Michael Zagaris/Getty Images Sport/Getty Images, Inc.

CHAPTER

13

Simple Linear Regression

Are the heights and weights of persons related? Does a person's weight depend on his/her height? If yes, what is the change in the weight of a person, on average, for every one inch increase in height? What is this rate of change for National Football League players? (See Case Study 13-1.)

This chapter considers the relationship between two variables in two ways: (1) by using regression analysis and (2) by computing the correlation coefficient. By using the regression model, we can evaluate the magnitude of change in one variable due to a certain change in another variable. For example, an economist can estimate the amount of change in food expenditure due to a certain change in the income of a household by using the regression model. A sociologist may want to estimate the increase in the crime rate due to a particular increase in the unemployment rate. Besides answering these questions, a regression model also helps predict the value of one variable for a given value of another variable. For example, by using the regression line, we can predict the (approximate) food expenditure of a household with a given income.

The correlation coefficient, on the other hand, simply tells us how strongly two variables are related. It does not provide any information about the size of the change in one variable as a result of a certain change in the other variable. For example, the correlation coefficient tells us how strongly income and food expenditure or crime rate and unemployment rate are related.

13.1 Simple Linear Regression

Only simple linear regression will be discussed in this chapter.[1] In the next two subsections the meaning of the words *simple* and *linear* as used in *simple linear regression* is explained.

13.1.1 Simple Regression

Let us return to the example of an economist investigating the relationship between food expenditure and income. What factors or variables does a household consider when deciding how much money it should spend on food every week or every month? Certainly, income of the household is one factor. However, many other variables also affect food expenditure. For instance, the assets owned by the household, the size of the household, the preferences and tastes of household members, and any special dietary needs of household members are some of the variables that influence a household's decision about food expenditure. These variables are called **independent** or **explanatory variables** because they all vary independently, and they explain the variation in food expenditures among different households. In other words, these variables explain why different households spend different amounts of money on food. Food expenditure is called the **dependent variable** because it depends on the independent variables. Studying the effect of two or more independent variables on a dependent variable using regression analysis is called **multiple regression**. However, if we choose only one (usually the most important) independent variable and study the effect of that single variable on a dependent variable, it is called a **simple regression**. Thus, a simple regression includes only two variables: one independent and one dependent. Note that whether it is a simple or a multiple regression analysis, it always includes one and only one dependent variable. It is the number of independent variables that changes in simple and multiple regressions.

Definition

Simple Regression A regression model is a mathematical equation that describes the relationship between two or more variables. A *simple regression* model includes only two variables: one independent and one dependent. The dependent variable is the one being explained, and the independent variable is the one that explains the variation in the dependent variable.

13.1.2 Linear Regression

The relationship between two variables in a regression analysis is expressed by a mathematical equation called a **regression equation** or **model**. A regression equation, when plotted, may assume one of many possible shapes, including a straight line. A regression equation that gives a straight-line relationship between two variables is called a **linear regression model**; otherwise, the model is called a **nonlinear regression model**. In this chapter, only linear regression models are studied.

Definition

Linear Regression A (simple) regression model that gives a straight-line relationship between two variables is called a *linear regression* model.

The two diagrams in Figure 13.1 show a linear and a nonlinear relationship between the dependent variable food expenditure and the independent variable income. A linear relationship

[1]The term *regression* was first used by Sir Francis Galton (1822–1911), who studied the relationship between the heights of children and the heights of their parents.

between income and food expenditure, shown in Figure 13.1*a*, indicates that as income increases, the food expenditure always increases at a constant rate. A nonlinear relationship between income and food expenditure, as depicted in Figure 13.1*b*, shows that as income increases, the food expenditure increases, although, after a point, the rate of increase in food expenditure is lower for every subsequent increase in income.

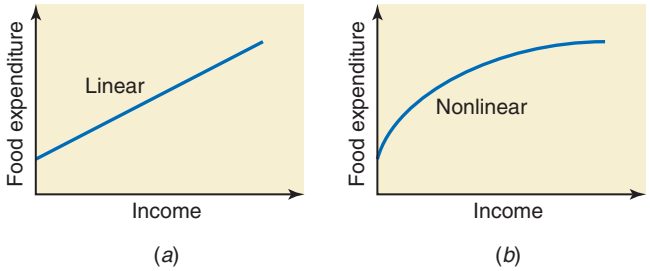


Figure 13.1 Relationship between food expenditure and income. (a) Linear relationship. (b) Nonlinear relationship.

The **equation of a linear relationship** between two variables x and y is written as

$$y = a + bx$$

Each set of values of a and b gives a different straight line. For instance, when $a = 50$ and $b = 5$, this equation becomes

$$y = 50 + 5x$$

To plot a straight line, we need to know two points that lie on that line. We can find two points on a line by assigning any two values to x and then calculating the corresponding values of y. For the equation $y = 50 + 5x$:

1. When $x = 0$, then $y = 50 + 5(0) = 50$.

2. When $x = 10$, then $y = 50 + 5(10) = 100$.

These two points are plotted in Figure 13.2. By joining these two points, we obtain the line representing the equation $y = 50 + 5x$.

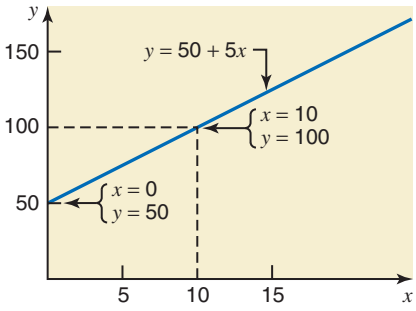


Figure 13.2 Plotting a linear equation.

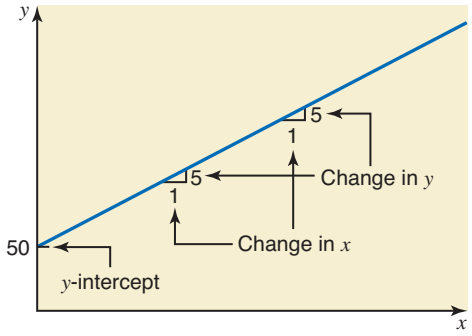
Note that in Figure 13.2 the line intersects the y (vertical) axis at 50. Consequently, 50 is called the ***y*-intercept**. The y-intercept is given by the constant term in the equation. It is the value of y when x is zero.

In the equation $y = 50 + 5x$, 5 is called the **coefficient of *x*** or the **slope** of the line. It gives the amount of change in y due to a change of one unit in x. For example:

$$\text{If } x = 10, \text{ then } y = 50 + 5(10) = 100.$$
$$\text{If } x = 11, \text{ then } y = 50 + 5(11) = 105.$$

Hence, as x increases by 1 unit (from 10 to 11), y increases by 5 units (from 100 to 105). This is true for any value of x. Such changes in x and y are shown in Figure 13.3.

Figure 13.3 *y*-intercept and slope of a line.



In general, when an equation is written in the form

$$y = a + bx$$

a gives the y-intercept and b represents the slope of the line. In other words, a represents the point where the line intersects the y-axis, and b gives the amount of change in y due to a change of one unit in x. Note that b is also called the coefficient of x.

13.1.3 Simple Linear Regression Model

In a regression model, the independent variable is usually denoted by x, and the dependent variable is usually denoted by y. The x variable, with its coefficient, is written on the right side of the = sign, whereas the y variable is written on the left side of the = sign. The y-intercept and the slope, which we earlier denoted by a and b, respectively, can be represented by any of the many commonly used symbols. Let us denote the y-intercept (which is also called the *constant term*) by $\boldsymbol{A}$, and the slope (or the coefficient of the x variable) by $\boldsymbol{B}$. Then, our simple linear regression model is written as

Constant term or y-intercept — Slope

$$y = A + Bx \tag{1}$$

Dependent variable — Independent variable

In model (1), A gives the value of y for $x = 0$, and B gives the change in y due to a change of one unit in x.

Model (1) is called a **deterministic model**. It gives an **exact relationship** between x and y. This model simply states that y is determined exactly by x, and for a given value of x there is one and only one (unique) value of y.

However, in many cases the relationship between variables is not exact. For instance, if y is food expenditure and x is income, then model (1) would state that food expenditure is determined by income only and that all households with the same income spend the same amount on food. As mentioned earlier, however, food expenditure is determined by many variables, only one of which is included in model (1). In reality, different households with the same income spend different amounts of money on food because of the differences in the sizes of the household, the assets they own, and their preferences and tastes. Hence, to take these variables into consideration and to make our model complete, we add another term to the right side of model (1). This term is called the **random error term**. It is denoted by ϵ (Greek letter *epsilon*). The complete regression model is written as

$$y = A + Bx + \epsilon \tag{2}$$

Random error term

The regression model (2) is called a **probabilistic model** or a **statistical relationship**.

Definition

Equation of a Regression Model In the *regression model* $y = A + Bx + \epsilon$, A is called the y-intercept or constant term, B is the slope, and ϵ is the random error term. The dependent and independent variables are y and x, respectively.

The random error term ϵ is included in the model to represent the following two phenomena:

1. *Missing or omitted variables*. As mentioned earlier, food expenditure is affected by many variables other than income. The random error term ϵ is included to capture the effect of all those missing or omitted variables that have not been included in the model.
2. *Random variation*. Human behavior is unpredictable. For example, a household may have many parties during one month and spend more than usual on food during that month. The same household may spend less than usual during another month because it spent quite a bit of money to buy furniture. The variation in food expenditure for such reasons may be called random variation.

In model (2), A and B are the **population parameters**. The regression line obtained for model (2) by using the population data is called the **population regression line**. The values of A and B in the population regression line are called the **true values of the y-intercept and slope**, respectively.

However, population data are difficult to obtain. As a result, we almost always use sample data to estimate model (2). The values of the y-intercept and slope calculated from sample data on x and y are called the **estimated values of A and B** and are denoted by a and b, respectively. Using a and b, we write the estimated regression model as

$$\hat{y} = a + bx \tag{3}$$

where $\hat{y}$ (read as y *hat*) is the **estimated** or **predicted value of y** for a given value of x. Equation (3) is called the **estimated regression model**; it gives the **regression of y on x**.

Definition

Estimates of A and B In the model $\hat{y} = a + bx$, a and b, which are calculated using sample data, are called the *estimates of A and B*, respectively.

13.1.4 Scatter Diagram

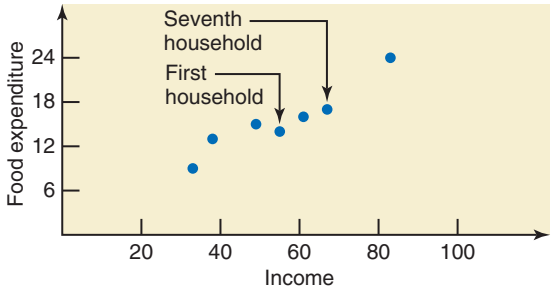
Suppose we take a sample of seven households from a small city and collect information on their incomes and food expenditures for the last month. The information obtained (in hundreds of dollars) is given in Table 13.1.

Table 13.1 Incomes and Food Expenditures of Seven Households

Income	Food Expenditure
55	14
83	24
38	13
61	16
33	9
49	15
67	17

In Table 13.1, we have a pair of observations for each of the seven households. Each pair consists of one observation on income and a second on food expenditure. For example, the first household's income for the last month was \$5500 and its food expenditure was \$1400. By plotting all seven pairs of values, we obtain a **scatter diagram** or **scatterplot**. Figure 13.4 gives the scatter diagram for the data of Table 13.1. Each dot in this diagram represents one household. A scatter diagram is helpful in detecting a relationship between two variables. For example, by looking at the scatter diagram of Figure 13.4, we can observe that there exists a strong linear relationship between food expenditure and income. If a straight line is drawn through the points, the points will be scattered closely around the line.

Figure 13.4 Scatter diagram.



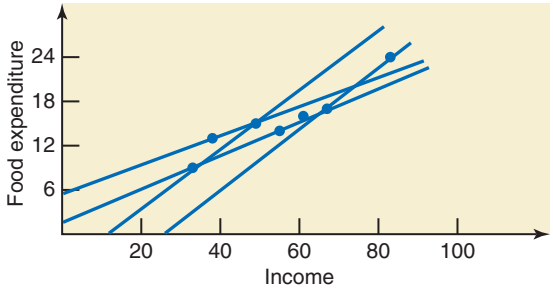
Definition

Scatter Diagram A plot of paired observations is called a *scatter diagram*.

As shown in Figure 13.5, a large number of straight lines can be drawn through the scatter diagram of Figure 13.4. Each of these lines will give different values for a and b of model (3).

In regression analysis, we try to find a line that best fits the points in the scatter diagram. Such a line provides a best possible description of the relationship between the dependent and independent variables. The **least squares method**, discussed in the next section, gives such a line. The line obtained by using the least squares method is called the **least squares regression line**.

Figure 13.5 Scatter diagram and straight lines.



13.1.5 Least Squares Regression Line

The value of y obtained for a member from the survey is called the **observed or actual value of y**. As mentioned earlier in this section, the value of y, denoted by $\hat{y}$, obtained for a given x by using the regression line is called the **predicted value of y**. The random error ϵ denotes the difference between the actual value of y and the predicted value of y for population data. For example, for a given household, ϵ is the difference between what this household actually spent on food during the last month and what is predicted using the population regression line. The ϵ is also called the *residual* because it measures the surplus (positive or negative) of actual food expenditure over what is predicted by using the regression model. If we estimate model (2) by

using sample data, the difference between the actual y and the predicted y based on this estimation cannot be denoted by ϵ. *The random error for the sample regression model is denoted by* e. Thus, e is an estimator of ϵ. If we estimate model (2) using sample data, then the value of e is given by

$$e = \text{Actual food expenditure} - \text{Predicted food expenditure} = y - \hat{y}$$

In Figure 13.6, e is the vertical distance between the actual position of a household and the point on the regression line. Note that in such a diagram, we always measure the dependent variable on the vertical axis and the independent variable on the horizontal axis.

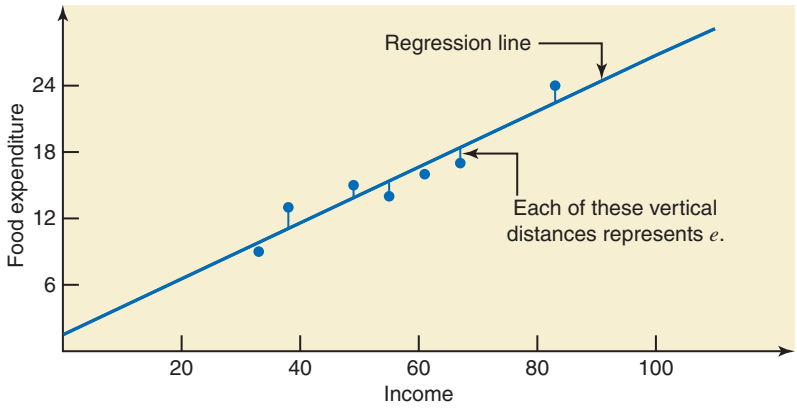


Figure 13.6 Regression line and random errors.

