.1167</td>
<td>-0.1160</td>
</tr>
<tr>
<td>15</td>
<td>-0.1076</td>
<td>0.1067</td>
<td>-0.1056</td>
</tr>
<tr>
<td>16</td>
<td>0.4130</td>
<td>-0.4063</td>
<td>0.3987</td>
</tr>
<tr>
<td>17</td>
<td>0.2131</td>
<td>-0.2048</td>
<td>0.1957</td>
</tr>
<tr>
<td>18</td>
<td>-0.0007</td>
<td>0.0012</td>
<td>-0.0016</td>
</tr>
<tr>
<td>19</td>
<td>0.0415</td>
<td>-0.0432</td>
<td>0.0449</td>
</tr>
<tr>
<td>20</td>
<td>0.0732</td>
<td>-0.0749</td>
<td>0.0766</td>
</tr>
<tr>
<td>21</td>
<td>-0.2107</td>
<td>0.2141</td>
<td>-0.2176</td>
</tr>
<tr>
<td>22</td>
<td>1.0656</td>
<td>-1.0793</td>
<td>1.0933</td>
</tr>
</tbody>
</table>
**Figure 76.37** Residual Statistics

<table>
<thead>
<tr>
<th>Sum of Residuals</th>
<th>-4.7569E-11</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sum of Squared Residuals</td>
<td>170.97193</td>
</tr>
<tr>
<td>Predicted Residual SS (PRESS)</td>
<td>237.71229</td>
</tr>
</tbody>
</table>

In Figure 76.36, observations 16, 17, and 19 exceed 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:

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

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 76.38.
Figure 76.38  Influence Diagnostics

Cook's D for Population

Observation

Cook's D

0.0

0.1

0.2

0.3

0.4

0.5

0.6

2000

1940

1950
Figure 76.38 continued
Figure 76.38  continued

Influence Diagnostics for Population

Observation

DFFITS

0.0

-0.5

-1.0

0  5  10  15  20

1940  1950  2000
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.
Chapter 76: The REG Procedure

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 request plots with the PARTIAL option.

The following statements use the fitness data in Example 76.2 with the PARTIAL option and ODS Graphics to produce the partial regression leverage plots. The plots are shown in Figure 76.39.

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

Figure 76.39  Partial Regression Leverage Plots

![Partial Regression Leverage Plots](image)
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 Class data (in the section “Getting Started: REG Procedure” on page 6308) 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 76.40.

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

**Figure 76.40** Full Model for Class Data, Residuals Shown

<table>
<thead>
<tr>
<th>Obs</th>
<th>Name</th>
<th>Weight</th>
<th>Age</th>
<th>Height</th>
<th>Predicted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alfred</td>
<td>112.5000</td>
<td>112</td>
<td>50.0000</td>
<td>124.8686</td>
<td>-12.3686</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>84.0000</td>
<td>84</td>
<td>0.0000</td>
<td>78.6273</td>
<td>5.3727</td>
</tr>
<tr>
<td>3</td>
<td>Barbara</td>
<td>98.0000</td>
<td>98</td>
<td>0.0000</td>
<td>110.2812</td>
<td>-12.2812</td>
</tr>
<tr>
<td>4</td>
<td>Carol</td>
<td>102.5000</td>
<td>102</td>
<td>50.0000</td>
<td>102.5670</td>
<td>-0.0670</td>
</tr>
<tr>
<td>5</td>
<td>Henry</td>
<td>102.5000</td>
<td>102</td>
<td>50.0000</td>
<td>105.0849</td>
<td>-2.5849</td>
</tr>
<tr>
<td>6</td>
<td>James</td>
<td>83.0000</td>
<td>83</td>
<td>0.0000</td>
<td>80.2266</td>
<td>2.7734</td>
</tr>
<tr>
<td>7</td>
<td>Jane</td>
<td>84.5000</td>
<td>84.5</td>
<td>0.0000</td>
<td>89.2191</td>
<td>-4.7191</td>
</tr>
<tr>
<td>8</td>
<td>Janet</td>
<td>112.5000</td>
<td>112</td>
<td>50.0000</td>
<td>102.7663</td>
<td>9.7337</td>
</tr>
<tr>
<td>9</td>
<td>Jeffrey</td>
<td>84.0000</td>
<td>84</td>
<td>0.0000</td>
<td>100.2095</td>
<td>-16.2095</td>
</tr>
<tr>
<td>10</td>
<td>John</td>
<td>99.5000</td>
<td>99.5</td>
<td>0.0000</td>
<td>86.3415</td>
<td>13.1585</td>
</tr>
<tr>
<td>11</td>
<td>Joyce</td>
<td>50.5000</td>
<td>50</td>
<td>5.0000</td>
<td>57.3660</td>
<td>-6.8660</td>
</tr>
<tr>
<td>12</td>
<td>Judy</td>
<td>90.0000</td>
<td>90</td>
<td>0.0000</td>
<td>107.9625</td>
<td>-17.9625</td>
</tr>
<tr>
<td>13</td>
<td>Louise</td>
<td>77.0000</td>
<td>77</td>
<td>0.0000</td>
<td>76.6295</td>
<td>0.3705</td>
</tr>
<tr>
<td>14</td>
<td>Mary</td>
<td>112.0000</td>
<td>112</td>
<td>0.0000</td>
<td>117.1544</td>
<td>-5.1544</td>
</tr>
<tr>
<td>15</td>
<td>Philip</td>
<td>150.0000</td>
<td>150</td>
<td>0.0000</td>
<td>138.2164</td>
<td>11.7836</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>128.0000</td>
<td>128</td>
<td>0.0000</td>
<td>107.2043</td>
<td>20.7957</td>
</tr>
<tr>
<td>17</td>
<td>Ronald</td>
<td>133.0000</td>
<td>133</td>
<td>0.0000</td>
<td>118.9529</td>
<td>14.0471</td>
</tr>
<tr>
<td>18</td>
<td>Thomas</td>
<td>85.0000</td>
<td>85</td>
<td>0.0000</td>
<td>79.6676</td>
<td>5.3324</td>
</tr>
<tr>
<td>19</td>
<td>William</td>
<td>112.0000</td>
<td>112</td>
<td>0.0000</td>
<td>117.1544</td>
<td>-5.1544</td>
</tr>
</tbody>
</table>

The REG Procedure
Model: MODEL1
Dependent Variable: Weight
Output Statistics

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 76.41 is produced.

**Figure 76.41** Model with Reweighted Observations

<table>
<thead>
<tr>
<th>Obs</th>
<th>Name</th>
<th>Weight Variable</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alfred</td>
<td>1.0000</td>
<td>112.5000</td>
<td>121.6250</td>
<td>-9.1250</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>1.0000</td>
<td>84.0000</td>
<td>79.9296</td>
<td>4.0704</td>
</tr>
<tr>
<td>3</td>
<td>Barbara</td>
<td>1.0000</td>
<td>98.0000</td>
<td>107.5484</td>
<td>-9.5484</td>
</tr>
<tr>
<td>4</td>
<td>Carol</td>
<td>1.0000</td>
<td>102.5000</td>
<td>102.1663</td>
<td>0.3337</td>
</tr>
<tr>
<td>5</td>
<td>Henry</td>
<td>1.0000</td>
<td>102.5000</td>
<td>104.3632</td>
<td>-1.8632</td>
</tr>
<tr>
<td>6</td>
<td>James</td>
<td>1.0000</td>
<td>83.0000</td>
<td>79.9762</td>
<td>3.0238</td>
</tr>
<tr>
<td>7</td>
<td>Jane</td>
<td>1.0000</td>
<td>84.5000</td>
<td>87.8225</td>
<td>-3.3225</td>
</tr>
<tr>
<td>8</td>
<td>Janet</td>
<td>1.0000</td>
<td>112.5000</td>
<td>103.6889</td>
<td>8.8111</td>
</tr>
<tr>
<td>9</td>
<td>Jeffrey</td>
<td>1.0000</td>
<td>84.0000</td>
<td>98.7606</td>
<td>-14.7606</td>
</tr>
<tr>
<td>10</td>
<td>John</td>
<td>1.0000</td>
<td>99.5000</td>
<td>85.3117</td>
<td>14.1883</td>
</tr>
<tr>
<td>11</td>
<td>Joyce</td>
<td>1.0000</td>
<td>50.5000</td>
<td>58.6811</td>
<td>-8.1811</td>
</tr>
<tr>
<td>12</td>
<td>Judy</td>
<td>0.5000</td>
<td>90.0000</td>
<td>106.8740</td>
<td>-16.8740</td>
</tr>
<tr>
<td>13</td>
<td>Louise</td>
<td>1.0000</td>
<td>77.0000</td>
<td>76.8377</td>
<td>0.1623</td>
</tr>
<tr>
<td>14</td>
<td>Mary</td>
<td>1.0000</td>
<td>112.0000</td>
<td>116.2429</td>
<td>-4.2429</td>
</tr>
<tr>
<td>15</td>
<td>Philip</td>
<td>1.0000</td>
<td>150.0000</td>
<td>135.9688</td>
<td>14.0312</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>0.5000</td>
<td>128.0000</td>
<td>103.5150</td>
<td>24.4850</td>
</tr>
<tr>
<td>17</td>
<td>Ronald</td>
<td>1.0000</td>
<td>133.0000</td>
<td>117.8121</td>
<td>15.1879</td>
</tr>
<tr>
<td>18</td>
<td>Thomas</td>
<td>1.0000</td>
<td>85.0000</td>
<td>78.1398</td>
<td>6.8602</td>
</tr>
<tr>
<td>19</td>
<td>William</td>
<td>1.0000</td>
<td>112.0000</td>
<td>116.2429</td>
<td>-4.2429</td>
</tr>
</tbody>
</table>

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
Reweighting Observations in an Analysis

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 76.42:

