CHAPTER 11

Linear Regression and Correlation

11.1 Introduction and Abstract of Research Study

The modeling of the relationship between a response variable and a set of explanatory variables is one of the most widely used of all statistical techniques. We refer to this type of modeling as regression analysis. A regression model provides the user with a functional relationship between the response variable and explanatory variables that allows the user to determine which of the explanatory variables have an effect on the response. The regression model allows the user to explore what happens to the response variable for specified changes in the explanatory variables. For example, financial officers must predict future cash flows based on specified values of interest rates, raw material costs, salary increases, and so on. When designing new training programs for employees, a company would want to study the relationship between employee efficiency and explanatory variables such as the results from employment tests, experience on similar jobs, educational background, and previous training. Medical researchers attempt to determine the factors which have an effect on cardiorespiratory fitness. Forest scientists study the relationship between the volume of wood in a tree to the diameter of the tree at a specified heights and the taper of the tree.

The basic idea of regression analysis is to obtain a model for the functional relationship between a **response variable** (often referred to as the dependent

variable) and one or more **explanatory variables** (often referred to as the independent variables). Regression models have a number of uses.

1. The model provides a description of the major features of the data set. In some cases, a subset of the explanatory variables will not affect the response variable and hence the researcher will not have to measure or control any of these variables in future studies. This may result in significant savings in future studies or experiments.
2. The equation relating the response variable to the explanatory variables produced from the regression analysis provides estimates of the response variable for values of the explanatory not observed in the study. For example, a clinical trial is designed to study the response of a subject to various dose levels of a new drug. Because of time and budgetary constraints, only a limited number of dose levels are used in the study. The regression equation will provide estimates of the subjects' response for dose levels not included in the study. The accuracy of these estimates will depend heavily on how well the final model fits the observed data.
3. In business applications, the prediction of future sales of a product is crucial to production planning. If the data provide a model that has a good fit in relating current sales to sales in previous months, prediction of sales in future months is possible. However, a crucial element in the accuracy of these predictions is that the business conditions during which model building data were collected remains fairly stable over the months for which the predictions are desired.
4. In some applications of regression analysis, the researcher is seeking a model which can accurately estimate the values of a variable that is difficult or expensive to measure using explanatory variables that are inexpensive to measure and obtain. If such a model is obtained, then in future applications it is possible to avoid having to obtain the values of the expensive variable by measuring the values of the inexpensive variables and using the regression equation to estimate the value of the expensive variable. For example, a physical fitness center wants to determine the physical well-being of its new clients. Maximal oxygen uptake is recognized as the single best measure of cardiorespiratory fitness but its measurement is expensive. Therefore, the director of the fitness center would want a model that provides accurate estimates of maximal oxygen uptake using easily measured variables such as weight, age, heart rate after 1-mile walk, time needed to walk 1 mile, and so on.

prediction versus explanation

We can distinguish between prediction (reference to future values) and explanation (reference to current or past values). Because of the virtues of hindsight, explanation is easier than prediction. However, it is often clearer to use the term *prediction* to include both cases. Therefore, in this book, we sometimes blur the distinction between prediction and explanation.

For prediction (or explanation) to make much sense, there must be some connection between the variable we're predicting (the dependent variable) and the variable we're using to make the prediction (the independent variable). No doubt, if you tried long enough, you could find 30 common stocks whose price changes over a year have been accurately predicted by the won–lost percentage of the 30 major league baseball teams on the fourth of July. However, such a prediction is absurd because there is no connection between the two variables. Prediction

unit of association

requires a **unit of association;** there should be an entity that relates the two variables. With time-series data, the unit of association may simply be time. The variables may be measured at the same time period or, for genuine prediction, the independent variable may be measured at a time period before the dependent variable. For cross-sectional data, an economic or physical entity should connect the variables. If we are trying to predict the change in market share of various soft drinks, we should consider the promotional activity for those drinks, not the advertising for various brands of spaghetti sauce. The need for a unit of association seems obvious, but many predictions are made for situations in which no such unit is evident.

simple regression

In this chapter, we consider simple linear regression analysis, in which there is a single independent variable and the equation for predicting a dependent variable y is a linear function of a given independent variable x. Suppose, for example, that the director of a county highway department wants to predict the cost of a resurfacing contract that is up for bids. We could reasonably predict the costs to be a function of the road miles to be resurfaced. A reasonable first attempt is to use a linear production function. Let y = total cost of a project in thousands of dollars, x = number of miles to be resurfaced, and $\hat{y}$ = the predicted cost, also in thousands of dollars. A prediction equation $\hat{y} = 2.0 + 3.0x$ (for example) is a linear equation. The constant term, such as the 2.0, is the **intercept** term and is interpreted as the predicted value of y when $x = 0$. In the road resurfacing example, we may interpret the intercept as the fixed cost of beginning the project. The coefficient of x, such as the 3.0, is the **slope** of the line, the predicted change in y when there is a one-unit change in x. In the road resurfacing example, if two projects differed by 1 mile in length, we would predict that the longer project cost 3 (thousand dollars) more than the shorter one. In general, we write the prediction equation as

intercept

slope

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

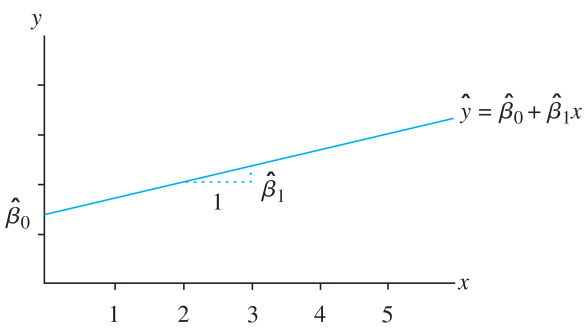
where $\hat{\beta}_0$ is the intercept and $\hat{\beta}_1$ is the slope. See Figure 11.1.

The basic idea of simple linear regression is to use data to fit a prediction line that relates a dependent variable y and a single independent variable x. The first assumption in simple regression is that the relation is, in fact, linear. According to the **assumption of linearity,** the slope of the equation does not change as x changes. In the road resurfacing example, we would assume that there were no (substantial) economies or diseconomies from projects of longer mileage. There is little point in using simple linear regression unless the linearity assumption makes sense (at least roughly).

assumption of linearity

Linearity is not always a reasonable assumption, on its face. For example, if we tried to predict y = number of drivers that are aware of a car dealer's midsummer

FIGURE 11.1
Linear prediction function



sale using x = number of repetitions of the dealer's radio commercial, the assumption of linearity means that the first broadcast of the commercial leads to no greater an increase in aware drivers than the thousand-and-first. (You've heard commercials like that.) We strongly doubt that such an assumption is valid over a wide range of x values. It makes far more sense to us that the effect of repetition would diminish as the number of repetitions got larger, so a straight-line prediction wouldn't work well.

