<table>
<thead>
<tr>
<th>Focus of this Chapter</th>
<th>Relationship to Previous Chapters</th>
<th>Relationship to Marketing Research Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Correlation</td>
<td>• Analytical Framework and Models (Chapter 2)</td>
<td>Problem Definition</td>
</tr>
<tr>
<td>• Regression</td>
<td>• Data Analysis Strategy (Chapter 15)</td>
<td>Approach to Problem</td>
</tr>
<tr>
<td></td>
<td>• General Procedure of Hypothesis Testing (Chapter 16)</td>
<td>Research Design</td>
</tr>
<tr>
<td></td>
<td>• Hypothesis Testing Related to Differences (Chapter 17)</td>
<td>Field Work</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Data Preparation and Analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Report Preparation and Presentation</td>
</tr>
</tbody>
</table>
Figure 18.2  Correlation and Regression: An Overview
Product Moment Correlation

- The *product moment correlation*, \( r \), summarizes the strength of association between two metric (interval or ratio scaled) variables, say \( X \) and \( Y \).

- It is an index used to determine whether a linear or straight-line relationship exists between \( X \) and \( Y \).

- As it was originally proposed by Karl Pearson, it is also known as the *Pearson correlation coefficient*. It is also referred to as *simple correlation*, *bivariate correlation*, or merely the *correlation coefficient*.
Division of the numerator and denominator by \((n-1)\) gives

\[
r = \frac{\sum_{i=1}^{n} \left( X_i - \bar{X} \right) \left( Y_i - \bar{Y} \right)}{\sqrt{\sum_{i=1}^{n} \left( X_i - \bar{X} \right)^2 \sum_{i=1}^{n} \left( Y_i - \bar{Y} \right)^2}}
\]

\[
r = \frac{\sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{n-1} \right) \left( \frac{Y_i - \bar{Y}}{n-1} \right)}{\sqrt{\sum_{i=1}^{n} \left( \frac{X_i - \bar{X}}{n-1} \right)^2 \sum_{i=1}^{n} \left( \frac{Y_i - \bar{Y}}{n-1} \right)^2}}
\]
Product Moment Correlation (Cont.)

- $r$ varies between -1.0 and +1.0.

- The correlation coefficient between two variables will be the same regardless of their underlying units of measurement.
**Table 18.1**  
Explaining Attitudes Towards Sports Cars

<table>
<thead>
<tr>
<th>Respondent No.</th>
<th>Attitude Toward Sports Cars</th>
<th>Duration of Sports Car Ownership</th>
<th>Importance Attached to Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>10</td>
<td>12</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>5</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>9</td>
<td>11</td>
<td>18</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>2</td>
<td>5</td>
</tr>
</tbody>
</table>

Copyright © 2011 Pearson Education, Inc.
Figure 18.3
Plot of Attitude with Duration

Duration of Car Ownership

Attitude
Product Moment Correlation

\[ \bar{X} = \frac{(10 + 12 + 12 + 4 + 12 + 6 + 8 + 2 + 18 + 9 + 17 + 2)}{12} = 9.333 \]

\[ \bar{Y} = \frac{(6 + 9 + 8 + 3 + 10 + 4 + 5 + 2 + 11 + 9 + 10 + 2)}{12} = 6.583 \]

\[ \sum_{i=1}^{n} (X_i - \bar{X})^2 = (10-9.33)^2 + (12-9.33)^2 + (12-9.33)^2 + (4-9.33)^2 \\
+ (12-9.33)^2 + (6-9.33)^2 + (8-9.33)^2 + (2-9.33)^2 \\
+ (18-9.33)^2 + (9-9.33)^2 + (17-9.33)^2 + (2-9.33)^2 \\
= 0.4489 + 7.1289 + 7.1289 + 28.4089 \\
+ 7.1289 + 11.0889 + 1.7689 + 53.7289 \\
+ 75.1689 + 0.1089 + 58.8289 + 53.7289 \\
= 304.6668 \]

\[ \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = (6-6.58)^2 + (9-6.58)^2 + (8-6.58)^2 + (3-6.58)^2 \\
+ (10-6.58)^2 + (4-6.58)^2 + (5-6.58)^2 + (2-6.58)^2 \\
+ (11-6.58)^2 + (9-6.58)^2 + (10-6.58)^2 + (2-6.58)^2 \\
= 0.3364 + 5.8564 + 2.0164 + 12.8164 \\
+ 11.6964 + 6.6564 + 2.4964 + 20.9764 \\
+ 19.5364 + 5.8564 + 11.6964 + 20.9764 \\
= 120.9168 \]

Thus, \( r = \frac{179.6668}{\sqrt{(304.6668)(120.9168)}} = 0.9361 \)
Decomposition of Total Variation

\[ r^2 = \frac{\text{Explained variation}}{\text{Total variation}} \]

\[ = \frac{SS_x}{SS_y} \]

\[ = \frac{\text{Total variation - Error variation}}{\text{Total variation}} \]

\[ = \frac{SS_y - SS_{error}}{SS_y} \]
Table 18.2
Calculation of the Product Moment Correlation

