Chapter 55
The REG Procedure

Chapter Table of Contents

Chapter 55
The REG Procedure

Overview

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 3, "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
- plots
  - plot model fit summary statistics and diagnostic statistics
  - produce normal quantile-quantile (Q-Q) and probability-probability (P-P) plots for statistics such as residuals
  - specify special shorthand options to plot ridge traces, confidence intervals, and prediction intervals
  - display the fitted model equation, summary statistics, and reference lines on the plot
  - control the graphics appearance with PLOT statement options and with global graphics statements including the TITLE, FOOTNOTE, NOTE, SYMBOL, and LEGEND statements

- "paint" or highlight line-printer scatter plots
- produce partial regression leverage line-printer plots

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 "Model-Selection Methods" section on page 2947.

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.

Getting Started

Simple Linear Regression

Suppose that a response variable Y can be predicted by a linear function of a regressor variable X. You can estimate β_0, the intercept, and β_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.

```
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.

```
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

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

The variable Weight is the response or dependent variable in this equation, and β_0 and β_1 are the unknown parameters to be estimated. The variable Height is the regressor or independent variable, and ϵ is the unknown error. The following commands invoke the REG procedure and fit this model to the data.

```
proc reg;
   model Weight = Height;
run;
```

Figure 55.1 includes some information concerning model fit.

Simple Linear Regression

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.24912	7193.24912	57.08	<.0001
Error	17	2142.48772	126.02869		
Corrected Total	18	9335.73684			

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

Figure 55.1. ANOVA Table

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, β_0 and β_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.

The "Parameter Estimates" table shown in Figure 55.2 contains the estimates of β_0 and β_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 ($t = -4.432$, $p = 0.0004$ and $t = 7.555$, $p < 0.0001$) indicate that the intercept and Height parameter estimates, respectively, are highly significant.

Simple Linear Regression

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Parameter Estimates

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

Figure 55.2. Parameter Estimates

From the parameter estimates, the fitted model is

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

The REG procedure can be used interactively. After you specify a model with the MODEL statement and submit the PROC REG statements, you can submit further statements without reinvoking the procedure. The following command can now be issued to request a plot of the residual versus the predicted values, as shown in Figure 55.3.

```
plot r.*p.;
run;
```

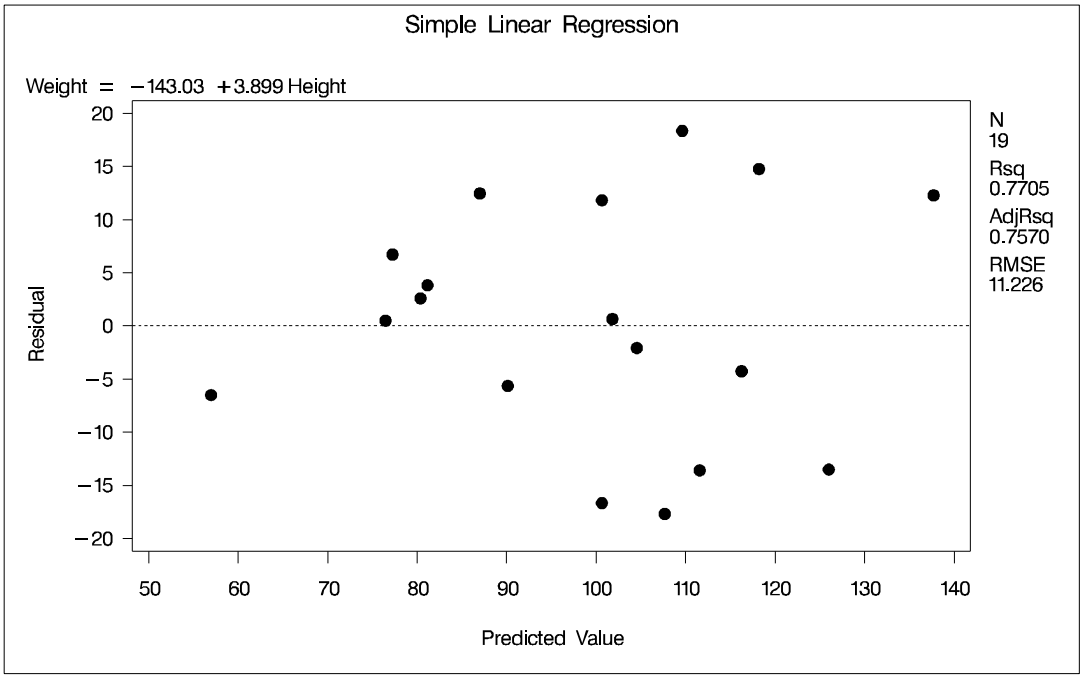


Figure 55.3. Plot of Residual vs. Predicted Values

A trend in the residuals would indicate nonconstant variance in the data. Figure 55.3 may indicate a slight trend in the residuals; they appear to increase slightly as the predicted values increase. A fan-shaped trend may indicate the need for a variance-stabilizing transformation. A curved trend (such as a semi-circle) may indicate the need for a quadratic term in the model. Since these residuals have no apparent trend, the analysis is considered to be acceptable.

Polynomial Regression

Consider a response variable Y that can be predicted by a polynomial function of a regressor variable X. You can estimate β_0, the intercept, β_1, the slope due to X, and β_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 1970 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:

```
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
;
```

The following statements begin the analysis. (Influence diagnostics and autocorrelation information for the full model are shown in Figure 55.43 on page 2972 and Figure 55.56 on page 2987.)

```
symbol1 c=blue;
proc reg data=USPopulation;
   var YearSq;
   model Population=Year / r cli clm;
   plot r.*p. / cframe=ligr;
run;
```

The DATA option ensures that the procedure uses the intended data set. Any variable that you might add to the model but that is not included in the first MODEL statement must appear in the VAR statement. 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. A plot of the residuals against the predicted values is requested by the PLOT statement.

The ANOVA table is displayed in Figure 55.4.

The REG Procedure
Model: MODEL1
Dependent Variable: Population

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	66336	66336	201.87	<.0001
Error	17	5586.29253	328.60544		
Corrected Total	18	71923			

Root MSE	18.12748	R-Square	0.9223
Dependent Mean	69.76747	Adj R-Sq	0.9178
Coeff Var	25.98271		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1958.36630	142.80455	-13.71	<.0001
Year	1	1.07879	0.07593	14.21	<.0001

Figure 55.4. ANOVA Table and Parameter Estimates

The Model F statistic is significant (F=201.873, 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

$$\text{Population} = -1958.37 + 1.08 \times \text{Year}$$

Figure 55.5 shows the confidence limits for both individual and expected values resulting from the CLM and CLI options.

The REG Procedure
Model: MODEL1
Dependent Variable: Population

Output Statistics

Obs	Dep Var Population	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	3.9290	-27.3240	7.9995	-44.2015	-10.4466	-69.1281	14.4800
2	5.3080	-16.5361	7.3615	-32.0674	-1.0048	-57.8150	24.7428
3	7.2390	-5.7481	6.7486	-19.9864	8.4901	-46.5582	35.0619
4	9.6380	5.0398	6.1684	-7.9744	18.0540	-35.3594	45.4390
5	12.8660	15.8277	5.6309	3.9475	27.7080	-24.2206	55.8761
6	17.0690	26.6157	5.1497	15.7509	37.4805	-13.1432	66.3746
7	23.1910	37.4036	4.7417	27.3996	47.4077	-2.1288	76.9360
8	31.4430	48.1916	4.4273	38.8508	57.5324	8.8218	87.5614
9	39.8180	58.9795	4.2275	50.0603	67.8987	19.7076	98.2514
10	50.1550	69.7675	4.1587	60.9933	78.5416	30.5283	109.0067
11	62.9470	80.5554	4.2275	71.6362	89.4746	41.2835	119.8273
12	75.9940	91.3434	4.4273	82.0026	100.6842	51.9736	130.7131
13	91.9720	102.1313	4.7417	92.1272	112.1354	62.5989	141.6637
14	105.7100	112.9193	5.1497	102.0544	123.7841	73.1603	152.6782
15	122.7750	123.7072	5.6309	111.8269	135.5875	83.6589	163.7555
16	131.6690	134.4951	6.1684	121.4810	147.5093	94.0959	174.8944
17	151.3250	145.2831	6.7486	131.0448	159.5214	104.4731	186.0931
18	179.3230	156.0710	7.3615	140.5397	171.6024	114.7921	197.3500
19	203.2110	166.8590	7.9995	149.9816	183.7364	125.0550	208.6630

Figure 55.5. Confidence Limits

The observed dependent variable is displayed for each observation along with its predicted value from the regression equation and the standard error of the mean predicted value. The 95% CL Mean columns are the confidence limits for the expected value of each observation. The 95% CL Predict columns are the confidence limits for the individual observations.

Figure 55.6 displays the residual analysis requested by the R option.

Output Statistics

Obs	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	31.2530	16.267	1.921	|*** |	0.446
2	21.8441	16.565	1.319	|** |	0.172
3	12.9871	16.824	0.772	|* |	0.048
4	4.5982	17.046	0.270	| |	0.005
5	-2.9617	17.231	-0.172	| |	0.002
6	-9.5467	17.381	-0.549	*| |	0.013
7	-14.2126	17.496	-0.812	*| |	0.024
8	-16.7486	17.579	-0.953	*| |	0.029
9	-19.1615	17.628	-1.087	**| |	0.034
10	-19.6125	17.644	-1.112	**| |	0.034
11	-17.6084	17.628	-0.999	*| |	0.029
12	-15.3494	17.579	-0.873	*| |	0.024
13	-10.1593	17.496	-0.581	*| |	0.012
14	-7.2093	17.381	-0.415	| |	0.008
15	-0.9322	17.231	-0.0541	| |	0.000
16	-2.8261	17.046	-0.166	| |	0.002
17	6.0419	16.824	0.359	| |	0.010
18	23.2520	16.565	1.404	|** |	0.195
19	36.3520	16.267	2.235	|**** |	0.604

Sum of Residuals	0
Sum of Squared Residuals	5586.29253
Predicted Residual SS (PRESS)	7619.90354

Figure 55.6. Residual Analysis

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 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 2 may indicate an inadequate model. The wave pattern seen in this plot is also an indication that the model is inadequate; a quadratic term may be needed or autocorrelation may be present in the data. 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. A fairly close agreement between the PRESS statistic (see Table 55.5 on page 2969) 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).

A plot of the residuals versus predicted values is shown in Figure 55.7.

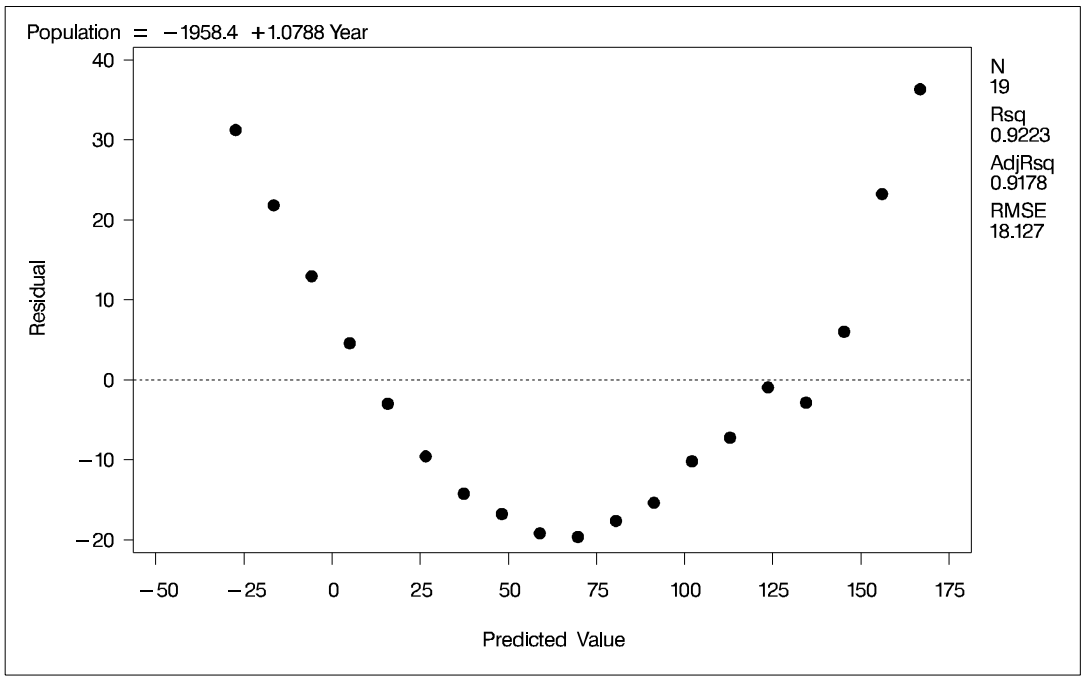


Figure 55.7. Plot of Residual vs. Predicted Values

The wave pattern of the studentized residual plot is seen here again. The semi-circle shape indicates an inadequate model; perhaps additional terms (such as the quadratic) are needed, or perhaps the data need to be transformed before analysis. If a model fits well, the plot of residuals against predicted values should exhibit no apparent trends.

Using the interactive feature of PROC REG, the following commands add the variable YearSq to the independent variables and refit the model.

```
add YearSq;
print;
plot / cframe=ligr;
run;
```

The ADD statement requests that YearSq be added to the model, and the PRINT command displays the ANOVA table for the new model. The PLOT statement with no variables recreates the most recent plot requested, in this case a plot of residual versus predicted values.

Figure 55.8 displays the ANOVA table and estimates for the new model.

The REG Procedure
Model: MODEL1.1
Dependent Variable: Population

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	71799	35900	4641.72	<.0001
Error	16	123.74557	7.73410		
Corrected Total	18	71923			

Root MSE	2.78102	R-Square	0.9983
Dependent Mean	69.76747	Adj R-Sq	0.9981
Coeff Var	3.98613		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20450	843.47533	24.25	<.0001
Year	1	-22.78061	0.89785	-25.37	<.0001
YearSq	1	0.00635	0.00023877	26.58	<.0001

Figure 55.8. ANOVA Table and Parameter Estimates

The overall F statistic is still significant (F=4641.719, p<0.0001). The R-square has increased from 0.9223 to 0.9983, indicating that the model now accounts for 99.8% of the variation in Population. All effects are significant with p<0.0001 for each effect in the model.

The fitted equation is now

$$\text{Population} = 20450 - 22.781 \times \text{Year} + 0.006 \times \text{Yearsq}$$

The confidence limits and residual analysis for the second model are displayed in Figure 55.9.

The REG Procedure
Model: MODEL1.1
Dependent Variable: Population

Output Statistics

Obs	Dep Var Population	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	3.9290	5.0384	1.7289	1.3734	8.7035	-1.9034	11.9803
2	5.3080	5.0389	1.3909	2.0904	7.9874	-1.5528	11.6306
3	7.2390	6.3085	1.1304	3.9122	8.7047	-0.0554	12.6724
4	9.6380	8.8472	0.9571	6.8182	10.8761	2.6123	15.0820
5	12.8660	12.6550	0.8721	10.8062	14.5037	6.4764	18.8335
6	17.0690	17.7319	0.8578	15.9133	19.5504	11.5623	23.9015
7	23.1910	24.0779	0.8835	22.2049	25.9509	17.8920	30.2638
8	31.4430	31.6931	0.9202	29.7424	33.6437	25.4832	37.9029
9	39.8180	40.5773	0.9487	38.5661	42.5885	34.3482	46.8065
10	50.1550	50.7307	0.9592	48.6972	52.7642	44.4944	56.9671
11	62.9470	62.1532	0.9487	60.1420	64.1644	55.9241	68.3823
12	75.9940	74.8448	0.9202	72.8942	76.7955	68.6350	81.0547
13	91.9720	88.8056	0.8835	86.9326	90.6785	82.6197	94.9915
14	105.7100	104.0354	0.8578	102.2169	105.8540	97.8658	110.2051
15	122.7750	120.5344	0.8721	118.6857	122.3831	114.3558	126.7130
16	131.6690	138.3025	0.9571	136.2735	140.3315	132.0676	144.5374
17	151.3250	157.3397	1.1304	154.9434	159.7360	150.9758	163.7036
18	179.3230	177.6460	1.3909	174.6975	180.5945	171.0543	184.2377
19	203.2110	199.2215	1.7289	195.5564	202.8865	192.2796	206.1633

Output Statistics

Obs	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	-1.1094	2.178	-0.509	| *| |	0.054
2	0.2691	2.408	0.112	| | |	0.001
3	0.9305	2.541	0.366	| | |	0.009
4	0.7908	2.611	0.303	| | |	0.004
5	0.2110	2.641	0.0799	| | |	0.000
6	-0.6629	2.645	-0.251	| | |	0.002
7	-0.8869	2.637	-0.336	| | |	0.004
8	-0.2501	2.624	-0.0953	| | |	0.000
9	-0.7593	2.614	-0.290	| | |	0.004
10	-0.5757	2.610	-0.221	| | |	0.002
11	0.7938	2.614	0.304	| | |	0.004
12	1.1492	2.624	0.438	| | |	0.008
13	3.1664	2.637	1.201	| |** |	0.054
14	1.6746	2.645	0.633	| |* |	0.014
15	2.2406	2.641	0.848	| |* |	0.026
16	-6.6335	2.611	-2.540	| *****| |	0.289
17	-6.0147	2.541	-2.367	| ****| |	0.370
18	1.6770	2.408	0.696	| |* |	0.054
19	3.9895	2.178	1.831	| |*** |	0.704

Sum of Residuals	-5.8175E-11
Sum of Squared Residuals	123.74557
Predicted Residual SS (PRESS)	188.54924

Figure 55.9. Confidence Limits and Residual Analysis

The plot of the studentized residuals shows that the wave structure is gone. The PRESS statistic is much closer to the Sum of Squared Residuals now, and both statistics have been dramatically reduced. Most of the Cook's D statistics have also been reduced.

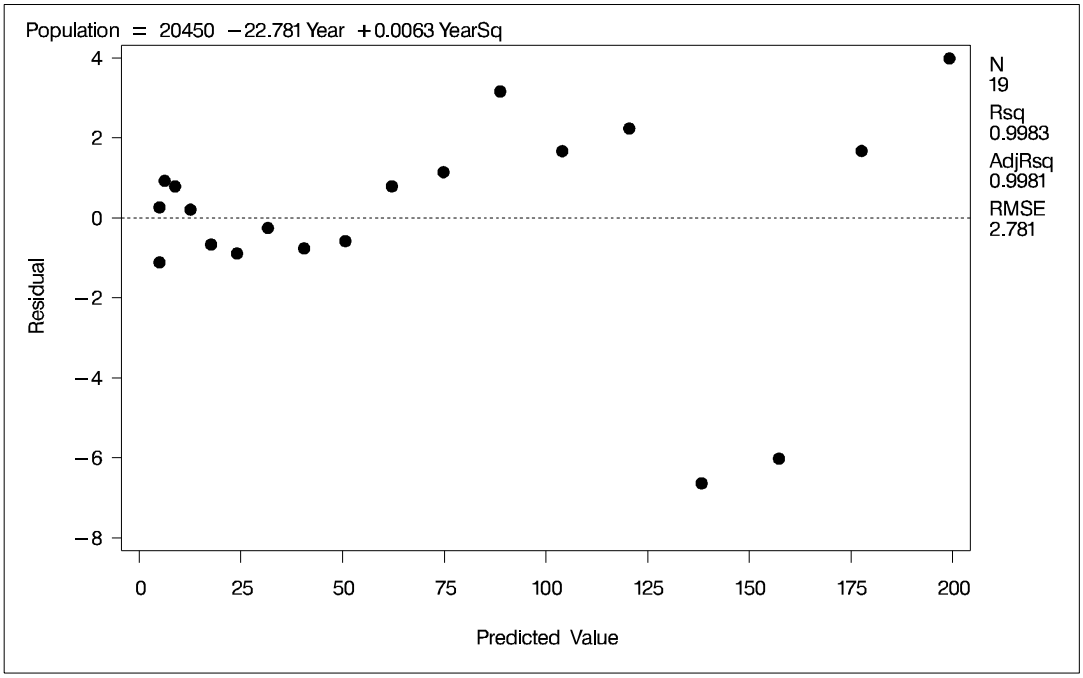


Figure 55.10. Plot of Residual vs. Predicted Values

The plot of residuals versus predicted values seen in Figure 55.10 has improved since a major trend is no longer visible.

To create a plot of the observed values, predicted values, and confidence limits against Year all on the same plot and to exert some control over the look of the resulting plot, you can submit the following statements.

```
symbol1 v=dot     c=yellow h=.3;
symbol2 v=square  c=red;
symbol3 f=simplex c=blue  h=2 v='-';
symbol4 f=simplex c=blue  h=2 v='-';
plot (Population predicted. u95. l95.)*Year
     / overlay cframe=ligr;
run;
```

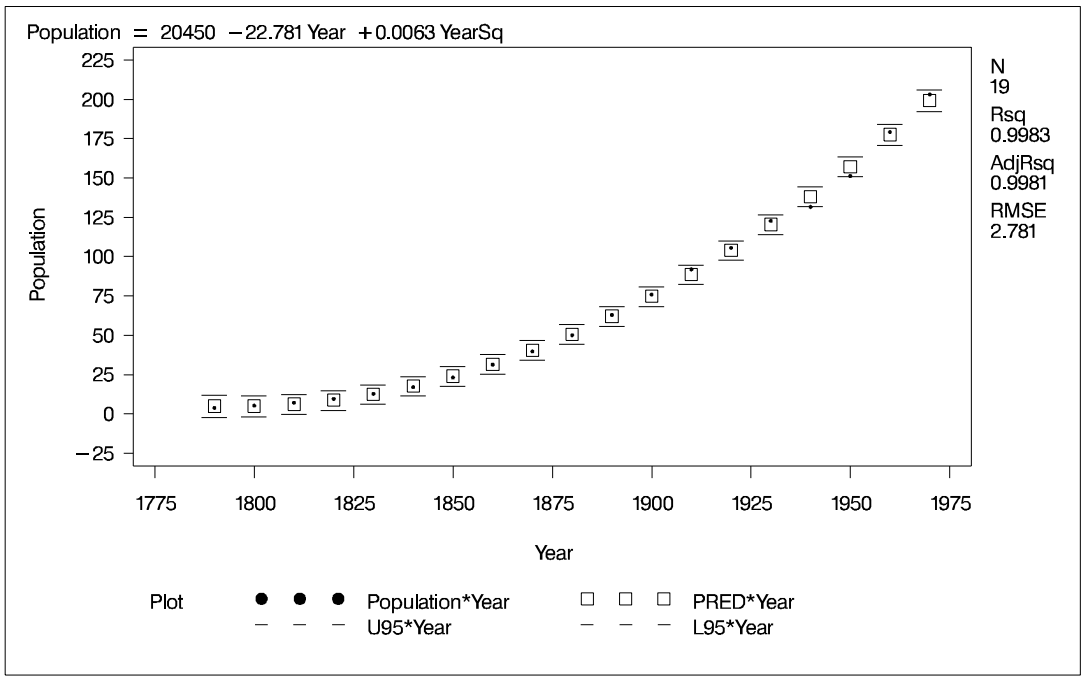


Figure 55.11. Plot of Population vs Year with Confidence Limits

The SYMBOL statements requests that the actual data be displayed as dots, the predicted values as squares, and the upper and lower 95% confidence limits for an individual value (sometimes called a *prediction interval*) as dashes. PROC REG provides the short-hand commands CONF and PRED to request confidence and prediction intervals for simple regression models; see the "PLOT Statement" section on page 2914 for details.

To complete an analysis of these data, you may want to examine influence statistics and, since the data are essentially time series data, examine the Durbin-Watson statistic. You might also want to examine other residual plots, such as the residuals vs. regressors.

Syntax

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**> **;**
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 "OUTSSCP= Data Sets" section on page 2942.) 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 TYPE=CORR, TYPE=COV, or TYPE=SSCP data sets are used as input data sets 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, INFLUENCE, and PARTIAL 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 55.1 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 55.1. PROC REG Statement Options

Option	Description
Data Set Options	
DATA=	names a data set to use for the regression
OUTEST=	outputs a data set that contains parameter estimates and other model fit summary statistics
OUTSSCP=	outputs a data set that contains sums of squares and crossproducts
COVOUT	outputs the covariance matrix for parameter estimates to the OUTEST= data set
EDF	outputs the number of regressors, the error degrees of freedom, and the model R^2 to the OUTEST= data set
OUTSTB	outputs standardized parameter estimates to the OUTEST= data set. Use only with the RIDGE= or PCOMIT= option.
OUTSEB	outputs standard errors of the parameter estimates to the OUTEST= data set
OUTVIF	outputs the variance inflation factors to the OUTEST= data set. Use only with the RIDGE= or PCOMIT= option.
PCOMIT=	performs incomplete principal component analysis and outputs estimates to the OUTEST= data set
PRESS	outputs the PRESS statistic to the OUTEST= data set
RIDGE=	performs ridge regression analysis and outputs estimates to the OUTEST= data set
RSQUARE	same effect as the EDF option
TABLEOUT	outputs standard errors, confidence limits, and associated test statistics of the parameter estimates to the OUTEST= data set
High Resolution Graphics Options	
ANNOTATE=	specifies an annotation data set
GOUT=	specifies the graphics catalog in which graphics output is saved

Table 55.1. (continued)

Option	Description
Display Options	
CORR	displays correlation matrix for variables listed in MODEL and VAR statements
SIMPLE	displays simple statistics for each variable listed in MODEL and VAR statements
USCCP	displays uncorrected sums of squares and crossproducts matrix
ALL	displays all statistics (CORR, SIMPLE, and USSCP)
NOPRINT	suppresses output
LINEPRINTER	creates plots requested as line printer plot
Other Options	
ALPHA=	sets significance value for confidence and prediction intervals and tests
SINGULAR=	sets criterion for checking for singularity

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 Software: Reference*. You can use this data set to add features to plots. 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 "OUTEST= Data Set" section on page 2938.

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 and some options in the MODEL and PRINT statements are not available. See Appendix A, "Special SAS Data Sets," for more information on 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=*graphics-catalog*

specifies the graphics catalog in which 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 plots requested as line printer plots. If you do not specify this option, requested plots are created on a high resolution graphics device. This option is required if plots are requested and you do not have SAS/GRAPH software.

NOPRINT

suppresses the normal display of results. Using this option in the PROC REG statement is equivalent to specifying NOPRINT in each MODEL statement. Note that this option temporarily disables the Output Delivery System (ODS); see Chapter 15, "Using the Output Delivery System," for more information.

OUTEST=*SAS-data-set*

requests that parameter estimates and optional model fit summary statistics be output to this data set. See the "OUTEST= Data Set" section on page 2938 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 on 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=*SAS-data-set*
requests that the sums of squares and crossproducts matrix be output to this TYPE=SSCP data set. See the "OUTSSCP= Data Sets" section on page 2942 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 on 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 components (IPC) analysis for each value *m* in the list. The procedure computes parameter estimates 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.

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

RIDGE=*list*
requests a ridge regression analysis and specifies the values of the ridge constant *k* (see the "Computations for Ridge Regression and IPC Analysis" section on page 2987). 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 55.10 on page 3023 illustrates this option.

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 "Computational Methods" section on page 2988.

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 α-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 "Interactive Analysis" section on page 2943 for an example.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC REG to obtain separate analyses on observations in groups defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in the order of the BY variables. The *variables* are one or more variables in the input data set.

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

- Sort the data using the SORT procedure with a similar BY statement.
- Specify the BY statement option NOTSORTED or DESCENDING 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 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 on the BY statement, refer to the discussion in *SAS Language Reference: Contents*. For more information on the DATASETS procedure, refer to the discussion in the *SAS Procedures Guide*.

DELETE Statement

DELETE *variables* **;**

The DELETE statement deletes independent 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 "Interactive Analysis" section on page 2943.

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 using a FREQ statement is the same as an analysis produced 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, or 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.

MODEL Statement

< *label:* > **MODEL** *dependents*=< *regressors*> < */ options* > **;**

After the keyword MODEL, the dependent (response) variables are specified, followed by an equal sign and the regressor variables. Variables specified in the MODEL statement must be numeric variables in the data set being analyzed. For example, if you want to specify a quadratic term for variable *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 55.2 lists the options available in the MODEL statement. Equations for the statistics available are given in the "Model Fit and Diagnostic Statistics" section on page 2968.