Assuming linearity, we would like to write y as a linear function of x: $y = \beta_0 + \beta_1 x$. However, according to such an equation, y is an exact linear function of x; no room is left for the inevitable errors (deviation of actual y values from their predicted values). Therefore, corresponding to each y we introduce a **random error term** ε_i and assume the model

random error term

$$y = \beta_0 + \beta_1 x + \varepsilon$$

We assume the random variable y to be made up of a predictable part (a linear function of x) and an unpredictable part (the random error ε_i). The coefficients β_0 and β_1 are interpreted as the true, underlying intercept and slope. The error term ε includes the effects of all other factors, known or unknown. In the road resurfacing project, unpredictable factors such as strikes, weather conditions, and equipment breakdowns would contribute to ε, as would factors such as hilliness or prerepair condition of the road—factors that might have been used in prediction but were not. The combined effects of unpredictable and ignored factors yield the random error terms ε.

For example, one way to predict the gas mileage of various new cars (the dependent variable) based on their curb weight (the independent variable) would be to assign each car to a different driver, say, for a 1-month period. What unpredictable and ignored factors might contribute to prediction error? Unpredictable (random) factors in this study would include the driving habits and skills of the drivers, the type of driving done (city versus highway), and the number of stoplights encountered. Factors that would be ignored in a regression analysis of mileage and weight would include engine size and type of transmission (manual versus automatic).

In regression studies, the values of the independent variable (the x_i values) are usually taken as predetermined constants, so the only source of randomness is the ε_i terms. Although most economic and business applications have fixed x_i values, this is not always the case. For example, suppose that x_i is the score of an applicant on an aptitude test and y_i is the productivity of the applicant. If the data are based on a random sample of applicants, x_i (as well as y_i) is a random variable. The question of fixed versus random in regard to x is not crucial for regression studies. If the x_is are random, we can simply regard all probability statements as conditional on the observed x_is.

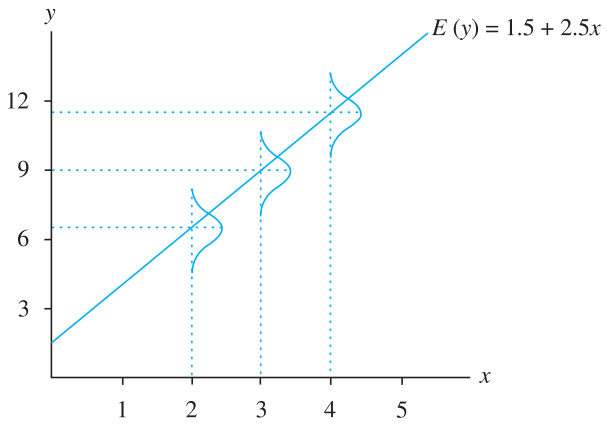
When we assume that the x_is are constants, the only random portion of the model for y_i is the random error term ε_i. We make the following formal assumptions.

DEFINITION 11.1

Formal assumptions of regression analysis:

1. The relation is, in fact, linear, so that the errors all have expected value zero: $E(\varepsilon_i) = 0$ for all i.
2. The errors all have the same variance: $\text{Var}(\varepsilon_i) = \sigma^2$ for all i.
3. The errors are independent of each other.
4. The errors are all normally distributed; ε_i is normally distributed for all i.

FIGURE 11.2
Theoretical distribution of y in regression



These assumptions are illustrated in Figure 11.2. The actual values of the dependent variable are distributed normally, with mean values falling on the regression line and the same standard deviation at all values of the independent variable. The only assumption not shown in the figure is independence from one measurement to another.

These are the formal assumptions, made in order to derive the significance tests and prediction methods that follow. We can begin to check these assumptions

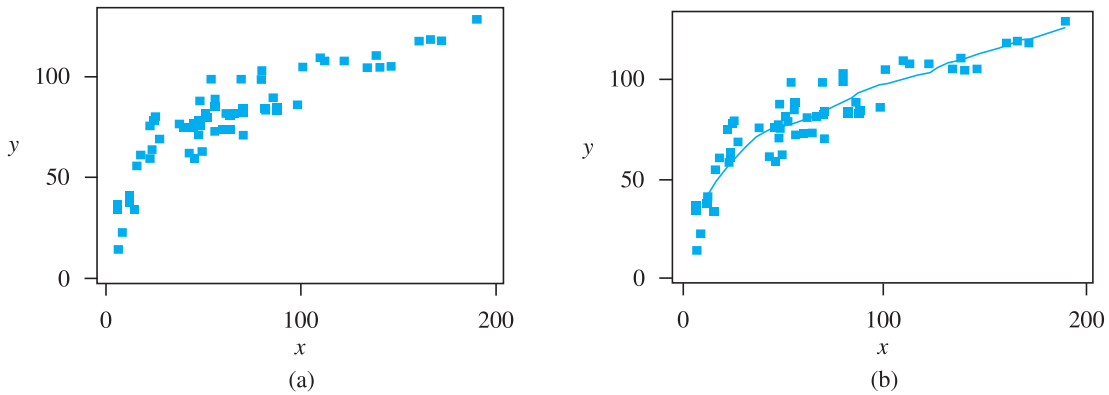
scatterplot

by looking at a **scatterplot** of the data. This is simply a plot of each (x, y) point, with the independent variable value on the horizontal axis, and the dependent variable value measured on the vertical axis. Look to see whether the points basically fall around a straight line or whether there is a definite curve in the pattern. Also look to see whether there are any evident outliers falling far from the general pattern of the data. A scatterplot is shown in part (a) of Figure 11.3.

smoothers

Recently, **smoothers** have been developed to sketch a curve through data without necessarily assuming any particular model. If such a smoother yields something close to a straight line, then linear regression is reasonable. One such method is called LOWESS (locally weighted scatterplot smoother). Roughly, a smoother takes a relatively narrow "slice" of data along the x axis, calculates

FIGURE 11.3 (a) Scatterplot and (b) LOWESS curve



a line that fits the data in that slice, moves the slice slightly along the x axis, recalculates the line, and so on. Then all the little lines are connected in a smooth curve. The width of the slice is called the *bandwidth*; this may often be controlled in the computer program that does the smoothing. The plain scatterplot (Figure 11.3a) is shown again (Figure 11.3b) with a LOWESS curve through it. The scatterplot shows a curved relation; the LOWESS curve confirms that impression.

spline fit

Another type of scatterplot smoother is the **spline fit.** It can be understood as taking a narrow slice of data, fitting a curve (often a cubic equation) to the slice, moving to the next slice, fitting another curve, and so on. The curves are calculated in such a way as to form a connected, continuous curve.

Many economic relations are not linear. For example, any diminishing returns pattern will tend to yield a relation that increases, but at a decreasing rate. If the scatterplot does not appear linear, by itself or when fitted with a LOWESS curve, it can often be "straightened out" by a **transformation** of either the independent variable or the dependent variable. A good statistical computer package or a spreadsheet program will compute such functions as the square root of each value of a variable. The transformed variable should be thought of as simply another variable.

transformation

For example, a large city dispatches crews each spring to patch potholes in its streets. Records are kept of the number of crews dispatched each day and the number of potholes filled that day. A scatterplot of the number of potholes patched and the number of crews and the same scatterplot with a LOWESS curve through it are shown in Figure 11.4. The relation is not linear. Even without the LOWESS curve, the decreasing slope is obvious. That's not surprising; as the city sends out more crews, they will be using less effective workers, the crews will have to travel farther to find holes, and so on. All these reasons suggest that diminishing returns will occur.

