

Chapter 3 Regression

3.1 Exploratory Data Analysis

Objectives

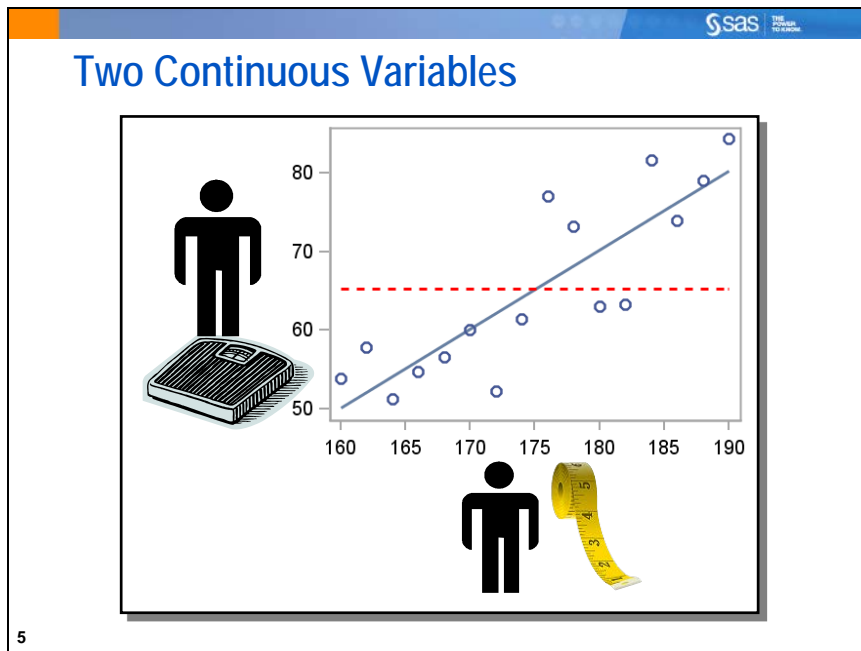
- Use a scatter plot to examine the relationship between two continuous variables.
- Use correlation statistics to quantify the degree of association between two continuous variables.
- Describe potential misuses of the correlation coefficient.
- Use the CORR procedure to obtain Pearson correlation coefficients.

3

Overview of Statistical Models

Type of Response \ Type of Predictors	Categorical	Continuous	Continuous and Categorical
Continuous	Analysis of Variance (ANOVA)	Ordinary Least Squares (OLS) Regression	Analysis of Covariance (ANCOVA)
Categorical	Contingency Table Analysis or Logistic Regression	Logistic Regression	Logistic Regression

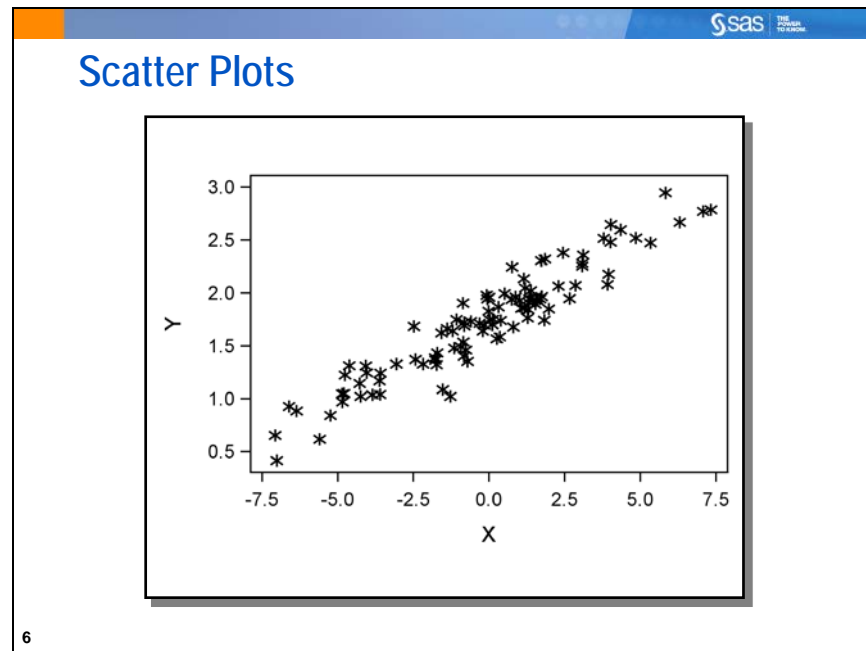
4



In the previous chapter, you learned that when you have a discrete predictor variable and a continuous outcome variable you use ANOVA to analyze your data. In this section, you have two continuous variables.

You use correlation analysis to examine and describe the relationship between two continuous variables. However, before you use correlation analysis, it is important to view the relationship between two continuous variables using a scatter plot.

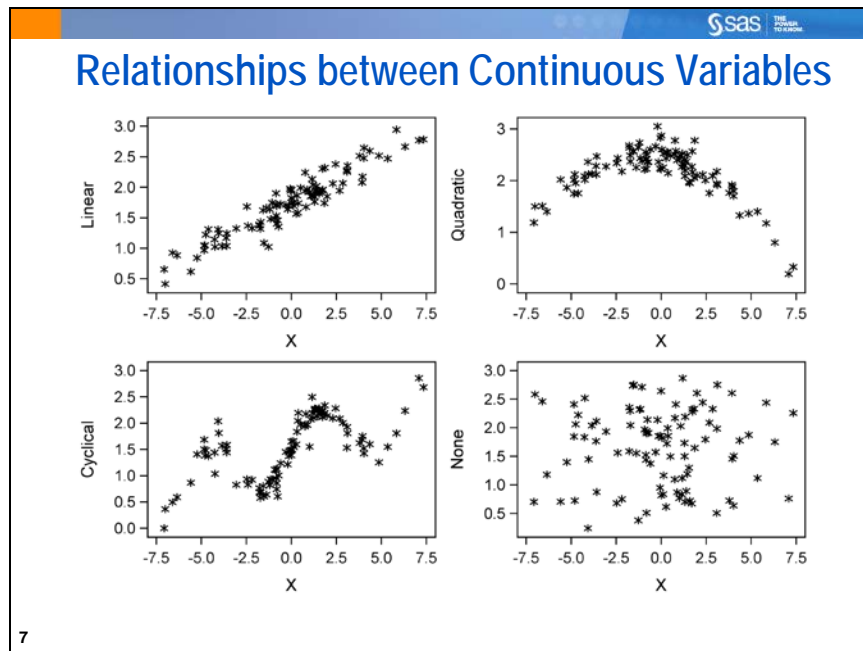
Example: A random sample of high school students is selected to determine the relationship between a person's height and weight. Height and weight are measured on a numeric scale. They have a large, potentially infinite number of possible values, rather than a few categories such as short, medium, and tall. Therefore, these variables are considered to be continuous.



Scatter plots are two-dimensional graphs produced by plotting one variable against another within a set of coordinate axes. The coordinates of each point correspond to the values of the two variables.

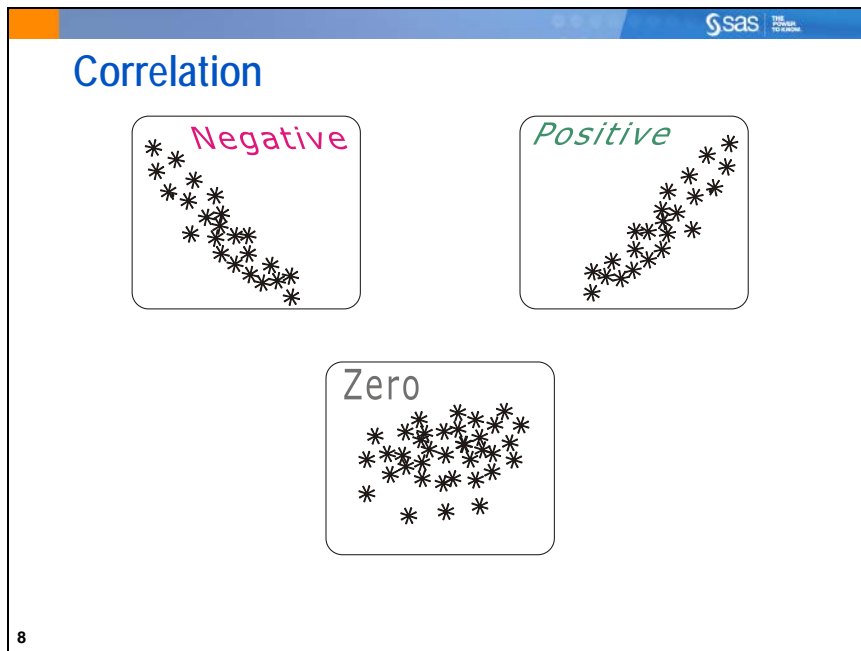
Scatter plots are useful to accomplish the following:

- explore the relationships between two variables
- locate outlying or unusual values
- identify possible trends
- identify a basic range of Y and X values
- communicate data analysis results



Describing the relationship between two continuous variables is an important first step in any statistical analysis. The scatter plot is the most important tool that you have in describing these relationships. The diagrams above illustrate some possible relationships.

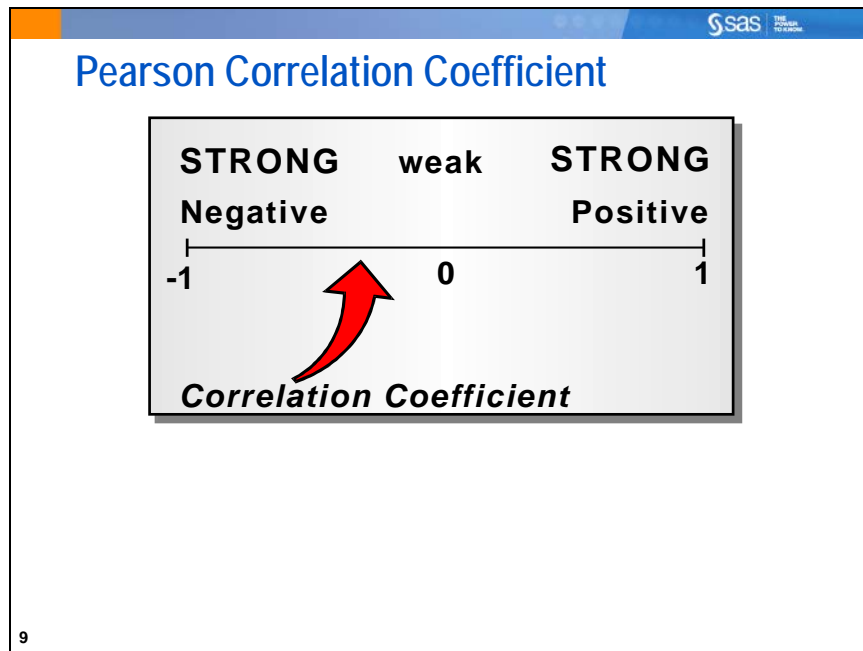
1. A straight line describes the relationship.
2. Curvature is present in the relationship.
3. There could be a cyclical pattern in the relationship. You might see this when the predictor is time.
4. There is no clear relationship between the variables.



As you examine the scatter plot, you can also quantify the relationship between two variables with correlation statistics. Two variables are correlated if there is a **linear** association between them. If not, the variables are uncorrelated.

You can classify correlated variables according to the type of correlation:

- | | |
|----------|---|
| Positive | One variable tends to increase in value as the other variable increases in value. |
| Negative | One variable tends to decrease in value as the other variable increases in value. |
| Zero | No linear relationship exists between the two variables (uncorrelated). |



Correlation statistics measure the degree of linear association between two variables. A common correlation statistic used for continuous variables is the Pearson correlation coefficient. Values of correlation statistics are as follows:


- between -1 and 1
- closer to either extreme if there is a high degree of linear association between the two variables
- close to 0 if there is no linear association between the two variables
- greater than 0 if there is a positive linear association
- less than 0 if there is a negative linear association

Hypothesis Test for a Correlation

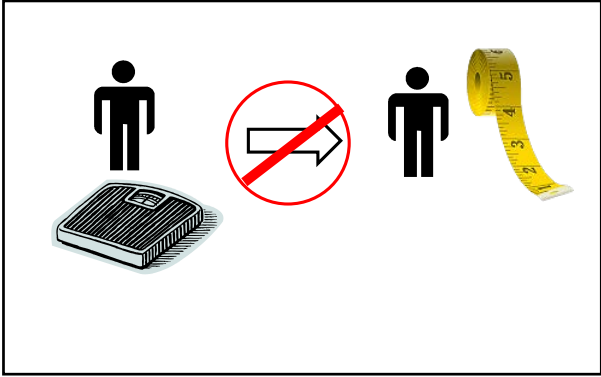
- The parameter representing correlation is ρ .
- ρ is estimated by the sample statistic r .
- $H_0: \rho=0$
- Rejecting H_0 indicates only great confidence that ρ is not exactly zero.
- A p -value does not measure the magnitude of the association.
- Sample size affects the p -value.

10

The null hypothesis for a test of a correlation coefficient is $\rho=0$. Rejecting the null hypothesis only means that you can be confident that the true population correlation is not 0. Small p -values can occur (as with many statistics) because of very large sample sizes. Even a correlation coefficient of 0.01 can be statistically significant with a large enough sample size. Therefore, it is important to also look at the value of r itself to see whether it is meaningfully large.



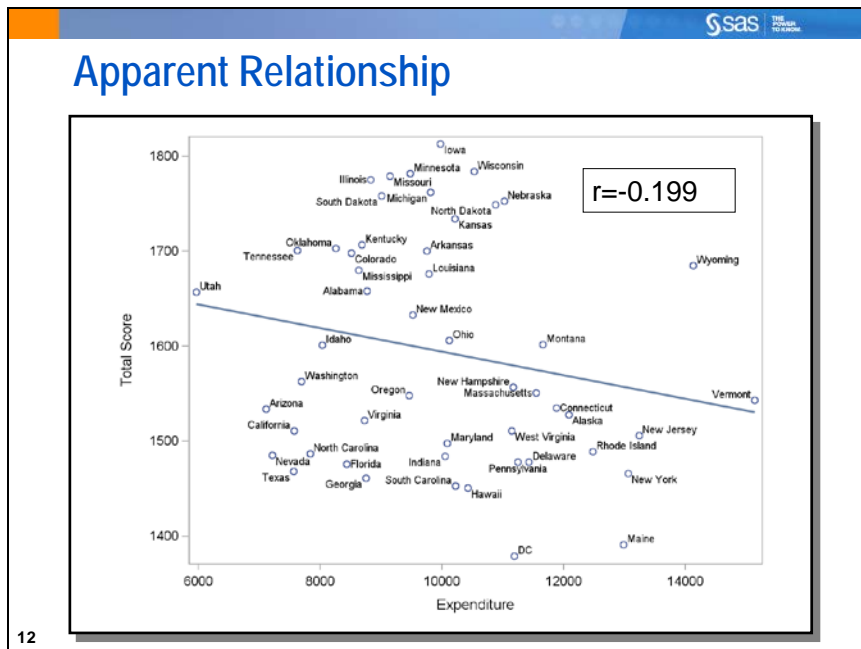
Correlation versus Causation



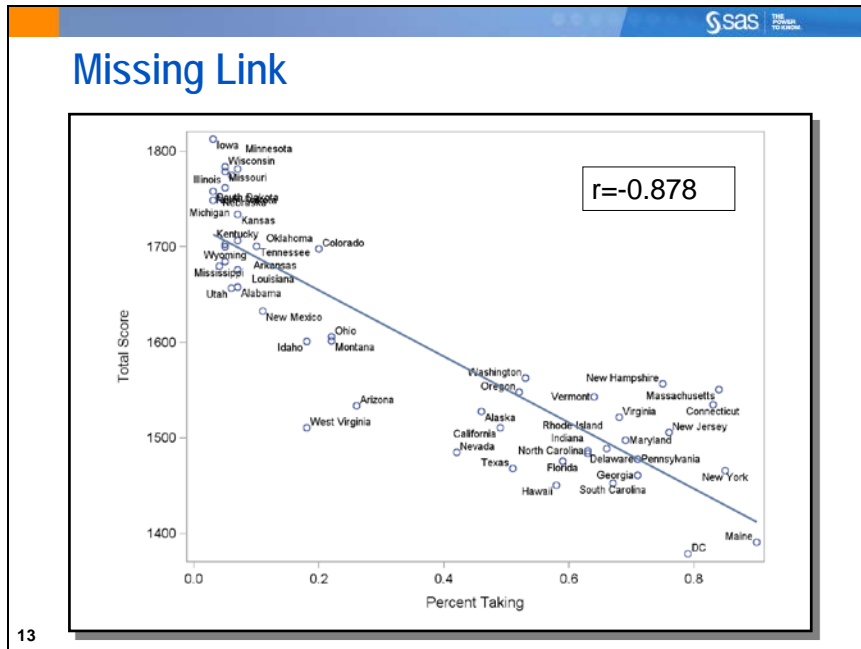
11

Common errors can be made when you interpret the correlation between variables. One example of this is using correlation coefficients to conclude a cause-and-effect relationship.

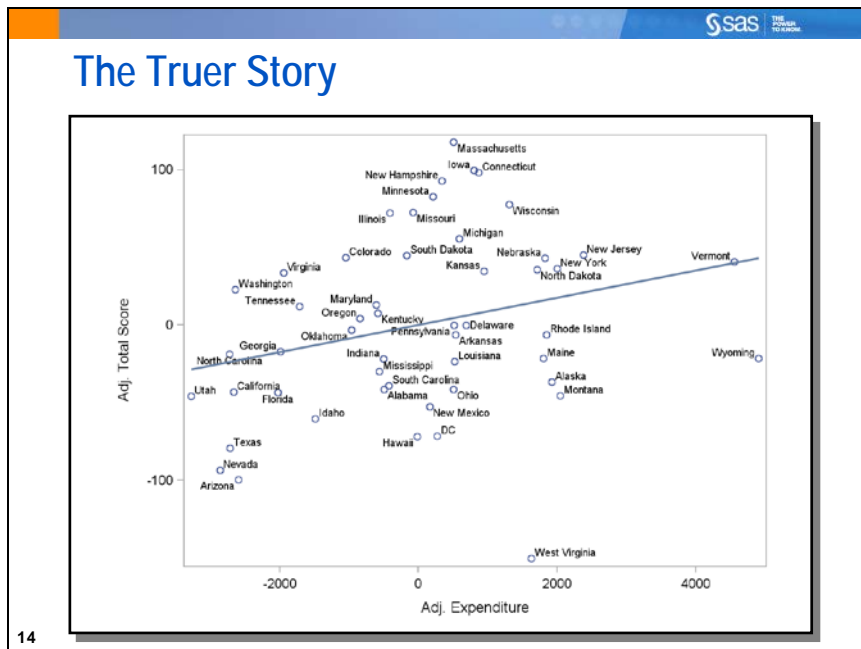
- A strong correlation between two variables does not mean change in one variable causes the other variable to change, or vice versa.
- Sample correlation coefficients can be large because of chance or because both variables are affected by other variables.
- “Correlation does not imply causation.”



An example of reaching errant conclusions comes from U.S. Department of Education data from the Scholastic Aptitude Test (SAT) from 2005. The scatter plot above shows each state's average total SAT score versus the average state expenditure in U.S. dollars per public school student. The correlation between the two variables is -0.199 . Looking at the plot and at this statistic, you might argue (and many argued) that more state spending does little or might even hurt student performance.

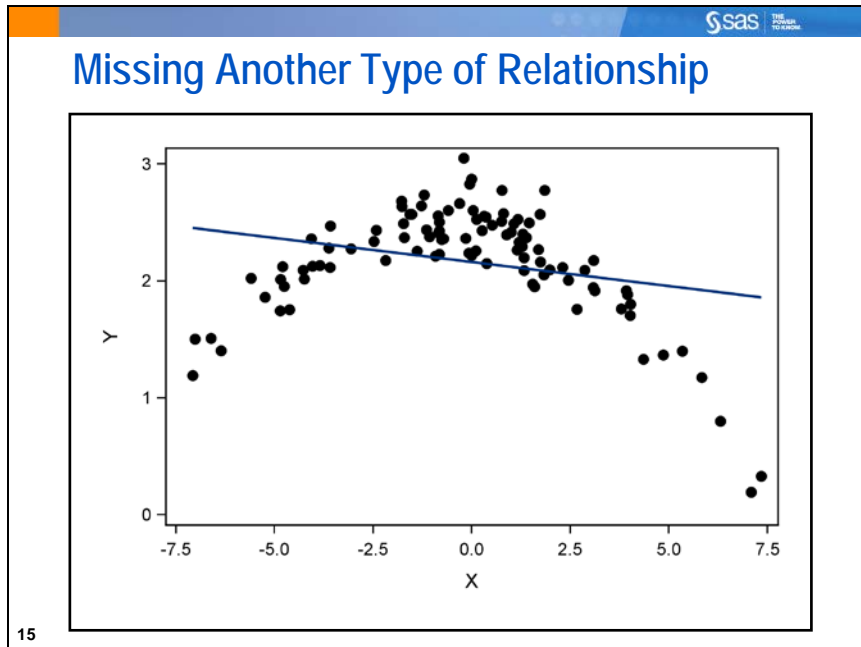


The 2005 report did not take into account the differences among the states in the percentage of students taking the SAT. There are many reasons for the varying participation rates. Some states have lower participation because their students primarily take the rival ACT standardized test. Others have rules requiring even non-college-bound students to take the test. In low participating states, often only the highest performing students choose to take the SAT. Another reported table shows the relationship between participation rate (percent taking the SAT) and average SAT total score. The correlation is -0.878 , indicating that states with lower participation rates tend to have higher average scores.



If you adjust for differences in participation rates, the conclusions about the effect of expenditures might change. In this case, there seems to be a slight positive linear relationship between expenditures and average total score on the SAT when you first adjust for participation rates. (These types of adjustments are described in greater detail in the sections about multiple regression.)

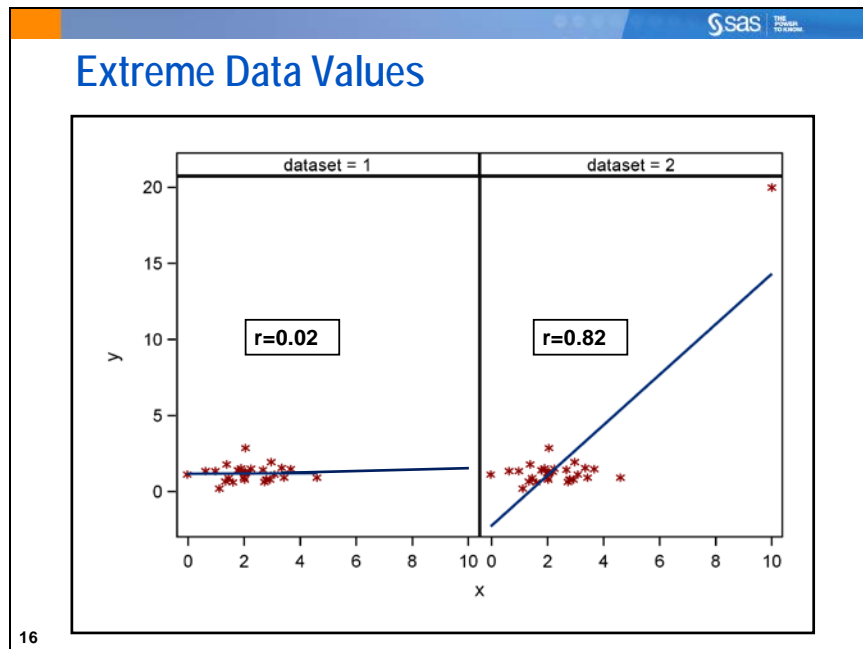
Simple correlations often do not tell the whole story.



In the scatter plot, the variables have a fairly low Pearson correlation coefficient. Why?

- Pearson correlation coefficients measure linear relationships.
- A Pearson correlation coefficient close to 0 indicates that there is not a strong linear relationship between two variables.
- A Pearson correlation coefficient close to 0 does not mean that there is no relationship of any kind between the two variables.

In this example, there is a curvilinear relationship between the two variables.



Correlation coefficients are highly affected by a few extreme values on either variable's range. The scatter plots show that the degree of linear relationship is mainly determined by one point. If you include the unusual point in the data set, the correlation is close to 1. If you do not include it, the correlation is close to 0.

In this situation, follow these steps:

1. Investigate the unusual data point to make sure it is valid.
2. If the data point is valid, collect more data between the unusual data point and the group of data points to see whether a linear relationship unfolds.
3. Try to replicate the unusual data point by collecting data at a fixed value of x (in this case, $x=10$). This determines whether the data point is unusual.
4. Compute two correlation coefficients, one with the unusual data point and one without it. This shows how influential the unusual data point is in the analysis. In this case, it is greatly influential.

The CORR Procedure

General form of the CORR procedure:


```
PROC CORR DATA=SAS-data-set <options>;  
  VAR variables;  
  WITH variables;  
  ID variables;  
RUN;
```

17

You can use the CORR procedure to produce correlation statistics and scatter plots for your data. By default, PROC CORR produces Pearson correlation statistics and corresponding p -values.

Selected CORR procedure statements:

- VAR** specifies variables for which to produce correlations. If a **WITH** statement is not specified, correlations are produced for each pair of variables in the **VAR** statement. If the **WITH** statement is specified, the **VAR** statement specifies the column variables in the correlation matrix.
- WITH** produces correlations for each variable in the **VAR** statement with all variables in the **WITH** statement. The **WITH** statement specifies the row variables in the correlation matrix.
- ID** specifies one or more additional tip variables to identify observations in scatter plots and scatter plot matrices.



The CORR Procedure

- Scatter plots and scatter plot matrices are available through ODS Graphics.
- The ID statement enables you to specify additional variables to identify observations in scatter plots and scatter plot matrices.

18

Exploratory analysis in preparation for multiple regression often involves looking at bivariate scatter plots and correlations between each of the predictor variables and the response variable. It is not suggested that exclusion or inclusion decisions be made on the basis of these analyses. The purpose is to explore the shape of the relationships (because linear regression assumes a linear shape to the relationship) and to screen for outliers. You also want to check for multivariate outliers when you test your multiple regression models later.

PROC CORR provides bivariate correlation tables. These tables are accompanied by ODS Statistical Graphics. An ID statement in the procedure helps identify outliers in the plots.

PROC CORR PLOTS OPTION: Syntax and Selected Sub-Options

- **PLOTS** <(ONLY)> <= (*plot-request* < *plot-request* >) >
 - ALL
 - **MATRIX** <(*matrix-options*)>
 - **SCATTER** <(*scatter-options*)>
 - **HIST** | **HISTOGRAM**
 - **NVAR=ALL** | *n*
 - **ELLIPSE=PREDICTION** | **CONFIDENCE** | **NONE**

19

Selected PLOTS= sub-options:

MATRIX <(*matrix-options*)> requests a scatter plot matrix for variables.

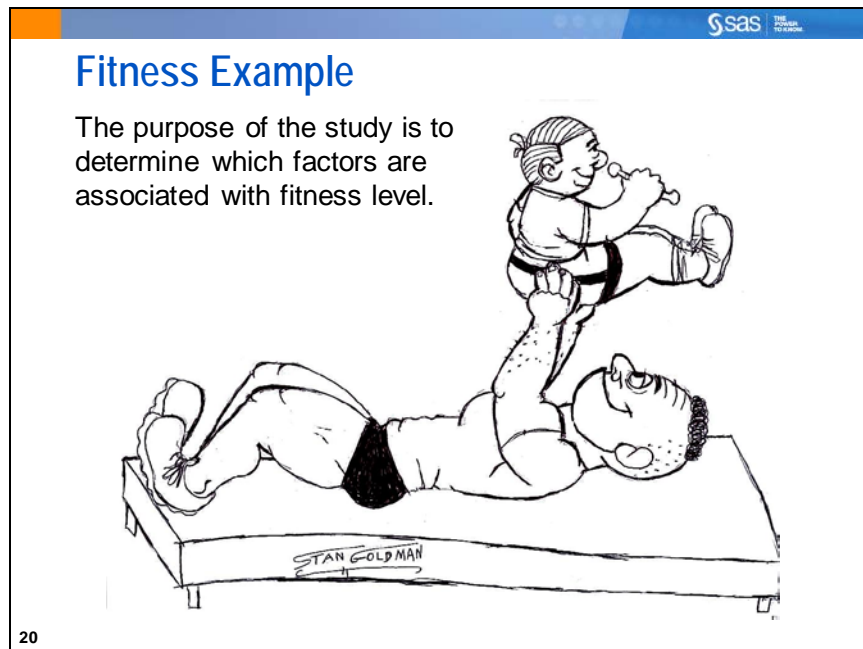
SCATTER <(*scatter-options*)> requests scatter plots for pairs of variables. When a scatter plot or a scatter plot matrix is requested, the Pearson correlations are also displayed.

The available *matrix-options* are as follows:

HIST | **HISTOGRAM** displays histograms of variables in the VAR list in the scatter plot matrix.

NVAR=ALL | *n* specifies the maximum number of variables in the VAR list to be displayed in the scatter plot matrix. By default, NVAR=5.

ELLIPSE= requests prediction ellipses for new observations (ELLIPSE=PREDICTION), confidence ellipses for the mean (ELLIPSE=CONFIDENCE), or no ellipses (ELLIPSE=NONE) to be created in the scatter plots. By default, ELLIPSE=PREDICTION.



20

In exercise physiology, an objective measure of aerobic fitness is how efficiently the body can absorb and use oxygen (oxygen consumption). Subjects participated in a predetermined exercise run of 1.5 miles. Measurements of oxygen consumption as well as several other continuous measurements such as age, pulse, and weight were recorded. The researchers are interested in determining whether any of these other variables can help predict oxygen consumption. These data are found in Rawlings (1998) but certain values of **Maximum_Pulse** and **Run_Pulse** were changed for illustration. **Name**, **Gender**, and **Performance** were also modified for illustration.

The **sasuser.fitness** data set contains the following variables:

Name	name of the member
Gender	gender of the member
RunTime	time to run 1.5 miles (in minutes)
Age	age of the member (in years)
Weight	weight of the member (in kilograms)
Oxygen_Consumption	a measure of the ability to use oxygen in the blood stream
Run_Pulse	pulse rate at the end of the run
Rest_Pulse	resting pulse rate
Maximum_Pulse	maximum pulse rate during the run
Performance	a measure of overall fitness



Data Exploration, Correlations, and Scatter Plots

Examine the relationships between **Oxygen_Consumption** and the continuous predictor variables in the data set. Use the CORR procedure.

```
/*st103d01.sas*/ /*Part A*/
ods graphics / reset=all imagemap;
proc corr data=sasuser.fitness rank
      plots(only)=scatter(nvar=all ellipse=none);
  var RunTime Age Weight Run_Pulse
      Rest_Pulse Maximum_Pulse Performance;
  with Oxygen_Consumption;
  id name;
  title "Correlations and Scatter Plots with Oxygen_Consumption";
run;
```



IMAGEMAP=ON in the ODS GRAPHICS statement enables tooltips to be used in HTML output. Tooltips are also functional in SAS Report output when you use SAS Enterprise Guide, starting with Version 4.3. Tooltips enable the user to identify data points by moving the cursor over observations in a plot. In PROC CORR, the variables used in the tooltips are the X axis and Y axis variables, the observation number, and any variable in the ID statement.