<table>
<thead>
<tr>
<th>Number</th>
<th>Attitude (Y)</th>
<th>Duration (X)</th>
<th>$X_i - \bar{X}$</th>
<th>$(X_i - \bar{X})^2$</th>
<th>$(Y_i - \bar{Y})$</th>
<th>$(Y_i - \bar{Y})^2$</th>
<th>$(X_i - \bar{X})(Y_i - \bar{Y})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>6</td>
<td>10</td>
<td>0.667</td>
<td>0.4489</td>
<td>-0.583</td>
<td>0.3364</td>
<td>-0.3886</td>
</tr>
<tr>
<td>2.</td>
<td>9</td>
<td>12</td>
<td>2.667</td>
<td>7.1289</td>
<td>2.417</td>
<td>5.8564</td>
<td>6.4614</td>
</tr>
<tr>
<td>3.</td>
<td>8</td>
<td>12</td>
<td>2.667</td>
<td>7.1289</td>
<td>1.417</td>
<td>2.0164</td>
<td>3.7914</td>
</tr>
<tr>
<td>4.</td>
<td>3</td>
<td>4</td>
<td>-5.333</td>
<td>28.4089</td>
<td>-3.583</td>
<td>12.8164</td>
<td>19.0814</td>
</tr>
<tr>
<td>5.</td>
<td>10</td>
<td>12</td>
<td>2.667</td>
<td>7.1289</td>
<td>3.417</td>
<td>11.6964</td>
<td>9.1314</td>
</tr>
<tr>
<td>6.</td>
<td>4</td>
<td>6</td>
<td>-3.333</td>
<td>11.0889</td>
<td>-2.583</td>
<td>6.5664</td>
<td>8.5914</td>
</tr>
<tr>
<td>7.</td>
<td>5</td>
<td>8</td>
<td>-1.333</td>
<td>1.7689</td>
<td>-1.583</td>
<td>2.4964</td>
<td>2.1014</td>
</tr>
<tr>
<td>8.</td>
<td>2</td>
<td>2</td>
<td>-7.333</td>
<td>53.7289</td>
<td>-4.583</td>
<td>20.9764</td>
<td>3.5714</td>
</tr>
<tr>
<td>9.</td>
<td>11</td>
<td>18</td>
<td>8.667</td>
<td>75.1689</td>
<td>7.417</td>
<td>19.5364</td>
<td>38.3214</td>
</tr>
<tr>
<td>10.</td>
<td>9</td>
<td>9</td>
<td>-10.333</td>
<td>0.1089</td>
<td>2.417</td>
<td>5.8564</td>
<td>-0.7986</td>
</tr>
<tr>
<td>11.</td>
<td>10</td>
<td>17</td>
<td>7.667</td>
<td>58.8289</td>
<td>3.417</td>
<td>11.6964</td>
<td>26.2314</td>
</tr>
<tr>
<td>12.</td>
<td>2</td>
<td>2</td>
<td>-7.333</td>
<td>53.7289</td>
<td>-4.583</td>
<td>20.9764</td>
<td>33.5714</td>
</tr>
<tr>
<td>Mean</td>
<td>6.583</td>
<td>9.333</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td>304.668</td>
<td>120.9168</td>
<td>179.6668</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Copyright © 2011 Pearson Education, Inc. Chapter 18 - 13
Calculation of $r$

$$r = \frac{179.6668}{\sqrt{(304.6668)(120.9168)}}$$

$$= \frac{179.6668}{(17.4547)(10.9962)}$$

$$= 0.9361$$
When it is computed for a population rather than a sample, the product moment correlation is denoted by $\rho$, the Greek letter rho. The coefficient $r$ is an estimator of $\rho$.

The statistical significance of the relationship between two variables measured by using $r$ can be conveniently tested. The hypotheses are:

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$
The test statistic is:

\[ t = r \left[ \frac{n-2}{1 - r^2} \right]^{1/2} \]

which has a \( t \) distribution with \( n - 2 \) degrees of freedom. For the correlation coefficient calculated based on the data given in Table 13.1,

\[ t = 0.9361 \left[ \frac{12-2}{1 - (0.9361)^2} \right]^{1/2} \]

\[ = 8.414 \]

and the degrees of freedom = 12 - 2 = 10. From the \( t \) distribution table (Table 4 in the Statistical Appendix), the critical value of \( t \) for a two-tailed test and \( \alpha = 0.05 \) is 2.228. Hence, the null hypothesis of no relationship between \( X \) and \( Y \) is rejected.
Figure 18.4
A Nonlinear Relationship for which $r = 0$
Regression Analysis

Regression analysis is used in the following ways:

- Determine whether the independent variables explain a significant variation in the dependent variable: *whether a relationship exists*.
- Determine how much of the variation in the dependent variable can be explained by the independent variables: *strength of the relationship*.
- Determine the structure or form of the relationship: *the mathematical equation relating the independent and dependent variables*.
- Predict the values of the dependent variable.
Regression Analysis (Cont.)