We can try several transformations of the independent variable to find a scatterplot in which the points more nearly fall along a straight line. Three common transformations are square root, natural logarithm, and inverse (one divided by the variable). We applied each of these transformations to the pothole repair data. The results are shown in Figure 11.5a–c, with LOWESS curves. The square root (a) and inverse transformations (c) didn't really give us a straight line. The

FIGURE 11.4 Scatterplots for pothole data

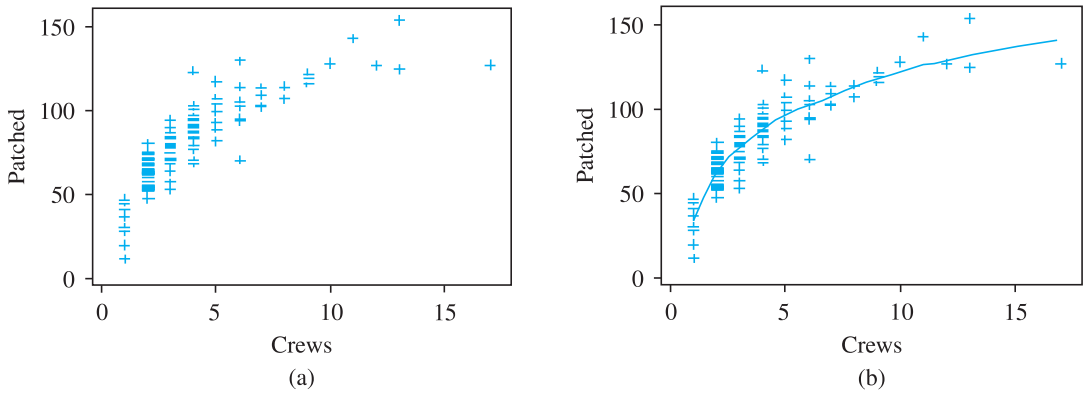
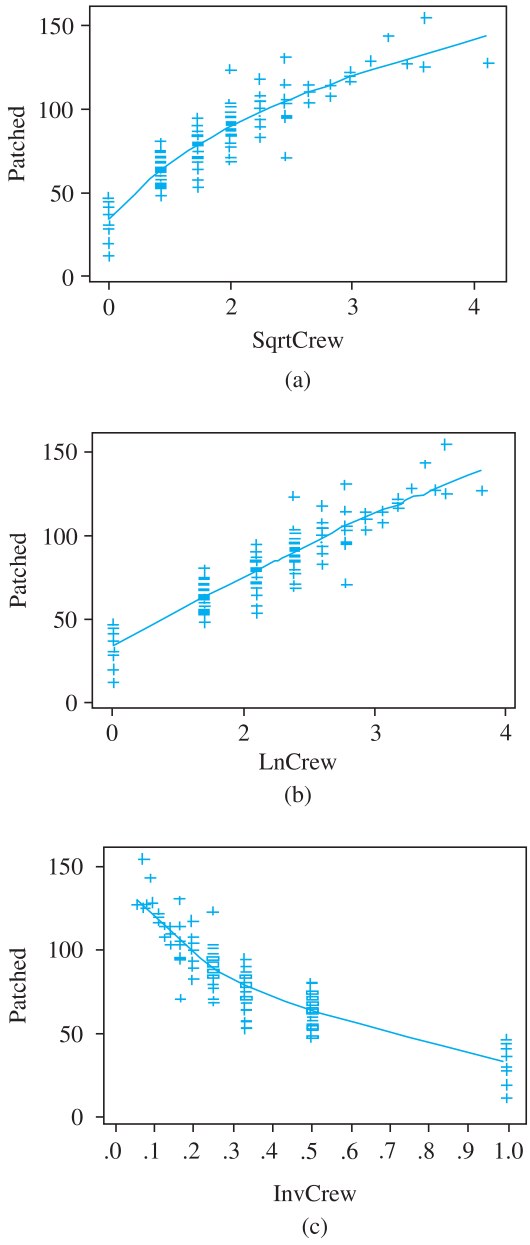


FIGURE 11.5
Scatterplots with transformed predictor



natural logarithm (b) worked very well, however. Therefore, we would use LnCrew as our independent variable.

Finding a good transformation often requires trial and error. Following are some suggestions to try for transformations. Note that there are *two* key features to look for in a scatterplot. First, is the relation nonlinear? Second, is there a pattern of increasing variability along the y (vertical) axis? If there is, the assumption of constant variance is questionable. These suggestions don't cover all the possibilities, but do include the most common problems.

DEFINITION 11.2

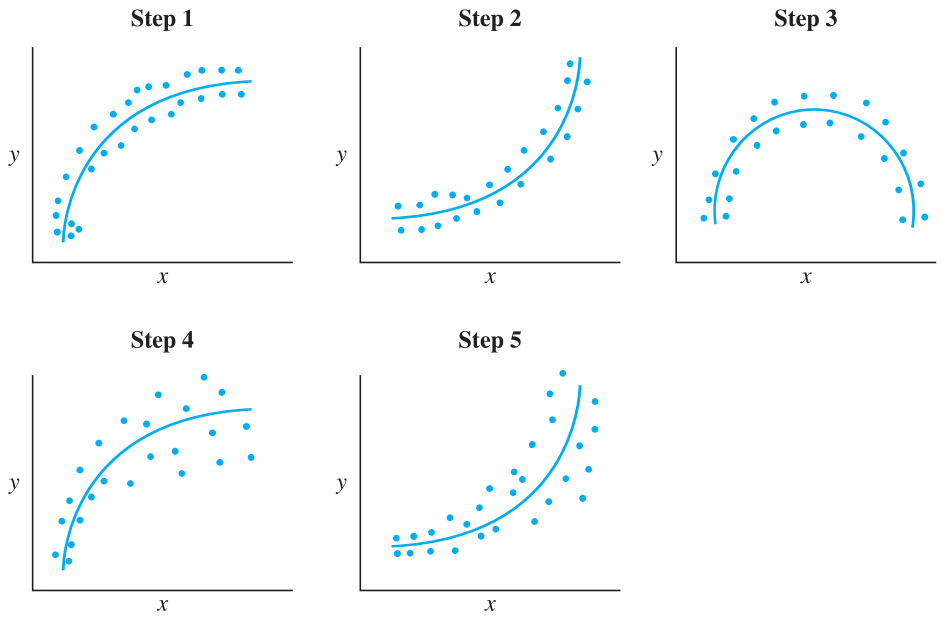
Steps for choosing a transformation:

1. If the plot indicates a relation that is increasing but at a decreasing rate, and if variability around the curve is roughly constant, transform x using square root, logarithm, or inverse transformations.
2. If the plot indicates a relation that is increasing at an increasing rate, and if variability is roughly constant, try using both x and x^2 as predictors. Because this method uses two variables, the multiple regression methods of the next two chapters are needed.
3. If the plot indicates a relation that increases to a maximum and then decreases, and if variability around the curve is roughly constant, again try using both x and x^2 as predictors.
4. If the plot indicates a relation that is increasing at a decreasing rate, and if variability around the curve increases as the predicted y value increases, try using y^2 as the dependent variable.
5. If the plot indicates a relation that is increasing at an increasing rate, and if variability around the curve increases as the predicted y value increases, try using $\ln(y)$ as the dependent variable. It sometimes may also be helpful to use $\ln(x)$ as the independent variable. Note that a change in a natural logarithm corresponds quite closely to a percentage change in the original variable. Thus, the slope of a transformed variable can be interpreted quite well as a percentage change.