Table 55.2. MODEL Statement Options

Option	Description
Model Selection and Details of Selection	
SELECTION=	specifies model selection method
BEST=	specifies maximum number of subset models displayed or output to the OUTEST= data set
DETAILS	produces summary statistics at each step
DETAILS=	specifies the display details for forward, backward, and stepwise methods
GROUPNAMES=	provides names for groups of variables
INCLUDE=	includes first n variables in the model
MAXSTEP=	specifies maximum number of steps that may be performed
NOINT	fits a model without the intercept term
PCOMIT=	performs incomplete principal component analysis and outputs estimates to the OUTEST= data set
SLE=	sets criterion for entry into model
RIDGE=	performs ridge regression analysis and outputs estimates to the OUTEST= data set
SLS=	sets criterion for staying in model
START=	specifies number of variables in model to begin the comparing and switching process

Table 55.2. (continued)

Option	Description
STOP=	stops selection criterion
Fit Statistics	
ADJRSQ	computes adjusted R^2
AIC	computes Akaike's information criterion
B	computes parameter estimates for each model
BIC	computes Sawa's Bayesian information criterion
CP	computes Mallows' C_p statistic
GMSEP	computes estimated MSE of prediction assuming multivariate normality
JP	computes J_p, the final prediction error
MSE	computes MSE for each model
PC	computes Amemiya's prediction criterion
RMSE	displays root MSE for each model
SBC	computes the SBC statistic
SP	computes S_p statistic for each model
SSE	computes error sum of squares for each model
Data Set Options	
EDF	outputs the number of regressors, the error degrees of freedom, and the model R^2 to the OUTEST= data set
OUTSEB	outputs standard errors of the parameter estimates to the OUTEST= data set
OUTSTB	outputs standardized parameter estimates to the OUTEST= data set. Use only with the RIDGE= or PCOMIT= option.
OUTVIF	outputs the variance inflation factors to the OUTEST= data set. Use only with the RIDGE= or PCOMIT= option.
PRESS	outputs the PRESS statistic to the OUTEST= data set
RSQUARE	has same effect as the EDF option
Regression Calculations	
I	displays inverse of sums of squares and crossproducts
XPX	displays sums-of-squares and crossproducts matrix
Details on Estimates	
ACOV	displays asymptotic covariance matrix of estimates assuming heteroscedasticity
COLLIN	produces collinearity analysis
COLLINOINT	produces collinearity analysis with intercept adjusted out
CORRB	displays correlation matrix of estimates
COVB	displays covariance matrix of estimates
PCORR1	displays squared partial correlation coefficients using Type I sums of squares
PCORR2	displays squared partial correlation coefficients using Type II sums of squares
SCORR1	displays squared semi-partial correlation coefficients using Type I sums of squares

Table 55.2. (continued)

Option	Description
SCORR2	displays squared semi-partial correlation coefficients using Type II sums of squares
SEQB	displays a sequence of parameter estimates during selection process
SPEC	tests that first and second moments of model are correctly specified
SS1	displays the sequential sums of squares
SS2	displays the partial sums of squares
STB	displays standardized parameter estimates
TOL	displays tolerance values for parameter estimates
VIF	computes variance-inflation factors
Predicted and Residual Values	
CLB	computes $100(1-\alpha)\%$ confidence limits for the parameter estimates
CLI	computes $100(1-\alpha)\%$ confidence limits for an individual predicted value
CLM	computes $100(1-\alpha)\%$ confidence limits for the expected value of the dependent variable
DW	computes a Durbin-Watson statistic
INFLUENCE	computes influence statistics
P	computes predicted values
PARTIAL	displays partial regression plots for each regressor
R	produces analysis of residuals
Display Options and Other Options	
ALL	requests the following options: ACOV, CLB, CLI, CLM, CORRB, COVB, I, P, PCORR1, PCORR2, R, SCORR1, SCORR2, SEQB, SPEC, SS1, SS2, STB, TOL, VIF, XPX
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. See the section "Testing for Heteroscedasticity" on page 2981 for more information.

ADJRSQ
computes R^2 adjusted for degrees of freedom for each model selected (Darlington 1968; Judge et al. 1980).

AIC
computes Akaike's information criterion for each model selected (Akaike 1969; Judge et al. 1980).

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

ALPHA=*number*
sets the significance level used for the construction of confidence intervals for the current MODEL statement. The value must be between 0 and 1; the default value of 0.05 results in 95% intervals. This option affects the MODEL options CLB, CLI, and CLM; the OUTPUT statement keywords LCL, LCLM, UCL, and UCLM; the PLOT statement keywords LCL., LCLM., UCL., and UCLM.; and the PLOT statement options CONF and PRED. Specifying this option in the MODEL statement 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
computes Sawa's Bayesian information criterion for each model selected (Sawa 1978; Judge et al. 1980).

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 α-level.

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 α-level. The confidence limits reflect variation in the error, as well as variation in the parameter estimates. See the "Predicted and Residual Values" section on page 2952 and Chapter 3, "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 α-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 2952 and Chapter 3 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 "Collinearity Diagnostics" section on page 2967.

COLLINOINT

requests the same analysis as the COLLIN option with the intercept variable adjusted out rather than included in the diagnostics. See the "Collinearity Diagnostics" section on page 2967.

CORRB

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

COVB

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

CP

computes Mallows' C_p statistic for each model selected (Mallows 1973; Hocking 1976). See the "Criteria Used in Model-Selection Methods" section on page 2949 for a discussion of the use of C_p.

DETAILS
DETAILS=*name*

specifies the level of detail produced when the BACKWARD, FORWARD or STEPWISE methods are 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.) The sample autocorrelation of the residuals is also produced. See the section "Autocorrelation in Time Series Data" on page 2986.

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

computes the estimated mean square error of prediction assuming that both independent and dependent variables are multivariate normal (Stein 1960; Darlington 1968). Note that Hocking's formula (1976, eq. 4.20) contains a misprint: "$n - 1$" should read "$n - 2$.")

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 2907) 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, . . ., GROUP*n* 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';
```

As another example,

```
model y={ht wgt age} bodyfat / selection=forward
  groupnames='htwgtage' 'bodyfat';
```

I

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

INCLUDE=*n*

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

INFLUENCE
requests a detailed analysis of the influence of each observation on the estimates and the predicted values. See the "Influence Diagnostics" section on page 2970 for details.

JP
computes J_p, the estimated mean square error of prediction for each model selected assuming that the values of the regressors are fixed and that the model is correct. The J_p statistic is also called the final prediction error (FPE) by Akaike (Nicholson 1948; Lord 1950; Mallows 1967; Darlington 1968; Rothman 1968; Akaike 1969; Hocking 1976; Judge et al. 1980).

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 15, "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 2952 for more information.

PARTIAL

requests partial regression leverage plots for each regressor. See the "Influence Diagnostics" section on page 2970 for more information.

PC

computes Amemiya's prediction criterion for each model selected (Amemiya 1976; Judge et al. 1980).

PCOMIT=*list*

requests an IPC analysis for each value *m* in the list. The procedure computes parameter estimates 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 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 using Type II sums of squares. These are calculated the same way as with the PCORR1 option, except that Type II SS are used instead of Type I SS.

PRESS

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

R

requests an analysis of the residuals. The results include everything requested by the P option plus the standard errors of the mean predicted and residual values, the studentized residual, and Cook's *D* statistic to measure the influence of each observation on the parameter estimates. See the section "Predicted and Residual Values" on page 2952 for more information.

RIDGE=*list*

requests a ridge regression analysis and specifies the values of the ridge constant *k* (see the "Computations for Ridge Regression and IPC Analysis" section on page 2987). 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 55.10 on page 3023 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

computes the SBC statistic for each model selected (Schwarz 1978; Judge et al. 1980).

SCORR1

displays the squared semi-partial correlation coefficients 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.

SCORR2

displays the squared semi-partial correlation coefficients 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.

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 "Model-Selection Methods" section on page 2947 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. Specifying this option in the MODEL statement 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 "Computational Methods" section on page 2988.

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
computes the S_p statistic for each model selected (Hocking 1976).

SPEC
performs a test that the first and second moments of the model are correctly specified. See the section "Testing for Heteroscedasticity" on page 2981 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 12, "The Four Types of Estimable Functions," for more information on 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 2967 for more detail.

VIF
produces variance inflation factors with the parameter estimates. Variance inflation is the reciprocal of tolerance. See the section "Collinearity Diagnostics" on page 2967 for more detail.

XPX
displays the $\mathbf{X}'\mathbf{X}$ crossproducts matrix for the model. The crossproducts matrix is bordered by the $\mathbf{X}'\mathbf{Y}$ and $\mathbf{Y}'\mathbf{Y}$ 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

$$(\mathbf{L}\beta - \mathbf{cj})\mathbf{M} = 0$$

where $\mathbf{L}$ is a linear function on the regressor side, β is a matrix of parameters, $\mathbf{c}$ is a column vector of constants, $\mathbf{j}$ is a row vector of ones, and $\mathbf{M}$ is a linear function on

the dependent side. The special case where the constants are zero is

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

See the section "Multivariate Tests" on page 2981 for more details.

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

As an example, consider the following statements:

```
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

```
mtest / canprint;
```

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

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

PRINT
displays the $\mathbf{H}$ and $\mathbf{E}$ 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 on 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 "Input Data Sets" section on page 2935 for more details.

The statistics created in the OUTPUT statement are described in this section. More details are contained in the "Predicted and Residual Values" section on page 2952 and the "Influence Diagnostics" section on page 2970. Also see Chapter 3, "Introduction to Regression Procedures," for definitions of the statistics available from the REG procedure.

You can specify the following options in the OUTPUT statement.

OUT=*SAS data set*
gives the name of the new data set. By default, the procedure uses the DATA*n* 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 SAS statements

```
proc reg data=a;
   model y z=x1 x2;
   output out=b
      p=yhat zhat
      r=yresid zresid;
run;
```

create an output data set named b. 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 "Model Fit and Diagnostic Statistics" section on page 2968 for computational formulas.

Keyword	Description
COOKD=*names*	Cook's D influence statistic
COVRATIO=*names*	standard influence of observation on covariance of betas, as discussed in the "Influence Diagnostics" section on page 2970
DFFITS=*names*	standard influence of observation on predicted value
H=*names*	leverage, $x_i(\mathbf{X}'\mathbf{X})^{-1}x_i'$
LCL=*names*	lower bound of a $100(1-\alpha)$% confidence interval for an individual prediction. This includes the variance of the error, as well as the variance of the parameter estimates.
LCLM=*names*	lower bound of a $100(1-\alpha)$% confidence interval for the expected value (mean) of the dependent variable
PREDICTED | P=*names*	predicted values
PRESS=*names*	ith residual divided by $(1-h)$, where h is the leverage, and where the model has been refit without the ith observation
RESIDUAL | R=*names*	residuals, calculated as ACTUAL minus PREDICTED
RSTUDENT=*names*	a studentized residual with the current observation deleted
STDI=*names*	standard error of the individual predicted value
STDP=*names*	standard error of the mean predicted value
STDR=*names*	standard error of the residual
STUDENT=*names*	studentized residuals, which are the residuals divided by their standard errors
UCL=*names*	upper bound of a $100(1-\alpha)$% confidence interval for an individual prediction
UCLM=*names*	upper bound of a $100(1-\alpha)$% confidence interval for the expected value (mean) of the dependent variable

PAINT Statement

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

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.

For detailed examples of painting scatter plots, see the section “Painting Scatter Plots” on page 2962.

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 can be used instead of the preceding symbols. Refer to the "Expressions" section in *SAS Language Reference: Concepts* for more information on 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 |.

Examples of the *variable compare value* form are

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

Examples of the *variable compare value logical variable compare value* form are

```
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 allobs / reset;
```

resets the symbols for all observations.

Options in the PAINT Statement

The following options can be used when either a condition is specified, the ALLOBS option is specified, or when 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,

```
paint temp lt 22 / symbol='c';
```

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

STATUS and UNDO

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

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

UNDO

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

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

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

```
paint / reset;
```

is used instead, observation 1 is unpainted.

PLOT Statement

PLOT < *yvariable*******xvariable* >< *=symbol* >
< . . . *yvariable*******xvariable* > < *=symbol* > < */ options* >**;**

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, high resolution graphics plots 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 high resolution graphics plots.

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.

The PLOT statement cannot be used when TYPE=CORR, TYPE=COV, or TYPE=SSCP data sets are 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. For detailed examples of using the PLOT statement and its options, see the section "Producing Scatter Plots" on page 2955.

Specifying Yvariables, Xvariables, and Symbol

More than one *yvariable*∗*xvariable* pair can be specified to request multiple plots. The *yvariables* and *xvariables* can be

- 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 55.3 on page 2917). 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 2987 and Example 55.8 on page 3020 for more information)
- keywords for model fit summary statistics available in the OUTEST= data set with _TYPE_= PARMS (see Table 55.3 on page 2917). 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 "Line Printer Plots" section on page 2924, apply across PLOT statements and remain in effect if they have been previously specified.

Options used for high resolution graphics plots are described in the following section; see for more information.

High Resolution Graphics Plots

The display of high resolution graphics plots is described in the following paragraphs, the options are summarized in Table 55.3 and described in the section "Dictionary of PLOT Statement Options" on page 2919, and the "Examples" section on page 2993 contains several examples of the graphics output.

Several line printer statements and options are not supported for high resolution graphics. In particular the PAINT statement is disabled, as are the PLOT statement options CLEAR, COLLECT, HPLOTS=, NOCOLLECT, SYMBOL=, and VPLOTS=. To display more than one plot per page or to collect plots from multiple PLOT statements, use the PROC GREPLAY statement (refer to *SAS/GRAPH Software: Reference*). Also note that high resolution 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 55.4 on page 2918 lists options to control the display of the equation. The "Examples" section on page 2993 illustrates the display of the model 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. (See Output 55.4.1 on page 3016.) 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 55.4 on page 2918.

A default reference line at zero is displayed if residuals are plotted; see Output 55.7.1 on page 3019. If the dependent variable is plotted against the independent variable in a simple linear regression model, the fitted regression line is displayed by default. (See Output 55.4.1 on page 3016.) 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" beginning on page 2919 and Example 55.10 on page 3023 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" beginning on page 2919 and in Example 55.9 on page 3022 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" beginning on page 2919 and Example 55.5 on page 3016 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 55.3. Keywords for PLOT Statement *xvariables* and *yvariables*

Keyword	Description
Diagnostic Statistics	
COOKD.	Cook's D influence statistics
COVRATIO.	standard influence of observation on covariance of betas
DFFITS.	standard influence of observation on predicted value
H.	leverage
LCL.	lower bound of $100(1-\alpha)\%$ confidence interval for individual prediction
LCLM.	lower bound of $100(1-\alpha)\%$ confidence interval for the mean of the dependent variable
PREDICTED. | PRED. | P.	predicted values
PRESS.	residuals from refitting the model with current observation deleted
RESIDUAL. | R.	residuals
RSTUDENT.	studentized residuals with the current observation deleted
STDI.	standard error of the individual predicted value
STDP.	standard error of the mean predicted value
STDR.	standard error of the residual
STUDENT.	residuals divided by their standard errors
UCL.	upper bound of $100(1-\alpha)\%$ confidence interval for individual prediction
UCLM.	upper bound of $100(1-\alpha)\%$ confidence interval for the mean of the dependent variables
Other Keywords used with Diagnostic Statistics	
NPP.	normal probability-probability plot
NQQ.	normal quantile-quantile plot
OBS.	observation number (cannot plot against OUTEST= statistics)
Model Fit Summary Statistics	
ADJRSQ.	adjusted R-square
AIC.	Akaike's information criterion
BIC.	Sawa's Bayesian information criterion
CP.	Mallows' C_p statistic

Table 55.3. (continued)

Keyword	Description
EDF.	error degrees of freedom
GMSEP.	estimated MSE of prediction, assuming multivariate normality
IN.	number of regressors in the model not including the intercept
JP.	final prediction error
MSE.	mean squared error
NP.	number of parameters in the model (including the intercept)
PC.	Amemiya's prediction criterion
RMSE.	root MSE
RSQ.	R-square
SBC.	SBC statistic
SP.	SP statistic
SSE.	error sum of squares

Summary of PLOT Statement Graphics Options

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" beginning on page 2919.

Table 55.4. High Resolution Graphics Options

Option	Description
General Graphics Options	
ANNOTATE=*SAS-data-set*	specifies the annotate data set
CHOCKING=*color*	requests a reference line for C_p model selection criteria
CMALLOWS=*color*	requests a reference line for the C_p model selection criterion
CONF	requests plots of $100(1-\alpha)\%$ confidence intervals for the mean
DESCRIPTION='*string*'	specifies a description for graphics catalog member
NAME='*string*'	names the plot in graphics catalog
OVERLAY	overlays plots from the same model
PRED	requests plots of $100(1-\alpha)\%$ prediction intervals for individual responses
RIDGEPLOT	requests the ridge trace for ridge regression
Axis and Legend Options	
LEGEND=*LEGENDn*	specifies LEGEND statement to be used
HAXIS=*values*	specifies tick mark values for horizontal axis
VAXIS=*values*	specifies tick mark values for vertical axis
Reference Line Options	
HREF=*values*	specifies reference lines perpendicular to horizontal axis
LHREF=*linetype*	specifies line style for HREF= lines
LLINE=*linetype*	specifies line style for lines displayed by default
LVREF=*linetype*	specifies line style for VREF= lines
NOLINE	suppresses display of any default reference line

Table 55.4. (continued)

Option	Description
VREF=*values*	specifies reference lines perpendicular to vertical axis
Color Options	
CAXIS=*color*	specifies color for axis line and tick marks
CFRAME=*color*	specifies color for frame
CHREF=*color*	specifies color for HREF= lines
CLINE=*color*	specifies color for lines displayed by default
CTEXT=*color*	specifies color for text
CVREF=*color*	specifies color for VREF= lines
Options for Displaying the Fitted Model Equation	
MODELFONT=*font*	specifies font of model equation and model label
MODELHT=*value*	specifies text height of model equation and model label
MODELLAB=*'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 the 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=*font*	specifies font of text displayed in the margin
STATHT=*value*	specifies height of text displayed in the margin

Dictionary of PLOT Statement Options

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

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

ANNOTATE=*SAS-data-set*
ANNO=*SAS-data-set*
specifies an input data set that contains appropriate variables for annotation. This applies only to displays created with the current PLOT statement. Refer to *SAS/GRAPH Software: 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 (see Example 55.5 on page 3016).

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 later in this section 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 (see Example 55.9 on page 3022). The ALPHA= option in the PROC REG or MODEL statement selects the significance level α, 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. (See Example 55.6 on page 3017.)

CVREF=*color*
CV=*color*
specifies the color for lines requested with the VREF= option.

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

EDF
displays the error degrees of freedom in the plot margin.

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

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

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

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

JP
displays the J_p statistic in the plot margin.

LEGEND=LEGEND*n*
specifies the LEGENDn statement to be used. The LEGENDn statement is a global graphics statement; refer to *SAS/GRAPH Software: Reference* for more information.

LHREF=*linetype*
LH=*linetype*
specifies the line style for lines requested with the HREF= option. The default *linetype* is 2. Note that LHREF=1 requests a solid line. Refer to *SAS/GRAPH Software: 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. Refer to *SAS/GRAPH Software: 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 explanation in the section "High Resolution Graphics Plots" beginning on page 2915 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.

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 the 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. See Example 55.11 on page 3024.

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 (see Example 55.9 on page 3022). The ALPHA= option in the PROC REG or MODEL statement selects the significance level α, 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 the MODEL statement. See Example 55.10 on page 3023.

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. Refer to *SAS/GRAPH Software: Reference* for tables of software fonts.

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

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 (left-most) 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 ten 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 on 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

```
plot residual.*predicted. y*x / collect;
run;
```

produce two plots. If these statements are then followed by

```
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
allows 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.

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 "MODEL Statement" section on page 2897 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 "Input Data Sets" section on page 2935 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, with these statements

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

the second PAINT statement paints any additional observations that meet the condition after deleting observations and refitting the model. The REFIT statement is used because the REWEIGHT statement does not cause the model to be recomputed. In this particular example, the same effect could be achieved by replacing the REFIT statement with a PLOT statement.

Most interactive statements can be used to implicitly refit the model; any plots or statistics produced by these statements reflect changes made to the model and changes made to the data used to compute the model. The two exceptions are the PAINT and REWEIGHT statements, which do not cause the model to be recomputed.

RESTRICT Statement

RESTRICT *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

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

is equivalent to imposing the two restrictions

```
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

$$c_1 \times variable_1 \pm c_2 \times variable_2 \pm \cdots \pm c_n \times variable_n$$

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

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

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

Examples of valid RESTRICT statements include the following:

```
restrict x1;
restrict a+b=l;
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 Raynor 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 γ measures the sensitivity of the SSE to the restriction constant. If the restriction constant is changed by a small amount ϵ, the SSE is changed by $2\gamma\epsilon$. The t ratio tests the significance of the restrictions. If γ is zero, the restricted estimates are the same as the unrestricted estimates, and a change in the restriction constant in either direction increases the SSE.

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

REWEIGHT Statement

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

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

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

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

the second REWEIGHT statement does not exclude any additional observations since the model is not recomputed after the first REWEIGHT statement. Use either 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 2974.

Specifying Condition

Condition is used to find observations to be reweighted. 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. 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 can be used instead of the preceding symbols. Refer to the "Expressions" chapter in *SAS Language Reference: Concepts* for more information on 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 |.

Examples of the *variable compare value* form are

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

Examples of the *variable compare value logical variable compare value* form are

```
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, with the 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';
 ...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 may be, but are not necessarily, reset. For example, in these statements

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

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

TEST Statement

< *label:* > **TEST** *equation* < , . . . , *equation* > < */ options* > **;**

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 allows 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

$$\mathbf{L}\beta = \mathbf{c}$$

then the numerator of the F test is

$$\mathbf{Q} = (\mathbf{Lb} - \mathbf{c})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}')^{-1}(\mathbf{Lb} - \mathbf{c})$$

divided by degrees of freedom, where $\mathbf{b}$ is the estimate of β. For example,

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

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

Note that, when the ACOV option is specified in the MODEL statement, tests are recomputed using the heteroscedasticity consistent covariance matrix (see the section "Testing for Heteroscedasticity" on page 2981).

One option can be specified in the TEST statement after a slash (/):

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

VAR Statement

VAR *variables* **;**

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

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

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

WEIGHT Statement

WEIGHT *variable* **;**

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

Values of the weight variable must be nonnegative. If an observation's weight is zero, the observation is deleted from the analysis. If a weight is negative or missing, it is set to zero, and the observation is excluded from the analysis. A more complete description of the WEIGHT statement can be found in Chapter 30, "The GLM Procedure."

Observation weights can be changed interactively with the REWEIGHT statement; see the section "REWEIGHT Statement" beginning on page 2929.

Details

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 may 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, or 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. See Appendix A, "Special SAS Data Sets," and the "SAS Files" section in *SAS Language Reference: Concepts* for more information on special SAS data sets.

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 percent 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 TYPE=CORR, TYPE=COV, or TYPE=SSCP data sets are 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

This example uses PROC CORR to produce an input data set for PROC REG. The fitness data for this analysis can be found in Example 55.1 on page 2993.

```
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 55.12. Figure 55.13 shows results from the regression using the TYPE=CORR data as an input data set.

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

Figure 55.12. TYPE=CORR Data Set Created by PROC CORR

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	656.27095	218.75698	30.27	<.0001
Error	27	195.11060	7.22632		
Corrected Total	30	851.38154			

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

Figure 55.13. Regression on TYPE=CORR Data Set

Example Using TYPE=SSCP Data Set

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 55.14 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 55.15 also shows the PROC REG results for the reduced model. (For the PROC REG results for the full model, see Figure 55.27 on page 2951.)

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

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

Figure 55.14. TYPE=SSCP Data Set Created by PROC CORR

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	656.27095	218.75698	30.27	<.0001
Error	27	195.11060	7.22632		
Corrected Total	30	851.38154			

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

Figure 55.15. Regression on TYPE=SSCP Data Set

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 8-byte 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 for the estimate, where n is the nearest integer to $100(1-\alpha)$ and α defaults to 0.05 or is set using the ALPHA= option in the PROC REG or MODEL statement
- UnB, the $100(1-\alpha)$ upper confidence 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
- _JP_, the J_p statistic, if the JP option is specified
- _PC_, the PC statistic, if the PC option is specified
- _GMSEP_, the GMSEP statistic, if the GMSEP option is specified
- _AIC_, the AIC statistic, if the AIC option is specified
- _BIC_, the BIC statistic, if the BIC option is specified
- _SBC_, the SBC statistic, if the SBC option is specified

The following is an example with a display of the OUTEST= data set. This example uses the population data given in the section "Polynomial Regression" beginning on page 2880. Figure 55.16 on page 2940 through Figure 55.18 on page 2941 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;
```

The REG Procedure
Model: M1
Dependent Variable: Population

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	66336	66336	201.87	<.0001
Error	17	5586.29253	328.60544		
Corrected Total	18	71923			

Root MSE	18.12748	R-Square	0.9223
Dependent Mean	69.76747	Adj R-Sq	0.9178
Coeff Var	25.98271		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1958.36630	142.80455	-13.71	<.0001
Year	1	1.07879	0.07593	14.21	<.0001

Figure 55.16. Regression Output for Model M1

The REG Procedure
Model: M2
Dependent Variable: Population

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	71799	35900	4641.72	<.0001
Error	16	123.74557	7.73410		
Corrected Total	18	71923			

Root MSE	2.78102	R-Square	0.9983
Dependent Mean	69.76747	Adj R-Sq	0.9981
Coeff Var	3.98613		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	20450	843.47533	24.25	<.0001
Year	1	-22.78061	0.89785	-25.37	<.0001
YearSq	1	0.00635	0.00023877	26.58	<.0001

Figure 55.17. Regression Output for Model M2

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Year	Population	YearSq
1	M1	PARMS	Population	18.1275	-1958.37	1.0788	-1	.
2	M2	PARMS	Population	2.7810	20450.43	-22.7806	-1	.006345585

Figure 55.18. OUTEST= Data Set

The following modification of the previous example uses the TABLEOUT and ALPHA= options to obtain additional information in the OUTEST= data set:

```
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 follows.

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Year	Population	YearSq
1	M1	PARMS	Population	18.1275	-1958.37	1.0788	-1	.
2	M1	STDERR	Population	18.1275	142.80	0.0759	.	.
3	M1	T	Population	18.1275	-13.71	14.2082	.	.
4	M1	PVALUE	Population	18.1275	0.00	0.0000	.	.
5	M1	L90B	Population	18.1275	-2206.79	0.9467	.	.
6	M1	U90B	Population	18.1275	-1709.94	1.2109	.	.
7	M2	PARMS	Population	2.7810	20450.43	-22.7806	-1	0.0063
8	M2	STDERR	Population	2.7810	843.48	0.8978	.	0.0002
9	M2	T	Population	2.7810	24.25	-25.3724	.	26.5762
10	M2	PVALUE	Population	2.7810	0.00	0.0000	.	0.0000
11	M2	L90B	Population	2.7810	18977.82	-24.3481	.	0.0059
12	M2	U90B	Population	2.7810	21923.04	-21.2131	.	0.0068

Figure 55.19. The OUTEST= Data Set When TABLEOUT is Specified

A slightly different OUTEST= data set is created when you use the RSQUARE selection method. This example requests only the "best" model for each subset size but asks for a variety of model selection statistics, as well as the estimated regression coefficients. An OUTEST= data set is created and displayed. See Figure 55.20 and Figure 55.21 for results.