The value of an error is positive if the point that gives the actual food expenditure is above the regression line and negative if it is below the regression line. *The sum of these errors is always zero*. In other words, the sum of the actual food expenditures for seven households included in the sample will be the same as the sum of the food expenditures predicted by the regression model. Thus,

$$\Sigma e = \Sigma(y - \hat{y}) = 0$$

Hence, to find the line that best fits the scatter of points, we cannot minimize the sum of errors. Instead, we minimize the **error sum of squares**, denoted by **SSE**, which is obtained by adding the squares of errors. Thus,

$$\text{SSE} = \Sigma e^2 = \Sigma(y - \hat{y})^2$$

The least squares method gives the values of a and b for model (3) such that the sum of squared errors (SSE) is minimum.

Error Sum of Squares (SSE) The *error sum of squares*, denoted by SSE, is

$$\text{SSE} = \Sigma e^2 = \Sigma(y - \hat{y})^2$$

The values of a and b that give the minimum SSE are called the *least squares estimates* of A and B, and the regression line obtained with these estimates is called the *least squares line*.

The Least Squares Line For the least squares regression line $\hat{y} = a + bx$,

$$b = \frac{\text{SS}_{xy}}{\text{SS}_{xx}} \quad \text{and} \quad a = \bar{y} - b\bar{x}$$

where $$\text{SS}_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} \quad \text{and} \quad \text{SS}_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n}$$

and SS stands for "sum of squares." The least squares regression line $\hat{y} = a + bx$ is also called the regression of y on x.

The least squares values of a and b are computed using the formulas just given.[2] These formulas are for estimating a sample regression line. Suppose we have access to a population data set. We can find the population regression line by using the same formulas with a little adaptation. If we have access to population data, we replace a by A, b by B, and n by N in these formulas, and use the values of Σx, Σy, Σxy, and Σx^2 calculated for population data to make the required computations. The population regression line is written as

$$\mu_{y|x} = A + Bx$$

where $\mu_{y|x}$ is read as "the mean value of y for a given x." When plotted on a graph, the points on this population regression line give the average values of y for the corresponding values of x. These average values of y are denoted by $\mu_{y|x}$.

Example 13–1 illustrates how to estimate a regression line for sample data.

EXAMPLE 13–1

Estimating the least squares regression line.

Find the least squares regression line for the data on incomes and food expenditures of the seven households given in Table 13.1. Use income as an independent variable and food expenditure as a dependent variable.

© Troels Graugaard/iStockphoto

Solution We are to find the values of a and b for the regression model $\hat{y} = a + bx$. Table 13.2 shows the calculations required for the computation of a and b. We denote the independent variable (income) by x and the dependent variable (food expenditure) by y, both in hundreds of dollars.

Table 13.2

Income x	Food Expenditure y	xy	x^2
55	14	770	3025
83	24	1992	6889
38	13	494	1444
61	16	976	3721
33	9	297	1089
49	15	735	2401
67	17	1139	4489
$\Sigma x = 386$	$\Sigma y = 108$	$\Sigma xy = 6403$	$\Sigma x^2 = 23{,}058$

The following steps are performed to compute a and b.

Step 1. *Compute* Σx, Σy, $\bar{x}$, *and* $\bar{y}$.

$$\Sigma x = 386, \qquad \Sigma y = 108$$
$$\bar{x} = \Sigma x/n = 386/7 = 55.1429$$
$$\bar{y} = \Sigma y/n = 108/7 = 15.4286$$

Step 2. *Compute* Σxy *and* Σx^2.

To calculate Σxy, we multiply the corresponding values of x and y. Then, we sum all the products. The products of x and y are recorded in the third column of Table 13.2. To compute Σx^2, we square each of the x values and then add them. The squared values of x are listed in the fourth column of Table 13.2. From these calculations,

$$\Sigma xy = 6403 \quad \text{and} \quad \Sigma x^2 = 23{,}058$$

[2]The values of SS_{xy} and SS_{xx} can also be obtained by using the following basic formulas:

$$SS_{xy} = \Sigma(x - \bar{x})(y - \bar{y}) \quad \text{and} \quad SS_{xx} = \Sigma(x - \bar{x})^2$$

However, these formulas take longer to make calculations.

Step 3. *Compute* SS_{xy} *and* SS_{xx}:

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 6403 - \frac{(386)(108)}{7} = 447.5714$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 23{,}058 - \frac{(386)^2}{7} = 1772.8571$$

Step 4. *Compute* a *and* b:

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{447.5714}{1772.8571} = .2525$$

$$a = \bar{y} - b\bar{x} = 15.4286 - (.2525)(55.1429) = 1.5050$$

Thus, our estimated regression model $\hat{y} = a + bx$ is

$$\boldsymbol{\hat{y} = 1.5050 + .2525x}$$

This regression line is called the least squares regression line. It gives the *regression of food expenditure on income*.

Note that we have rounded all calculations to four decimal places. We can round the values of a and b in the regression equation to two decimal places, but we do not do this here because we will use this regression equation for prediction and estimation purposes later. ■

Using this estimated regression model, we can find the predicted value of y for any specific value of x. For instance, suppose we randomly select a household whose monthly income is \$6100, so that $x = 61$ (recall that x denotes income in hundreds of dollars). The predicted value of food expenditure for this household is

$$\hat{y} = 1.5050 + (.2525)(61) = \$16.9075 \text{ hundred} = \$1690.75$$

In other words, based on our regression line, we predict that a household with a monthly income of \$6100 is expected to spend \$1690.75 per month on food. This value of $\hat{y}$ can also be interpreted as a point estimator of the mean value of y for $x = 61$. Thus, we can state that, on average, all households with a monthly income of \$6100 spend about \$1690.75 per month on food.

In our data on seven households, there is one household whose income is \$6100. The actual food expenditure for that household is \$1600 (see Table 13.1). The difference between the actual and predicted values gives the error of prediction. Thus, the error of prediction for this household, which is shown in Figure 13.7, is

$$e = y - \hat{y} = 16 - 16.9075 = -\$.9075 \text{ hundred} = -\$90.75$$

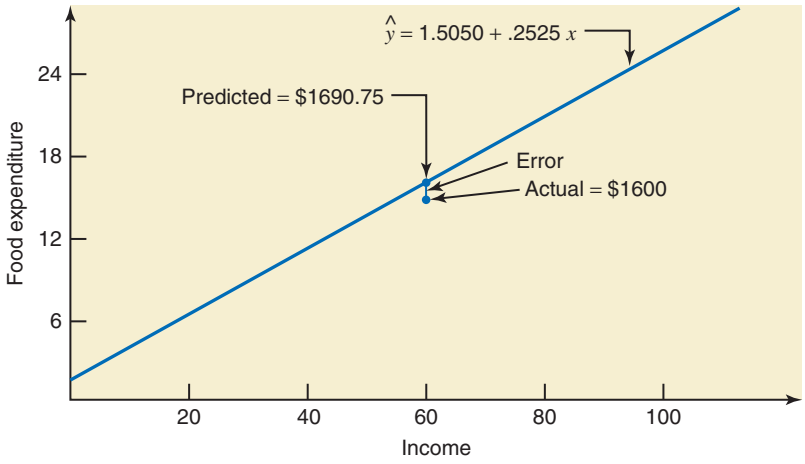


Figure 13.7 Error of prediction.

Therefore, the error of prediction is −\$90.75. The negative error indicates that the predicted value of y is greater than the actual value of y. Thus, if we use the regression model, this household's food expenditure is overestimated by \$90.75.

13.1.6 Interpretation of *a* and *b*

How do we interpret $a = 1.5050$ and $b = .2525$ obtained in Example 13–1 for the regression of food expenditure on income? A brief explanation of the y-intercept and the slope of a regression line was given in Section 13.1.2. Below we explain the meaning of a and b in more detail.

Interpretation of *a*

Consider a household with zero income. Using the estimated regression line obtained in Example 13–1, we get the predicted value of y for $x = 0$ as

$$\hat{y} = 1.5050 + .2525(0) = \$1.5050 \text{ hundred} = \$150.50$$

Thus, we can state that a household with no income is expected to spend \$150.50 per month on food. Alternatively, we can also state that the point estimate of the average monthly food expenditure for all households with zero income is \$150.50. Note that here we have used $\hat{y}$ as a point estimate of $\mu_{y|x}$. Thus, $a = 150.50$ gives the predicted or mean value of y for $x = 0$ based on the regression model estimated for the sample data.

However, we should be very careful when making this interpretation of a. In our sample of seven households, the incomes vary from a minimum of \$3300 to a maximum of \$8300. (Note that in Table 13.1, the minimum value of x is 33 and the maximum value is 83.) Hence, our regression line is valid only for the values of x between 33 and 83. If we predict y for a value of x outside this range, the prediction usually will not hold true. Thus, since $x = 0$ is outside the range of household incomes that we have in the sample data, the prediction that a household with zero income spends \$150.50 per month on food does not carry much credibility. The same is true if we try to predict y for an income greater than \$8300, which is the maximum value of x in Table 13.1.

Interpretation of *b*

The value of b in a regression model gives the change in y (dependent variable) due to a change of one unit in x (independent variable). For example, by using the regression equation obtained in Example 13–1, we see:

$$\text{When } x = 50, \quad \hat{y} = 1.5050 + .2525(50) = 14.1300$$
$$\text{When } x = 51, \quad \hat{y} = 1.5050 + .2525(51) = 14.3825$$

Hence, when x increased by one unit, from 50 to 51, $\hat{y}$ increased by $14.3825 - 14.1300 = .2525$, which is the value of b. Because our unit of measurement is hundreds of dollars, we can state that, on average, a \$100 increase in income will result in a \$25.25 increase in food expenditure. We can also state that, on average, a \$1 increase in income of a household will increase the food expenditure by \$.2525. Note the phrase "on average" in these statements. The regression line is seen as a measure of the mean value of y for a given value of x. If one household's income is increased by \$100, that household's food expenditure may or may not increase by \$25.25. However, if the incomes of all households are increased by \$100 each, the average increase in their food expenditures will be very close to \$25.25.

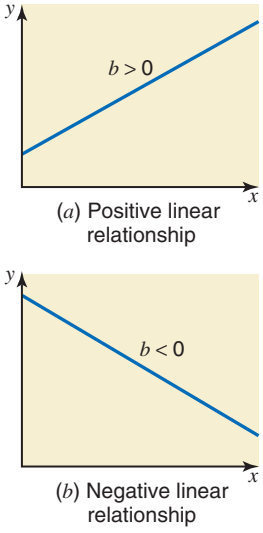


Figure 13.8 Positive and negative linear relationships between x and y.

Note that when b is positive, an increase in x will lead to an increase in y, and a decrease in x will lead to a decrease in y. In other words, when b is positive, the movements in x and y are in the same direction. Such a relationship between x and y is called a **positive linear relationship**. The regression line in this case slopes upward from left to right. On the other hand, if the value of b is negative, an increase in x will lead to a decrease in y, and a decrease in x will cause an increase in y. The changes in x and y in this case are in opposite directions. Such a relationship between x and y is called a **negative linear relationship**. The regression line in this case slopes downward from left to right. The two diagrams in Figure 13.8 show these two cases.

Remember ▶ For a regression model, b is computed as $b = SS_{xy}/SS_{xx}$. The value of SS_{xx} is always positive, and that of SS_{xy} can be positive or negative. Hence, the sign of b depends on the sign of SS_{xy}. If SS_{xy} is positive (as in our example on the incomes and food expenditures of seven households), then b will be positive, and if SS_{xy} is negative, then b will be negative.

Case Study 13–1 illustrates the difference between the population regression line and a sample regression line.

CASE STUDY 13–1

REGRESSION OF WEIGHTS ON HEIGHTS FOR NFL PLAYERS

Data Set III that accompanies this text lists many characteristics of National Football League (NFL) players who were on the rosters of all NFL teams as of October 31, 2011. These data comprise the population of NFL players for that point in time. We postulate the following simple linear regression model for these data:

$$y = A + Bx + \epsilon$$

where y is the weight (in pounds) and x is the height (in inches) of an NFL player.

Using the population data that contain 1874 players, we obtain the following regression line:

$$\mu_{y|x} = -690 + 12.7x$$

This equation gives the population regression line because it is obtained by using the population data. (Note that in the population regression line we write $\mu_{y|x}$ instead of $\hat{y}$.) Thus, the true values of A and B are, respectively,

$$A = -690 \quad \text{and} \quad B = 12.7$$

The value of B indicates that for every 1-inch increase in the height of an NFL player, weight increases on average by 12.7 pounds. However, $A = -690$ does not make any sense. It states that the weight of a player with zero height is -690 pounds. (Recall from Section 13.1.6 that we must be very careful if and when we apply the regression equation to predict y for values of x outside the range of data used to find the regression line.) Figure 13.9 gives the scatter diagram and the regression line for the heights and weights of all NFL players.

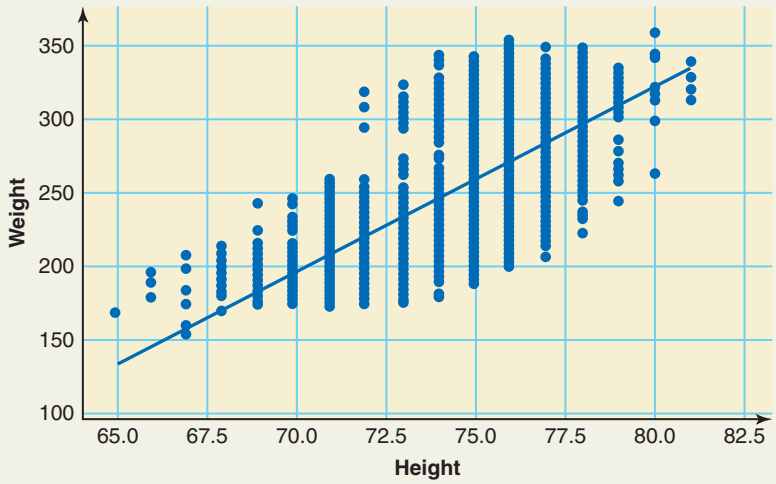


Figure 13.9 Scatter diagram for the data on heights and weights of all NFL players.

Next, we selected a random sample of 50 players and estimated the regression model for this sample. The estimated regression line for this sample is

$$\hat{y} = -739 + 13.3x$$

The values of a and b are

$$a = -739 \quad \text{and} \quad b = 13.3$$

These values of a and b give the estimates of A and B based on sample data. The scatter diagram and the regression line for the sample observations on heights and weights is given in Figure 13.10. Note that this figure does not show exactly 50 dots because some points/dots may be exactly the same or very close to each other.

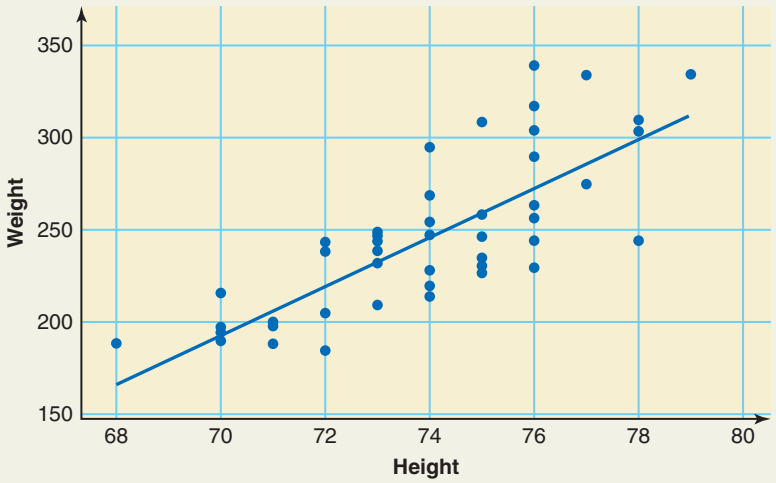


Figure 13.10 Scatter diagram for the data on heights and weights of 50 NFL players.