The plots in Figure 11.6 correspond to the descriptions given in Definition 11.2.

There are symmetric recommendations for the situations where the relation is decreasing at a decreasing rate, use Step 1 or Step 4 transformations or if the relation is decreasing at an increasing rate use Step 2 or Step 5 transformations.

FIGURE 11.6
Plots corresponding to steps in Definition 11.2



EXAMPLE 11.1

An airline has seen a very large increase in the number of free flights used by participants in its frequent flyer program. To try to predict the trend in these flights in the near future, the director of the program assembled data for the last 72 months. The dependent variable y is the number of thousands of free flights; the independent variable x is month number. A scatterplot with a LOWESS smoother, done using Minitab, is shown in Figure 11.7. What transformation is suggested?

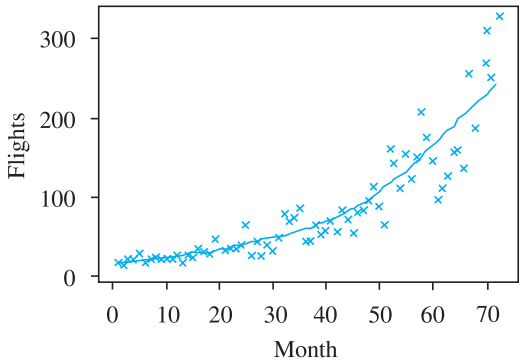


FIGURE 11.7
Frequent flyer free flights by month

Solution The pattern shows flights increasing at an increasing rate. The LOWESS curve is definitely turning upward. In addition, variation (up and down) around the curve is increasing. The points around the high end of the curve (on the right, in this case) scatter much more than the ones around the low end of the curve. The increasing variability suggests transforming the y variable. A natural logarithm (ln) transformation often works well. Minitab computed the logarithms and replotted the data, as shown in Figure 11.8. The pattern is much closer to a straight line, and the scatter around the line is much closer to constant.

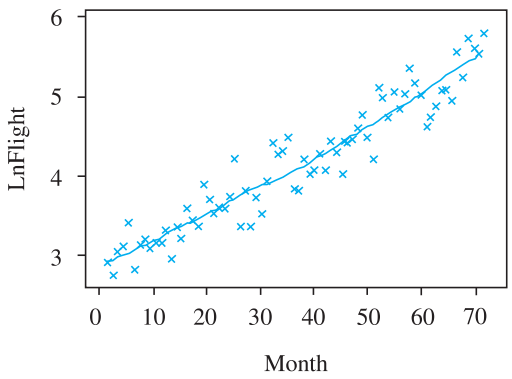


FIGURE 11.8
Result of logarithm transformation

We will have more to say about checking assumptions in Chapter 12. For a simple regression with a single predictor, careful checking of a scatterplot, ideally with a smooth curve fit through it, will help avoid serious blunders.

Once we have decided on any mathematical transformations, we must estimate the actual equation of the regression line. In practice, only sample data are available. The population intercept, slope, and error variance all have to be estimated from limited sample data. The assumptions we made in this section allow us to make inferences about the true parameter values from the sample data.

Abstract of Research Study: Two Methods for Detecting *E. coli*

The case study in Chapter 7 described a new microbial method for the detection of *E. coli,* Petrifilm HEC test. The researcher wanted to evaluate the agreement of the results obtained using the HEC test with results obtained from an elaborate laboratory-based procedure, hydrophobic grid membrane filtration (HGMF). The HEC test is easier to inoculate, more compact to incubate, and safer to handle than conventional procedures. However, prior to using the HEC procedure it was necessary to compare the readings from the HEC test to readings from the HGMF procedure obtained on the same meat sample to determine whether the two procedures were yielding the same readings. If the readings differed but an equation could be obtained that could closely relate the HEC reading to the HGMF reading, then the researchers could calibrate the HEC readings to predict what readings would have been obtained using the HGMF test procedure. If the HEC test results were unrelated to the HGMF test procedure results, then the HEC test could not be used in the field in detecting *E. coli*. The necessary regression analysis to answer these questions will be given at the end of this chapter.

11.2 Estimating Model Parameters

The intercept β_0 and slope β_1 in the regression model

$$y = \beta_0 + \beta_1 x + \varepsilon$$

are population quantities. We must estimate these values from sample data. The error variance σ_ε^2 is another population parameter that must be estimated. The first regression problem is to obtain estimates of the slope, intercept, and variance: we discuss how to do so in this section.

The road resurfacing example of Section 11.1 is a convenient illustration. Suppose the following data for similar resurfacing projects in the recent past are available. Note that we do have a unit of association: The connection between a particular cost and mileage is that they're based on the same project.

Cost y_i (in thousands of dollars):	6.0	14.0	10.0	14.0	26.0
Mileage x_i (in miles):	1.0	3.0	4.0	5.0	7.0

A first step in examining the relation between y and x is to plot the data as a scatterplot. Remember that each point in such a plot represents the (x, y) coordinates of one data entry, as in Figure 11.9. The plot makes it clear that there is

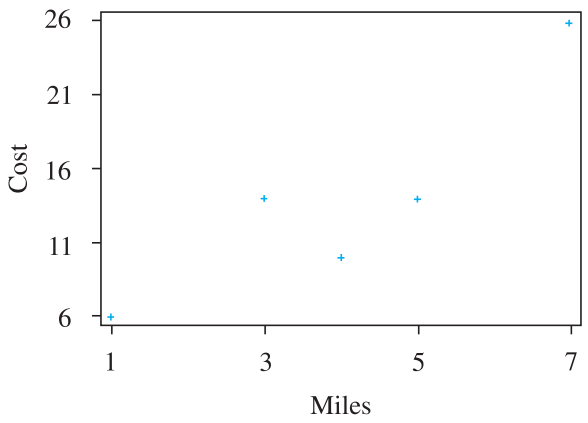
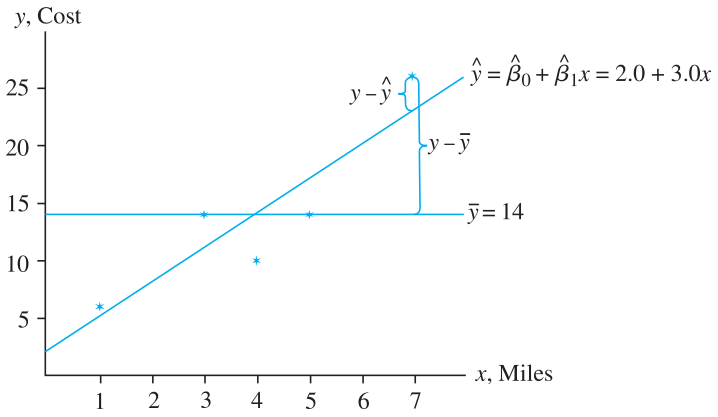


FIGURE 11.9
Scatterplot of cost versus mileage

FIGURE 11.10
Deviations from the least-squares line from the mean



an imperfect but generally increasing relation between x and y. A straight-line relation appears plausible; there is no evident transformation with such limited data.