```
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;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

R-Square Selection Method

Number in Model	R-Square	C(p)	AIC	BIC	Estimated MSE of Prediction	J(p)	MSE	SBC
1	0.7434	13.6988	64.5341	65.4673	8.0546	8.0199	7.53384	67.40210
2	0.7642	12.3894	63.9050	64.8212	7.9478	7.8621	7.16842	68.20695
3	0.8111	6.9596	59.0373	61.3127	6.8583	6.7253	5.95669	64.77326
4	0.8368	4.8800	56.4995	60.3996	6.3984	6.2053	5.34346	63.66941
5	0.8480	5.1063	56.2986	61.5667	6.4565	6.1782	5.17634	64.90250
6	0.8487	7.0000	58.1616	64.0748	6.9870	6.5804	5.36825	68.19952

Number in Model	R-Square	Parameter Estimates: Intercept	Age	Weight	RunTime	RunPulse	RestPulse	MaxPulse
1	0.7434	82.42177	.	.	-3.31056	.	.	.
2	0.7642	88.46229	-0.15037	.	-3.20395	.	.	.
3	0.8111	111.71806	-0.25640	.	-2.82538	-0.13091	.	.
4	0.8368	98.14789	-0.19773	.	-2.76758	-0.34811	.	0.27051
5	0.8480	102.20428	-0.21962	-0.07230	-2.68252	-0.37340	.	0.30491
6	0.8487	102.93448	-0.22697	-0.07418	-2.62865	-0.36963	-0.02153	0.30322

Figure 55.20. PROC REG Output for Physical Fitness Data: Best Models

Obs	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	Age	Weight	RunTime	RunPulse	RestPulse	Max Pulse
1	MODEL1	PARMS	Oxygen	2.74478	82.422	.	.	-3.31056	.	.	.
2	MODEL1	PARMS	Oxygen	2.67739	88.462	-0.15037	.	-3.20395	.	.	.
3	MODEL1	PARMS	Oxygen	2.44063	111.718	-0.25640	.	-2.82538	-0.13091	.	.
4	MODEL1	PARMS	Oxygen	2.31159	98.148	-0.19773	.	-2.76758	-0.34811	.	0.27051
5	MODEL1	PARMS	Oxygen	2.27516	102.204	-0.21962	-0.072302	-2.68252	-0.37340	.	0.30491
6	MODEL1	PARMS	Oxygen	2.31695	102.934	-0.22697	-0.074177	-2.62865	-0.36963	-0.021534	0.30322

Obs	Oxygen	_IN_	_P_	_EDF_	_MSE_	_RSQ_	_CP_	_JP_	_GMSEP_	_AIC_	_BIC_	_SBC_
1	-1	1	2	29	7.53384	0.74338	13.6988	8.01990	8.05462	64.5341	65.4673	67.4021
2	-1	2	3	28	7.16842	0.76425	12.3894	7.86214	7.94778	63.9050	64.8212	68.2069
3	-1	3	4	27	5.95669	0.81109	6.9596	6.72530	6.85833	59.0373	61.3127	64.7733
4	-1	4	5	26	5.34346	0.83682	4.8800	6.20531	6.39837	56.4995	60.3996	63.6694
5	-1	5	6	25	5.17634	0.84800	5.1063	6.17821	6.45651	56.2986	61.5667	64.9025
6	-1	6	7	24	5.36825	0.84867	7.0000	6.58043	6.98700	58.1616	64.0748	68.1995

Figure 55.21. PROC PRINT Output for Physical Fitness Data: OUTEST= Data Set

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 may need in a VAR statement or include all the variables that you may need in a MODEL statement.

The following example uses the fitness data from Example 55.1 on page 2993 to produce an output data set with the OUTSSCP= option. The resulting output is shown in Figure 55.22.

```
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.

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

Figure 55.22. SSCP Data Set Created with OUTSSCP= Option: REG Procedure

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. 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 2974. Extra features for line printer scatter plots are discussed in the section "Line Printer Scatter Plot Features" on page 2955.

The following example illustrates the usefulness of the interactive features. First, the full regression model is fit to the class data (see the "Getting Started" section on page 2877), and Figure 55.23 is produced.

```
proc reg data=Class;
   model Weight=Age Height;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	7215.63710	3607.81855	27.23	<.0001
Error	16	2120.09974	132.50623		
Corrected Total	18	9335.73684			

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

Figure 55.23. Interactive Analysis: Full Model

Next, the regression model is reduced by the following statements, and Figure 55.24 is produced.

```
delete age;
print;
run;
```

```
                          The REG Procedure
                           Model: MODEL1.1
                     Dependent Variable: Weight

                         Analysis of Variance

                                    Sum of           Mean
Source                   DF        Squares         Square    F Value    Pr > F

Model                     1     7193.24912     7193.24912      57.08    <.0001
Error                    17     2142.48772      126.02869
Corrected Total          18     9335.73684


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


                        Parameter Estimates

                     Parameter       Standard
    Variable     DF     Estimate          Error    t Value    Pr > |t|

    Intercept     1   -143.02692       32.27459      -4.43      0.0004
    Height        1      3.89903        0.51609       7.55      <.0001
```

Figure 55.24. Interactive Analysis: Reduced Model

Note that the MODEL label has been changed from MODEL1 to MODEL1.1, as the original MODEL has been changed by the delete statement.

The following statements generate a scatter plot of the residuals against the predicted values from the full model. Figure 55.25 is produced, and the scatter plot shows a possible outlier.

```
add age;
plot r.*p. / cframe=ligr;
run;
```

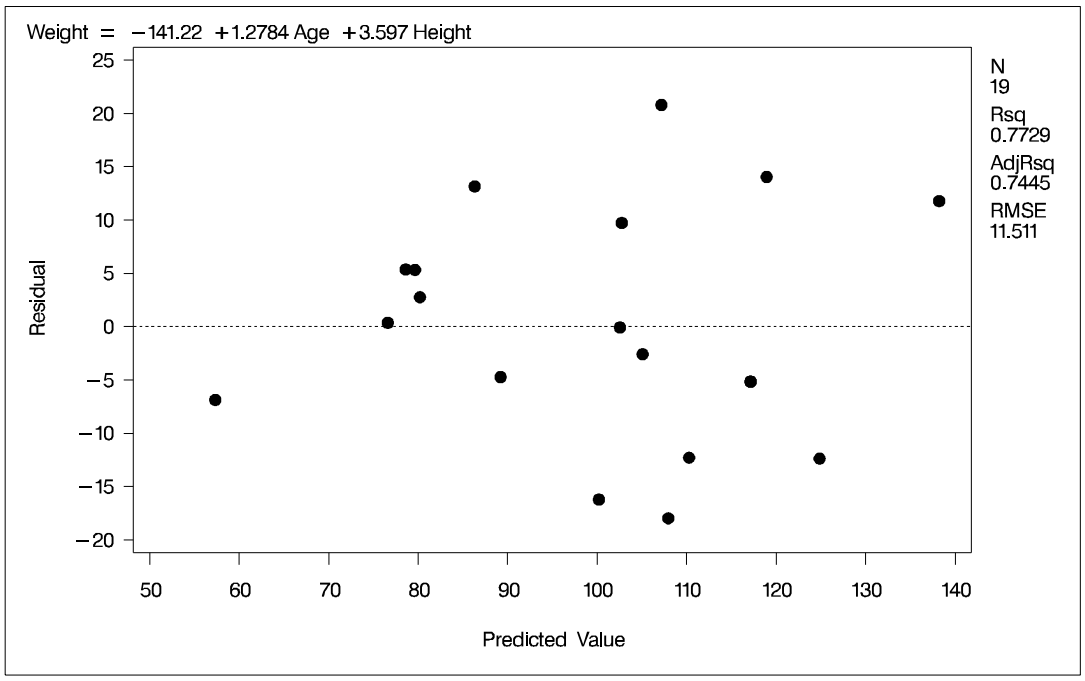


Figure 55.25. Interactive Analysis: Scatter Plot

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 55.26 shows the new scatter plot.

```
reweight r.>20;
plot / cframe=ligr;
run;
```

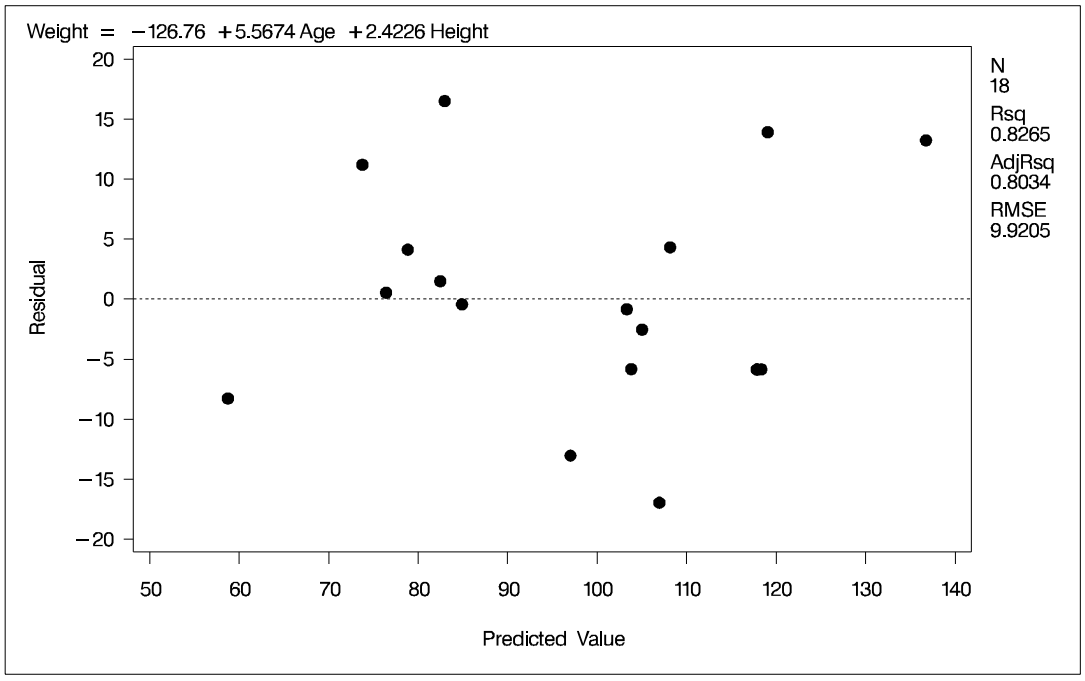


Figure 55.26. Interactive Analysis: Scatter Plot for Refitted Model

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 may be the only method you need.

Forward Selection (FORWARD)

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

Backward Elimination (BACKWARD)

The backward elimination technique begins by calculating F statistics for a model, including 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 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 may be removed without considering what adding the "best" remaining variable might accomplish. The MAXR method may 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 may 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 on 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 percent, the overall probability of rejecting at least one true null hypothesis is much larger than 5 percent. 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 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 percent to 25 percent.

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 methods, with a large value of the BEST= option, adding one more variable to the list from which regressors are selected may 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 55.1 on page 2993. Figure 55.28 shows the parameter estimates and the tables from the SS1, SS2, STB, CLB, COVB, and CORRB options:

```
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 55.27). The F statistic for the overall model is significant, indicating that the model explains a significant portion of the variation in the data.

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	722.54361	120.42393	22.43	<.0001
Error	24	128.83794	5.36825		
Corrected Total	30	851.38154			

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

Figure 55.27. ANOVA Table

The procedure next displays Parameter Estimates and some associated statistics (Figure 55.28). First, the estimates are shown, followed by their Standard Errors. The next two columns of the table contain the t statistics and the corresponding probabilities for testing the null hypothesis that the parameter is not significantly different from zero. These probabilities are usually referred to as p-values. For example, the Intercept term in the model is estimated to be 102.9 and is significantly different from zero. The next two columns of the table are the result of requesting the SS1 and SS2 options, and they show sequential and partial Sums of Squares (SS) associated with each variable. The Standardized Estimates (produced by the STB option) are the parameter estimates that result when all variables are standardized to a mean of 0 and a variance of 1. These estimates are computed by multiplying the original estimates by the standard deviation of the regressor (independent) variable and then dividing by the standard deviation of the dependent variable. The CLB option adds the upper and lower 95% confidence limits for the parameter estimates; the α level can be changed by specifying the ALPHA= option in the PROC REG or MODEL statement.

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Standardized Estimate	95% Confidence Limits	
Intercept	1	102.93448	12.40326	8.30	<.0001	69578	369.72831	0	77.33541	128.53355
RunTime	1	-2.62865	0.38456	-6.84	<.0001	632.90010	250.82210	-0.68460	-3.42235	-1.83496
Age	1	-0.22697	0.09984	-2.27	0.0322	17.76563	27.74577	-0.22204	-0.43303	-0.02092
Weight	1	-0.07418	0.05459	-1.36	0.1869	5.60522	9.91059	-0.11597	-0.18685	0.03850
RunPulse	1	-0.36963	0.11985	-3.08	0.0051	38.87574	51.05806	-0.71133	-0.61699	-0.12226
MaxPulse	1	0.30322	0.13650	2.22	0.0360	26.82640	26.49142	0.52161	0.02150	0.58493
RestPulse	1	-0.02153	0.06605	-0.33	0.7473	0.57051	0.57051	-0.03080	-0.15786	0.11480

Figure 55.28. SS1, SS2, STB, CLB, COVB, and CORRB Options: Parameter Estimates

The final two tables are produced as a result of requesting the COVB and CORRB options (Figure 55.29). These tables show the estimated covariance matrix of the parameter estimates, and the estimated correlation matrix of the estimates.

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Covariance of Estimates

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

Correlation of Estimates

Variable	Intercept	RunTime	Age	Weight	RunPulse	MaxPulse	RestPulse
Intercept	1.0000	0.1610	-0.7285	-0.2632	0.1889	-0.4919	-0.1806
RunTime	0.1610	1.0000	-0.3696	-0.2104	-0.1963	0.0881	-0.4297
Age	-0.7285	-0.3696	1.0000	0.1875	-0.1006	0.2629	0.2259
Weight	-0.2632	-0.2104	0.1875	1.0000	0.1474	-0.1842	0.1054
RunPulse	0.1889	-0.1963	-0.1006	0.1474	1.0000	-0.9140	-0.0966
MaxPulse	-0.4919	0.0881	0.2629	-0.1842	-0.9140	1.0000	0.0380
RestPulse	-0.1806	-0.4297	0.2259	0.1054	-0.0966	0.0380	1.0000

Figure 55.29. SS1, SS2, STB, CLB, COVB, and CORRB Options: Covariances and Correlations

For further discussion of the parameters and statistics, see the "Displayed Output" section on page 2989, and Chapter 3, "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 α level can be specified with the ALPHA= option in the PROC REG or MODEL statement.

You can use these statistics in PLOT and PAINT statements. This is useful in performing a variety of regression diagnostics. For definitions of the statistics produced by these options, see Chapter 3, "Introduction to Regression Procedures."

The following example uses the US population data found on the section "Polynomial Regression" beginning on page 2880.

```
data USPop2;
   input Year @@;
   YearSq=Year*Year;
   datalines;
1980 1990 2000
;
data USPop2;
   set USPopulation USPop2;

proc reg data=USPop2;
   id Year;
   model Population=Year YearSq / r cli clm;
run;
```

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

                          Analysis of Variance

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      2         71799          35900    4641.72    <.0001
Error                     16     123.74557        7.73410
Corrected Total           18         71923


            Root MSE              2.78102    R-Square     0.9983
            Dependent Mean       69.76747    Adj R-Sq     0.9981
            Coeff Var             3.98613


                          Parameter Estimates

                       Parameter       Standard
    Variable      DF       Estimate          Error    t Value    Pr > |t|

    Intercept      1          20450      843.47533      24.25      <.0001
    Year           1      -22.78061        0.89785     -25.37      <.0001
    YearSq         1        0.00635     0.00023877      26.58      <.0001
```

Figure 55.30. Regression Using the R, CLI, and CLM Options

The REG Procedure
Model: MODEL1
Dependent Variable: Population

Output Statistics

Obs	Year	Dep Var Population	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict	
1	1790	3.9290	5.0384	1.7289	1.3734	8.7035	-1.9034	11.9803
2	1800	5.3080	5.0389	1.3909	2.0904	7.9874	-1.5528	11.6306
3	1810	7.2390	6.3085	1.1304	3.9122	8.7047	-0.0554	12.6724
4	1820	9.6380	8.8472	0.9571	6.8182	10.8761	2.6123	15.0820
5	1830	12.8660	12.6550	0.8721	10.8062	14.5037	6.4764	18.8335
6	1840	17.0690	17.7319	0.8578	15.9133	19.5504	11.5623	23.9015
7	1850	23.1910	24.0779	0.8835	22.2049	25.9509	17.8920	30.2638
8	1860	31.4430	31.6931	0.9202	29.7424	33.6437	25.4832	37.9029
9	1870	39.8180	40.5773	0.9487	38.5661	42.5885	34.3482	46.8065
10	1880	50.1550	50.7307	0.9592	48.6972	52.7642	44.4944	56.9671
11	1890	62.9470	62.1532	0.9487	60.1420	64.1644	55.9241	68.3823
12	1900	75.9940	74.8448	0.9202	72.8942	76.7955	68.6350	81.0547
13	1910	91.9720	88.8056	0.8835	86.9326	90.6785	82.6197	94.9915
14	1920	105.7100	104.0354	0.8578	102.2169	105.8540	97.8658	110.2051
15	1930	122.7750	120.5344	0.8721	118.6857	122.3831	114.3558	126.7130
16	1940	131.6690	138.3025	0.9571	136.2735	140.3315	132.0676	144.5374
17	1950	151.3250	157.3397	1.1304	154.9434	159.7360	150.9758	163.7036
18	1960	179.3230	177.6460	1.3909	174.6975	180.5945	171.0543	184.2377
19	1970	203.2110	199.2215	1.7289	195.5564	202.8865	192.2796	206.1633
20	1980	.	222.0660	2.1348	217.5404	226.5916	214.6338	229.4983
21	1990	.	246.1797	2.6019	240.6639	251.6955	238.1062	254.2532
22	2000	.	271.5625	3.1257	264.9363	278.1887	262.6932	280.4317

Output Statistics

Obs	Year	Residual	Std Error Residual	Student Residual	-2-1 0 1 2	Cook's D
1	1790	-1.1094	2.178	-0.509	| *| |	0.054
2	1800	0.2691	2.408	0.112	| | |	0.001
3	1810	0.9305	2.541	0.366	| | |	0.009
4	1820	0.7908	2.611	0.303	| | |	0.004
5	1830	0.2110	2.641	0.0799	| | |	0.000
6	1840	-0.6629	2.645	-0.251	| | |	0.002
7	1850	-0.8869	2.637	-0.336	| | |	0.004
8	1860	-0.2501	2.624	-0.0953	| | |	0.000
9	1870	-0.7593	2.614	-0.290	| | |	0.004
10	1880	-0.5757	2.610	-0.221	| | |	0.002
11	1890	0.7938	2.614	0.304	| | |	0.004
12	1900	1.1492	2.624	0.438	| | |	0.008
13	1910	3.1664	2.637	1.201	| |** |	0.054
14	1920	1.6746	2.645	0.633	| |* |	0.014
15	1930	2.2406	2.641	0.848	| |* |	0.026
16	1940	-6.6335	2.611	-2.540	| *****| |	0.289
17	1950	-6.0147	2.541	-2.367	| ****| |	0.370
18	1960	1.6770	2.408	0.696	| |* |	0.054
19	1970	3.9895	2.178	1.831	| |*** |	0.704
20	1980	.	.	.		.
21	1990	.	.	.		.
22	2000	.	.	.		.

Sum of Residuals	-5.8175E-11
Sum of Squared Residuals	123.74557
Predicted Residual SS (PRESS)	188.54924

Figure 55.31. Regression Using the R, CLI, and CLM Options

After producing the usual Analysis of Variance and Parameter Estimates tables (Figure 55.30), the procedure displays the results of requesting the options for predicted and residual values (Figure 55.31). 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 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, a large number of observations with absolute values greater than two indicates an inadequate model. A version of the studentized residual plot can be created on a high-resolution graphics device; see Example 55.7 on page 3019 for a similar example.

Line Printer Scatter Plot Features

This section discusses the special options available with line printer scatter plots. Detailed examples of high resolution graphics plots and options are given in the "55.6" section on page 3017.

Producing Scatter Plots

The interactive PLOT statement available in PROC REG enables you to look at scatter plots of data and diagnostic statistics. These plots can help you to evaluate the model and detect outliers in your data. Several options enable you to place multiple plots on a single page, superimpose plots, and collect plots to be overlaid by later plots. The PAINT statement can be used to highlight points on a plot. See the section "Painting Scatter Plots" on page 2962 for more information on painting.

The Class data set introduced in is used in the following examples.

You can superimpose several plots with the OVERLAY option. With the following statements, a plot of Weight against Height is overlaid with plots of the predicted values and the 95% prediction intervals. The model on which the statistics are based is the full model including Height and Age. These statements produce Output 55.32:

```
proc reg data=Class lineprinter;
   model Weight=Height Age / noprint;
   plot (ucl. lcl. p.)*Height='-' Weight*Height
        / overlay symbol='o';
run;
```

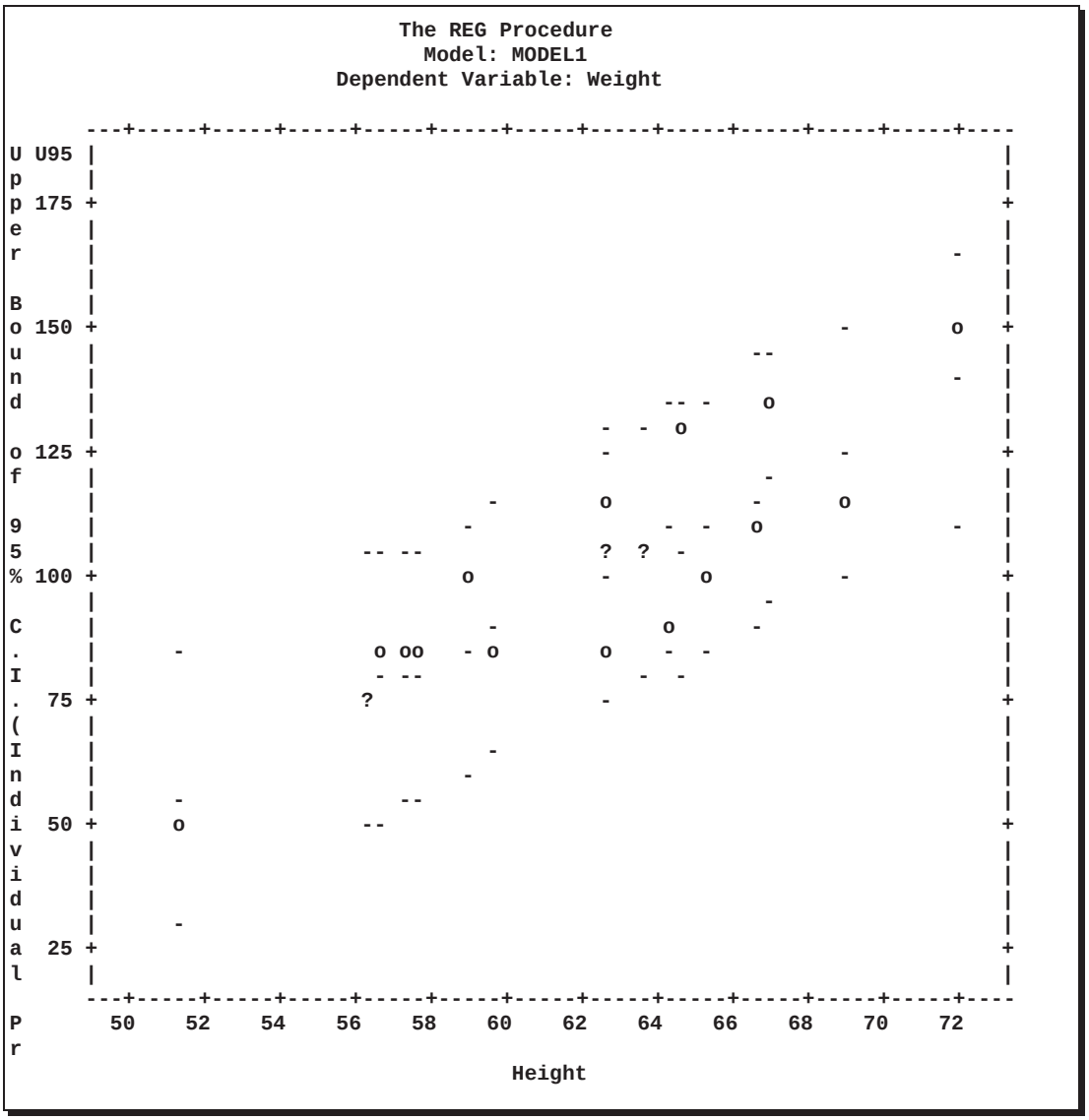


Figure 55.32. Scatter Plot Showing Data, Predicted Values, and Confidence Limits

In this plot, the data values are marked with the symbol 'o' and the predicted values and prediction interval limits are labeled with the symbol '-'. The plot is scaled to accommodate the points from all plots. This is an important difference from the COLLECT option, which does not rescale plots after the first plot or plots are collected. You could separate the overlaid plots by using the following statements:

```
plot;
run;
```

This places each of the four plots on a separate page, while the statements

```
plot / overlay;
run;
```

repeat the previous overlaid plot. In general, the statement

```
plot;
```

is equivalent to respecifying the most recent PLOT statement without any options. However, the COLLECT, HPLOTS=, SYMBOL=, and VPLOTS= options apply across PLOT statements and remain in effect.

The next example shows how you can overlay plots of statistics before and after a change in the model. For the full model involving Height and Age, the ordinary residuals and the studentized residuals are plotted against the predicted values. The COLLECT option causes these plots to be collected or retained for re-display later. The option HPLOTS=2 allows the two plots to appear side by side on one page. The symbol 'f' is used on these plots to identify them as resulting from the full model. These statements produce Figure 55.33:

```
plot r.*p. student.*p. / collect hplots=2 symbol='f';
run;
```

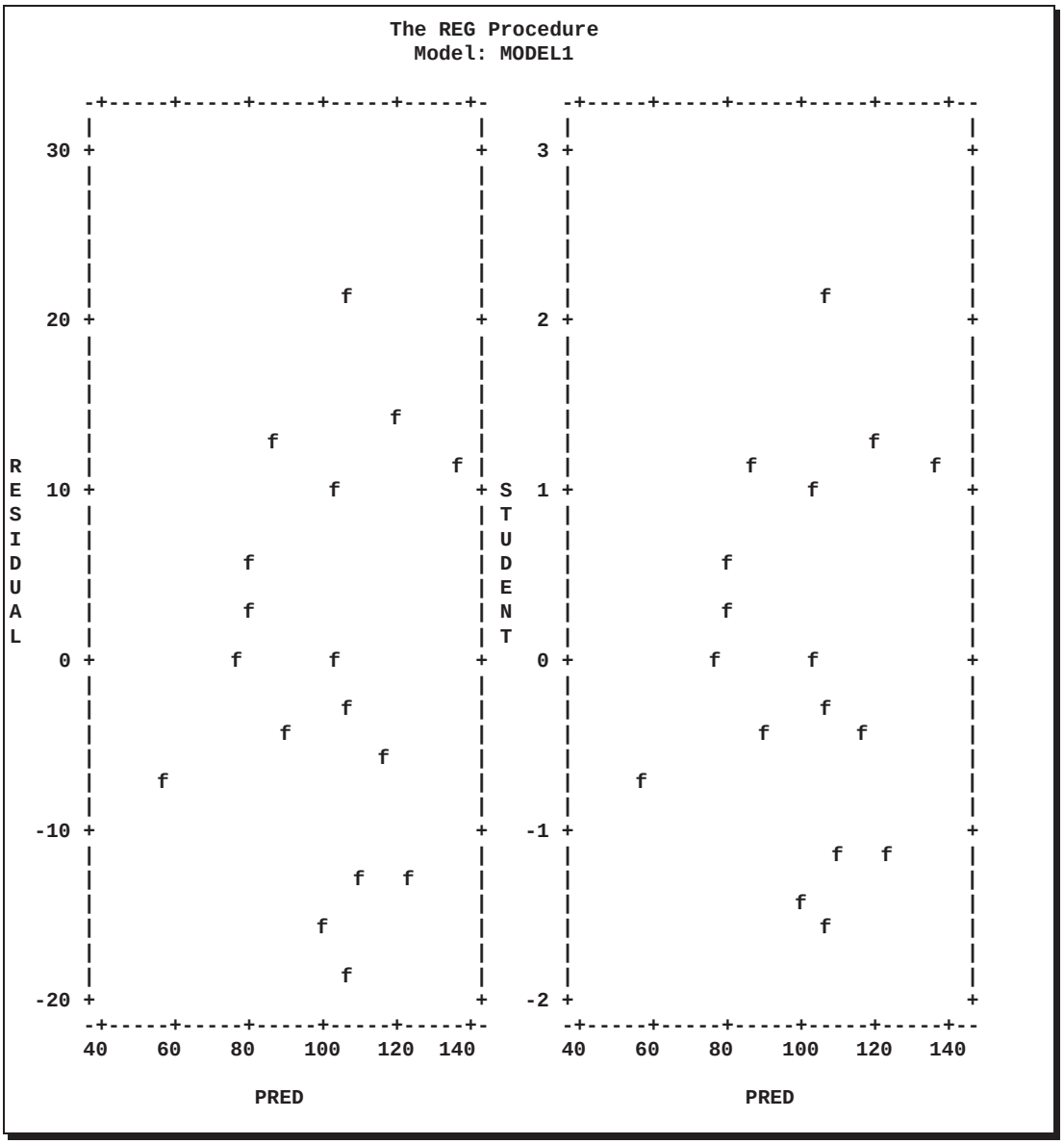


Figure 55.33. Collecting Residual Plots for the Full Model

Note that these plots are not overlaid. The COLLECT option does not overlay the plots in one PLOT statement but retains them so that they can be overlaid by later plots. When the COLLECT option appears in a PLOT statement, the plots in that statement become the first plots in the collection.