Source: www.sportscity.com/NFL-salaries and www.nfl.com/teams

As we can observe from Figures 13.9 and 13.10, the scatter diagrams for population and sample data both show a (positive) linear relationship between the heights and weights of NFL players, although not a very strong positive relationship.

13.1.7 Assumptions of the Regression Model

Like any other theory, the linear regression analysis is also based on certain assumptions. Consider the population regression model

$$y = A + Bx + \epsilon \tag{4}$$

Four assumptions are made about this model. These assumptions are explained next with reference to the example on incomes and food expenditures of households. Note that these assumptions are made about the population regression model and not about the sample regression model.

Assumption 1: The random error term ϵ has a mean equal to zero for each x. In other words, among all households with the same income, some spend more than the predicted food expenditure (and, hence, have positive errors) and others spend less than the predicted food expenditure (and, consequently, have negative errors). This assumption simply states that the sum of the positive errors is equal to the sum of the negative errors, so that the mean of errors for all households with the same income is zero. Thus, when the mean value of ϵ is zero, the mean value of y for a given x is equal to $A + Bx$, and it is written as

$$\mu_{y|x} = A + Bx$$

As mentioned earlier in this chapter, $\mu_{y|x}$ is read as "the mean value of y for a given value of x." When we find the values of A and B for model (4) using the population data, the points on the regression line give the average values of y, denoted by $\mu_{y|x}$, for the corresponding values of x.

Assumption 2: The errors associated with different observations are independent. According to this assumption, the errors for any two households in our example are independent. In other words, all households decide independently how much to spend on food.

Assumption 3: For any given x, the distribution of errors is normal. The corollary of this assumption is that the food expenditures for all households with the same income are normally distributed.

Assumption 4: The distribution of population errors for each x has the same (constant) standard deviation, which is denoted by σ_ϵ. This assumption indicates that the spread of points around the regression line is similar for all x values.

Figure 13.11 illustrates the meanings of the first, third, and fourth assumptions for households with incomes of \$4000 and \$7500 per month. The same assumptions hold true for any other income level. In the population of all households, there will be many households with a monthly income of \$4000. Using the population regression line, if we calculate the errors for all these households and prepare the distribution of these errors, it will look like the distribution given in Figure 13.11*a*. Its standard deviation will be σ_ϵ. Similarly, Figure 13.11*b* gives the distribution of errors for all those households in the population whose monthly income is \$7500. Its standard deviation is also σ_ϵ. Both of these distributions are identical. Note that the mean of both of these distributions is $E(\epsilon) = 0$.

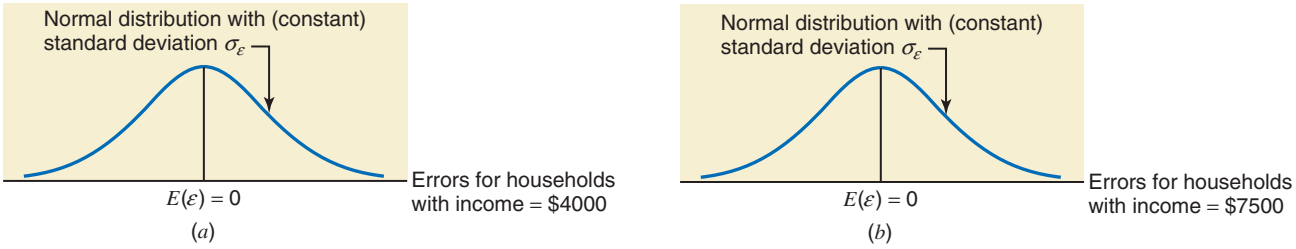


Figure 13.11 (*a*) Errors for households with an income of \$4000 per month. (*b*) Errors for households with an income of \$7500 per month.

Figure 13.12 shows how the distributions given in Figure 13.11 look when they are plotted on the same diagram with the population regression line. The points on the vertical line through $x = 40$ give the food expenditures for various households in the population, each of which has the same monthly income of \$4000. The same is true about the vertical line through $x = 75$ or any other vertical line for some other value of x.

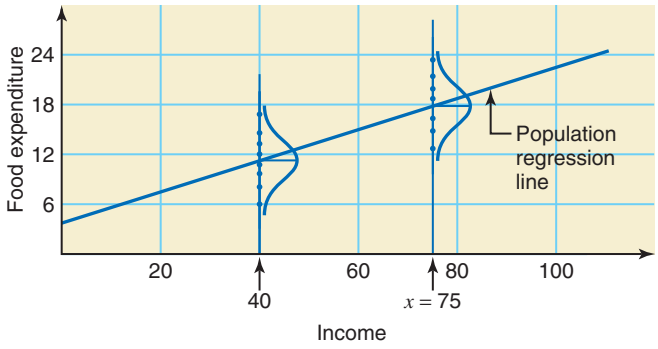


Figure 13.12 Distribution of errors around the population regression line.

13.1.8 Cautions in Using Regression

When carefully applied, regression is a very helpful technique for making predictions and estimations about one variable for a certain value of another variable. However, we need to be cautious when using the regression analysis, for it can give us misleading results and predictions. The following are the two most important points to remember when using regression.

(a) A Note on the Use of Simple Linear Regression

We should apply linear regression with caution. When we use simple linear regression, we assume that the relationship between two variables is described by a straight line. In the real world, the relationship between variables may not be linear. Hence, before we use a simple linear regression, it is better to construct a scatter diagram and look at the plot of the data points. We should estimate a linear regression model only if the scatter diagram indicates such a relationship. The scatter diagrams of Figure 13.13 give two examples for which the relationship between x and y is not linear. Consequently, using linear regression in such cases would be wrong.

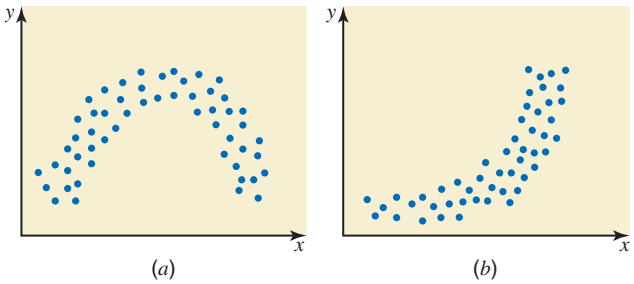


Figure 13.13 Nonlinear relationship between x and y.

(b) Extrapolation

The regression line estimated for the sample data is reliable only for the range of x values observed in the sample. For example, the values of x in our example on incomes and food expenditures vary from a minimum of 33 to a maximum of 83. Hence, our estimated regression line is applicable only for values of x between 33 and 83; that is, we should use this regression line to estimate the mean food expenditure or to predict the food expenditure of a single household only for income levels between \$3300 and \$8300. If we estimate or predict y for a value of x either less than 33 or greater than 83, it is called *extrapolation*. This does not mean that we should never use the regression line for extrapolation. Instead, we should interpret such predictions cautiously and not attach much importance to them.

Similarly, if the data used for the regression estimation are time-series data (see Exercises 13.100 and 13.101), the predicted values of y for periods outside the time interval used for the estimation of the regression line should be interpreted very cautiously. When using the estimated regression line for extrapolation, we are assuming that the same linear relationship between the two variables holds true for values of x outside the given range. It is possible that the relationship between the two variables may not be linear outside that range. Nonetheless, even if it is linear, adding a few more observations at either end will probably give a new estimation of the regression line.

EXERCISES

CONCEPTS AND PROCEDURES

13.1 Explain the meaning of the words *simple* and *linear* as used in *simple linear regression*.

13.2 Explain the meaning of independent and dependent variables for a regression model.

13.3 Explain the difference between exact and nonexact relationships between two variables. Give one example of each.

13.4 Explain the difference between linear and nonlinear relationships between two variables.

13.5 Explain the difference between a simple and a multiple regression model.

13.6 Briefly explain the difference between a deterministic and a probabilistic regression model.

13.7 Why is the random error term included in a regression model?

13.8 Explain the least squares method and least squares regression line. Why are they called by these names?

13.9 Explain the meaning and concept of SSE. You may use a graph for illustration purposes.

13.10 Explain the difference between y and $\hat{y}$.

13.11 Two variables x and y have a positive linear relationship. Explain what happens to the value of y when x increases. Give one example of a positive relationship between two variables.

13.12 Two variables x and y have a negative linear relationship. Explain what happens to the value of y when x increases. Give one example of a negative relationship between two variables.

13.13 Explain the following.

- **a.** Population regression line
- **b.** Sample regression line
- **c.** True values of A and B
- **d.** Estimated values of A and B that are denoted by a and b, respectively

13.14 Briefly explain the assumptions of the population regression model.

13.15 Plot the following straight lines. Give the values of the y-intercept and slope for each of these lines and interpret them. Indicate whether each of the lines gives a positive or a negative relationship between x and y.

a. $y = 80 - 3x$ **b.** $y = 250 + 8x$

13.16 Plot the following straight lines. Give the values of the y-intercept and slope for each of these lines and interpret them. Indicate whether each of the lines gives a positive or a negative relationship between x and y.

a. $y = 100 + 6x$ **b.** $y = -40 + 9x$

13.17 A population data set produced the following information.

$$N = 300, \quad \Sigma x = 9880, \quad \Sigma y = 1324, \quad \Sigma xy = 85{,}080, \quad \Sigma x^2 = 500{,}350$$

Find the population regression line.

13.18 A population data set produced the following information.

$$N = 420, \quad \Sigma x = 3920, \quad \Sigma y = 2840, \quad \Sigma xy = 28{,}350, \quad \Sigma x^2 = 48{,}530$$

Find the population regression line.

13.19 The following information is obtained from a sample data set.

$$n = 10, \quad \Sigma x = 100, \quad \Sigma y = 280, \quad \Sigma xy = 3120, \quad \Sigma x^2 = 1140$$

Find the estimated regression line.

13.20 The following information is obtained from a sample data set.

$$n = 12, \quad \Sigma x = 66, \quad \Sigma y = 588, \quad \Sigma xy = 2104, \quad \Sigma x^2 = 350$$

Find the estimated regression line.

APPLICATIONS

13.21 A car rental company charges \$80 a day and 10 cents per mile for renting a car. Let y be the total rental charges (in dollars) for a car for one day and x be the miles driven. The equation for the relationship between x and y is

$$y = 80 + .10x$$

a. How much will a person pay who rents a car for one day and drives it 100 miles?
b. Suppose each of 20 persons rents a car from this agency for one day and drives it 100 miles. Will each of them pay the same amount for renting a car for a day or do you expect each person to pay a different amount? Explain.
c. Is the relationship between x and y exact or nonexact?

13.22 Bob's Pest Removal Service specializes in removing wild creatures (skunks, bats, reptiles, etc.) from private homes. He charges \$50 to go to a house plus \$30 per hour for his services. Let y be the total amount (in dollars) paid by a household using Bob's services and x the number of hours Bob spends capturing and removing the animal(s). The equation for the relationship between x and y is

$$y = 50 + 30x$$

a. Bob spent 4 hours removing a coyote from under Alice's house. How much will he be paid?
b. Suppose nine persons called Bob for assistance during a week. Strangely enough, each of these jobs required exactly 4 hours. Will each of these clients pay Bob the same amount, or do you expect each one to pay a different amount? Explain.
c. Is the relationship between x and y exact or nonexact?

13.23 A researcher took a sample of 25 electronics companies and found the following relationship between x and y, where x is the amount of money (in millions of dollars) spent on advertising by a company in 2009 and y represents the total gross sales (in millions of dollars) of that company for 2009.

$$\hat{y} = 3.6 + 11.75x$$

a. An electronics company spent \$3 million on advertising in 2009. What are its expected gross sales for 2009?
b. Suppose four electronics companies spent \$3 million each on advertising in 2009. Do you expect these four companies to have the same actual gross sales for 2009? Explain.
c. Is the relationship between x and y exact or nonexact?

13.24 A researcher took a sample of 20 years and found the following relationship between x and y, where x is the number of major natural calamities (such as tornadoes, hurricanes, earthquakes, floods, etc.) that occurred during a year and y represents the average annual total profits (in millions of dollars) of a sample of insurance companies in the United States.

$$\hat{y} = 342.6 - 2.10x$$

a. A randomly selected year had 24 major calamities. What are the expected average profits of U.S. insurance companies for that year?

b. Suppose the number of major calamities was the same for each of 3 years. Do you expect the average profits for all U.S. insurance companies to be the same for each of these 3 years? Explain.
c. Is the relationship between x and y exact or nonexact?

13.25 An auto manufacturing company wanted to investigate how the price of one of its car models depreciates with age. The research department at the company took a sample of eight cars of this model and collected the following information on the ages (in years) and prices (in hundreds of dollars) of these cars.

Age	8	3	6	9	2	5	6	2
Price	38	220	95	33	267	134	112	245

a. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between ages and prices of cars?
b. Find the regression line with price as a dependent variable and age as an independent variable.
c. Give a brief interpretation of the values of a and b calculated in part b.
d. Plot the regression line on the scatter diagram of part a and show the errors by drawing vertical lines between scatter points and the regression line.
e. Predict the price of a 7-year-old car of this model.
f. Estimate the price of an 18-year-old car of this model. Comment on this finding.

13.26 The following table gives information on the amount of sugar (in grams) and the calorie count in one serving of a sample of 13 varieties of Kellogg's cereal.

Sugar (grams)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	180	80	120	190	190	100	120

Source: kelloggs.com.

a. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between the amount of sugar and the number of calories per serving?
b. Find the predictive regression equation of the number of calories on the amount of sugar.
c. Give a brief interpretation of the values of a and b calculated in part b.
d. Plot the predictive regression line on the scatter diagram of part a and show the errors by drawing vertical lines between scatter points and the predictive regression line.
e. Calculate the predicted calorie count for a cereal with 16 grams of sugar per serving.
f. Estimate the calorie count for a cereal with 52 grams of sugar per serving. Comment on this finding.

13.27 The following table contains information on the amount of time that each of 12 students spends each day (on average) on social networks (Facebook, Twitter, etc.) and the Internet for social or entertainment purposes and his or her grade point average (GPA).

Time (hours per day)	4.4	6.2	4.2	1.6	4.7	5.4	1.3	2.1	6.1	3.3	4.4	3.5
GPA	3.22	2.21	3.13	3.69	2.7	2.2	3.69	3.25	2.66	2.89	2.71	3.36

a. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between grade point average and time spent on social networks and the Internet?
b. Find the predictive regression line of GPA on time.
c. Give a brief interpretation of the values of a and b calculated in part b.
d. Plot the predictive regression line on the scatter diagram of part a, and show the errors by drawing vertical lines between scatter points and the predictive regression line.
e. Calculate the predicted GPA for a college student who spends 3.8 hours per day on social networks and the Internet for social or entertainment purposes.
f. Calculate the predicted GPA for a college student who spends 16 hours per day on social networks and the Internet for social or entertainment purposes. Comment on this finding.