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

**Figure 76.42** Observations Excluded from Analysis, Model Refitted, and Observations Reweighted

```
The REG Procedure
Model: MODEL1.5
Dependent Variable: Weight

Output Statistics

<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>1.0000</td>
<td>112.5000</td>
<td>120.9716</td>
<td>-8.4716</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>1.0000</td>
<td>84.0000</td>
<td>79.5342</td>
<td>4.4658</td>
</tr>
<tr>
<td>3</td>
<td>Barbara</td>
<td>1.0000</td>
<td>98.0000</td>
<td>107.0746</td>
<td>-9.0746</td>
</tr>
<tr>
<td>4</td>
<td>Carol</td>
<td>1.0000</td>
<td>102.5000</td>
<td>101.5681</td>
<td>0.9319</td>
</tr>
<tr>
<td>5</td>
<td>Henry</td>
<td>1.0000</td>
<td>102.5000</td>
<td>103.7588</td>
<td>-1.2588</td>
</tr>
<tr>
<td>6</td>
<td>James</td>
<td>1.0000</td>
<td>83.0000</td>
<td>79.7204</td>
<td>3.2796</td>
</tr>
<tr>
<td>7</td>
<td>Jane</td>
<td>1.0000</td>
<td>84.5000</td>
<td>87.5443</td>
<td>-3.0443</td>
</tr>
<tr>
<td>8</td>
<td>Janet</td>
<td>1.0000</td>
<td>112.5000</td>
<td>102.9467</td>
<td>9.5533</td>
</tr>
<tr>
<td>9</td>
<td>Jeffrey</td>
<td>1.0000</td>
<td>84.0000</td>
<td>98.3117</td>
<td>-14.3117</td>
</tr>
<tr>
<td>10</td>
<td>John</td>
<td>1.0000</td>
<td>99.5000</td>
<td>85.0407</td>
<td>14.4593</td>
</tr>
<tr>
<td>11</td>
<td>Joyce</td>
<td>1.0000</td>
<td>50.5000</td>
<td>58.6253</td>
<td>-8.1253</td>
</tr>
<tr>
<td>12</td>
<td>Judy</td>
<td>1.0000</td>
<td>90.0000</td>
<td>106.2625</td>
<td>-16.2625</td>
</tr>
<tr>
<td>13</td>
<td>Louise</td>
<td>1.0000</td>
<td>77.0000</td>
<td>76.5908</td>
<td>0.4092</td>
</tr>
<tr>
<td>14</td>
<td>Mary</td>
<td>1.0000</td>
<td>112.0000</td>
<td>115.4651</td>
<td>-3.4651</td>
</tr>
<tr>
<td>15</td>
<td>Philip</td>
<td>1.0000</td>
<td>150.0000</td>
<td>134.9953</td>
<td>15.0047</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>0.5000</td>
<td>128.0000</td>
<td>103.1923</td>
<td>24.8077</td>
</tr>
<tr>
<td>17</td>
<td>Ronald</td>
<td>0</td>
<td>133.0000</td>
<td>117.0299</td>
<td>15.9701</td>
</tr>
<tr>
<td>18</td>
<td>Thomas</td>
<td>1.0000</td>
<td>85.0000</td>
<td>78.0288</td>
<td>6.9712</td>
</tr>
<tr>
<td>19</td>
<td>William</td>
<td>1.0000</td>
<td>112.0000</td>
<td>115.4651</td>
<td>-3.4651</td>
</tr>
</tbody>
</table>

Sum of Residuals 0
Sum of Squared Residuals 1637.81879
Predicted Residual SS (PRESS) 2473.87984
```

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 76.43:
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 76.44:
The resulting model reflects changes made only by the first REWEIGHT statement since the third REWEIGHT statement negates the effect of the second REWEIGHT statement. Observations 1, 3, 9, 10, 12, 16, and 17 have their weights changed to 0.75. Note that the label MODEL1.9 reflects the use of nine REWEIGHT statements for the current model.

Now suppose you want to reset the observations selected by the most recent REWEIGHT statement to their original weights. Use the REWEIGHT statement with the RESET option to do this. The following statements produce Figure 76.45:

```
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 76.45** 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 Variable</th>
<th>Dependent Variable</th>
<th>Predicted Value</th>
<th>Residual</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alfred</td>
<td>0.7500</td>
<td>112.5000</td>
<td>126.0076</td>
<td>-13.5076</td>
</tr>
<tr>
<td>2</td>
<td>Alice</td>
<td>1.0000</td>
<td>84.0000</td>
<td>77.8727</td>
<td>6.1273</td>
</tr>
<tr>
<td>3</td>
<td>Barbara</td>
<td>0.7500</td>
<td>98.0000</td>
<td>111.2805</td>
<td>-13.2805</td>
</tr>
<tr>
<td>4</td>
<td>Carol</td>
<td>1.0000</td>
<td>102.5000</td>
<td>102.4703</td>
<td>0.0297</td>
</tr>
<tr>
<td>5</td>
<td>Henry</td>
<td>1.0000</td>
<td>102.5000</td>
<td>105.1278</td>
<td>-2.6278</td>
</tr>
<tr>
<td>6</td>
<td>James</td>
<td>1.0000</td>
<td>83.0000</td>
<td>80.2290</td>
<td>2.7710</td>
</tr>
<tr>
<td>7</td>
<td>Jane</td>
<td>1.0000</td>
<td>84.5000</td>
<td>89.7199</td>
<td>-5.2199</td>
</tr>
<tr>
<td>8</td>
<td>Janet</td>
<td>1.0000</td>
<td>112.5000</td>
<td>102.0122</td>
<td>10.4878</td>
</tr>
<tr>
<td>9</td>
<td>Jeffrey</td>
<td>0.7500</td>
<td>94.5000</td>
<td>100.6507</td>
<td>-16.6507</td>
</tr>
<tr>
<td>10</td>
<td>John</td>
<td>0.7500</td>
<td>99.5000</td>
<td>86.6828</td>
<td>12.8172</td>
</tr>
<tr>
<td>11</td>
<td>Joyce</td>
<td>1.0000</td>
<td>50.5000</td>
<td>56.7703</td>
<td>-6.2703</td>
</tr>
<tr>
<td>12</td>
<td>Judy</td>
<td>1.0000</td>
<td>90.0000</td>
<td>108.1649</td>
<td>-18.1649</td>
</tr>
<tr>
<td>13</td>
<td>Louise</td>
<td>1.0000</td>
<td>77.0000</td>
<td>76.4327</td>
<td>0.5673</td>
</tr>
<tr>
<td>14</td>
<td>Mary</td>
<td>1.0000</td>
<td>112.0000</td>
<td>117.1975</td>
<td>-5.1975</td>
</tr>
<tr>
<td>15</td>
<td>Philip</td>
<td>1.0000</td>
<td>150.0000</td>
<td>138.7581</td>
<td>11.2419</td>
</tr>
<tr>
<td>16</td>
<td>Robert</td>
<td>1.0000</td>
<td>128.0000</td>
<td>108.7016</td>
<td>19.2984</td>
</tr>
<tr>
<td>17</td>
<td>Ronald</td>
<td>0.7500</td>
<td>133.0000</td>
<td>119.0957</td>
<td>13.9043</td>
</tr>
<tr>
<td>18</td>
<td>Thomas</td>
<td>1.0000</td>
<td>85.0000</td>
<td>80.3076</td>
<td>4.6924</td>
</tr>
<tr>
<td>19</td>
<td>William</td>
<td>1.0000</td>
<td>112.0000</td>
<td>117.1975</td>
<td>-5.1975</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 $E(\epsilon'\epsilon) = \sigma^2I$. 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

$$e_i = y_i - x_i\beta$$

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.

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

where

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

is the leverage of the $i$th observation.

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

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 $\text{HCCMETHOD}=0,1,2,$ or 3 in the $\text{MODEL}$ statement to select a heteroscedasticity-consistent covariance matrix estimator, with $HC_0$ being the default. The $\text{ACOV}$ option in the $\text{MODEL}$
Chapter 76: The REG Procedure

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, refer to 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 76.6). 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.
Multivariate Tests

The MTEST statement described in the section “MTEST Statement” on page 6351 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:

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

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):