Selected PROC CORR statement options:

RANK	orders the correlations from highest to lowest in absolute value.
PLOTS	creates scatter plots and scatter plot matrices using ODS GRAPHICS.

Selected PROC CORR statement:

ID	when used in HTML output with IMAGEMAP, adds the listed variables to the information available with tooltips.
----	---

Suboptions for the PLOTS option:

SCATTER	generates scatter plots for pairs of variables.
---------	---

Suboptions for the SCATTER sub-option:

NVAR=<k>	specifies the maximum number of variables in the VAR list to be displayed in the matrix plot. If NVAR=ALL is specified, then all variables in the VAR list (up to a limit of 10) are displayed.
----------	---

ELLIPSE=NONE	suppresses the drawing of confidence ellipses on scatter plots.
--------------	---

The tabular output from PROC CORR is shown below. By default, the analysis generates a table of univariate statistics for the analysis variables and then a table of correlations and *p*-values.

PROC CORR Output

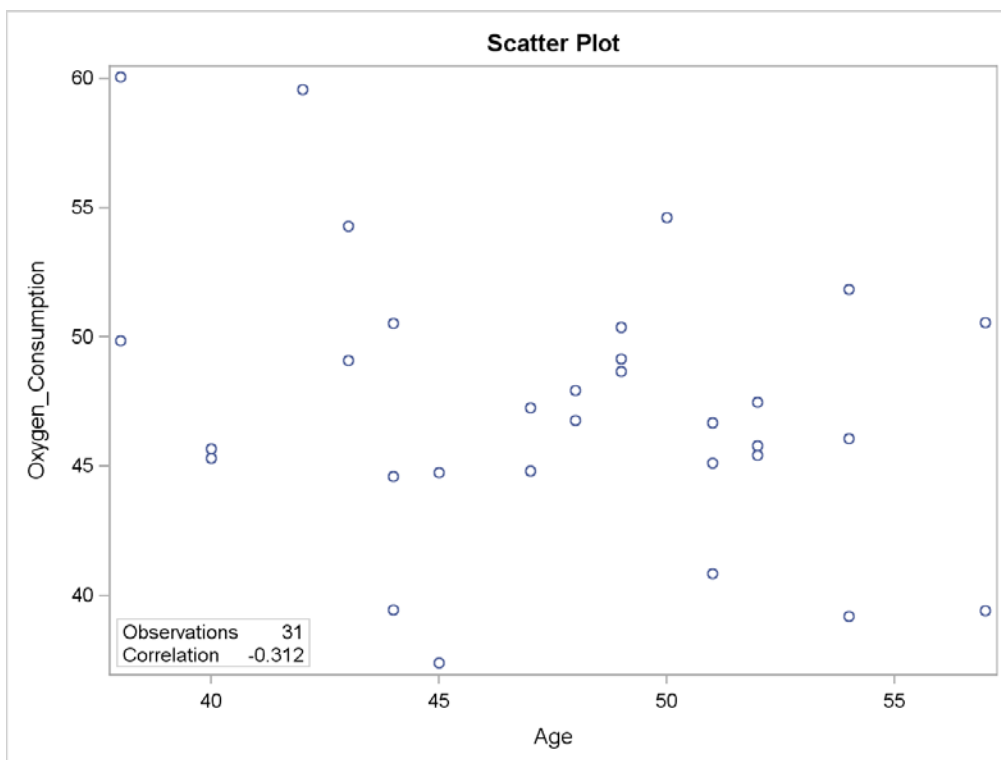
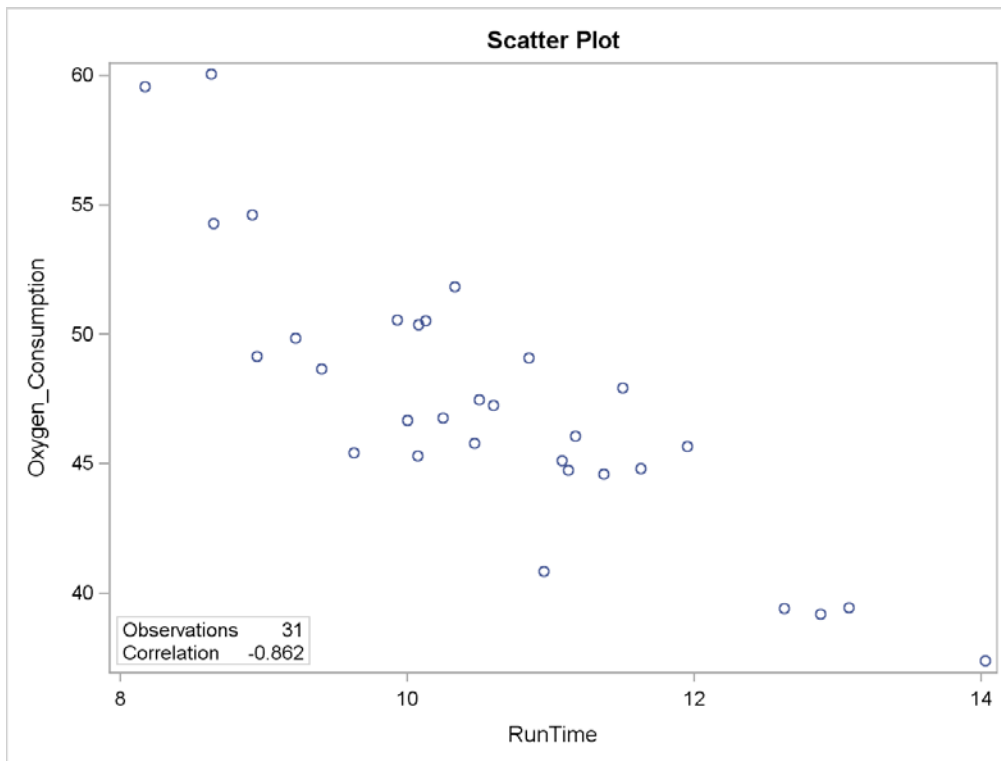
1 With Variables:	Oxygen_Consumption
7 Variables:	RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance

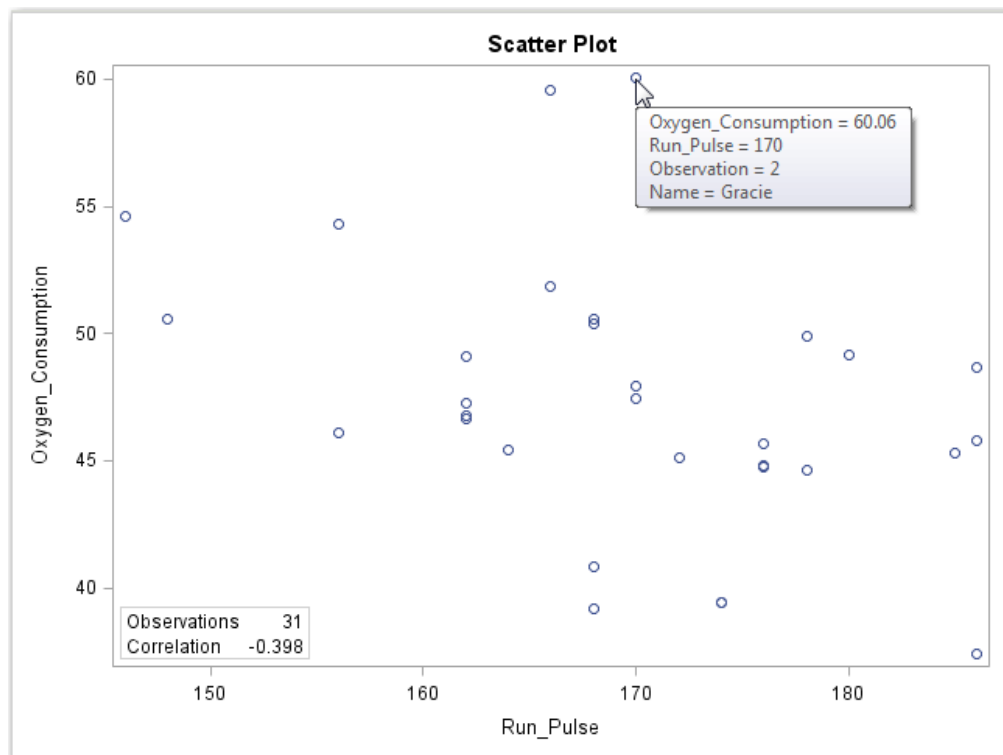
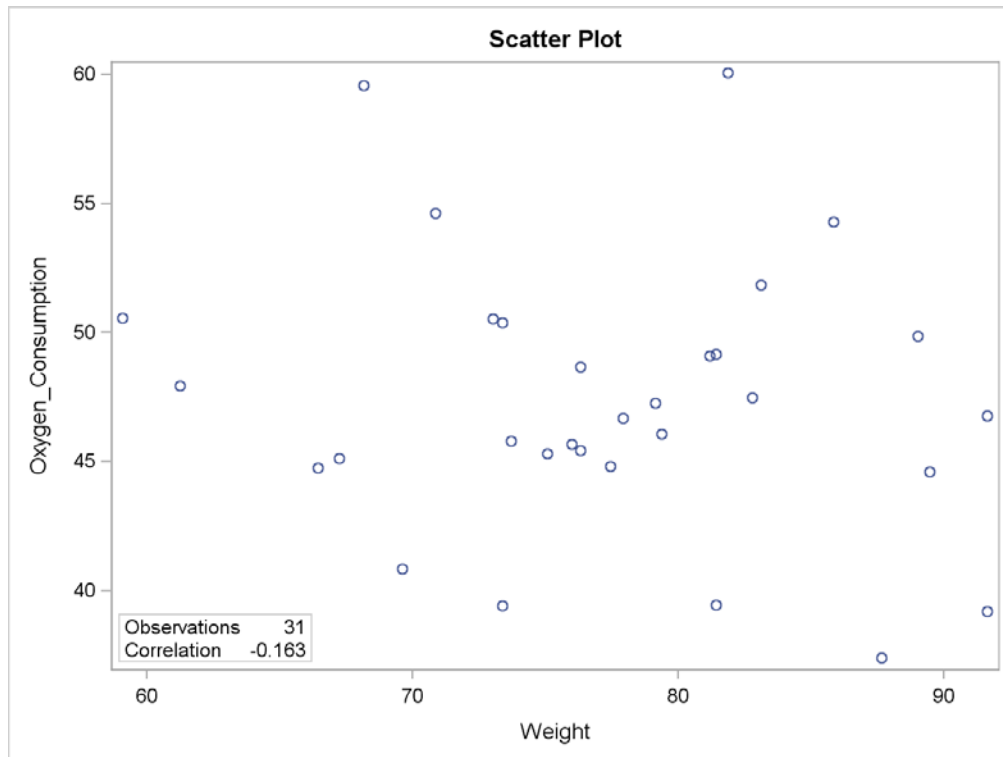
Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Oxygen_Consumption	31	47.37581	5.32777	1469	37.39000	60.06000
RunTime	31	10.58613	1.38741	328.17000	8.17000	14.03000
Age	31	47.67742	5.26236	1478	38.00000	57.00000
Weight	31	77.44452	8.32857	2401	59.08000	91.63000
Run_Pulse	31	169.64516	10.25199	5259	146.00000	186.00000
Rest_Pulse	31	53.45161	7.61944	1657	40.00000	70.00000
Maximum_Pulse	31	173.77419	9.16410	5387	155.00000	192.00000
Performance	31	56.64516	18.32584	1756	20.00000	94.00000

Pearson Correlation Coefficients, N = 31 Prob > r under H0: Rho=0							
Oxygen_Consumption	RunTime	Performance	Rest_Pulse	Run_Pulse	Age	Maximum_Pulse	Weight
	-0.86219	0.77890	-0.39935	-0.39808	-0.31162	-0.23677	-0.16289
	<.0001	<.0001	0.0260	0.0266	0.0879	0.1997	0.3813

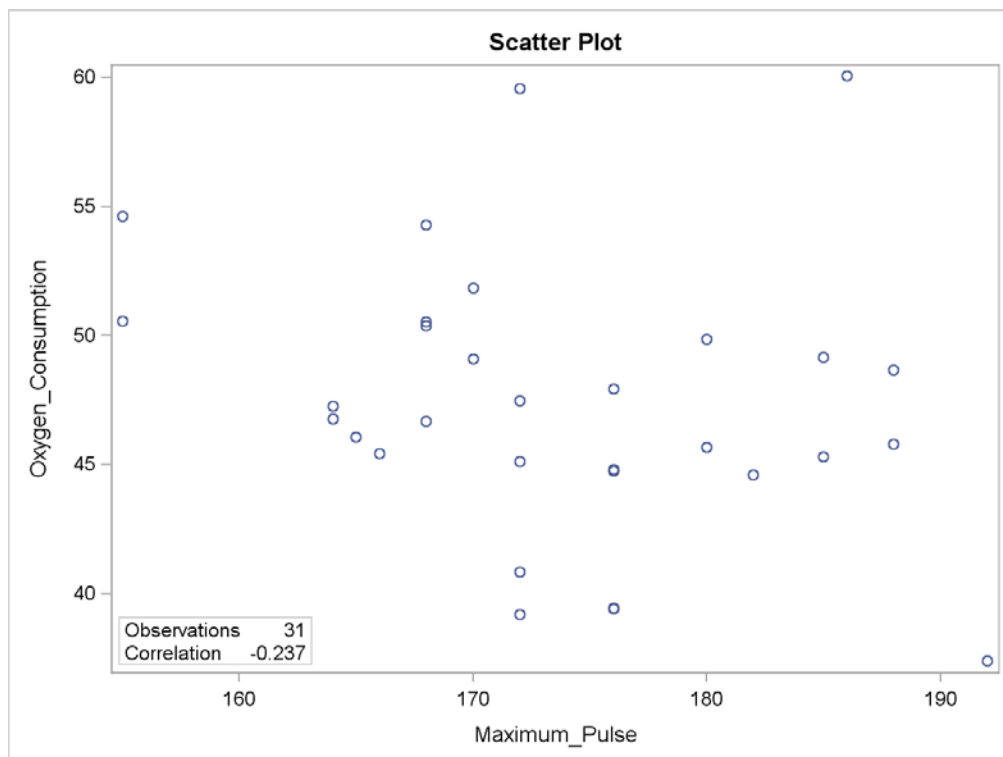
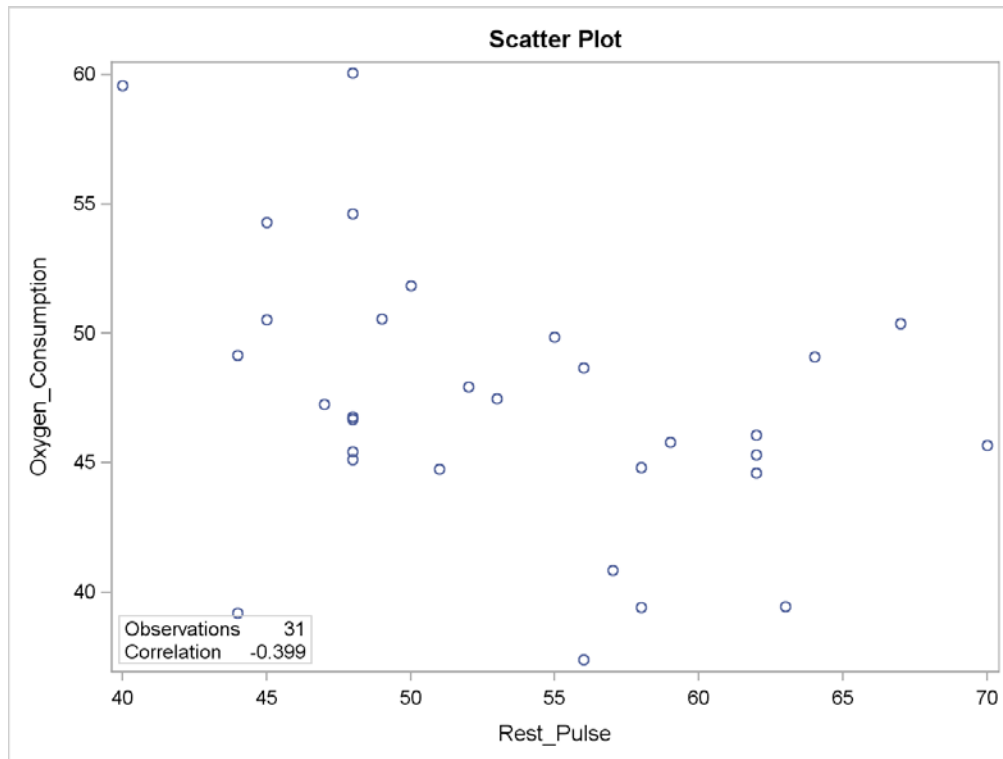
The correlation coefficient between **Oxygen_Consumption** and **RunTime** is -0.86219. The p -value is small, which indicates that the population correlation coefficient (ρ) is likely different from 0. The second largest correlation coefficient, in absolute value, is **Performance**, at 0.77890.

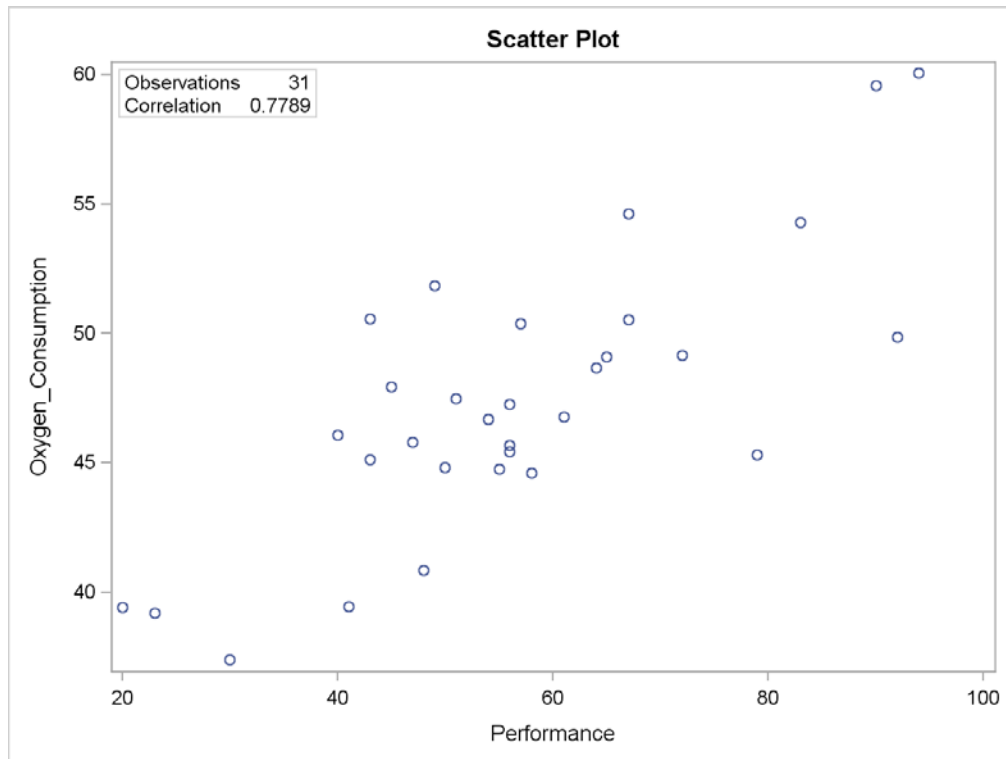
Scatter plots associated with these correlations are shown below.





If you want to explore an observation further, you can move the cursor over the observation and information is displayed in a floating box. You can only do this in an HTML file with IMAGEMAP turned on. The coordinate values, observation number, and ID variable values are displayed.





The correlation and scatter plot analyses indicate that several variables might be good predictors for **Oxygen_Consumption**.

When you prepare to conduct a regression analysis, it is always good practice to examine the correlations among the potential predictor variables. PROC CORR can be used to generate a matrix of correlation coefficients. To ensure that the imagemap feature used in the previous demonstration is deactivated, we will include a RESET=ALL option in the ODS statement.

```
/*st103d01.sas*/ /*Part B*/
ods graphics / reset=all;
proc corr data=sasuser.fitness nosimple
      plots=matrix(nvar=all histogram);
  var RunTime Age Weight Run_Pulse
      Rest_Pulse Maximum_Pulse Performance;
  title "Correlations and Scatter Plot Matrix of Fitness Predictors";
run;
```

Selected PROC CORR statement option:

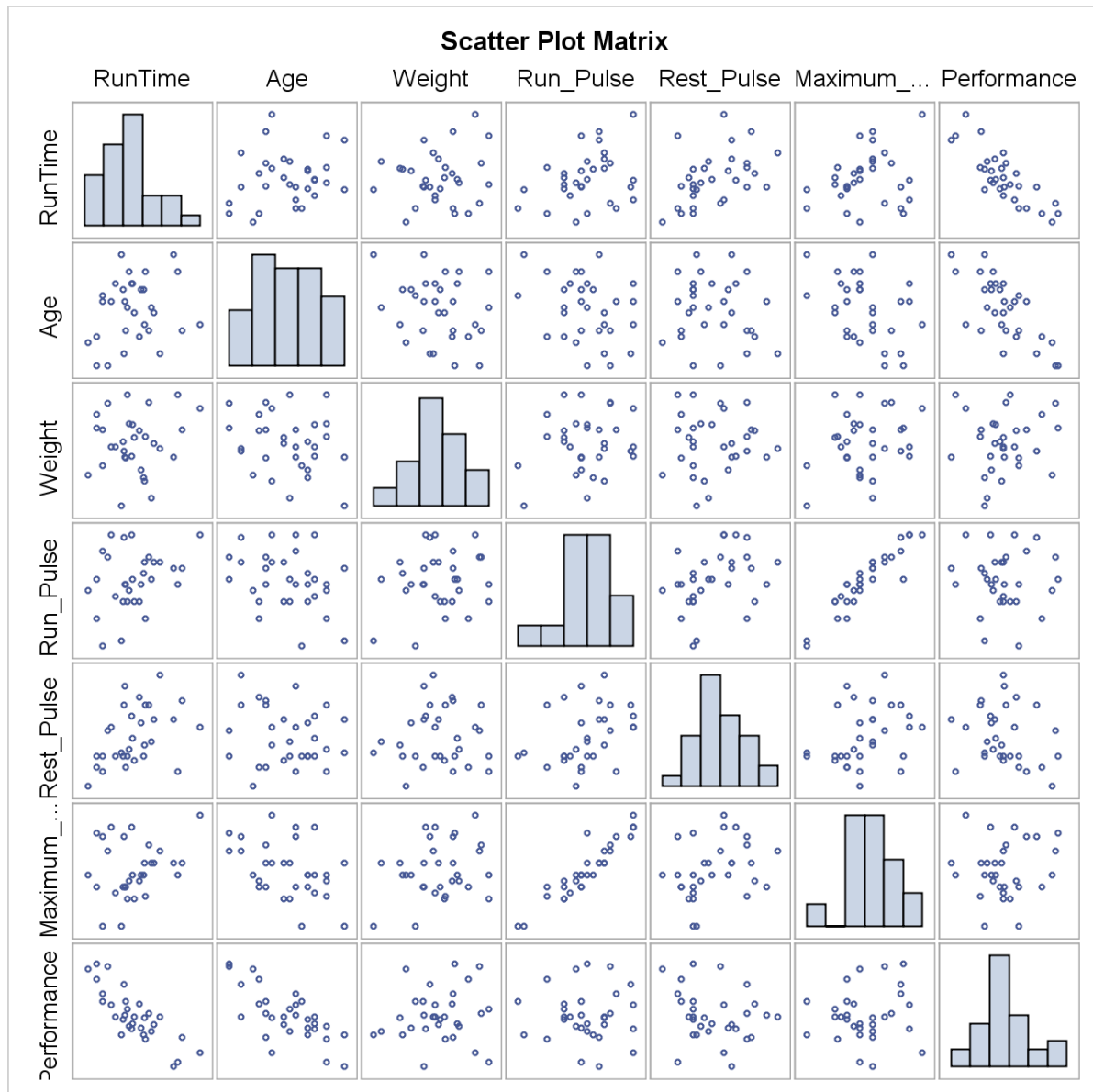
NOSIMPLE suppresses printing simple descriptive statistics for each variable.

PROC CORR Output

7 Variables: RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse Performance

Pearson Correlation Coefficients, N = 31 Prob > r under H0: Rho=0							
	RunTime	Age	Weight	Run_Pulse	Rest_Pulse	Maximum_Pulse	Performance
RunTime	1.00000	0.19523 0.2926	0.14351 0.4412	0.31365 0.0858	0.45038 0.0110	0.22610 0.2213	-0.82049 <.0001
Age	0.19523 0.2926	1.00000	-0.24050 0.1925	-0.31607 0.0832	-0.15087 0.4178	-0.41490 0.0203	-0.71257 <.0001
Weight	0.14351 0.4412	-0.24050 0.1925	1.00000	0.18152 0.3284	0.04397 0.8143	0.24938 0.1761	0.08974 0.6312
Run_Pulse	0.31365 0.0858	-0.31607 0.0832	0.18152 0.3284	1.00000	0.35246 0.0518	0.92975 <.0001	-0.02943 0.8751
Rest_Pulse	0.45038 0.0110	-0.15087 0.4178	0.04397 0.8143	0.35246 0.0518	1.00000	0.30512 0.0951	-0.22560 0.2224
Maximum_Pulse	0.22610 0.2213	-0.41490 0.0203	0.24938 0.1761	0.92975 <.0001	0.30512 0.0951	1.00000	0.09002 0.6301
Performance	-0.82049 <.0001	-0.71257 <.0001	0.08974 0.6312	-0.02943 0.8751	-0.22560 0.2224	0.09002 0.6301	1.00000

There are strong correlations between **Run_Pulse** and **Maximum_Pulse** (0.92975) and between **RunTime** and **Performance** (-0.82049). These associations are seen in more detail in the matrix of scatter plots.



The following correlation table was created from the matrix by choosing small p -values. The table is in descending order, based on the absolute value of the correlation. It provides a summary of the correlation analysis of the independent variables.

<u>Row Variable</u>	<u>Column Variable</u>	<u>Pearson's r</u>	<u>Prob > r</u>
Run_Pulse	Maximum_Pulse	0.92975	<.0001
RunTime	Performance	-0.82049	<.0001
Performance	Age	-0.71257	<.0001
RunTime	Rest_Pulse	0.45038	0.0110
Age	Maximum_Pulse	-0.41490	0.0203
Run_Pulse	Rest_Pulse	0.35246	0.0518



Exercises

1. Describing the Relationship between Continuous Variables

Percentage of body fat, age, weight, height, and 10 body circumference measurements (for example, abdomen) were recorded for 252 men by Dr. Roger W. Johnson of Calvin College in Minnesota. The data are in the **sasuser.BodyFat2** data set. Body fat, one measure of health, was accurately estimated by an underwater weighing technique. There are two measures of percentage body fat in this data set. The following variables are in the data set:

Case	Case Number
PctBodyFat1	Percent body fat using Brozek's equation, $457/\text{Density} - 414.2$
PctBodyFat2	Percent body fat using Siri's equation, $495/\text{Density} - 450$
Density	Density (gm/cm^3)
Age	Age (yrs)
Weight	Weight (lbs)
Height	Height (inches)
Adiposity	Adiposity index = $\text{Weight}/\text{Height}^2$ (kg/m^2)
FatFreeWt	Fat Free Weight = $(1 - \text{fraction of body fat}) * \text{Weight}$, using Brozek's formula (lbs)
Neck	Neck circumference (cm)
Chest	Chest circumference (cm)
Abdomen	Abdomen circumference (cm) "at the umbilicus and level with the iliac crest"
Hip	Hip circumference (cm)
Thigh	Thigh circumference (cm)
Knee	Knee circumference (cm)
Ankle	Ankle circumference (cm)
Biceps	Extended biceps circumference (cm)
Forearm	Forearm circumference (cm)
Wrist	Wrist circumference (cm) "distal to the styloid processes"

- a. Generate scatter plots and correlations for the VAR variables **Age**, **Weight**, **Height**, and the circumference measures versus the WITH variable, **PctBodyFat2**.



Important! ODS Graphics in PROC CORR limits you to 10 VAR variables at a time, so for this exercise, look at the relationships with **Age**, **Weight**, and **Height** separately from the circumference variables (**Neck Chest Abdomen Hip Thigh Knee Ankle Biceps Forearm Wrist**).



This limitation exists only on the graphics obtained from ODS. The correlation table will display all variables in the VAR statement by default.

- 1) Can straight lines adequately describe the relationships?
 - 2) Are there any outliers that you should investigate?
 - 3) What variable has the highest correlation with **PctBodyFat2**?
 - a) What is the p -value for the coefficient?
 - b) Is the correlation statistically significant at the 0.05 level?
- b. Generate correlations among all of the variables in the previously mentioned variables minus **PctBodyFat2**. Are there any notable relationships?

3.01 Multiple Choice Poll

The correlation between tuition and rate of graduation at U.S. colleges is 0.55. What does this mean?

- a. The way to increase graduation rates at your college is to raise tuition.
- b. Increasing graduation rates is expensive, causing tuition to rise.
- c. Students who are richer tend to graduate more often than poorer students.
- d. None of the above.

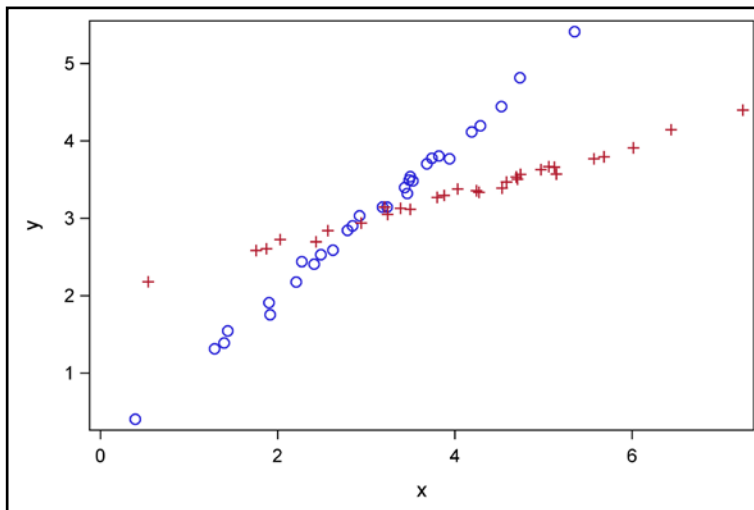
3.2 Simple Linear Regression

Objectives

- Explain the concepts of simple linear regression.
- Fit a simple linear regression using the REG procedure.
- Produce predicted values and confidence intervals.

28

Overview



29

In the last section, you used correlation analysis to quantify the linear relationships between continuous response variables. Two pairs of variables can have the same correlation, but very different linear relationships. In this section, you use simple linear regression to define the linear relationship between a response variable and a predictor variable.

- The *response variable* is the variable of primary interest.
- The *predictor variable* is used to explain the variability in the response variable.