13.28 While browsing through the magazine rack at a bookstore, a statistician decides to examine the relationship between the price of a magazine and the percentage of the magazine space that contains advertisements. The data are given in the following table.

Percentage containing ads	37	43	55	49	70	28	65	32
Price ($)	5.50	7.00	4.95	5.75	3.95	9.00	5.50	6.50

a. Construct a scatter diagram for these data. Does the scatter diagram exhibit a linear relationship between the percentage of a magazine's space containing ads and the price of the magazine?

b. Find the estimated regression equation of price on the percentage containing ads.
c. Give a brief interpretation of the values of a and b calculated in part b.
d. Plot the estimated regression line on the scatter diagram of part a, and show the errors by drawing vertical lines between scatter points and the predictive regression line.
e. Predict the price of a magazine with 50% of its space containing ads.
f. Estimate the price of a magazine with 99% of its space containing ads. Comment on this finding.

13.29 The following table gives the total payroll (in millions of dollars) on the opening day of the 2011 season and the percentage of games won during the 2011 season by each of the National League baseball teams.

Team	Total Payroll (millions of dollars)	Percentage of Games Won
Arizona Diamondbacks	53.60	58.0
Atlanta Braves	87.00	54.9
Chicago Cubs	125.50	43.8
Cincinnati Reds	76.20	48.8
Colorado Rockies	88.00	45.1
Houston Astros	70.70	34.6
Los Angeles Dodgers	103.80	50.9
Miami Marlins	56.90	44.4
Milwaukee Brewers	85.50	59.3
New York Mets	120.10	47.5
Philadelphia Phillies	173.00	63.0
Pittsburgh Pirates	46.00	44.4
San Diego Padres	45.90	43.8
San Francisco Giants	118.20	53.1
St. Louis Cardinals	105.40	55.6
Washington Nationals	63.70	49.7

Source: http://baseball.about.com/od/newsrumors/a/2011-Baseball-Team-Payrolls.htm.

a. Find the least squares regression line with total payroll as the independent variable and percentage of games won as the dependent variable.
b. Is the equation of the regression line obtained in part a the population regression line? Why or why not? Do the values of the y-intercept and the slope of the regression line give A and B or a and b?
c. Give a brief interpretation of the values of the y-intercept and the slope obtained in part a.
d. Predict the percentage of games won by a team with a total payroll of $100 million.

13.30 The following table gives the total payroll (in millions of dollars) on the opening day of the 2011 season and the percentage of games won during the 2011 season by each of the American League baseball teams.

Team	Total Payroll (millions of dollars)	Percentage of Games Won
Baltimore Orioles	85.30	42.6
Boston Red Sox	161.40	55.6
Chicago White Sox	129.30	48.8
Cleveland Indians	49.20	49.4
Detroit Tigers	105.70	58.6
Kansas City Royals	36.10	43.8
Los Angeles Angels	139.00	53.1
Minnesota Twins	112.70	38.9
New York Yankees	201.70	59.9
Oakland Athletics	66.60	45.7
Seattle Mariners	86.40	41.4
Tampa Bay Rays	41.90	56.2
Texas Rangers	92.30	59.3
Toronto Blue Jays	62.50	50.0

Source: http://baseball.about.com/od/newsrumors/a/2011-Baseball-Team-Payrolls.htm.

a. Find the least squares regression line with total payroll as the independent variable and percentage of games won as the dependent variable.
b. Is the equation of the regression line obtained in part a the population regression line? Why or why not? Do the values of the y-intercept and the slope of the regression line give A and B or a and b?
c. Give a brief interpretation of the values of the y-intercept and the slope obtained in part a.
d. Predict the percentage of games won by a team with a total payroll of \$100 million.

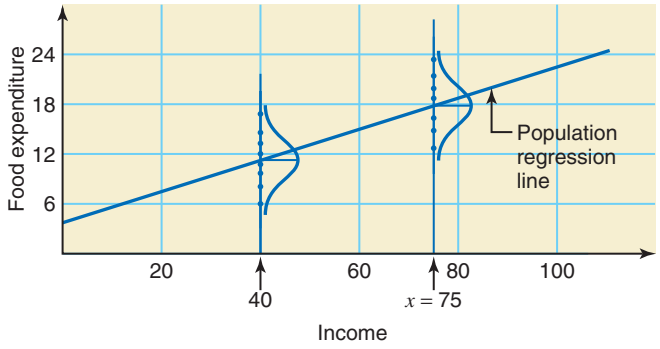
13.2 Standard Deviation of Errors and Coefficient of Determination

In this section we discuss two concepts related to regression analysis. First we discuss the concept of the standard deviation of random errors and its calculation. Then we learn about the concept of the coefficient of determination and its calculation.

13.2.1 Standard Deviation of Errors

When we consider incomes and food expenditures, all households with the same income are expected to spend different amounts on food. Consequently, the random error ϵ will assume different values for these households. The standard deviation σ_ϵ measures the spread of these errors around the population regression line. The **standard deviation of errors** tells us how widely the errors and, hence, the values of y are spread for a given x. In Figure 13.12, which is reproduced as Figure 13.14, the points on the vertical line through $x = 40$ give the monthly food expenditures for all households with a monthly income of \$4000. The distance of each dot from the point on the regression line gives the value of the corresponding error. The standard deviation of errors σ_ϵ measures the spread of such points around the population regression line. The same is true for $x = 75$ or any other value of x.

Figure 13.14 Spread of errors for $x = 40$ and $x = 75$.



Note that σ_ϵ denotes the standard deviation of errors for the population. However, usually σ_ϵ is unknown. In such cases, it is estimated by s_e, which is the standard deviation of errors for the sample data. The following is the basic formula to calculate s_e:

$$s_e = \sqrt{\frac{\text{SSE}}{n-2}} \qquad \text{where} \qquad \text{SSE} = \Sigma(y - \hat{y})^2$$

In this formula, $n - 2$ represents the **degrees of freedom** for the regression model. The reason $df = n - 2$ is that we lose one degree of freedom to calculate $\bar{x}$ and one for $\bar{y}$.

Degrees of Freedom for a Simple Linear Regression Model The *degrees of freedom* for a simple linear regression model are

$$df = n - 2$$

For computational purposes, it is more convenient to use the following formula to calculate the standard deviation of errors s_e.

Standard Deviation of Errors The *standard deviation of errors* is calculated as[3]

$$s_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}}$$

where

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

The calculation of SS_{xy} was discussed earlier in this chapter.[4]

Like the value of SS_{xx}, the value of SS_{yy} is always positive.

Example 13–2 illustrates the calculation of the standard deviation of errors for the data of Table 13.1.

EXAMPLE 13–2

Calculating the standard deviation of errors.

Compute the standard deviation of errors s_e for the data on monthly incomes and food expenditures of the seven households given in Table 13.1.

Solution To compute s_e, we need to know the values of SS_{yy}, SS_{xy}, and b. In Example 13–1, we computed SS_{xy} and b. These values are

$$SS_{xy} = 447.5714 \quad \text{and} \quad b = .2525$$

To compute SS_{yy}, we calculate Σy^2 as shown in Table 13.3.

Table 13.3

Income x	Food Expenditure y	y^2
55	14	196
83	24	576
38	13	169
61	16	256
33	9	81
49	15	225
67	17	289
$\Sigma x = 386$	$\Sigma y = 108$	$\Sigma y^2 = 1792$

The value of SS_{yy} is

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 1792 - \frac{(108)^2}{7} = 125.7143$$

Hence, the standard deviation of errors is

$$s_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}} = \sqrt{\frac{125.7143 - .2525(447.5714)}{7-2}} = \mathbf{1.5939}$$

■

13.2.2 Coefficient of Determination

We may ask the question: How good is the regression model? In other words: How well does the independent variable explain the dependent variable in the regression model? The *coefficient of determination* is one concept that answers this question.

[3]If we have access to population data, the value of σ_ϵ is calculated using the formula

$$\sigma_\epsilon = \sqrt{\frac{SS_{yy} - B\,SS_{xy}}{N}}$$

[4]The basic formula to calculate SS_{yy} is $\Sigma(y - \bar{y})^2$.

For a moment, assume that we possess information only on the food expenditures of households and not on their incomes. Hence, in this case, we cannot use the regression line to predict the food expenditure for any household. As we did in earlier chapters, in the absence of a regression model, we use $\bar{y}$ to estimate or predict every household's food expenditure. Consequently, the error of prediction for each household is now given by $y - \bar{y}$, which is the difference between the actual food expenditure of a household and the mean food expenditure. If we calculate such errors for all households in the sample and then square and add them, the resulting sum is called the **total sum of squares** and is denoted by **SST**. Actually SST is the same as SS_{yy} and is defined as

$$\text{SST} = SS_{yy} = \Sigma(y - \bar{y})^2$$

However, for computational purposes, SST is calculated using the following formula.

Total Sum of Squares (SST) The *total sum of squares*, denoted by SST, is calculated as

$$\text{SST} = \Sigma y^2 - \frac{(\Sigma y)^2}{n}$$

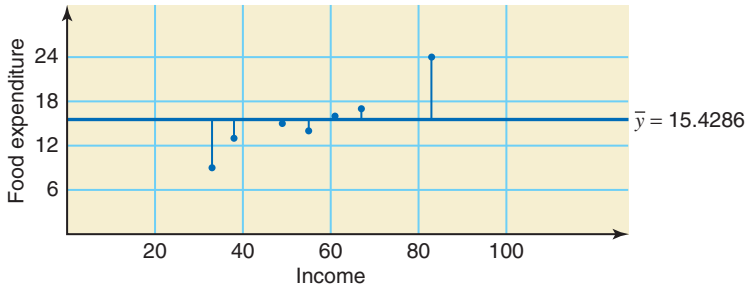
Note that this is the same formula that we used to calculate SS_{yy}.

The value of SS_{yy}, which is 125.7143, was calculated in Example 13–2. Consequently, the value of SST is

$$\text{SST} = 125.7143$$

From Example 13–1, $\bar{y} = 15.4286$. Figure 13.15 shows the error for each of the seven households in our sample using the scatter diagram of Figure 13.4 and using $\bar{y}$.

Figure 13.15 Total errors.



Now suppose we use the simple linear regression model to predict the food expenditure of each of the seven households in our sample. In this case, we predict each household's food expenditure by using the regression line we estimated earlier in Example 13–1, which is

$$\hat{y} = 1.5050 + .2525x$$

The predicted food expenditures, denoted by $\hat{y}$, for the seven households are listed in Table 13.4. Also given are the errors and error squares.

Table 13.4

x	y	$\hat{y} = 1.5050 + .2525x$	$e = y - \hat{y}$	$e^2 = (y - \hat{y})^2$
55	14	15.3925	−1.3925	1.9391
83	24	22.4625	1.5375	2.3639
38	13	11.1000	1.9000	3.6100
61	16	16.9075	−.9075	.8236
33	9	9.8375	−.8375	.7014
49	15	13.8775	1.1225	1.2600
67	17	18.4225	−1.4225	2.0235
				$\Sigma e^2 = \Sigma(y - \hat{y})^2 = 12.7215$

We calculate the values of $\hat{y}$ (given in the third column of Table 13.4) by substituting the values of x in the estimated regression model. For example, the value of x for the first household is 55. Substituting this value of x in the regression equation, we obtain

$$\hat{y} = 1.5050 + .2525(55) = 15.3925$$

Similarly we find the other values of $\hat{y}$. The error sum of squares SSE is given by the sum of the fifth column in Table 13.4. Thus,

$$\text{SSE} = \Sigma(y - \hat{y})^2 = 12.7215$$

The errors of prediction for the regression model for the seven households are shown in Figure 13.16.

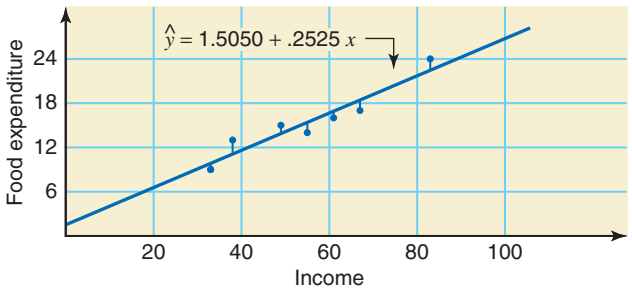


Figure 13.16 Errors of prediction when regression model is used.

Thus, from the foregoing calculations,

$$\text{SST} = 125.7143 \quad \text{and} \quad \text{SSE} = 12.7215$$

These values indicate that the sum of squared errors decreased from 125.7143 to 12.7215 when we used $\hat{y}$ in place of $\bar{y}$ to predict food expenditures. This reduction in squared errors is called the **regression sum of squares** and is denoted by **SSR**. Thus,

$$\text{SSR} = \text{SST} - \text{SSE} = 125.7143 - 12.7215 = 112.9928$$

The value of SSR can also be computed by using the formula

$$\text{SSR} = \Sigma(\hat{y} - \bar{y})^2$$

Regression Sum of Squares (SSR) The *regression sum of squares*, denoted by SSR, is

$$\text{SSR} = \text{SST} - \text{SSE}$$

Thus, SSR is the portion of SST that is explained by the use of the regression model, and SSE is the portion of SST that is not explained by the use of the regression model. The sum of SSR and SSE is always equal to SST. Thus,

$$\text{SST} = \text{SSR} + \text{SSE}$$

The ratio of SSR to SST gives the **coefficient of determination**. The coefficient of determination calculated for population data is denoted by $\boldsymbol{\rho^2}$ (ρ is the Greek letter *rho*), and the one calculated for sample data is denoted by $\boldsymbol{r^2}$. The coefficient of determination gives the proportion of SST that is explained by the use of the regression model. The value of the coefficient of determination always lies in the range zero to one. The coefficient of determination can be calculated by using the formula

$$r^2 = \frac{\text{SSR}}{\text{SST}} \quad \text{or} \quad \frac{\text{SST} - \text{SSE}}{\text{SST}}$$

However, for computational purposes, the following formula is more efficient to use to calculate the coefficient of determination.

Coefficient of Determination The *coefficient of determination*, denoted by r^2, represents the proportion of SST that is explained by the use of the regression model. The computational formula for r^2 is[5]

$$r^2 = \frac{b\,\text{SS}_{xy}}{\text{SS}_{yy}}$$

and

$$0 \le r^2 \le 1$$

Example 13–3 illustrates the calculation of the coefficient of determination for a sample data set.

EXAMPLE 13–3

Calculating the coefficient of determination.

For the data of Table 13.1 on monthly incomes and food expenditures of seven households, calculate the coefficient of determination.