```plaintext
* 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
```

```plaintext
```
The following statements perform a multivariate analysis of variance and produce Figure 76.46 through Figure 76.49:

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

**Figure 76.46** Multivariate Analysis of Variance: REG Procedure

<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>316.00000</td>
<td>63.20000</td>
<td>12.04</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>94.50000</td>
<td>5.25000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>23</td>
<td>410.50000</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE: 2.29129, R-Square: 0.7698, Dependent Mean: 9.75000, Adj R-Sq: 0.7058, Coeff Var: 23.50039

| Variable   | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|------------|----|--------------------|----------------|---------|------|--------|
| Intercept  | 1  | 9.75000            | 0.46771        | 20.85   | <.0001 |
| sexcode    | 1  | 0.16667            | 0.46771        | 0.36    | 0.7257 |
| drug1      | 1  | -2.75000           | 0.66144        | -4.16   | 0.0006 |
| drug2      | 1  | -2.25000           | 0.66144        | -3.40   | 0.0032 |
| sexdrug1   | 1  | -0.66667           | 0.66144        | -1.01   | 0.3269 |
| sexdrug2   | 1  | -0.41667           | 0.66144        | -0.63   | 0.5366 |
Figure 76.47  Multivariate Analysis of Variance: REG Procedure

<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>69.33333</td>
<td>13.86667</td>
<td>2.19</td>
<td>0.1008</td>
</tr>
<tr>
<td>Error</td>
<td>18</td>
<td>114.00000</td>
<td>6.33333</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>23</td>
<td>183.33333</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE   2.51661    R-Square  0.3782
Dependent Mean 8.66667    Adj R-Sq 0.2055
Coeff Var 29.03782

Parameter Estimates

| 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 76.48  Multivariate Analysis of Variance: First Test

The REG Procedure
Model: MODELL1
Multivariate Test: y1y2drug

Multivariate Statistics and Exact F Statistics

<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>0.28053917</td>
<td>23.08</td>
<td>2</td>
<td>18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Pillai's Trace</td>
<td>0.71946083</td>
<td>23.08</td>
<td>2</td>
<td>18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Hotelling-Lawley Trace</td>
<td>2.56456456</td>
<td>23.08</td>
<td>2</td>
<td>18</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Roy's Greatest Root</td>
<td>2.56456456</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 76.49  Multivariate Analysis of Variance: Second Test

The REG Procedure  
Model: MODEL1  
Multivariate Test: drugshow

Error Matrix (E)

94.5 76.5  
76.5 114

Hypothesis Matrix (H)

301 97.5  
97.5 36.333333333

Adjusted Approximate Squared
Canonical Canonical Standard Canonical
Correlation Correlation Error Correlation

1 0.905903 0.899927 0.040101 0.820661  
2 0.244371 . 0.210254 0.059717

Eigenvalues of Inv(E)*H  
= CanRsq/(1-CanRsq)

Eigenvalue Difference Proportion Cumulative

1 4.5760 4.5125 0.9863 0.9863  
2 0.0635 0.0137 1.0000

Test of H0: The canonical correlations in the  
current row and all that follow are zero

Likelihood Approximate  
Ratio F Value Num DF Den DF Pr > F

1 0.16862952 12.20 4 34 <.0001  
2 0.94028273 1.14 1 18 0.2991

Multivariate Statistics and F Approximations

S=2  M=-0.5  N=7.5

Statistic Value F Value Num DF Den DF Pr > F

Wilks' Lambda 0.16862952 12.20 4 34 <.0001  
Pillai's Trace 0.88037810 7.08 4 36 0.0003  
Hotelling-Lawley Trace 4.63953666 19.40 4 19.407 <.0001  
Roy's Greatest Root 4.57602675 41.18 2 18 <.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.  
NOTE: F Statistic for Wilks' Lambda is exact.

The four multivariate test statistics are all highly significant, giving strong evidence that the coefficients of drug1 and drug2 are not zero for both dependent variables.
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, refer to 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 76.50). 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 76.50 Regression Using DW Option

The REG Procedure
Model: MODEL1
Dependent Variable: Population

Durbin-Watson D 1.191
Pr < DW 0.0050
Pr > DW 0.9950
Number of Observations 22
1st Order Autocorrelation 0.323

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, refer to Rawlings (1988), Neter, Wasserman, and Kutner (1990), and Marquardt and Snee (1975). Unlike Rawlings (1988), 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 $1E^{-7}$ 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 6378 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 6407 and the section “Influence Statistics” on page 6409.

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 $R^2$, is a version of $R^2$ 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 \( R^2 \). Results for other models with the same number of variables are then shown in order of decreasing \( R^2 \), 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 \( R^2 \) and \( C_p \), respectively.

For each model considered, the RSQUARE, ADJRSQ, and CP methods display the following:

• 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
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 76.10  ODS Tables Produced by PROC REG

<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 ALL, ACOV</td>
<td></td>
</tr>
<tr>
<td>ACovTestANOVA</td>
<td>Test ANOVA using ACOV estimates</td>
<td>TEST ACOV (MODEL statement)</td>
<td></td>
</tr>
<tr>
<td>ANOVA</td>
<td>Model ANOVA table</td>
<td>MODEL Default</td>
<td></td>
</tr>
<tr>
<td>CanCorr</td>
<td>Canonical correlations for hypothesis combinations</td>
<td>MTEST CANPRINT</td>
<td></td>
</tr>
<tr>
<td>CollinDiag</td>
<td>Collinearity Diagnostics table</td>
<td>MODEL COLLIN</td>
<td></td>
</tr>
<tr>
<td>CollinDiagNoInt</td>
<td>Collinearity Diagnostics for no intercept model</td>
<td>MODEL COLLINOINT</td>
<td></td>
</tr>
<tr>
<td>ConditionBounds</td>
<td>Bounds on condition number</td>
<td>MODEL (SELECTION=BACKWARD</td>
<td>FORWARD</td>
</tr>
<tr>
<td>Corr</td>
<td>Correlation matrix for analysis variables</td>
<td>PROC ALL, CORR</td>
<td></td>
</tr>
<tr>
<td>CorrB</td>
<td>Correlation of estimates</td>
<td>MODEL CORRB</td>
<td></td>
</tr>
<tr>
<td>CovB</td>
<td>Covariance of estimates</td>
<td>MODEL COVB</td>
<td></td>
</tr>
<tr>
<td>CrossProducts</td>
<td>Bordered model $X'X$ matrix</td>
<td>MODEL ALL, XPX</td>
<td></td>
</tr>
<tr>
<td>DWStatistic</td>
<td>Durbin-Watson statistic</td>
<td>MODEL ALL, DW</td>
<td></td>
</tr>
<tr>
<td>DependenceEquations</td>
<td>Linear dependence equations</td>
<td>MODEL Default if needed</td>
<td></td>
</tr>
<tr>
<td>Eigenvalues</td>
<td>MTest eigenvalues</td>
<td>MTEST CANPRINT</td>
<td></td>
</tr>
<tr>
<td>Eigenvectors</td>
<td>MTest eigenvectors</td>
<td>MTEST CANPRINT</td>
<td></td>
</tr>
<tr>
<td>EntryStatistics</td>
<td>Entry statistics for selection methods</td>
<td>MODEL (SELECTION=BACKWARD</td>
<td>FORWARD</td>
</tr>
<tr>
<td>ErrorPlusHypothesis</td>
<td>MTest error plus hypothesis matrix $H+E$</td>
<td>MTEST PRINT</td>
<td></td>
</tr>
<tr>
<td>ErrorSSCP</td>
<td>MTest error matrix $E$</td>
<td>MTEST PRINT</td>
<td></td>
</tr>
<tr>
<td>FitStatistics</td>
<td>Model fit statistics</td>
<td>MODEL Default</td>
<td></td>
</tr>
<tr>
<td>HypothesisSSCP</td>
<td>MTest hypothesis matrix</td>
<td>MTEST PRINT</td>
<td></td>
</tr>
<tr>
<td>InvMTestCov</td>
<td>$\text{Inv}(L \text{Ginv}(X'X) L')$ and $\text{Inv}(Lb-c)$</td>
<td>MTEST DETAILS</td>
<td></td>
</tr>
<tr>
<td>ODS Table Name</td>
<td>Description</td>
<td>Statement</td>
<td>Option</td>
</tr>
<tr>
<td>------------------</td>
<td>------------------------------------------------------------------------------</td>
<td>-----------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>InvTestCov</td>
<td>$\text{Inv}(\mathbf{L} \ \text{Ginv}(\mathbf{X}'\mathbf{X}) \ \mathbf{L}')$ and $\text{Inv}(\mathbf{Lb-c})$</td>
<td>TEST</td>
<td>PRINT</td>
</tr>
<tr>
<td>InvXPX</td>
<td>Bordered $\mathbf{X}'\mathbf{X}$ inverse matrix</td>
<td>MODEL</td>
<td>I</td>
</tr>
<tr>
<td>MTestCov</td>
<td>$\mathbf{L} \ \text{Ginv}(\mathbf{X}'\mathbf{X}) \mathbf{L}'$ and $\mathbf{Lb-c}$</td>
<td>MTEST</td>
<td>DETAILS</td>
</tr>
<tr>
<td>MTransform</td>
<td>MTest matrix $\mathbf{M}$, 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>Default</td>
<td></td>
</tr>
<tr>
<td>OutputStatistics</td>
<td>Output statistics table</td>
<td>MODEL</td>
<td>ALL, CLI, CLM, INFERENCE, 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, INFERENCE, 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 R-square, Adj-RSq, 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>$\mathbf{L} \ \text{Ginv}(\mathbf{X}'\mathbf{X}) \mathbf{L}'$ and $\mathbf{Lb-c}$</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, with the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” on page 609 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 608 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 76.8 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 form normality.
Following Rawlings (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 76.1.6 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 76.13 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 6409 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 76.5 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|CP`), 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 76.2 provides several examples.

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 76.11.

<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>CooksDPlot</td>
<td>Cook’s $D$ statistic versus observation number</td>
<td>COOKSD</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>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 scatter plot of data</td>
<td>FIT</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>
</tbody>
</table>
### Table 76.11  \textit{continued}

<table>
<thead>
<tr>
<th>ODS Graph Name</th>
<th>Plot Description</th>
<th>PLOTS Option</th>
</tr>
</thead>
<tbody>
<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>Plot of residuals versus regressor</td>
<td>RESIDUALS</td>
</tr>
<tr>
<td>RFPPlot</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>VIFPlot</td>
<td>Plot of VIF traces</td>
<td>RIDGE(UNPACK)</td>
</tr>
</tbody>
</table>

---

**Examples: REG Procedure**

**Example 76.1: 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 following 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 (\textit{Sports Illustrated}, April 20, 1987) are for the 1987 season and the performance measures are from 1986 (Collier Books, \textit{The 1987 Baseball Encyclopedia Update}).
data baseball;
  length name $ 18;
  length team $ 12;
  input name $ 1-18 no_atbat no_hits no_home no_runs no_rbi no_bb yr_major
  cr_atbat cr_hits cr_home cr_runs cr_rbi cr_bb league $ division $ team $ position $ no_outs no_assts no_error salary;
  logSalary = log10(salary);
  label name="Player's Name"
  no_hits="Hits in 1986"
  no_runs="Runs in 1986"
  no_rbi="RBIs in 1986"
  no_bb="Walks in 1986"
  yr_major="Years in MLB"
  cr_hits="Career Hits"
  salary="1987 Salary in $ Thousands"
  logSalary = "log10(Salary)";
  datalines;
  Allanson, Andy 293 66 1 30 29 14
  1 293 66 1 30 29 14
  American East Cleveland C 446 33 20 .
  Ashby, Alan 315 81 7 24 38 39
  14 3449 835 69 321 414 375
  National West Houston C 632 43 10 475
  Davis, Alan 479 130 18 66 72 76
  3 1624 457 63 224 266 263
  American West Seattle 1B 880 82 14 480
  Dawson, Andre 496 141 20 65 78 37
  11 5628 1575 225 828 838 354
  National East Montreal RF 200 11 3 500
  Galarraga, Andres 321 87 10 39 42 30
  2 396 101 12 48 46 33
  National East Montreal 1B 805 40 4 91.5
  Griffin, Alfredo 594 169 4 74 51 35
  11 4408 1133 19 501 336 194
  ... more lines ...
  Wilson, Willie 631 170 9 77 44 31
  11 4908 1457 30 775 357 249
  American West KansasCity CF 408 4 3 1000
; 

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:

ods graphics on;

proc reg data=baseball;
  id name team league;
  model logSalary = no_hits no_runs no_rbi no_bb yr_major cr_hits;
run;
Example 76.1: Modeling Salaries of Major League Baseball Players

Output 76.1.1 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 76.1.1 Default Output from PROC REG**

<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>22.92208</td>
<td>3.82035</td>
<td>60.56</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>256</td>
<td>16.14954</td>
<td>0.06308</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>262</td>
<td>39.07162</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.25117  R-Square 0.5867  Dependent Mean 2.57416  Adj R-Sq 0.5770  Coeff Var 9.75719

| Variable     | Label          | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|--------------|----------------|----|--------------------|----------------|---------|------|
| Intercept    | Intercept      | 1  | 1.80065            | 0.05912        | 30.46   | <.0001|
| no_hits      | Hits in 1986   | 1  | 0.00288            | 0.00091244     | 3.15    | 0.0018|
| no_runs      | Runs in 1986   | 1  | 0.00008638         | 0.00173        | 0.05    | 0.9602|
| no_rbi       | RBIs in 1986   | 1  | 0.00054382         | 0.00102        | 0.53    | 0.5947|
| no_bb        | Walks in 1986  | 1  | 0.000292           | 0.00104        | 2.81    | 0.0054|
| yr_major     | Years in MLB   | 1  | 0.03087            | 0.00836        | 3.69    | 0.0003|
| cr_hits      | Career Hits    | 1  | 0.00010384         | 0.00006328     | 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 76.1.2 Fit Diagnostics

Output 76.1.2 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:

- Residual vs Predicted Value
- RStudent vs Predicted Value
- RStudent vs Leverage
- Quantile vs Predicted Value
- log10(Salary) vs Predicted Value
- Cook’s D vs Observation
- Percent vs Residual
- Fit-Mean vs Residual
- Proportion Less vs Prop

Observations 263
Parameters 7
Error DF 256
MSE 0.0631
R-Square 0.5867
Adj R-Square 0.577
Example 76.1: Modeling Salaries of Major League Baseball Players