Simple Linear Regression Analysis

The objectives of simple linear regression are as follows:

- assess the significance of the predictor variable in explaining the variability or behavior of the response variable
- predict the values of the response variable given the values of the predictor variable

30

In simple linear regression, the values of the predictor variable are assumed to be fixed. Thus, you try to explain the variability of the response variable given the values of the predictor variable.

Fitness Example

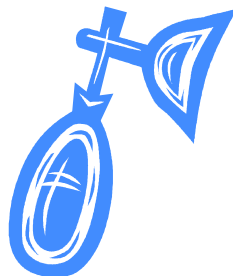
PREDICTOR

RunTime



RESPONSE

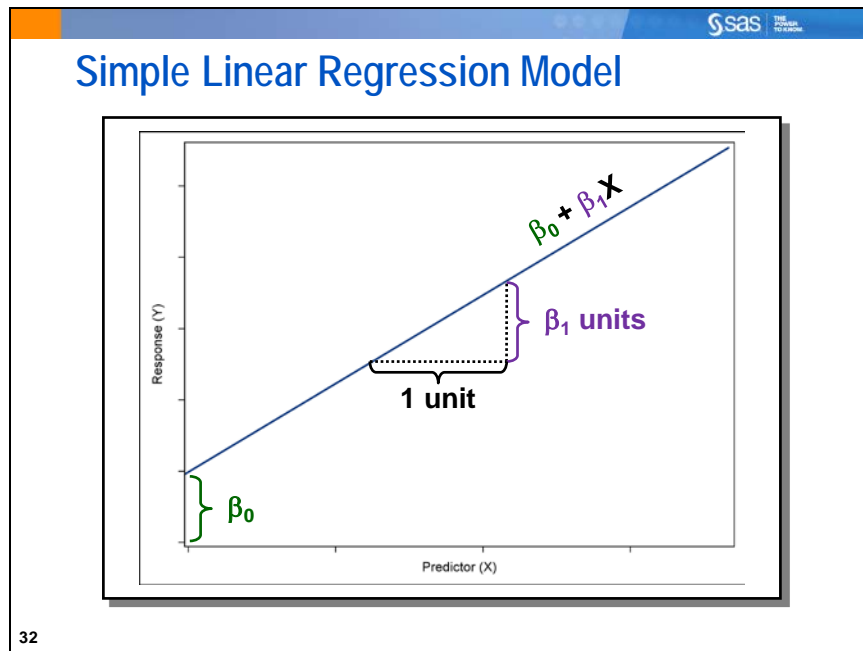
Oxygen_Consumption



31

The analyst noted that the running time measure has the highest correlation with the oxygen consumption capacity of the club members. Consequently, she wants to further explore the relationship between **Oxygen_Consumption** and **RunTime**.

She decides to run a simple linear regression of **Oxygen_Consumption** versus **RunTime**.



The relationship between the response variable and the predictor variable can be characterized by the equation $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1, \dots, n$

where

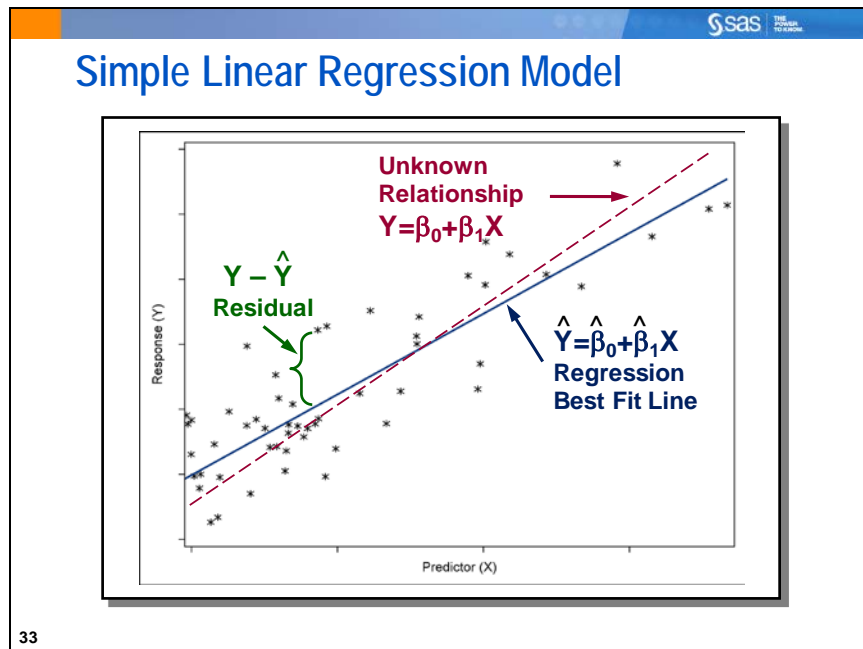
y_i is the response variable.

x_i is the predictor variable.

β_0 is the intercept parameter, which corresponds to the value of the response variable when the predictor is 0.

β_1 is the slope parameter, which corresponds to the magnitude of change in the response variable given a one unit change in the predictor variable.

ε_i is the error term representing deviations of y_i about $\beta_0 + \beta_1 x_i$.



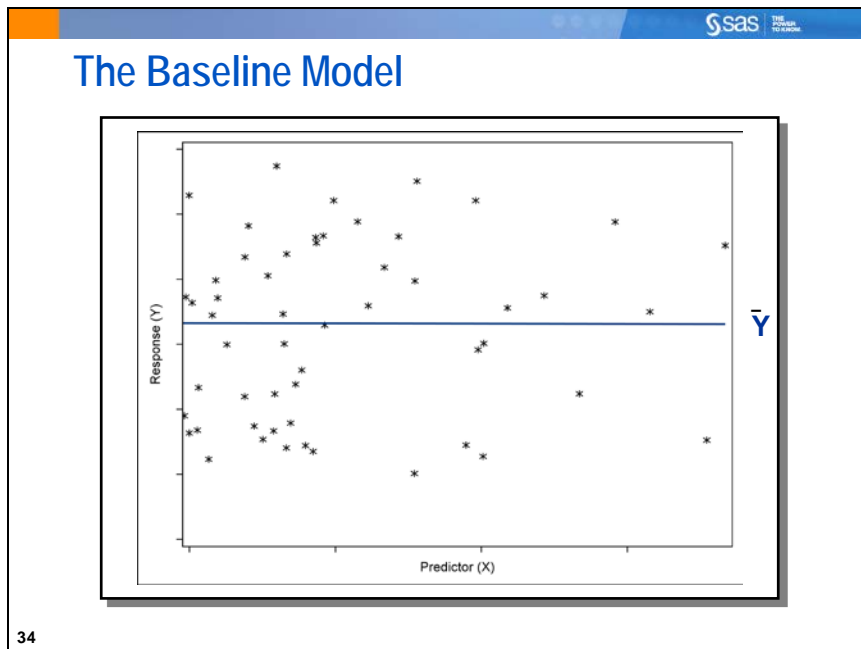
Because your goal in simple linear regression is usually to characterize the relationship between the response and predictor variables in your population, you begin with a sample of data. From this sample, you estimate the unknown population parameters (β_0 , β_1) that define the assumed relationship between your response and predictor variables.

Estimates of the unknown population parameters β_0 and β_1 are obtained by the *method of least squares*. This method provides the estimates by determining the line that minimizes the sum of the squared vertical distances between the observations and the fitted line. In other words, the fitted or regression line is as close as possible to all the data points.

The method of least squares produces parameter estimates with certain optimum properties. If the assumptions of simple linear regression are valid, the least squares estimates are unbiased estimates of the population parameters and have minimum variance (efficiency). The least squares estimators are often called BLUE (Best Linear Unbiased Estimators). The term *best* is used because of the minimum variance property.

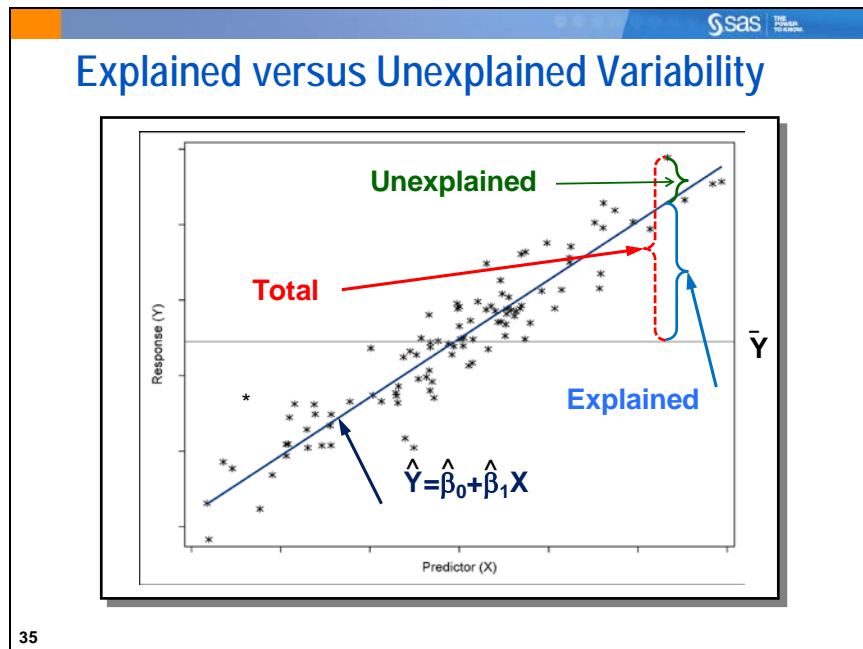
Because of these optimum properties, the method of least squares is used by many data analysts to investigate the relationship between continuous predictor and response variables.

With a large and representative sample, the fitted regression line should be a good approximation of the relationship between the response and predictor variables in the population. The estimated parameters obtained using the method of least squares should be good approximations of the true population parameters.



To determine whether the predictor variable explains a significant amount of variability in the response variable, the simple linear regression model is compared to the baseline model. The fitted regression line in a baseline model is a horizontal line across all values of the predictor variable. The slope of the regression line is 0 and the intercept is the sample mean of the response variable, (\bar{Y}).

In a baseline model, there is no association between the response variable and the predictor variable. Therefore, knowing the value of the predictor variable does not improve predictions of the response over simply using the unconditional mean (the mean calculated disregarding the predictor variables) of the response variable.



To determine whether a simple linear regression model is better than the baseline model, compare the explained variability to the unexplained variability.


Explained variability is related to the difference between the regression line and the mean of the response variable. The model sum of squares (SSM) is the amount of variability explained by your model. The model sum of squares is equal to $\sum(\hat{Y}_i - \bar{Y})^2$.

Unexplained variability is related to the difference between the observed values and the regression line. The error sum of squares (SSE) is the amount of variability unexplained by your model. The error sum of squares is equal to $\sum(Y_i - \hat{Y}_i)^2$.

Total variability is related to the difference between the observed values and the mean of the response variable. The corrected total sum of squares is the sum of the explained and unexplained variability. The corrected total sum of squares is equal to $\sum(Y_i - \bar{Y})^2$.



Remember that the relationship of the following: total=unexplained+explained applies for sums of squares over all observations and not necessarily for any individual observation.



Model Hypothesis Test

Null Hypothesis:

- The simple linear regression model does **not** fit the data better than the baseline model.
- $\beta_1=0$

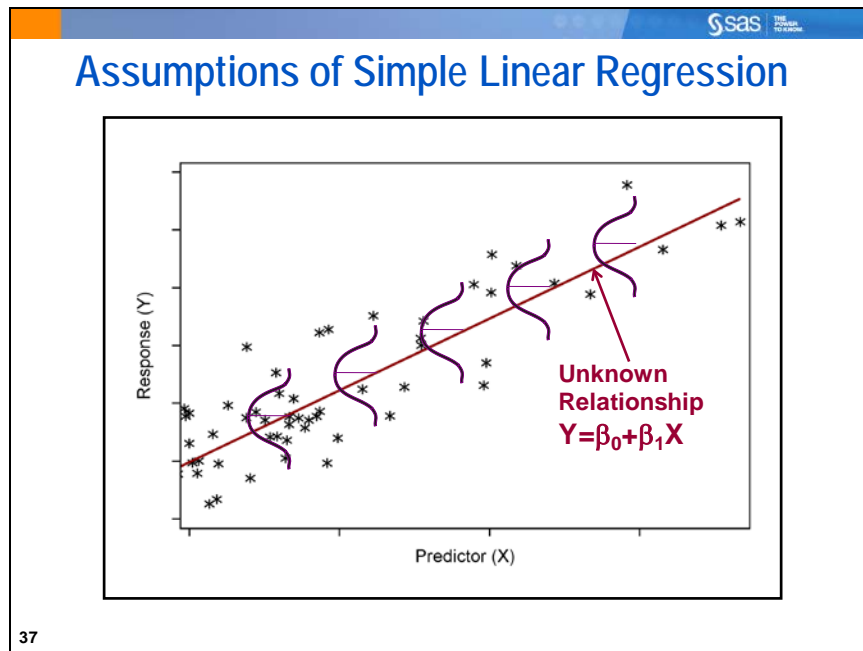
Alternative Hypothesis:

- The simple linear regression model does fit the data better than the baseline model.
- $\beta_1 \neq 0$

36

If the estimated simple linear regression model does **not** fit the data better than the baseline model, you fail to reject the null hypothesis. Thus, you do **not** have enough evidence to say that the slope of the regression line in the population differs from zero.

If the estimated simple linear regression model **does** fit the data better than the baseline model, you reject the null hypothesis. Thus, you **do** have enough evidence to say that the slope of the regression line in the population differs from zero and that the predictor variable explains a significant amount of variability in the response variable.



One of the assumptions of simple linear regression is that the mean of the response variable is linearly related to the value of the predictor variable. In other words, a straight line connects the means of the response variable at each value of the predictor variable.

The other assumptions are the same as the assumptions for ANOVA, that is, the error is normally distributed and has constant variance across the range of the predictor variable, and observations are independent.



The verification of these assumptions is discussed in a later chapter.

The REG Procedure

General form of the REG procedure:

```
PROC REG DATA=SAS-data-set <options>;  
    MODEL dependent(s)=regressor(s) </ options>;  
RUN;  
QUIT;
```

38

The REG procedure enables you to fit regression models to your data.

Selected REG procedure statement:

MODEL specifies the response and predictor variables. The variables must be numeric.



PROC REG supports RUN-group processing, which means that the procedure stays active until a PROC, DATA, or QUIT statement is encountered. This enables you to submit additional statements followed by another RUN statement without resubmitting the PROC statement.



Performing Simple Linear Regression

Example: Because there is an apparent linear relationship between **Oxygen_Consumption** and **RunTime**, perform a simple linear regression analysis with **Oxygen_Consumption** as the response variable.

```
/*st103d02.sas*/
proc reg data=sasuser.fitness;
  model Oxygen_Consumption=RunTime;
  title 'Predicting Oxygen_Consumption from RunTime';
run;
quit;
```

PROC REG Output

Number of Observations Read	31
Number of Observations Used	31

The Number of Observations Read and the Number of Observations Used are the same, which indicates that no missing values were detected for **Oxygen_Consumption** and **RunTime**.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	633.01458	633.01458	84.00	<.0001
Error	29	218.53997	7.53586		
Corrected Total	30	851.55455			

The Analysis of Variance (ANOVA) table provides an analysis of the variability observed in the data and the variability explained by the regression line.

The ANOVA table for simple linear regression is divided into six columns:

Source	labels the source of variability.
DF	is the degrees of freedom associated with each source of variability.
Sum of Squares	is the amount of variability associated with each source of variability.
Mean Square	is the ratio of the sum of squares and the degrees of freedom. This value corresponds to the amount of variability associated with each degree of freedom for each source of variation.
F Value	is the ratio of the mean square for the model and the mean square for the error. This ratio compares the variability explained by the regression line to the variability unexplained by the regression line.
Pr>F	is the p -value associated with the F value.

Each of the column measurements are applied to the following sources of variation:

Model is the variability explained by your model (Between Group).

Error is the variability unexplained by your model (Within Group).

Corrected Total is the total variability in the data (Total).

The F value tests whether the slope of the predictor variable is equal to 0. The p -value is small (less than 0.05), so you have enough evidence at the 0.05 significance level to reject the null hypothesis. Thus, you can conclude that the simple linear regression model fits the data better than the baseline model. In other words, **RunTime** explains a significant amount of variability of **Oxygen_Consumption**.

The third part of the output provides summary measures of fit for the model.

Root MSE	2.74515	R Square	0.7434
Dependent Mean	47.37581	Adj R Sq	0.7345
Coeff Var	5.79442		

Root MSE The root mean square error is an estimate of the standard deviation of the response variable at each value of the predictor variable. It is the square root of the MSE.

Dependent Mean The overall mean of the response variable is \bar{Y} .

Coeff Var The coefficient of variation is the size of the standard deviation relative to the mean. The coefficient of variation is

- calculated as $\left(\frac{RootMSE}{\bar{Y}} \right) * 100$
- a unitless measure, so it can be used to compare data that has different units of measurement or different magnitudes of measurement.

R Square The coefficient of determination is also referred to as the R-square value. This value is

- between 0 and 1.
- the proportion of variability observed in the data explained by the regression line. In this example, the value is 0.7434, which means that the regression line explains 74% of the total variation in the response values.
- the square of the multiple correlation between Y and the Xs.



The R square is the squared value of the correlation that you saw earlier between **RunTime** and **Oxygen_Consumption** (0.86219). This is no coincidence. For simple regression, the R-square value is the square of the value of the bivariate Pearson correlation coefficient.

Adj R Sq The adjusted R square is adjusted for the number of parameters in the model. This statistic is useful in multiple regression and is discussed in a later section.

The Parameter Estimates table defines the model for your data.

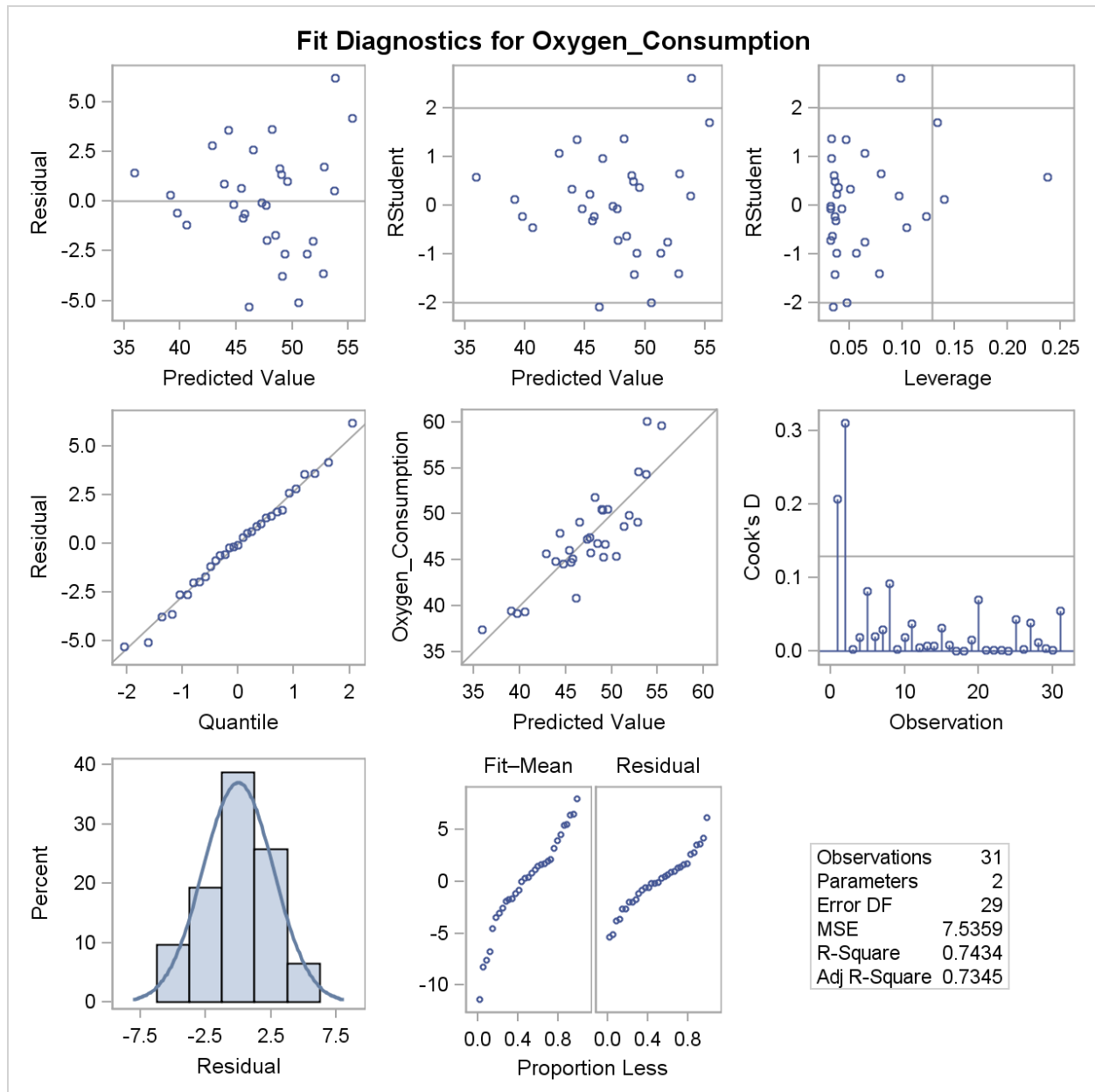
Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	82.42494	3.85582	21.38	<.0001
RunTime	1	-3.31085	0.36124	-9.17	<.0001

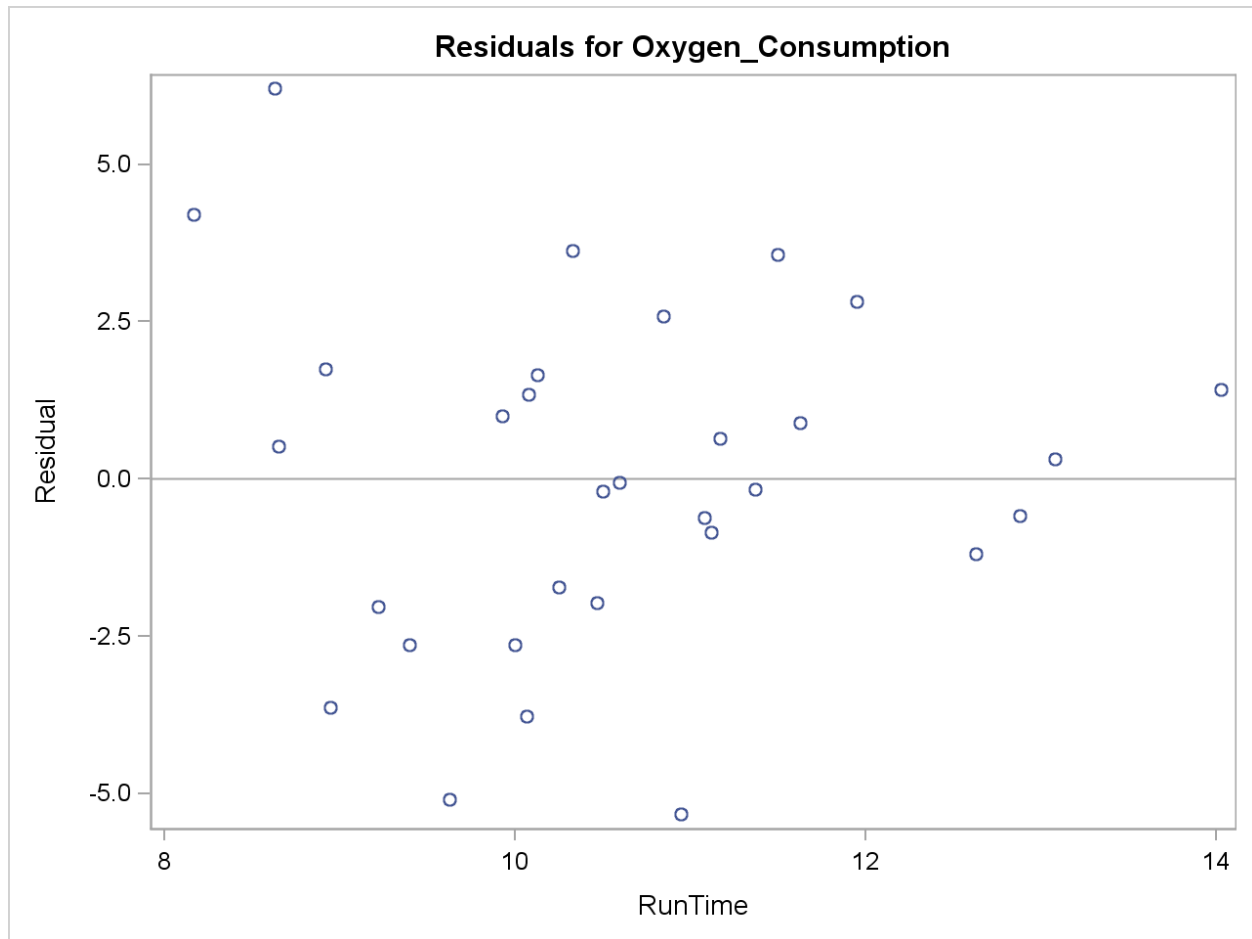
DF	represents the degrees of freedom associated with each term in the model.
Parameter Estimate	is the estimated value of the parameters associated with each term in the model.
Standard Error	is the standard error of each parameter estimate.
t Value	is the t statistic, which is calculated by dividing the parameter estimates by their corresponding standard error estimates.
Pr > t	is the p -value associated with the t statistic. It tests whether the parameter associated with each term in the model is different from 0. For this example, the slope for the predictor variable is statistically different from 0. Thus, you can conclude that the predictor variable explains a significant portion of variability in the response variable.

Because the estimate of $\beta_0=82.42494$ and $\beta_1=-3.31085$, the estimated regression equation is given by **Oxygen_Consumption** $=82.42494-3.31085*(\text{RunTime})$.

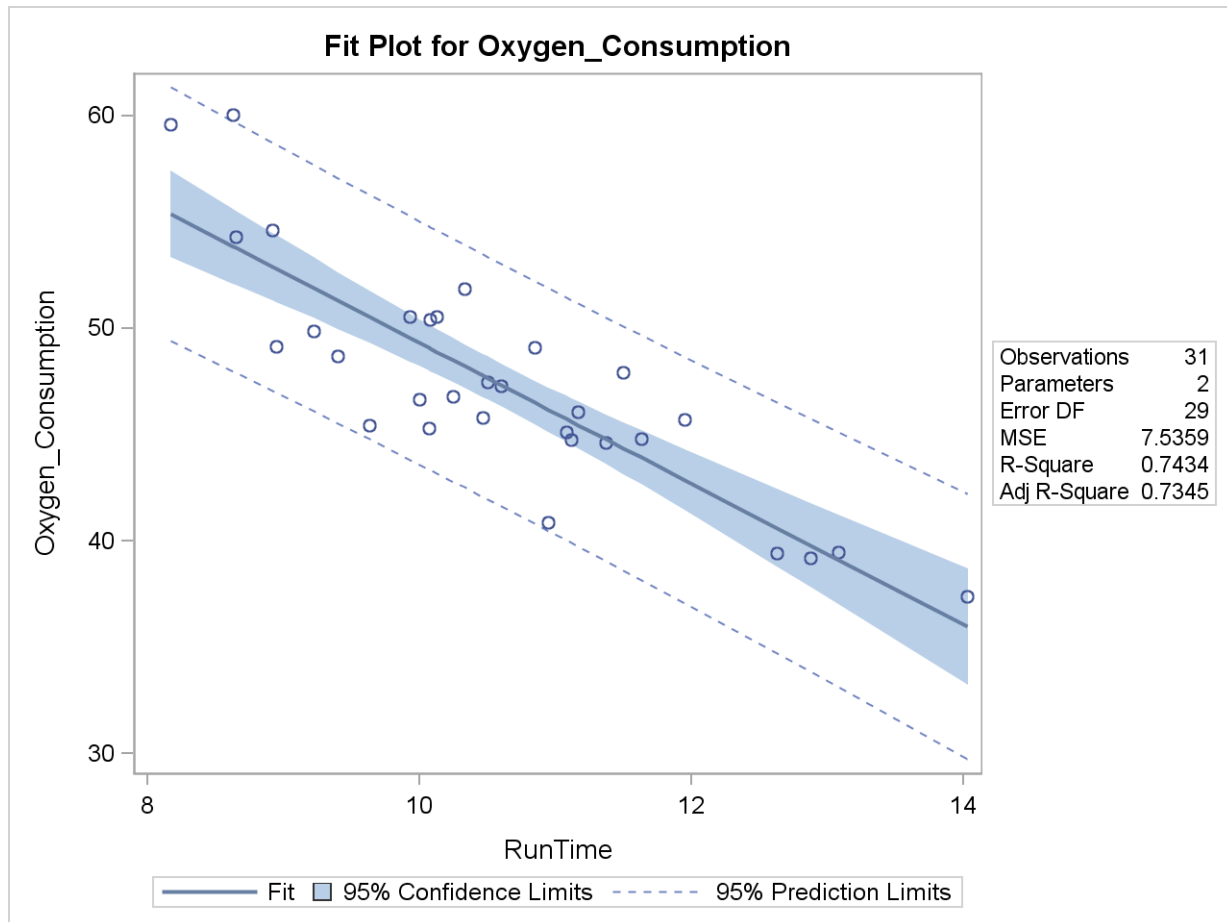
The model indicates that a one-unit greater value for **RunTime** is associated with a 3.31085 lower value for **Oxygen_Consumption**. However, *extrapolation of the model beyond the range of your predictor variables is inappropriate*. You cannot assume that the relationship maintains in areas that were not sampled from.

The parameter estimates table also shows that the intercept parameter is not equal to 0. However, the test for the intercept parameter only has practical significance when the range of values for the predictor variable includes 0. In this example, the test could not have practical significance because **RunTime** $=0$ (running at the speed of light) is not inside the range of observed values.





The diagnostics table and the residuals by **RunTime** table show a variety of plots designed to help with an assessment of the data's fulfillment of statistical assumptions and influential outliers. These plots are explored in detail in a later chapter.



The Fit plot produced by ODS Graphics shows the predicted regression line superimposed over a scatter plot of the data.

To assess the level of precision around the mean estimates of **Oxygen_Consumption**, you can produce *confidence intervals around the means*. This is represented in the shaded area in the plot.

- A 95% confidence interval for the mean says that you are 95% confident that your interval contains the population mean of Y for a particular X.
- Confidence intervals become wider as you move away from the mean of the independent variable. This reflects the fact that your estimates become more variable as you move away from the means of X and Y.

Suppose that the mean **Oxygen_Consumption** at a fixed value of **RunTime** is not the focus. If you are interested in establishing an inference on a future single observation, you need a *prediction interval around the individual observations*. This is represented by the area between the broken lines in the plot.

- A 95% prediction interval is one that you are 95% confident contains a new observation.
- Prediction intervals are wider than confidence intervals because single observations have more variability than sample means.