Regression analysis is used in the following ways: (Cont.)

- Control for other independent variables when evaluating the contributions of a specific variable or set of variables.
- Regression analysis is concerned with the nature and degree of association between variables and does not imply or assume any causality.
**Statistics: Bivariate Regression Analysis**

- **Bivariate regression model.** The basic regression equation is \( Y_i = \beta_0 + \beta_1 X_i + e_i \), where \( Y \) = dependent or criterion variable, \( X \) = independent or predictor variable, \( \beta_0 \) = intercept of the line, \( \beta_1 \) = slope of the line, and \( e_i \) is the error term associated with the \( i \) th observation.

- **Coefficient of determination.** The strength of association is measured by the coefficient of determination, \( r^2 \). It varies between 0 and 1 and signifies the proportion of the total variation in \( Y \) that is accounted for by the variation in \( X \).

- **Estimated or predicted value.** The estimated or predicted value of \( Y \) is \( \hat{Y}_i = a + b x \), where \( \hat{Y}_i \) is the predicted value of \( Y_i \), and \( a \) and \( b \) are estimators of \( \beta_0 \) and \( \beta_1 \), respectively.
Regression coefficient. The estimated parameter b is usually referred to as the non-standardized regression coefficient.

Scattergram. A scatter diagram, or scattergram, is a plot of the values of two variables for all the cases or observations.

Standard error of estimate. This statistic, SEE, is the standard deviation of the actual Y values from the predicted \( \hat{Y} \) values.

Standard error. The standard deviation of b, SE\( b \), is called the standard error.
Statistics: Bivariate Regression Analysis (Cont.)

- **Standardized regression coefficient.** Also termed the beta coefficient or beta weight, this is the slope obtained by the regression of \( Y \) on \( X \) when the data are standardized.

- **Sum of squared errors.** The distances of all the points from the regression line are squared and added together to arrive at the sum of squared errors, which is a measure of total error, \( \sum e_j^2 \).

- **t statistic.** A t statistic with \( n - 2 \) degrees of freedom can be used to test the null hypothesis that no linear relationship exists between \( X \) and \( Y \), or \( H_0: \beta_1 = 0 \), where

\[
t = \frac{b}{SE_b}
\]
Conducting Bivariate Regression Analysis: Plot the Scatter Diagram

- **A scatter diagram**, or **scattergram**, is a plot of the values of two variables for all the cases or observations.
- The most commonly used technique for fitting a straight line to a scattergram is the **least-squares procedure**.
- In fitting the line, the least-squares procedure minimizes the sum of squared errors, $\sum e_j^2$. 
Figure 18.5
Conducting Bivariate Regression Analysis

- Plot the Scatter Diagram
- Formulate the General Model
- Estimate the Parameters
- Estimate Standardized Regression Coefficients
- Test for Significance
- Determine the Strength and Significance of Association
- Check Prediction Accuracy
- Examine the Residuals
- Refine the Model
Formulate the Bivariate Regression Model

In the bivariate regression model, the general form of a straight line is:

\[ Y = \beta_0 + \beta_1 X \]

where

- \( Y \) = dependent or criterion variable
- \( X \) = independent or predictor variable
- \( \beta_0 \) = intercept of the line
- \( \beta_1 \) = slope of the line

The regression procedure adds an error term to account for the probabilistic or stochastic nature of the relationship:

\[ Y_i = \beta_0 + \beta_1 X + e_i \]
Figure 18.6 Bivariate Regression
Estimate the Parameters

In most cases, $\beta_0$ and $\beta_1$ are unknown and are estimated from the sample using the equation:

$$\hat{Y}_i = a + b X_i$$

where, $\hat{Y}_i$ is the estimated or predicted value of $Y_i$, and $a$ and $b$ are estimators of $\beta_0$ and $\beta_1$, respectively.

$$b = \frac{\text{COV}_{xy}}{S_x^2}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

$$= \frac{\sum_{i=1}^{n}X_iY_i - n\bar{X}\bar{Y}}{\sum_{i=1}^{n}X_i^2 - n\bar{X}^2}$$
Estimate the Parameters (Cont.)

The intercept, \( a \), may then be calculated using:

\[
a = \bar{Y} - b\bar{X}
\]

For the data in Table 18.1, the estimation of parameters may be estimated as follows:

\[
\sum_{i=1}^{12} X_i Y_j = (10) (6) + (12) (9) + (12) (8) + (4) (3) + (12) (10) + (6) (4)
\]
\[
+ (8) (5) + (2) (2) + (18) (11) + (9) (9) + (17) (10) + (2) (2)
\]
\[
= 917
\]

\[
\sum_{i=1}^{12} X_i^2 = 102 + 122 + 122 + 42 + 122 + 62
\]
\[
+ 82 + 22 + 182 + 92 + 172 + 22
\]
\[
= 1350
\]
Estimate the Parameters (Cont.)