```sas
proc reg data=baseball
  plots(only label)=(RStudentByLeverage CooksD);
  id name team league;
  model logSalary = no_hits no_runs no_rbi no_bb yr_major cr_hits;
run;
```

Output 76.1.3 and Output 76.1.4 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 76.1.3** 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 6419 for details about reweighting.

```sas
proc reg data=baseball
   plots=(RStudentByLeverage(label) residuals(smooth));
   where name^="Rose, Pete";
   id name team league;
   model logSalary = no_hits no_runs no_rbi no_bb yr_major cr_hits;
run;
```

Output 76.1.5 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 76.1.5  Fit Diagnostics

Output 76.1.6 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 yr_major and cr_hits. 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:

```sas
data baseball;
  set baseball(where=(name^="Rose, Pete");
  yr_major2 = yr_major*yr_major;
  cr_hits2 = cr_hits*cr_hits;
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 = no_hits no_runs no_rbi no_bb yr_major cr_hits 
    yr_major2 cr_hits2;
run;
ods graphics off;
```

Output 76.1.6 Residuals by Regressors

Residual by Regressors for logSalary
With LOESS Smooths

You can address this lack of fit by using polynomials of degree 2 for these two variables as shown in the following statements:
Output 76.1.7 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 76.1.1.

Output 76.1.7  Output from PROC REG

<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>30.69367</td>
<td>3.83671</td>
<td>117.13</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Error</td>
<td>253</td>
<td>8.28706</td>
<td>0.03276</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Corrected Total</td>
<td>261</td>
<td>38.98073</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Root MSE 0.18098  R-Square 0.7874  Dependent Mean 2.57301  Adj R-Sq 0.7807  Coeff Var 7.03393

Parameter Estimates

| Variable     | Label          | DF | Estimate | Standard Error | t Value | Pr > |t| |
|--------------|----------------|----|----------|----------------|---------|------|---|
| Intercept    | Intercept      | 1  | 1.64564  | 0.05030        | 32.72   | <.0001|
| no_hits      | Hits in 1986   | 1  | -0.000005539 | 0.00069200 | -0.08  | 0.9363|
| no_runs      | Runs in 1986   | 1  | 0.00093586 | 0.00125       | 0.75   | 0.4549|
| no_rbi       | RBIs in 1986   | 1  | 0.00187   | 0.00074649    | 2.51   | 0.0127|
| no_bb        | Walks in 1986  | 1  | 0.00218   | 0.00075057    | 2.90   | 0.0040|
| yr_major     | Years in MLB   | 1  | 0.10383   | 0.01495       | 6.94   | <.0001|
| cr_hits      | Career Hits    | 1  | 0.00073955 | 0.00011970   | 6.18   | <.0001|
| yr_major2    |                | 1  | -0.00625  | 0.00071687    | -8.73  | <.0001|
| cr_hits2     |                | 1  | -1.44072E-7 | 4.348471E-8  | -3.31  | 0.0011|

The plots of residuals by regressors in Output 76.1.8 and Output 76.1.9 show that the strong pattern in the plots for cr_majors and cr_hits 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 76.1.8 Residuals by Regressors

Residual by Regressors for logSalary
With LOESS Smooths

Residual by Regressors for logSalary
With LOESS Smooths

Residual by Regressors for logSalary
With LOESS Smooths

Residual by Regressors for logSalary
With LOESS Smooths

Residual by Regressors for logSalary
With LOESS Smooths

Residual by Regressors for logSalary
With LOESS Smooths
Output 76.1.9 Residuals by Regressors

Output 76.1.10 show the diagnostics plots; three of the plots, with points of interest labeled, are shown individually in Output 76.1.11, Output 76.1.12, and Output 76.1.13. 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 76.1.10.
Output 76.1.10  Fit Diagnostics
Output 76.1.11  Outlier and Leverage Diagnostics
Output 76.1.12  Observed by Predicted Values
The RStudent by leverage plot in Output 76.1.11 and the Cook’s $D$ plot in Output 76.1.13 show that there are still a number of influential observations. By specifying the DFFITS and DFBETAS suboptions of the PLOTS= option, you obtain additional influence diagnostics plots shown in Output 76.1.14 and Output 76.1.15. See “Influence Statistics” on page 6409 for details about the interpretation DFFITS and DFBETAS statistics.
Output 76.1.14  DFFITS

Influence Diagnostics for log10(Salary)
Output 76.1.15 DFBETAS

You can continue this analysis by investigating how the influential observations 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.
Chapter 76: The REG Procedure

Example 76.2: 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  54.297 8.65  45  156  168  42  68.15  59.571 8.17  40  166  172  
  38  89.02  49.874 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 8.63  48  170  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  67.25  45.118 11.08  48  172  172  54  91.63  39.203 12.88  44  168  172  
  51  73.71  45.790 12.47  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 RunPulse 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 76.2.1 shows the sequence of models produced by the FORWARD model-selection method.
Output 76.2.1  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

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>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>

Parameter Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>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

Forward Selection: 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>

Parameter Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>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>
Output 76.2.1 continued

Bounds on condition number: 1.0369, 4.1478

<table>
<thead>
<tr>
<th>Forward Selection: Step 3</th>
</tr>
</thead>
</table>

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 Squares</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 Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>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

<table>
<thead>
<tr>
<th>Forward Selection: Step 4</th>
</tr>
</thead>
</table>

Variable MaxPulse Entered: R-Square = 0.8368 and C(p) = 4.8800

Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Squares</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>

Parameter Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>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>
Output 76.2.1 continued

<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>

Parameter Standard Error Type II SS F Value Pr > F

<table>
<thead>
<tr>
<th>Variable</th>
<th>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 76.2.2 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.
### Chapter 76: The REG Procedure

#### Output 76.2.2 Backward Selection Method: PROC REG

**Backward Elimination: Step 0**

All Variables Entered: R-Square = 0.8487 and C(p) = 7.0000

**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>

**Parameter Estimates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>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

**Backward Elimination: Step 1**

Variable RestPulse Removed: R-Square = 0.8480 and C(p) = 5.1063

**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>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>

**Parameter Estimates**

<table>
<thead>
<tr>
<th>Variable</th>
<th>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>
### Output 76.2.2 continued

Bounds on condition number: 8.7312, 104.83
--------------------------------------------------------------------------------
Backward Elimination: Step 2

Variable Weight Removed: R-Square = 0.8368 and C(p) = 4.8800

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>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>

Parameter Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>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

The MAXR method tries to find the “best” one-variable model, the “best” two-variable model, and so on. Output 76.2.3 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.

### Output 76.2.3 Maximum R-Square Improvement Selection Method: PROC REG

Maximum R-Square Improvement: Step 1

Variable RunTime Entered: R-Square = 0.7434 and C(p) = 13.6988

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>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>
Chapter 76: The REG Procedure

Output 76.2.3  continued

<table>
<thead>
<tr>
<th>Parameter</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>3.85530</td>
<td>3443.36654</td>
<td>457.05</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>RunTime</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>

Parameter Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>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

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>
### Output 76.2.3 continued

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>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

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

#### Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Mean Squares</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>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>
</tr>
<tr>
<td>Corrected Total</td>
<td>30</td>
<td>851.38154</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>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>
Output 76.2.3 continued

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

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>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>

Parameter Standard

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimate</th>
<th>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>

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

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>
Example 76.2: Aerobic Fitness Prediction

Example 76.2: Aerobic Fitness Prediction

Output 76.2.3 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>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 76.2.4) and produce the plots shown in Output 76.2.5 and Output 76.2.6.

ods graphics on;

proc reg data=fitness plots=(criteria sbc);
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
       / selection=forward details=summary;
run;

Output 76.2.4 Forward Selection Summary

<table>
<thead>
<tr>
<th>Step</th>
<th>Variable</th>
<th>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>
<td></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>
<td></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>
<td></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>
<td></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>
<td></td>
</tr>
</tbody>
</table>

Output 76.2.5 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 76.2.6 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 76.2.5  Fit Criteria

[Graph showing fit criteria for oxygen with step numbers and best model evaluated star]
Example 76.2: Aerobic Fitness Prediction

Output 76.2.6 SBC Criterion

Next, the RSQUARE model-selection method is used to request $R^2$ and $C_p$ statistics for all possible combinations of the six independent variables. The following statements produce Output 76.2.7:

```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 76.2.7 All Models by the RSQUARE Method: PROC REG

<table>
<thead>
<tr>
<th>Model</th>
<th>Number in Index</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 RunPulse</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>
</tbody>
</table>
### Output 76.2.7 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>Model Number in Index</th>
<th>R-Square</th>
<th>C(p)</th>
<th>Variables in Model</th>
</tr>
</thead>
<tbody>
<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>
<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>0.8138</td>
<td>9.3948</td>
<td>Age Weight RunTime RunPulse RestPulse</td>
</tr>
<tr>
<td>60</td>
<td>5</td>
<td>0.8150</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 RunTime 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 76.2.7 are arranged first by the number of variables in the model and then by the magnitude of $R^2$ for the model.

Output 76.2.8 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 76.2.7.
Output 76.2.8  Fit Criteria

Output 76.2.9 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.
Output 76.2.9 $C_p$ Criterion

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 76.2.10 shows the summary table listing the regressors in the 20 models that yield the smallest $C_p$ values, and Output 76.2.11 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 6331 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 76.2.10  $C_p$ Selection Summary: PROC REG

The REG Procedure  
Model: MODEL1  
Dependent Variable: Oxygen  

$C_p$ Selection Method

<table>
<thead>
<tr>
<th>Model Index</th>
<th>Model Number in Index</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 $R^2$, the statistics are biased and the $p$-values for the parameter estimates are not valid.

**Example 76.3: 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:

```plaintext
*-------------------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
   ... 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 76.3.1 through Output 76.3.4:
   proc reg outest=est1 outsscp=ssscp1 rsquare;
        by sex;
        eq1: model weight=height;
        eq2: model weight=height age;

   proc print data=ssscp1;
        title2 'SSCP type data set';

   proc print data=est1;
        title2 'EST type data set';
   run;
Output 76.3.1  Height and Weight Data: Submodel for Female Children

<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

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|
| Intercept| 1  | -153.12891         | 21.24814       | -7.21   | <.0001|
| height   | 1  | 4.16361            | 0.35052        | 11.88   | <.0001|
Output 76.3.2  Height and Weight Data: Full Model for Female Children

```
--- sex=f -------------------------------------

The REG Procedure
Model: eq2
Dependent Variable: weight

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>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

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|-------------------|----------------|---------|------|---|
| Intercept| 1  | -150.59698        | 20.76730       | -7.25   | <.0001|
| height   | 1  | 3.60378           | 0.40777        | 8.84    | <.0001|
| age      | 1  | 1.90703           | 0.75543        | 2.52    | 0.0130|
Output 76.3.3  Height and Weight Data: Submodel for Male Children

The REG Procedure
Model: eq1
Dependent Variable: weight

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>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>

Root MSE 12.28504  R-Square 0.6245
Dependent Mean 103.44841  Adj R-Sq 0.6215
Coeff Var 11.87552

Parameter Estimates

| 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 76.3.4 Height and Weight Data: Full Model for Male Children

<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>

Root MSE 11.70979
R-Square 0.6616
Dependent Mean 103.44841
Adj R-Sq 0.6561
Coeff Var 11.31945

| 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

$$weight = -150.57 + 3.60 \times height + 1.91 \times age$$

and for males, the full model is

$$weight = -113.71 + 2.68 \times height + 3.08 \times age$$

The OUTSSCP= data set is shown in Output 76.3.5. 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 76.3.5  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</td>
<td>f</td>
<td>SSCP</td>
<td>Intercept</td>
<td>111.0</td>
<td>6718.40</td>
<td>10975.50</td>
<td>1824.90</td>
</tr>
<tr>
<td>2</td>
<td>f</td>
<td>SSCP</td>
<td>height</td>
<td>6718.4</td>
<td>407879.32</td>
<td>669469.85</td>
<td>110818.32</td>
</tr>
<tr>
<td>3</td>
<td>f</td>
<td>SSCP</td>
<td>weight</td>
<td>10975.5</td>
<td>669469.85</td>
<td>1123360.75</td>
<td>182444.95</td>
</tr>
<tr>
<td>4</td>
<td>f</td>
<td>SSCP</td>
<td>age</td>
<td>1824.9</td>
<td>110818.32</td>
<td>182444.95</td>
<td>30363.81</td>
</tr>
<tr>
<td>5</td>
<td>f</td>
<td>N</td>
<td></td>
<td>111.0</td>
<td>111.00</td>
<td>111.00</td>
<td>111.00</td>
</tr>
<tr>
<td>6</td>
<td>m</td>
<td>SSCP</td>
<td>Intercept</td>
<td>126.0</td>
<td>7825.00</td>
<td>13034.50</td>
<td>2072.10</td>
</tr>
<tr>
<td>7</td>
<td>m</td>
<td>SSCP</td>
<td>height</td>
<td>7825.0</td>
<td>488243.60</td>
<td>817919.60</td>
<td>129432.57</td>
</tr>
<tr>
<td>8</td>
<td>m</td>
<td>SSCP</td>
<td>weight</td>
<td>13034.5</td>
<td>817919.60</td>
<td>1398238.75</td>
<td>217717.45</td>
</tr>
<tr>
<td>9</td>
<td>m</td>
<td>SSCP</td>
<td>age</td>
<td>2072.1</td>
<td>129432.57</td>
<td>217717.45</td>
<td>34515.95</td>
</tr>
<tr>
<td>10</td>
<td>m</td>
<td>N</td>
<td></td>
<td>126.0</td>
<td>126.00</td>
<td>126.00</td>
<td>126.00</td>
</tr>
</tbody>
</table>

The OUTEST= data set is displayed in Output 76.3.6; 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 $R^2$ 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 76.3.6  OUTEST Data Set

<table>
<thead>
<tr>
<th>Eq</th>
<th>sex</th>
<th><em>MODEL</em></th>
<th>PARMS</th>
<th>weight</th>
<th>12.3461</th>
<th>-153.129</th>
<th>4.16361</th>
<th>-1</th>
<th>1 2 109 0.56416</th>
</tr>
</thead>
<tbody>
<tr>
<td>1f</td>
<td>eq1</td>
<td>PARMS</td>
<td>weight</td>
<td>12.0527</td>
<td>-150.597</td>
<td>3.60378</td>
<td>1.90703</td>
<td>2 3 108 0.58845</td>
<td></td>
</tr>
<tr>
<td>2f</td>
<td>eq2</td>
<td>PARMS</td>
<td>weight</td>
<td>12.2850</td>
<td>-125.698</td>
<td>3.68977</td>
<td>1 124 0.62451</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3m</td>
<td>eq1</td>
<td>PARMS</td>
<td>weight</td>
<td>11.7098</td>
<td>-113.713</td>
<td>2.68075</td>
<td>3.08167</td>
<td>2 3 123 0.66161</td>
<td></td>
</tr>
</tbody>
</table>
Example 76.4: 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, $X_1_i$ is a quantitative variable, $X_2_i$ is a qualitative variable taking on values 0 or 1, and $X_1_iX_2_i$ is the interaction. The variable $X_2_i$ is called a dummy, binary, or indicator variable. With values 0 or 1, it distinguishes between two populations. The model is of the form

$$Y_i = \beta_0 + \beta_1X_1_i + \beta_2X_2_i + \beta_3X_1_iX_2_i + \epsilon_i$$

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 $X_1_i$ is the size of the firm, measured in total assets of the firm. The variable $X_2_i$ 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 $X_2_i$. If $X_2_i = 0$, the model is

$$Y_i = \beta_0 + \beta_1X_1_i + \epsilon_i$$

This is the model for a mutual company. If $X_2_i = 1$, the model for a stock firm is

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1_i + \epsilon_i$$
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*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
;
run;
```

The following statements begin the analysis:

```plaintext
proc reg data=insurance;
  model time = size type sizetype;
run;
```

The ANOVA table is displayed in Output 76.4.1.

**Output 76.4.1** 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>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>

Root MSE: 3.32021
Dependent Mean: 19.40000
Coefficient of Variation: 17.11450

---

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:

```
   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 76.4.2.