Solution From earlier calculations made in Examples 13–1 and 13–2,

$$b = .2525, \quad \text{SS}_{xy} = 447.5714, \quad \text{and} \quad \text{SS}_{yy} = 125.7143$$

Hence,

$$r^2 = \frac{b\,\text{SS}_{xy}}{\text{SS}_{yy}} = \frac{(.2525)(447.5714)}{125.7143} = .8990 = \mathbf{.90}$$

Thus, we can state that SST is reduced by approximately 90% (from 125.7143 to 12.7215) when we use $\hat{y}$, instead of $\bar{y}$, to predict the food expenditures of households. Note that r^2 is usually rounded to two decimal places. ■

The total sum of squares SST is a measure of the total variation in food expenditures, the regression sum of squares SSR is the portion of total variation explained by the regression model (or by income), and the error sum of squares SSE is the portion of total variation not explained by the regression model. Hence, for Example 13–3 we can state that 90% of the total variation in food expenditures of households occurs because of the variation in their incomes, and the remaining 10% is due to randomness and other variables.

Usually, the higher the value of r^2, the better is the regression model. This is so because if r^2 is larger, a greater portion of the total errors is explained by the included independent variable, and a smaller portion of errors is attributed to other variables and randomness.

EXERCISES

CONCEPTS AND PROCEDURES

13.31 What are the degrees of freedom for a simple linear regression model?

13.32 Explain the meaning of coefficient of determination.

13.33 Explain the meaning of SST and SSR. You may use graphs for illustration purposes.

13.34 A population data set produced the following information.

$$N = 250, \quad \Sigma x = 9680, \quad \Sigma y = 1456, \quad \Sigma xy = 82{,}050,$$
$$\Sigma x^2 = 485{,}870, \quad \text{and} \quad \Sigma y^2 = 140{,}895$$

Find the values of σ_ϵ and ρ^2.

[5]If we have access to population data, the value of ρ^2 is calculated using the formula

$$\rho^2 = \frac{B\,\text{SS}_{xy}}{\text{SS}_{yy}}$$

The values of SS_{xy} and SS_{yy} used here are calculated for the population data set.

13.35 A population data set produced the following information.

$$N = 460, \quad \Sigma x = 3920, \quad \Sigma y = 2650, \quad \Sigma xy = 26{,}570,$$
$$\Sigma x^2 = 48{,}530, \quad \text{and} \quad \Sigma y^2 = 39{,}347$$

Find the values of σ_ϵ and ρ^2.

13.36 The following information is obtained from a sample data set.

$$n = 10, \quad \Sigma x = 100, \quad \Sigma y = 240, \quad \Sigma xy = 3840,$$
$$\Sigma x^2 = 1140, \quad \text{and} \quad \Sigma y^2 = 28{,}400$$

Find the values of s_e and r^2.

13.37 The following information is obtained from a sample data set.

$$n = 15, \quad \Sigma x = 82, \quad \Sigma y = 602, \quad \Sigma xy = 2534,$$
$$\Sigma x^2 = 412, \quad \text{and} \quad \Sigma y^2 = 60{,}123$$

Find the values of s_e and r^2.

APPLICATIONS

13.38 The following table gives information on the calorie count and grams of fat for the 11 types of bagels produced by Panera Bread.

Bagel	Calories	Fat (grams)
Asiago Cheese	330	6.5
Blueberry	330	1.5
Chocolate Chip	370	5.5
Cinnamon Crunch	430	8.0
Cinnamon Swirl & Raisin	320	3.0
Everything	300	2.5
French Toast	350	5.0
Jalapeno & Cheddar	310	3.0
Plain	290	1.0
Sesame	310	3.0
Sweet Onion & Poppyseed	390	7.0

With calories as the dependent variable and fat content as the independent variable, find the following:

a. SS_{xx}, SS_{yy}, and SS_{xy} **b.** Standard deviation of errors
c. SST, SSE, and SSR **d.** Coefficient of determination

13.39 The following table provides information on the speed at takeoff (in meters per second) and distance traveled (in meters) by a random sample of 10 world-class long jumpers.

Speed	8.5	8.8	9.3	8.9	8.2	8.6	8.7	9.0	8.7	9.1
Distance	7.72	7.91	8.33	7.93	7.39	7.65	7.95	8.28	7.86	8.14

With distance traveled as the dependent variable and speed at takeoff as the independent variable, find the following:

a. SS_{xx}, SS_{yy}, and SS_{xy} **b.** Standard deviation of errors
c. SST, SSE, and SSR **d.** Coefficient of determination

13.40 Refer to Exercise 13.25. The following table, which gives the ages (in years) and prices (in hundreds of dollars) of eight cars of a specific model, is reproduced from that exercise.

Age	8	3	6	9	2	5	6	2
Price	38	220	95	33	267	134	112	245

a. Calculate the standard deviation of errors.
b. Compute the coefficient of determination and give a brief interpretation of it.

13.41 The following table, reproduced from Exercise 13.26, gives information on the amount of sugar (in grams) and the calorie count in one serving of a sample of 13 varieties of Kellogg's cereal.

Sugar (grams)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	190	100	120	190	190	110	120

Source: kelloggs.com.

a. Determine the standard deviation of errors.
b. Find the coefficient of determination and give a brief interpretation of it.

13.42 The following table, reproduced from Exercise 13.27, contains information on the amount of time spent each day (on average) on social networks and the Internet for social or entertainment purposes and the grade point average for a random sample of 12 college students.

Time (hours per day)	4.4	6.2	4.2	1.6	4.7	5.4	1.3	2.1	6.1	3.3	4.4	3.5
GPA	3.22	2.21	3.13	3.69	2.7	2.2	3.69	3.25	2.66	2.89	2.71	3.36

a. Calculate the standard deviation of errors.
b. Compute the coefficient of determination, and give a brief interpretation of it. What percentage of the variation in GPA is explained by the least squares regression line of GPA on time? What percentage is not explained?

13.43 The following table, reproduced from Exercise 13.28, lists the percentages of space for eight magazines that contain advertisements and the prices of these magazines.

Percentage containing ads	37	43	55	49	70	28	65	32
Price ($)	5.50	7.00	4.95	5.75	3.95	9.00	5.50	6.50

a. Find the standard deviation of errors.
b. Compute the coefficient of determination. What percentage of the variation in price is explained by the least squares regression of price on the percentage of magazine space containing ads? What percentage of this variation is not explained?

13.44 Refer to data given in Exercise 13.29 on the total 2011 payroll and the percentage of games won during the 2011 season by each of the National League baseball teams.
a. Find the standard deviation of errors, σ_ϵ. (Note that this data set belongs to a population.)
b. Compute the coefficient of determination, ρ^2.

13.45 Refer to data given in Exercise 13.30 on the total 2011 payroll and the percentage of games won during the 2011 season by each of the American League baseball teams.
a. Find the standard deviation of errors, σ_ϵ. (Note that this data set belongs to a population.)
b. Compute the coefficient of determination, ρ^2.

13.3 Inferences About *B*

This section is concerned with estimation and tests of hypotheses about the population regression slope B. We can also make confidence intervals and test hypotheses about the y-intercept A of the population regression line. However, making inferences about A is beyond the scope of this text.

13.3.1 Sampling Distribution of *b*

One of the main purposes for determining a regression line is to find the true value of the slope B of the population regression line. However, in almost all cases, the regression line is estimated using sample data. Then, based on the sample regression line, inferences are made about the population regression line. The slope b of a sample regression line is a point estimator of the slope B of the population regression line. The different sample regression lines estimated for different samples taken from the same population will give different values of b. If only one sample is taken and the regression line for that sample is estimated, the value of b will depend on which elements are included in the sample. Thus, b is a random variable, and it possesses a

probability distribution that is more commonly called its sampling distribution. The shape of the sampling distribution of b, its mean, and standard deviation are given next.

Mean, Standard Deviation, and Sampling Distribution of *b* Because of the assumption of normally distributed random errors, the sampling distribution of b is normal. The mean and standard deviation of b, denoted by μ_b and σ_b, respectively, are

$$\mu_b = B \quad \text{and} \quad \sigma_b = \frac{\sigma_\epsilon}{\sqrt{SS_{xx}}}$$

However, usually the standard deviation of population errors σ_ϵ is not known. Hence, the sample standard deviation of errors s_e is used to estimate σ_ϵ. In such a case, when σ_ϵ is unknown, the standard deviation of b is estimated by s_b, which is calculated as

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

If σ_ϵ is known, the normal distribution can be used to make inferences about B. However, if σ_ϵ is not known, the normal distribution is replaced by the t distribution to make inferences about B.

13.3.2 Estimation of *B*

The value of b obtained from the sample regression line is a point estimate of the slope B of the population regression line. As mentioned in Section 13.3.1, if σ_ϵ is not known, the t distribution is used to make a confidence interval for B.

Confidence Interval for *B* The $(1 - \alpha)100\%$ *confidence interval for* B is given by

$$b \pm ts_b$$

where

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}}$$

and the value of t is obtained from the t distribution table for $\alpha/2$ area in the right tail of the t distribution and $n - 2$ degrees of freedom.

Example 13–4 describes the procedure for making a confidence interval for B.

■ EXAMPLE 13–4

Constructing a confidence interval for B.

Construct a 95% confidence interval for B for the data on incomes and food expenditures of seven households given in Table 13.1.

Solution From the given information and earlier calculations in Examples 13–1 and 13–2,

$$n = 7, \quad b = .2525, \quad SS_{xx} = 1772.8571, \quad \text{and} \quad s_e = 1.5939$$

The confidence level is 95%. We have

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{1.5939}{\sqrt{1772.8571}} = .0379$$

$$df = n - 2 = 7 - 2 = 5$$

$$\alpha/2 = (1 - .95)/2 = .025$$

From the t distribution table, the value of t for 5 df and .025 area in the right tail of the t distribution curve is 2.571. The 95% confidence interval for B is

$$b \pm ts_b = .2525 \pm 2.571(.0379) = .2525 \pm .0974 = \textbf{.155 to .350}$$

Thus, we are 95% confident that the slope B of the population regression line is between .155 and .350. ■

13.3.3 Hypothesis Testing About *B*

Testing a hypothesis about B when the null hypothesis is $B = 0$ (that is, the slope of the regression line is zero) is equivalent to testing that x does not determine y and that the regression line is of no use in predicting y for a given x. However, we should remember that we are testing for a linear relationship between x and y. It is possible that x may determine y nonlinearly. Hence, a nonlinear relationship may exist between x and y.

To test the hypothesis that x does not determine y linearly, we will test the null hypothesis that the slope of the regression line is zero; that is, $B = 0$. The alternative hypothesis can be: (1) x determines y, that is, $B \neq 0$; (2) x determines y positively, that is, $B > 0$; or (3) x determines y negatively, that is, $B < 0$.

The procedure used to make a hypothesis test about B is similar to the one used in earlier chapters. It involves the same five steps. Of course, use can use the p-value approach too.

Test Statistic for *b* The value of the *test statistic t for b* is calculated as

$$t = \frac{b - B}{s_b}$$

The value of B is substituted from the null hypothesis.

Example 13–5 illustrates the procedure for testing a hypothesis about B.

Conducting a test of hypothesis about B.

■ EXAMPLE 13–5

Test at the 1% significance level whether the slope of the regression line for the example on incomes and food expenditures of seven households is positive.

Solution From the given information and earlier calculations in Examples 13–1 and 13–4,

$$n = 7, \quad b = .2525, \quad \text{and} \quad s_b = .0379$$

Step 1. *State the null and alternative hypotheses.*

We are to test whether or not the slope B of the population regression line is positive. Hence, the two hypotheses are

$$H_0: B = 0 \quad \text{(The slope is zero)}$$

$$H_1: B > 0 \quad \text{(The slope is positive)}$$

Note that we can also write the null hypothesis as H_0: $B \leq 0$, which states that the slope is either zero or negative.

Step 2. *Select the distribution to use.*

Here, σ_ϵ is not known. All assumptions for the population regression model are assumed to hold true. Hence, we will use the t distribution to make the test about B.

Step 3. *Determine the rejection and nonrejection regions.*

The significance level is .01. The $>$ sign in the alternative hypothesis indicates that the test is right-tailed. Therefore,

$$\text{Area in the right tail of the } t \text{ distribution} = \alpha = .01$$

$$df = n - 2 = 7 - 2 = 5$$

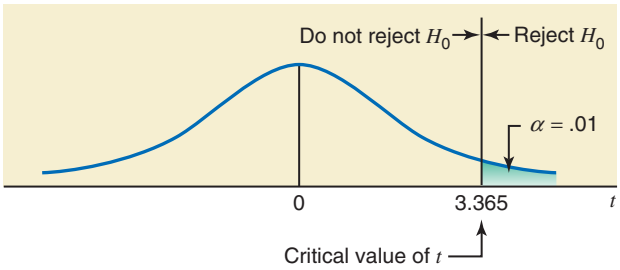


Figure 13.17 Rejection and nonrejection regions.

From the t distribution table, the critical value of t for 5 df and .01 area in the right tail of the t distribution is 3.365, as shown in Figure 13.17.

Step 4. *Calculate the value of the test statistic.*

The value of the test statistic t for b is calculated as follows:

$$t = \frac{b - B}{s_b} = \frac{.2525 - 0}{.0379} = 6.662$$

(From H_0)

Step 5. *Make a decision.*

The value of the test statistic $t = 6.662$ is greater than the critical value of $t = 3.365$, and it falls in the rejection region. Hence, we reject the null hypothesis and conclude that x (income) determines y (food expenditure) positively. That is, food expenditure increases with an increase in income and it decreases with a decrease in income.

Using the *p*-Value to Make a Decision

We can find the range for the p-value (as we did in Chapters 9 and 10) from the t distribution table, Table V of Appendix C, and make a decision by comparing that p-value with the significance level. For this example, $df = 5$, and the observed value of t is 6.662. From Table V (the t distribution table) in the row of $df = 5$, the largest value of t is 5.893 for which the area in the right tail of the t distribution is .001. Since our observed value of $t = 6.662$ is larger than 5.893, the p-value for $t = 6.662$ is less than .001, that is,

$$p\text{-value} < .001$$

Note that if we use technology to find this p-value, we will obtain a p-value of .000. Thus, we can state that for any α equal to or higher than .001 (the upper limit of the p-value range), we will reject the null hypothesis. For our example, $\alpha = .01$, which is larger than the p-value of .001. As a result, we reject the null hypothesis. ■

Note that the null hypothesis does not always have to be $B = 0$. We may test the null hypothesis that B is equal to a certain value. See Exercises 13.47 to 13.50, 13.54, 13.57, and 13.58 for such cases.

A Note on Regression and Causality

The regression line does not prove causality between two variables; that is, it does not predict that a change in y is *caused* by a change in x. The information about causality is based on theory or common sense. A regression line describes only whether or not a significant quantitative relationship between x and y exists. Significant relationship means that we reject the null hypothesis H_0: $B = 0$ at a given significance level. The estimated regression line gives the change in y due to a change of one unit in x. Note that it does not indicate that the reason y has changed is that x has changed. In our example on incomes and food expenditures, it is economic theory and common sense, not the regression line, that tell us that food expenditure depends on income. The regression analysis simply helps determine whether or not this dependence is significant.

EXERCISES

CONCEPTS AND PROCEDURES

13.46 Describe the mean, standard deviation, and shape of the sampling distribution of the slope b of the simple linear regression model.