It may be recalled from earlier calculations of the simple correlation that

\[ \bar{X} = 9.333 \]
\[ \bar{Y} = 6.583 \]

Given \( n=12 \), \( b \) can be calculated as

\[
b = \frac{917 - (12)(9.333)(6.583)}{1350 - (12)(9.333)^2} = 0.5897
\]

\[
a = \bar{Y} - b\bar{X}
= 6.583 - (0.5897)(9.333)
= 1.0793
\]
Standardization is the process by which the raw data are transformed into new variables that have a mean of 0 and a variance of 1 (Chapter 13).

When the data are standardized, the intercept assumes a value of 0.

The term beta coefficient or beta weight is used to denote the standardized regression coefficient.

\[ B_{yx} = B_{xy} = r_{xy} \]

There is a simple relationship between the standardized and non-standardized regression coefficients:

\[ B_{yx} = b_{yx} \left( \frac{S_x}{S_y} \right) \]
Test for Significance

The statistical significance of the linear relationship between \( X \) and \( Y \) may be tested by examining the hypotheses:

\[
H_0 : \beta_1 = 0 \\
H_1 : \beta_1 \neq 0
\]

A \( t \) statistic with \( n - 2 \) degrees of freedom can be used, where

\[
t = \frac{b}{SE_b}
\]

\( SE_b \) denotes the standard deviation of \( b \) and is called the **standard error**.
Test for Significance

Using a computer program, the regression of attitude on duration of residence, using the data shown in Table 18.1, yielded the results shown in Table 18.2. The intercept, \( a \), equals 1.0793, and the slope, \( b \), equals 0.5897. Therefore, the estimated equation is:

\[
\text{Attitude(\hat{y})} = 1.0793 + 0.5897 \times \text{(Duration of Car Ownership)}
\]

The standard error, or standard deviation of \( b \) is estimated as 0.07008, and the value of the \( t \) statistic as \( t = \frac{0.5897}{0.0700} = 8.414 \), with \( n - 2 = 10 \) degrees of freedom.

From Table 4 in the Statistical Appendix, we see that the critical value of \( t \) with 10 degrees of freedom and \( \alpha = 0.05 \) is 2.228 for a two-tailed test. Since the calculated value of \( t \) is larger than the critical value, the null hypothesis is rejected.
Test for Significance

Using a computer program, the regression of attitude on duration of residence, using the data shown in Table 18.1, yielded the results shown in Table 18.2. The intercept, $a$, equals 1.0793, and the slope, $b$, equals 0.5897. Therefore, the estimated equation is:

$$\text{Attitude}(Y) = 1.0793 + 0.5897 \ (\text{Duration of Car Ownership})$$

The standard error, or standard deviation of $b$ is estimated as 0.07008, and the value of the $t$ statistic as $t = 0.5897/0.0700 = 8.414$, with $n - 2 = 10$ degrees of freedom.

From Table 4 in the Statistical Appendix, we see that the critical value of $t$ with 10 degrees of freedom and $\alpha = 0.05$ is 2.228 for a two-tailed test. Since the calculated value of $t$ is larger than the critical value, the null hypothesis is rejected.
Determine Strength and Significance of Association

The total variation, $SS_y$, may be decomposed into the variation accounted for by the regression line, $SS_{reg}$, and the error or residual variation, $SS_{error}$ or $SS_{res}$, as follows:

$$SS_y = SS_{reg} + SS_{res}$$

where

$$SS_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

$$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

$$SS_{res} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$
Figure 18.7
Decomposition of the Total Variation In Bivariate Regression

\[ Y \]

\[ \bar{Y} \]

Total variation, \( SS_Y \)

Residual variation, \( SS_{RES} \)

Explained variation, \( SS_{REG} \)

\[ X \]
To illustrate the calculations of $r^2$, let us consider again the effect of attitude toward the city on the duration of residence. It may be recalled from earlier calculations of the simple correlation coefficient that:

$$SS_y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 120.9168$$
Determine Strength and Significance of Association (Cont.)

The predicted values ($\hat{Y}$) can be calculated using the regression equation:

$$\text{Attitude (}$\hat{Y}$\text{)} = 1.0793 + 0.5897 \text{ (Duration of Car Ownership)}$$

For the first observation in Table 18.1, this value is:

$$\hat{Y} = 1.0793 + 0.5897 \times 10 = 6.9763.$$  

For each successive observation, the predicted values are, in order, 8.1557, 8.1557, 3.4381, 8.1557, 4.6175, 5.7969, 2.2587, 11.6939, 6.3866, 11.1042, and 2.2587.
Determine Strength and Significance of Association (Cont.)