The regression analysis problem is to find the best straight-line prediction. The most common criterion for "best" is based on squared prediction error. We find the equation of the prediction line—that is, the slope $\hat{\beta}_1$ and intercept $\hat{\beta}_0$ that minimize the total squared prediction error. The method that accomplishes this goal is called the **least-squares method** because it chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the quantity.

least-squares method

$$\sum_i (y_i - \hat{y}_i)^2 = \sum_i [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

The prediction errors are shown on the plot of Figure 11.10 as vertical deviations from the line. The deviations are taken as vertical distances because we're trying to predict y values, and errors should be taken in the y direction. For these data, the least-squares line can be shown to be $\hat{y} = 2.0 + 3.0x$; one of the deviations from it is indicated by the smaller brace. For comparison, the mean $\bar{y} = 14.0$ is also shown; deviation from the mean is indicated by the larger brace. The least-squares principle leads to some fairly long computations for the slope and intercept. Usually, these computations are done by computer.

DEFINITION 11.3

The **least-squares estimates of slope and intercept** are obtained as follows:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where

$$S_{xy} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) \quad \text{and} \quad S_{xx} = \sum_i (x_i - \bar{x})^2$$

Thus, S_{xy} is the sum of x deviations times y deviations and S_{xx} is the sum of x deviations squared.

For the road resurfacing data, $n = 5$ and

$$\sum x_i = 1.0 + \cdots + 7.0 = 20.0$$

so $\bar{x} = \dfrac{20.0}{5} = 4.0$. Similarly,

$$\sum y_i = 70.0, \bar{y} = \frac{70.0}{5} = 14.0$$

Also,

$$\begin{aligned} S_{xx} &= \sum (x_i - \bar{x})^2 \\ &= (1.0 - 4.0)^2 + \cdots + (7.0 - 4.0)^2 \\ &= 20.00 \end{aligned}$$

and

$$\begin{aligned} S_{xy} &= \sum (x_i - \bar{x})(y_i - \bar{y}) \\ &= (1.0 - 4.0)(6.0 - 14.0) + \cdots + (7.0 - 4.0)(26.0 - 14.0) \\ &= 60.0 \end{aligned}$$

Thus,

$$\hat{\beta}_1 = \frac{60.0}{20.0} = 3.0 \quad \text{and} \quad \hat{\beta}_0 = 14.0 - (3.0)(4.0) = 2.0$$

From the value $\hat{\beta}_1 = 3$, we can conclude that the estimated average increase in cost for each additional mile is \$3,000.

EXAMPLE 11.2

Data from a sample of 10 pharmacies are used to examine the relation between prescription sales volume and the percentage of prescription ingredients purchased directly from the supplier. The sample data are shown in Table 11.1.

TABLE 11.1
Data for Example 11.2

Pharmacy	Sales Volume, y (in \$1,000)	% of Ingredients Purchased Directly, x
1	25	10
2	55	18
3	50	25
4	75	40
5	110	50
6	138	63
7	90	42
8	60	30
9	10	5
10	100	55

a. Find the least-squares estimates for the regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
b. Predict sales volume for a pharmacy that purchases 15% of its prescription ingredients directly from the supplier.
c. Plot the (x, y) data and the prediction equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.
d. Interpret the value of $\hat{\beta}_1$ in the context of the problem.

Solution

a. The equation can be calculated by virtually any statistical computer package; for example, here is abbreviated Minitab output:

```
MTB > Regress 'Sales' on 1 variable 'Directly'

The regression equation is
Sales = 4.70 + 1.97 Directly

Predictor     Coef     Stdev   t-ratio       p
Constant     4.698     5.952      0.79   0.453
Directly    1.9705    0.1545     12.75   0.000
```

To see how the computer does the calculations, you can obtain the least-squares estimates from Table 11.2.

TABLE 11.2
Calculations for obtaining least-squares estimates

	y	x	$y-\bar{y}$	$x-\bar{x}$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
	25	10	−46.3	−23.8	1,101.94	566.44
	55	18	−16.3	−15.8	257.54	249.64
	50	25	−21.3	−8.8	187.44	77.44
	75	40	3.7	6.2	22.94	38.44
	110	50	38.7	16.2	626.94	262.44
	138	63	66.7	29.2	1,947.64	852.64
	90	42	18.7	8.2	153.34	67.24
	60	30	−11.3	−3.8	42.94	14.44
	10	5	−61.3	−28.8	1,765.44	829.44
	100	55	28.7	21.2	608.44	449.44
Totals	713	338	0	0	6,714.60	3,407.60
Means	71.3	33.8				

$$S_{xx} = \sum (x - \bar{x})^2 = 3{,}407.6$$

$$S_{xy} = \sum (x - \bar{x})(y - \bar{y}) = 6{,}714.6$$

Substituting into the formulas for $\hat{\beta}_0$ and $\hat{\beta}_1$,

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{6{,}714.6}{3{,}407.6} = 1.9704778 \qquad \text{rounded to } 1.97$$

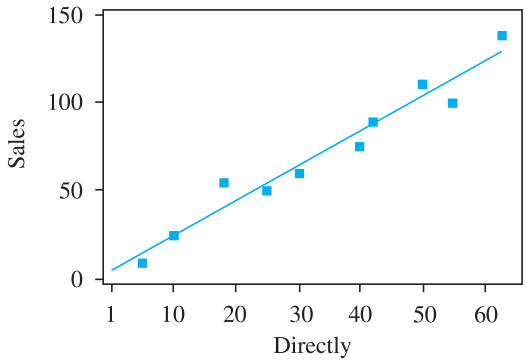
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 71.3 - 1.9704778(33.8) = 4.6978519 \qquad \text{rounded to } 4.70$$

b. When $x = 15\%$, the predicted sales volume is $\hat{y} = 4.70 + 1.97(15) = 34.25$ (that is, \$34,250).

c. The (x, y) data and prediction equation are shown in Figure 11.11.

d. From $\hat{\beta}_1 = 1.97$, we conclude that if a pharmacy would increase by 1% the percentage of ingredients purchased directly, then the estimated increase in average sales volume would be \$1,970.