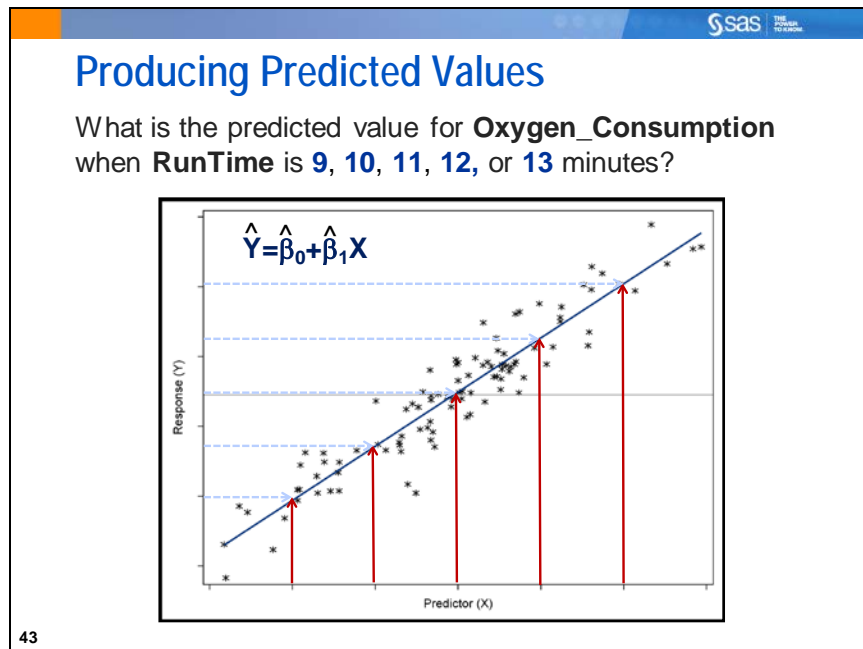
Printed tables for the confidence and prediction intervals at each observed data point can be obtained by adding the CLM and CLI options to the MODEL statement.

3.02 Multiple Choice Poll

Run PROC REG with this MODEL statement:

`model y=x1;` If the parameter estimate (slope) of x1 is 0, then the best guess (predicted value) of y when x1=13 is which of the following?


- a. 13
- b. the mean of y
- c. a random number
- d. the mean of x1
- e. 0



One objective in regression analysis is to predict values of the response variable given values of the predictor variables. You can obviously use the estimated regression equation to produce predicted values, but if you want a large number of predictions, this can be cumbersome.

To produce predicted values in PROC REG, follow these steps:

1. Create a data set with the values of the independent variable for which you want to make predictions.
2. Concatenate the data in the step above with the original data set.
3. Fit a simple linear regression model to the new data set and specify the P option in the MODEL statement. Because the observations added in the previous step contain missing values for the response variable, PROC REG does not include these observations when fitting the regression model. However, PROC REG does produce predicted values for these observations.



The SCORE Procedure

General form of the SCORE procedure:

```
PROC SCORE DATA=SAS-data-set  
              <SCORE=SAS-data-set>  
              <OUT=SAS-data-set>  
              <other options>;  
      VAR variables;  
RUN;
```

44

The SCORE procedure multiplies values from two SAS data sets, one containing coefficients (for example, factor-scoring coefficients or regression coefficients) and the other containing raw data to be scored using the coefficients from the first data set. The result of this multiplication is a SAS data set that contains linear combinations of the coefficients and the raw data values.

Many statistical procedures output coefficients that PROC SCORE can apply to raw data to produce scores. The new score variable is formed as a linear combination of raw data and scoring coefficients. For each observation in the raw data set, PROC SCORE multiplies the value of a variable in the raw data set by the matching scoring coefficient from the data set of scoring coefficients. This multiplication process is repeated for each variable in the VAR statement. The resulting products are then summed to produce the value of the new score variable. This entire process is repeated for each observation in the raw data set. In other words, PROC SCORE cross multiplies part of one data set with another.



Producing Predicted Values

Example: Produce predicted values of **Oxygen_Consumption** when **RunTime** is 9, 10, 11, 12, or 13.

Produce predicted values by outputting the parameter estimates from PROC REG into a data set and then scoring the new observations in PROC SCORE. Here is an example program to create the data set containing the observations to be scored.

```
/*st103d03.sas*/
data Need_Predictions;
    input RunTime @@;
    datalines;
9 10 11 12 13
;
run;
```

The regression model is submitted, as usual, but with an OUTEST= option for scoring (predicting the values of) new observations.

The MODEL statement below is preceded by an alphanumeric string followed by a colon (:). This string is the label of the model and is used as the name of the variable containing the predictions from a subsequent run of PROC SCORE.



The default model label is MODEL n , where n is the ordered value of the n^{th} MODEL statement in one run of PROC REG. That label is eventually used by PROC SCORE to name the variable that contains predicted values for the raw data set (the one to be scored).

```
proc reg data=sasuser.fitness noprint outest=Betas;
    PredOxy: model Oxygen_Consumption=RunTime;
run;
quit;

proc print data=Betas;
    title "OUTEST= Data Set from PROC REG";
run;
```

Selected PROC REG statement option:

OUTEST= outputs parameter estimates and model information to a SAS data set.

Obs	MODEL_	_TYPE_	_DEPVAR_	_RMSE_	Intercept	RunTime	Oxygen_Consumption
1	PredOxy	PARMS	Oxygen_Consumption	2.74515	82.4249	-3.31085	-1

Notice the variable **_TYPE_**; its value of that variable is important when you run PROC SCORE.

In the second part of this example, PROC SCORE is used to score a new data set, **Need_Predictions**. For PROC SCORE, the TYPE= specification is PARMS, and the names of the score variables are found in the variable **_MODEL_**, which gets its values from the model label.

```

proc score data=Need_Predictions score=Betas
          out=Scored type=parms;
    var RunTime;
run;

proc print data=Scored;
    title "Scored New Observations";
run;

```

Selected PROC SCORE statement options:

DATA= names the data set with the observations to be scored.

SCORE= names the data set with parameter estimates.

OUT= names the data set to which scored observations are to be written.

TYPE= tells PROC SCORE what type of data the SCORE= data set contains.

Obs	RunTime	PredOxy
1	9	52.6272
2	10	49.3164
3	11	46.0055
4	12	42.6947
5	13	39.3838

The predicted value for **Oxygen_Consumption** when **RunTime** is 9 is 52.6272.



Choose only values within or near the range of the predictor variable when you are predicting new values for the response variable. For this example, the values of the variable **RunTime** range from 8.17 to 14.03 minutes. Therefore, it is unwise to predict the value of **Oxygen_Consumption** for a **RunTime** of 18. The reason is that the relationship between the predictor variable and the response variable might be different beyond the range of your data.



Obtaining Predicted Values Using the P Option in the MODEL Statement (Self-Study)

If the data set used to produce the model is small, then that data set can be concatenated with the data set containing the data to be scored. You can then use the P option in the MODEL statement to produce predicted values.

```
/*st103d03.sas*/  /*Self Study*/
data Need_Predictions;
    input RunTime @@;
    datalines;
9 10 11 12 13
;
run;

data Predict;
    set Need_Predictions
        sasuser.fitness;
run;

ods graphics off;

proc reg data=Predict;
    model Oxygen_Consumption=RunTime / p;
    id RunTime;
    title 'Oxygen_Consumption=RunTime with Predicted Values';
run;
quit;
```

Selected REG procedure statement:

ID specifies a variable to label observations in the output produced by certain MODEL statement options.

Selected MODEL statement option:

P prints the values of the response variable, the predicted values, and the residual values.

PROC REG Output

Number of Observations Read	36
Number of Observations Used	31
Number of Observations with Missing Values	5

Notice that 36 observations were read; 31 were used and 5 had missing values. The observations in **Need_Predictions** had missing values for **Oxygen_Consumption**, so they were eliminated from the analysis.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	633.01458	633.01458	84.00	<.0001
Error	29	218.53997	7.53586		
Corrected Total	30	851.55455			

Root MSE	2.74515	R-Square	0.7434
Dependent Mean	47.37581	Adj R-Sq	0.7345
Coeff Var	5.79442		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	82.42494	3.85582	21.38	<.0001
RunTime	1	-3.31085	0.36124	-9.17	<.0001

The model output is not affected by the extra five observations, because they were not used in any calculations, due to missing values.

Partial Output

Output Statistics				
Obs	RunTime	Dependent Variable	Predicted Value	Residual
1	9.00	.	52.6272	.
2	10.00	.	49.3164	.
3	11.00	.	46.0055	.
4	12.00	.	42.6947	.
5	13.00	.	39.3838	.
6	8.17	59.5700	55.3753	4.1947
7	8.63	60.0600	53.8523	6.2077
8	8.65	54.3000	53.7860	0.5140
9	8.92	54.6300	52.8921	1.7379
10	8.95	49.1600	52.7928	-3.6328

Because you specified **RunTime** in the ID statement, the values of this variable appear in the first column after **Obs**.

The output shows that the estimated value of **Oxygen_Consumption** is 52.6272 when **RunTime** equals 9. This is identical to the value produced in PROC SCORE.




Exercises

2. Fitting a Simple Linear Regression Model

Use the **sasuser.BodyFat2** data set for this exercise.

- a. Perform a simple linear regression model with **PctBodyFat2** as the response variable and **Weight** as the predictor.
 - 1) What is the value of the F statistic and the associated p -value? How would you interpret this with regard to the null hypothesis?
 - 2) Write the predicted regression equation.
 - 3) What is the value of the R-square statistic? How would you interpret this?
- b. Produce predicted values for **PctBodyFat2** when **Weight** is 125, 150, 175, 200, and 225.

What are the predicted values?

The SAS logo, consisting of the letters 'sas' in a stylized font, with the tagline 'The Way to Success' in smaller text to the right.

3.03 Multiple Choice Poll

What is the predicted value for **PctBodyFat2** when **Weight** is 150?

- a. 0.17439
- b. 150
- c. 14.1067

50

3.3 Concepts of Multiple Regression

Objectives

- Explain the mathematical model for multiple regression.
- Describe the main advantage of multiple regression versus simple linear regression.
- Explain the standard output from the REG procedure.
- Describe common pitfalls of multiple linear regression.

53

Multiple Linear Regression with Two Variables

Consider the two-variable model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

where

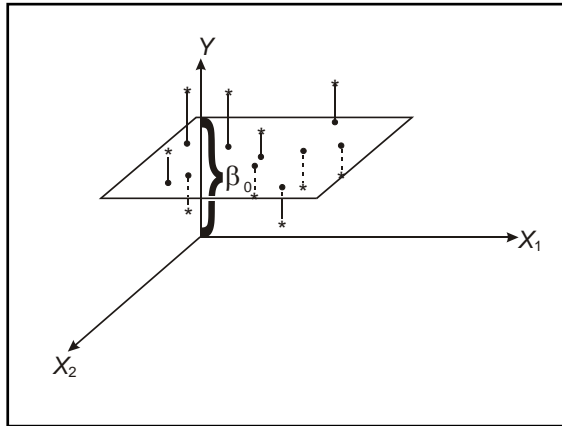
Y	is the dependent variable.
X_1 and X_2	are the independent or predictor variables.
ε	is the error term.
β_0 , β_1 , and β_2	are unknown parameters.

54

In simple linear regression, you can model the relationship between the two variables (two dimensions) with a line (one dimension).

For the two-variable model, you can model the relationship of three variables (three dimensions) with a plane (two dimensions).

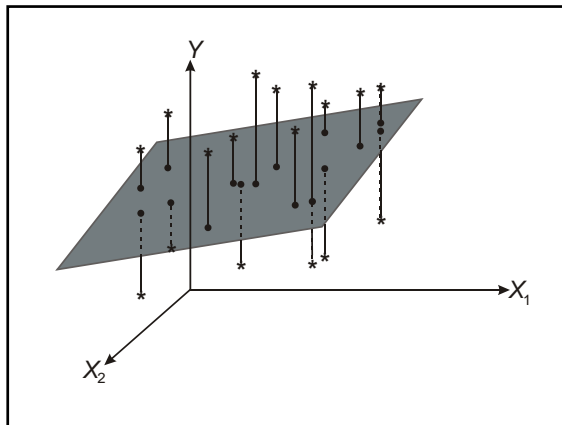
Picturing the Model: No Relationship



55

If there is no relationship among Y and X_1 and X_2 , the model is a horizontal plane passing through the point $(Y=\beta_0, X_1=0, X_2=0)$.


Picturing the Model: A Relationship



56

If there is a relationship among Y and X_1 and X_2 , the model is a sloping plane passing through three points:

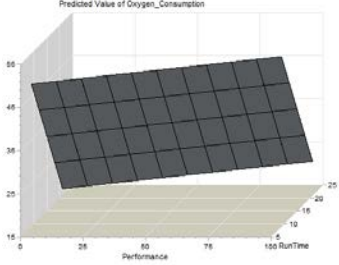
- $(Y=\beta_0, X_1=0, X_2=0)$
- $(Y=\beta_0+\beta_1, X_1=1, X_2=0)$
- $(Y=\beta_0+\beta_2, X_1=0, X_2=1)$



The Multiple Linear Regression Model

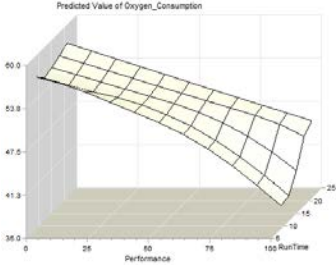
In general, you model the dependent variable, Y , as a linear function of k independent variables, X_1 through X_k :

$$Y = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k + \varepsilon$$



$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$

Linear?



$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2 + \varepsilon$

Nonlinear?

57

You investigate the relationship among $k+1$ variables (k predictors+1 response) using a k -dimensional surface for prediction.

The multiple general linear model is not restricted to modeling only planar relationships. By using higher order terms, such as quadratic or cubic powers of the X s or cross products of one X with another, surfaces more complex than planes can be modeled.


In the examples, the models are limited to relatively simple surfaces.



The model has $p=k+1$ parameters (the β s), including the intercept, β_0 .

sas THE POWER OF DATA

Multiple Regression Example

PREDICTORS		RESPONSE
Performance		
RunTime		
Age		Oxygen_Consumption
Weight		
Run_Pulse		
Rest_Pulse		
Maximum_Pulse		

67

sas THE POWER OF DATA

Model Hypothesis Test

Null Hypothesis:

- The regression model does **not** fit the data better than the baseline model.
- $\beta_1 = \beta_2 = \dots = \beta_k = 0$

Alternative Hypothesis:

- The regression model does fit the data better than the baseline model.
- Not all β_i s equal zero.

58

If the estimated linear regression model does **not** fit the data better than the baseline model, you fail to reject the null hypothesis. Thus, you do **not** have enough evidence to say that all of the slopes of the regression in the population differ from zero. The predictor variables do not explain a significant amount of variability in the response variable.

If the estimated linear regression model **does** fit the data better than the baseline model, you reject the null hypothesis. Thus, you **do** have enough evidence to say that at least one slope of the regression in the population differs from zero. At least one predictor variable explains a significant amount of variability in the response variable.

3.04 Multiple Choice Poll

Which statistic in the ANOVA table is used to test the overall model hypotheses?

- a. F
- b. t
- c. R square
- d. Adjusted R square


60

Assumptions for Linear Regression

- The mean of the Ys is accurately modeled by a linear function of the Xs.
- The random error term, ε , is assumed to have a normal distribution with a mean of zero.
- The random error term, ε , is assumed to have a constant variance, σ^2 .
- The errors are independent.

62

Techniques to evaluate the validity of these assumptions are discussed in a later chapter.



Multiple Linear Regression versus Simple Linear Regression

Main Advantage
Multiple linear regression enables you to investigate the relationship among Y and several independent variables simultaneously.


Main Disadvantages
Increased complexity makes it more difficult to do the following:

- ascertain which model is “best”
- interpret the models

63

The advantage of performing multiple linear regression over a series of simple linear regression models far outweighs the disadvantages. In practice, many responses depend on multiple factors that might interact in some way.

SAS tools help you decide upon a “best” model, a choice that might depend on the purposes of the analysis, as well as subject-matter expertise.



Common Applications

Multiple linear regression is a powerful tool for the following tasks:

- Prediction – to develop a model to predict future values of a response variable (Y) based on its relationships with other predictor variables (Xs)
- Analytical or Explanatory Analysis – to develop an understanding of the relationships between the response variable and predictor variables

64

Even though multiple linear regression enables you to analyze many experimental designs, ranging from simple to complex, you focus on applications for analytical studies and predictive modeling. Other SAS procedures, such as GLM, are better suited for analyzing experimental data.

The distinction between using multiple regression for an analytic analysis and prediction modeling is somewhat artificial. A model developed for prediction is probably a good analytic model. Conversely, a model developed for an analytic study is probably a good prediction model.

Myers (1999) refers to four applications of regression:

- prediction
- variable screening
- model specifications
- parameter estimation

The term *analytical analysis* is similar to Myers' parameter estimation application and variable screening.

Prediction

- The terms in the model, the values of their coefficients, and their statistical significance are of secondary importance.
- The focus is on producing a model that is the best at predicting future values of Y as a function of the X s. The predicted value of Y is given by this formula:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_k X_k$$

65

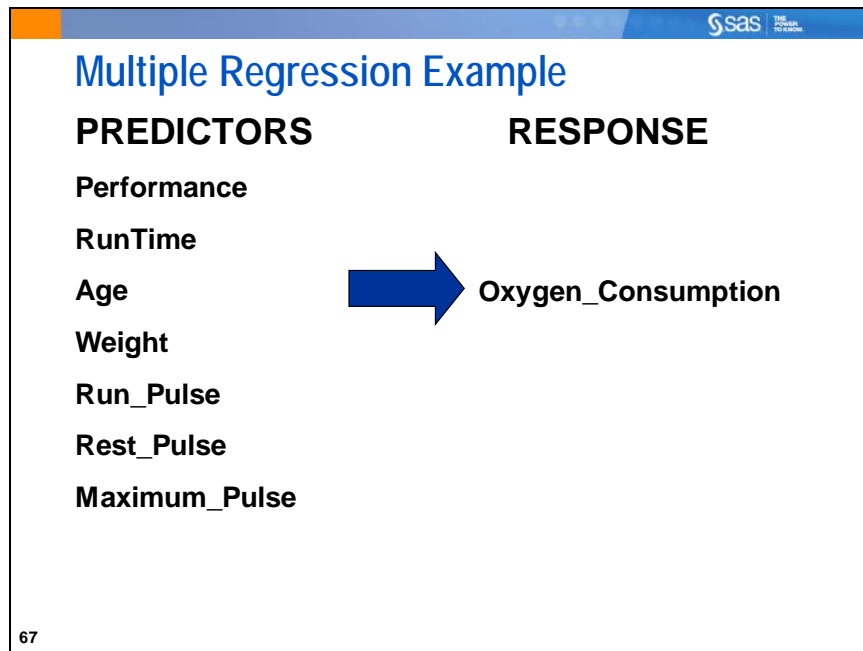
Most investigators whose main goal is prediction do not ignore the terms in the model (the X s), the values of their coefficients (the β s), or their statistical significance (the p -values). They use these statistics to help choose among models with different numbers of terms and predictive capabilities.

Analytical or Explanatory Analysis

- The focus is on understanding the relationship between the dependent variable and the independent variables.
- Consequently, the statistical significance of the coefficients is important as well as the magnitudes and signs of the coefficients.

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_k X_k$$

66




An analyst knows from doing a simple linear regression that the measure of performance is an important variable in explaining the oxygen consumption capability of a club member.

The analyst is interested in investigating other information to ascertain whether other variables are important in explaining the oxygen consumption capability.

Recall that you did a simple linear regression on **Oxygen_Consumption** with **RunTime** as the predictor variable.

The R square for this model was 0.7434, which suggests that 25.64% of the variation in **Oxygen_Consumption** is still unexplained.

Consequently, adding other variables to the model, such as **Performance** or **Age**, might provide a significantly better model.



Adjusted R Square

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

68

The R square always increases or stays the same as you include more terms in the model. Therefore, choosing the “best” model is not as simple as just making the R square as large as possible.

The adjusted R square is a measure similar to R square, but it takes into account the number of terms in the model. It can be thought of as a penalized version of R square with the penalty increasing with each parameter added to the model.



Fitting a Multiple Linear Regression Model

Example: Invoke PROC REG and perform a multiple linear regression analysis of **Oxygen_Consumption** on **Performance** and **RunTime**. Interpret the output for the two-variable model.

```
/*st103d04.sas*/
ods graphics off;
proc reg data=sasuser.fitness;
    model Oxygen_Consumption=Performance RunTime;
    title 'Multiple Linear Regression for Fitness Data';
run;
quit;
ods graphics on;
```

The only required statement for PROC REG is the MODEL statement.

General form of the MODEL statement:

```
MODEL Y=X1 X2 ... Xk;
```

where

Y is the dependent variable.

X1 X2 ... Xk

is a list of the independent variables that are included in the model.

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	646.33101	323.16550	44.09	<.0001
Error	28	205.22355	7.32941		
Corrected Total	30	851.55455			

PROC REG Output

Model DF	is 2, the number of parameters minus 1.
Error DF	is 28, the total numbers of observations (31) minus the number of parameters in the model (3).
Corrected Total DF	is 30, the number of observations minus 1.
Model Sum of Squares	is the total variation in the Y explained by the model.
Error Sum of Squares	is the variation in the Y <i>not</i> explained by the model.
Corrected Total Sum of Squares	is the total variation in the Y.
Model Mean Square	is the Model Sum of Squares divided by the Model DF – also known as model variance.
Mean Square Error	is the Error Sum of Squares divided by the Error DF and is an estimate of σ^2 , the variance of the random error term – also known as error variance.
F Value	is the (Mean Square Model)/(Mean Square Error).

$\text{Pr}>F$ is small. Therefore, you reject $H_0: \beta_1=\beta_2=0$ and conclude that at least one $\beta_i \neq 0$.

Root MSE	2.70729	R-Square	0.7590
Dependent Mean	47.37581	Adj R-Sq	0.7418
Coeff Var	5.71450		

The R square for this model, 0.7590, is only slightly larger than the R square for the model in which **RunTime** is the only predictor variable, 0.7434.

The adjusted R square for this model is 0.7418, slightly higher than the adjusted R square of 0.7345 for the **RunTime** only model. This suggests, although mildly, that adding **Performance** does improve the model predicting **Oxygen_Consumption**.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	71.52626	8.93520	8.00	<.0001
Performance	1	0.06360	0.04718	1.35	0.1885
RunTime	1	-2.62163	0.62320	-4.21	0.0002

Using the estimates for β_0 , β_1 , and β_2 above, this model can be written as the following:

$$\text{Oxygen_Consumption} = 71.52626 + 0.06360 * \text{Performance} - 2.62163 * \text{RunTime}$$

The p -value for **Performance** is large, which suggests that the slope is not significantly different from 0. The correlation that you saw between **Performance** and **Oxygen_Consumption** was large and statistically significant ($r = .77890$, $p < .0001$). The test for $\beta_i = 0$ is conditioned on the other terms in the model. That is the reason that neither **Performance** nor **RunTime** have the same p -values (or parameter estimates) when used alone as when used in a model that includes both. The test for $\beta_1 = 0$ (for **Performance**) is conditional on (or adjusted for) X_2 (**RunTime**). Similarly, the test for $\beta_2 = 0$ is conditional on X_1 (**Performance**).

The significance level of the test does *not* depend on the order in which you list the independent variables in the MODEL statement, but it does depend on the variables included in the MODEL statement.

In a later section, you look at the difficulties involved with analyzing and selecting the best models due to the relationships among predictor variables.



Exercises

3. Performing Multiple Regression Using the REG Procedure

- a. Using the **sasuser.BodyFat2** data set, run a regression of **PctBodyFat2** on the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.
 - 1) Compare the ANOVA table with that from the model with only **Weight** in the previous exercise. What is different?
 - 2) How do the R square and the adjusted R square compare with these statistics for the **Weight** regression demonstration?
 - 3) Did the estimate for the intercept change? Did the estimate for the coefficient of **Weight** change?

4. Simplifying the Model

- a. Rerun the model in **3a**, but eliminate the variable with the highest p -value. Compare the output with the Exercise **3a** model.
- b. Did the p -value for the model change notably?
- c. Did the R square and adjusted R square change notably?
- d. Did the parameter estimates and their p -values change notably?

5. More Simplifying of the Model

- a. Rerun the model in Exercise **4a**, but drop the variable with the highest p -value.
- b. How did the output change from the previous model?
- c. Did the number of parameters with a p -value less than 0.05 change?

3.05 Multiple Choice Poll

When **Oxygen_Consumption** is regressed on **RunTime**, **Age**, **Run_Pulse**, and **Maximum_Pulse**, the parameter estimate for **Age** is -2.78. What does this mean?

- a. For each year older, the predicted value of oxygen consumption is 2.78 greater.
 - b. For each year older, the predicted value of oxygen consumption is 2.78 lower.
 - c. For every 2.78 years older, oxygen consumption doubles.
 - d. For every 2.78 years younger, oxygen consumption doubles.
- * Assume that the values of all other predictors are held constant.

3.4 Model Building and Interpretation

Objectives

- Explain the REG procedure options for model selection.
- Describe model selection options and interpret output to evaluate the fit of several models.

Model Selection

Eliminating one variable at a time manually for small data sets is a reasonable approach.

However, eliminating one variable at a time manually for large data sets can take an extreme amount of time.

77

A process for selecting models might be to start with all the variables in the **sasuser.fitness** data set and eliminate the least significant terms, based on p -values.

For a small data set, a final model can be developed in a reasonable amount of time. If you start with a large model, however, eliminating one variable at a time can take an extreme amount of time. You would have to continue this process until only terms with p -values lower than some threshold value, such as 0.05 or 0.10, remain.

Model Selection Options

The **SELECTION=** option in the **MODEL** statement of **PROC REG** supports these model selection techniques:

Stepwise selection methods

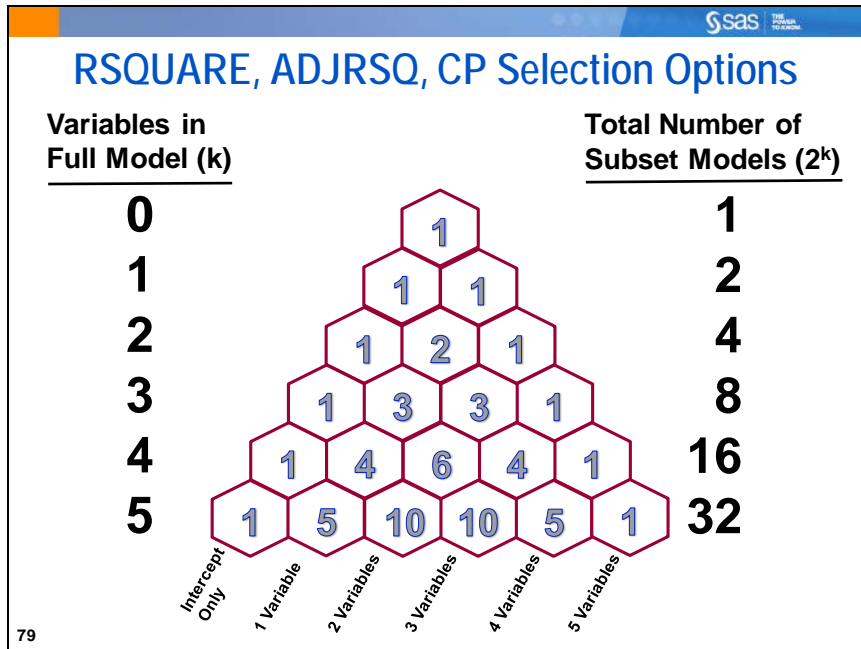
- **STEPWISE**, **FORWARD**, or **BACKWARD**

All-possible regressions ranked using

- **RSQUARE**, **ADJRSQ**, or **CP**

SELECTION=NONE is the default.

78



In the **sasuser.fitness** data set, there are seven possible independent variables. Therefore, there are $2^7=128$ possible regression models. There are seven possible one-variable models, 21 possible two-variable models, 35 possible three-variable models, and so on.

You can choose to only look at the best n (as measured by the model R^2 for $k=1, 2, 3, \dots, 7$) by using the **BEST=** option on the model statement. The **BEST=** option only reduces the output. All regressions are still calculated.

If there were 20 possible independent variables, there would be more than 1,000,000 models. In a later demonstration, you see another technique that does not have to examine all the models to help you choose a set of candidate models.

Mallows' C_p

- Mallows' C_p is a simple indicator of effective variable selection within a model.
- Look for models with $C_p \leq p$, where p equals the number of parameters in the model, including the intercept.

Mallows recommends choosing the first (fewest variables) model where C_p approaches p .

80

Mallows' C_p (1973) is estimated by
$$C_p = p + \frac{(\text{MSE}_p - \text{MSE}_{\text{full}})(n - p)}{\text{MSE}_{\text{full}}}$$

where


MSE_p is the mean squared error for the model with p parameters.

MSE_{full} is the mean squared error for the full model used to estimate the true residual variance.

n is the number of observations.

p is the number of parameters, including an intercept parameter, if estimated.

The choice of the best model based on C_p is debatable, as will be shown in the slide about Hocking's criterion. Many choose the model with the smallest C_p value. However, Mallows recommended that the best model will have a C_p value approximating p . The most parsimonious model that fits that criterion is generally considered to be a good choice, although subject-matter knowledge should also be a guide in the selection from among competing models.



Hocking's Criterion versus Mallows' C_p

Hocking (1976) suggests selecting a model based on the following:

- $C_p \leq p$ for prediction
- $C_p \leq 2p - p_{\text{full}} + 1$ for parameter estimation

81

Hocking suggested the use of the C_p statistic, but with alternative criteria, depending on the purpose of the analysis. His suggestion of $(C_p \leq 2p - p_{\text{full}} + 1)$ is included in the REG procedure's calculations of criteria reference plots for best models.



Automated Model Selection

Example: Invoke PROC REG to produce a regression of **Oxygen_Consumption** on all the other variables in the **fitness** data set.

```
/*st103d05.sas*/ /*Part A*/
ods graphics / imagemap=on;
proc reg data=sasuser.fitness plots(only)=(rsquare adjrsq cp);
  ALL_REG: model oxygen_consumption=
                Performance RunTime Age Weight
                Run_Pulse Rest_Pulse Maximum_Pulse
  / selection=rsquare adjrsq cp;
  title 'Best Models Using All-Regression Option';
run;
quit;
```

Selected MODEL statement options:

SELECTION= enables you to choose the different selection methods – RSQUARE, ADJRSQ, and CP. The first listed method is the one that determines the sorting order in the output.

Selected SELECTION= option methods:

RSQUARE tells PROC REG to use the model R square to rank the model from best to worst for a given number of variables.

ADJRSQ prints the adjusted R square for each model.