Therefore,

\[ SS_{reg} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \]

\[ = (6.9763 - 6.5833)^2 + (8.1557 - 6.5833)^2 \]
\[ + (8.1557 - 6.5833)^2 + (3.4381 - 6.5833)^2 \]
\[ + (8.1557 - 6.5833)^2 + (4.6175 - 6.5833)^2 \]
\[ + (5.7969 - 6.5833)^2 + (2.2587 - 6.5833)^2 \]
\[ + (11.6939 - 6.5833)^2 + (6.3866 - 6.5833)^2 \]
\[ + (11.1042 - 6.5833)^2 + (2.2587 - 6.5833)^2 \]

\[ = 0.1544 + 2.4724 + 2.4724 + 9.8922 \]
\[ + 2.4724 + 3.8643 + 0.6184 + 18.7021 \]
\[ + 26.1182 + 0.0387 + 20.4385 + 18.7021 \]

\[ = 105.9524 \]
Determine Strength and Significance of Association (Cont.)

\[ SS_{res} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = (6-6.9763)^2 + (9-8.1557)^2 + (8-8.1557)^2 \]
\[ + (3-3.4381)^2 + (10-8.1557)^2 + (4-4.6175)^2 \]
\[ + (5-5.7969)^2 + (2-2.2587)^2 + (11-11.6939)^2 \]
\[ + (9-6.3866)^2 + (10-11.1042)^2 + (2-2.2587)^2 \]
\[ = 14.9644 \]

It can be seen that \( SS_y = SS_{reg} + SS_{res} \). Furthermore,

\[ r^2 = \frac{SS_{reg}}{SS_y} \]
\[ = \frac{105.9524}{120.9168} \]
\[ = 0.8762 \]
Determine Strength and Significance of Association (Cont.)

Another equivalent test for examining the significance of the linear relationship between $X$ and $Y$ (significance of $b$) is the test for the significance of the coefficient of determination. The hypotheses in this case are:

$$H_0: \quad R^2_{pop} = 0$$

$$H_1: \quad R^2_{pop} > 0$$
Determine Strength and Significance of Association (Cont.)

The appropriate test statistic is the $F$ statistic, 
\[ F = \frac{SS_{\text{reg}}}{SS_{\text{res}}/(n-2)} \]

which has an $F$ distribution with 1 and $n - 2$ degrees of freedom. The $F$ test is a generalized form of the $t$ test (see Chapter 17). If a random variable is $t$ distributed with $n$ degrees of freedom, then $t^2$ is $F$ distributed with 1 and $n$ degrees of freedom. Hence, the $F$ test for testing the significance of the coefficient of determination is equivalent to testing the following hypotheses:

\[
\begin{align*}
H_0: \beta_1 &= 0 & \text{or} & \quad H_0: \rho &= 0 \\
H_0: \beta_1 &\neq 0 & H_0: \rho &\neq 0
\end{align*}
\]
From Table 18.2, it can be seen that:

\[ r^2 = \frac{105.9522}{105.9522 + 14.9644} = 0.8762 \]

which is the same as the value calculated earlier.

The value of the \( F \) statistic is:

\[ F = \frac{105.9522}{14.9644/10} = 70.8027 \quad \text{with 1 and 10 degrees of freedom.} \]

The calculated \( F \) statistic exceeds the critical value of 4.96 determined from Table 5 in the Statistical Appendix. Therefore, the relationship is significant, corroborating the results of the \( t \) test.
### Table 18.3 Bivariate Regression

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple $R$</td>
<td>.9361</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.8762</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>.8639</td>
</tr>
<tr>
<td>Standard Error</td>
<td>1.2233</td>
</tr>
</tbody>
</table>

#### Analysis of Variance

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
<td>Sum of Squares</td>
<td>Mean Square</td>
</tr>
<tr>
<td>Regression</td>
<td>1</td>
<td>105.9522</td>
</tr>
<tr>
<td>Residual</td>
<td>10</td>
<td>14.9644</td>
</tr>
</tbody>
</table>

$F = 70.8027$  
Significance of $F = .0000$

#### VARIABLES IN THE EQUATION

<table>
<thead>
<tr>
<th>Variable</th>
<th>$b$</th>
<th>SE $b$</th>
<th>Beta ($B$)</th>
<th>$T$</th>
<th>Sig. $T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duration</td>
<td>.5897</td>
<td>.0700</td>
<td>.9361</td>
<td>8.414</td>
<td>.0000</td>
</tr>
<tr>
<td>(Constant)</td>
<td>1.0793</td>
<td>.7434</td>
<td>1.452</td>
<td>.1772</td>
<td></td>
</tr>
</tbody>
</table>
Check Prediction Accuracy

To estimate the accuracy of predicted values, $\hat{Y}$, it is useful to calculate the standard error of estimate, $SEE$.

$$SEE = \sqrt{\frac{SS_{res}}{n-2}}$$

From Table 18.2,

$$SEE = \sqrt{14.9644/(12-2)} = 1.2233$$
Assumptions

- The error term is normally distributed. For each fixed value of $X$, the distribution of $Y$ is normal.
- The means of all these normal distributions of $Y$, given $X$, lie on a straight line with slope $b$.
- The mean of the error term is 0.
- The variance of the error term is constant. This variance does not depend on the values assumed by $X$.
- The error terms are uncorrelated. In other words, the observations have been drawn independently.
Multiple Regression

The general form of the multiple regression model is as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_k X_k + e$$

which is estimated by the following equation:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \cdots + b_k X_k$$

As before, the coefficient $a$ represents the intercept, but the $b$'s are now the partial regression coefficients.
Statistics Associated with Multiple Regression (Cont.)