Next, the model is reduced by deleting the Age variable. The PLOT statement requests the same plots as before but labels the points with the symbol 'r' denoting the reduced model. The following statements produce Figure 55.34:

```
delete Age;
plot r.*p. student.*p. / symbol='r';
run;
```

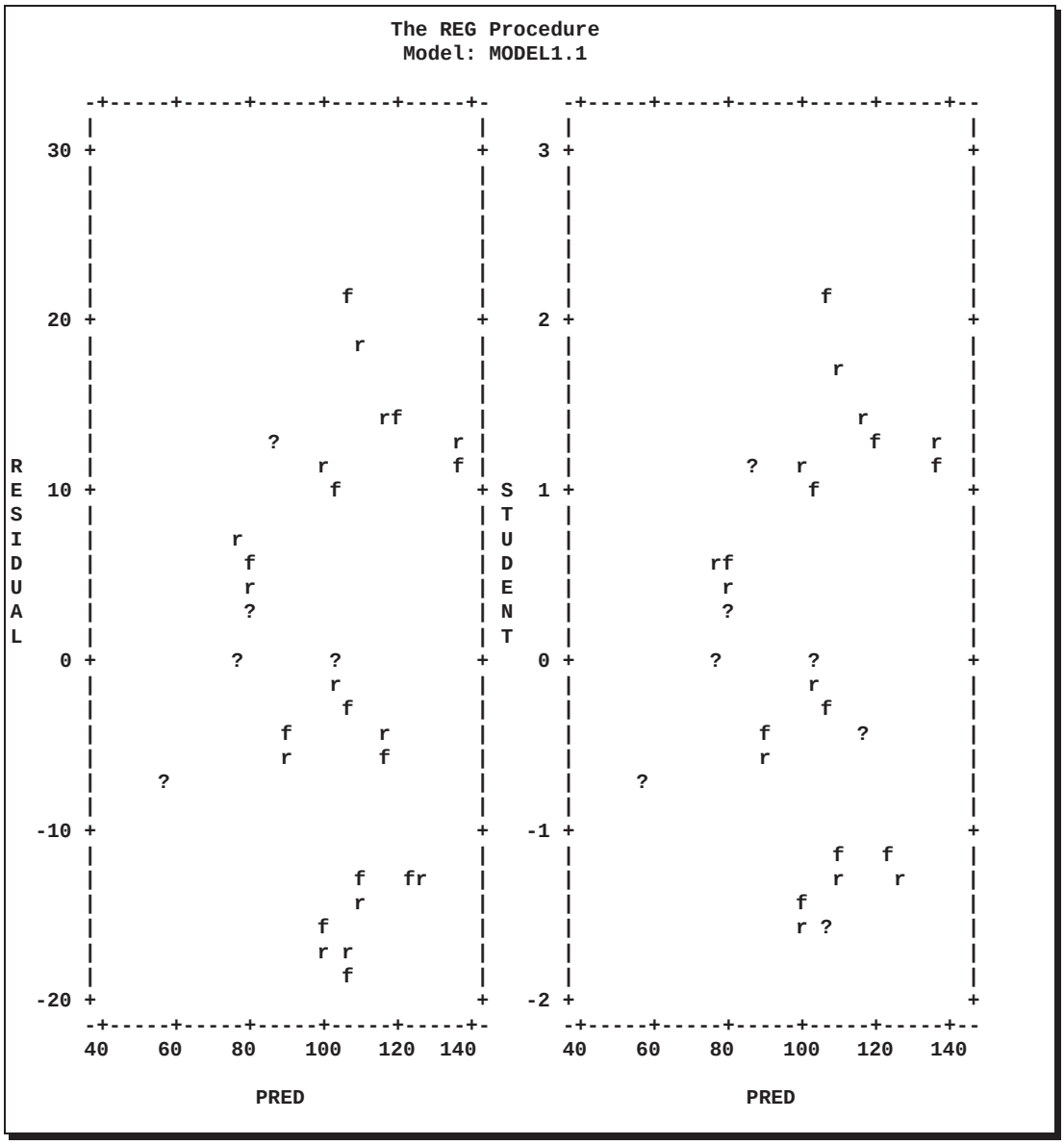


Figure 55.34. Overlaid Residual Plots for Full and Reduced Models

Notice that the COLLECT option causes the corresponding plots to be overlaid. Also notice that the DELETE statement causes the model label to be changed from MODEL1 to MODEL1.1. The points labeled 'f' are from the full model, and points labeled 'r' are from the reduced model. Positions labeled '?' contain at least one point from each model. In this example, the OVERLAY option cannot be used because all of the plots to be overlaid cannot be specified in one PLOT statement. With the COLLECT option, any changes to the model or the data used to fit the model do not affect plots collected before the changes. Collected plots are always reproduced exactly as they first appear. (Similarly, a PAINT statement does not affect plots collected before the PAINT statement is issued.)

The previous example overlays the residual plots for two different models. You may prefer to see them side by side on the same page. This can also be done with the

COLLECT option by using a blank plot. Continuing from the last example, the COLLECT, HPLOTS=2, and SYMBOL='r' options are still in effect. In the following PLOT statement, the CLEAR option deletes the collected plots and allows the specified plot to begin a new collection. The plot created is the residual plot for the reduced model. These statements produce Figure 55.35:

```
plot r.*p. / clear;
run;
```

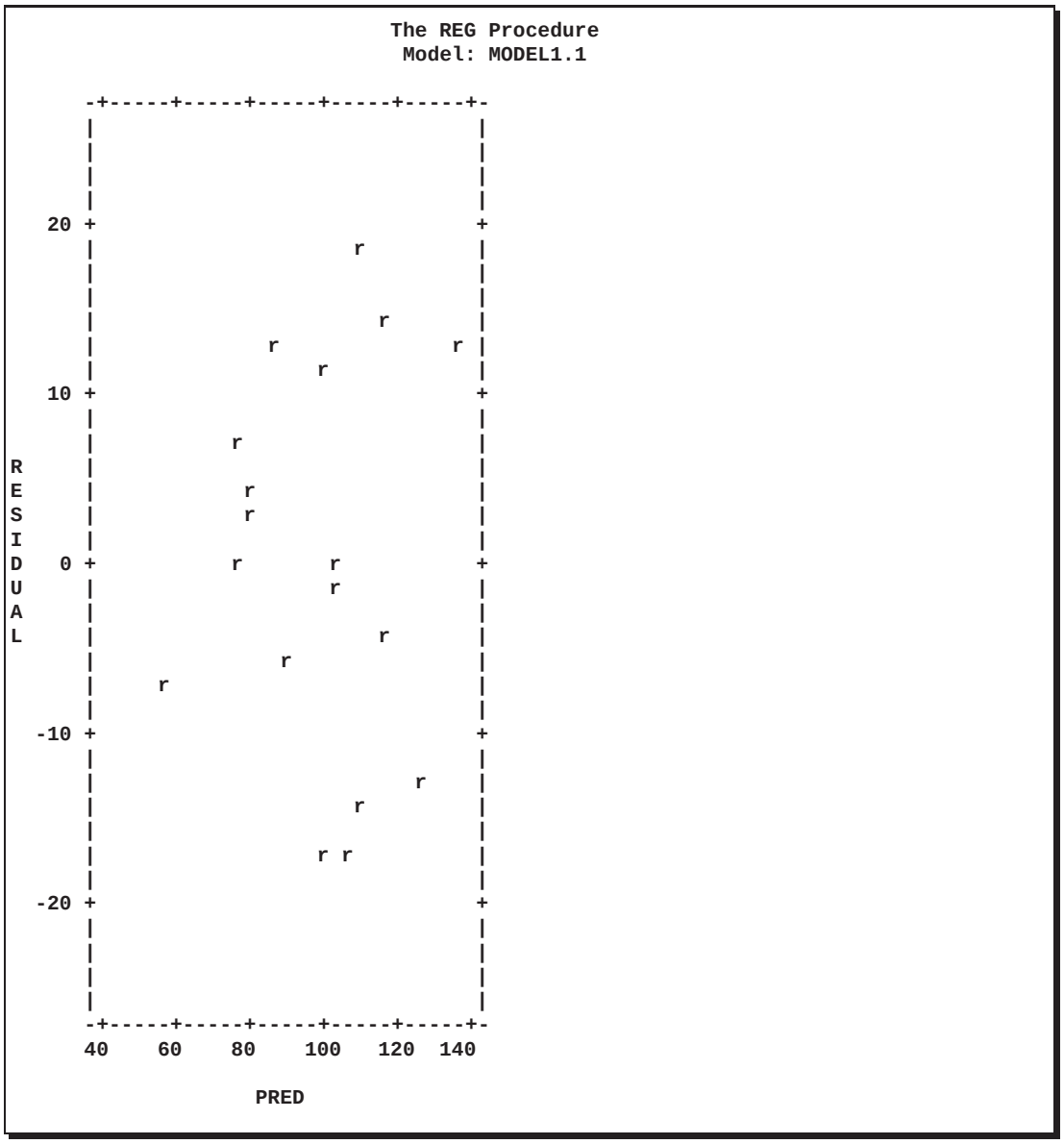


Figure 55.35. Residual Plot for Reduced Model Only

The next statements add the variable AGE to the model and place the residual plot for the full model next to the plot for the reduced model. Notice that a blank plot is created in the first plot request by placing nothing between the quotes. Since the

COLLECT option is in effect, this plot is superimposed on the residual plot for the reduced model. The residual plot for the full model is created by the second request. The result is the desired side-by-side plots. The NOCOLLECT option turns off the collection process after the specified plots are added and displayed. Any PLOT statements that follow show only the newly specified plots. These statements produce Figure 55.36:

```
add Age;
plot r.*p.='' r.*p.='f' / nocollect;
run;
```

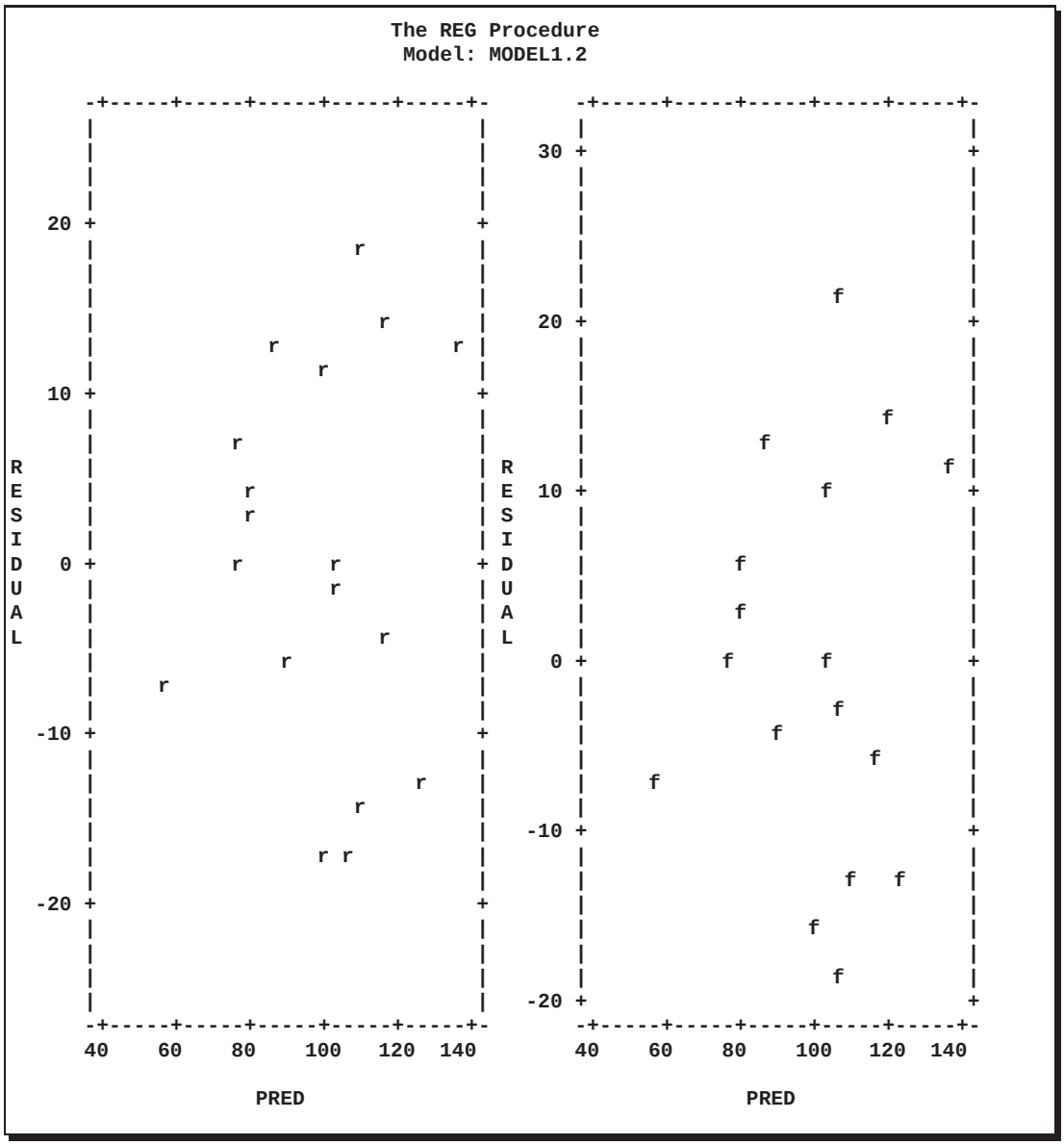


Figure 55.36. Side-by-Side Residual Plots for the Full and Reduced Models

Frequently, when the COLLECT option is in effect, you want the current and following PLOT statements to show only the specified plots. To do this, use both the CLEAR and NOCOLLECT options in the current PLOT statement.

Painting Scatter Plots

Painting scatter plots is a useful interactive tool that enables you to mark points of interest in scatter plots. Painting can be used to identify extreme points in scatter plots or to reveal the relationship between two scatter plots. The CLASS data (from the "Simple Linear Regression" section on page 2877) is used to illustrate some of these applications. First, a scatter plot of the studentized residuals against the predicted values is generated. This plot is shown in Figure 55.37.

```
proc reg data=Class lineprinter;
   model Weight=Age Height / noprint;
   plot student.*p.;
run;
```

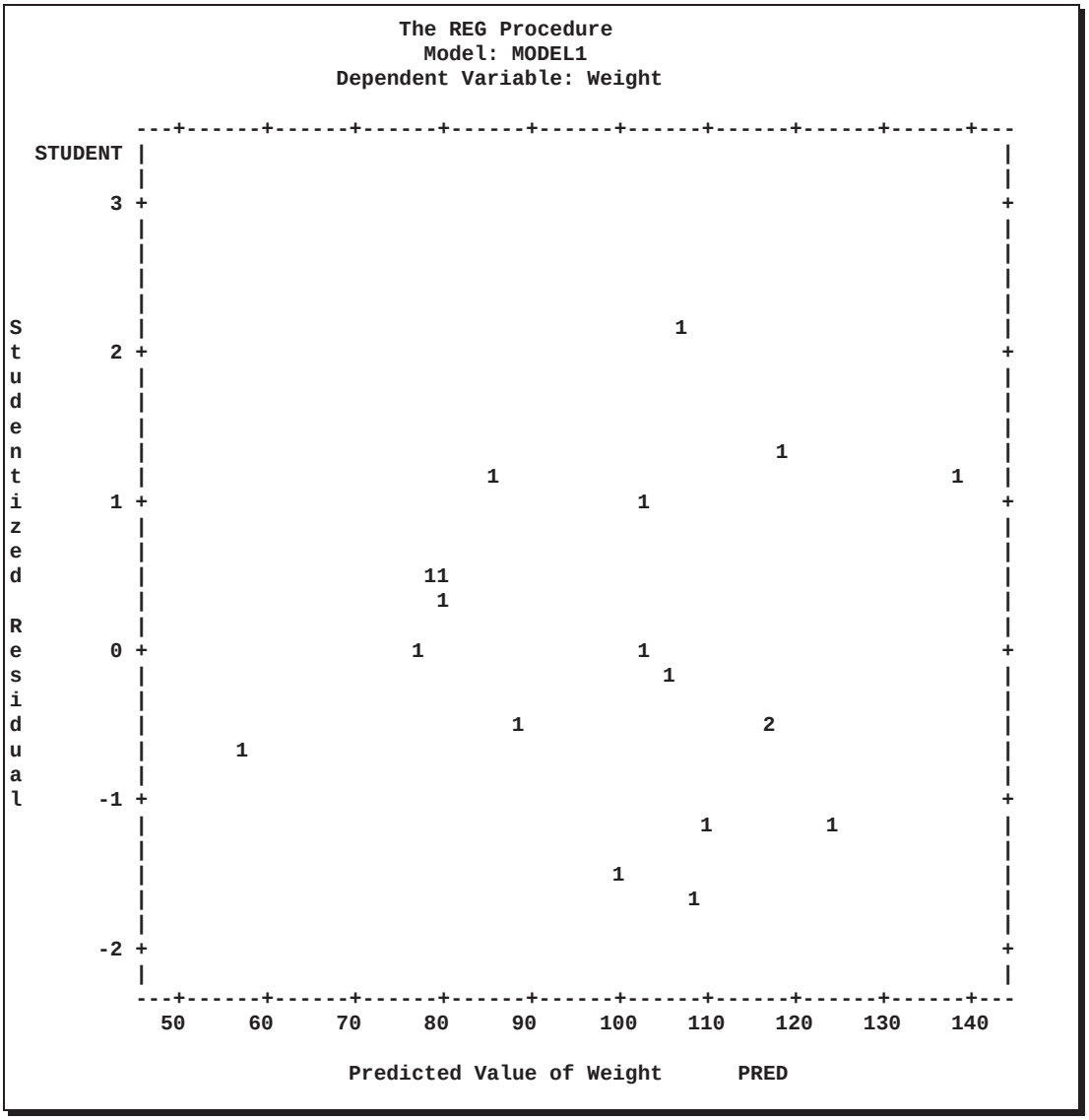


Figure 55.37. Plotting Studentized Residuals Against Predicted Values

Then, the following statements identify the observation 'Henry' in the scatter plot and produce Figure 55.38:

```
paint Name='Henry' / symbol = 'H';
plot;
run;
```

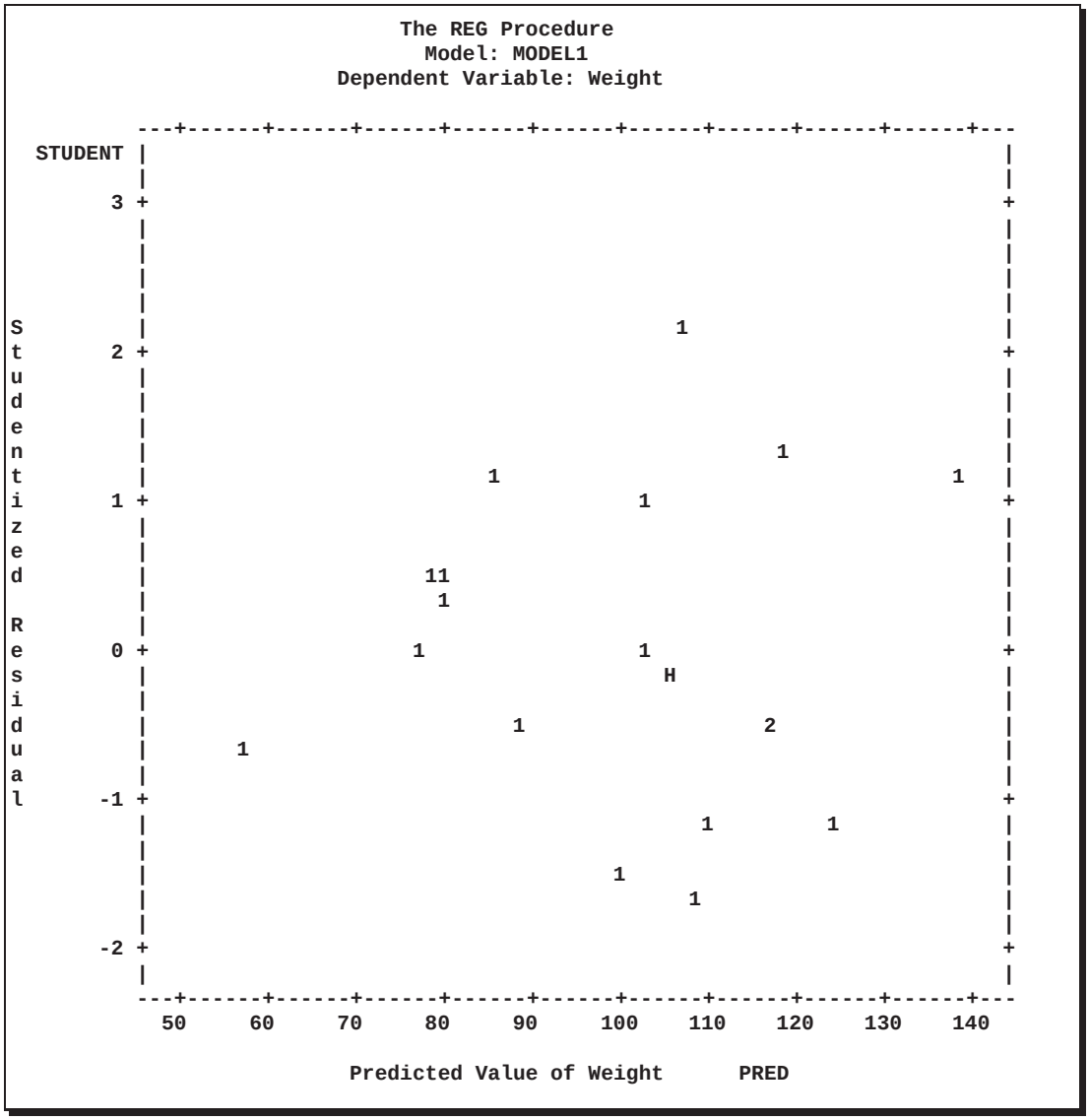


Figure 55.38. Painting One Observation

Next, the following statements identify observations with large absolute residuals:

```
paint student.>=2 or student.<=-2 / symbol='s';
plot;
run;
```

The log shows the observation numbers found with these conditions and gives the painting symbol and the number of observations found. Note that the previous PAINT

statement is also used in the PLOT statement. Figure 55.39 shows the scatter plot produced by the preceding statements.

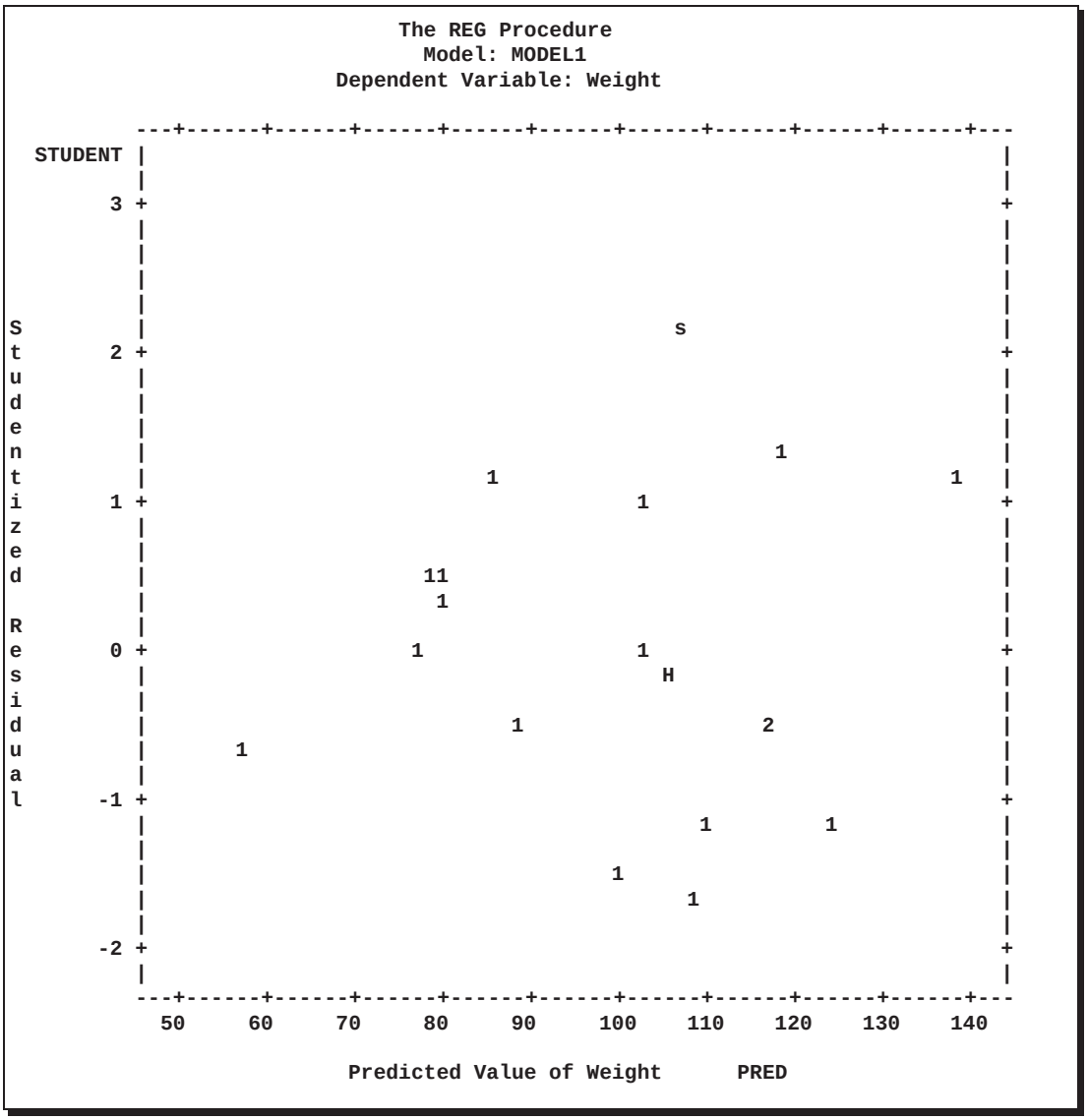


Figure 55.39. Painting Several Observations

The following statements relate two different scatter plots. These statements produce Figure 55.40.

```
paint student.>=1 / symbol='p';
paint student.<1 and student.>-1 / symbol='s';
paint student.<=-1 / symbol='n';
plot student. * p. cookd. * h. / hplots=2;
run;
```

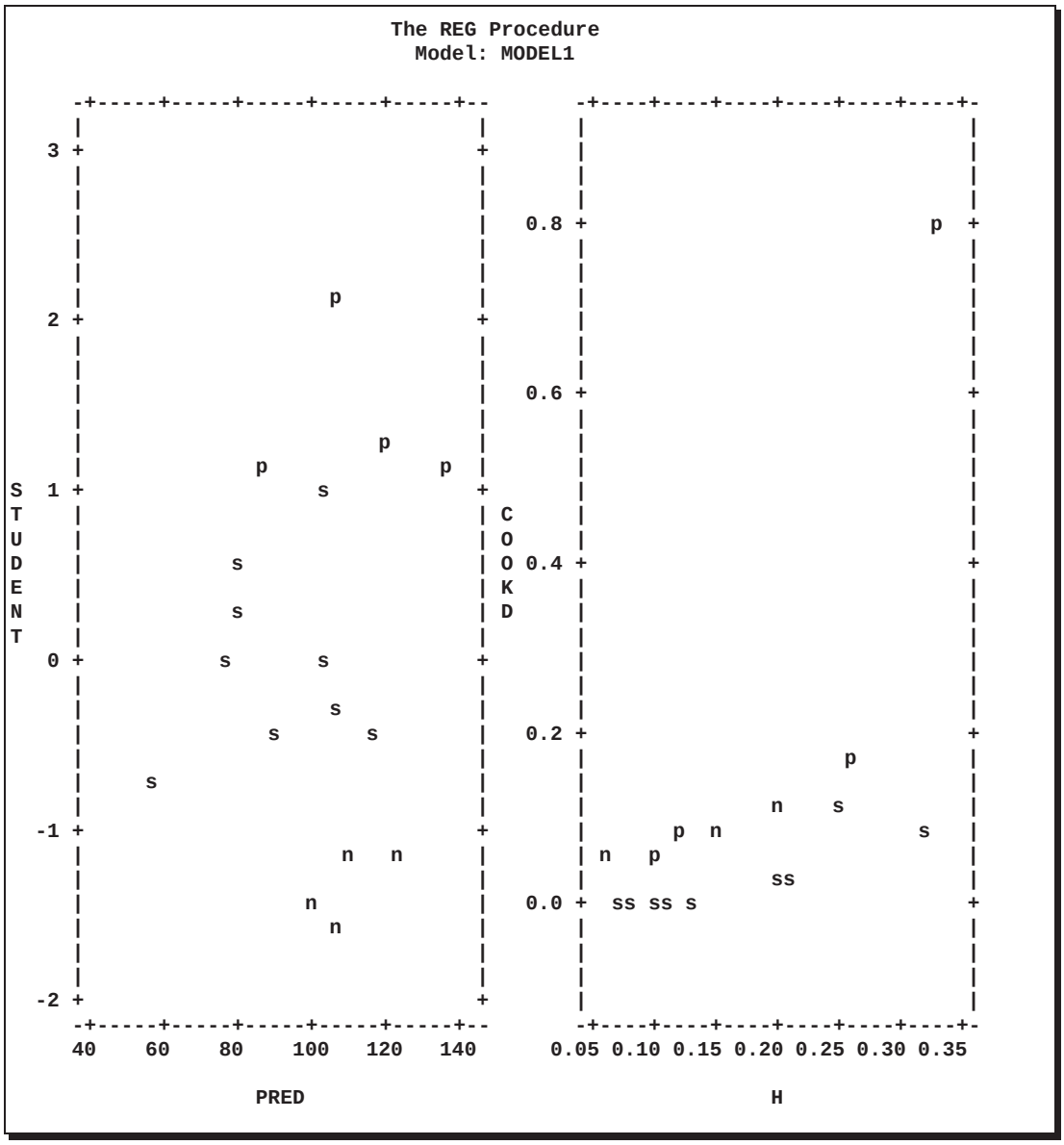


Figure 55.40. Painting Observations on More than One Plot

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 $\mathbf{X}$ multiplied by the true parameters:

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

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 next example uses the fitness data from Example 55.1 on page 2993. 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 55.41 shows how this problem is diagnosed.

```
data fit2;
   set fitness;
   Dif=RunPulse-RestPulse;
proc reg data=fit2;
   model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse Dif;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	722.54361	120.42393	22.43	<.0001
Error	24	128.83794	5.36825		
Corrected Total	30	851.38154			

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

NOTE: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.
NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Dif = RunPulse - RestPulse

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	102.93448	12.40326	8.30	<.0001
RunTime	1	-2.62865	0.38456	-6.84	<.0001
Age	1	-0.22697	0.09984	-2.27	0.0322
Weight	1	-0.07418	0.05459	-1.36	0.1869
RunPulse	B	-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	.	.	.

Figure 55.41. Model that is Not Full Rank: REG Procedure

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, $\mathbf{X}'\mathbf{X}$ 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 $\mathbf{X}'\mathbf{X}$ rather than singular values of $\mathbf{X}$. The eigenvalues of $\mathbf{X}'\mathbf{X}$ are the squares of the singular values of $\mathbf{X}$.

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 $\mathbf{X}$ matrix. Belsey, Kuh, and Welsch (1980) suggest that, when this number is around 10, weak dependencies may be starting to affect the regression estimates. When this number is larger than 100, the estimates may 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 (dependent) 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 55.1 on page 2993. The following statements produce Figure 55.42.