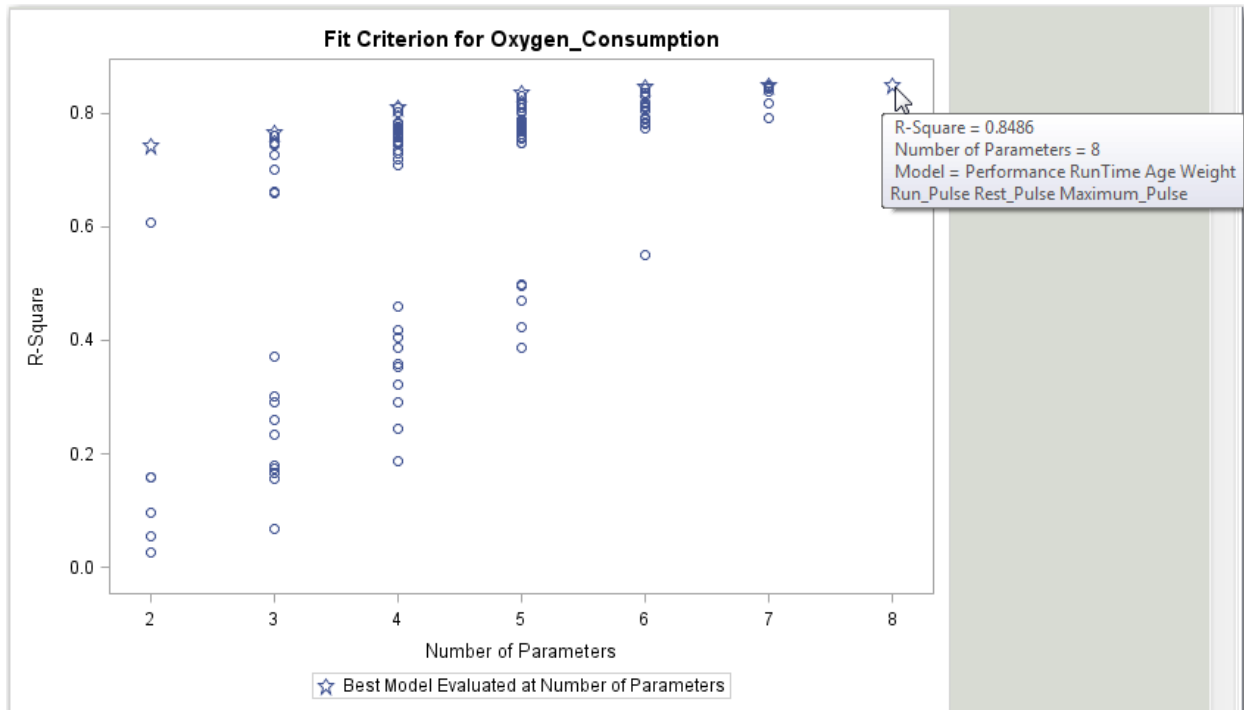
CP prints Mallows' C_p statistic for each model.

Partial HTML Output

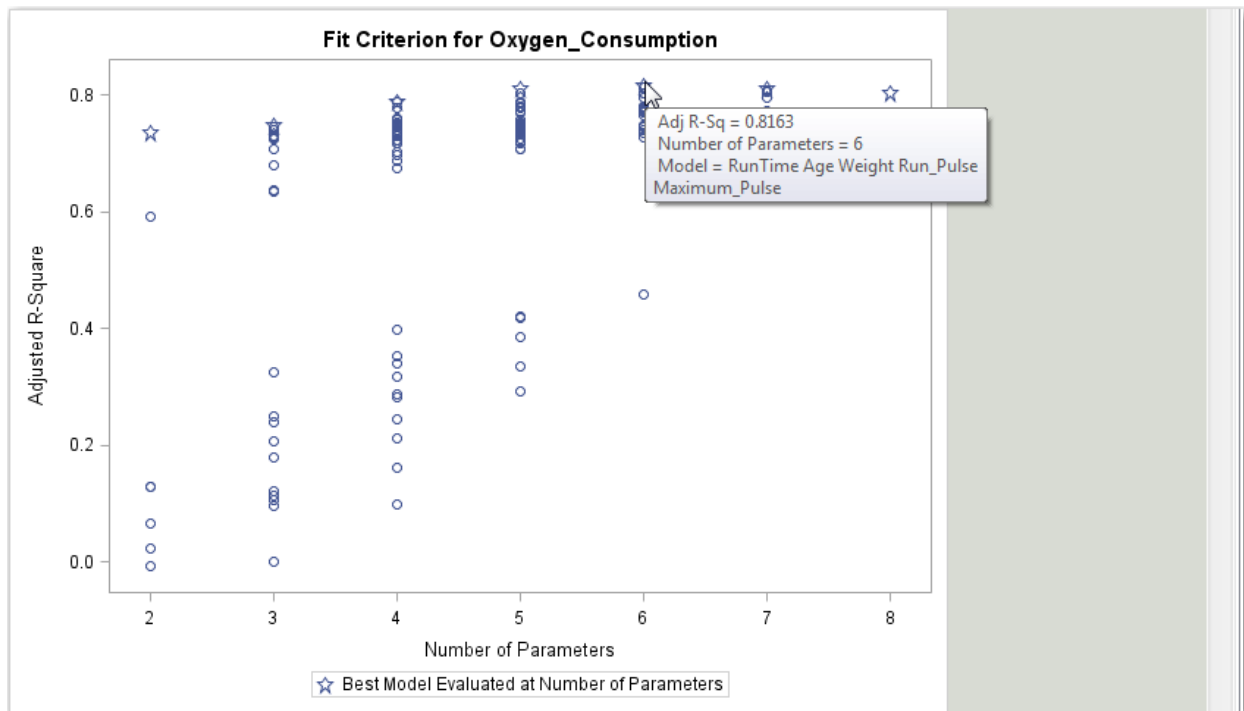
Number of Observations Read	31
Number of Observations Used	31

Model Index	Number in Model	R-Square	Adjusted R-Square	C(p)	Variables in Model
1	1	0.7434	0.7345	11.9967	RunTime
2	1	0.6067	0.5931	32.7650	Performance
3	1	0.1595	0.1305	100.7200	Rest_Pulse
4	1	0.1585	0.1294	100.8736	Run_Pulse
5	1	0.0971	0.0660	110.1977	Age
6	1	0.0561	0.0235	116.4349	Maximum_Pulse
7	1	0.0265	-0.0070	120.9214	Weight
8	2	0.7647	0.7479	10.7530	RunTime Age
9	2	0.7614	0.7444	11.2503	RunTime Run_Pulse
10	2	0.7590	0.7418	11.6205	Performance RunTime
11	2	0.7475	0.7295	13.3606	Performance Run_Pulse
12	2	0.7452	0.7270	13.7166	RunTime Maximum_Pulse
13	2	0.7449	0.7267	13.7588	RunTime Weight
14	2	0.7435	0.7252	13.9735	RunTime Rest_Pulse

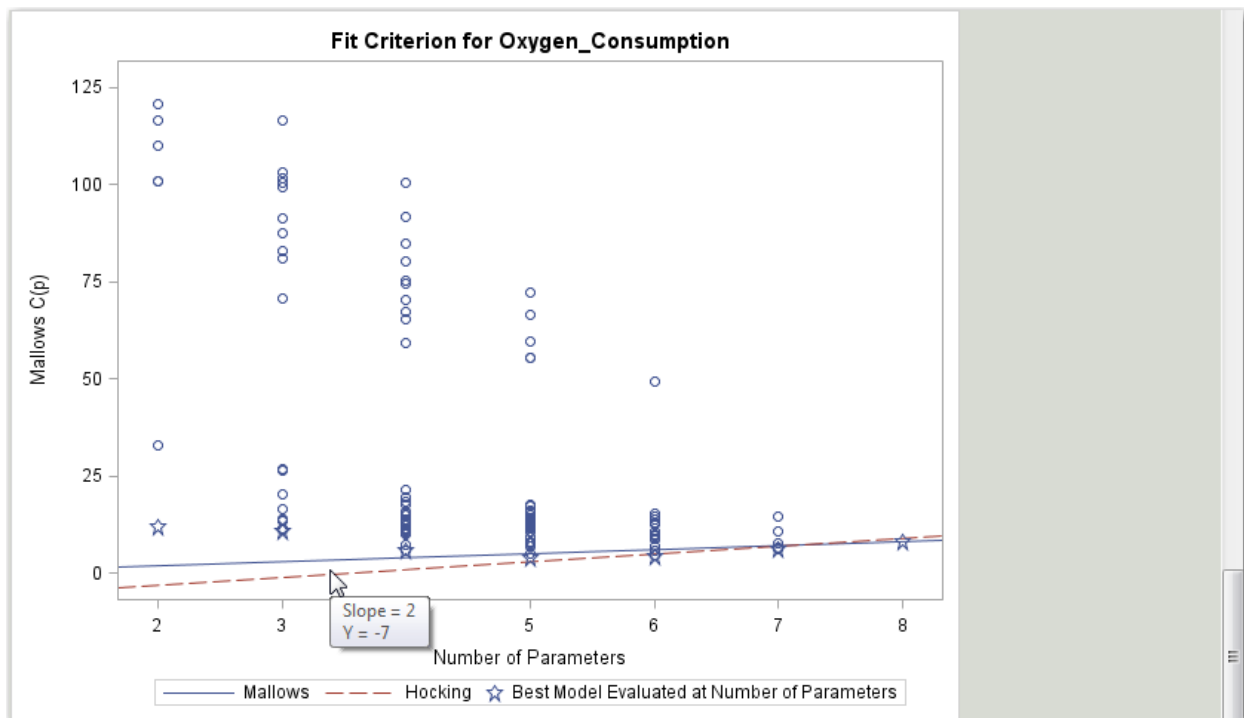
There are many models to compare. It would be unwieldy to try to determine the best model by viewing the output tables. Therefore, it is advisable to look at the ODS plots.



The R-square plot compares all models based on their R-square values. As noted earlier, adding variables to a model always increases R-square, and therefore the full model is always best. Therefore, you can only use the R-square value to compare models of equal numbers of parameters.



The adjusted R square does not have the problem that the R square has. You can compare models of different sizes. In this case, it is difficult to see which model has the higher adjusted R square, the starred model for six parameters or seven parameters.



The line $C_p=p$ is plotted to help you identify models that satisfy the criterion $C_p \leq p$ for prediction. The lower line is plotted to help identify which models satisfy Hocking's criterion $C_p \leq 2p - p_{\text{full}} + 1$ for parameter estimation.

Use the graph and review the output to select a relatively short list of models that satisfy the criterion appropriate for your objective. The first model to fall below the line for Mallows' criterion has five parameters. The first model to fall below Hocking's criterion has six parameters.

It is often the case that the best model is difficult to see because of the range of C_p values at the high end. These models are clearly not the best and therefore you can focus on the models near the bottom of the range of C_p .

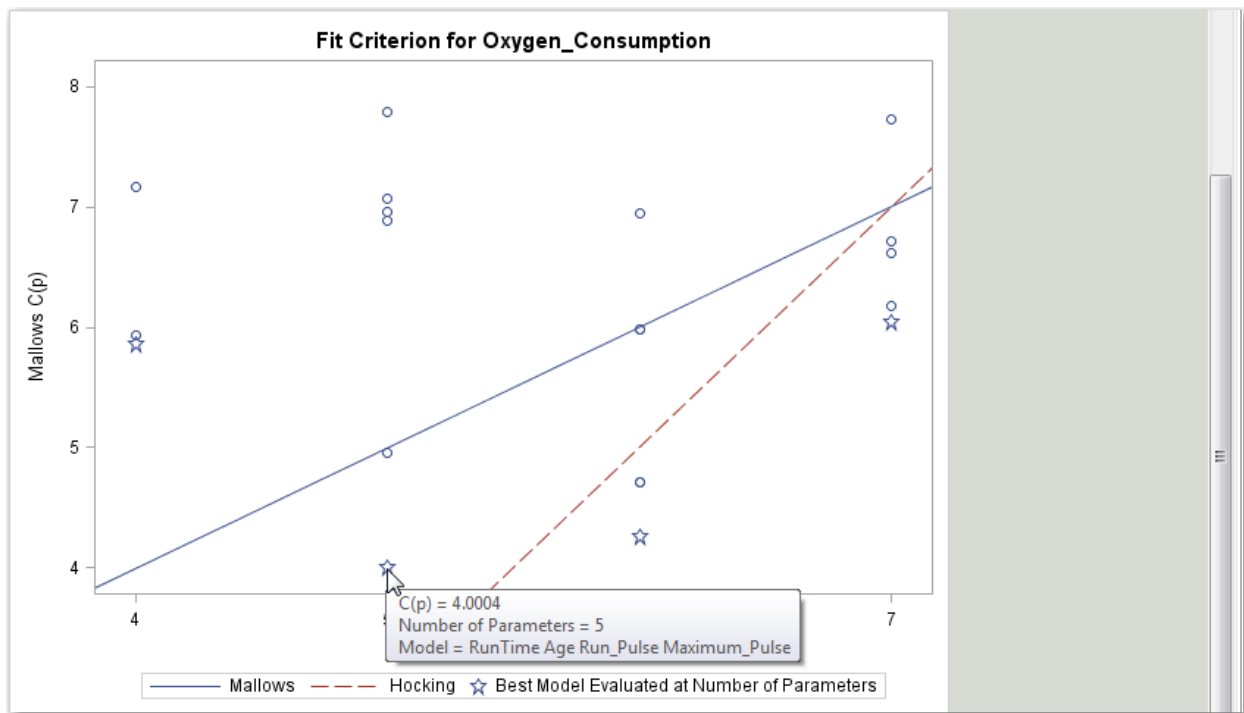
```
/*st103d05.sas*/ /*Part B*/
ods graphics / imagemap=on;
proc reg data=sasuser.fitness plots(only)=(cp);
    ALL_REG: model oxygen_consumption=
        Performance RunTime Age Weight
        Run_Pulse Rest_Pulse Maximum_Pulse
    / selection=cp rsquare adjrsq best=20;
    title 'Best Models Using All-Regression Option';
run;
quit;
```

Selected SELECTION= option methods:

BEST= n limits the output to only the best n models.

Model Index	Number in Model	C(p)	R-Square	Adjusted R-Square	Variables in Model
1	4	4.0004	0.8355	0.8102	RunTime Age Run_Pulse Maximum_Pulse
2	5	4.2598	0.8469	0.8163	RunTime Age Weight Run_Pulse Maximum_Pulse
3	5	4.7158	0.8439	0.8127	Performance RunTime Weight Run_Pulse Maximum_Pulse
4	5	4.7168	0.8439	0.8127	Performance RunTime Age Run_Pulse Maximum_Pulse
5	4	4.9567	0.8292	0.8029	Performance RunTime Run_Pulse Maximum_Pulse
6	3	5.8570	0.8101	0.7890	RunTime Run_Pulse Maximum_Pulse
7	3	5.9367	0.8096	0.7884	RunTime Age Run_Pulse
8	5	5.9783	0.8356	0.8027	RunTime Age Run_Pulse Rest_Pulse Maximum_Pulse
9	5	5.9856	0.8356	0.8027	Performance Age Weight Run_Pulse Maximum_Pulse
10	6	6.0492	0.8483	0.8104	Performance RunTime Age Weight Run_Pulse Maximum_Pulse
11	6	6.1758	0.8475	0.8094	RunTime Age Weight Run_Pulse Rest_Pulse Maximum_Pulse
12	6	6.6171	0.8446	0.8057	Performance RunTime Weight Run_Pulse Rest_Pulse Maximum_Pulse
13	6	6.7111	0.8440	0.8049	Performance RunTime Age Run_Pulse Rest_Pulse Maximum_Pulse
14	4	6.8865	0.8165	0.7882	Performance RunTime Age Run_Pulse
15	5	6.9446	0.8293	0.7951	Performance RunTime Run_Pulse Rest_Pulse Maximum_Pulse
16	4	6.9623	0.8160	0.7877	RunTime Weight Run_Pulse Maximum_Pulse
17	4	7.0752	0.8152	0.7868	RunTime Age Weight Run_Pulse
18	3	7.1734	0.8014	0.7794	Performance RunTime Run_Pulse
19	6	7.7279	0.8373	0.7966	Performance Age Weight Run_Pulse Rest_Pulse Maximum_Pulse
20	4	7.7942	0.8105	0.7814	RunTime Run_Pulse Rest_Pulse Maximum_Pulse

Investigate the plot of Mallows' $C(p)$.



In this example the number of variables in the full model, p_{full} , equals 8 (seven variables plus the intercept).

The smallest model with an observation below the Mallows line has $p=5$ (which matches to Number in Model of 4 in the previous table). The model with the star at five parameters and the model above it are considered “best,” based on Mallows’ original criterion. The starred model has a $C_p=4.004$, satisfying Mallows’ criterion (**Oxygen_Consumption=RunTime Age Run_Pulse Maximum_Pulse**) and the one above has a value of 4.9567 (**Oxygen_Consumption=Performance RunTime Run_Pulse Maximum_Pulse**). The only difference between the two models is that the first includes **Age** and the second includes **Performance**. By the strictest definition, the second model should be selected, because its C_p value is closest to p .

The smallest model that falls under the Hocking line has $p=6$. The model with the smaller C_p value will be considered the “best” explanatory model. The table shows that the first model with $p=6$ is **Oxygen_Consumption=RunTime Age Weight Run_Pulse Maximum_Pulse**, with a C_p value of 4.2598. Two other models that are also below the Hocking line are **Oxygen_Consumption=Performance RunTime Weight Run_Pulse Maximum_Pulse** and **Oxygen_Consumption=Performance RunTime Age Run_Pulse Maximum_Pulse**. (They are nearly on top of one another in the plot.)

"Best" Models – Prediction


The two best candidate models based on Mallows' original criterion includes these regressor variables:

$p=5$	$C_p=4.0004$ $R^2=0.8355$ Adj. $R^2=0.8102$	RunTime, Age, Run_Pulse, Maximum_Pulse
$p=5$	$C_p=4.9567$ $R^2=0.8292$ Adj. $R^2=0.8029$	Performance, RunTime, Run_Pulse, Maximum_Pulse

83

Some models might be essentially equivalent based on their C_p , R square, or other measures. When, as in this case, there are several candidate "best" models, it is the responsibility of the investigator to determine which model makes the most sense based on theory and experience. The choice between these two models is essentially the choice between **Age** and **Performance**. Because age is much easier to measure than the subjective measure of performance, the first model is selected here.

A limitation of the evaluation that you did thus far is that you do not know the magnitude and signs of the coefficients of the candidate models or their statistical significance.



“Best” Models – Parameter Estimation

The three best candidate models for analytic purposes, according to Hocking, include those listed below:

$p=6$	$C_p=4.2598$ $R^2=0.8469$ Adj. $R^2=0.8163$	RunTime, Age, Weight, Run_Pulse, Maximum_Pulse
$p=6$	$C_p=4.7158$ $R^2=0.8439$ Adj. $R^2=0.8127$	Performance, RunTime, Weight, Run_Pulse, Maximum_Pulse
$p=6$	$C_p=4.7168$ $R^2=0.8439$ Adj. $R^2=0.8127$	Performance, RunTime, Age, Run_Pulse, Maximum_Pulse

84

The variables **RunTime**, **Run_Pulse**, and **Maximum_Pulse** once again appear in all candidate models. The choice of models depends on the selection of pairs from **Performance**, **Age**, and **Weight**. You again choose a model with objective measures, **Age** and **Weight**. That is the top model in the list. Your choice might differ.



Estimating and Testing the Coefficients for the Selected Models

Example: Invoke PROC REG to compare the ANOVA tables and parameter estimates for the two-candidate models in the **fitness** data set.

```
/*st103d06.sas*/
ods graphics off;
proc reg data=sasuser.fitness;
  PREDICT: model Oxygen_Consumption=
    RunTime Age Run_Pulse Maximum_Pulse;
  EXPLAIN: model Oxygen_Consumption=
    RunTime Age Weight Run_Pulse Maximum_Pulse;
  title 'Check "Best" Two Candidate Models';
run;
quit;
ods graphics on;
```

PROC REG can have more than one MODEL statement. You can assign a label to each MODEL statement to identify the output generated for each model.

Output for the PREDICT Model

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	711.45087	177.86272	33.01	<.0001
Error	26	140.10368	5.38860		
Corrected Total	30	851.55455			

Root MSE	2.32134	R-Square	0.8355
Dependent Mean	47.37581	Adj R-Sq	0.8102
Coeff Var	4.89984		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	97.16952	11.65703	8.34	<.0001
RunTime	1	-2.77576	0.34159	-8.13	<.0001
Age	1	-0.18903	0.09439	-2.00	0.0557
Run_Pulse	1	-0.34568	0.11820	-2.92	0.0071
Maximum_Pulse	1	0.27188	0.13438	2.02	0.0534

The R square and adjusted R square are the same as calculated during the model selection program. If there are missing values in the data set, however, this might not be true.

The model F is large and highly significant. **Age** and **Maximum_Pulse** are not significant at the 0.05 level of significance. However, all terms are significant at $\alpha=0.10$.

The adjusted R square is close to the R square, which suggests that there are not too many variables in the model.

Output for the EXPLAIN Model

Number of Observations Read	31
Number of Observations Used	31

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	721.20532	144.24106	27.66	<.0001
Error	25	130.34923	5.21397		
Corrected Total	30	851.55455			

Root MSE	2.28341	R-Square	0.8469
Dependent Mean	47.37581	Adj R-Sq	0.8163
Coeff Var	4.81978		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	101.33835	11.86474	8.54	<.0001
RunTime	1	-2.68846	0.34202	-7.86	<.0001
Age	1	-0.21217	0.09437	-2.25	0.0336
Weight	1	-0.07332	0.05360	-1.37	0.1836
Run_Pulse	1	-0.37071	0.11770	-3.15	0.0042
Maximum_Pulse	1	0.30603	0.13452	2.28	0.0317

The adjusted R square is slightly larger than in the PREDICT model and very close to the R square.


The model F is large, but smaller than in the PREDICT model. However, it is still highly significant. All terms included in the model are significant except **Weight**. The p -values for **Age**, **Run_Pulse**, and **Maximum_Pulse** are smaller in this model than they were in the PREDICT model.

Including the additional variable in the model changes the coefficients of the other terms and changes the t statistics for all.




3.06 Multiple Choice Poll

Which value tends to increase (can never decrease) as you add predictor variables to your regression model?

- a. R square
- b. Adjusted R square
- c. Mallows' C_p
- d. Both a and b
- e. F statistic
- f. All of the above


THE
SAS INSTITUTE

Stepwise Selection Methods

	<p>FORWARD SELECTION</p>
	<p>BACKWARD ELIMINATION</p>
	<p>STEPWISE SELECTION</p>

89

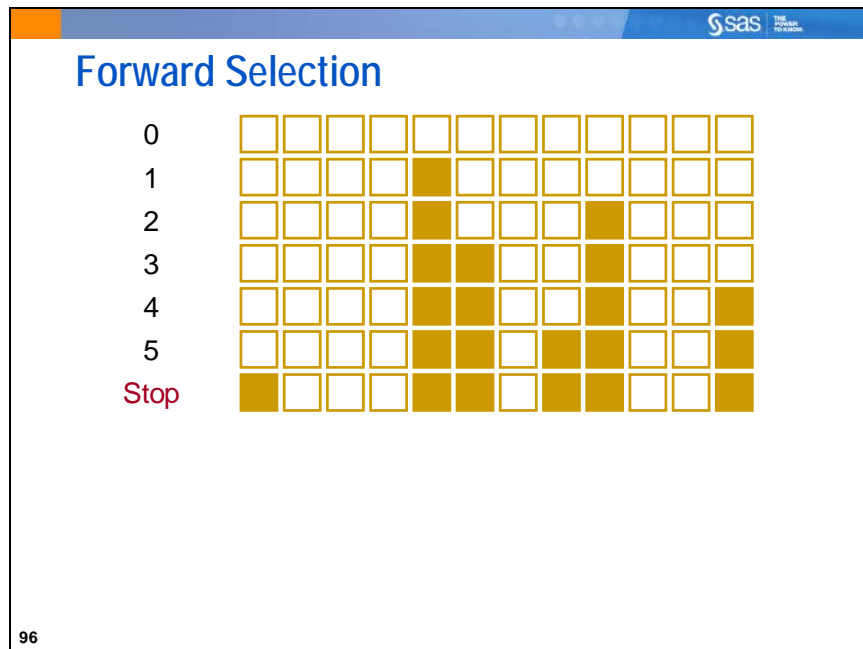
The all-possible regression technique that was discussed can be computer intensive, especially if there are a large number of potential independent variables.

PROC REG also offers the following stepwise SELECTION= options:

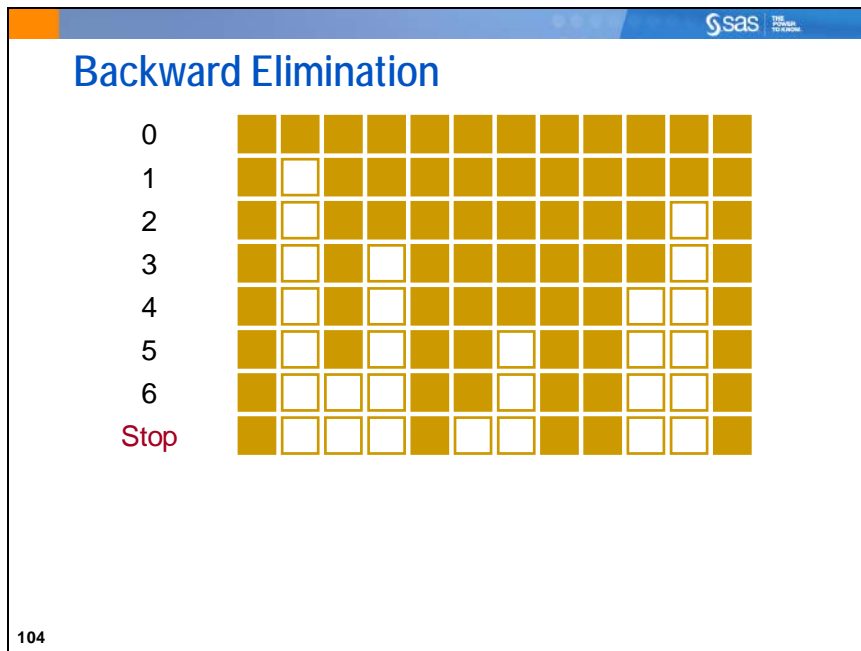
- FORWARD** first selects the best one-variable model. Then it selects the best two variables among those that contain the first selected variable. FORWARD continues this process, but stops when it reaches the point where no additional variables have p -value levels less than some stopping criterion (0.50, by default).
- BACKWARD** starts with the full model. Next, the variable that is least significant, given the other variables, is removed from the model. BACKWARD continues this process until all of the remaining variables have p -values less than a stopping criterion value (0.10, by default).
- STEPWISE** works like a combination of the FORWARD and BACKWARD method. The default entry p -value is 0.15 and the default stay p -value is also 0.15.



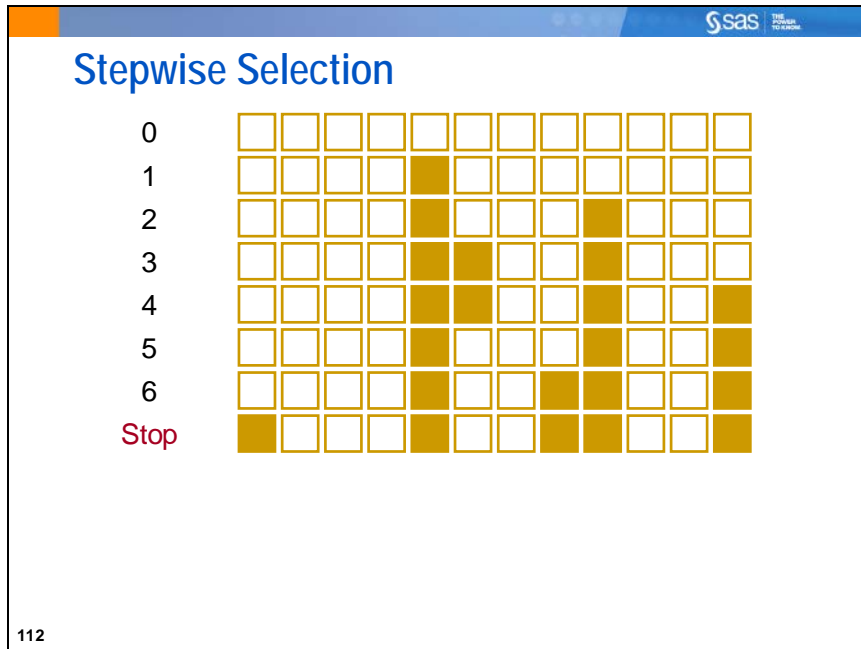
The SLENTY= (for forward step stopping criteria) and SLSTAY= (for backward step stopping criteria) options can be used to change the default stopping values.



Forward selection starts with an empty model. The method computes an F statistic for each predictor variable not in the model and examines the largest of these statistics. If it is significant at a specified significance level (specified by the `SLENTY=` option), the corresponding variable is added to the model. After a variable is entered in the model, it is never removed from the model. The process is repeated until none of the remaining variables meets the specified level for entry. By default, `SLENTY=0.50`.



Backward elimination starts off with the full model. Results of the F test for individual parameter estimates are examined, and the least significant variable that falls above the specified significance level (specified by the SLSTAY= option) is removed. After a variable is removed from the model, it remains excluded. The process is repeated until no other variable in the model meets the specified significance level for removal. By default, SLSTAY=0.10.



Stepwise selection is similar to forward selection in that it starts with an empty model and incrementally builds a model one variable at a time. However, the method differs from forward selection in that variables already in the model do not necessarily remain. The backward component of the method removes variables from the model that do not meet the significance criteria specified in the SLSTAY= option. The stepwise selection process terminates if no further variables can be added to the model or if the variable entered into the model is the only variable removed in the subsequent backward elimination. By default, SLENTY=0.15 and SLSTAY=0.15.

Stepwise selection (Forward, Backward, and Stepwise) has some serious shortcomings. Simulation studies (Derksen and Keselman 1992) evaluating variable selection techniques found the following:

1. The degree of collinearity among the predictor variables affected the frequency with which authentic predictor variables found their way into the final model.
2. The number of candidate predictor variables affected the number of noise variables that gained entry to the model.
3. The size of the sample was of little practical importance in determining the number of authentic variables contained in the final model.

One recommendation is to use the variable selection methods to create several candidate models, and then use subject-matter knowledge to select the variables that result in the best model within the scientific or business context of the problem. Therefore, you are simply using these methods as a useful tool in the model-building process (Hosmer and Lemeshow 2000).

Are p -values and Parameter Estimates Correct?

Automated model selection results in the following:

- biases in parameter estimates, predictions, and standard errors
- incorrect calculation of degrees of freedom
- p -values that tend to err on the side of overestimating significance (increasing Type I Error probability)

113

Statisticians give warnings and cautions about the appropriate interpretation of p -values from models chosen using any automated variable selection technique. Refitting many submodels in terms of an optimum fit to the data distorts the significance levels of conventional statistical tests. However, many researchers and users of statistical software neglect to report that the models that they ended up with were chosen using automated methods. They report statistical quantities such as standard errors, confidence limits, p -values, and R square as if the resulting model were entirely prespecified. These inferences are inaccurate, tending to err on the side of overstating the significance of predictors and making predictions with overly optimistic confidence. This problem is very evident when there are many iterative stages in model building. When there are many variables and you use stepwise selection to find a small subset of variables, inferences become less accurate (Chatfield 1995, Raftery 1994, Freedman 1983).