FIGURE 11.11
Sample data and least-squares prediction equation



EXAMPLE 11.3

In Chapter 3 we discussed a study which related the crime rate in a major city to the number of casino employees in that city. The study was attempting to associate an increase in crime rate with increasing levels of casino gambling which is reflected in the number of people employed in the gambling industry. Use the information in Table 3.17 on page 107 to calculate the least-squares estimates of the intercept and slope of the line relating crime rate to number of casino employees. Use the following Minitab output to confirm your calculations.

Solution From Table 3.17 on page 107, we have the following summary statistics for y crime rate (number of crimes per 1000 population) and x the number of casino employees (in thousands):

$$\bar{x} = \frac{318}{10} = 31.80, \quad \bar{y} = \frac{27.85}{10} = 2.785,$$

$$S_{xx} = 485.60, \quad S_{yy} = 7.3641, \quad S_{xy} = 55.810$$

Thus,

$$\hat{\beta}_1 = \frac{55.810}{485.60} = .11493 \quad \text{and} \quad \hat{\beta}_0 = 2.785 - (.11493)(31.80) = -.8698$$

The Minitab output is given here

```
The regression equation is
CrimeRate = -0.870 + 0.115 Employees


Predictor      Coef  SE Coef       T      P
Constant   -0.8698   0.5090   -1.71  0.126
Employees  0.11493  0.01564    7.35  0.000


S = 0.344566  R-Sq = 87.1%  R-Sq(adj) = 85.5%


Analysis of Variance

Source          DF       SS      MS      F      P
Regression       1   6.4142  6.4142  54.03  0.000
Residual Error   8   0.9498  0.1187
Total            9   7.3641
```

From the previous output, the values calculated are the same as the values from Minitab. We would interpret the value of the estimated slope $\hat{\beta}_1 = .11493$ as follows. For an increase of 1,000 employees in the casino industry, the average crime rate would increase .115. It is important to note that these types of social relationships are much more complex than this simple relationship. Also, it would be a major mistake to place much credence in this type of conclusion because of all the other factors that may have an effect on the crime rate.

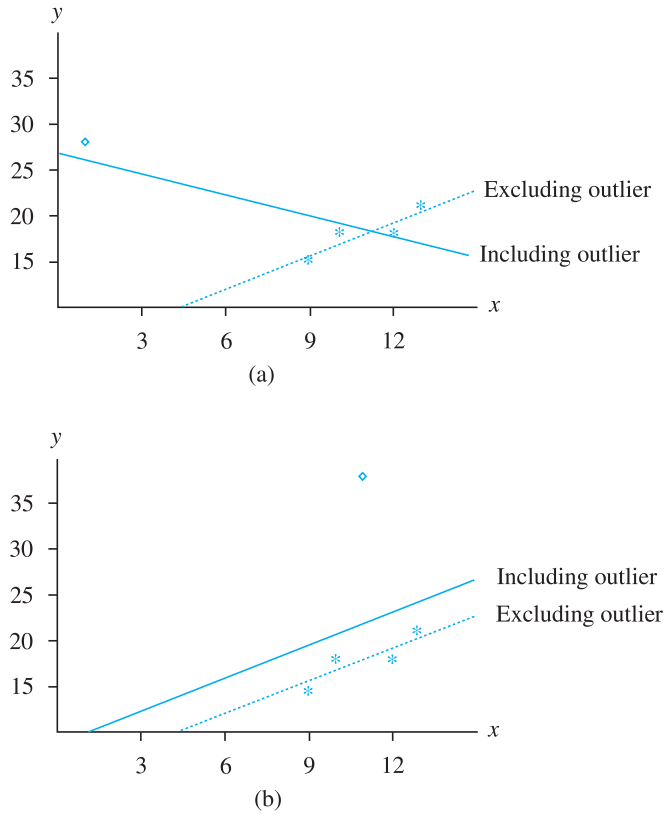
high leverage point

high influence point

The estimate of the regression slope can potentially be greatly affected by **high leverage points.** These are points that have very high or very low values of the independent variable—outliers in the x direction. They carry great weight in the estimate of the slope. A high leverage point that also happens to correspond to a y outlier is a **high influence point.** It will alter the slope and twist the line badly.

A point has high influence if omitting it from the data will cause the regression line to change substantially. To have high influence, a point must first have high leverage and, in addition, must fall outside the pattern of the remaining points. Consider the two scatterplots in Figure 11.12. In plot (a), the point in the upper left corner is far to the left of the other points; it has a much lower x value and therefore has high leverage. If we drew a line through the other points, the line would fall far below this point, so the point is an outlier in the y direction as well. Therefore, it also has high influence. Including this point would change the slope of the line greatly. In contrast, in plot (b), the y outlier point corresponds to an x value very near the mean, having low leverage. Including this point would pull the line

FIGURE 11.12
(a) High influence and (b) low influence points



upward, increasing the intercept, but it wouldn't increase or decrease the slope much at all. Therefore, it does not have great influence.

A high leverage point indicates only a *potential* distortion of the equation. Whether or not including the point will "twist" the equation depends on its influence (whether or not the point falls near the line through the remaining points). A point must have *both* high leverage and an outlying y value to qualify as a high influence point.

Mathematically, the effect of a point's leverage can be seen in the S_{xy} term that enters into the slope calculation. One of the many ways this term can be written is

$$S_{xy} = \sum (x_i - \bar{x}) y_i$$

We can think of this equation as a weighted sum of y values. The weights are large positive or negative numbers when the x value is far from its mean and has high leverage. The weight is almost 0 when x is very close to its mean and has low leverage.

diagnostic measures

Most computer programs that perform regression analyses will calculate one or another of several **diagnostic measures** of leverage and influence. We won't try to summarize all of these measures. We only note that very large values of any of these measures correspond to very high leverage or influence points. The distinction between high leverage (x outlier) and high influence (x outlier and y outlier) points is not universally agreed upon yet. Check the program's documentation to see what definition is being used.

The standard error of the slope $\hat{\beta}_1$ is calculated by all statistical packages. Typically, it is shown in output in a column to the right of the coefficient column. Like any standard error, it indicates how accurately one can estimate the correct population or process value. The quality of estimation of $\hat{\beta}_1$ is influenced by two quantities: the error variance σ_ε^2 and the amount of variation in the independent variable S_{xx}:

$$\sigma_{\hat{\beta}_1} = \frac{\sigma_\varepsilon}{\sqrt{S_{xx}}}$$

The greater the variability σ_ε of the y value for a given value of x, the larger $\sigma_{\hat{\beta}_1}$ is. Sensibly, if there is high variability around the regression line, it is difficult to estimate that line. Also, the smaller the variation in x values (as measured by S_{xx}), the larger $\sigma_{\hat{\beta}_1}$ is. The slope is the predicted change in y per unit change in x; if x changes very little in the data, so that S_{xx} is small, it is difficult to estimate the rate of change in y accurately. If the price of a brand of diet soda has not changed for years, it is obviously hard to estimate the change in quantity demanded when price changes.

$$\sigma_{\hat{\beta}_0} = \sigma_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

The standard error of the estimated intercept $\hat{\beta}_0$ is influenced by n, naturally, and also by the size of the square of the sample mean, $\bar{x}^2$, relative to S_{xx}. The intercept is the predicted y value when $x = 0$; if all the x_i are, for instance, large positive numbers, predicting y at $x = 0$ is a huge extrapolation from the actual data. Such extrapolation magnifies small errors, and the standard error of $\hat{\beta}_0$ is large. The ideal situation for estimating $\hat{\beta}_0$ is when $\bar{x} = 0$.