```
proc reg data=fitness;
   model Oxygen=RunTime Age Weight RunPulse MaxPulse RestPulse
         / tol vif collin;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	722.54361	120.42393	22.43	<.0001
Error	24	128.83794	5.36825		
Corrected Total	30	851.38154			

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

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	1	102.93448	12.40326	8.30	<.0001	.	0
RunTime	1	-2.62865	0.38456	-6.84	<.0001	0.62859	1.59087
Age	1	-0.22697	0.09984	-2.27	0.0322	0.66101	1.51284
Weight	1	-0.07418	0.05459	-1.36	0.1869	0.86555	1.15533
RunPulse	1	-0.36963	0.11985	-3.08	0.0051	0.11852	8.43727
MaxPulse	1	0.30322	0.13650	2.22	0.0360	0.11437	8.74385
RestPulse	1	-0.02153	0.06605	-0.33	0.7473	0.70642	1.41559

Collinearity Diagnostics

Number	Eigenvalue	Condition Index	Proportion of Variation: Intercept	RunTime	Age	Weight	RunPulse	MaxPulse	RestPulse
1	6.94991	1.00000	0.00002326	0.00021086	0.00015451	0.00019651	0.00000862	0.00000634	0.00027850
2	0.01868	19.29087	0.00218	0.02522	0.14632	0.01042	0.00000244	0.00000743	0.39064
3	0.01503	21.50072	0.00061541	0.12858	0.15013	0.23571	0.00119	0.00125	0.02809
4	0.00911	27.62115	0.00638	0.60897	0.03186	0.18313	0.00149	0.00123	0.19030
5	0.00607	33.82918	0.00133	0.12501	0.11284	0.44442	0.01506	0.00833	0.36475
6	0.00102	82.63757	0.79966	0.09746	0.49660	0.10330	0.06948	0.00561	0.02026
7	0.00017947	196.78560	0.18981	0.01455	0.06210	0.02283	0.91277	0.98357	0.00568

Figure 55.42. Regression Using the TOL, VIF, and COLLIN Options

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 $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, and the parameter estimate is denoted by $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$. The subscript i denotes values for the ith observation, the parenthetical subscript (i) means that the statistic is computed using all observations except the ith observation, and the subscript jj indicates the jth diagonal matrix entry. The ALPHA= option in the PROC REG or MODEL statement is used to set the α value for the t statistics.

Table 55.5 contains the summary statistics for assessing the fit of the model.

Table 55.5. Formulas and Definitions for Model Fit Summary Statistics

MODEL Option or Statistic	Definition or Formula
n	the number of observations
p	the number of parameters including the intercept
i	1 if there is an intercept, 0 otherwise
$\hat{\sigma}^2$	the estimate of pure error variance from the SIGMA= option or from fitting the full model
SST_0	the uncorrected total sum of squares for the dependent variable
SST_1	the total sum of squares corrected for the mean for the dependent variable
SSE	the error sum of squares
MSE	$\frac{\text{SSE}}{n-p}$
R^2	$1-\frac{\text{SSE}}{\text{SST}_i}$
ADJRSQ	$1-\frac{(n-i)(1-R^2)}{n-p}$
AIC	$n\ln\left(\frac{\text{SSE}}{n}\right)+2p$
BIC	$n\ln\left(\frac{\text{SSE}}{n}\right)+2(p+2)q-2q^2$ where $q=\frac{n\hat{\sigma}^2}{\text{SSE}}$
CP (C_p)	$\frac{\text{SSE}}{\hat{\sigma}^2}+2p-n$
GMSEP	$\frac{\text{MSE}(n+1)(n-2)}{n(n-p-1)}=\frac{1}{n}S_p(n+1)(n-2)$
JP (J_p)	$\frac{n+p}{n}\text{MSE}$
PC	$\frac{n+p}{n-p}(1-R^2)=J_p\left(\frac{n}{\text{SST}_i}\right)$
PRESS	the sum of squares of $predr_i$ (see Table 55.6)
RMSE	$\sqrt{\text{MSE}}$
SBC	$n\ln\left(\frac{\text{SSE}}{n}\right)+p\ln(n)$
SP (S_p)	$\frac{\text{MSE}}{n-p-1}$

Table 55.6 contains the diagnostic statistics and their formulas; these formulas and further information can be found in Chapter 3, "Introduction to Regression Procedures," and in the "Influence Diagnostics" section on page 2970. Each statistic is computed for each observation.

Table 55.6. Formulas and Definitions for Diagnostic Statistics

MODEL Option or Statistic	Formula
PRED ($\widehat{Y}_i$)	$\mathbf{X}_i\mathbf{b}$
RES (r_i)	$\mathbf{Y}_i - \widehat{Y}_i$
H (h_i)	$x_i(\mathbf{X}'\mathbf{X})^{-}\mathbf{x}_i'$
STDP	$\sqrt{h_i\widehat{\sigma}^2}$
STDI	$\sqrt{(1+h_i)\widehat{\sigma}^2}$
STDR	$\sqrt{(1-h_i)\widehat{\sigma}^2}$
LCL	$\widehat{Y}_i - t_{\frac{\alpha}{2}}\text{STDP}$
LCLM	$\widehat{Y}_i - t_{\frac{\alpha}{2}}\text{STDI}$
UCL	$\widehat{Y}_i + t_{\frac{\alpha}{2}}\text{STDP}$
UCLM	$\widehat{Y}_i + t_{\frac{\alpha}{2}}\text{STDI}$
STUDENT	$\frac{r_i}{\text{STDR}_i}$
RSTUDENT	$\frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$
COOKD	$\frac{1}{p}\text{STUDENT}^2(\frac{\text{STDP}}{\text{STDR}^2})$
COVRATIO	$\frac{\det(\hat{\sigma}^2_{(i)}(\mathbf{X}'_{(i)}\mathbf{x}_{(i)})^{-1}}{\det(\hat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1})}$
DFFITS	$\frac{(\widehat{Y}_i - \widehat{Y}_{(i)})}{(\hat{\sigma}_{(i)}\sqrt{h_i})}$
DFBETAS$_j$	$\frac{\mathbf{b}_j - \mathbf{b}_{(i)j}}{\hat{\sigma}_{(i)}\sqrt{(\mathbf{X}'\mathbf{X})_{jj}}}$
PRESS($predr_i$)	$\frac{r_i}{1-h_i}$

Influence Diagnostics

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

$$h_i = \mathbf{x}_i(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i'$$

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

For each observation, PROC REG first displays the residual, the studentized residual (RSTUDENT), and the h_i. The studentized residual RSTUDENT differs slightly from STUDENT since the error variance is estimated by $s^2_{(i)}$ without the ith 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 may need some attention.

The COVRATIO statistic measures the change in the determinant of the covariance matrix of the estimates by deleting the ith observation:

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

Belsley, Kuh, and Welsch 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 ith observation and is calculated by deleting the ith observation. A large value indicates that the observation is very influential in its neighborhood of the $\mathbf{X}$ 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 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 2952.

The DFBETAS statistics are the scaled measures of the change in each parameter estimate and are calculated by deleting the ith observation:

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

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

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

Figure 55.43 shows the tables produced by the INFLUENCE option for the population example (the section "Polynomial Regression" beginning on page 2880). See Figure 55.30 for the fitted regression equation.

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

The REG Procedure
Model: MODEL1
Dependent Variable: Population

Output Statistics

Obs	Residual	RStudent	Hat Diag H	Cov Ratio	DFFITS	DFBETAS Intercept	DFBETAS Year	DFBETAS YearSq
1	-1.1094	-0.4972	0.3865	1.8834	-0.3946	-0.2842	0.2810	-0.2779
2	0.2691	0.1082	0.2501	1.6147	0.0625	0.0376	-0.0370	0.0365
3	0.9305	0.3561	0.1652	1.4176	0.1584	0.0666	-0.0651	0.0636
4	0.7908	0.2941	0.1184	1.3531	0.1078	0.0182	-0.0172	0.0161
5	0.2110	0.0774	0.0983	1.3444	0.0256	-0.0030	0.0033	-0.0035
6	-0.6629	-0.2431	0.0951	1.3255	-0.0788	0.0296	-0.0302	0.0307
7	-0.8869	-0.3268	0.1009	1.3214	-0.1095	0.0609	-0.0616	0.0621
8	-0.2501	-0.0923	0.1095	1.3605	-0.0324	0.0216	-0.0217	0.0218
9	-0.7593	-0.2820	0.1164	1.3519	-0.1023	0.0743	-0.0745	0.0747
10	-0.5757	-0.2139	0.1190	1.3650	-0.0786	0.0586	-0.0587	0.0587
11	0.7938	0.2949	0.1164	1.3499	0.1070	-0.0784	0.0783	-0.0781
12	1.1492	0.4265	0.1095	1.3144	0.1496	-0.1018	0.1014	-0.1009
13	3.1664	1.2189	0.1009	1.0168	0.4084	-0.2357	0.2338	-0.2318
14	1.6746	0.6207	0.0951	1.2430	0.2013	-0.0811	0.0798	-0.0784
15	2.2406	0.8407	0.0983	1.1724	0.2776	-0.0427	0.0404	-0.0380
16	-6.6335	-3.1845	0.1184	0.2924	-1.1673	-0.1531	0.1636	-0.1747
17	-6.0147	-2.8433	0.1652	0.3989	-1.2649	-0.4843	0.4958	-0.5076
18	1.6770	0.6847	0.2501	1.4757	0.3954	0.2240	-0.2274	0.2308
19	3.9895	1.9947	0.3865	0.9766	1.5831	1.0902	-1.1025	1.1151

Sum of Residuals	-5.8175E-11
Sum of Squared Residuals	123.74557
Predicted Residual SS (PRESS)	188.54924

Figure 55.43. Regression Using the INFLUENCE Option

In Figure 55.43, 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.

The PARTIAL Option

The PARTIAL option in the MODEL statement produces partial regression leverage plots. This option requires the use of the LINEPRINTER option in the PROC REG statement since high resolution partial regression plots are not currently supported. One plot is created for each regressor in the full, current 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.

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.

In the plot, points are marked by the number of replicates appearing at one position. The symbol '*' is used if there are ten or more replicates. If an ID statement is specified, the left-most nonblank character in the value of the ID variable is used as the plotting symbol.

The following statements use the fitness data in Example 55.1 on page 2993 with the PARTIAL option to produce the partial regression leverage plots in the OUTPUT window. The plots are not shown.

```
proc reg data=fitness lineprinter;
   model Oxygen=RunTime Weight Age / partial;
run;
```

The following statements create one of the partial regression plots on a high resolution graphics device for the fitness data; all four plots (created by regressing Oxygen and one of the variables on the remaining variables) are displayed in Figure 55.44. Notice that the Int variable is explicitly added to be used as the intercept term.

```
data fitness2;
   set fitness;
   Int=1;
proc reg data=fitness2 noprint;
   model Oxygen Int = RunTime Weight Age / noint;
   output out=temp r=ry rx;
symbol1 c=blue;
```

```
proc gplot data=temp;
   plot ry*rx / cframe=ligr;
   label ry='Oxygen'
         rx='Intercept';
run;
```

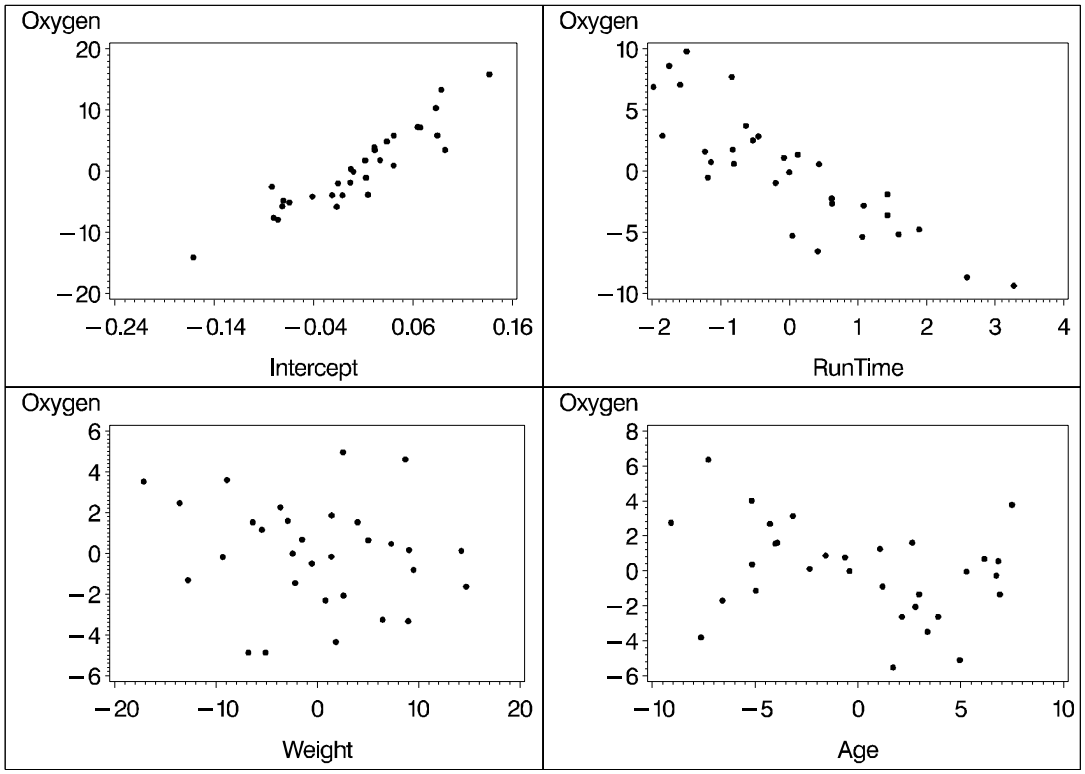
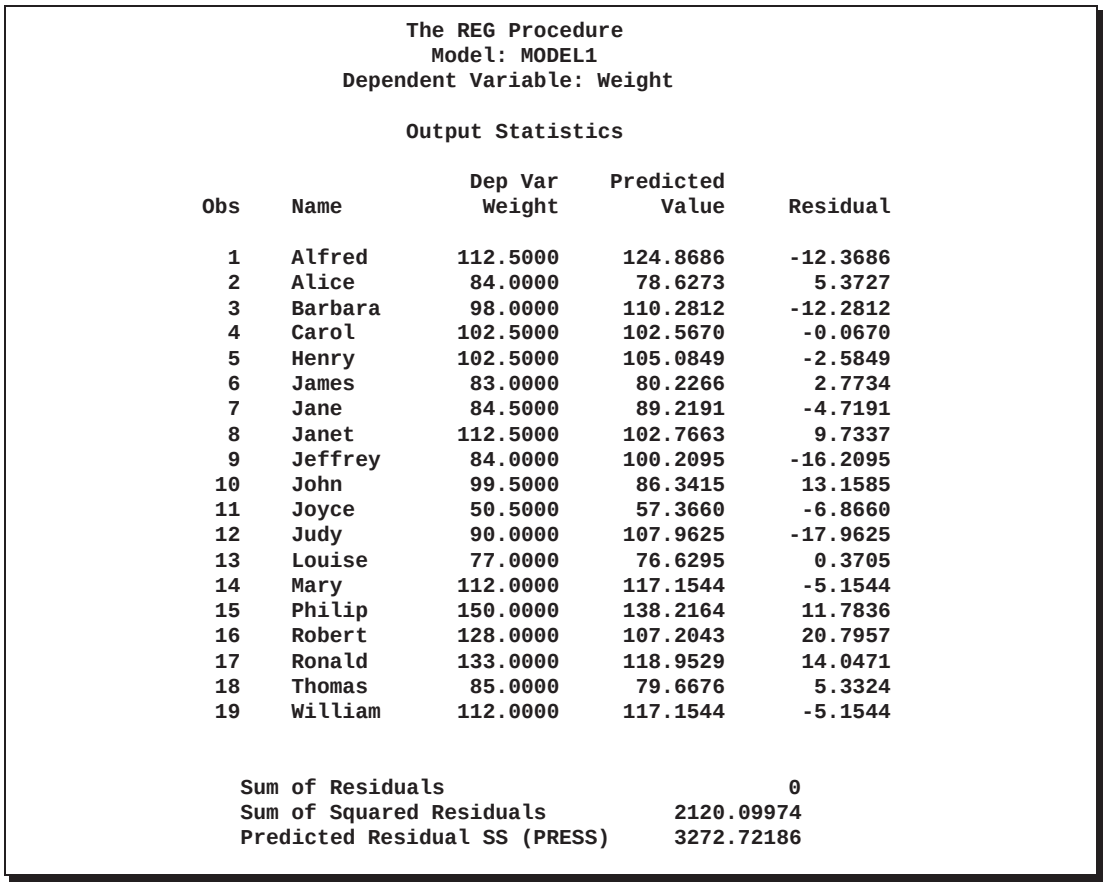


Figure 55.44. Partial Regression Leverage Plots

Reweighting Observations in an Analysis

Reweighting observations is an interactive feature of PROC REG that enables you to change the weights of observations used in computing the regression equation. Observations can also be deleted from the analysis (not from the data set) by changing their weights to zero. The Class data (in the "Getting Started" section on page 2877) 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 55.45.

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



The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Output Statistics

Obs	Name	Dep Var Weight	Predicted Value	Residual
1	Alfred	112.5000	124.8686	-12.3686
2	Alice	84.0000	78.6273	5.3727
3	Barbara	98.0000	110.2812	-12.2812
4	Carol	102.5000	102.5670	-0.0670
5	Henry	102.5000	105.0849	-2.5849
6	James	83.0000	80.2266	2.7734
7	Jane	84.5000	89.2191	-4.7191
8	Janet	112.5000	102.7663	9.7337
9	Jeffrey	84.0000	100.2095	-16.2095
10	John	99.5000	86.3415	13.1585
11	Joyce	50.5000	57.3660	-6.8660
12	Judy	90.0000	107.9625	-17.9625
13	Louise	77.0000	76.6295	0.3705
14	Mary	112.0000	117.1544	-5.1544
15	Philip	150.0000	138.2164	11.7836
16	Robert	128.0000	107.2043	20.7957
17	Ronald	133.0000	118.9529	14.0471
18	Thomas	85.0000	79.6676	5.3324
19	William	112.0000	117.1544	-5.1544

Sum of Residuals	0
Sum of Squared Residuals	2120.09974
Predicted Residual SS (PRESS)	3272.72186

Figure 55.45. Full Model for CLASS Data, Residuals Shown

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.

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

At this point, a message (on the log) appears that tells you which observations have been reweighted and what the new weights are. Figure 55.46 is produced.

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

Output Statistics

Obs	Name	Weight Variable	Dep Var Weight	Predicted Value	Residual
1	Alfred	1.0000	112.5000	121.6250	-9.1250
2	Alice	1.0000	84.0000	79.9296	4.0704
3	Barbara	1.0000	98.0000	107.5484	-9.5484
4	Carol	1.0000	102.5000	102.1663	0.3337
5	Henry	1.0000	102.5000	104.3632	-1.8632
6	James	1.0000	83.0000	79.9762	3.0238
7	Jane	1.0000	84.5000	87.8225	-3.3225
8	Janet	1.0000	112.5000	103.6889	8.8111
9	Jeffrey	1.0000	84.0000	98.7606	-14.7606
10	John	1.0000	99.5000	85.3117	14.1883
11	Joyce	1.0000	50.5000	58.6811	-8.1811
12	Judy	0.5000	90.0000	106.8740	-16.8740
13	Louise	1.0000	77.0000	76.8377	0.1623
14	Mary	1.0000	112.0000	116.2429	-4.2429
15	Philip	1.0000	150.0000	135.9688	14.0312
16	Robert	0.5000	128.0000	103.5150	24.4850
17	Ronald	0	133.0000	117.8121	15.1879
18	Thomas	1.0000	85.0000	78.1398	6.8602
19	William	1.0000	112.0000	116.2429	-4.2429

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

NOTE: The above statistics use observation weights or frequencies.

Figure 55.46. Model with Reweighted Observations

The first REWEIGHT statement excludes observation 17, and the second REWEIGHT statement reweights observations 12 and 16 to 0.5. An important feature to note from this example is that the model is not refit until after the PRINT statement. REWEIGHT statements do not cause the model to be refit. This is so that multiple REWEIGHT statements can be applied to a subsequent model.

In this example, since the intent is to reweight observations with large residuals, the observation that was mistakenly included in the analysis should be deleted; then, the model should be fit for those remaining observations, and the observations with large residuals should be reweighted. To accomplish this, use the REFIT statement. Note that the model label has been changed from MODEL1 to MODEL1.2 as two REWEIGHT statements have been used. These statements produce Figure 55.47:

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

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

Output Statistics

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

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

NOTE: The above statistics use observation weights or frequencies.

Figure 55.47. Observations Excluded from Analysis, Model Refitted and Observations Reweighted

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 55.48:

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

The REG Procedure
Model: MODEL1.6
Dependent Variable: Weight

Output Statistics

Obs	Name	Dep Var Weight	Predicted Value	Residual
1	Alfred	112.5000	124.8686	-12.3686
2	Alice	84.0000	78.6273	5.3727
3	Barbara	98.0000	110.2812	-12.2812
4	Carol	102.5000	102.5670	-0.0670
5	Henry	102.5000	105.0849	-2.5849
6	James	83.0000	80.2266	2.7734
7	Jane	84.5000	89.2191	-4.7191
8	Janet	112.5000	102.7663	9.7337
9	Jeffrey	84.0000	100.2095	-16.2095
10	John	99.5000	86.3415	13.1585
11	Joyce	50.5000	57.3660	-6.8660
12	Judy	90.0000	107.9625	-17.9625
13	Louise	77.0000	76.6295	0.3705
14	Mary	112.0000	117.1544	-5.1544
15	Philip	150.0000	138.2164	11.7836
16	Robert	128.0000	107.2043	20.7957
17	Ronald	133.0000	118.9529	14.0471
18	Thomas	85.0000	79.6676	5.3324
19	William	112.0000	117.1544	-5.1544

Sum of Residuals	0
Sum of Squared Residuals	2120.09974
Predicted Residual SS (PRESS)	3272.72186

Figure 55.48. Restoring Weights of All Observations

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

Now suppose you want only to undo the changes made by the most recent REWEIGHT statement. Use REWEIGHT UNDO for this. The following statements produce Figure 55.49:

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

```
                          The REG Procedure
                            Model: MODEL1.9
                      Dependent Variable: Weight

                            Output Statistics

                          Weight      Dep Var    Predicted
        Obs    Name       Variable     Weight       Value     Residual

          1    Alfred       0.7500   112.5000    125.1152    -12.6152
          2    Alice        1.0000    84.0000     78.7691      5.2309
          3    Barbara      0.7500    98.0000    110.3236    -12.3236
          4    Carol        1.0000   102.5000    102.8836     -0.3836
          5    Henry        1.0000   102.5000    105.3936     -2.8936
          6    James        1.0000    83.0000     80.1133      2.8867
          7    Jane         1.0000    84.5000     89.0776     -4.5776
          8    Janet        1.0000   112.5000    103.3322      9.1678
          9    Jeffrey      0.7500    84.0000    100.2835    -16.2835
         10    John         0.7500    99.5000     86.2090     13.2910
         11    Joyce        1.0000    50.5000     57.0745     -6.5745
         12    Judy         0.7500    90.0000    108.2622    -18.2622
         13    Louise       1.0000    77.0000     76.5275      0.4725
         14    Mary         1.0000   112.0000    117.6752     -5.6752
         15    Philip       1.0000   150.0000    138.9211     11.0789
         16    Robert       0.7500   128.0000    107.0063     20.9937
         17    Ronald       0.7500   133.0000    119.4681     13.5319
         18    Thomas       1.0000    85.0000     79.3061      5.6939
         19    William      1.0000   112.0000    117.6752     -5.6752


                  Sum of Residuals                          0
                  Sum of Squared Residuals           1694.87114
                  Predicted Residual SS (PRESS)      2547.22751

NOTE: The above statistics use observation weights or frequencies.
```

Figure 55.49. Example of UNDO in REWEIGHT Statement

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 55.50:

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

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

Output Statistics

Obs	Name	Weight Variable	Dep Var Weight	Predicted Value	Residual
1	Alfred	0.7500	112.5000	126.0076	-13.5076
2	Alice	1.0000	84.0000	77.8727	6.1273
3	Barbara	0.7500	98.0000	111.2805	-13.2805
4	Carol	1.0000	102.5000	102.4703	0.0297
5	Henry	1.0000	102.5000	105.1278	-2.6278
6	James	1.0000	83.0000	80.2290	2.7710
7	Jane	1.0000	84.5000	89.7199	-5.2199
8	Janet	1.0000	112.5000	102.0122	10.4878
9	Jeffrey	0.7500	84.0000	100.6507	-16.6507
10	John	0.7500	99.5000	86.6828	12.8172
11	Joyce	1.0000	50.5000	56.7703	-6.2703
12	Judy	1.0000	90.0000	108.1649	-18.1649
13	Louise	1.0000	77.0000	76.4327	0.5673
14	Mary	1.0000	112.0000	117.1975	-5.1975
15	Philip	1.0000	150.0000	138.7581	11.2419
16	Robert	1.0000	128.0000	108.7016	19.2984
17	Ronald	0.7500	133.0000	119.0957	13.9043
18	Thomas	1.0000	85.0000	80.3076	4.6924
19	William	1.0000	112.0000	117.1975	-5.1975

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

NOTE: The above statistics use observation weights or frequencies.