One solution to this problem is to split your data. One part can be used for finding the regression model and the other part can be used for inference. Another solution is to use bootstrapping methods to obtain the correct standard errors and p -values. *Bootstrapping* is a resampling method that tries to approximate the distribution of the parameter estimates to estimate the standard error.



Stepwise Regression

Example: Select a model for predicting **Oxygen_Consumption** in the **fitness** data set by using the FORWARD, BACKWARD, and STEPWISE methods.

```
/*st103d07.sas*/
proc reg data=sasuser.fitness plots(only)=adjrsq;
  FORWARD: model oxygen_consumption=
              Performance RunTime Age Weight
              Run_Pulse Rest_Pulse Maximum_Pulse
  / selection=forward;
  BACKWARD: model oxygen_consumption=
              Performance RunTime Age Weight
              Run_Pulse Rest_Pulse Maximum_Pulse
  / selection=backward;
  STEPWISE: model oxygen_consumption=
              Performance RunTime Age Weight
              Run_Pulse Rest_Pulse Maximum_Pulse
  / selection=stepwise;
  title 'Best Models Using Stepwise Selection';
run;
quit;
```

Partial PROC REG Output

Number of Observations Read	31
Number of Observations Used	31

Forward Selection: Step 1

Variable RunTime Entered: R-Square = 0.7434 and C(p) = 11.9967

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	633.01458	633.01458	84.00	<.0001
Error	29	218.53997	7.53586		
Corrected Total	30	851.55455			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	82.42494	3.85582	3443.63138	456.97	<.0001
RunTime	-3.31085	0.36124	633.01458	84.00	<.0001

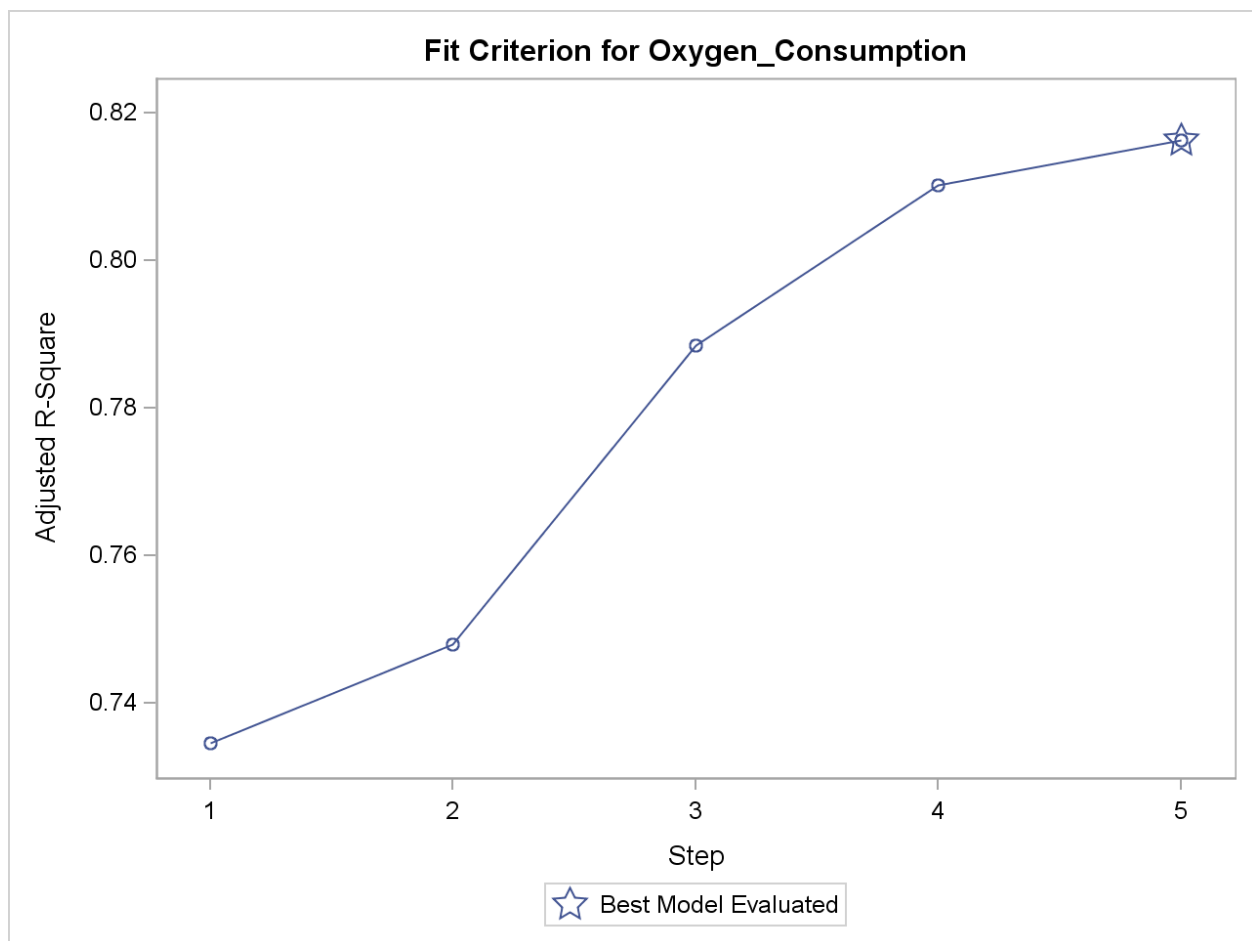
...

Partial PROC REG Output (Continued)

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	11.9967	84.00	<.0001
2	Age	2	0.0213	0.7647	10.7530	2.54	0.1222
3	Run_Pulse	3	0.0449	0.8096	5.9367	6.36	0.0179
4	Maximum_Pulse	4	0.0259	0.8355	4.0004	4.09	0.0534
5	Weight	5	0.0115	0.8469	4.2598	1.87	0.1836

The model selected at each step is printed and a summary of the sequence of steps is given at the end of the output. In the summary, the variables are listed in the order in which they were selected. The partial R square shows the increase in the model R square as each term was added.

The model that FORWARD selected has the same variables as the model chosen using the all-regressions techniques with the Hocking criterion. This will not always be the case.



The Adjusted R-Square plot shows the progression of that statistic at each step. The star denotes the best model of the five that were tested. This is not necessarily the highest adjusted R-square value of all possible subsets, but is the best of the five tested in the Forward model.

Partial PROC REG Output (Continued)

Backward Elimination: Step 0**All Variables Entered: R-Square = 0.8486 and C(p) = 8.0000**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	722.66124	103.23732	18.42	<.0001
Error	23	128.89331	5.60406		
Corrected Total	30	851.55455			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	131.78249	72.20754	18.66607	3.33	0.0810
Performance	-0.12619	0.30097	0.98519	0.18	0.6789
RunTime	-3.86019	2.93659	9.68350	1.73	0.2016
Age	-0.46082	0.58660	3.45842	0.62	0.4401
Weight	-0.05812	0.06892	3.98514	0.71	0.4078
Run_Pulse	-0.36207	0.12324	48.37354	8.63	0.0074
Rest_Pulse	-0.01512	0.06817	0.27581	0.05	0.8264
Maximum_Pulse	0.30102	0.13981	25.97886	4.64	0.0420

Bounds on condition number: 162.85, 2262.9**Backward Elimination: Step 1****Variable Rest_Pulse Removed: R-Square = 0.8483 and C(p) = 6.0492**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	722.38543	120.39757	22.37	<.0001
Error	24	129.16912	5.38205		
Corrected Total	30	851.55455			

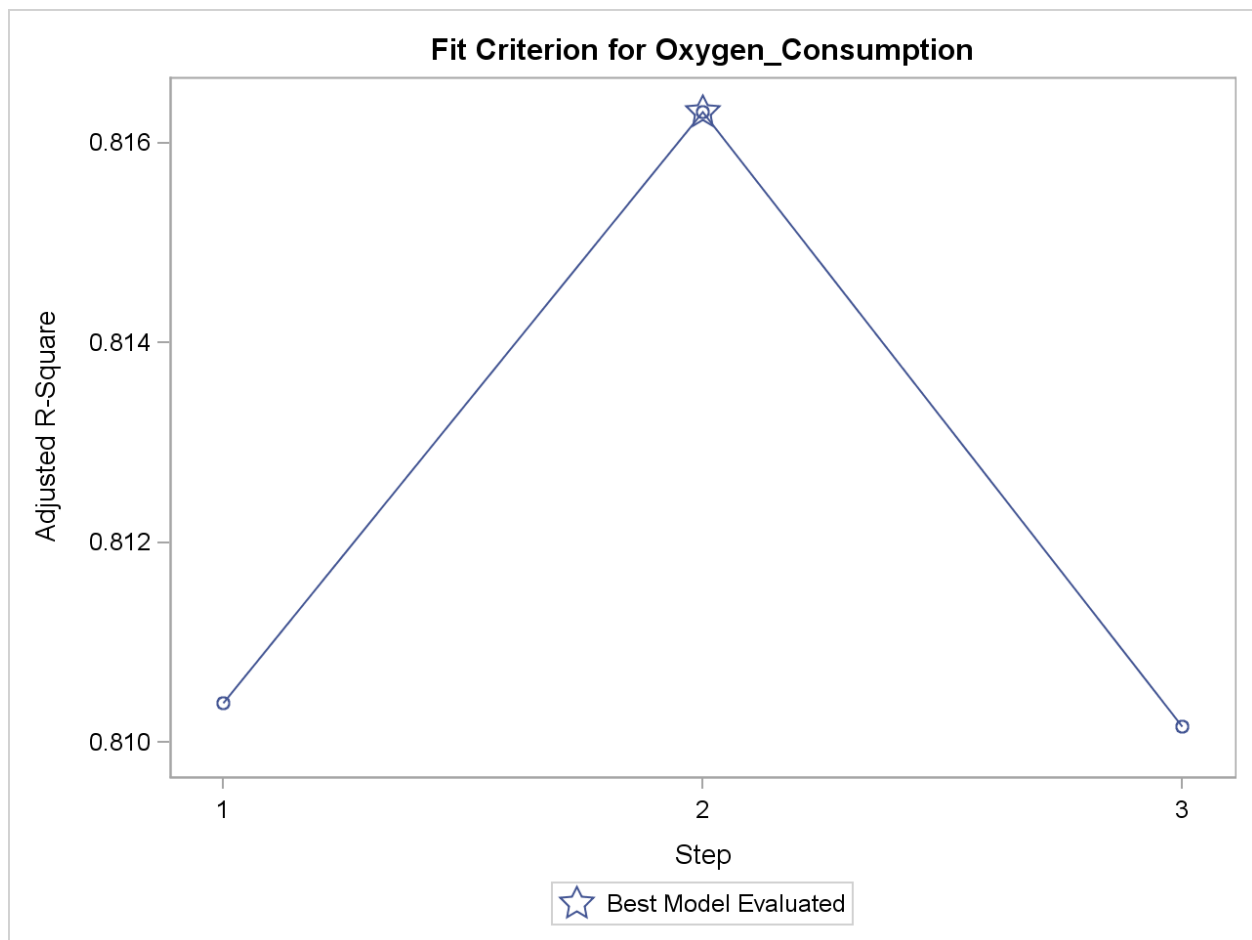
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	133.73795	70.23358	19.51494	3.63	0.0689
Performance	-0.13647	0.29144	1.18011	0.22	0.6438
RunTime	-3.99624	2.81438	10.85139	2.02	0.1685
Age	-0.47577	0.57106	3.73583	0.69	0.4130
Weight	-0.05545	0.06650	3.74132	0.70	0.4126
Run_Pulse	-0.36430	0.12037	49.29878	9.16	0.0058
Maximum_Pulse	0.30184	0.13696	26.13890	4.86	0.0374

...

Partial PROC REG Output (Continued)

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Rest_Pulse	6	0.0003	0.8483	6.0492	0.05	0.8264
2	Performance	5	0.0014	0.8469	4.2598	0.22	0.6438
3	Weight	4	0.0115	0.8355	4.0004	1.87	0.1836

Using the BACKWARD elimination option and the default p -value, three independent variables were eliminated. By coincidence the final model is the same as the one considered best based on C_p , using the Mallows criterion.

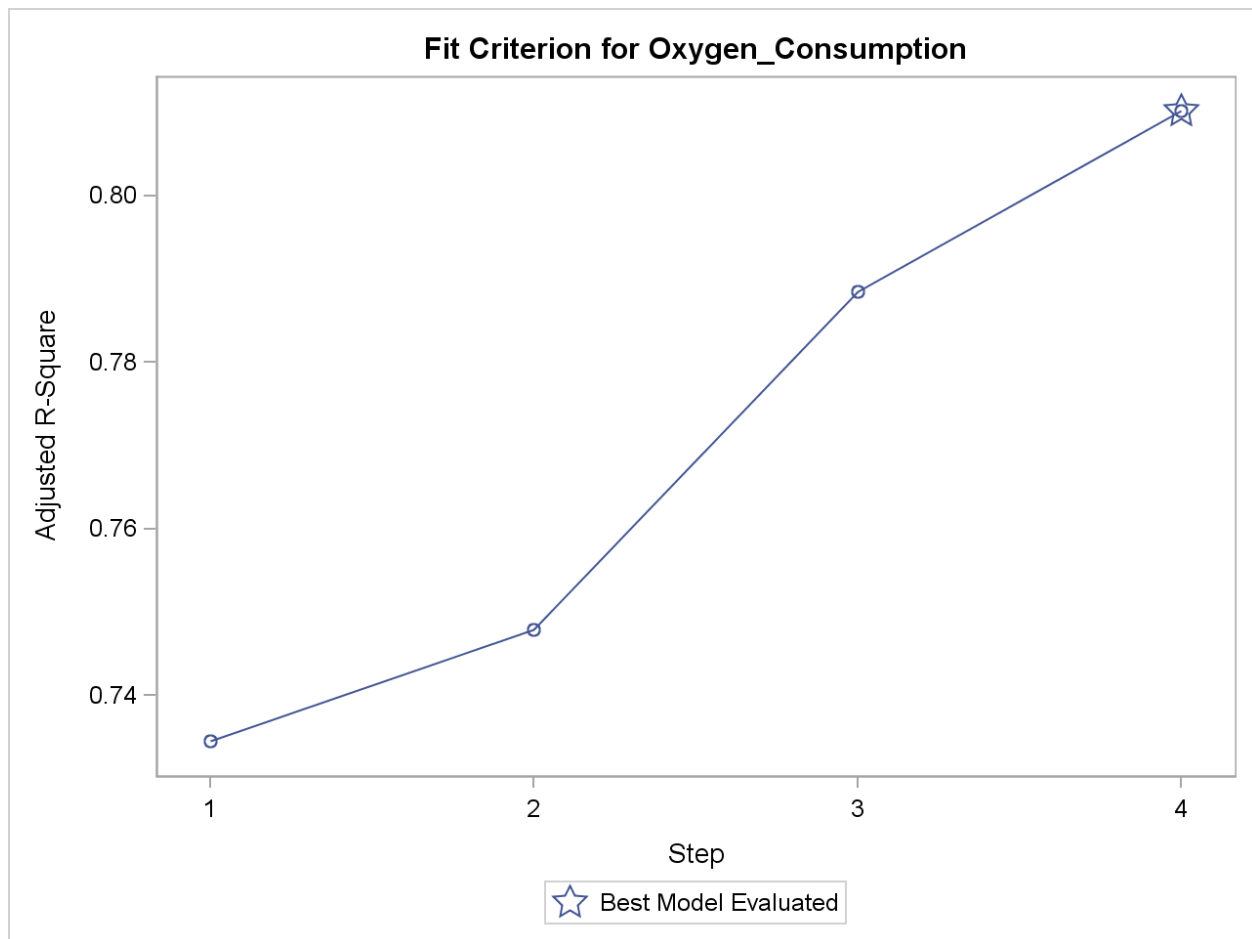


The adjusted R-square for the model at step 2 (before **Weight** was removed) was greatest of the three tested.

Partial PROC REG Output (Continued)

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	RunTime		1	0.7434	0.7434	11.9967	84.00	<.0001
2	Age		2	0.0213	0.7647	10.7530	2.54	0.1222
3	Run_Pulse		3	0.0449	0.8096	5.9367	6.36	0.0179
4	Maximum_Pulse		4	0.0259	0.8355	4.0004	4.09	0.0534

Using the STEPWISE option and the default variable entry and removal p -value criteria, the same subset resulted as that using the BACKWARD option.



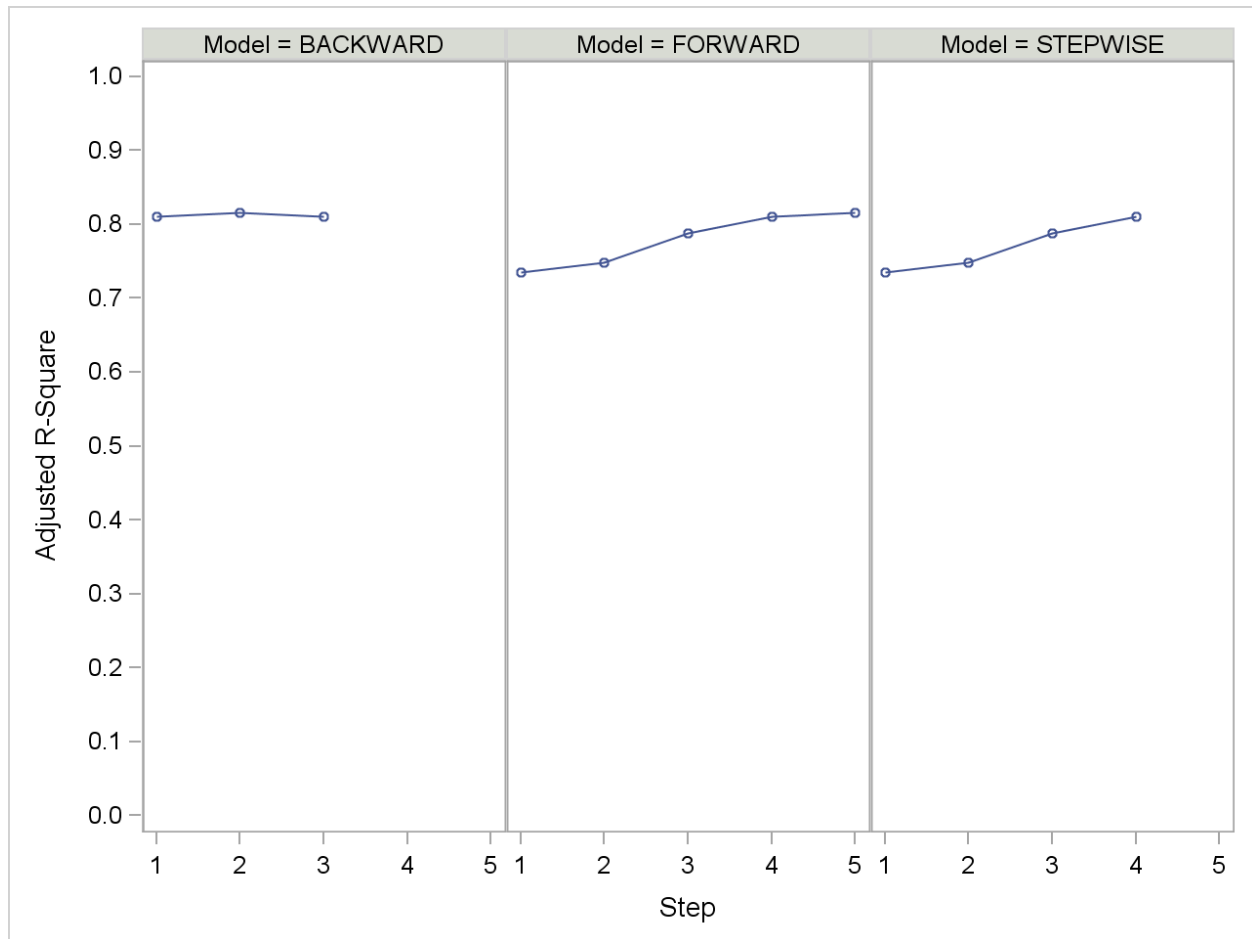
The SLENTY= default criterion is $p < 0.50$ for the FORWARD method and $p < .15$ for the STEPWISE method. After **RunTime** was entered into the model, **Age** was entered at step 2 with a p -value of 0.1222. If the SLENTY= criterion were set to something less than 0.10, the final model would be quite different. It would include only one variable, **RunTime**. This underscores the precariousness of relying on one stepwise method for defining a “best” model.



The scale of the default Y axes in these plots might give misleading information about the effect of adding or removing variables. The same plots displayed side-by-side and using a common y-scale of 0 to 1 is shown below. The differences do not look nearly as great.



The “Bounds on the condition number” reported at each step of the output for the STEPWISE selection methods refer to a measurement of *collinearity* (correlation among predictor variables). (The concept of collinearity is discussed in a later chapter.)



Stepwise Regression Models*	
FORWARD	RunTime, Age, Weight, Run_Pulse, Maximum_Pulse
BACKWARD	RunTime, Age, Run_Pulse, Maximum_Pulse
STEPWISE	RunTime, Age, Run_Pulse, Maximum_Pulse
* Using default values of SLENTY and SLSTAY	


115

The final models obtained using the default SLENTY= and SLSTAY= criteria are displayed. It is important to note that the choice of criterion levels can greatly affect the final models that are selected using stepwise methods. Some analysts use the defaults to get models to a manageable size then do manual reduction instead of using low values for SLENTY and SLSTAY.

Stepwise Models, Alternative Criteria	
FORWARD (slentry=0.05)	RunTime
BACKWARD (slstay=0.05)	RunTime, Run_Pulse, Maximum_Pulse
STEPWISE (slentry=0.05, slstay=0.05)	RunTime

116

The final models using 0.05 as the forward and backward step criteria resulted in very different models than those chosen using the default criteria.



Comparison of Selection Methods

Stepwise regression	uses fewer computer resources.
All-possible regression	generates more candidate models that might have nearly equal R^2 statistics and C_p statistics.

117

The stepwise regression methods have an advantage when there are a large number of independent variables.

With the all-possible regression techniques, you can compare essentially equivalent models and use your knowledge of the data set and subject area to select a model that is more easily interpreted.



Exercises

6. Using All-Regression Techniques

Use the `sasuser.BodyFat2` data set to identify a set of “best” models.

- a. With the `SELECTION=CP` option, use an all-possible regression technique to identify a set of candidate models that predict **PctBodyFat2** as a function of the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.

Hint: Select only the best 60 models based on C_p to compare.

- b. Use a stepwise regression method to select a candidate model. Try `FORWARD`, `STEPWISE`, and `BACKWARD`.
- c. How many variables would result from a model using `FORWARD` selection and a significance level for entry criterion of 0.05, instead of the default `SLENTRY` of 0.50?

3.07 Poll

The `STEPWISE`, `BACKWARD`, and `FORWARD` strategies result in the same final model if the same significance levels are used in all three.

- ☐ True
- ☐ False

3.5 Solutions

Solutions to Exercises

1. Describing the Relationships between Continuous Variables

- a. Generate scatter plots and correlations for the VAR variables **Age**, **Weight**, **Height**, and the circumference measures versus the WITH variable, **PctBodyFat2**.



Important! ODS Graphics in PROC CORR limits you to 10 VAR variables at a time, so for this exercise, look at the relationships with **Age**, **Weight**, and **Height** separately from the other variables.



Correlation tables can be created using more than 10 VAR variables at a time.

```
/*st103s01.sas*/ /*Part A*/
proc corr data=sasuser.BodyFat2 rank
    plots(only)=scatter(nvar=all ellipse=none);
    var Age Weight Height;
    with PctBodyFat2;
    title "Correlations and Scatter Plots with Body Fat %";
run;

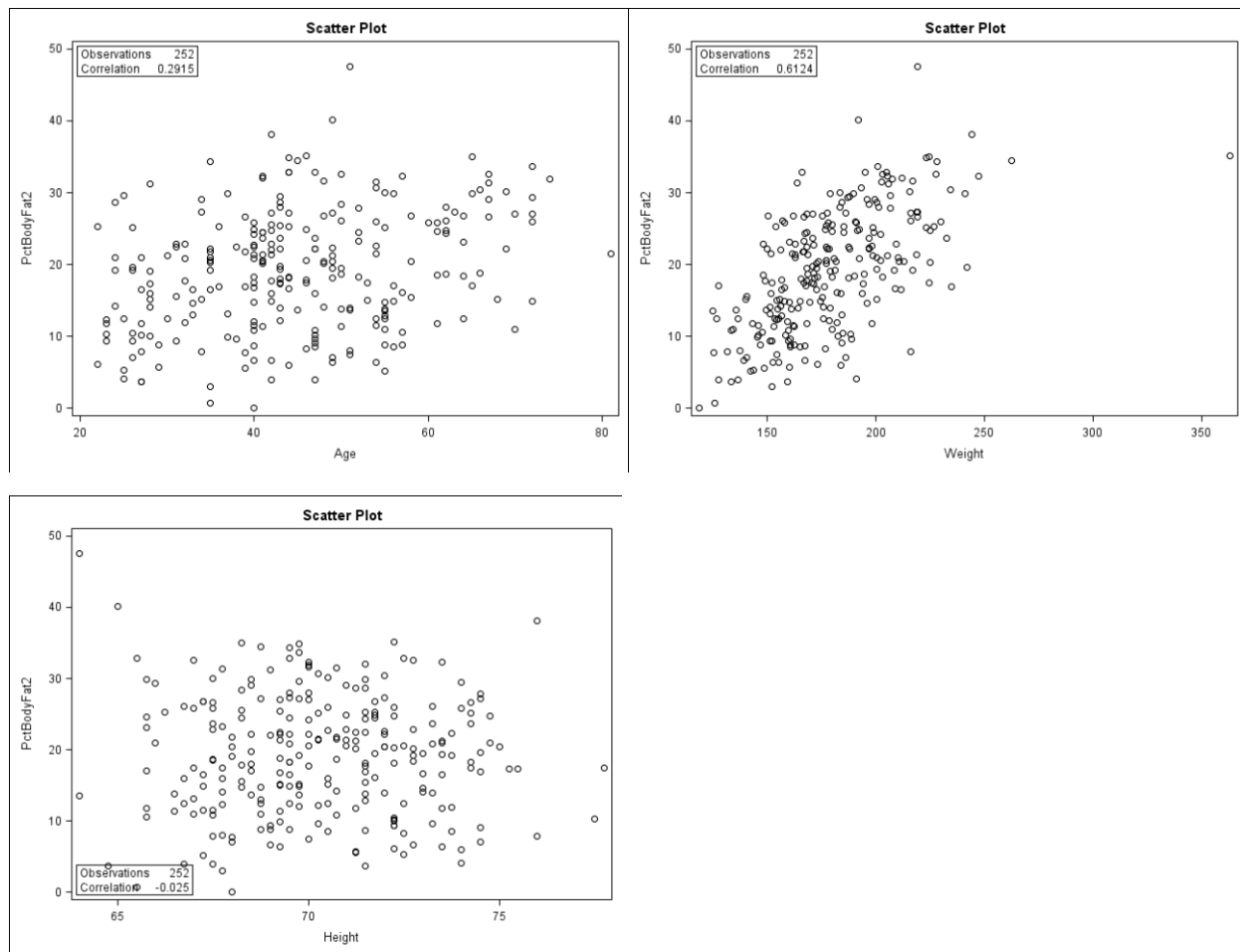
proc corr data=sasuser.BodyFat2 rank
    plots(only)=scatter(nvar=all ellipse=none);
    var Neck Chest Abdomen Hip Thigh
        Knee Ankle Biceps Forearm Wrist;
    with PctBodyFat2;
    title "Correlations and Scatter Plots with Body Fat %";
run;
```

1 With Variables:	PctBodyFat2		
3 Variables:	Age	Weight	Height

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PctBodyFat2	252	19.15079	8.36874	4826	0	47.50000
Age	252	44.88492	12.60204	11311	22.00000	81.00000
Weight	252	178.92440	29.38916	45089	118.50000	363.15000
Height	252	70.30754	2.60958	17718	64.00000	77.75000

Pearson Correlation Coefficients, N = 252
Prob > |r| under H0: Rho=0

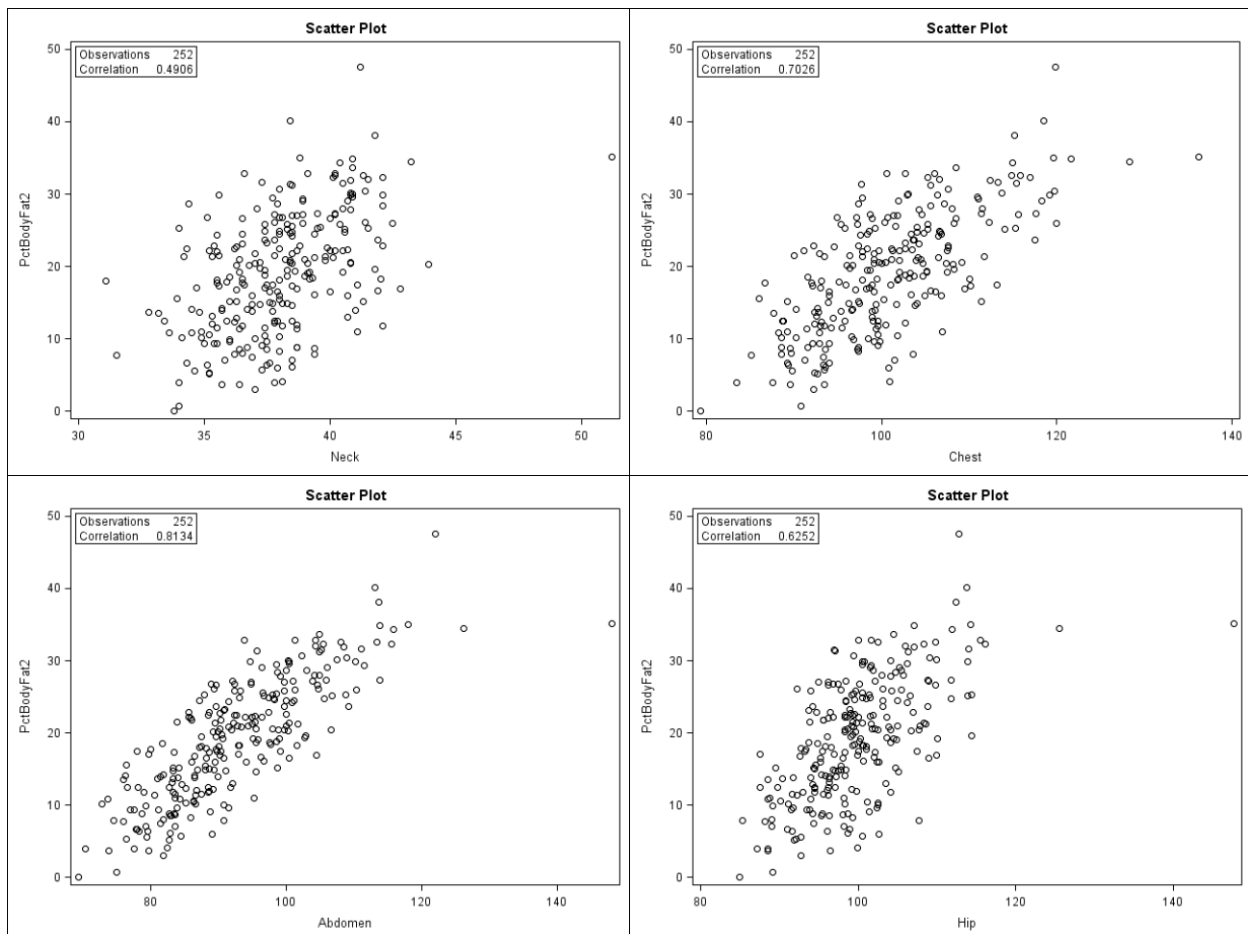
PctBodyFat2	Weight	Age	Height
	0.61241	0.29146	-0.02529
	<.0001	<.0001	0.6895