13.47 The following information is obtained for a sample of 16 observations taken from a population.

$$SS_{xx} = 340.700, \quad s_e = 1.951, \quad \text{and} \quad \hat{y} = 12.45 + 6.32x$$

a. Make a 99% confidence interval for B.
b. Using a significance level of .025, can you conclude that B is positive?
c. Using a significance level of .01, can you conclude that B is different from zero?
d. Using a significance level of .02, test whether B is different from 4.50. (*Hint:* The null hypothesis here will be H_0: $B = 4.50$, and the alternative hypothesis will be H_1: $B \neq 4.50$. Notice that the value of $B = 4.50$ will be used to calculate the value of the test statistic t.)

13.48 The following information is obtained for a sample of 36 observations taken from a population.

$$SS_{xx} = 274.600, \quad s_e = .953, \quad \text{and} \quad \hat{y} = 280.56 - 3.77x$$

a. Make a 95% confidence interval for B.
b. Using a significance level of .01, test whether B is negative.
c. Testing at the 5% significance level, can you conclude that B is different from zero?
d. Test if B is different from -5.20. Use $\alpha = .01$.

13.49 The following information is obtained for a sample of 100 observations taken from a population.

$$SS_{xx} = 524.884 \quad s_e = 1.464, \quad \text{and} \quad \hat{y} = 5.48 + 2.50x$$

a. Make a 98% confidence interval for B.
b. Test at the 2.5% significance level whether B is positive.
c. Can you conclude that B is different from zero? Use $\alpha = .01$.
d. Using a significance level of .01, test whether B is greater than 1.75.

13.50 The following information is obtained for a sample of 80 observations taken from a population.

$$SS_{xx} = 380.592, \quad s_e = .961, \quad \text{and} \quad \hat{y} = 160.24 - 2.70x$$

a. Make a 97% confidence interval for B.
b. Test at the 1% significance level whether B is negative.
c. Can you conclude that B is different from zero? Use $\alpha = .01$.
d. Using a significance level of .025, test whether B is less than -1.25.

APPLICATIONS

13.51 Refer to Exercise 13.25. The data on ages (in years) and prices (in hundreds of dollars) for eight cars of a specific model are reproduced from that exercise.

Age	8	3	6	9	2	5	6	2
Price	38	220	95	33	267	134	112	245

a. Construct a 95% confidence interval for B. You can use results obtained in Exercises 13.25 and 13.40 here.
b. Test at the 5% significance level whether B is negative.

13.52 The data given in the table below are the midterm scores in a course for a sample of 10 students and the scores of student evaluations of the instructor. (In the instructor evaluation scores, 1 is the lowest and 4 is the highest score.)

Instructor score	3	2	3	1	2	4	3	4	4	2
Midterm score	90	75	97	64	47	99	75	88	93	81

a. Find the regression of instructor scores on midterm scores.
b. Construct a 99% confidence interval for B.
c. Test at the 1% significance level whether B is positive.

13.53 The following data give the experience (in years) and monthly salaries (in hundreds of dollars) of nine randomly selected secretaries.

Experience	14	3	5	6	4	9	18	5	16
Monthly salary	62	29	37	43	35	60	67	32	60

a. Find the least squares regression line with experience as an independent variable and monthly salary as a dependent variable.
b. Construct a 98% confidence interval for B.
c. Test at the 2.5% significance level whether B is greater than zero.

13.54 The following table, reproduced from Exercise 13.26, gives information on the amount of sugar (in grams) and the calorie count in one serving of a sample of 13 varieties of Kellogg's cereal.

Sugar (grams)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	180	80	120	190	190	100	120

Source: kelloggs.com.

a. Make a 95% confidence interval for B. You can use the calculations made in Exercises 13.26 and 13.41 here.
b. It is well known that each additional gram of carbohydrate adds 4 calories. Sugar is one type of carbohydrate. Using regression equation for the data in the table, test at the 1% significance level whether B is different from 4.

13.55 Refer to Exercise 13.27. The following table, reproduced from that exercise, contains information on the amount of time spent each day (on average) on social networks and the Internet for social or entertainment purposes and the grade point average for a random sample of 12 college students.

Time (hours per day)	4.4	6.2	4.2	1.6	4.7	5.4	1.3	2.1	6.1	3.3	4.4	3.5
GPA	3.22	2.21	3.13	3.69	2.7	2.2	3.69	3.25	2.66	2.89	2.71	3.36

a. Construct a 98% confidence interval for B. You can use the results obtained in Exercises 13.27 and 13.42.
b. Test at the 1% significance level whether B is negative.

13.56 The following table, reproduced from Exercise 13.28, lists the percentages of space for eight magazines that contain advertisements and the prices of these magazines.

Percentage containing ads	37	43	55	49	70	28	65	32
Price ($)	5.50	7.00	4.95	5.75	3.95	9.00	5.50	6.50

a. Construct a 98% confidence interval for B. You can use the calculations made in Exercises 13.28 and 13.43 here.
b. Testing at the 5% significance level, can you conclude that B is different from zero?

13.57 The following table, reproduced from Exercise 13.38, gives information on the calorie count and grams of fat for the 11 types of bagels produced by Panera Bread.

Bagel	**Calories**	**Fat (grams)**
Asiago Cheese	330	6.0
Blueberry	330	1.5
Chocolate Chip	370	6.0
Cinnamon Crunch	430	8.0
Cinnamon Swirl & Raisin	320	2.5
Everything	300	2.5
French Toast	350	5.0
Jalapeno & Cheddar	310	3.0
Plain	290	1.5
Sesame	310	3.0
Sweet Onion & Poppyseed	390	7.0

a. Find the least squares regression line with calories as the dependent variable and fat content as the independent variable.
b. Make a 95% confidence interval for B. You may use the results obtained in Exercise 13.38.
c. Test at the 5% significance level whether B is different from 14.

13.58 The following table, reproduced from Exercise 13.39, provides information on the speed at takeoff (in meters per second) and distance traveled (in meters) by a random sample of 10 world-class long jumpers.

Speed	8.5	8.8	9.3	8.9	8.2	8.6	8.7	9.0	8.7	9.1
Distance	7.72	7.91	8.33	7.93	7.39	7.65	7.95	8.28	7.86	8.14

a. Find the predictive regression line of the distance traveled on the speed at takeoff.
b. Make a 98% confidence interval for B. You may use the results obtained in Exercise 13.39.
c. Test at the 1% significance level whether B is less than 1.2.

13.4 Linear Correlation

This section describes the meaning and calculation of the linear correlation coefficient and the procedure to conduct a test of hypothesis about it.

13.4.1 Linear Correlation Coefficient

Another measure of the relationship between two variables is the correlation coefficient. This section describes the simple linear correlation, for short **linear correlation**, which measures the strength of the linear association between two variables. In other words, the linear correlation coefficient measures how closely the points in a scatter diagram are spread around the regression line. The correlation coefficient calculated for the population data is denoted by $\boldsymbol{\rho}$ (Greek letter *rho*) and the one calculated for sample data is denoted by $\boldsymbol{r}$. (Note that the square of the correlation coefficient is equal to the coefficient of determination.)

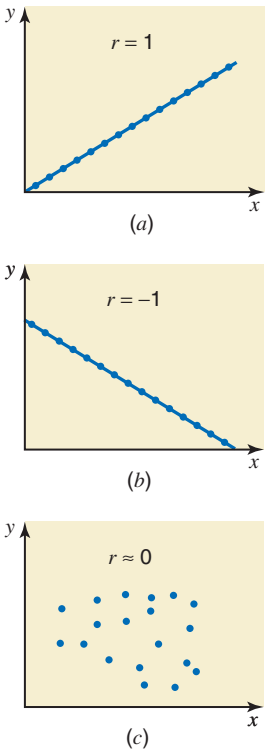


Figure 13.18 Linear correlation between two variables. (*a*) Perfect positive linear correlation, $r = 1$. (*b*) Perfect negative linear correlation, $r = -1$. (*c*) No linear correlation, $r \approx 0$.

Value of the Correlation Coefficient The *value of the correlation coefficient* always lies in the range -1 to 1; that is,

$$-1 \le \rho \le 1 \quad \text{and} \quad -1 \le r \le 1$$

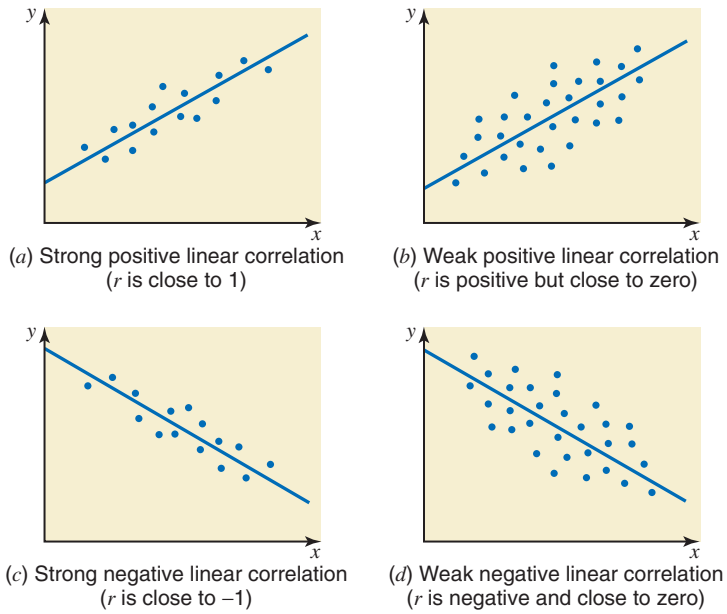
Although we can explain the linear correlation using the population correlation coefficient ρ, we will do so using the sample correlation coefficient r.

If $r = 1$, it is said to be a *perfect positive linear correlation*. In such a case, all points in the scatter diagram lie on a straight line that slopes upward from left to right, as shown in Figure 13.18*a*. If $r = -1$, the correlation is said to be a *perfect negative linear correlation*. In this case, all points in the scatter diagram fall on a straight line that slopes downward from left to right, as shown in Figure 13.18*b*. If the points are scattered all over the diagram, as shown in Figure 13.18*c*, then there is *no linear correlation* between the two variables, and consequently r is close to 0. Note that here r is *not* equal to zero but is very *close* to zero.

We do not usually encounter an example with perfect positive or perfect negative correlation (unless the relationship between variables is exact). What we observe in real-world problems is either a positive linear correlation with $0 < r < 1$ (that is, the correlation coefficient is greater than zero but less than 1) or a negative linear correlation with $-1 < r < 0$ (that is, the correlation coefficient is greater than -1 but less than zero).

If the correlation between two variables is positive and close to 1, we say that the variables have a *strong positive linear correlation*. If the correlation between two variables is positive but close to zero, then the variables have a *weak positive linear correlation*. In contrast, if the correlation between two variables is negative and close to -1, then the variables are said to have a *strong negative linear correlation*. If the correlation between two variables is negative but close to zero, there exists a *weak negative linear correlation* between the variables. Graphically, a strong correlation indicates that the points in the scatter diagram are very close to the regression line,

and a weak correlation indicates that the points in the scatter diagram are widely spread around the regression line. These four cases are shown in Figure 13.19*a–d*.



(*a*) Strong positive linear correlation (*r* is close to 1)

(*b*) Weak positive linear correlation (*r* is positive but close to zero)

(*c*) Strong negative linear correlation (*r* is close to –1)

(*d*) Weak negative linear correlation (*r* is negative and close to zero)

Figure 13.19 Linear correlation between two variables.

The linear correlation coefficient is calculated by using the following formula. (This correlation coefficient is also called the *Pearson product moment correlation coefficient*.)

Linear Correlation Coefficient The *simple linear correlation coefficient*, denoted by r, measures the strength of the linear relationship between two variables for a sample and is calculated as[6]

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\, SS_{yy}}}$$

Because both SS_{xx} and SS_{yy} are always positive, the sign of the correlation coefficient r depends on the sign of SS_{xy}. If SS_{xy} is positive, then r will be positive, and if SS_{xy} is negative, then r will be negative. Another important observation to remember is that *r and b, calculated for the same sample, will always have the same sign. That is, both r and b are either positive or negative*. This is so because both r and b provide information about the relationship between x and y. Likewise, the corresponding population parameters ρ and B will always have the same sign.

Example 13–6 illustrates the calculation of the linear correlation coefficient r.

■ EXAMPLE 13–6

Calculating the linear correlation coefficient.

Calculate the correlation coefficient for the example on incomes and food expenditures of seven households.

Solution From earlier calculations made in Examples 13–1 and 13–2,

$$SS_{xy} = 447.5714, \quad SS_{xx} = 1772.8571, \quad \text{and} \quad SS_{yy} = 125.7143$$

[6]If we have access to population data, the value of ρ is calculated using the formula

$$\rho = \frac{SS_{xy}}{\sqrt{SS_{xx}\, SS_{yy}}}$$

Here the values of SS_{xy}, SS_{xx}, and SS_{yy} are calculated using the population data.

Substituting these values in the formula for r, we obtain

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\,SS_{yy}}} = \frac{447.5714}{\sqrt{(1772.8571)(125.7143)}} = .9481 = \mathbf{.95}$$

Thus, the linear correlation coefficient is .95. The correlation coefficient is usually rounded to two decimal places. ■

The linear correlation coefficient simply tells us how strongly the two variables are (linearly) related. The correlation coefficient of .95 for incomes and food expenditures of seven households indicates that income and food expenditure are very strongly and positively correlated. This correlation coefficient does not, however, provide us with any more information.

The square of the correlation coefficient gives the coefficient of determination, which was explained in Section 13.4. Thus, $(.95)^2$ is .90, which is the value of r^2 calculated in Example 13–3.

Sometimes the calculated value of r may indicate that the two variables are very strongly linearly correlated, but in reality they may not be. For example, if we calculate the correlation coefficient between the price of haircut and the size of families in the United States using data for the last 30 years, we will find a strong negative linear correlation. Over time, the price of haircut has increased and the size of families has decreased. This finding does not mean that family size and the price of haircut are related. As a result, before we calculate the correlation coefficient, we must seek help from a theory or from common sense to postulate whether or not the two variables have a causal relationship.

Another point to note is that in a simple regression model, one of the two variables is categorized as an independent (also known as an explanatory or predictor) variable and the other is classified as a dependent (also known as a response) variable. However, no such distinction is made between the two variables when the correlation coefficient is calculated.

13.4.2 Hypothesis Testing About the Linear Correlation Coefficient

This section describes how to perform a test of hypothesis about the population correlation coefficient ρ using the sample correlation coefficient r. We can use the t distribution to make this test. However, to use the t distribution, both variables should be normally distributed.

Usually (although not always), the null hypothesis is that the linear correlation coefficient between the two variables is zero, that is, $\rho = 0$. The alternative hypothesis can be one of the following: (1) the linear correlation coefficient between the two variables is less than zero, that is, $\rho < 0$; (2) the linear correlation coefficient between the two variables is greater than zero, that is, $\rho > 0$; or (3) the linear correlation coefficient between the two variables is not equal to zero, that is, $\rho \neq 0$.