Figure 55.50. REWEIGHT Statement with RESET option

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 twelve REWEIGHT statements have been applied to the original model.

Testing for Heteroscedasticity

The regression model is specified as $y_i = \mathbf{x}_i\beta + \epsilon_i$, where the ϵ_i's are identically and independently distributed: $E(\epsilon) = 0$ and $E(\epsilon'\epsilon) = \sigma^2\mathbf{I}$. If the ϵ_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, the ACOV option provides a consistent estimate of the covariance matrix. If the regression data are from a simple random sample, the ACOV option produces the covariance matrix. This matrix is

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

where

$$e_i = y_i - \mathbf{x}_i\mathbf{b}$$

The SPEC option performs a model specification test. The null hypothesis for this test maintains that the errors are homoscedastic, 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 option, tests listed in the TEST statement are performed with both the usual covariance matrix and the heteroscedasticity consistent covariance matrix. 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.

Multivariate Tests

The MTEST statement described in the "MTEST Statement" section on page 2907 can test hypotheses involving several dependent variables in the form

$$(\mathbf{L}\beta - \mathbf{cj})\mathbf{M} = 0$$

where $\mathbf{L}$ is a linear function on the regressor side, β is a matrix of parameters, $\mathbf{c}$ is a column vector of constants, $\mathbf{j}$ is a row vector of ones, and $\mathbf{M}$ is a linear function on the dependent side. The special case where the constants are zero is

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

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

$$\begin{aligned}\mathbf{H} &= \mathbf{M}'(\mathbf{LB}-\mathbf{cj})'(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-}\mathbf{L}')^{-1}(\mathbf{LB}-\mathbf{cj})\mathbf{M}\\ \mathbf{E} &= \mathbf{M}'(\mathbf{Y}'\mathbf{Y}-\mathbf{B}'(\mathbf{X}'\mathbf{X})\mathbf{B})\mathbf{M}\end{aligned}$$

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

Four test statistics based on the eigenvalues of $\mathbf{E}^{-1}\mathbf{H}$ or $(\mathbf{E}+\mathbf{H})^{-1}\mathbf{H}$ are formed. These are Wilks' Lambda, Pillai's Trace, the Hotelling-Lawley Trace, and Roy's maximum root. These test statistics are discussed in Chapter 3, "Introduction to Regression Procedures."

The following statements perform a multivariate analysis of variance and produce Figures 55.51 through 55.55:

```
* Manova Data from Morrison (1976, 190);
  data a;
     input sex $ drug $ @;
     do rep=1 to 4;
        input y1 y2 @;
        sexcode=(sex='m')-(sex='f');
        drug1=(drug='a')-(drug='c');
        drug2=(drug='b')-(drug='c');
        sexdrug1=sexcode*drug1;
        sexdrug2=sexcode*drug2;
        output;
     end;
     datalines;
  m a  5  6  5  4  9  9  7  6
  m b  7  6  7  7  9 12  6  8
  m c 21 15 14 11 17 12 12 10
  f a  7 10  6  6  9  7  8 10
  f b 10 13  8  7  7  6  6  9
  f c 16 12 14  9 14  8 10  5
  ;
  proc reg;
     model y1 y2=sexcode drug1 drug2 sexdrug1 sexdrug2;
     y1y2drug: mtest y1=y2, drug1,drug2;
     drugshow: mtest drug1, drug2 / print canprint;
  run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	316.00000	63.20000	12.04	<.0001
Error	18	94.50000	5.25000		
Corrected Total	23	410.50000			

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

Parameter Estimates

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 55.51. Multivariate Analysis of Variance: REG Procedure

The REG Procedure
Model: MODEL1
Dependent Variable: y2

Analysis of Variance

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

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 55.52. Multivariate Analysis of Variance: REG Procedure

The REG Procedure
Model: MODEL1
Multivariate Test: Y1Y2DRUG

Multivariate Statistics and Exact F Statistics

S=1 M=0 N=8

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.28053917	23.08	2	18	<.0001
Pillai's Trace	0.71946083	23.08	2	18	<.0001
Hotelling-Lawley Trace	2.56456456	23.08	2	18	<.0001
Roy's Greatest Root	2.56456456	23.08	2	18	<.0001

Figure 55.53. Multivariate Analysis of Variance: First Test

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.

```
                          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
```

Figure 55.54. Multivariate Analysis of Variance: Second Test

```
                          The REG Procedure
                            Model: MODEL1
                        Multivariate Test: DRUGSHOW

                  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.
```

Figure 55.55. Multivariate Analysis of Variance: Second Test (continued)

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.

Autocorrelation in Time Series Data

When regression is performed on time series data, the errors may 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 may 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 DW option for the US population data (see Figure 55.56):

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

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

              Durbin-Watson D                1.264
              Number of Observations            19
              1st Order Autocorrelation      0.299
```

Figure 55.56. Regression Using DW Option

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

$$\mathbf{D}^{-\frac{1}{2}}(\mathbf{Z}'\mathbf{Z} + k\mathbf{I}_p)^{-1}\mathbf{Z}'\mathbf{Y}$$

where $\mathbf{Z} = \mathbf{X}\mathbf{D}^{-\frac{1}{2}}$ and $\mathbf{I}_p$ is a $p \times p$ identity matrix.

For IPC analysis, the smallest *m* eigenvalues of $\mathbf{Z}'\mathbf{Z}$ (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 μ is the mean of the nonmissing x-values and σ 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 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 $\mathbf{X}'\mathbf{X}$ 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 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. Add an additional nk^2 for one of the R, CLM, or CLI options 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 “Input Data Sets” section on page 2935 for information on 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 3, "Introduction to Regression Procedures," while other formulas can be found inthe section "Model Fit and Diagnostic Statistics" on page 2968 and the "Influence Diagnostics" section on page 2970.

The analysis-of-variance table includes

- 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-Sq, the adjusted R^2, is a version of R^2 that has been adjusted for degrees of freedom. It is calculated as

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

  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, or CP are used, the analysis-of-variance table and the parameter estimates with associated statistics are produced at each step. Also displayed are

- 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

ODS Table Names

PROC REG assigns a name to each table it creates. You can use these names to reference the table when using the Output Delivery System (ODS) to select tables and create output data sets. These names are listed in the following table. For more information on ODS, see Chapter 15, "Using the Output Delivery System."

Table 55.7. ODS Tables Produced in PROC REG

ODS Table Name	Description	Statement	Option
ACovEst	Consistent covariance of estimates matrix	MODEL	ALL, ACOV
ACovTestANOVA	Test ANOVA using ACOV estimates	TEST	ACOV (MODEL statement)
ANOVA	Model ANOVA table	MODEL	default
CanCorr	Canonical correlations for hypothesis combinations	MTEST	CANPRINT
CollinDiag	Collinearity Diagnostics table	MODEL	COLLIN
CollinDiagNoInt	Collinearity Diagnostics for no intercept model	MODEL	COLLINOINT
ConditionBounds	Bounds on condition number	MODEL	(SELECTION=BACKWARD | FORWARD | STEPWISE | MAXR | MINR) and DETAILS
Corr	Correlation matrix for analysis variables	PROC	ALL, CORR
CorrB	Correlation of estimates	MODEL	CORRB
CovB	Covariance of estimates	MODEL	COVB
CrossProducts	Bordered model **X'X** matrix	MODEL	ALL, XPX
DWStatistic	Durbin-Watson statistic	MODEL	ALL, DW
DependenceEquations	Linear dependence equations	MODEL	default if needed
Eigenvalues	MTest eigenvalues	MTEST	CANPRINT
Eigenvectors	MTest eigenvectors	MTEST	CANPRINT

Table 55.7. (continued)

ODS Table Name	Description	Statement	Option
EntryStatistics	Entry statistics for selection methods	MODEL	(SELECTION=BACKWARD | FORWARD | STEPWISE | MAXR | MINR) and DETAILS
ErrorPlusHypothesis	MTest error plus hypothesis matrix **H**+**E**	MTEST	PRINT
ErrorSSCP	MTest error matrix **E**	MTEST	PRINT
FitStatistics	Model fit statistics	MODEL	default
HypothesisSSCP	MTest hypothesis matrix	MTEST	PRINT
InvMTestCov	Inv(**L** Ginv(**X'X**) **L'**) and Inv(**Lb-c**)	MTEST	DETAILS
InvTestCov	Inv(**L** Ginv(**X'X**) **L'**) and Inv(**Lb-c**)	TEST	PRINT
InvXPX	Bordered **X'X** inverse matrix	MODEL	I
MTestCov	**L** Ginv(**X'X**) **L'** and **Lb-c**	MTEST	DETAILS
MTransform	MTest matrix **M**, across dependents	MTEST	DETAILS
MultStat	Multivariate test statistics	MTEST	default
OutputStatistics	Output statistics table	MODEL	ALL, CLI, CLM, INFLUENCE, P, R
ParameterEstimates	Model parameter estimates	MODEL	default
RemovalStatistics	Removal statistics for selection methods	MODEL	(SELECTION=BACKWARD | STEPWISE | MAXR | MINR) and DETAILS
ResidualStatistics	Residual statistics and PRESS statistic	MODEL	ALL, CLI, CLM, INFLUENCE, P, R
SelParmEst	Parameter estimates for selection methods	MODEL	SELECTION=BACKWARD | FORWARD | STEPWISE | MAXR | MINR
SelectionSummary	Selection summary for forward, backward and stepwise methods	MODEL	SELECTION=BACKWARD | FORWARD | STEPWISE
SeqParmEst	Sequential parameter estimates	MODEL	SEQB
SimpleStatistics	Simple statistics for analysis variables	PROC	ALL, SIMPLE
SpecTest	White's heteroscedasticity test	MODEL	ALL, SPEC
SubsetSelSummary	Selection summary for R-Square, Adj-RSq and Cp methods	MODEL	SELECTION=RSQUARE | ADJRSQ | CP
TestANOVA	Test ANOVA table	TEST	default
TestCov	**L** Ginv(**X'X**) **L'** and **Lb-c**	TEST	PRINT
USSCP	Uncorrected SSCP matrix for analysis variables	PROC	ALL, USSCP

Examples

Example 55.1. 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. The following statements produce Output 55.1.1 through Output 55.1.5. (Collinearity diagnostics for the full model are shown in Figure 55.42 on page 2968.)

```
*------------------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 10.47 59 186 188   57 59.08 50.545  9.93 49 148 155
49 76.32 48.673  9.40 56 186 188   48 61.24 47.920 11.50 52 170 176
52 82.78 47.467 10.50 53 170 172
;
proc reg data=fitness;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=forward;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=backward;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=maxr;
run;
```

The FORWARD model-selection method begins with no variables in the model and adds RunTime, then Age,...

Output 55.1.1. 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

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      1     632.90010      632.90010      84.01    <.0001
Error                     29     218.48144        7.53384
Corrected Total           30     851.38154


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept    82.42177      3.85530   3443.36654   457.05  <.0001
      RunTime      -3.31056      0.36119    632.90010    84.01  <.0001

                     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

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      2     650.66573      325.33287      45.38    <.0001
Error                     28     200.71581        7.16842
Corrected Total           30     851.38154


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept    88.46229      5.37264   1943.41071   271.11  <.0001
      Age          -0.15037      0.09551     17.76563     2.48  0.1267
      RunTime      -3.20395      0.35877    571.67751    79.75  <.0001

                  Bounds on condition number: 1.0369, 4.1478
--------------------------------------------------------------------------------
```

...then RunPulse, then MaxPulse,...

Forward Selection: Step 3

Variable RunPulse Entered: R-Square = 0.8111 and C(p) = 6.9596

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	690.55086	230.18362	38.64	<.0001
Error	27	160.83069	5.95669		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	111.71806	10.23509	709.69014	119.14	<.0001
Age	-0.25640	0.09623	42.28867	7.10	0.0129
RunTime	-2.82538	0.35828	370.43529	62.19	<.0001
RunPulse	-0.13091	0.05059	39.88512	6.70	0.0154

Bounds on condition number: 1.3548, 11.597

Forward Selection: Step 4

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

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	98.14789	11.78569	370.57373	69.35	<.0001
Age	-0.19773	0.09564	22.84231	4.27	0.0488
RunTime	-2.76758	0.34054	352.93570	66.05	<.0001
RunPulse	-0.34811	0.11750	46.90089	8.78	0.0064
MaxPulse	0.27051	0.13362	21.90067	4.10	0.0533

Bounds on condition number: 8.4182, 76.851

...and finally, Weight. 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.

Forward Selection: Step 5

Variable Weight Entered: R-Square = 0.8480 and C(p) = 5.1063

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.97309	144.39462	27.90	<.0001
Error	25	129.40845	5.17634		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	102.20428	11.97929	376.78935	72.79	<.0001
Age	-0.21962	0.09550	27.37429	5.29	0.0301
Weight	-0.07230	0.05331	9.52157	1.84	0.1871
RunTime	-2.68252	0.34099	320.35968	61.89	<.0001
RunPulse	-0.37340	0.11714	52.59624	10.16	0.0038
MaxPulse	0.30491	0.13394	26.82640	5.18	0.0316

Bounds on condition number: 8.7312, 104.83

No other variable met the 0.5000 significance level for entry into the model.

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RunTime	1	0.7434	0.7434	13.6988	84.01	<.0001
2	Age	2	0.0209	0.7642	12.3894	2.48	0.1267
3	RunPulse	3	0.0468	0.8111	6.9596	6.70	0.0154
4	MaxPulse	4	0.0257	0.8368	4.8800	4.10	0.0533
5	Weight	5	0.0112	0.8480	5.1063	1.84	0.1871

The BACKWARD model-selection method begins with the full model.

Output 55.1.2. Backward Selection Method: PROC REG

```
                          The REG Procedure
                            Model: MODEL2
                       Dependent Variable: Oxygen

                       Backward Elimination: Step 0

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


                             Analysis of Variance

                                      Sum of           Mean
Source                    DF         Squares         Square    F Value    Pr > F

Model                      6      722.54361      120.42393      22.43    <.0001
Error                     24      128.83794        5.36825
Corrected Total           30      851.38154


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept   102.93448     12.40326    369.72831    68.87  <.0001
      Age          -0.22697      0.09984     27.74577     5.17  0.0322
      Weight       -0.07418      0.05459      9.91059     1.85  0.1869
      RunTime      -2.62865      0.38456    250.82210    46.72  <.0001
      RunPulse     -0.36963      0.11985     51.05806     9.51  0.0051
      RestPulse    -0.02153      0.06605      0.57051     0.11  0.7473
      MaxPulse      0.30322      0.13650     26.49142     4.93  0.0360

                  Bounds on condition number: 8.7438, 137.13
--------------------------------------------------------------------------------
```

RestPulse is the first variable deleted,...

```
                       Backward Elimination: Step 1


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


                             Analysis of Variance

                                      Sum of           Mean
Source                    DF         Squares         Square    F Value    Pr > F

Model                      5      721.97309      144.39462      27.90    <.0001
Error                     25      129.40845        5.17634
Corrected Total           30      851.38154


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept   102.20428     11.97929    376.78935    72.79  <.0001
      Age          -0.21962      0.09550     27.37429     5.29  0.0301
      Weight       -0.07230      0.05331      9.52157     1.84  0.1871
      RunTime      -2.68252      0.34099    320.35968    61.89  <.0001
      RunPulse     -0.37340      0.11714     52.59624    10.16  0.0038
      MaxPulse      0.30491      0.13394     26.82640     5.18  0.0316

                  Bounds on condition number: 8.7312, 104.83
--------------------------------------------------------------------------------
```

...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.

Backward Elimination: Step 2

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

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	712.45153	178.11288	33.33	<.0001
Error	26	138.93002	5.34346		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	98.14789	11.78569	370.57373	69.35	<.0001
Age	-0.19773	0.09564	22.84231	4.27	0.0488
RunTime	-2.76758	0.34054	352.93570	66.05	<.0001
RunPulse	-0.34811	0.11750	46.90089	8.78	0.0064
MaxPulse	0.27051	0.13362	21.90067	4.10	0.0533

Bounds on condition number: 8.4182, 76.851

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RestPulse	5	0.0007	0.8480	5.1063	0.11	0.7473
2	Weight	4	0.0112	0.8368	4.8800	1.84	0.1871

The MAXR method tries to find the "best" one-variable model, the "best" two-variable model, and so on. For the fitness data, the one-variable model contains RunTime; the two-variable model contains RunTime and Age...

Output 55.1.3. Maximum R-Square Improvement Selection Method: PROC REG

```
                          The REG Procedure
                            Model: MODEL3
                      Dependent Variable: Oxygen

                  Maximum R-Square Improvement: Step 1

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


                          Analysis of Variance

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      1      632.90010      632.90010      84.01    <.0001
Error                     29      218.48144        7.53384
Corrected Total           30      851.38154


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept    82.42177      3.85530   3443.36654   457.05  <.0001
      RunTime      -3.31056      0.36119    632.90010    84.01  <.0001

                     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

                                    Sum of           Mean
Source                    DF       Squares         Square    F Value    Pr > F

Model                      2      650.66573      325.33287      45.38    <.0001
Error                     28      200.71581        7.16842
Corrected Total           30      851.38154


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept    88.46229      5.37264   1943.41071   271.11  <.0001
      Age          -0.15037      0.09551     17.76563     2.48  0.1267
      RunTime      -3.20395      0.35877    571.67751    79.75  <.0001

                  Bounds on condition number: 1.0369, 4.1478
--------------------------------------------------------------------------------


            The above model is the best  2-variable model found.
```

...the three-variable model contains RunTime, Age, and RunPulse; the four-variable model contains Age, RunTime, RunPulse, and MaxPulse...

```
                    Maximum R-Square Improvement: Step 3

        Variable RunPulse Entered: R-Square = 0.8111 and C(p) = 6.9596

                              Analysis of Variance

                                       Sum of           Mean
Source                     DF         Squares         Square    F Value    Pr > F

Model                       3      690.55086      230.18362      38.64    <.0001
Error                      27      160.83069        5.95669
Corrected Total            30      851.38154


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept   111.71806     10.23509    709.69014   119.14  <.0001
      Age          -0.25640      0.09623     42.28867     7.10  0.0129
      RunTime      -2.82538      0.35828    370.43529    62.19  <.0001
      RunPulse     -0.13091      0.05059     39.88512     6.70  0.0154

                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

                                       Sum of           Mean
Source                     DF         Squares         Square    F Value    Pr > F

Model                       4      712.45153      178.11288      33.33    <.0001
Error                      26      138.93002        5.34346
Corrected Total            30      851.38154


                  Parameter     Standard
      Variable     Estimate        Error   Type II SS  F Value  Pr > F

      Intercept    98.14789     11.78569    370.57373    69.35  <.0001
      Age          -0.19773      0.09564     22.84231     4.27  0.0488
      RunTime      -2.76758      0.34054    352.93570    66.05  <.0001
      RunPulse     -0.34811      0.11750     46.90089     8.78  0.0064
      MaxPulse      0.27051      0.13362     21.90067     4.10  0.0533

                Bounds on condition number: 8.4182, 76.851
--------------------------------------------------------------------------------

             The above model is the best  4-variable model found.
```

...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.

Maximum R-Square Improvement: Step 5

Variable Weight Entered: R-Square = 0.8480 and C(p) = 5.1063

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.97309	144.39462	27.90	<.0001
Error	25	129.40845	5.17634		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	102.20428	11.97929	376.78935	72.79	<.0001
Age	-0.21962	0.09550	27.37429	5.29	0.0301
Weight	-0.07230	0.05331	9.52157	1.84	0.1871
RunTime	-2.68252	0.34099	320.35968	61.89	<.0001
RunPulse	-0.37340	0.11714	52.59624	10.16	0.0038
MaxPulse	0.30491	0.13394	26.82640	5.18	0.0316

Bounds on condition number: 8.7312, 104.83

The above model is the best 5-variable model found.

Maximum R-Square Improvement: Step 6

Variable RestPulse Entered: R-Square = 0.8487 and C(p) = 7.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	722.54361	120.42393	22.43	<.0001
Error	24	128.83794	5.36825		
Corrected Total	30	851.38154			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	102.93448	12.40326	369.72831	68.87	<.0001
Age	-0.22697	0.09984	27.74577	5.17	0.0322
Weight	-0.07418	0.05459	9.91059	1.85	0.1869
RunTime	-2.62865	0.38456	250.82210	46.72	<.0001
RunPulse	-0.36963	0.11985	51.05806	9.51	0.0051
RestPulse	-0.02153	0.06605	0.57051	0.11	0.7473
MaxPulse	0.30322	0.13650	26.49142	4.93	0.0360

Bounds on condition number: 8.7438, 137.13

The above model is the best 6-variable model found.

No further improvement in R-Square is possible.

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, BACKWARDS and FORWARD selection methods, you can control the amount of detail displayed by using the DETAILS option. For example, the following statements display only the selection summary table for the FORWARD selection method.

```
proc reg data=fitness;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=forward details=summary;
run;
```

Output 55.1.4. Forward Selection Summary

The REG Procedure
Model: MODEL1
Dependent Variable: Oxygen

Summary of Forward Selection

Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RunTime	1	0.7434	0.7434	13.6988	84.01	<.0001
2	Age	2	0.0209	0.7642	12.3894	2.48	0.1267
3	RunPulse	3	0.0468	0.8111	6.9596	6.70	0.0154
4	MaxPulse	4	0.0257	0.8368	4.8800	4.10	0.0533
5	Weight	5	0.0112	0.8480	5.1063	1.84	0.1871

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 55.1.5

```
model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
      / selection=rsquare cp;
title 'Physical fitness data: all models';
run;
```

Output 55.1.5. All Models by the RSQUARE Method: PROC REG

```
                    Physical fitness data: all models

                           The REG Procedure
                             Model: MODEL2
                        Dependent Variable: Oxygen

                            R-Square Selection Method

Number in
  Model    R-Square      C(p)  Variables in Model

      1      0.7434   13.6988  RunTime
      1      0.1595  106.3021  RestPulse
      1      0.1584  106.4769  RunPulse
      1      0.0928  116.8818  Age
      1      0.0560  122.7072  MaxPulse
      1      0.0265  127.3948  Weight
-------------------------------------------------------------------------------
      2      0.7642   12.3894  Age RunTime
      2      0.7614   12.8372  RunTime RunPulse
      2      0.7452   15.4069  RunTime MaxPulse
      2      0.7449   15.4523  Weight RunTime
      2      0.7435   15.6746  RunTime RestPulse
      2      0.3760   73.9645  Age RunPulse
      2      0.3003   85.9742  Age RestPulse
      2      0.2894   87.6951  RunPulse MaxPulse
      2      0.2600   92.3638  Age MaxPulse
      2      0.2350   96.3209  RunPulse RestPulse
      2      0.1806  104.9523  Weight RestPulse
      2      0.1740  105.9939  RestPulse MaxPulse
      2      0.1669  107.1332  Weight RunPulse
      2      0.1506  109.7057  Age Weight
      2      0.0675  122.8881  Weight MaxPulse
-------------------------------------------------------------------------------
```

```
       3     0.8111    6.9596  Age RunTime RunPulse
       3     0.8100    7.1350  RunTime RunPulse MaxPulse
       3     0.7817   11.6167  Age RunTime MaxPulse
       3     0.7708   13.3453  Age Weight RunTime
       3     0.7673   13.8974  Age RunTime RestPulse
       3     0.7619   14.7619  RunTime RunPulse RestPulse
       3     0.7618   14.7729  Weight RunTime RunPulse
       3     0.7462   17.2588  Weight RunTime MaxPulse
       3     0.7452   17.4060  RunTime RestPulse MaxPulse
       3     0.7451   17.4243  Weight RunTime RestPulse
       3     0.4666   61.5873  Age RunPulse RestPulse
       3     0.4223   68.6250  Age RunPulse MaxPulse
       3     0.4091   70.7102  Age Weight RunPulse
       3     0.3900   73.7424  Age RestPulse MaxPulse
       3     0.3568   79.0013  Age Weight RestPulse
       3     0.3538   79.4891  RunPulse RestPulse MaxPulse
       3     0.3208   84.7216  Weight RunPulse MaxPulse
       3     0.2902   89.5693  Age Weight MaxPulse
       3     0.2447   96.7952  Weight RunPulse RestPulse
       3     0.1882  105.7430  Weight RestPulse MaxPulse
-------------------------------------------------------------------------------
       4     0.8368    4.8800  Age RunTime RunPulse MaxPulse
       4     0.8165    8.1035  Age Weight RunTime RunPulse
       4     0.8158    8.2056  Weight RunTime RunPulse MaxPulse
       4     0.8117    8.8683  Age RunTime RunPulse RestPulse
       4     0.8104    9.0697  RunTime RunPulse RestPulse MaxPulse
       4     0.7862   12.9039  Age Weight RunTime MaxPulse
       4     0.7834   13.3468  Age RunTime RestPulse MaxPulse
       4     0.7750   14.6788  Age Weight RunTime RestPulse
       4     0.7623   16.7058  Weight RunTime RunPulse RestPulse
       4     0.7462   19.2550  Weight RunTime RestPulse MaxPulse
       4     0.5034   57.7590  Age Weight RunPulse RestPulse
       4     0.5025   57.9092  Age RunPulse RestPulse MaxPulse
       4     0.4717   62.7830  Age Weight RunPulse MaxPulse
       4     0.4256   70.0963  Age Weight RestPulse MaxPulse
       4     0.3858   76.4100  Weight RunPulse RestPulse MaxPulse
-------------------------------------------------------------------------------
       5     0.8480    5.1063  Age Weight RunTime RunPulse MaxPulse
       5     0.8370    6.8461  Age RunTime RunPulse RestPulse MaxPulse
       5     0.8176    9.9348  Age Weight RunTime RunPulse RestPulse
       5     0.8161   10.1685  Weight RunTime RunPulse RestPulse MaxPulse
       5     0.7887   14.5111  Age Weight RunTime RestPulse MaxPulse
       5     0.5541   51.7233  Age Weight RunPulse RestPulse MaxPulse
-------------------------------------------------------------------------------
       6     0.8487    7.0000  Age Weight RunTime RunPulse RestPulse MaxPulse
```

The models in Output 55.1.5 are arranged first by the number of variables in the model and second by the magnitude of R^2 for the model. 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 55.2. Predicting Weight by Height and Age

In this example, the weights of school children are modeled as a function of their heights and ages. Modeling is performed separately for boys and girls. 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. 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 55.2.1 through Output 55.2.4:

```
*------------Data on Age, Weight, and Height of Children-------*
| Age (months), height (inches), and weight (pounds) were      |
| recorded for a group of school children.                     |
| From Lewis and Taylor (1967).                                |
*--------------------------------------------------------------*;

data htwt;
   input sex $ age :3.1 height weight @@;
   datalines;
f 143 56.3  85.0 f 155 62.3 105.0 f 153 63.3 108.0 f 161 59.0  92.0
f 191 62.5 112.5 f 171 62.5 112.0 f 185 59.0 104.0 f 142 56.5  69.0
f 160 62.0  94.5 f 140 53.8  68.5 f 139 61.5 104.0 f 178 61.5 103.5
f 157 64.5 123.5 f 149 58.3  93.0 f 143 51.3  50.5 f 145 58.8  89.0
f 191 65.3 107.0 f 150 59.5  78.5 f 147 61.3 115.0 f 180 63.3 114.0
f 141 61.8  85.0 f 140 53.5  81.0 f 164 58.0  83.5 f 176 61.3 112.0
f 185 63.3 101.0 f 166 61.5 103.5 f 175 60.8  93.5 f 180 59.0 112.0
f 210 65.5 140.0 f 146 56.3  83.5 f 170 64.3  90.0 f 162 58.0  84.0
f 149 64.3 110.5 f 139 57.5  96.0 f 186 57.8  95.0 f 197 61.5 121.0
f 169 62.3  99.5 f 177 61.8 142.5 f 185 65.3 118.0 f 182 58.3 104.5
f 173 62.8 102.5 f 166 59.3  89.5 f 168 61.5  95.0 f 169 62.0  98.5
f 150 61.3  94.0 f 184 62.3 108.0 f 139 52.8  63.5 f 147 59.8  84.5
f 144 59.5  93.5 f 177 61.3 112.0 f 178 63.5 148.5 f 197 64.8 112.0
f 146 60.0 109.0 f 145 59.0  91.5 f 147 55.8  75.0 f 145 57.8  84.0
f 155 61.3 107.0 f 167 62.3  92.5 f 183 64.3 109.5 f 143 55.5  84.0
f 183 64.5 102.5 f 185 60.0 106.0 f 148 56.3  77.0 f 147 58.3 111.5
f 154 60.0 114.0 f 156 54.5  75.0 f 144 55.8  73.5 f 154 62.8  93.5
f 152 60.5 105.0 f 191 63.3 113.5 f 190 66.8 140.0 f 140 60.0  77.0
f 148 60.5  84.5 f 189 64.3 113.5 f 143 58.3  77.5 f 178 66.5 117.5
f 164 65.3  98.0 f 157 60.5 112.0 f 147 59.5 101.0 f 148 59.0  95.0
f 177 61.3  81.0 f 171 61.5  91.0 f 172 64.8 142.0 f 190 56.8  98.5
f 183 66.5 112.0 f 143 61.5 116.5 f 179 63.0  98.5 f 186 57.0  83.5
f 182 65.5 133.0 f 182 62.0  91.5 f 142 56.0  72.5 f 165 61.3 106.5
f 165 55.5  67.0 f 154 61.0 122.5 f 150 54.5  74.0 f 155 66.0 144.5
f 163 56.5  84.0 f 141 56.0  72.5 f 147 51.5  64.0 f 210 62.0 116.0
f 171 63.0  84.0 f 167 61.0  93.5 f 182 64.0 111.5 f 144 61.0  92.0
f 193 59.8 115.0 f 141 61.3  85.0 f 164 63.3 108.0 f 186 63.5 108.0
f 169 61.5  85.0 f 175 60.3  86.0 f 180 61.3 110.5 m 165 64.8  98.0
m 157 60.5 105.0 m 144 57.3  76.5 m 150 59.5  84.0 m 150 60.8 128.0
m 139 60.5  87.0 m 189 67.0 128.0 m 183 64.8 111.0 m 147 50.5  79.0
m 146 57.5  90.0 m 160 60.5  84.0 m 156 61.8 112.0 m 173 61.3  93.0
m 151 66.3 117.0 m 141 53.3  84.0 m 150 59.0  99.5 m 164 57.8  95.0
m 153 60.0  84.0 m 206 68.3 134.0 m 250 67.5 171.5 m 176 63.8  98.5
m 176 65.0 118.5 m 140 59.5  94.5 m 185 66.0 105.0 m 180 61.8 104.0
m 146 57.3  83.0 m 183 66.0 105.5 m 140 56.5  84.0 m 151 58.3  86.0
m 151 61.0  81.0 m 144 62.8  94.0 m 160 59.3  78.5 m 178 67.3 119.5
m 193 66.3 133.0 m 162 64.5 119.0 m 164 60.5  95.0 m 186 66.0 112.0
m 143 57.5  75.0 m 175 64.0  92.0 m 175 68.0 112.0 m 175 63.5  98.5
```

```
m 173 69.0 112.5 m 170 63.8 112.5 m 174 66.0 108.0 m 164 63.5 108.0
m 144 59.5  88.0 m 156 66.3 106.0 m 149 57.0  92.0 m 144 60.0 117.5
m 147 57.0  84.0 m 188 67.3 112.0 m 169 62.0 100.0 m 172 65.0 112.0
m 150 59.5  84.0 m 193 67.8 127.5 m 157 58.0  80.5 m 168 60.0  93.5
m 140 58.5  86.5 m 156 58.3  92.5 m 156 61.5 108.5 m 158 65.0 121.0
m 184 66.5 112.0 m 156 68.5 114.0 m 144 57.0  84.0 m 176 61.5  81.0
m 168 66.5 111.5 m 149 52.5  81.0 m 142 55.0  70.0 m 188 71.0 140.0
m 203 66.5 117.0 m 142 58.8  84.0 m 189 66.3 112.0 m 188 65.8 150.5
m 200 71.0 147.0 m 152 59.5 105.0 m 174 69.8 119.5 m 166 62.5  84.0
m 145 56.5  91.0 m 143 57.5 101.0 m 163 65.3 117.5 m 166 67.3 121.0
m 182 67.0 133.0 m 173 66.0 112.0 m 155 61.8  91.5 m 162 60.0 105.0
m 177 63.0 111.0 m 177 60.5 112.0 m 175 65.5 114.0 m 166 62.0  91.0
m 150 59.0  98.0 m 150 61.8 118.0 m 188 63.3 115.5 m 163 66.0 112.0
m 171 61.8 112.0 m 162 63.0  91.0 m 141 57.5  85.0 m 174 63.0 112.0
m 142 56.0  87.5 m 148 60.5 118.0 m 140 56.8  83.5 m 160 64.0 116.0
m 144 60.0  89.0 m 206 69.5 171.5 m 159 63.3 112.0 m 149 56.3  72.0
m 193 72.0 150.0 m 194 65.3 134.5 m 152 60.8  97.0 m 146 55.0  71.5
m 139 55.0  73.5 m 186 66.5 112.0 m 161 56.8  75.0 m 153 64.8 128.0
m 196 64.5  98.0 m 164 58.0  84.0 m 159 62.8  99.0 m 178 63.8 112.0
m 153 57.8  79.5 m 155 57.3  80.5 m 178 63.5 102.5 m 142 55.0  76.0
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
;

title '----- Data on age, weight, and height of children ------';
proc reg outest=est1 outsscp=sscp1 rsquare;
   by sex;
   eq1: model  weight=height;
   eq2: model  weight=height age;
proc print data=sscp1;
   title2 'SSCP type data set';
proc print data=est1;
   title2 'EST type data set';
run;
```

Output 55.2.1. Height and Weight Data: Female Children

```
          ----- Data on age, weight, and height of children ------

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

                            The REG Procedure
                                Model: EQ1
                        Dependent Variable: weight

                            Analysis of Variance

                                      Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F

Model                       1          21507          21507     141.09    <.0001
Error                     109          16615      152.42739
Corrected Total           110          38121


              Root MSE             12.34615    R-Square     0.5642
              Dependent Mean       98.87838    Adj R-Sq     0.5602
              Coeff Var            12.48620


                            Parameter Estimates

                         Parameter       Standard
      Variable     DF       Estimate          Error    t Value    Pr > |t|

      Intercept     1     -153.12891       21.24814      -7.21      <.0001
      height        1        4.16361        0.35052      11.88      <.0001
```

```
          ----- Data on age, weight, and height of children ------

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

                            The REG Procedure
                                Model: EQ2
                        Dependent Variable: weight

                            Analysis of Variance

                                      Sum of           Mean
Source                     DF        Squares         Square    F Value    Pr > F

Model                       2          22432          11216      77.21    <.0001
Error                     108          15689      145.26700
Corrected Total           110          38121


              Root MSE             12.05268    R-Square     0.5884
              Dependent Mean       98.87838    Adj R-Sq     0.5808
              Coeff Var            12.18939


                            Parameter Estimates

                         Parameter       Standard
      Variable     DF       Estimate          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 55.2.2. Height and Weight Data: Male Children

```
          ----- Data on age, weight, and height of children ------

---------------------------------- sex=m -----------------------------------

                            The REG Procedure
                                Model: EQ1
                        Dependent Variable: weight

                            Analysis of Variance

                                     Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F

Model                      1          31126          31126     206.24    <.0001
Error                    124          18714      150.92222
Corrected Total          125          49840


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


                            Parameter Estimates

                         Parameter       Standard
     Variable     DF       Estimate          Error    t Value    Pr > |t|

     Intercept     1     -125.69807       15.99362      -7.86      <.0001
     height        1        3.68977        0.25693      14.36      <.0001
```

```
          ----- Data on age, weight, and height of children ------

---------------------------------- sex=m -----------------------------------

                            The REG Procedure
                                Model: EQ2
                        Dependent Variable: weight

                            Analysis of Variance

                                     Sum of           Mean
Source                    DF        Squares         Square    F Value    Pr > F

Model                      2          32975          16487     120.24    <.0001
Error                    123          16866      137.11922
Corrected Total          125          49840


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


                            Parameter Estimates

                         Parameter       Standard
     Variable     DF       Estimate          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 females and males, the overall F statistics for both models are significant, indicating that the model explains a significant portion of the variation in the data. For females, the full model is

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

and, for males, the full model is

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

Output 55.2.3. SSCP Matrix

----- Data on age, weight, and height of children ------
SSCP type data set

Obs	sex	_TYPE_	_NAME_	Intercept	height	weight	age
1	f	SSCP	Intercept	111.0	6718.40	10975.50	1824.90
2	f	SSCP	height	6718.4	407879.32	669469.85	110818.32
3	f	SSCP	weight	10975.5	669469.85	1123360.75	182444.95
4	f	SSCP	age	1824.9	110818.32	182444.95	30363.81
5	f	N		111.0	111.00	111.00	111.00
6	m	SSCP	Intercept	126.0	7825.00	13034.50	2072.10
7	m	SSCP	height	7825.0	488243.60	817919.60	129432.57
8	m	SSCP	weight	13034.5	817919.60	1398238.75	217717.45
9	m	SSCP	age	2072.1	129432.57	217717.45	34515.95
10	m	N		126.0	126.00	126.00	126.00

The OUTSSCP= data set is shown in Output 55.2.3. 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 55.2.4. OUTEST Data Set

----- Data on age, weight, and height of children ------
EST type data set

Obs	sex	_MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	height	weight	age	_IN_	_P_	_EDF_	_RSQ_
1	f	EQ1	PARMS	weight	12.3461	-153.129	4.16361	-1	.	1	2	109	0.56416
2	f	EQ2	PARMS	weight	12.0527	-150.597	3.60378	-1	1.90703	2	3	108	0.58845
3	m	EQ1	PARMS	weight	12.2850	-125.698	3.68977	-1	.	1	2	124	0.62451
4	m	EQ2	PARMS	weight	11.7098	-113.713	2.68075	-1	3.08167	2	3	123	0.66161

The OUTEST= data set is displayed in Output 55.2.4; 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.

Example 55.3. 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, $X1_i$ is a quantitative variable, $X2_i$ is a qualitative variable taking on values 0 or 1, and $X1_iX2_i$ is the interaction. The variable $X2_i$ is called a dummy, binary, or indicator variable. With values 0 or 1, it distinguishes between two populations. The model is of the form

$$Y_i = \beta_0 + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X1_i X2_i + \epsilon_i$$

for the observations $i = 1, 2, \ldots, n$. The parameters to be estimated are β_0, β_1, β_2, and β_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 they implement new insurance innovations. He believes that the type of firm may affect this relationship and suspects that there may be some interaction between the size and type of firm. The dummy variable in the model allows the two firms to have different intercepts. The interaction term allows 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 ith firm. The variable $X1_i$ is the size of the firm, measured in total assets of the firm. The variable $X2_i$ denotes the firm type and is 0 if the firm is a mutual fund company and 1 if the firm is a stock company. The dummy variable allows 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 $X2_i$. If $X2_i = 0$, the model is

$$Y_i = \beta_0 + \beta_1 X1_i + \epsilon_i$$

This is the model for a mutual company. If $X2_i = 1$, the model for a stock firm is

$$Y_i = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X1_i + \epsilon_i$$

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.

```
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:

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

The ANOVA table is displayed in Output 55.3.1.

*From Neter, J. et al., *Applied Linear Statistical Models*, Third Edition, Copyright (c) 1990, Richard D. Irwin. Reprinted with permission of The McGraw-Hill Companies.

Output 55.3.1. ANOVA Table and Parameter Estimates

Regression With Quantitative and Qualitative Variables

The REG Procedure
Model: MODEL1
Dependent Variable: time

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	1504.41904	501.47301	45.49	<.0001
Error	16	176.38096	11.02381		
Corrected Total	19	1680.80000			

Root MSE	3.32021	R-Square	0.8951
Dependent Mean	19.40000	Adj R-Sq	0.8754
Coeff Var	17.11450		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	33.83837	2.44065	13.86	<.0001
size	1	-0.10153	0.01305	-7.78	<.0001
type	1	8.13125	3.65405	2.23	0.0408
sizetype	1	-0.00041714	0.01833	-0.02	0.9821

The overall F statistic is significant (F=45.490, p<0.0001). The interaction term is not significant (t=−0.023, p=0.9821). Hence, this term should be removed and the model re-fitted, as shown in the following statements.

```
   delete sizetype;
   print;
run;
```

The DELETE statement removes the interaction term (sizetype) from the model. The new ANOVA table is shown in Output 55.3.2.

Output 55.3.2. ANOVA Table and Parameter Estimates

Regression With Quantitative and Qualitative Variables

The REG Procedure
Model: MODEL1.1
Dependent Variable: time

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1504.41333	752.20667	72.50	<.0001
Error	17	176.38667	10.37569		
Corrected Total	19	1680.80000			

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 ($X2_i = 0$) is

$$\text{time} = 33.87 - 0.102 \times \text{size}$$

and the fitted model for a stock company ($X2_i = 1$) is

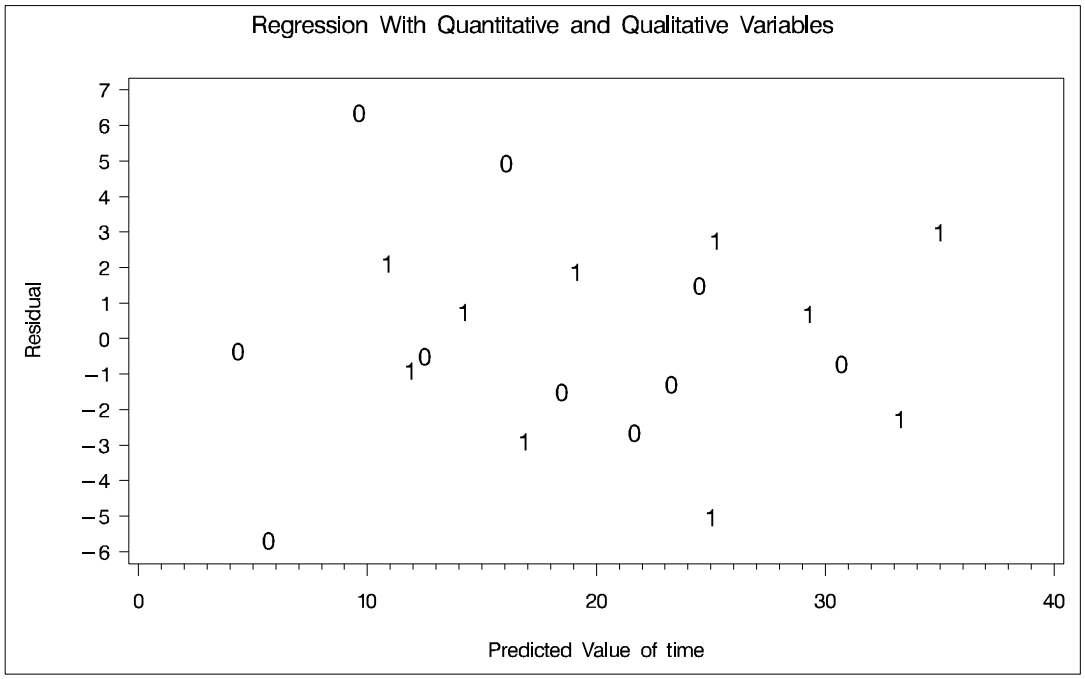
$$\text{time} = (33.87 + 8.055) - 0.102 \times \text{size}$$

So the two models have different intercepts but the same slope.

Now plot the residual versus predicted values using the firm type as the plot symbol (PLOT=TYPE); this can be useful in determining if the firm types have different residual patterns. PROC REG does not support the `plot y*x=type` syntax for high-resolution graphics, so use PROC GPLOT to create Output 55.3.3. First, the OUTPUT statement saves the residuals and predicted values from the new model in the OUT= data set.

```
   output out=out r=r p=p;
run;
symbol1 v='0' c=blue   f=swissb;
symbol2 v='1' c=yellow f=swissb;
axis1 label=(angle=90);
proc gplot data=out;
   plot r*p=type    / nolegend vaxis=axis1 cframe=ligr;
   plot p*size=type / nolegend vaxis=axis1 cframe=ligr;
run;
```

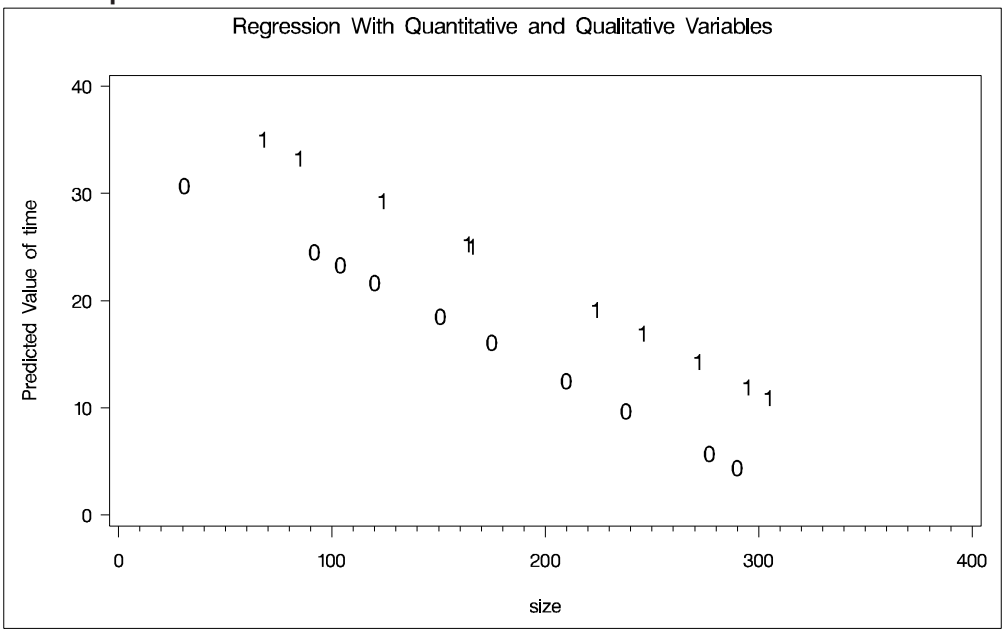
Output 55.3.3. Plot of Residual vs. Predicted Values



The residuals show no major trend. Neither firm type by itself shows a trend either. This indicates that the model is satisfactory.

A plot of the predicted values versus size appears in Output 55.3.4, where the firm type is again used as the plotting symbol.

Output 55.3.4. Plot of Predicted vs. Size



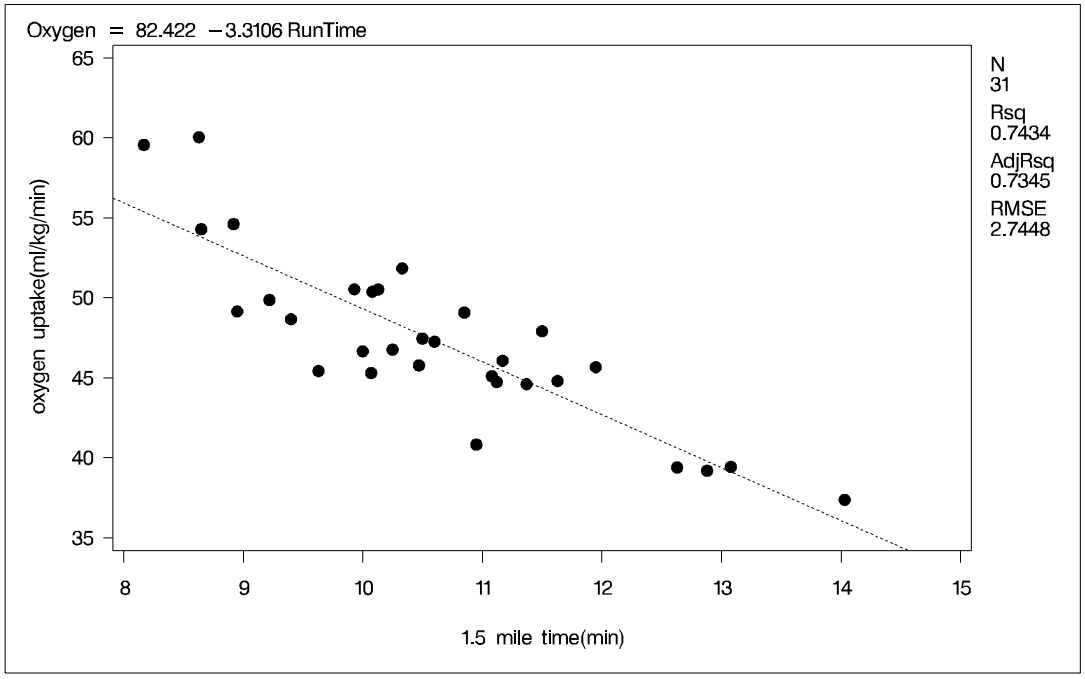
The different intercepts are very evident in this plot.

Example 55.4. Displaying Plots for Simple Linear Regression

This example introduces the basic PROC REG graphics syntax used to produce a standard plot of data from the aerobic fitness data set (Example 55.1 on page 2993). A simple linear regression of Oxygen on RunTime is performed, and a plot of Oxygen∗RunTime is requested. The fitted model, the regression line, and the four default statistics are also displayed in Output 55.4.1.

```
data fitness;
   set fitness;
   label Age      =’age(years)’
         Weight   =’weight(kg)’
         Oxygen   =’oxygen uptake(ml/kg/min)’
         RunTime  =’1.5 mile time(min)’
         RestPulse=’rest pulse’
         RunPulse =’running pulse’
         MaxPulse =’maximum running pulse’;
proc reg data=fitness;
   model Oxygen=RunTime;
   plot Oxygen*RunTime / cframe=ligr;
run;
```

Output 55.4.1. Simple Linear Regression

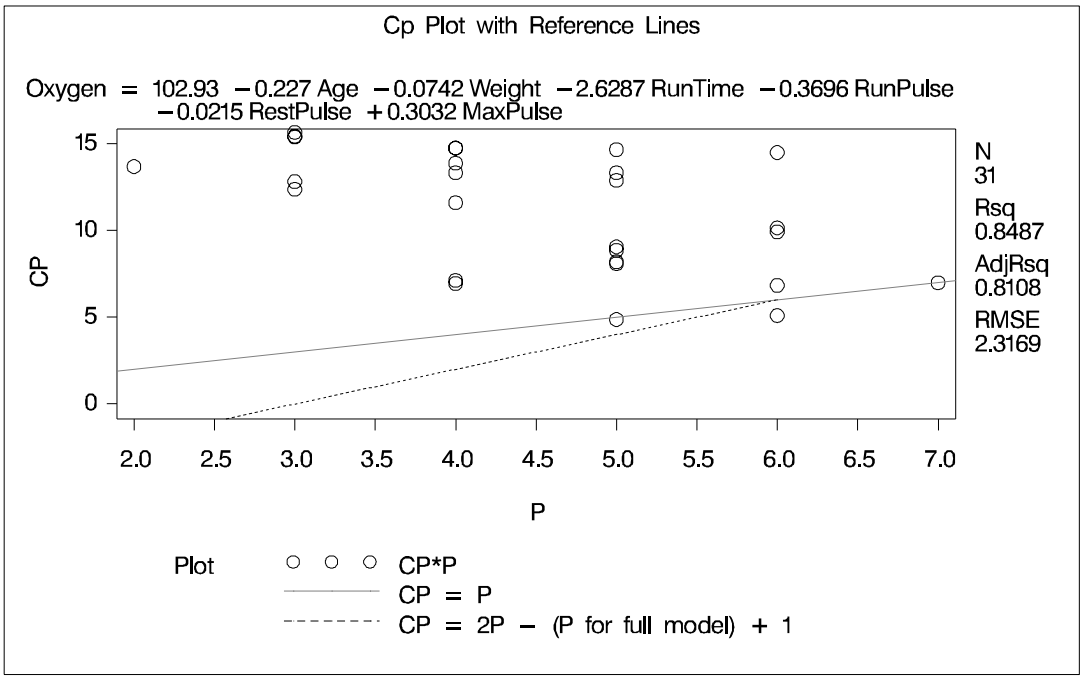


Example 55.5. Creating a C_p Plot

The C_p statistics for model selection are plotted against the number of parameters in the model, and the CHOCKING= and CMALLOWS= options draw useful reference lines. Note the four default statistics in the plot margin, the default model equation, and the default legend in Output 55.5.1.

```
title 'Cp Plot with Reference Lines';
symbol1 c=green;
proc reg data=fitness;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=rsquare noprint;
   plot cp.*np.
        / chocking=red  cmallows=blue
          vaxis=0 to 15 by 5 cframe=ligr;
run;
```

Output 55.5.1. C_p Plot



Using the criteria suggested by Hocking (1976) (see the section "Dictionary of PLOT Statement Options" beginning on page 2919), Output 55.5.1 indicates that a 6-variable model is a reasonable choice for doing parameter estimation, while a 5-variable model may be suitable for doing prediction.

Example 55.6. Controlling Plot Appearance with Graphics Options

This example uses model fit summary statistics from the OUTEST= data set to create a plot for a model selection analysis. Global graphics statements and PLOT statement options are used to control the appearance of the plot.

```
goptions ctitle=black   htitle=3.5pct ftitle=swiss
         ctext =magenta htext =3.0pct ftext =swiss
         cback =ligr    border;
symbol1 v=circle c=red h=1 w=2;
title1 'Selection=Rsquare';
title2 'plot Rsquare versus the number of parameters P in '
       'each model';

proc reg data=fitness;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse
         / selection=rsquare noprint;
   plot rsq.*np.
        / aic bic edf gmsep jp np pc sbc sp
          haxis=2 to 7 by 1
          caxis=red cframe=white ctext=blue
```

```
            modellab='Full Model' modelht=2.4
            statht=2.4;
run;
```

In the GOPTIONS statement,

BORDER frames the entire display

CBACK= specifies the background color

CTEXT= selects the default color for the border and all text, including titles, footnotes, and notes

CTITLE= specifies the title, footnote, note, and border color

HTEXT= specifies the height for all text in the display

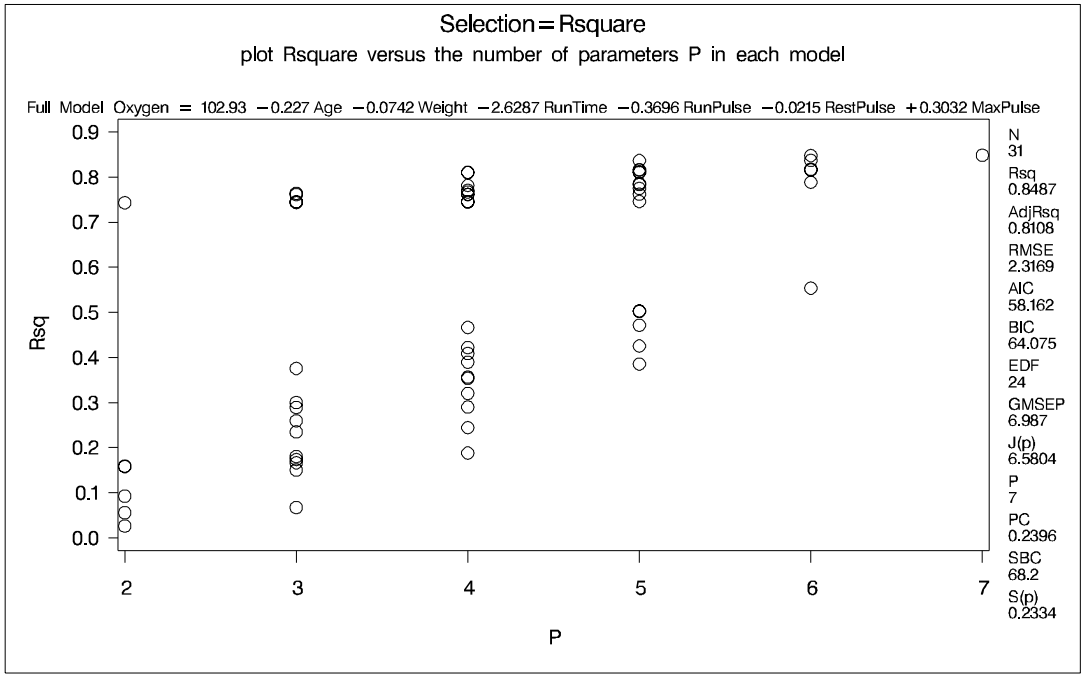
HTITLE= specifies the height for the first title line

FTEXT= selects the default font for all text, including titles, footnotes, notes, the model label and equation, the statistics, the axis labels, the tick values, and the legend

FTITLE= specifies the first title font

For more information on the GOPTIONS statement and other global graphics statements, refer to *SAS/GRAPH Software: Reference*.

Output 55.6.1. Controlling Plot Appearance and Plotting OUTEST= Statistics



In Output 55.6.1, note the following:

- The PLOT statement option CTEXT= affects all text not controlled by the CTITLE= option in the GOPTIONS statement. Hence, the GOPTIONS statement option CTEXT=MAGENTA has no effect. Therefore, the color of the title is black and all other text is blue.