- **Adjusted \( R^2 \).** \( R^2 \), coefficient of multiple determination, is adjusted for the number of independent variables and the sample size to account for the diminishing returns. After the first few variables, the additional independent variables do not make much contribution.

- **Coefficient of multiple determination.** The strength of association in multiple regression is measured by the square of the multiple correlation coefficient, \( R^2 \), which is also called the coefficient of multiple determination.
Statistics Associated with Multiple Regression (Cont.)

- **F test.** The F test is used to test the null hypothesis that the coefficient of multiple determination in the population, $R^2_{pop}$, is zero. The test statistic has an F distribution with k and $(n - k - 1)$ degrees of freedom.

- **Partial F test.** The significance of a partial regression coefficient, $\beta_i$, of $X_i$ may be tested using an incremental F statistic. The incremental F statistic is based on the increment in the explained sum of squares resulting from the addition of the independent variable $X_i$ to the regression equation after all the other independent variables have been included.
Statistics Associated with Multiple Regression (Cont.)

- **Partial regression coefficient.** The partial regression coefficient, $b_1$, denotes the change in the predicted value, $Y$, per unit change in $X_1$ when the other independent variables, $X_2$ to $X_k$, are held constant.
Partial Regression Coefficients

\[ \hat{Y} = a + b_1 X_1 + b_2 X_2 \]

First, note that the relative magnitude of the partial regression coefficient of an independent variable is, in general, different from that of its bivariate regression coefficient.

The interpretation of the partial regression coefficient, \( b_1 \), is that it represents the expected change in \( Y \) when \( X_1 \) is changed by one unit but \( X_2 \) is held constant or otherwise controlled. Likewise, \( b_2 \) represents the expected change in \( Y \) for a unit change in \( X_2 \), when \( X_1 \) is held constant. Thus, calling \( b_1 \) and \( b_2 \) partial regression coefficients is appropriate.
Partial Regression Coefficients (Cont.)

- It can also be seen that the combined effects of $X_1$ and $X_2$ on $Y$ are additive. In other words, if $X_1$ and $X_2$ are each changed by one unit, the expected change in $Y$ would be $(b_1+b_2)$.

- Extension to the case of $k$ variables is straightforward. The partial regression coefficient, $b_1$, represents the expected change in $Y$ when $X_1$ is changed by one unit and $X_2$ through $X_k$ are held constant. It can also be interpreted as the bivariate regression coefficient, $b$, for the regression of $Y$ on the residuals of $X_1$, when the effect of $X_2$ through $X_k$ has been removed from $X_1$. 
Partial Regression Coefficients (Cont.)

The relationship of the standardized to the non-standardized coefficients remains the same as before:

\[ B_1 = b_1 \left( \frac{S_{x_1}}{S_y} \right) \]

\[ B_k = b_k \left( \frac{S_{x_k}}{S_y} \right) \]

The estimated regression equation is:

\[ \hat{Y} = 0.33732 + 0.48108 \, X_1 + 0.28865 \, X_2 \]

or

Attitude
\[ = 0.33732 + 0.48108 \, \text{(Duration)} + 0.28865 \, \text{(Importance)} \]
Table 18.4 Multivariate Regression

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple $R$</td>
<td>.9721</td>
</tr>
<tr>
<td>$R^2$</td>
<td>.9450</td>
</tr>
<tr>
<td>Adjusted $R^2$</td>
<td>.9330</td>
</tr>
<tr>
<td>Standard Error</td>
<td>.8597</td>
</tr>
</tbody>
</table>

Analysis of Variance

<table>
<thead>
<tr>
<th></th>
<th>df</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>2</td>
<td>114.2643</td>
<td>57.1321</td>
</tr>
<tr>
<td>Residual</td>
<td>9</td>
<td>6.6524</td>
<td>.7392</td>
</tr>
</tbody>
</table>

$F = 77.2936$  
Significance of $F = .0000$

VARIABLES IN THE EQUATION

<table>
<thead>
<tr>
<th>Variable</th>
<th>$b$</th>
<th>SE $b$</th>
<th>Beta ($B$)</th>
<th>$T$</th>
<th>Sig. $T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Importance</td>
<td>.2887</td>
<td>.08608</td>
<td>.3138</td>
<td>3.353</td>
<td>.0085</td>
</tr>
<tr>
<td>Duration</td>
<td>.4811</td>
<td>.05895</td>
<td>.7636</td>
<td>8.160</td>
<td>.0000</td>
</tr>
<tr>
<td>(Constant)</td>
<td>.3373</td>
<td>.56736</td>
<td>.595</td>
<td>.595</td>
<td>.5668</td>
</tr>
</tbody>
</table>
Examination of Residuals

- A **residual** is the difference between the observed value of $Y_i$ and the value predicted by the estimated regression equation, $\hat{Y}_i$.