To this point, we have considered only the estimates of intercept and slope. We also have to estimate the true error variance σ_ε^2. We can think of this quantity as "variance around the line," or as the mean squared prediction error. The estimate of σ_ε^2 is based on the **residuals** $y_i - \hat{y}_i$, which are the prediction errors in the sample.

residuals

The estimate of σ_ε^2 based on the sample data is the sum of squared residuals divided by $n - 2$, the degrees of freedom. The estimated variance is often shown in computer output as MS(Error) or MS(Residual). Recall that MS stands for "mean square" and is always a sum of squares divided by the appropriate degrees of freedom:

$$s_\varepsilon^2 = \frac{\Sigma_i (y_i - \hat{y}_i)^2}{n - 2} = \frac{\text{SS(Residual)}}{n - 2}$$

In the computer output for Example 11.3, SS(Residual) is shown to be 0.9498.

Just as we divide by $n - 1$ rather than by n in the ordinary sample variance s^2 (in Chapter 3), we divide by $n - 2$ in s_ε^2, the estimated variance around the line. The reduction from n to $n - 2$ occurs because in order to estimate the variability around the regression line, we must first estimate the two parameters β_0 and β_1 to obtain the estimated line. The effective sample size for estimating σ_ε^2 is thus $n - 2$. In our definition, s_ε^2 is undefined for $n = 2$, as it should be. Another argument is that dividing by $n - 2$ makes s_ε^2 an unbiased estimator of σ_ε^2. In the computer output of Example 11.3, $n - 2 = 10 - 2 = 8$ is shown as DF (degrees of freedom) for RESIDUAL and $s_\varepsilon^2 = 0.1187$ is shown as MS for RESIDUAL.

residual standard deviation

The square root s_ε of the sample variance is called the **sample standard deviation around the regression line,** the **standard error of estimate,** or the **residual standard deviation.** Because s_ε estimates σ_ε, the standard deviation of y_i, σ_ε estimates the standard deviation of the population of y values associated with a given value of the independent variable x. The output in Example 11.3 labels s_ε as S with $S = 0.344566$.

Like any other standard deviation, the residual standard deviation may be interpreted by the Empirical Rule. About 95% of the prediction errors will fall within ± 2 standard deviations of the mean error; the mean error is always 0 in the least-squares regression model. Therefore, a residual standard deviation of 0.345 means that about 95% of prediction errors will be less than $\pm 2(0.345) = \pm 0.690$.

The estimates $\hat{\beta}_0, \hat{\beta}_1$, and s_ε are basic in regression analysis. They specify the regression line and the probable degree of error associated with y values for a given value of x. The next step is to use these sample estimates to make inferences about the true parameters.

EXAMPLE 11.4

Forest scientists are concerned with the decline in forest growth throughout the world. One aspect of this decline is the possible effect of emissions from coal-fired power plants. The scientists in particular are interested in the pH level of the soil and the resulting impact on tree growth retardation. The scientists study various forests which are likely to be exposed to these emissions. They measure various aspects of growth associated with trees in a specified region and the soil pH in the same region. The forest scientists then want to determine impact on tree growth as the soil becomes more acidic. An index of growth retardation is constructed from the various measurements taken on the trees with a high value indicating greater retardation in tree growth. A higher value of soil pH indicates a more acidic soil. Twenty tree stands which are exposed to the power plant emissions are selected for study. The values of the growth retardation index and average soil pH are recorded in Table 11.3.

TABLE 11.3
Forest growth retardation data

Stand	Soil pH	Grow Ret	Stand	Soil pH	Grow Ret
1	3.3	17.78	11	3.9	14.95
2	3.4	21.59	12	4.0	15.87
3	3.4	23.84	13	4.1	17.45
4	3.5	15.13	14	4.2	14.35
5	3.6	23.45	15	4.3	14.64
6	3.6	20.87	16	4.4	17.25
7	3.7	17.78	17	4.5	12.57
8	3.7	20.09	18	5.0	7.15
9	3.8	17.78	19	5.1	7.50
10	3.8	12.46	20	5.2	4.34

The scientists expect that as the soil pH increases within an acceptable range, the trees will have a lower value for growth retardation index.

Using the above data and analysis using Minitab, do the following:

1. Examine the scatterplot and decide whether a straight line is a reasonable model.
2. Identify least-squares estimates for β_0 and β_1 in the model $y = \beta_0 + \beta_1 x + \varepsilon$, where y is the index of growth retardation and x is the soil pH.
3. Predict the growth retardation for a soil pH of 4.0.
4. Identify s_ε, the sample standard deviation about the regression line.
5. Interpret the value of $\hat{\beta}_1$.

```
Regression Analysis: GrowthRet versus SoilpH
The regression equation is
GrowthRet = 47.5 - 7.86 SoilpH


Predictor      Coef    SE Coef         T       P
Constant     47.475      4.428     10.72   0.000
SoilpH       -7.859      1.090     -7.21   0.000


S = 2.72162   R-Sq = 74.3%   R-Sq(adj) = 72.9%



Analysis of Variance

Source           DF         SS        MS      F        P
Regression        1     385.28    385.28  52.01    0.000
Residual Error   18     133.33      7.41
Total            19     518.61
```

Solution

1. A scatterplot drawn by the Minitab package is shown in Figure 11.13. The data appear to fall approximately along a downward-sloping line. There does not appear to be a need for using a more complex model.