Test Statistic for *r* If both variables are normally distributed and the null hypothesis is H_0: $\rho = 0$, then the value of the test statistic t is calculated as

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Here $n - 2$ are the degrees of freedom.

Example 13–7 describes the procedure to perform a test of hypothesis about the linear correlation coefficient.

Performing a test of hypothesis about the correlation coefficient.

■ EXAMPLE 13–7

Using a 1% level of significance and the data from Example 13–1, test whether the linear correlation coefficient between incomes and food expenditures is positive. Assume that the populations of both variables are normally distributed.

Solution From Examples 13–1 and 13–6,

$$n = 7 \quad \text{and} \quad r = .9481$$

Below we use the five steps to perform this test of hypothesis.

Step 1. *State the null and alternative hypotheses.*

We are to test whether the linear correlation coefficient between incomes and food expenditures is positive. Hence, the null and alternative hypotheses are, respectively,

H_0: $\rho = 0$ (The linear correlation coefficient is zero.)

H_1: $\rho > 0$ (The linear correlation coefficient is positive.)

Step 2. *Select the distribution to use.*

The population distributions for both variables are normally distributed. Hence, we can use the t distribution to perform this test about the linear correlation coefficient.

Step 3. *Determine the rejection and nonrejection regions.*

The significance level is 1%. From the alternative hypothesis we know that the test is right-tailed. Hence,

$$\text{Area in the right tail of the } t \text{ distribution} = .01$$

$$df = n - 2 = 7 - 2 = 5$$

From the t distribution table, the critical value of t is 3.365. The rejection and nonrejection regions for this test are shown in Figure 13.20.

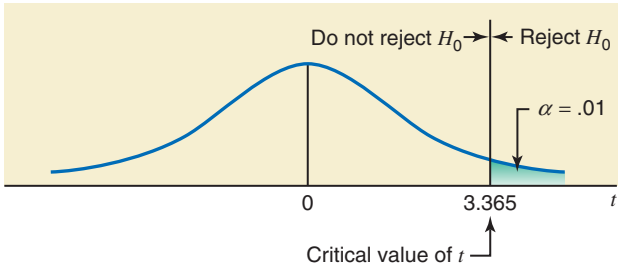


Figure 13.20 Rejection and nonrejection regions.

Step 4. *Calculate the value of the test statistic.*

The value of the test statistic t for r is calculated as follows:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = .9481\sqrt{\frac{7-2}{1-(.9481)^2}} = 6.667$$

Step 5. *Make a decision.*

The value of the test statistic $t = 6.667$ is greater than the critical value of $t = 3.365$, and it falls in the rejection region. Hence, we reject the null hypothesis and conclude that there is a positive linear relationship between incomes and food expenditures.

Using the *p*-Value to Make a Decision

We can find the range for the p-value from the t distribution table (Table V of Appendix C) and make a decision by comparing that p-value with the significance level. For this example, $df = 5$, and the observed value of t is 6.667. From Table V (the t distribution table) in the row of $df = 5$, the largest value of t is 5.893, for which the area in the right tail of the t distribution is .001. Since our observed value of $t = 6.667$ is larger than 5.893, the p-value for $t = 6.803$ is less than .001, that is,

$$p\text{-value} < .001$$

Thus, we can state that for any α equal to or greater than .001 (the upper limit of the p-value range), we will reject the null hypothesis. For our example, $\alpha = .01$, which is greater than the p-value of .001. As a result, we reject the null hypothesis. ■

EXERCISES

CONCEPTS AND PROCEDURES

13.59 What does a linear correlation coefficient tell about the relationship between two variables? Within what range can a correlation coefficient assume a value?

13.60 What is the difference between ρ and r? Explain.

13.61 Explain each of the following concepts. You may use graphs to illustrate each concept.

a. Perfect positive linear correlation
b. Perfect negative linear correlation
c. Strong positive linear correlation
d. Strong negative linear correlation
e. Weak positive linear correlation
f. Weak negative linear correlation
g. No linear correlation

13.62 Can the values of B and ρ calculated for the same population data have different signs? Explain.

13.63 For a sample data set, the linear correlation coefficient r has a positive value. Which of the following is true about the slope b of the regression line estimated for the same sample data?

a. The value of b will be positive.
b. The value of b will be negative.
c. The value of b can be positive or negative.

13.64 For a sample data set, the slope b of the regression line has a negative value. Which of the following is true about the linear correlation coefficient r calculated for the same sample data?

a. The value of r will be positive.
b. The value of r will be negative.
c. The value of r can be positive or negative.

13.65 For a sample data set on two variables, the value of the linear correlation coefficient is (close to) zero. Does this mean that these variables are not related? Explain.

13.66 Will you expect a positive, zero, or negative linear correlation between the two variables for each of the following examples?

a. Grade of a student and hours spent playing video games.
b. Incomes and entertainment expenditures of households
c. Ages of women and makeup expenses per month
d. Price of a computer and consumption of Coca-Cola
e. Price and consumption of wine

13.67 Will you expect a positive, zero, or negative linear correlation between the two variables for each of the following examples?

a. SAT scores and GPAs of students
b. Stress level and blood pressure of individuals
c. Amount of fertilizer used and yield of corn per acre
d. Ages and prices of houses
e. Heights of husbands and incomes of their wives

13.68 A population data set produced the following information.

$$N = 250, \quad \Sigma x = 9880, \quad \Sigma y = 1456, \quad \Sigma xy = 90{,}980,$$
$$\Sigma x^2 = 485{,}870, \quad \text{and} \quad \Sigma y^2 = 140{,}875$$

Find the linear correlation coefficient ρ.

13.69 A population data set produced the following information.

$$N = 460, \quad \Sigma x = 3920, \quad \Sigma y = 2650, \quad \Sigma xy = 26{,}570,$$
$$\Sigma x^2 = 48{,}530, \quad \text{and} \quad \Sigma y^2 = 39{,}347$$

Find the linear correlation coefficient ρ.

13.70 A sample data set produced the following information.

$$n = 10, \quad \Sigma x = 100, \quad \Sigma y = 240, \quad \Sigma xy = 3860,$$
$$\Sigma x^2 = 1140, \quad \text{and} \quad \Sigma y^2 = 30{,}400$$

a. Calculate the linear correlation coefficient r.
b. Using a 2% significance level, can you conclude that ρ is different from zero?

13.71 A sample data set produced the following information.

$$n = 12, \quad \Sigma x = 66, \quad \Sigma y = 588, \quad \Sigma xy = 2244,$$
$$\Sigma x^2 = 396, \quad \text{and} \quad \Sigma y^2 = 58{,}734$$

a. Calculate the linear correlation coefficient r.
b. Using a 1% significance level, can you conclude that ρ is negative?

APPLICATIONS

13.72 Refer to Exercise 13.25. The data on ages (in years) and prices (in hundreds of dollars) for eight cars of a specific model are reproduced from that exercise.

Age	8	3	6	9	2	5	6	2
Price	38	220	95	33	267	134	112	245

a. Do you expect the ages and prices of cars to be positively or negatively related? Explain.
b. Calculate the linear correlation coefficient.
c. Test at the 2.5% significance level whether ρ is negative.

13.73 The following table, reproduced from Exercise 13.53, gives the experience (in years) and monthly salaries (in hundreds of dollars) of nine randomly selected secretaries.

Experience	14	3	5	6	4	9	18	5	16
Monthly salary	62	29	37	43	35	60	67	32	60

a. Do you expect the experience and monthly salaries to be positively or negatively related? Explain.
b. Compute the linear correlation coefficient.
c. Test at a 5% significance level whether ρ is positive.

13.74 The following table lists the midterm and final exam scores for seven students in a statistics class.

Midterm score	79	95	81	66	87	94	59
Final exam score	85	97	78	76	94	84	67

a. Do you expect the midterm and final exam scores to be positively or negatively related?
b. Plot a scatter diagram. By looking at the scatter diagram, do you expect the correlation coefficient between these two variables to be close to zero, 1, or -1?
c. Find the correlation coefficient. Is the value of r consistent with what you expected in parts a and b?
d. Using a 1% significance level, test whether the linear correlation coefficient is positive.

13.75 The following data give the ages (in years) of husbands and wives for six couples.

Husband's age	43	57	28	19	35	39
Wife's age	37	51	32	20	33	38

a. Do you expect the ages of husbands and wives to be positively or negatively related?
b. Plot a scatter diagram. By looking at the scatter diagram, do you expect the correlation coefficient between these two variables to be close to zero, 1, or -1?
c. Find the correlation coefficient. Is the value of r consistent with what you expected in parts a and b?
d. Using a 5% significance level, test whether the correlation coefficient is different from zero.

13.76 The following table, reproduced from Exercise 13.26, gives information on the amount of sugar (in grams) and the calorie count in one serving of a sample of 13 varieties of Kellogg's cereal.

Sugar (grams)	4	15	12	11	8	6	7	2	7	14	20	3	13
Calories	120	200	140	110	120	80	180	80	120	190	190	100	120

Source: kelloggs.com.

a. Find the correlation coefficient. Is its sign the same as that of b calculated in Exercise 13.26?
b. Test at a 1% significance level whether the linear correlation coefficient between the two variables listed in the table is positive.

13.77 The following table, reproduced from Exercises 13.38 and 13.57, gives information on the calorie count and grams of fat for 11 types of bagels produced by Panera Bread.

Bagel	Calories	Fat (grams)
Asiago Cheese	330	6.0
Blueberry	330	1.5
Chocolate Chip	370	6.0
Cinnamon Crunch	430	8.0
Cinnamon Swirl & Raisin	320	2.5
Everything	300	2.5
French Toast	350	5.0
Jalapeno & Cheddar	310	3.0
Plain	290	1.5
Sesame	310	3.0
Sweet Onion & Poppyseed	390	7.0

a. Find the correlation coefficient. Is the sign of the correlation coefficient the same as that of b calculated in Exercise 13.57?

b. Test at a 1% significance level whether ρ is different from zero.

13.78 Refer to data given in Exercise 13.29 on the total 2011 payroll and the percentage of games won during the 2011 season by each of the National League baseball teams. Compute the linear correlation coefficient, ρ. Does it make sense to make a confidence interval and to test a hypothesis about ρ here? Explain.

13.79 Refer to data given in Exercise 13.30 on the total 2011 payroll and the percentage of games won during the 2011 season by each of the American League baseball teams. Compute the linear correlation coefficient, ρ. Does it make sense to make a confidence interval and to test a hypothesis about ρ here? Explain.

13.5 Regression Analysis: A Complete Example

This section works out an example that includes all the topics we have discussed so far in this chapter.

EXAMPLE 13–8

A complete example of regression analysis.

A random sample of eight drivers selected from a small town insured with a company and having similar minimum required auto insurance policies was selected. The following table lists their driving experiences (in years) and monthly auto insurance premiums (in dollars):

Driving Experience (years)	Monthly Auto Insurance Premium ($)
5	64
2	87
12	50
9	71
15	44
6	56
25	42
16	60

(a) Does the insurance premium depend on the driving experience, or does the driving experience depend on the insurance premium? Do you expect a positive or a negative relationship between these two variables?

(b) Compute SS_{xx}, SS_{yy}, and SS_{xy}.

(c) Find the least squares regression line by choosing appropriate dependent and independent variables based on your answer in part a.

(d) Interpret the meaning of the values of a and b calculated in part c.

(e) Plot the scatter diagram and the regression line.

(f) Calculate r and r^2, and explain what they mean.

(g) Predict the monthly auto insurance premium for a driver with 10 years of driving experience.

(h) Compute the standard deviation of errors.

(i) Construct a 90% confidence interval for B.

(j) Test at a 5% significance level whether B is negative.

(k) Using $\alpha = .05$, test whether ρ is different from zero.

Solution

(a) Based on theory and intuition, we expect the insurance premium to depend on driving experience. Consequently, the insurance premium is a dependent variable (variable y) and driving experience is an independent variable (variable x) in the regression model. A new driver is considered a high risk by the insurance companies, and he or she has to pay a higher premium for auto insurance. On average, the insurance premium is expected to decrease with an increase in the years of driving experience. Therefore, we expect a negative relationship between these two variables. In other words, both the population correlation coefficient ρ and the population regression slope B are expected to be negative.

(b) Table 13.5 shows the calculation of Σx, Σy, Σxy, Σx^2, and Σy^2.

Table 13.5

Experience x	Premium y	xy	x^2	y^2
5	64	320	25	4096
2	87	174	4	7569
12	50	600	144	2500
9	71	639	81	5041
15	44	660	225	1936
6	56	336	36	3136
25	42	1050	625	1764
16	60	960	256	3600
$\Sigma x = 90$	$\Sigma y = 474$	$\Sigma xy = 4739$	$\Sigma x^2 = 1396$	$\Sigma y^2 = 29{,}642$

The values of $\bar{x}$ and $\bar{y}$ are

$$\bar{x} = \Sigma x/n = 90/8 = 11.25$$

$$\bar{y} = \Sigma y/n = 474/8 = 59.25$$

The values of SS_{xy}, SS_{xx}, and SS_{yy} are computed as follows:

$$SS_{xy} = \Sigma xy - \frac{(\Sigma x)(\Sigma y)}{n} = 4739 - \frac{(90)(474)}{8} = \mathbf{-593.5000}$$

$$SS_{xx} = \Sigma x^2 - \frac{(\Sigma x)^2}{n} = 1396 - \frac{(90)^2}{8} = \mathbf{383.5000}$$

$$SS_{yy} = \Sigma y^2 - \frac{(\Sigma y)^2}{n} = 29{,}642 - \frac{(474)^2}{8} = \mathbf{1557.5000}$$

(c) To find the regression line, we calculate a and b as follows:

$$b = \frac{SS_{xy}}{SS_{xx}} = \frac{-593.5000}{383.5000} = -1.5476$$

$$a = \bar{y} - b\bar{x} = 59.25 - (-1.5476)(11.25) = 76.6605$$

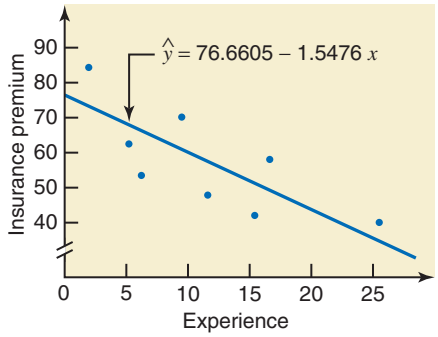
Thus, our estimated regression line $\hat{y} = a + bx$ is

$$\hat{y} = 76.6605 - 1.5476x$$

(d) The value of $a = 76.6605$ gives the value of $\hat{y}$ for $x = 0$; that is, it gives the monthly auto insurance premium for a driver with no driving experience. However, as mentioned earlier in this chapter, we should not attach much importance to this statement because the sample contains drivers with only 2 or more years of experience. The value of b gives the change in $\hat{y}$ due to a change of one unit in x. Thus, $b = -1.5476$ indicates that, on average, for every extra year of driving experience, the monthly auto insurance premium decreases by \$1.55. Note that when b is negative, y decreases as x increases.

(e) Figure 13.21 shows the scatter diagram and the regression line for the data on eight auto drivers. Note that the regression line slopes downward from left to right. This result is consistent with the negative relationship we anticipated between driving experience and insurance premium.