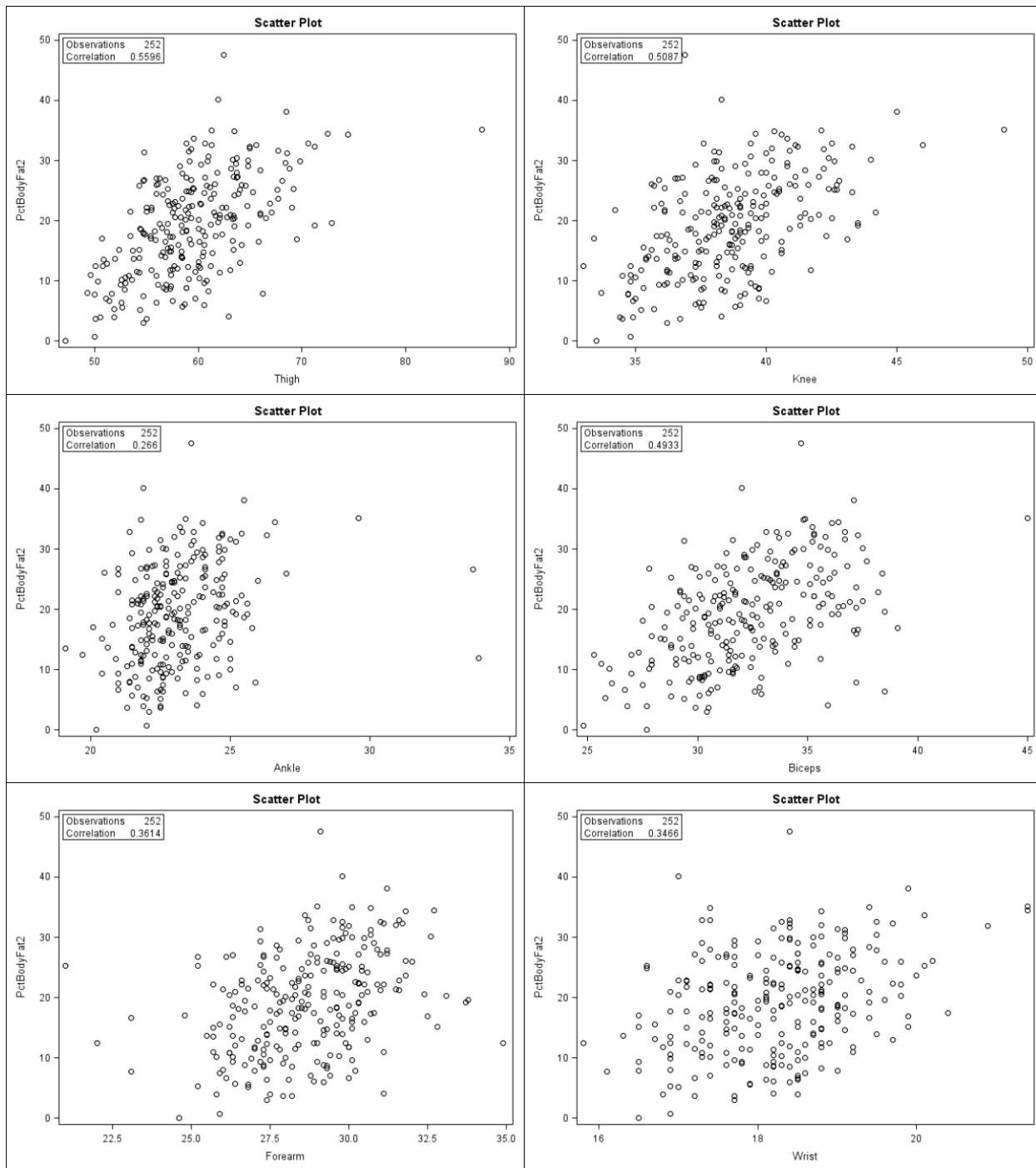


1 With Variables:	PctBodyFat2									
10 Variables:	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
PctBodyFat2	252	19.15079	8.36874	4826	0	47.50000
Neck	252	37.99206	2.43091	9574	31.10000	51.20000
Chest	252	100.82421	8.43048	25408	79.30000	136.20000
Abdomen	252	92.55595	10.78308	23324	69.40000	148.10000
Hip	252	99.90476	7.16406	25176	85.00000	147.70000
Thigh	252	59.40595	5.24995	14970	47.20000	87.30000
Knee	252	38.59048	2.41180	9725	33.00000	49.10000
Ankle	252	23.10238	1.69489	5822	19.10000	33.90000
Biceps	252	32.27341	3.02127	8133	24.80000	45.00000
Forearm	252	28.66389	2.02069	7223	21.00000	34.90000
Wrist	252	18.22976	0.93358	4594	15.80000	21.40000

Pearson Correlation Coefficients, N = 252 Prob > r under H0: Rho=0											
PctBodyFat2	Abdomen	Chest	Hip	Thigh	Knee	Biceps	Neck	Forearm	Wrist	Ankle	
	0.81343	0.70262	0.62520	0.55961	0.50867	0.49327	0.49059	0.36139	0.34657	0.26597	
	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	<.0001	





1) Can straight lines adequately describe the relationships?

Height seems to be the only variable that shows no real linear relationship. Age and Ankle show little linear trend.

2) Are there any outliers that you should investigate?

The Weight outlier is present again, as well as Neck, Abdomen, Hip, Knee, and Biceps. There are two outliers for Ankle.

3) What variable has the highest correlation with **PctBodyFat2**?

Abdomen, with 0.81343, is the variable with the highest correlation with **PctBodyFat2**.

a) What is the p -value for the coefficient?

<.0001

b) Is it statistically significant at the 0.05 level?

Yes

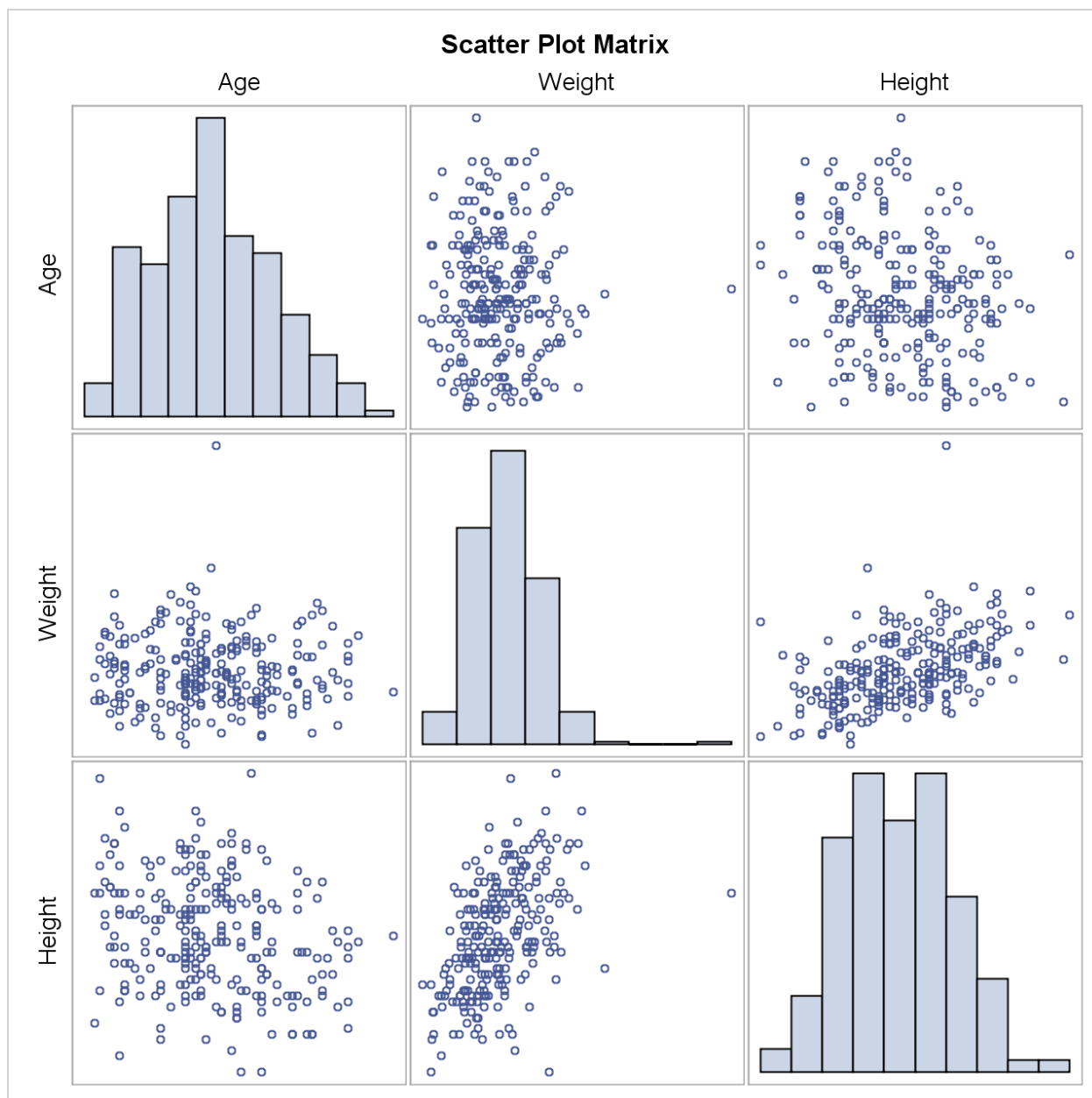
b. Generate correlations among all of the VAR variables (**Age, Weight, Height**) among one another and among the circumference measures. Are there any notable relationships?

```
/*st103s01.sas*/  /*Part B*/
proc corr data=sasuser.BodyFat2 nosimple
      plots=matrix(nvar=all histogram);
  var Age Weight Height;
  title "Correlations and Scatter Plot Matrix of Basic Measures";
run;

proc corr data=sasuser.BodyFat2 nosimple
      plots=matrix(nvar=all histogram);
  var Neck Chest Abdomen Hip Thigh
      Knee Ankle Biceps Forearm Wrist;
  title "Correlations and Scatter Plot Matrix of Circumferences";
run;

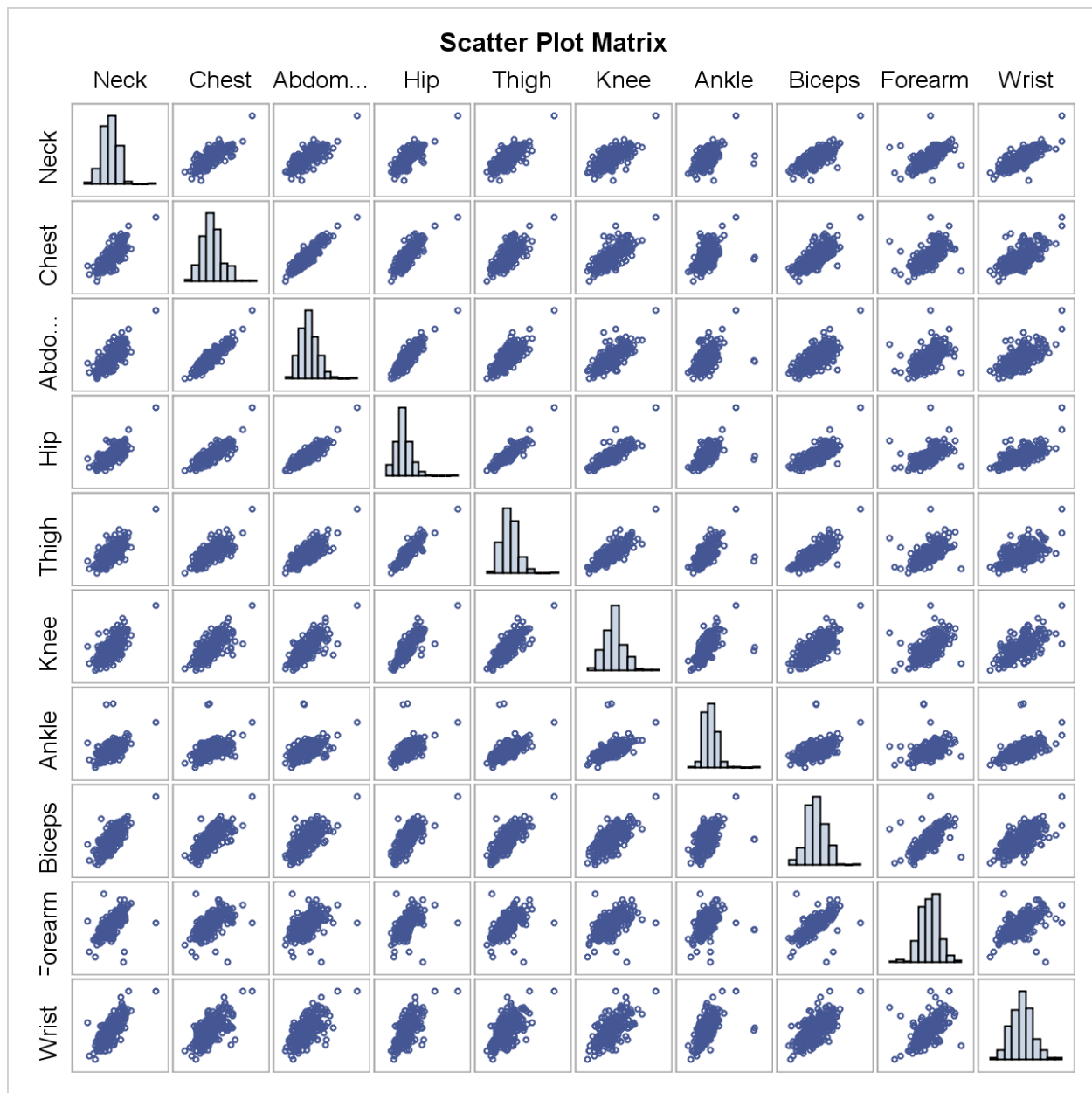
proc corr data=sasuser.BodyFat2 nosimple
      plots=matrix(nvar=all histogram);
  var Neck Chest Abdomen Hip Thigh
      Knee Ankle Biceps Forearm Wrist;
  with Age Weight Height;
  title "Correlations and Scatter Plot Matrix of Circumferences";
run;
```

Pearson Correlation Coefficients, N = 252 Prob > r under H0: Rho=0			
	Age	Weight	Height
Age	1.00000	-0.01275 0.8404	-0.24521 <.0001
Weight	-0.01275 0.8404	1.00000	0.48689 <.0001
Height	-0.24521 <.0001	0.48689 <.0001	1.00000



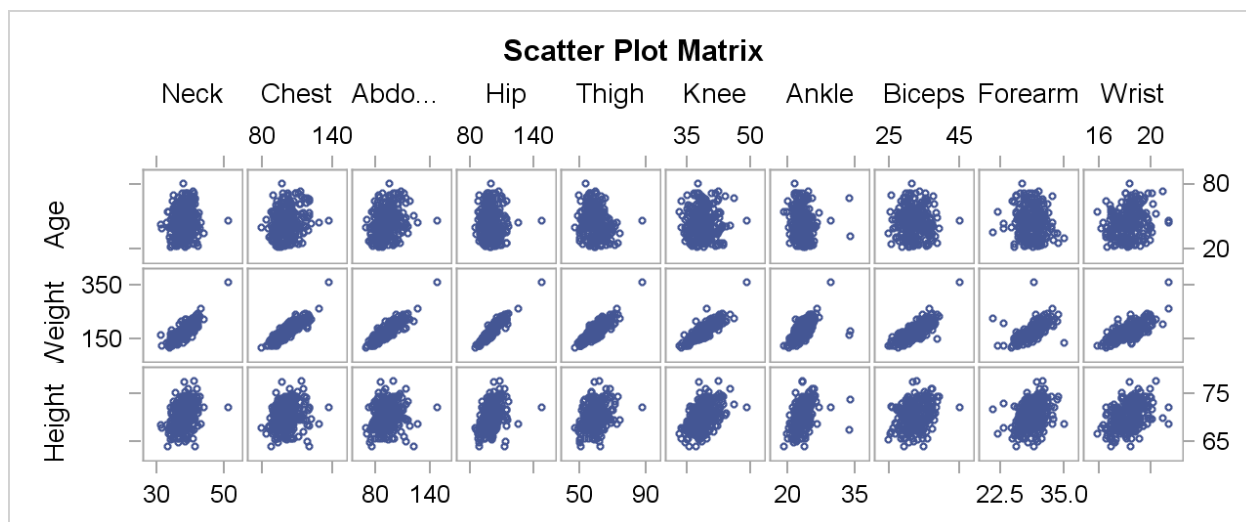
Height and Weight seem to correlate relatively strongly. The outlier might affect the measurement of the relationship.

Pearson Correlation Coefficients, N = 252 Prob > r under H0: Rho=0										
	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Neck	1.00000	0.78484 <.0001	0.75408 <.0001	0.73496 <.0001	0.69570 <.0001	0.67240 <.0001	0.47789 <.0001	0.73115 <.0001	0.62366 <.0001	0.74483 <.0001
Chest	0.78484 <.0001	1.00000	0.91583 <.0001	0.82942 <.0001	0.72986 <.0001	0.71950 <.0001	0.48299 <.0001	0.72791 <.0001	0.58017 <.0001	0.66016 <.0001
Abdomen	0.75408 <.0001	0.91583 <.0001	1.00000	0.87407 <.0001	0.76662 <.0001	0.73718 <.0001	0.45322 <.0001	0.68498 <.0001	0.50332 <.0001	0.61983 <.0001
Hip	0.73496 <.0001	0.82942 <.0001	0.87407 <.0001	1.00000	0.89641 <.0001	0.82347 <.0001	0.55839 <.0001	0.73927 <.0001	0.54501 <.0001	0.63009 <.0001
Thigh	0.69570 <.0001	0.72986 <.0001	0.76662 <.0001	0.89641 <.0001	1.00000	0.79917 <.0001	0.53980 <.0001	0.76148 <.0001	0.56684 <.0001	0.55868 <.0001
Knee	0.67240 <.0001	0.71950 <.0001	0.73718 <.0001	0.82347 <.0001	0.79917 <.0001	1.00000	0.61161 <.0001	0.67871 <.0001	0.55590 <.0001	0.66451 <.0001
Ankle	0.47789 <.0001	0.48299 <.0001	0.45322 <.0001	0.55839 <.0001	0.53980 <.0001	0.61161 <.0001	1.00000	0.48485 <.0001	0.41905 <.0001	0.56619 <.0001
Biceps	0.73115 <.0001	0.72791 <.0001	0.68498 <.0001	0.73927 <.0001	0.76148 <.0001	0.67871 <.0001	0.48485 <.0001	1.00000	0.67826 <.0001	0.63213 <.0001
Forearm	0.62366 <.0001	0.58017 <.0001	0.50332 <.0001	0.54501 <.0001	0.56684 <.0001	0.55590 <.0001	0.41905 <.0001	0.67826 <.0001	1.00000	0.58559 <.0001
Wrist	0.74483 <.0001	0.66016 <.0001	0.61983 <.0001	0.63009 <.0001	0.55868 <.0001	0.66451 <.0001	0.56619 <.0001	0.63213 <.0001	0.58559 <.0001	1.00000



There are several relationships that appear to have high correlations (such as those among Hip, Thigh, and Knee).

Pearson Correlation Coefficients, N = 252 Prob > r under H0: Rho=0										
	Neck	Chest	Abdomen	Hip	Thigh	Knee	Ankle	Biceps	Forearm	Wrist
Age	0.11351 0.0721	0.17645 0.0050	0.23041 0.0002	-0.05033 0.4263	-0.20010 0.0014	0.01752 0.7820	-0.10506 0.0961	-0.04116 0.5154	-0.08506 0.1783	0.21353 0.0006
Weight	0.83072 <.0001	0.89419 <.0001	0.88799 <.0001	0.94088 <.0001	0.86869 <.0001	0.85317 <.0001	0.61369 <.0001	0.80042 <.0001	0.63030 <.0001	0.72977 <.0001
Height	0.32114 <.0001	0.22683 0.0003	0.18977 0.0025	0.37211 <.0001	0.33856 <.0001	0.50050 <.0001	0.39313 <.0001	0.31851 <.0001	0.32203 <.0001	0.39778 <.0001



Weight seems to correlate highly with all circumference variables.

2. Fitting a Simple Linear Regression Model

Use the **sasuser.BodyFat2** data set for this exercise.

- a. Perform a simple linear regression model with **PctBodyFat2** as the response variable and **Weight** as the predictor.

```
/*st103s02.sas*/ /*Part A*/
ods graphics off;
proc reg data=sasuser.BodyFat2;
  model PctBodyFat2=Weight;
  title "Regression of % Body Fat on Weight";
run;
quit;
ods graphics on;
```

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6593.01614	6593.01614	150.03	<.0001
Error	250	10986	43.94389		
Corrected Total	251	17579			

Root MSE	6.62902	R-Square	0.3751
Dependent Mean	19.15079	Adj R-Sq	0.3726
Coeff Var	34.61485		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-12.05158	2.58139	-4.67	<.0001
Weight	1	0.17439	0.01424	12.25	<.0001

- 1) What is the value of the F statistic and the associated p -value? How would you interpret this with regard to the null hypothesis?

The F value is 150.03 and the p -value is <.0001. You would reject the null hypothesis of no relationship.

- 2) Write the predicted regression equation.

**From the parameter estimates table, the predicted value equation is as follows:
PctBodyFat2=-12.05158+0.17439*Weight.**

- 3) What is the value of the R-square statistic? How would you interpret this?

The R-square value of 0.3751 can be interpreted to mean that 37.51% of the variability in PctBodyFat2 can be explained by Weight.

- b. Produce predicted values for PctBodyFat2 when Weight is 125, 150, 175, 200, and 225.

```

/*st103s02.sas*/  /*Part B*/
ods graphics off;
proc reg data=sasuser.BodyFat2 outest=Betas;
    PredBodyFat: model PctBodyFat2=Weight;
    title "Regression of % Body Fat on Weight";
run;
quit;
ods graphics on;

data ToScore;
    input Weight @@;
    datalines;
125 150 175 200 225
;
run;

proc score data=ToScore score=Betas
    out=Scored type=parms;
    var Weight;
run;

proc print data=Scored;
    title "Predicted % Body Fat from Weight 125 150 175 200 225";
run;

```

Obs	Weight	PredBodyFat
1	125	9.7470
2	150	14.1067
3	175	18.4664
4	200	22.8261
5	225	27.1859

What are the predicted values?

The predicted values are as listed in the output above under **PredBodyFat**.

3. Performing Multiple Regression Using the REG Procedure

- a. Using the **sasuser.BodyFat2** data set, run a regression of **PctBodyFat2** on the variables **Age**, **Weight**, **Height**, **Neck**, **Chest**, **Abdomen**, **Hip**, **Thigh**, **Knee**, **Ankle**, **Biceps**, **Forearm**, and **Wrist**.

- 1) Compare the ANOVA table with that from the model with only **Weight** in the previous exercise. What is different?

```
/*st103s03.sas*/ /*Part A*/
proc reg data=sasuser.BodyFat2;
  model PctBodyFat2=Age Weight Height
        Neck Chest Abdomen Hip Thigh
        Knee Ankle Biceps Forearm Wrist;
  title 'Regression of PctBodyFat2 on All '
        'Predictors';
run;
quit;
```

PROC REG Output

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	13	13159	1012.22506	54.50	<.0001
Error	238	4420.06401	18.57170		
Corrected Total	251	17579			

Root MSE	4.30949	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7348
Coeff Var	22.50293		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-21.35323	22.18616	-0.96	0.3368
Age	1	0.06457	0.03219	2.01	0.0460
Weight	1	-0.09638	0.06185	-1.56	0.1205
Height	1	-0.04394	0.17870	-0.25	0.8060
Neck	1	-0.47547	0.23557	-2.02	0.0447
Chest	1	-0.01718	0.10322	-0.17	0.8679
Abdomen	1	0.95500	0.09016	10.59	<.0001
Hip	1	-0.18859	0.14479	-1.30	0.1940
Thigh	1	0.24835	0.14617	1.70	0.0906
Knee	1	0.01395	0.24775	0.06	0.9552
Ankle	1	0.17788	0.22262	0.80	0.4251
Biceps	1	0.18230	0.17250	1.06	0.2917
Forearm	1	0.45574	0.19930	2.29	0.0231
Wrist	1	-1.65450	0.53316	-3.10	0.0021

There are key differences between the ANOVA table for this model and the Simple Linear Regression model.

- The degrees of freedom for the model are much higher, 13 versus 1.
 - The Mean Square model and the F ratio are much smaller.
- 2) How do the R square and the adjusted R square compare with these statistics for the **Weight** regression demonstration?

Both the R square and adjusted R square for the full models are larger than the simple linear regression. The multiple regression model explains almost 75% of the variation in the **PctBodyFat2** variable versus only about 37.5% explained by the simple linear regression model.

- 3) Did the estimate for the intercept change? Did the estimate for the coefficient of **Weight** change?

Yes, including the other variables in the model changed the estimates both of the intercept and the slope for **Weight**. Also, the p -values for both changed dramatically. The slope of **Weight** is now not significantly different from zero.

4. Simplifying the Model

- a. Rerun the model in **3a.**, but eliminate the variable with the highest p -value. Compare the output with the Exercise **3a.** model.

This program reruns the regression with **Knee** removed because it has the largest p -value (0.9552).

```
/*st103s03.sas*/ /*Part B*/
proc reg data=sasuser.BodyFat2;
  model PctBodyFat2=Age Weight Height
        Neck Chest Abdomen Hip Thigh
        Ankle Biceps Forearm Wrist;
  title 'Remove Knee';
run;
quit;
```

PROC REG Output

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	12	13159	1096.57225	59.29	<.0001
Error	239	4420.12286	18.49424		
Corrected Total	251	17579			

Root MSE	4.30049	R-Square	0.7486
Dependent Mean	19.15079	Adj R-Sq	0.7359
Coeff Var	22.45595		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-21.30204	22.12123	-0.96	0.3365
Age	1	0.06503	0.03108	2.09	0.0374
Weight	1	-0.09602	0.06138	-1.56	0.1191
Height	1	-0.04166	0.17369	-0.24	0.8107
Neck	1	-0.47695	0.23361	-2.04	0.0423
Chest	1	-0.01732	0.10298	-0.17	0.8666
Abdomen	1	0.95497	0.08998	10.61	<.0001
Hip	1	-0.18801	0.14413	-1.30	0.1933
Thigh	1	0.25089	0.13876	1.81	0.0719
Ankle	1	0.18018	0.21841	0.82	0.4102
Biceps	1	0.18182	0.17193	1.06	0.2913
Forearm	1	0.45667	0.19820	2.30	0.0221
Wrist	1	-1.65227	0.53057	-3.11	0.0021

- b. Did the p -value for the model change notably?

The p -value for the model did not change out to four decimal places.

- c. Did the R square and adjusted R square change notably?

The R square showed essentially no change. The adjusted R square increased from 0.7348 to 0.7359. When an adjusted R square increases by removing a variable from the model, it strongly implies that the removed variable was not necessary.

- d. Did the parameter estimates and their p -values change notably?

Some of the parameter estimates and their p -values changed slightly, none to any large degree.

5. More Simplifying of the Model

- a. Rerun the model in Exercise 4a, but drop the variable with the highest p -value.

This program reruns the regression with **Chest** removed, because it is the variable with the highest p -value in the previous model.

```
/*st103s03.sas*/ /*Part C*/
proc reg data=sasuser.BodyFat2;
  model PctBodyFat2=Age Weight Height
        Neck Abdomen Hip Thigh
        Ankle Biceps Forearm Wrist;
  title 'Remove Knee and Chest';
run;
quit;
```

PROC REG Output

Number of Observations Read	252
Number of Observations Used	252

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	13158	1196.21310	64.94	<.0001
Error	240	4420.64572	18.41936		
Corrected Total	251	17579			

Root MSE	4.29178	R-Square	0.7485
Dependent Mean	19.15079	Adj R-Sq	0.7370
Coeff Var	22.41044		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t
Intercept	1	-23.13736	19.20171	-1.20	0.2294
Age	1	0.06488	0.03100	2.09	0.0374
Weight	1	-0.10095	0.05380	-1.88	0.0618
Height	1	-0.03120	0.16185	-0.19	0.8473
Neck	1	-0.47631	0.23311	-2.04	0.0421
Abdomen	1	0.94965	0.08406	11.30	<.0001
Hip	1	-0.18316	0.14092	-1.30	0.1950
Thigh	1	0.25583	0.13534	1.89	0.0599
Ankle	1	0.18215	0.21765	0.84	0.4035
Biceps	1	0.18055	0.17141	1.05	0.2933
Forearm	1	0.45262	0.19634	2.31	0.0220
Wrist	1	-1.64984	0.52930	-3.12	0.0020

- b. How did the output change from the previous model?

The ANOVA table did not change greatly. The R square remained essentially unchanged. The adjusted R square increased again, which confirms that the variable Chest did not contribute to explaining the variation in PctBodyFat2 when the other variables are in the model.

- c. Did the number of parameters with p -values less than 0.05 change?