### Output 76.4.2 ANOVA Table and Parameter Estimates

#### 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>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

#### Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > |t| |
|----------|----|--------------------|----------------|---------|-------|-----|
| Intercept| 1  | 33.87407           | 1.81386        | 18.68   | <.0001 |
| size     | 1  | -0.10174           | 0.00889        | -11.44  | <.0001 |
| type     | 1  | 8.05547            | 1.45911        | 5.52    | <.0001 |

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 $R^2$ did not change with the omission of the interaction term.
The fitted model is

\[
\text{time} = 33.87 - 0.102 \times \text{size} + 8.055 \times \text{type}
\]

The fitted model for a mutual fund company \((X_2 = 0)\) is

\[
\text{time} = 33.87 - 0.102 \times \text{size}
\]

and the fitted model for a stock company \((X_2 = 1)\) is

\[
\text{time} = (33.87 + 8.055) - 0.102 \times \text{size}
\]

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 76.4.3, 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.

```sas
output out=out r=r p=p;
run;

proc sgplot data=out;
  scatter x=p y=r / markerchar=type group=type;
run;
```
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 76.4.4, 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 76.4.4  Plot of Predicted vs. Size

The different intercepts are very evident in this plot.

Example 76.5: 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:

\begin{verbatim}
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';
\end{verbatim}
Chapter 76: The REG Procedure

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 76.5.1.

Output 76.5.1 Ridge Regression and VIF Traces
The OUTVIF option outputs the variance inflation factors to the OUTEST= data set that is shown in Output 76.5.2.

**Output 76.5.2** OUTEST Data Set Showing VIF Values

<table>
<thead>
<tr>
<th>Obs</th>
<th>MODEL_</th>
<th><em>TYPE</em></th>
<th><em>DEPVAR</em></th>
<th>RIDGE_</th>
<th>PCOMIT_</th>
<th>RMSE_</th>
<th>Intercept</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>MODEL1</td>
<td>PARMS</td>
<td>x4</td>
<td>.</td>
<td>.</td>
<td>1.15596</td>
<td>390.538</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>
</tr>
<tr>
<td>3</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.000</td>
<td>.</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>
</tr>
<tr>
<td>5</td>
<td>MODEL1</td>
<td>RIDGEVIF</td>
<td>x4</td>
<td>0.002</td>
<td>.</td>
<td>2.69721</td>
<td>-103.388</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>
</tr>
<tr>
<td>7</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.004</td>
<td>.</td>
<td>3.22340</td>
<td>-93.797</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>
</tr>
<tr>
<td>9</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.006</td>
<td>.</td>
<td>3.47752</td>
<td>-87.687</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>
</tr>
<tr>
<td>11</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.008</td>
<td>.</td>
<td>3.62677</td>
<td>-83.593</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>
</tr>
<tr>
<td>13</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.010</td>
<td>.</td>
<td>3.72505</td>
<td>-80.603</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>
</tr>
<tr>
<td>15</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.012</td>
<td>.</td>
<td>3.79477</td>
<td>-78.276</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>
</tr>
<tr>
<td>17</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.014</td>
<td>.</td>
<td>3.84693</td>
<td>-76.381</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>
</tr>
<tr>
<td>19</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.016</td>
<td>.</td>
<td>3.88750</td>
<td>-74.785</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>
</tr>
<tr>
<td>21</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.018</td>
<td>.</td>
<td>3.92004</td>
<td>-73.407</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>
</tr>
<tr>
<td>23</td>
<td>MODEL1</td>
<td>RIDGE</td>
<td>x4</td>
<td>0.020</td>
<td>.</td>
<td>3.94679</td>
<td>-72.193</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Obs</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>0.78</td>
<td>10.174</td>
<td>-121.626</td>
<td>-0.008</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>2</td>
<td>7682.37</td>
<td>320.022</td>
<td>53.525</td>
<td>344.545</td>
<td>6643.32</td>
<td>-1</td>
</tr>
<tr>
<td>3</td>
<td>-0.78</td>
<td>10.174</td>
<td>-121.626</td>
<td>-0.008</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>4</td>
<td>11.18</td>
<td>58.731</td>
<td>10.744</td>
<td>63.208</td>
<td>11.22</td>
<td>-1</td>
</tr>
<tr>
<td>5</td>
<td>0.05</td>
<td>4.404</td>
<td>-9.065</td>
<td>-0.003</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>6</td>
<td>4.36</td>
<td>23.939</td>
<td>9.996</td>
<td>25.744</td>
<td>5.15</td>
<td>-1</td>
</tr>
<tr>
<td>7</td>
<td>0.06</td>
<td>2.839</td>
<td>-21.338</td>
<td>-0.002</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>8</td>
<td>2.93</td>
<td>13.011</td>
<td>9.383</td>
<td>13.976</td>
<td>3.81</td>
<td>-1</td>
</tr>
<tr>
<td>9</td>
<td>0.06</td>
<td>2.110</td>
<td>-28.447</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>10</td>
<td>2.36</td>
<td>8.224</td>
<td>8.838</td>
<td>8.821</td>
<td>3.23</td>
<td>-1</td>
</tr>
<tr>
<td>11</td>
<td>0.06</td>
<td>1.689</td>
<td>-33.377</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>12</td>
<td>2.04</td>
<td>5.709</td>
<td>8.343</td>
<td>6.112</td>
<td>2.89</td>
<td>-1</td>
</tr>
<tr>
<td>13</td>
<td>0.06</td>
<td>1.414</td>
<td>-37.177</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>14</td>
<td>1.84</td>
<td>4.226</td>
<td>7.891</td>
<td>4.514</td>
<td>2.65</td>
<td>-1</td>
</tr>
<tr>
<td>15</td>
<td>0.06</td>
<td>1.221</td>
<td>-40.297</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>16</td>
<td>1.69</td>
<td>3.279</td>
<td>7.476</td>
<td>3.493</td>
<td>2.46</td>
<td>-1</td>
</tr>
<tr>
<td>17</td>
<td>0.06</td>
<td>1.078</td>
<td>-42.965</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>18</td>
<td>1.57</td>
<td>2.637</td>
<td>7.094</td>
<td>2.801</td>
<td>2.31</td>
<td>-1</td>
</tr>
<tr>
<td>19</td>
<td>0.06</td>
<td>0.968</td>
<td>-45.309</td>
<td>-0.001</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>20</td>
<td>1.47</td>
<td>2.182</td>
<td>6.741</td>
<td>2.310</td>
<td>2.18</td>
<td>-1</td>
</tr>
<tr>
<td>21</td>
<td>0.06</td>
<td>0.880</td>
<td>-47.407</td>
<td>-0.000</td>
<td>0.00</td>
<td>-1</td>
</tr>
<tr>
<td>22</td>
<td>1.39</td>
<td>1.847</td>
<td>6.415</td>
<td>1.949</td>
<td>2.06</td>
<td>-1</td>
</tr>
<tr>
<td>23</td>
<td>0.06</td>
<td>0.809</td>
<td>-49.310</td>
<td>-0.000</td>
<td>0.00</td>
<td>-1</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 76.5.1 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.

```latex
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 76.5.3 and Output 76.5.4.

**Output 76.5.3** VIF Traces
Example 76.6: 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 6426 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 76.6.1 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 76.6.1 Analysis of Variance**

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

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>

Root MSE = 2.43923, R-Square = 0.9652, Dependent Mean = 41.65833, Adj R-Sq = 0.9362, Coeff Var = 5.85533
```
References


Subject Index

A
adjusted R² selection (REG), 6395
alpha level
REG procedure, 6326
annotate
global data set (REG), 6326
local data set (REG), 6364
autocorrelation
REG procedure, 6431

B
backward elimination
REG procedure, 6307, 6394

C
collinearity
REG procedure, 6405
correlation
matrix (REG), 6326
covariance matrix
REG procedure, 6326
COVRATIO statistic, 6410
crossproducts matrix
REG procedure, 6433

d
delete variables (REG), 6339
deleting observations
REG procedure, 6419
DFBETAS statistic (REG), 6411
DFFITS statistic
REG procedure, 6410
diagnostic statistics
REG procedure, 6407, 6408

F
fit diagnostics
examples (REG), 6441
forward selection
REG procedure, 6307, 6393

G
graphics
keywords (REG), 6361
options (REG), 6362
traditional plots (REG), 6360

H
hat matrix, 6409
heteroscedasticity
testing (REG), 6425
hypothesis tests
multivariate (REG), 6427
REG procedure, 6351, 6376

I
incomplete principal components
REG procedure, 6328, 6348
influence diagnostics
examples (REG), 6441
influence statistics
REG procedure, 6409
IPC analysis
REG procedure, 6328, 6348, 6432

L
lack of fit
examples (REG), 6491
lack-of-fit
testing (REG), 6426
line printer plots
REG procedure, 6368
LMSELECT procedure
ODS Graphics, 6328

M
Mallows’ Cₚ selection
REG procedure, 6395
model
fit summary (REG), 6407
model building
examples (REG), 6441
model selection
examples (REG), 6458
REG procedure, 6307, 6393, 6396, 6397
multicollinearity
REG procedure, 6405
multivariate tests
   REG procedure, 6427

N
non-full-rank models
   REG procedure, 6403

O
ODS graph names
   REG procedure, 6440
ODS GRAPHICS
   examples (REG), 6441
ODS Graphics
   LMSELECT procedure, 6328

P
P-P plots
   REG procedure, 6432
parameter estimates
   example (REG), 6397
   REG procedure, 6434
partial regression leverage plots
   REG procedure, 6417
plots
   keywords (REG), 6361
   line printer (REG), 6368
   options (REG), 6362, 6363
   traditional (REG), 6360
polynomial regression
   REG procedure, 6312
predicted values
   REG procedure, 6396, 6400
   prediction
   example (REG), 6458