Figure 13.21 Scatter diagram and the regression line.



(f) The values of r and r^2 are computed as follows:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}\,SS_{yy}}} = \frac{-593.5000}{\sqrt{(383.5000)(1557.5000)}} = -.7679 = \mathbf{-.77}$$

$$r^2 = \frac{b\,SS_{xy}}{SS_{yy}} = \frac{(-1.5476)(-593.5000)}{1557.5000} = .5897 = \mathbf{.59}$$

The value of $r = -.77$ indicates that the driving experience and the monthly auto insurance premium are negatively related. The (linear) relationship is strong but not very strong. The value of $r^2 = .59$ states that 59% of the total variation in insurance premiums is explained by years of driving experience, and 41% is not. The low value of r^2 indicates that there may be many other important variables that contribute to the determination of auto insurance premiums. For example, the premium is expected to depend on the driving record of a driver and the type and age of the car.

(g) Using the estimated regression line, we find the predicted value of y for $x = 10$ as:

$$\hat{y} = 76.6605 - 1.5476x = 76.6605 - 1.5476(10) = \mathbf{\$61.18}$$

Thus, we expect the monthly auto insurance premium of a driver with 10 years of driving experience to be \$61.18.

(h) The standard deviation of errors is

$$s_e = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}} = \sqrt{\frac{1557.5000 - (-1.5476)(-593.5000)}{8-2}} = \mathbf{10.3199}$$

(i) To construct a 90% confidence interval for B, first we calculate the standard deviation of b:

$$s_b = \frac{s_e}{\sqrt{SS_{xx}}} = \frac{10.3199}{\sqrt{383.5000}} = .5270$$

For a 90% confidence level, the area in each tail of the t distribution is

$$\alpha/2 = (1 - .90)/2 = .05$$

The degrees of freedom are

$$df = n - 2 = 8 - 2 = 6$$

From the t distribution table, the t value for .05 area in the right tail of the t distribution and 6 df is 1.943. The 90% confidence interval for B is

$$b \pm ts_b = -1.5476 \pm 1.943(.5270)$$
$$= -1.5476 \pm 1.0240 = \mathbf{-2.57 \text{ to } -.52}$$

Thus, we can state with 90% confidence that B lies in the interval -2.57 to $-.52$. That is, on average, the monthly auto insurance premium of a driver decreases by an amount between \$.52 and \$2.57 for every extra year of driving experience.

(j) We perform the following five steps to test the hypothesis about B.

Step 1. *State the null and alternative hypotheses.*

The null and alternative hypotheses are, respectively,

$$H_0: B = 0 \quad (B \text{ is not negative.})$$
$$H_1: B < 0 \quad (B \text{ is negative.})$$

Note that the null hypothesis can also be written as $H_0: B \geq 0$.

Step 2. *Select the distribution to use.*

Because σ_ϵ is not known, we use the t distribution to make the hypothesis test.

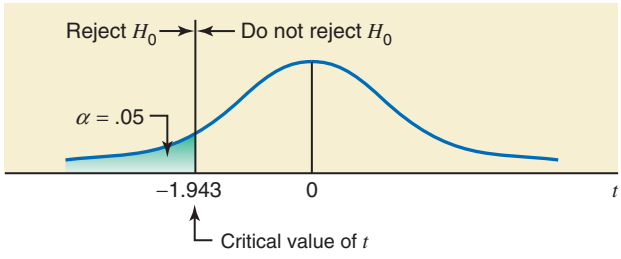
Step 3. *Determine the rejection and nonrejection regions.*

The significance level is .05. The $<$ sign in the alternative hypothesis indicates that it is a left-tailed test.

$$\text{Area in the left tail of the } t \text{ distribution} = \alpha = .05$$
$$df = n - 2 = 8 - 2 = 6$$

From the t distribution table, the critical value of t for .05 area in the left tail of the t distribution and 6 df is -1.943, as shown in Figure 13.22.

Figure 13.22 Rejection and nonrejection regions.



Step 4. *Calculate the value of the test statistic.*

The value of the test statistic t for b is calculated as follows:

$$t = \frac{b - B}{s_b} = \frac{-1.5476 - 0}{.5270} = -2.937$$

(From H_0)

Step 5. *Make a decision.*

The value of the test statistic $t = -2.937$ falls in the rejection region. Hence, we reject the null hypothesis and conclude that B is negative. That is, the monthly auto insurance premium decreases with an increase in years of driving experience.

Using the *p*-Value to Make a Decision

We can find the range for the p-value from the t distribution table (Table V of Appendix C) and make a decision by comparing that p-value with the significance level. For this example, $df = 6$ and the observed value of t is -2.937. From Table V (the t distribution table) in the row of $df = 6$, 2.937 is between 2.447 and 3.143. The corresponding areas in the right tail of the t distribution are .025 and .01, respectively. Our test is left-tailed, however, and the observed value of t is negative. Thus, $t = -2.937$ lies between -2.447 and -3.143. The corresponding areas in the left tail of the t distribution are .025 and .01. Therefore the range of the p-value is

$$.01 < p\text{-value} < .025$$

Thus, we can state that for any α equal to or greater than .025 (the upper limit of the p-value range), we will reject the null hypothesis. For our example, $\alpha = .05$, which is greater than the upper limit of the p-value of .025. As a result, we reject the null hypothesis.

Note that if we use technology to find this p-value, we will obtain a p-value of .013. Then we can reject the null hypothesis for any $\alpha \geq .013$.

(k) We perform the following five steps to test the hypothesis about the linear correlation coefficient ρ.

Step 1. *State the null and alternative hypotheses.*

The null and alternative hypotheses are, respectively,

H_0: $\rho = 0$ (The linear correlation coefficient is zero.)

H_1: $\rho \neq 0$ (The linear correlation coefficient is different from zero.)

Step 2. *Select the distribution to use.*

Assuming that variables x and y are normally distributed, we will use the t distribution to perform this test about the linear correlation coefficient.

Step 3. *Determine the rejection and nonrejection regions.*

The significance level is 5%. From the alternative hypothesis we know that the test is two-tailed. Hence,

$$\text{Area in each tail of the } t \text{ distribution} = .05/2 = .025$$

$$df = n - 2 = 8 - 2 = 6$$

From the t distribution table, Table V of Appendix C, the critical values of t are -2.447 and 2.447. The rejection and nonrejection regions for this test are shown in Figure 13.23.

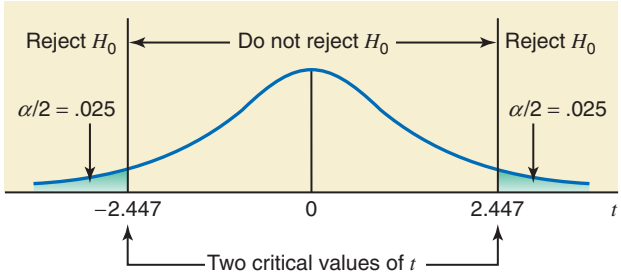


Figure 13.23 Rejection and nonrejection regions.

Step 4. *Calculate the value of the test statistic.*

The value of the test statistic t for r is calculated as follows:

$$t = r\sqrt{\frac{n-2}{1-r^2}} = (-.7679)\sqrt{\frac{8-2}{1-(-.7679)^2}} = -2.936$$

Step 5. *Make a decision.*

The value of the test statistic $t = -2.936$ falls in the rejection region. Hence, we reject the null hypothesis and conclude that the linear correlation coefficient between driving experience and auto insurance premium is different from zero.

Using the *p*-Value to Make a Decision

We can find the range for the p-value from the t distribution table and make a decision by comparing that p-value with the significance level. For this example, $df = 6$ and the observed value of t is -2.936. From Table V (the t distribution table) in the row of $df = 6$, $t = 2.936$ is between 2.447 and 3.143. The corresponding areas in the right tail of the t distribution curve are .025 and .01, respectively. Since the test is two tailed, the range of the p-value is

$$2(.01) < p\text{-value} < 2(.025) \quad \textbf{or} \quad .02 < p\text{-value} < .05$$

Thus, we can state that for any α equal to or greater than .05 (the upper limit of the p-value range), we will reject the null hypothesis. For our example, $\alpha = .05$, which is equal to the upper limit of the p-value. As a result, we reject the null hypothesis. ■

EXERCISES

APPLICATIONS

13.80 The owner of a small factory that produces working gloves is concerned about the high cost of air conditioning in the summer but is afraid that keeping the temperature in the factory too high will lower productivity. During the summer, he experiments with temperature settings from 68ºF to 81ºF and measures each day's productivity. The following table gives the temperature and the number of pairs of gloves (in hundreds) produced on each of the 8 randomly selected days.

Temperature (ºF)	72	71	78	75	81	77	68	76
Pairs of gloves	37	37	32	36	33	35	39	34

a. Do the pairs of gloves produced depend on temperature, or does temperature depend on pairs of gloves produced? Do you expect a positive or a negative relationship between these two variables?
b. Taking temperature as an independent variable and pairs of gloves produced as a dependent variable, compute SS_{xx}, SS_{yy}, and SS_{xy}.
c. Find the least squares regression line.
d. Interpret the meaning of the values of a and b calculated in part c.
e. Plot the scatter diagram and the regression line.
f. Calculate r and r^2, and explain what they mean.
g. Compute the standard deviation of errors.
h. Predict the number of pairs of gloves produced when $x = 74$.
i. Construct a 99% confidence interval for B.
j. Test at a 5% significance level whether B is negative.
k. Using $\alpha = .01$ can you conclude that ρ is negative?

13.81 The following table gives information on the limited tread warranties (in thousands of miles) and the prices of 12 randomly selected tires at a national tire retailer as of July 2012.

Warranty (thousands of miles)	60	70	75	50	80	55	65	65	70	65	60	65
Price per tire ($)	95	135	94	90	121	70	140	80	92	125	160	155

a. Taking warranty length as an independent variable and price per tire as a dependent variable, compute SS_{xx}, SS_{yy}, and SS_{xy}.

b. Find the regression of price per tire on warranty length.
c. Briefly explain the meaning of the values of a and b calculated in part b.
d. Calculate r and r^2 and explain what they mean.
e. Plot the scatter diagram and the regression line.
f. Predict the price of a tire with a warranty length of 73,000 miles.
g. Compute the standard deviation of errors.
h. Construct a 95% confidence interval for B.
i. Test at a 5% significance level if B is positive.
j. Using $\alpha = .025$, can you conclude that the linear correlation coefficient is positive?

13.82 The recommended air pressure in a basketball is between 7 and 9 pounds per square inch (psi). When dropped from a height of 6 feet, a properly inflated basketball should bounce upward between 52 and 56 inches (http://www.bestsoccerbuys.com/balls-basketball.html). The basketball coach at a local high school purchased 10 new basketballs for the upcoming season, inflated the balls to pressures between 7 and 9 psi, and performed the *bounce test* mentioned above. The data obtained are given in the following table.

Pressure (psi)	7.8	8.1	8.3	7.4	8.9	7.2	8.6	7.5	8.1	8.5
Bounce height (inches)	54.1	54.3	55.2	53.3	55.4	52.2	55.7	54.6	54.8	55.3

a. With the pressure as an independent variable and bounce height as a dependent variable, compute SS_{xx}, SS_{yy}, and SS_{xy}.
b. Find the least squares regression line.
c. Interpret the meaning of the values of a and b calculated in part b.
d. Calculate r and r^2 and explain what they mean.
e. Compute the standard deviation of errors.
f. Predict the bounce height of a basketball for $x = 8.0$.
g. Construct a 98% confidence interval for B.
h. Test at a 5% significance level whether B is different from zero.
i. Using $\alpha = .05$, can you conclude that ρ is different from zero?

13.83 The following table gives information on the incomes (in thousands of dollars) and charitable contributions (in hundreds of dollars) for the last year for a random sample of 10 households.

Income	Charitable Contributions
76	15
57	4
140	42
97	33
75	5
107	32
65	10
77	18
102	28
53	4

a. With income as an independent variable and charitable contributions as a dependent variable, compute SS_{xx}, SS_{yy}, and SS_{xy}.
b. Find the regression of charitable contributions on income.
c. Briefly explain the meaning of the values of a and b.
d. Calculate r and r^2 and briefly explain what they mean.
e. Compute the standard deviation of errors.
f. Construct a 99% confidence interval for B.
g. Test at a 1% significance level whether B is positive.
h. Using a 1% significance level, can you conclude that the linear correlation coefficient is different from zero?

13.84 The following data give information on the average ticket prices (in U.S. dollars) and the average percentage of capacity filled for seven hockey teams during the 2011–2012 National Hockey League regular season. (Note: Capacity levels exceeding 100.0% imply standing-room-only attendees.)

Team	Anaheim	Vancouver	Dallas	Edmonton	New Jersey	Toronto	Philadelphia
Average ticket price ($)	36.94	68.38	29.95	70.13	45.86	123.27	66.89
Percentage capacity filled	86.4	102.5	76.8	100.0	87.4	103.7	107.4

Source: http://espn.go.com/blog/dallas/stars/post/_/id/13315/stars-have-cheapest-ticket-in-nhl and http://espn.go.com/nhl/attendance.

a. Taking average ticket price as an independent variable and percentage of capacity filled as a dependent variable, compute SS_{xx}, SS_{yy}, and SS_{xy}.
b. Find the least squares regression line.
c. Briefly explain the meaning of the values of a and b calculated in part b.
d. Calculate r and r^2 and briefly explain what they mean.
e. Compute the standard deviation of errors.
f. Construct a 95% confidence interval for B.
g. Test at a 2.5% significance level whether B is positive.
h. Using a 2.5% significance level, test whether ρ is positive.

13.85 The following table gives information on GPAs and starting salaries (rounded to the nearest thousand dollars) of seven recent college graduates.

GPA	2.90	3.81	3.20	2.42	3.94	2.05	2.25
Starting salary	48	53	50	37	65	32	37

a. With GPA as an independent variable and starting salary as a dependent variable, compute SS_{xx}, SS_{yy}, and SS_{xy}.
b. Find the least squares regression line.
c. Interpret the meaning of the values of a and b calculated in part b.
d. Calculate r and r^2 and briefly explain what they mean.
e. Compute the standard deviation of errors.
f. Construct a 95% confidence interval for B.
g. Test at a 1% significance level whether B is different from zero.
h. Test at a 1% significance level whether ρ is positive.

13.6 Using the Regression Model

Let us return to the example on incomes and food expenditures to discuss two major uses of a regression model:

1. Estimating the mean value of y for a given value of x. For instance, we can use our food expenditure regression model to estimate the mean food expenditure of all households with a specific income (say, $5500 per month).
2. Predicting a particular value of y for a given value of x. For instance, we can determine the expected food expenditure of a randomly selected household with a particular monthly income (say, $5500) using our food expenditure regression model.

13.6.1 Using the Regression Model for Estimating the Mean Value of y

Our population regression model is

$$y = A + Bx + \epsilon$$

As mentioned earlier in this chapter, the mean value of y for a given x is denoted by $\mu_{y|x}$, read as "the mean value of y for a given value of x." Because of the assumption that the mean value of ϵ is zero, the mean value of y is given by

$$\mu_{y|x} = A + Bx$$