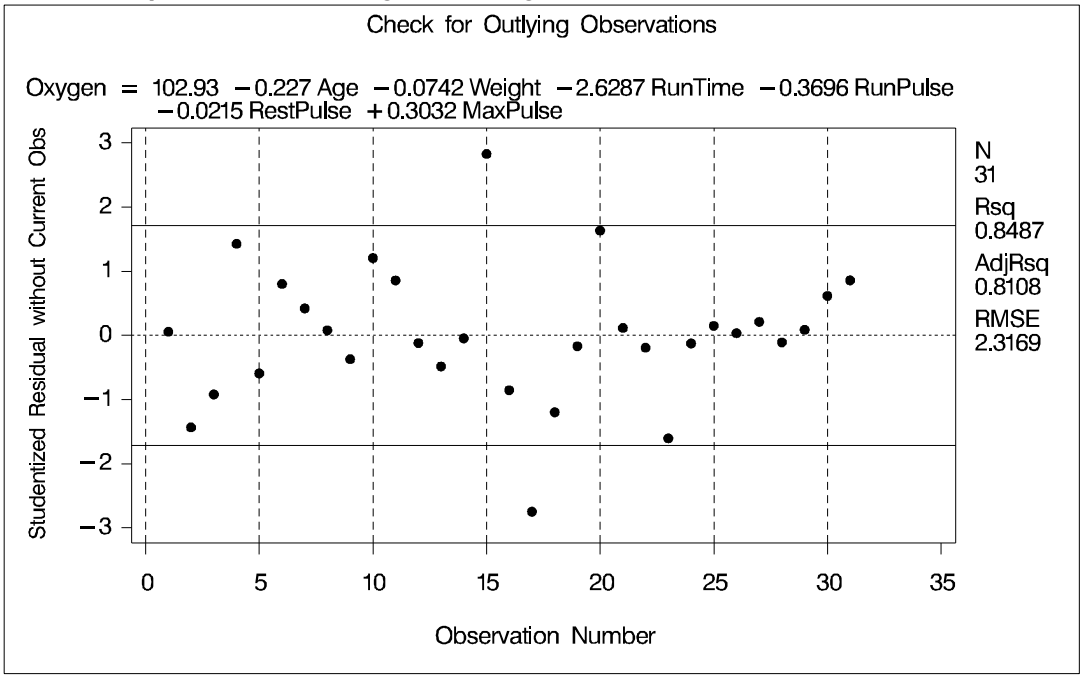
- The area enclosed by the axes and the frame has a white background, while the background outside the plot area is gray.
- The MODELHT= option allows the entire model equation to fit on one line.
- The STATHT= option allows the statistics in the margin to fit in one column.
- The displayed statistics and the fitted model equation refer to the selected model. See the "High Resolution Graphics Plots" section beginning on page 2915 for more information.

Example 55.7. Plotting Model Diagnostic Statistics

This example illustrates how you can display diagnostics for checking the adequacy of a regression model. The following statements plot the studentized deleted residuals against the observation number for the full model. Vertical reference lines at $\pm\text{tinv}(.95, n-p-1) = \pm 1.714$ are added to identify possible outlying Oxygen values. A vertical reference line is displayed at zero by default when the RSTUDENT option is specified. The graph is shown in Output 55.7.1. Observations 15 and 17 are indicated as possible outliers.

```
title 'Check for Outlying Observations';
symbol v=dot h=1 c=green;

proc reg data=fitness;
   model Oxygen=Age Weight RunTime RunPulse RestPulse MaxPulse;
   plot rstudent.*obs.
        / vref= -1.714 1.714 cvref=blue lvref=1
          href= 0 to 30 by 5 chref=red  cframe=ligr;
run;
```

Output 55.7.1. Plotting Model Diagnostic Statistics

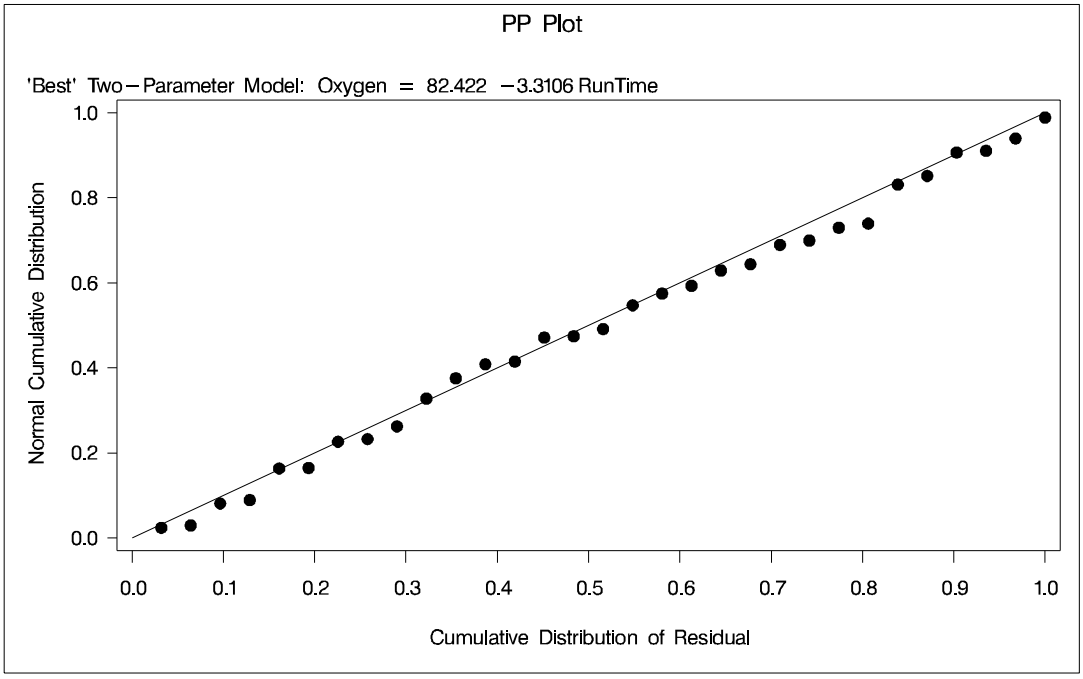


Example 55.8. Creating PP and QQ Plots

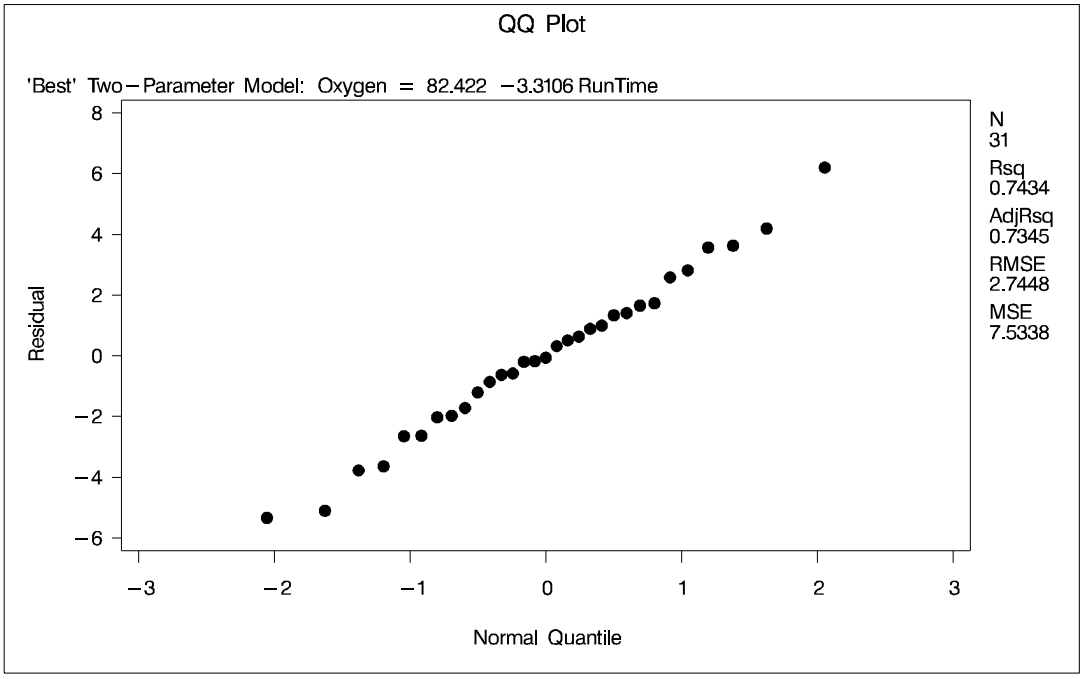
The following program creates probability-probability plots and quantile-quantile plots of the residuals (Output 55.8.1 and Output 55.8.2, respectively). An annotation data set is created to produce the (0,0)−(1,1) reference line for the PP-plot. Note that the NOSTAT option for the PP-plot suppresses the statistics that would be displayed in the margin.

```
data annote1;
   length function color $8;
   retain ysys xsys '2' color 'black';
   function='move';
      x=0;
      y=0;
      output;
   function='draw';
      x=1;
      y=1;
      output;
run;

symbol1 c=blue;
proc reg data=fitness;
   title 'PP Plot';
   model Oxygen=RunTime / noprint;
   plot npp.*r.
        / annotate=annote1 nostat cframe=ligr
          modellab="'Best' Two-Parameter Model:";
run;
   title 'QQ Plot';
   plot r.*nqq.
        / noline mse cframe=ligr
          modellab="'Best' Two-Parameter Model:";
run;
```

Output 55.8.1. Normal Probability-Probability Plot for the Residuals



Output 55.8.2. Normal Quantile-Quantile Plot for the Residuals

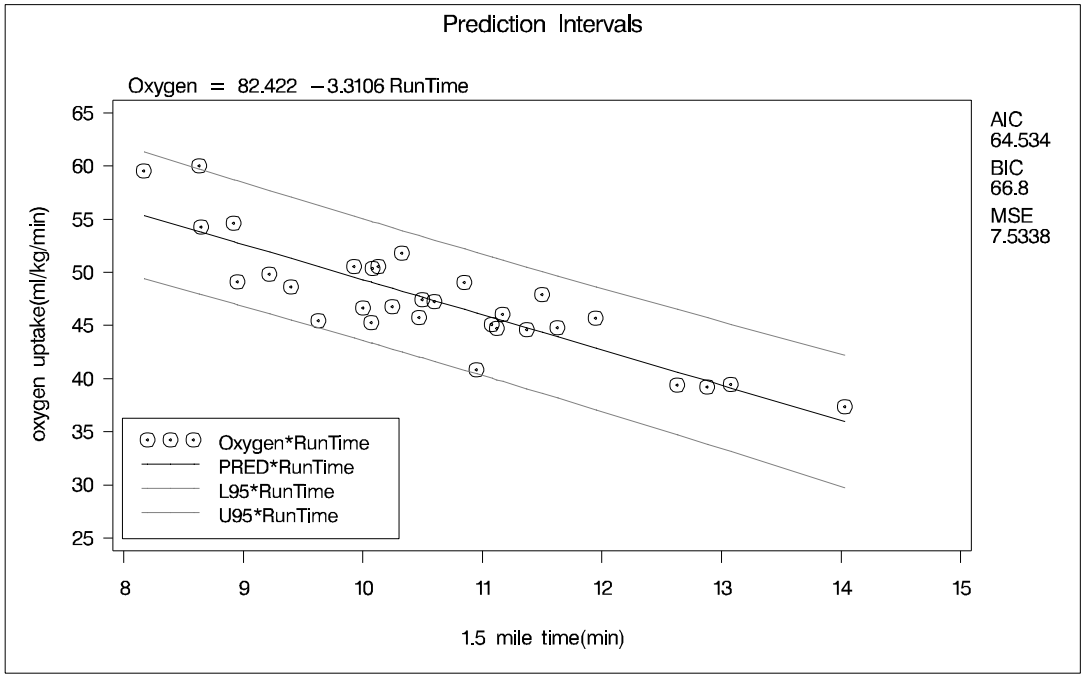


Example 55.9. Displaying Confidence and Prediction Intervals

This example illustrates how you can use shorthand commands to plot the dependent variable, the predicted value, and the 95% confidence or prediction intervals against a regressor. The following statements use the PRED option to create a plot with prediction intervals; the CONF option works similarly. Results are displayed in Output 55.9.1. Note that the statistics displayed by default in the margin are suppressed while three other statistics are exhibited.

```
legend1 position=(bottom left inside)
        across=1 cborder=red offset=(0,0)
        shape=symbol(3,1) label=none
        value=(height=.8);
title 'Prediction Intervals';
symbol1 c=yellow v=- h=1;
symbol2 c=red;
symbol3 c=blue;
symbol4 c=blue;

proc reg data=fitness;
   model Oxygen=RunTime / noprint;
   plot Oxygen*RunTime / pred nostat mse aic bic
        caxis=red ctext=blue cframe=ligr
        legend=legend1 modellab='         ';
run;
```

Output 55.9.1. Prediction Intervals



Plots can be produced with both confidence and prediction intervals using the following statement.

```
plot Oxygen*RunTime / conf pred;
```

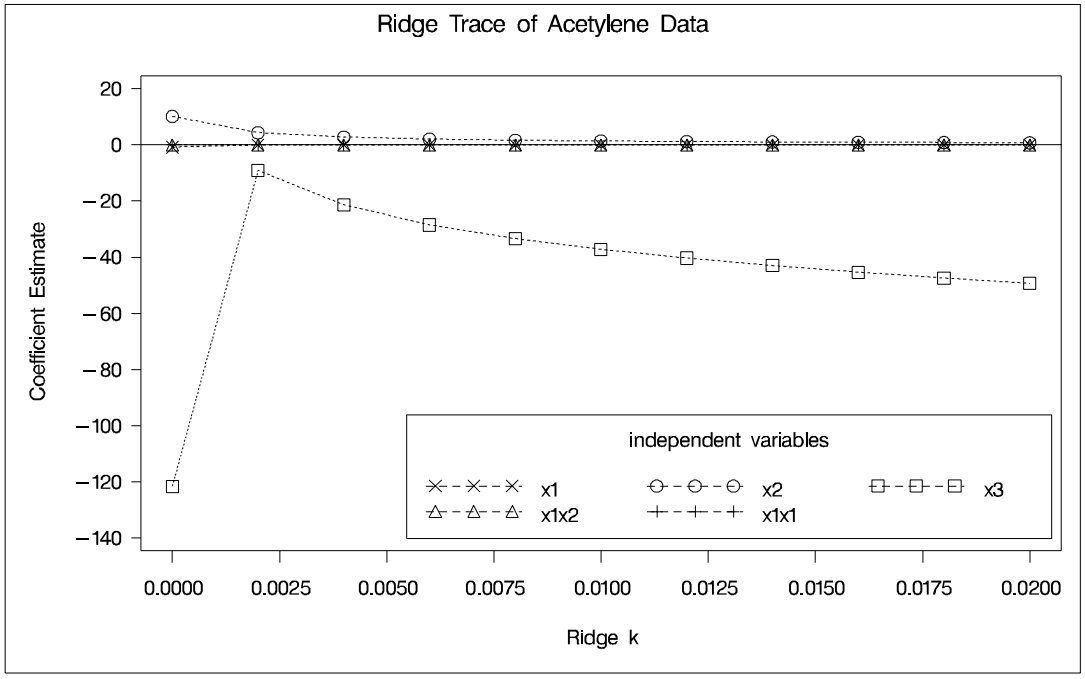
Example 55.10. Displaying the Ridge Trace for Acetylene Data

This example and Example 55.11 use the acetylene data in Marquardt and Snee (1975) to illustrate the RIDGEPLOT and OUTVIF options.

```
data acetyl;
   input x1-x4 @@;
   x1x2 = x1 * x2;
   x1x1 = x1 * x1;
   label x1  = 'reactor temperature(celsius)'
         x2  = 'h2 to n-heptone ratio'
         x3  = 'contact time(sec)'
         x4  = 'conversion percentage'
         x1x2= 'temperature-ratio interaction'
         x1x1= 'squared temperature';
   datalines;
1300  7.5 .012 49   1300  9   .012  50.2 1300 11 .0115 50.5
1300 13.5 .013 48.5 1300 17   .0135 47.5 1300 23 .012  44.5
1200  5.3 .04  28   1200  7.5 .038  31.5 1200 11 .032  34.5
1200 13.5 .026 35   1200 17   .034  38   1200 23 .041  38.5
1100  5.3 .084 15   1100  7.5 .098  17   1100 11 .092  20.5
1100 17   .086 29.5
;

title 'Ridge Trace of Acetylene Data';
symbol1 v=x        c=blue;
symbol2 v=circle   c=yellow;
symbol3 v=square   c=cyan;
symbol4 v=triangle c=green;
symbol5 v=plus     c=orange;
legend2 position=(bottom right inside)
        across=3 cborder=black offset=(0,0)
        label=(color=blue position=(top center)
               'independent variables') cframe=white;

proc reg data=acetyl outvif
         outest=b ridge=0 to 0.02 by .002;
   model x4=x1 x2 x3 x1x2 x1x1/noprint;
   plot / ridgeplot nomodel legend=legend2 nostat
          vref=0 lvref=1 cvref=blue cframe=ligr;
run;
```

The results produced by the RIDGEPLOT option are shown in Output 55.10.1. The OUTVIF option outputs the variance inflation factors to the OUTEST= data set, which is used in Example 55.11.

Output 55.10.1. Using the RIDEGPLOT Option for Ridge Regression



Example 55.11. Plotting Variance Inflation Factors

This example uses the REG procedure to create plots from a data set. The variance inflation factors (output by the OUTVIF option in the previous example) are plotted against the ridge regression control values *k*. The following statements create Output 55.11.1:

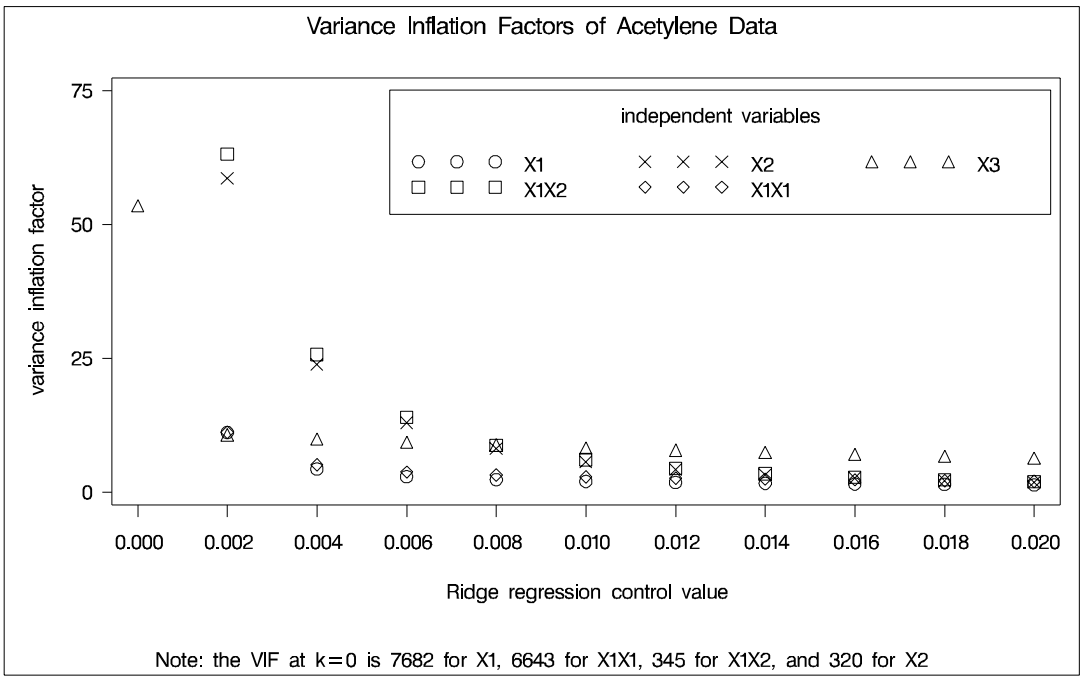
```
data b (keep=_RIDGE_ x1-x3 x1x2 x1x1);
   set b;
   if _TYPE_='RIDGEVIF';
   label x1='variance inflation factor';
run;

legend3 position=(top right inside) across=3
        cborder=black cframe=white
        label=(color=blue position=(top center)
               'independent variables')
        value=('X1' 'X2' 'X3' 'X1X2' 'X1X1');
symbol1 c=blue   /*v=circle  */;
symbol2 c=yellow /*v=x       */;
symbol3 c=cyan   /*v=triangle*/;
symbol4 c=green  /*v=square  */;
symbol5 c=orange /*v=diamond */;
title 'Variance Inflation Factors of Acetylene Data';

proc reg data=b;
   var _RIDGE_ x3 x1x2 x1x1;
   model x1=x2 / noprint;
```

```
   plot (x1 x2 x3 x1x2 x1x1)*_RIDGE_
        / nomodel nostat legend=legend3 overlay
          vaxis = 0 to 75 by 25 cframe=ligr
          haxis = 0 to .02 by .002;
   footnote "Note: the VIF at k=0 is 7682 for X1, "
            "6643 for X1X1, 345 for X1X2, and 320 for X2";
run;
```

The GPLOT procedure can create the same plot with the following statements. The resulting display is not shown in this report.

```
axis1 label=(a=90 r=0 'variance inflation factor')
      order=(0 to 75 by 25) minor=none;
proc gplot data=b;
   plot (x1 x2 x3 x1x2 x1x1)*_RIDGE_
        / legend=legend3 overlay frame
          vaxis = axis1
          haxis = 0 to .02 by .002 hminor=0;
   footnote "Note: the VIF at k=0 is 7682 for X1, "
            "6643 for X1X1, 345 for X1X2, and 320 for X2";
run;
```

Output 55.11.1. Using PROC REG to Plot the VIFs



References

Akaike, H. (1969), "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics*, 21, 243–247.

Allen, D.M. (1971), "Mean Square Error of Prediction as a Criterion for Selecting Variables," *Technometrics*, 13, 469–475.

Allen, D.M. and Cady, F.B. (1982), *Analyzing Experimental Data by Regression*, Belmont, CA: Lifetime Learning Publications.

Amemiya, T. (1976), "Selection of Regressors," Technical Report No. 225, Stanford, CA: Stanford University.

Belsley, D.A., Kuh, E., and Welsch, R.E. (1980), *Regression Diagnostics*, New York: John Wiley & Sons, Inc.

Berk, K.N. (1977), "Tolerance and Condition in Regression Computations," *Journal of the American Statistical Association*, 72, 863–866.

Bock, R.D. (1975), *Multivariate Statistical Methods in Behavioral Research*, New York: McGraw-Hill Book Co.

Box, G.E.P. (1966), "The Use and Abuse of Regression," *Technometrics*, 8, 625–629.

Cook, R.D. (1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15–18.

Cook, R.D. (1979), "Influential Observations in Linear Regression," *Journal of the American Statistical Association*, 74, 169–174.

Daniel, C. and Wood, F. (1980), *Fitting Equations to Data*, Revised Edition, New York: John Wiley & Sons, Inc.

Darlington, R.B. (1968), "Multiple Regression in Psychological Research and Practice," *Psychological Bulletin*, 69, 161–182.

Draper, N. and Smith, H. (1981), *Applied Regression Analysis*, Second Edition, New York: John Wiley & Sons, Inc.

Durbin, J. and Watson, G.S. (1951), "Testing for Serial Correlation in Least Squares Regression," *Biometrika*, 37, 409–428.

Freund, R.J. and Littell, R.C. (1986), *SAS System for Regression, 1986 Edition*, Cary, NC: SAS Institute Inc.

Furnival, G.M. and Wilson, R.W. (1974), "Regression by Leaps and Bounds," *Technometrics*, 16, 499–511.

Gauss, K.F. (1809), *Werke*, 4, 1–93.

Goodnight, J.H. (1979), "A Tutorial on the SWEEP Operator," *The American Statistician*, 33, 149–158. (Also available as *The Sweep Operator: Its Importance in Statistical Computing*, SAS Technical Report R-106.)

Grunfeld, Y. (1958), "The Determinants of Corporate Investment," unpublished thesis, Chicago, discussed in Boot, J.C.G. (1960), "Investment Demand: An Empirical Contribution to the Aggregation Problem," *International Economic Review*, 1, 3–30.

Hocking, R.R. (1976), "The Analysis and Selection of Variables in Linear Regression," *Biometrics*, 32, 1–50.

Johnston, J. (1972), *Econometric Methods*, New York: McGraw-Hill Book Co.

Judge, G.G., Griffiths, W.E., Hill, R.C., and Lee, T. (1980), *The Theory and Practice of Econometrics*, New York: John Wiley & Sons, Inc.

Judge, G.G., Griffiths, W.E., Hill, R.C., Lutkepohl, H., and Lee, T.C. (1985), "The Theory and Practice of Econometrics," Second Edition, New York: John Wiley & Sons, Inc.

Kennedy, W.J. and Gentle, J.E. (1980), *Statistical Computing*, New York: Marcel Dekker, Inc.

Lewis, T. and Taylor, L.R. (1967), *Introduction to Experimental Ecology*, New York: Academic Press, Inc.

LaMotte, L.R. (1994), "A Note on the Role of Independence in t Statistics Constructed From Linear Statistics in Regression Models," *The American Statistician*, 48, 238–240.

Lord, F.M. (1950), "Efficiency of Prediction when a Progression Equation from One Sample is Used in a New Sample," Research Bulletin No. 50-40, Princeton, NJ: Educational Testing Service.

Mallows, C.L. (1967), "Choosing a Subset Regression," unpublished report, Bell Telephone Laboratories.

Mallows, C.L. (1973), "Some Comments on C_p," *Technometrics*, 15, 661–675.

Mardia, K.V., Kent, J.T., and Bibby, J.M. (1979), *Multivariate Analysis*, London: Academic Press, Inc.

Markov, A.A. (1900), *Wahrscheinlichkeitsrechnung*, Tebrer, Leipzig.

Marquardt, D.W. and Snee, R.D. (1975), "Ridge Regression in Practice," *American Statistician*, 29 (1), 3–20.

Morrison, D.F. (1976), *Multivariate Statistical Methods*, Second Edition, New York: McGraw-Hill, Inc.

Mosteller, F. and Tukey, J.W. (1977), *Data Analysis and Regression*, Reading, MA: Addison-Wesley Publishing Co., Inc.

Neter, J., Wasserman, W., and Kutner, M.H. (1990), *Applied Linear Statistical Models*, Homewood, Illinois: Richard D. Irwin, Inc.

Neter, J., Wasserman, W., and Kutner, M. H. (1990), *Applied Linear Statistical Models*, Third Edition, Homewood, IL: Irwin.

Nicholson, G.E., Jr. (1948), "The Application of a Regression Equation to a New Sample," unpublished Ph.D. dissertation, University of North Carolina at Chapel Hill.

Pillai, K.C.S. (1960), *Statistical Table for Tests of Multivariate Hypotheses*, Manila: The Statistical Center, University of the Philippines.

Pindyck, R.S. and Rubinfeld, D.L. (1981), *Econometric Models and Econometric Forecasts*, Second Edition, New York: McGraw-Hill Book Co.

Pringle, R.M. and Raynor, A.A. (1971), *Generalized Inverse Matrices with Applications to Statistics*, New York: Hafner Publishing Company.

Rao, C.R. (1973), *Linear Statistical Inference and Its Applications*, Second Edition, New York: John Wiley & Sons, Inc.

Rawlings, J.O. (1988), *Applied Regression Analysis: A Research Tool*, Belmont, California: Wadsworth, Inc.

Rothman, D. (1968), Letter to the editor, *Technometrics*, 10, 432.

Sall, J.P. (1981), *SAS Regression Applications*, Revised Edition, SAS Technical Report A-102, Cary, NC: SAS Institute Inc.

Sawa, T. (1978), "Information Criteria for Discriminating Among Alternative Regression Models," *Econometrica*, 46, 1273–1282.

Schwarz, G. (1978), "Estimating the Dimension of a Model," *Annals of Statistics*, 6, 461–464.

Stein, C. (1960), "Multiple Regression," in *Contributions to Probability and Statistics*, eds. I. Olkin et al., Stanford, CA: Stanford University Press.

Timm, N.H. (1975), *Multivariate Analysis with Applications in Education and Psychology*, Monterey, CA: Brooks-Cole Publishing Co.

Weisberg, S. (1980), *Applied Linear Regression*, New York: John Wiley & Sons, Inc.

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrics*, 48, 817–838.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc., *SAS/STAT® User's Guide, Version 8*, Cary, NC: SAS Institute Inc., 1999.

SAS/STAT® User's Guide, Version 8

ISBN 1–58025–494–2

SAS Institute Inc., SAS Campus Drive, Cary, North Carolina 27513.

1st printing, October 1999