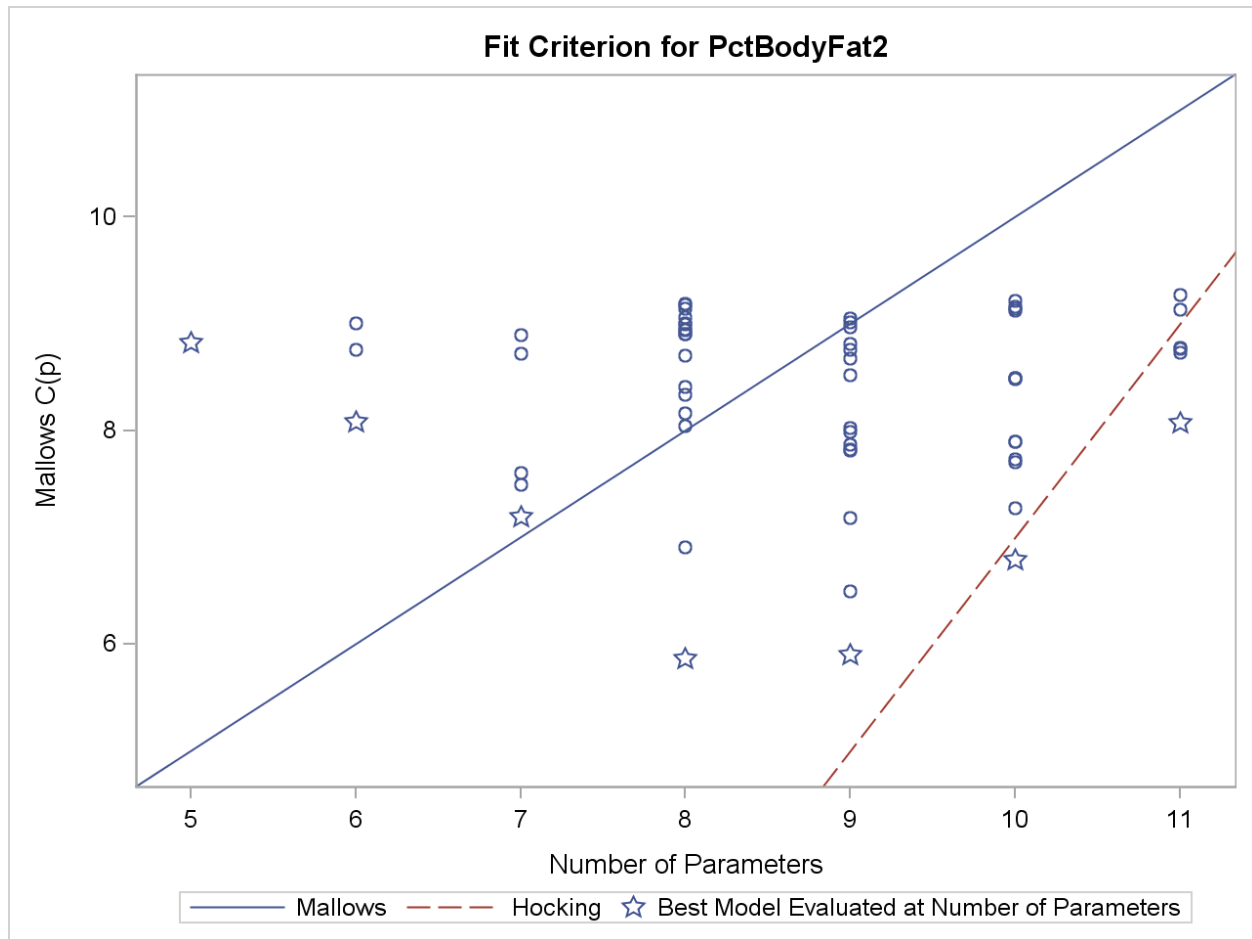
The p -value for Weight changed more than any other and is now just above 0.05. The p -values and parameter estimates for other variables changed much less. There are no more variables in this model with p -values below 0.05, compared with the previous one.

6. Using All-Regression Techniques

- a. With the SELECTION=CP option, use an all-possible regression technique to identify a set of candidate models that predict PctBodyFat2 as a function of the variables Age, Weight, Height, Neck, Chest, Abdomen, Hip, Thigh, Knee, Ankle, Biceps, Forearm, and Wrist. Hint: Select only the best 60 models based on C_p to compare.

```
/*st103s04.sas*/ /*Part A*/
ods graphics / imagemap=on;

proc reg data=sasuser.BodyFat2 plots(only)=(cp);
  model PctBodyFat2=Age Weight Height
        Neck Chest Abdomen Hip Thigh
        Knee Ankle Biceps Forearm Wrist
        / selection=cp best=60;
  title "Using Mallows Cp for Model Selection";
run;
quit;
```



The plot indicates that the best model according to Mallows' criterion is an eight-parameter (seven variables plus an intercept) model. The best model according to Hocking's criterion has 10 parameters (including the intercept).

A partial table of the 60 models, their $C(p)$ values, and the numbers of variables in the models is displayed.

Model Index	Number in Model	C(p)	R-Square	Variables in Model
1	7	5.8653	0.7445	Age Weight Neck Abdomen Thigh Forearm Wrist
2	8	5.8986	0.7466	Age Weight Neck Abdomen Hip Thigh Forearm Wrist
3	8	6.4929	0.7459	Age Weight Neck Abdomen Thigh Biceps Forearm Wrist
4	9	6.7834	0.7477	Age Weight Neck Abdomen Hip Thigh Biceps Forearm Wrist
5	7	6.9017	0.7434	Age Weight Neck Abdomen Biceps Forearm Wrist
6	8	7.1778	0.7452	Age Weight Neck Abdomen Thigh Ankle Forearm Wrist
7	6	7.1860	0.7410	Age Weight Abdomen Thigh Forearm Wrist
8	9	7.2729	0.7472	Age Weight Neck Abdomen Hip Thigh Ankle Forearm Wrist
9	6	7.4937	0.7406	Age Weight Neck Abdomen Forearm Wrist
10	6	7.6018	0.7405	Weight Neck Abdomen Biceps Forearm Wrist
11	9	7.7067	0.7468	Age Weight Neck Abdomen Thigh Ankle Biceps Forearm Wrist
12	9	7.7282	0.7467	Age Weight Height Neck Abdomen Hip Thigh Forearm Wrist
13	8	7.8146	0.7445	Age Weight Height Neck Abdomen Thigh Forearm Wrist
14	8	7.8246	0.7445	Age Weight Neck Chest Abdomen Thigh Forearm Wrist
15	8	7.8651	0.7445	Age Weight Neck Abdomen Thigh Knee Forearm Wrist
16	9	7.8966	0.7466	Age Weight Neck Abdomen Hip Thigh Knee Forearm Wrist
17	9	7.8986	0.7466	Age Weight Neck Chest Abdomen Hip Thigh Forearm Wrist
18	8	7.9907	0.7443	Age Weight Neck Abdomen Ankle Biceps Forearm Wrist



Number in Model does not include the intercept in this table.

The best MALLOWS model is either the eight-parameter models, number 1 (includes the variables Age, Weight, Neck, Abdomen, Thigh, Forearm, and Wrist) or number 5 (includes the variables Age, Weight, Neck, Abdomen, Biceps, Forearm, and Wrist).

The best HOCKING model is number 4. It includes Hip, along with the variables in the best MALLOWS models listed above.

- b. Use a stepwise regression method to select a candidate model. Try FORWARD, STEPWISE, and BACKWARD.

```

/*st103s04.sas*/ /*Part B*/
proc reg data=sasuser.BodyFat2 plots(only)=adjrsq;
  FORWARD: model PctBodyFat2=Age Weight Height
    Neck Chest Abdomen Hip Thigh
    Knee Ankle Biceps Forearm Wrist
    / selection=forward;
  BACKWARD: model PctBodyFat2=Age Weight Height
    Neck Chest Abdomen Hip Thigh
    Knee Ankle Biceps Forearm Wrist
    / selection=backward;
  STEPWISE: model PctBodyFat2=Age Weight Height
    Neck Chest Abdomen Hip Thigh
    Knee Ankle Biceps Forearm Wrist
    / selection=stepwise;
  title "Using Stepwise Methods for Model Selection";
run;
quit;

```


Partial Output

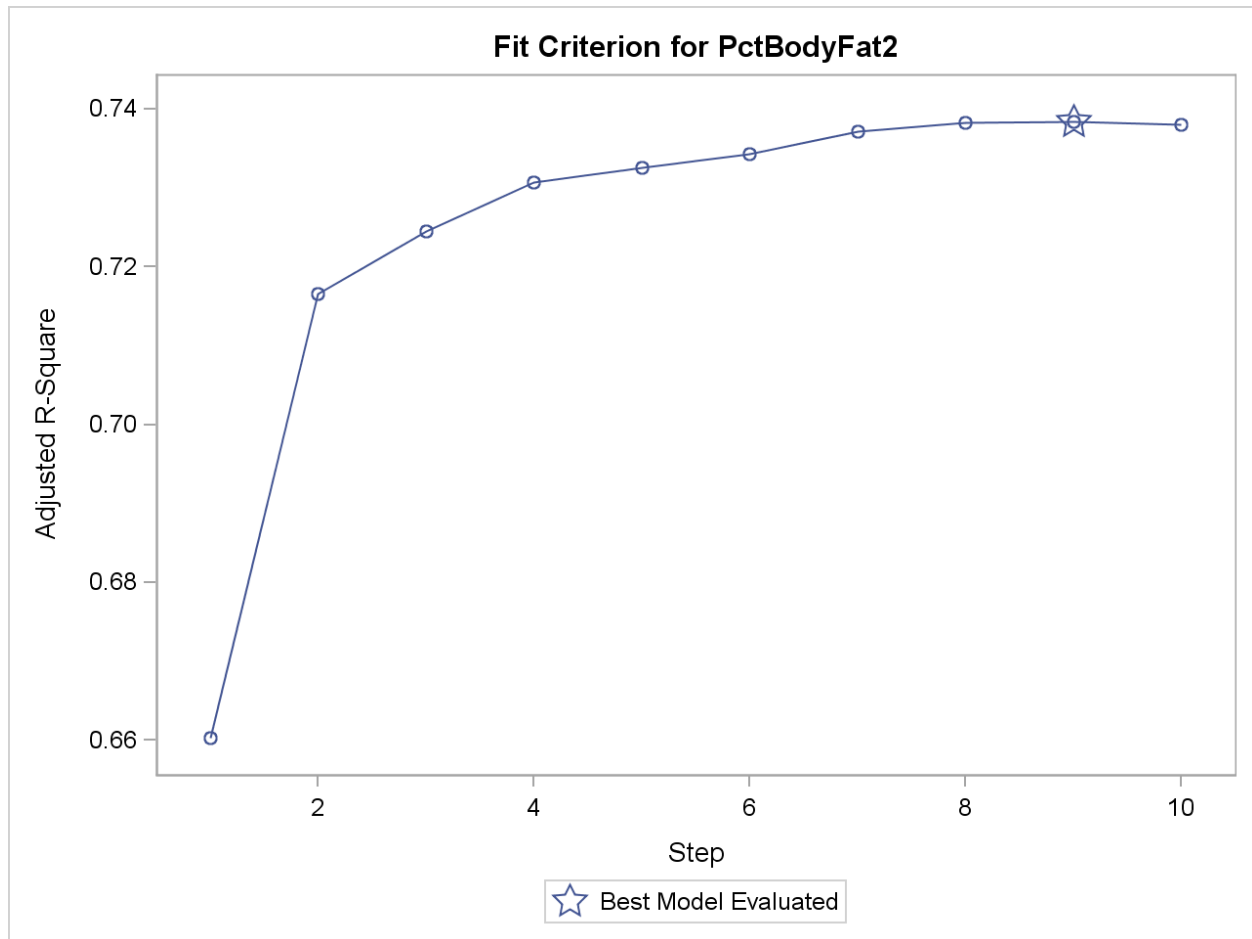
Forward Selection: Step 10**Variable Ankle Entered: R-Square = 0.7485 and C(p) = 8.0682**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	10	13158	1315.76595	71.72	<.0001
Error	241	4421.33035	18.34577		
Corrected Total	251	17579			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-25.99962	12.15316	83.96376	4.58	0.0334
Age	0.06509	0.03092	81.31425	4.43	0.0363
Weight	-0.10740	0.04207	119.56769	6.52	0.0113
Neck	-0.46749	0.22812	77.05006	4.20	0.0415
Abdomen	0.95772	0.07276	3178.52750	173.26	<.0001
Hip	-0.17912	0.13908	30.42960	1.66	0.1990
Thigh	0.25926	0.13389	68.78441	3.75	0.0540
Ankle	0.18453	0.21686	13.28232	0.72	0.3957
Biceps	0.18617	0.16858	22.37399	1.22	0.2705
Forearm	0.45303	0.19593	98.08072	5.35	0.0216
Wrist	-1.65666	0.52706	181.25142	9.88	0.0019

Bounds on condition number: 20.913, 668.17**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	Abdomen	1	0.6617	0.6617	72.2434	488.93	<.0001
2	Weight	2	0.0571	0.7188	20.1709	50.58	<.0001
3	Wrist	3	0.0089	0.7277	13.7069	8.15	0.0047
4	Forearm	4	0.0073	0.7350	8.8244	6.78	0.0098
5	Neck	5	0.0029	0.7379	8.0748	2.73	0.1000
6	Age	6	0.0027	0.7406	7.4937	2.58	0.1098
7	Thigh	7	0.0038	0.7445	5.8653	3.66	0.0569
8	Hip	8	0.0021	0.7466	5.8986	1.99	0.1594
9	Biceps	9	0.0012	0.7477	6.7834	1.13	0.2888
10	Ankle	10	0.0008	0.7485	8.0682	0.72	0.3957



The **FORWARD** final model is the same model as the best model using the **HOCKING** criterion plus Ankle (Abdomen, Weight, Wrist, Forearm, Neck, Age, Thigh, Hip, Biceps, and Ankle). The Criterion plot shows that the increase in adjusted R square is best for the model in Step 9. The increase is rather modest after about Step 4.

Backward Elimination: Step 6

Variable Hip Removed: R-Square = 0.7445 and C(p) = 5.8653

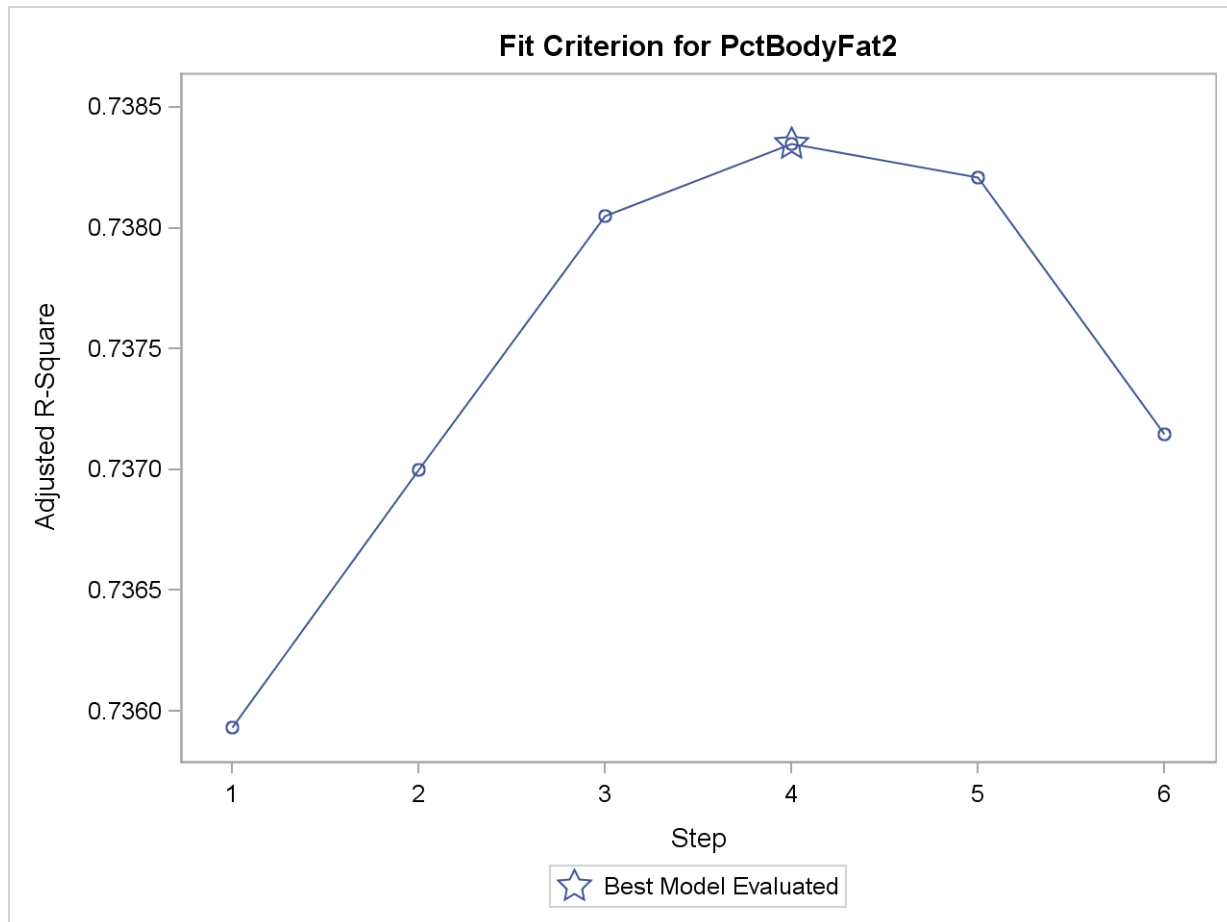
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	13087	1869.59160	101.56	<.0001
Error	244	4491.84861	18.40922		
Corrected Total	251	17579			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-33.25799	9.00681	251.00658	13.63	0.0003
Age	0.06817	0.03079	90.22018	4.90	0.0278
Weight	-0.11944	0.03403	226.84802	12.32	0.0005
Neck	-0.40380	0.22062	61.67131	3.35	0.0684
Abdomen	0.91788	0.06950	3211.14250	174.43	<.0001
Thigh	0.22196	0.11601	67.38659	3.66	0.0569
Forearm	0.55314	0.18479	164.95134	8.96	0.0030
Wrist	-1.53240	0.51041	165.93323	9.01	0.0030

Bounds on condition number: 13.634, 261.24

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	Knee	12	0.0000	0.7486	12.0032	0.00	0.9552
2	Chest	11	0.0000	0.7485	10.0313	0.03	0.8666
3	Height	10	0.0000	0.7485	8.0682	0.04	0.8473
4	Ankle	9	0.0008	0.7477	6.7834	0.72	0.3957
5	Biceps	8	0.0012	0.7466	5.8986	1.13	0.2888
6	Hip	7	0.0021	0.7445	5.8653	1.99	0.1594



The final model using the **BACKWARD** option is the same model as the one suggested by Mallows' criterion (Age, Weight, Neck, Abdomen, Thigh, Forearm, and Wrist).

The Criterion plot shows that the adjusted R square was best at Step 4. Be careful not to over-interpret this difference. The Y-axis only ranges from approximately 0.7360 to 0.7385. The differences are all minor.

Stepwise Selection: Step 7

Variable Thigh Entered: R-Square = 0.7445 and C(p) = 5.8653

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	13087	1869.59160	101.56	<.0001
Error	244	4491.84861	18.40922		
Corrected Total	251	17579			

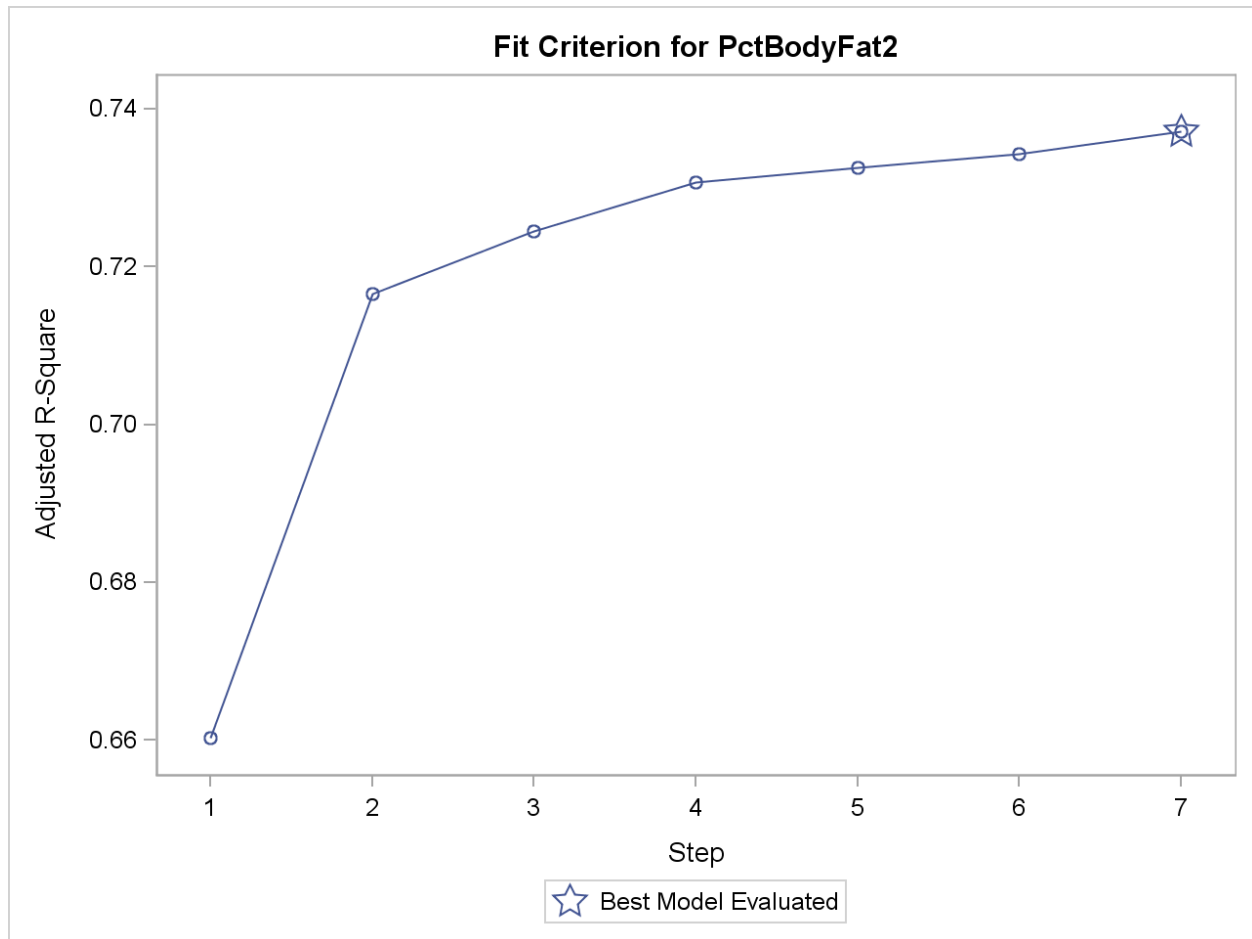
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-33.25799	9.00681	251.00658	13.63	0.0003
Age	0.06817	0.03079	90.22018	4.90	0.0278
Weight	-0.11944	0.03403	226.84802	12.32	0.0005
Neck	-0.40380	0.22062	61.67131	3.35	0.0684
Abdomen	0.91788	0.06950	3211.14250	174.43	<.0001
Thigh	0.22196	0.11601	67.38659	3.66	0.0569
Forearm	0.55314	0.18479	164.95134	8.96	0.0030
Wrist	-1.53240	0.51041	165.93323	9.01	0.0030

Bounds on condition number: 13.634, 261.24

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Abdomen		1	0.6617	0.6617	72.2434	488.93	<.0001
2	Weight		2	0.0571	0.7188	20.1709	50.58	<.0001
3	Wrist		3	0.0089	0.7277	13.7069	8.15	0.0047
4	Forearm		4	0.0073	0.7350	8.8244	6.78	0.0098
5	Neck		5	0.0029	0.7379	8.0748	2.73	0.1000
6	Age		6	0.0027	0.7406	7.4937	2.58	0.1098
7	Thigh		7	0.0038	0.7445	5.8653	3.66	0.0569



The model using the STEPWISE option results in the same model as that using the BACKWARD option (Age, Weight, Neck, Abdomen, Thigh, Forearm, and Wrist).

- c. How many variables would result from a model using FORWARD selection and a significance level for entry criterion of 0.05, instead of the default SLENTY of 0.50?

```
/*st103s04.sas*/ /*Part C*/
proc reg data=sasuser.BodyFat2 plots(only)=adjrsq;
  FORWARD05:model PctBodyFat2=Age Weight Height
    Neck Chest Abdomen Hip Thigh
    Knee Ankle Biceps Forearm Wrist
    / selection=forward slentry=0.05;
  title "Using Forward Stepwise with SLENTY=0.05";
run;
quit;
```

Partial Output

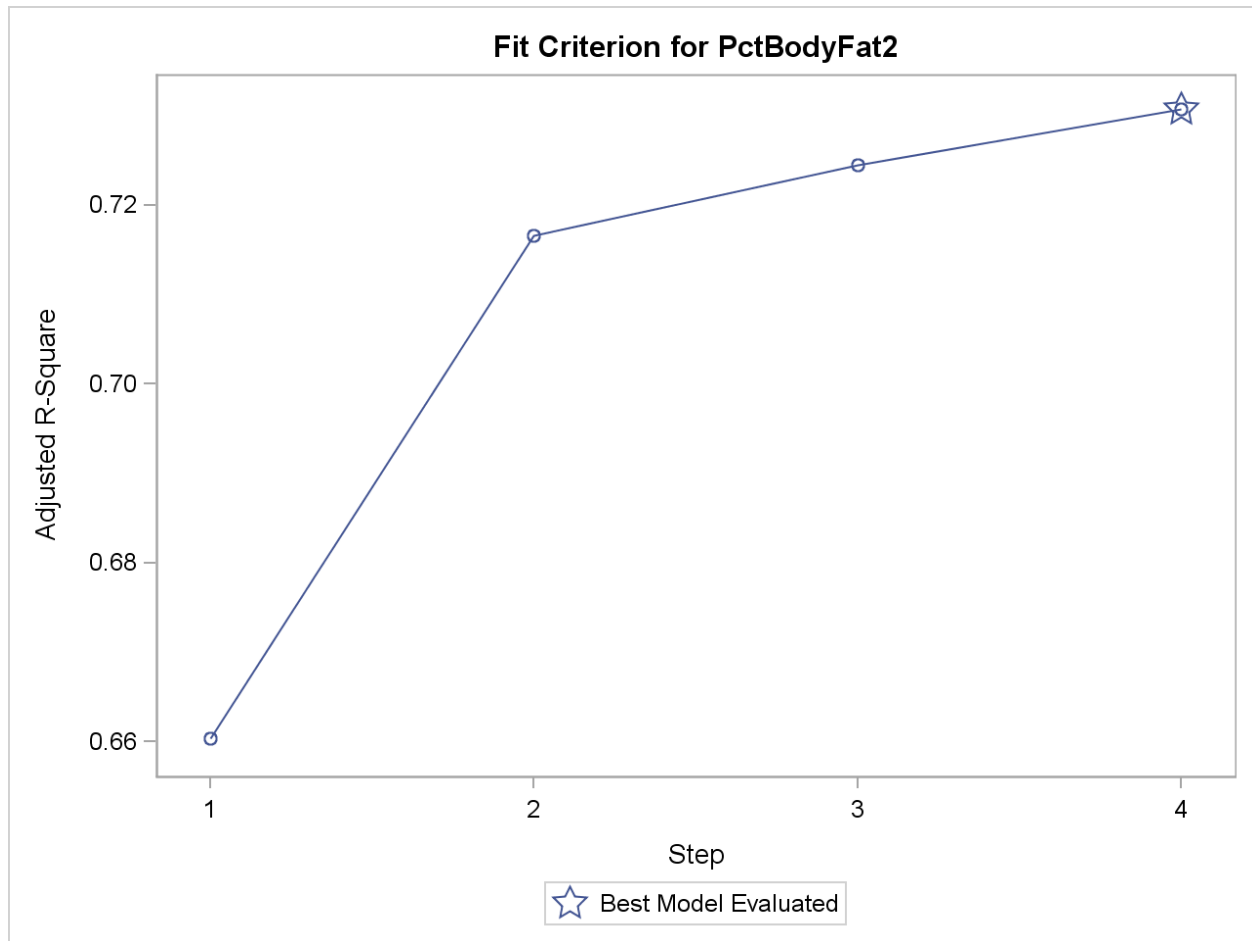
Forward Selection: Step 4**Variable Forearm Entered: R-Square = 0.7350 and C(p) = 8.8244**

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	12921	3230.18852	171.28	<.0001
Error	247	4658.23577	18.85925		
Corrected Total	251	17579			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-34.85407	7.24500	436.46987	23.14	<.0001
Weight	-0.13563	0.02475	566.43299	30.03	<.0001
Abdomen	0.99575	0.05607	5948.85562	315.43	<.0001
Forearm	0.47293	0.18166	127.81846	6.78	0.0098
Wrist	-1.50556	0.44267	218.15750	11.57	0.0008

Bounds on condition number: 7.0408, 63.886**No other variable met the 0.0500 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	Abdomen	1	0.6617	0.6617	72.2434	488.93	<.0001
2	Weight	2	0.0571	0.7188	20.1709	50.58	<.0001
3	Wrist	3	0.0089	0.7277	13.7069	8.15	0.0047
4	Forearm	4	0.0073	0.7350	8.8244	6.78	0.0098



The model using SLENTY=0.05 has substantially fewer (4) variables than the default SELECTION=FORWARD final model (Weight, Abdomen, Forearm, and Wrist).

The Criterion plot, showing adjusted R square at each step, is also produced.

Solutions to Student Activities (Polls/Quizzes)

3.01 Multiple Choice Poll – Correct Answer

The correlation between tuition and rate of graduation at U.S. colleges is 0.55. What does this mean?

- a. The way to increase graduation rates at your college is to raise tuition.
- b. Increasing graduation rates is expensive, causing tuition to rise.
- c. Students who are richer tend to graduate more often than poorer students.
- ☒ d. None of the above.

26

3.02 Multiple Choice Poll – Correct Answer

Run PROC REG with this MODEL statement:

`model y=x1;` If the parameter estimate (slope) of x1 is 0, then the best guess (predicted value) of y when x1=13 is which of the following?

- a. 13
- ☒ b. the mean of y
- c. a random number
- d. the mean of x1
- e. 0

42

3.03 Multiple Choice Poll – Correct Answer

What is the predicted value for **PctBodyFat2** when **Weight** is 150?

- a. 0.17439
- b. 150
- ☒ c. 14.1067

51

3.04 Multiple Choice Poll – Correct Answer

Which statistic in the ANOVA table is used to test the overall model hypotheses?

- ☒ a. F
- b. t
- c. R square
- d. Adjusted R square

61

3.05 Multiple Choice Poll – Correct Answer

When **Oxygen_Consumption** is regressed on **RunTime**, **Age**, **Run_Pulse**, and **Maximum_Pulse**, the parameter estimate for **Age** is -2.78. What does this mean?

- a. For each year older, the predicted value of oxygen consumption is 2.78 greater.
 - ☒ b. For each year older, the predicted value of oxygen consumption is 2.78 lower.
 - c. For every 2.78 years older, oxygen consumption doubles.
 - d. For every 2.78 years younger, oxygen consumption doubles.
- * Assume that the values of all other predictors are held constant.

74

3.06 Multiple Choice Poll – Correct Answer

Which value tends to increase (can never decrease) as you add predictor variables to your regression model?

- ☒ a. R square
- b. Adjusted R square
- c. Mallows' C_p
- d. Both a and b
- e. F statistic
- f. All of the above

88

3.07 Poll – Correct Answer

The STEPWISE, BACKWARD, and FORWARD strategies result in the same final model if the same significance levels are used in all three.

- ☐ True
- ☒ False