- A plot of residuals against time, or the sequence of observations, will throw some light on the assumption that the error terms are uncorrelated.
Examination of Residuals (Cont.)

- Examine the histogram of standardized residuals. Compare the frequency of residuals to the normal distribution and if the difference is small then the normality assumption may be reasonably met.

- Examine the normal probability plot of standardized residuals. The normal probability plot shows the observed standardized residuals compared to expected standardized residuals from a normal distribution. If the observed residuals are normally distributed, they will fall on the 45 degree line.
Examination of Residuals (Cont.)

- Examine the plot of standardized residuals versus standardized predicted values. This plot should be random with no discernible pattern. This will provide an indication on the assumptions of linearity and constant variance for the error term.

- Look at the table of residual statistics and identify any standardized predicted values or standardized residuals that are more than plus or minus 3 standard deviations. Values larger than this may indicate the presence of outliers in the data.
Examination of Residuals (Cont.)

- Plotting the residuals against the independent variables provides evidence of the appropriateness. Again, the plot should result in a random pattern.

- To examine whether any additional variables should be included in the regression equation, one could run a regression of the residuals on the proposed variables.
Figure 18.8 A Concept Map for Product Moment Correlation

Metric (interval- or ratio-scaled)

\[ \text{Strength of Linear Association measures} \]

\[ r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}} \]

\[ \text{Pearson Correlation, Simple Correlation, Bivariate Correlation, or the Correlation Coefficient} \]

\[ r^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} = \frac{SS_x}{SS_y} = \frac{SS_y - SS_{\text{error}}}{SS_y} \]

\[ \text{It's square varies between} \ -1.0 \text{ and } +1.0 \]

\[ \text{Independent of Units of Measurement} \]
Figure 18.9 A Concept Map for Conducting Bivariate Regression

- **Y** on the vertical axis, **X** on the horizontal axis
  - Scatter Diagram
  - indicates
  - Patterns or problems in the data
  - has the form
    - General Model
    - assumes
    - Error terms are independent and normally distributed with zero mean and a constant variance
    - $Y_i = \beta_0 + \beta_1 X + \epsilon$
    - $b = \frac{COV_{xy}}{S_x^2}$
    - $a = \bar{Y} - b \bar{X}$
  - Regression Parameters
  - calculate
  - $B_{yx} = b_{yx} \frac{S_x}{S_y}$
  - $t = \frac{b}{SE_b}$
  - $F = \frac{SS_{reg}}{SS_{res}/(n-2)}$
  - $r^2 = \frac{SS_{reg}}{SS_y} = \frac{SS_y - SS_{res}}{SS_y}$
  - Strength and Significance of Association
  - assessed by
  - Conducting Significance Testing
  - Significance Testing hypotheses are if significant, examine
  - $H_0: \beta_1 = 0$
  - $H_1: \beta_1 \neq 0$
  - Significance Testing
    - use F test
    - use t-test
    - alternatively use F test
    - $t = \frac{b}{SE_b}$
  - Prediction Accuracy
    - calculate
    - Standard Error of Estimate (SEE)
    - test assumptions by
    - Examination of Residuals
      - examine
        - Table of Residual Statistics
        - Standardized Residuals
          - compare to normal distribution
          - plot of
            - Normal Probability Plot
            - Standardized Residuals Versus Standardized Predicted Values
        - Histogram
          - compare to expected standardized residuals

SPSS Windows

The CORRELATE program computes Pearson product moment correlations and partial correlations with significance levels. Univariate statistics, covariance, and cross-product deviations may also be requested.

To select these procedures using SPSS for Windows, click:

Analyze > Correlate > Bivariate …

Scatterplots can be obtained by clicking:

Graphs > Scatter … > Simple Scatter > Define
REGRESSION calculates bivariate and multiple regression equations, associated statistics, and plots. It allows for an easy examination of residuals. This procedure can be run by clicking:

**Analyze > Regression > Linear …**

The detailed steps are illustrated using the data of Table 18.1.
SPSS Detailed Steps: Correlation

1. Select ANALYZE from the SPSS menu bar.
2. Click CORRELATE and then BIVARIATE.
3. Move "Attitude Towards Sports Cars” and “Duration of Car Ownership" into the VARIABLES box.
4. Check PEARSON under CORRELATION COEFFICIENTS.
5. Check ONE-TAILED under TEST OF SIGNIFICANCE.
6. Check FLAG SIGNIFICANT CORRELATIONS.
7. Click OK.
SPSS Detailed Steps: Bivariate Regression

1. Select ANALYZE from the SPSS menu bar.
2. Click REGRESSION and then LINEAR.
4. Move "Duration of Car Ownership" into the INDEPENDENT(S) box.
5. Select ENTER in the METHOD box (default option).
6. Click STATISTICS and check ESTIMATES under REGRESSION COEFFICIENTS.

7. Check MODEL FIT.

8. Click CONTINUE.

9. Click PLOTS.

10. In the LINEAR REGRESSION:PLOTS box, move *ZRESID into the Y: box and *ZPRED into the X: box.
11. Check HISTOGRAM and NORMAL PROBABILITY PLOT in the STANDARDIZED RESIDUALS PLOTS.