Q
Q-Q plots
   REG procedure, 6432
qualitative variables
   REG procedure, 6482

R
R^2 improvement
   REG procedure, 6394, 6395
R^2 selection
   REG procedure, 6395
refitting models
   REG procedure, 6420
REG procedure
   adding variables, 6338
   adjusted R^2 selection, 6395
   alpha level, 6326
   annotations, 6326, 6364
   ANOVA table, 6434
   autocorrelation, 6431
   backward elimination, 6307, 6394
   collinearity, 6405
   computational methods, 6433
   correlation matrix, 6326
   covariance matrix, 6326
   crossproducts matrix, 6433
   delete variables, 6339
   deleting observations, 6419
   diagnostic statistics, 6407, 6408
   dictionary of options, 6363
   fit diagnostics, 6441
   forward selection, 6307, 6393
   graphics keywords and options, 6361, 6362
   graphics plots, traditional, 6360
   heteroscedasticity, testing, 6425
   hypothesis tests, 6351, 6376
   incomplete principal components, 6328, 6348
   influence diagnostics, 6411
   influence statistics, 6409
   input data sets, 6378
   interactive analysis, 6322, 6389
   introductory example, 6308
   IPC analysis, 6328, 6348, 6432
   lack of fit, 6491
   lack-of-fit, testing, 6426
   line printer plots, 6368
   Mallows' C_p selection, 6395
   missing values, 6378
   model building, 6441
   model fit summary statistics, 6407
   model selection, 6307, 6393, 6396, 6397, 6458
   multicollinearity, 6405
   multivariate tests, 6427
   new regressors, 6378
   non-full-rank models, 6403
   ODS graph names, 6440
   ODS GRAPHICS, 6441
   ODS table names, 6436
   output data sets, 6382, 6388
   P-P plots, 6432
   parameter estimates, 6397, 6434
   partial regression leverage plots, 6417
   plot keywords and options, 6361–6363
   plots, traditional, 6360
   polynomial regression, 6312
   predicted values, 6396, 6400, 6458
   Q-Q plots, 6432
   qualitative variables, 6482
R² improvement, 6394, 6395
R² selection, 6395
refitting models, 6420
residual values, 6400
restoring weights, 6421
reweighting observations, 6419
ridge regression, 6337, 6349, 6367, 6432, 6487
singularities, 6433
stepwise selection, 6307, 6394
summary statistics, 6407
sweep algorithm, 6433
time series data, 6431
variance inflation factors (VIF), 6328

regression analysis (REG), 6306
residuals
  REG procedure, 6400
restoring weights
  REG procedure, 6421
reweighting observations
  REG procedure, 6419
ridge regression
  REG procedure, 6337, 6349, 6367, 6432, 6487

S

singularities
  REG procedure, 6433
stepwise selection
  REG procedure, 6307, 6394
studentized residual, 6410
summary statistics
  REG procedure, 6407
sweep algorithm
  REG procedure, 6433

T
time series data
  REG procedure, 6431

V

variance inflation factors (VIF)
  REG procedure, 6328
VIF, see variance inflation factors
Syntax Index

A

ACOV option
  MODEL statement (REG), 6343
ACOVMETHOD= option
  MODEL statement (REG), 6343
ADD statement, REG procedure, 6338
ADJRSQ option
  MODEL statement (REG), 6343
AIC option
  MODEL statement (REG), 6343
  PLOT statement (REG), 6363
ALL option
  MODEL statement (REG), 6343
  PROC REG statement, 6326
ALLOBS option
  PAINT statement (REG), 6357
  REWEIGHT statement (REG), 6374
ALPHA= option
  MODEL statement (REG), 6343
  PROC REG statement, 6326
ANNOTATE= option
  PLOT statement (REG), 6364
  PROC REG statement, 6326
B

B option
  MODEL statement (REG), 6344
BEST= option
  MODEL statement (REG), 6344
BIC option
  MODEL statement (REG), 6344
  PLOT statement (REG), 6364
BY statement
  REG procedure, 6339
C

CANPRINT option
  MTEST statement (REG), 6352
CAXIS= option
  PLOT statement (REG), 6364
CFRAME= option
  PLOT statement (REG), 6364
CHOCKING= option
  PLOT statement (REG), 6364
CHREF= option
  PLOT statement (REG), 6364
CLB option
  MODEL statement (REG), 6344
CLEAR option
  PLOT statement (REG), 6368
CLI option
  MODEL statement (REG), 6344
CLINE= option
  PLOT statement (REG), 6364
CLM option
  MODEL statement (REG), 6344
CMALLOWS= option
  PLOT statement (REG), 6364
COLLECT option
  PLOT statement (REG), 6368
COLLIN option
  MODEL statement (REG), 6345
COLLNOINT option
  MODEL statement (REG), 6345
CONF option
  PLOT statement (REG), 6364
CORR option
  PROC REG statement, 6326
CORRB option
  MODEL statement (REG), 6345
COVB option
  MODEL statement (REG), 6345
COVOUT option
  PROC REG statement, 6326
CP option
  MODEL statement (REG), 6345
  PLOT statement (REG), 6365
CTEXT= option
  PLOT statement (REG), 6365
CVREF= option
  PLOT statement (REG), 6365
DATA= option
  PROC REG statement, 6326
DELETE statement, REG procedure, 6339
DESCRIPTION= option
  PLOT statement (REG), 6365
DETAILS option
  MODEL statement (REG), 6345
  MTEST statement (REG), 6352
DW option
MODEL statement (REG), 6345
DWPROB option
MODEL statement (REG), 6345

E

EDF option
MODEL statement (REG), 6345
PLOT statement (REG), 6365
PROC REG statement, 6327

F

FREQ statement
REG procedure, 6339

G

GMSEP option
MODEL statement (REG), 6345
PLOT statement (REG), 6365
GOUT= option
PROC REG statement, 6327
GROUPNAMES= option
MODEL statement (REG), 6346

H

HAXIS= option
MODEL statement (REG), 6346
PLOT statement (REG), 6365
HCC option
MODEL statement (REG), 6346
HCCMETHOD= option
MODEL statement (REG), 6346
HPLOTS= option
PLOT statement (REG), 6369
HREF= option
PLOT statement (REG), 6365

I

I option
MODEL statement (REG), 6346
ID statement
REG procedure, 6340
IN option
PLOT statement (REG), 6365
INCLUDE= option
MODEL statement (REG), 6346
INFLUENCE option
MODEL statement (REG), 6346

J

JP option

K

keyword= option
OUTPUT statement (REG), 6353

L

LACKFIT option
MODEL statement (REG), 6347
LEGEND= option
PLOT statement (REG), 6365
LHREF= option
PLOT statement (REG), 6366
LINEPRINTER option
PROC REG statement, 6327
LLINE= option
PLOT statement (REG), 6366
LVREF= option
PLOT statement (REG), 6366

M

MAXSTEP option
MODEL statement (REG), 6347
MODEL statement
REG procedure, 6340
MODELFONT option
PLOT statement (REG), 6366
MODELHT option
PLOT statement (REG), 6366
MODELLAB option
PLOT statement (REG), 6366
MSEP option
MODEL statement (REG), 6347
PLOT statement (REG), 6366
MSTAT= option
MTEST statement (REG), 6353
MTEST statement
REG procedure, 6351