12. Click CONTINUE.

13. Click OK.

The steps for running multiple regression are similar, except for Step 4. In Step 4, move "Duration of Car Ownership and "Importance Attached to Performance" into the INDEPENDENT(S) box.
Excel

**Correlations** can be determined in EXCEL by using the DATA > DATA ANALYSIS > CORRELATION function. Use the Correlation Worksheet Function when a correlation coefficient for two cell ranges is needed.

**Regression** can be accessed from the DATA > DATA ANALYSIS menu. Depending on the features selected, the output can consist of a summary output table, including an ANOVA table, a standard error of y estimate, coefficients, standard error of coefficients, $R^2$ and adjusted $R^2$ values, and the number of observations.
In addition, the function computes a residual output table, a residual plot, a line fit plot, normal probability plot, and a two-column probability data output table. The detailed steps are illustrated using the data of Table 18.1.
Excel Detailed Steps: Correlation

1. Select DATA tab.
2. In the ANALYSIS tab, select DATA ANALYSIS.
3. The DATA ANALYSIS Window pops up.
4. Select CORRELATION from the DATA ANALYSIS Window.
5. Click OK.
6. The CORRELATION pop-up window appears on screen.
7. The CORRELATION window has two portions
   1. INPUT
   2. OUTPUT OPTIONS
Excel Detailed Steps: Correlation (Cont.)

8. The INPUT portion asks for the following inputs:
   a. Click in the INPUT RANGE box. Select (highlight) all the rows of data under ATTITUDE and DURATION.
      $B$2:$C$13 should appear in INPUT RANGE.
   b. Select COLUMNS beside GROUPED BY.
   c. Leave LABELS IN FIRST ROW as blank.

9. In the OUTPUT OPTIONS window, select NEW WORKBOOK Options.

10. Click OK.
Excel Detailed Steps: Bivariate Regression

1. Select DATA tab.
2. In the ANALYSIS tab, select DATA ANALYSIS.
3. DATA ANALYSIS Window pops up.
4. Select REGRESSION from the DATA ANALYSIS Window.
5. Click OK.
6. The REGRESSION pop-up window appears on screen.
Excel Detailed Steps: Bivariate Regression (Cont.)

7. The REGRESSION window has four portions:
   a. INPUT
   b. OUTPUT OPTIONS
   c. RESIDUALS
   d. NORMAL PROBABILITY

8. The INPUT portion asks for the following inputs:
   a. Click in the INPUT Y RANGE box. Select (highlight) all the rows of data under ATTITUDE. $B$2:$B$13 should appear on INPUT Y RANGE.
Excel Detailed Steps: Bivariate Regression (Cont.)

b. Click in the INPUT X RANGE box. Select (highlight) all the rows of data under DURATION. $C$2:$C$13 should appear on INPUT X RANGE.

c. Leave LABELS and CONSTANT IS ZERO as blanks. CONFIDENCE LEVEL should be 95% (default).

9. In the OUTPUT OPTIONS pop-up window, select NEW WORKBOOK Options.
10. Under RESIDUALS check RESIDUAL PLOTS.
11. Under NORMAL PROBABILITY check NORMAL PROBABILITY PLOTS.
12. Click OK.
Excel Detailed Steps: Multiple Regression

The steps for running multiple regression are similar, except for Step 8b. In Step 8b, click in the INPUT X RANGE box.

Select (highlight) all the rows of data under Duration and Importance. $C$2:$D$13 should appear on INPUT X RANGE.
MINITAB

Correlation can be computed using STAT > BASIC STATISTICS > CORRELATION function. It calculates Pearson’s product moment using all the columns.

SAS

For a point-and-click approach for performing metric correlations, use the ANALYZE task within SAS Enterprise Guide. The MULTIVARIATE > CORRELATIONS task offers Pearson product moment correlations, as well as some other measures.
Exhibit 18.2
Other Computer Programs for Regression

**MINITAB**

Regression analysis under the STAT > REGRESSION function can perform simple and multiple analysis. The output includes a linear regression equation, table of coefficients $R^2$, $R^2$ adjusted, analysis of variance table, a table of fits and residuals that provide unusual observations and residual plots.

**SAS**

For a point-and-click approach for performing regression analysis, use the ANALYZE task within SAS Enterprise Guide. The REGRESSION > LINEAR task calculates bivariate and multiple regression equations, associated statistics, and plots. It allows for an easy examination of residuals.
Acronym: Regression

The main features of regression analysis may be summarized by the acronym REGRESSION:

R - residual analysis is useful
E - estimation of parameters: solution of simultaneous equations
G - general model is linear
R\(^2\) - strength of association
E - error terms are independent and \(N(0, s^2)\)
S - standardized regression coefficients
S - standard error of estimate: prediction accuracy
I - individual coefficients and overall \(F\)-tests
O - optimal: minimizes total error
N - unstandardized regression coefficients