N

NAME= option
PLOT statement (REG), 6366
NOCOLLECT option
PLOT statement (REG), 6369
NOINT option
MODEL statement (REG), 6347
NOLEGEND option
PLOT statement (REG), 6366
NOLINE option
PLOT statement (REG), 6366
NOLIST option
  PAINT statement (REG), 6357
  REWEIGHT statement (REG), 6375
NOMODEL option
  PLOT statement (REG), 6366
NOPRINT option
  MODEL statement (REG), 6347
  PROC REG statement, 6328
NOSTAT option
  PLOT statement (REG), 6366
NP option
  PLOT statement (REG), 6367
OUT= option
  OUTPUT statement (REG), 6353
OUTEST= option
  PROC REG statement, 6327
OUTPUT statement
  REG procedure, 6353
OUTSEB option
  MODEL statement (REG), 6347
  PROC REG statement, 6327
OUTSSCP= option
  PROC REG statement, 6327
OUTSTB option
  MODEL statement (REG), 6347
  PROC REG statement, 6328
OUTVIF option
  MODEL statement (REG), 6347
  PROC REG statement, 6328
OVERLAY option
  PLOT statement (REG), 6367, 6369
P option
  MODEL statement (REG), 6347
  REG procedure, 6355
PAINT statement
  REG procedure, 6355
PARTIAL option
  MODEL statement (REG), 6348
PARTIALDATA option
  MODEL statement (REG), 6348
PARTIALR2 option
  MODEL statement (REG), 6348
PC option
  MODEL statement (REG), 6348
  PLOT statement (REG), 6367
PCOMIT= option
  MODEL statement (REG), 6348
  PROC REG statement, 6328
PCORR1 option
  MODEL statement (REG), 6348
PCORR2 option
  MODEL statement (REG), 6348
PLOT option
  PROC REG statement, 6328
PLOT statement
  REG procedure, 6358
PLOTS option
  PROC REG statement, 6328
PRED option
  PLOT statement (REG), 6367
PRESS option
  MODEL statement (REG), 6348
  PROC REG statement, 6337
PRINT option
  MTEST statement (REG), 6353
  TEST statement (REG), 6377
PRINT statement, REG procedure, 6370
PROC REG statement, see REG procedure
R option
  MODEL statement (REG), 6348
  REG procedure, 6338
  REG procedure, ADD statement, 6338
  REG procedure, BY statement, 6339
  REG procedure, DELETE statement, 6339
  REG procedure, FREQ statement, 6339
  REG procedure, ID statement, 6340
  REG procedure, MODEL statement, 6340
ACOV option, 6343
ACOVMETHOD= option, 6343
ADJRSQ option, 6343
ALPHA= option, 6343
AIC option, 6343
ALL option, 6343
ALPHA= option, 6343
B option, 6344
BEST= option, 6344
BIC option, 6344
CLB option, 6344
CLI option, 6344
CLM option, 6344
COLLIN option, 6345
COLLINOINT option, 6345
CORRB option, 6345
COVB option, 6345
CP option, 6345
DETAILS option, 6345
DW option, 6345
DWPROB option, 6345
EDF option, 6345
GMSEP option, 6345
GROUPNAMES= option, 6346
HCC option, 6346
HCCMETHOD= option, 6346
I option, 6346
INCLUDE= option, 6346
INFLUENCE option, 6346
JP option, 6347
LACKFIT option, 6347
MAXSTEP option, 6347
MSE option, 6347
Noint option, 6347
Noprint option, 6347
OUTSEB option, 6347
OUTSTB option, 6347
OUTVIF option, 6347
P option, 6347
PARTIAL option, 6348
PARTIALDATA option, 6348
PARTIALR2 option, 6348
PC option, 6348
PCOMIT= option, 6348
PCORR1 option, 6348
PCORR2 option, 6348
PRESS option, 6348
R option, 6348
RIDGE= option, 6349
RMSE option, 6349
RSQUARE option, 6349
SBC option, 6349
SCORR1 option, 6349
SCORR2 option, 6349
SELECTION= option, 6307, 6350
SEQB option, 6350
SIGMA= option, 6350
SINGULAR= option, 6350
SLENTRY= option, 6350
SLSTAY= option, 6350
SP option, 6350
SPEC option, 6350
SS1 option, 6350
SS2 option, 6350
SSE option, 6350
START= option, 6351
STB option, 6351
STOP= option, 6351
TOL option, 6351
VIF option, 6351
WHITE option, 6351
Xpx option, 6351
REG procedure, MTEST statement, 6351
CANPRINT option, 6352
DETAILS option, 6352
MSTAT= option, 6353
PRINT option, 6353
REG procedure, OUTPUT statement, 6353
keyword= option, 6353
OUT= option, 6353
REG procedure, PAINT statement, 6355
ALLOBS option, 6357
Nolist option, 6357
RESET option, 6357
STATUS option, 6358
SYMBOL= option, 6357
UNDO option, 6358
REG procedure, PLOT statement, 6358
AIC option, 6363
ANNOTATE= option, 6364
BIC option, 6364
CAXIS= option, 6364
CFRAME= option, 6364
CHOCKING= option, 6364
CHREF= option, 6364
CLEAR option, 6368
CLINE= option, 6364
CMALLOWS= option, 6364
COLLECT option, 6368
CONF option, 6364
CP option, 6365
CTEXT= option, 6365
CVREF= option, 6365
DESCRIPTION= option, 6365
EDF option, 6365
GMSEP option, 6365
HAXIS= option, 6365
HPLOTS= option, 6369
HREF= option, 6365
IN option, 6365
JP option, 6365
LEGEND= option, 6365
LHREF= option, 6366
LLINE= option, 6366
LVREF= option, 6366
MODELFONT option, 6366
MODELHT option, 6366
MODELLAB option, 6366
MSE option, 6366
NAME= option, 6366
NOCOLLECT option, 6369
NOLENGEN option, 6366
NOLINE option, 6366
NOMODEL option, 6366
NOSTAT option, 6366
NP option, 6367
OVERLAY option, 6367, 6369
PC option, 6367
PRED option, 6367
RIDGEPLOT option, 6367
SBC option, 6367
SP option, 6367
SSE option, 6367
STATFONT option, 6367
STATHT option, 6367
summary of options, 6361, 6362
SYMBOL= option, 6369
USEALL option, 6367
VAXIS= option, 6367
VPLOTS= option, 6370
VREF= option, 6367
REG procedure, PRINT statement, 6370
REG procedure, PROC REG statement, 6325
    ALL option, 6326
    ALPHA= option, 6326
    ANNOTATE= option, 6326
    CORR option, 6326
    COVOUT option, 6326
    DATA= option, 6326
    EDF option, 6327
    GOUT= option, 6327
    LINEPRINTER option, 6327
    NOPRINT option, 6327
    OUTEST= option, 6327
    OUTSEB option, 6327
    OUTSSCP= option, 6327
    OUTSTB option, 6327
    OUTVIF option, 6328
    PCOMIT= option, 6328
    PLOT option, 6328
    PLOTS option, 6328
    PRESS option, 6337
    RIDGE= option, 6337
    RSQUARE option, 6338
    SIMPLE option, 6338
    SINGULAR= option, 6338
    TABLEOUT option, 6338
    USSCP option, 6338
REG procedure, REFIT statement, 6371
REG procedure, RESTRICT statement, 6371
REG procedure, REWEIGHT statement, 6373
    ALLOBS option, 6374
    NOLIST option, 6375
    RESET option, 6375
    STATUS option, 6376
    UNDO option, 6376
    WEIGHT= option, 6375
REG procedure, TEST statement, 6376
PRINT option, 6377
REG procedure, VAR statement, 6377
REG procedure, WEIGHT statement, 6377
RESET option
    PAINT statement (REG), 6357
REG procedure, PRINT statement, 6370
REG procedure, PROC REG statement, 6325
    ALL option, 6326
    ALPHA= option, 6326
    ANNOTATE= option, 6326
    CORR option, 6326
    COVOUT option, 6326
    DATA= option, 6326
    EDF option, 6327
    GOUT= option, 6327
    LINEPRINTER option, 6327
    NOPRINT option, 6327
    OUTEST= option, 6327
    OUTSEB option, 6327
    OUTSSCP= option, 6327
    OUTSTB option, 6327
    OUTVIF option, 6328
    PCOMIT= option, 6328
    PLOT option, 6328
    PLOTS option, 6328
    PRESS option, 6337
    RIDGE= option, 6337
    RSQUARE option, 6338
    SIMPLE option, 6338
    SINGULAR= option, 6338
    TABLEOUT option, 6338
    USSCP option, 6338
REG procedure, REFIT statement, 6371
REG procedure, RESTRICT statement, 6371
REG procedure, REWEIGHT statement, 6373
    ALLOBS option, 6374
    NOLIST option, 6375
    RESET option, 6375
    STATUS option, 6376
    UNDO option, 6376
    WEIGHT= option, 6375
REG procedure, TEST statement, 6376
PRINT option, 6377
REG procedure, VAR statement, 6377
REG procedure, WEIGHT statement, 6377
RESET option
    PAINT statement (REG), 6357
REWEIGHT statement (REG), 6375
RESTRICT statement
    REG procedure, 6371
REG procedure, REWEIGHT statement, 6373
RIDGEPLOT option
    MODEL statement (REG), 6349
    PROC REG statement, 6337
REG procedure, PRINT statement, 6370
    PLOT statement (REG), 6367
SCORR1 option
    MODEL statement (REG), 6349
SCORR2 option
    MODEL statement (REG), 6349
SELECTION= option
    MODEL statement (REG), 6350
REG procedure, MODEL statement, 6307
SEQB option
    MODEL statement (REG), 6350
SIGMA= option
    MODEL statement (REG), 6350
SIMPLE option
    PROC REG statement, 6338
SINGULAR= option
    MODEL statement (REG), 6350
PROC REG statement, 6338
SLENTRY= option
    MODEL statement (REG), 6350
SLSTAY= option
    MODEL statement (REG), 6350
SPEC option
    MODEL statement (REG), 6350
SS1 option
    MODEL statement (REG), 6350
SS2 option
    MODEL statement (REG), 6350
SSE option
    MODEL statement (REG), 6350
DIFF option
    MODEL statement (REG), 6351

STATFONT option
   PLOT statement (REG), 6367

STATHT option
   PLOT statement (REG), 6367

STATUS option
   PAINT statement (REG), 6358
   REWEIGHT statement (REG), 6376

STB option
   MODEL statement (REG), 6351

STOP= option
   MODEL statement (REG), 6351

SYMBOL= option
   PAINT statement (REG), 6357
   PLOT statement (REG), 6369

T

TABLEOUT option
   PROC REG statement, 6338

TEST statement
   REG procedure, 6376

TOL option
   MODEL statement (REG), 6351

U

UNDO option
   PAINT statement (REG), 6358
   REWEIGHT statement (REG), 6376

USEALL option
   PLOT statement (REG), 6367

USSCP option
   PROC REG statement, 6338

V

VAR statement
   REG procedure, 6377

VAXIS= option
   PLOT statement (REG), 6367

VIF option
   MODEL statement (REG), 6351

VPLOTS= option
   PLOT statement (REG), 6370

VREF= option
   PLOT statement (REG), 6367

W

WEIGHT statement
   REG procedure, 6377

WEIGHT= option
   REWEIGHT statement (REG), 6375

WHITE option
We welcome your feedback.

- If you have comments about this book, please send them to yourturn@sas.com. Include the full title and page numbers (if applicable).
- If you have comments about the software, please send them to suggest@sas.com.
Whether you are new to the work force or an experienced professional, you need to distinguish yourself in this rapidly changing and competitive job market. SAS® Publishing provides you with a wide range of resources to help you set yourself apart. Visit us online at support.sas.com/bookstore.

**SAS® Press**

Need to learn the basics? Struggling with a programming problem? You’ll find the expert answers that you need in example-rich books from SAS Press. Written by experienced SAS professionals from around the world, SAS Press books deliver real-world insights on a broad range of topics for all skill levels.

[support.sas.com/saspress](support.sas.com/saspress)

**SAS® Documentation**

To successfully implement applications using SAS software, companies in every industry and on every continent all turn to the one source for accurate, timely, and reliable information: SAS documentation. We currently produce the following types of reference documentation to improve your work experience:

- Online help that is built into the software.
- Tutorials that are integrated into the product.
- Reference documentation delivered in HTML and PDF – free on the Web.
- Hard-copy books.

[support.sas.com/publishing](support.sas.com/publishing)

**SAS® Publishing News**

Subscribe to SAS Publishing News to receive up-to-date information about all new SAS titles, author podcasts, and new Web site features via e-mail. Complete instructions on how to subscribe, as well as access to past issues, are available at our Web site.

[support.sas.com/spn](support.sas.com/spn)