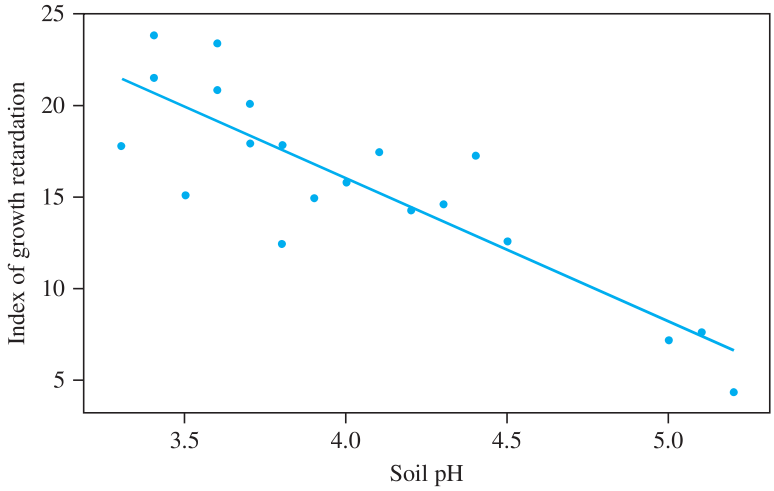


FIGURE 11.13
Scatterplot of growth retardation versus soil pH

2. The output shows the coefficients twice, with differing numbers of digits. The estimated intercept (constant) is $\hat{\beta}_0 = 47.475$ and the estimated slope (Soil pH) is $\hat{\beta}_1 = -7.859$. Note that the negative slope corresponds to a downward-sloping line.
3. The least-squares prediction when $x = 4.0$ is

$$\hat{y} = 47.475 - 7.859(4.0) = 16.04$$

4. The standard deviation around the fitted line (the residual standard deviation) is shown as $S = 2.72162$. Therefore, about 95% of the prediction errors should be less than $\pm 1.96(2.72162) = \pm 5.334$.
5. From $\hat{\beta}_1 = -7.859$, we conclude that for a 1 unit increase in soil pH, there is an estimated decrease of 7.859 in the average value of the growth retardation index.

11.3 Inferences about Regression Parameters

The slope, intercept, and residual standard deviation in a simple regression model are all estimates based on limited data. As with all other statistical quantities, they are affected by random error. In this section, we consider how to allow for that random error. The concepts of hypothesis tests and confidence intervals that we have applied to means and proportions apply equally well to regression summary figures.

t test for β_1

The t distribution can be used to make significance tests and confidence intervals for the true slope and intercept. One natural null hypothesis is that the true slope β_1 equals 0. If this H_0 is true, a change in x yields no predicted change in y, and it follows that x has no value in predicting y. We know from the previous section that the sample slope $\hat{\beta}_1$ has the expected value β_1 and standard error

$$\sigma_{\hat{\beta}_1} = \sigma_\varepsilon \sqrt{\frac{1}{S_{xx}}}$$

In practice, σ_ε is not known and must be estimated by s_ε, the residual standard deviation. In almost all regression analysis computer outputs, the estimated standard

error is shown next to the coefficient. A test of this null hypothesis is given by the t statistic

$$t = \frac{\hat{\beta}_1 - \beta_1}{\text{estimated standard error } (\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1}{s_\varepsilon \sqrt{1/S_{xx}}}$$

The most common use of this statistic is shown in the following summary.

Summary of a Statistical Test for β_1

Hypotheses:

Case 1. H_0: $\beta_1 \le 0$ vs. H_a: $\beta_1 > 0$
Case 2. H_0: $\beta_1 \ge 0$ vs. H_a: $\beta_1 < 0$
Case 3. H_0: $\beta_1 = 0$ vs. H_a: $\beta_1 \ne 0$

T.S.: $t = \dfrac{\hat{\beta}_1 - 0}{s_\varepsilon / \sqrt{S_{xx}}}$

R.R.: For df $= n - 2$ and Type I error α,
1. Reject H_0 if $t > t_\alpha$.
2. Reject H_0 if $t < -t_\alpha$.
3. Reject H_0 if $|t| > t_{\alpha/2}$.

Check assumptions and draw conclusions.

All regression analysis outputs show this t value.

In most computer outputs, this test is indicated after the standard error and labeled as T TEST or T STATISTIC. Often, a p-value is also given, which eliminates the need for looking up the t value in a table.

EXAMPLE 11.5

Use the computer output of Example 11.4 (reproduced here) to locate the value of the t statistic for testing H_0: $\beta_1 = 0$ in the tree growth retardation example. Give the observed level of significance for the test.

```
Predictor      Coef    SE Coef        T      P
Constant     47.475      4.428    10.72  0.000
SoilpH       -7.859      1.090    -7.21  0.000


S = 2.72162   R-Sq = 74.3%   R-Sq(adj) = 72.9%


Analysis of Variance

Source           DF        SS      MS      F      P
Regression        1    385.28  385.28  52.01  0.000
Residual Error   18    133.33    7.41
Total            19    518.61
```

Solution From the Minitab output, the value of the test statistic is $t = -7.21$. The p-value for the two-tailed alternative H_a: $\beta_1 \ne 0$, labelled as P, is .000. In fact,

the value is given by p-value $= 2Pr[t_{18} > 7.21] = .000000521$ which indicates that the value given on the computer output should be interpreted as p-value $< .0001$. Because the value is so small, we can reject the hypothesis that tree growth retardation is not associated with soil pH.

EXAMPLE 11.6

The following data show mean ages of executives of 15 firms in the food industry and the previous year's percentage increase in earnings per share of the firms. Use the Systat output shown to test the hypothesis that executive age has no predictive value for change in earnings. Should a one-sided or two-sided alternative be used?

Mean age	x:	38.2	40.0	42.5	43.4	44.6	44.9	45.0	45.4
Change, earnings per share	y:	8.9	13.0	4.7	−2.4	12.5	18.4	6.6	13.5
	x:	46.0	47.3	47.3	48.0	49.1	50.5	51.6	
	y:	8.5	15.3	18.9	6.0	10.4	15.9	17.1	

```
DEP VAR: CHGEPS N: 15 MULTIPLE R: 0.383 SQUARED MULTIPLE R: 0.147
STANDARD ERROR OF ESTIMATE: 5.634

VARIABLE   COEFFICIENT   STD ERROR   STD COEF    T     P(2 TAIL)
CONSTANT      -16.991      18.866      0.000   0.901     0.384
MEANAGE         0.617       0.413      0.383   1.496     0.158

                    ANALYSIS OF VARIANCE

SOURCE       SUM-OF-SQUARES   DF   MEAN-SQUARE   F-RATIO    P
REGRESSION           71.055    1        71.055     2.239   0.158
RESIDUAL            412.602   13        31.739
```

Solution In the model $y = \beta_0 + \beta_1 x + \varepsilon$, the null hypothesis is H_0: $\beta_1 = 0$. The myth in American business is that younger managers tend to be more aggressive and harder driving, but it is also possible that the greater experience of the older executives leads to better decisions. Therefore, there is a good reason to choose a two-sided research hypothesis, H_a: $\beta_1 \neq 0$. The t statistic is shown in the output column marked T, reasonably enough. It shows $t = 1.496$, with a (two-sided) p-value of .158. There is not enough evidence to conclude that there is any relation between age and change in earnings.

In passing, note that the interpretation of $\hat{\beta}_0$ is rather interesting in this example; it would be the predicted change in earnings of a firm with mean age of its managers equal to 0. Hmm.

It is also possible to calculate a confidence interval for the true slope. This is an excellent way to communicate the likely degree of inaccuracy in the estimate of that slope. The confidence interval once again is simply the estimate plus or minus a t table value times the standard error.

Confidence Interval for Slope β_1

$$\hat{\beta}_1 - t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{S_{xx}}} \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{S_{xx}}}$$

The required degrees of freedom for the table value $t_{\alpha/2}$ is $n - 2$, the error df.

EXAMPLE 11.7

Compute a 95% confidence interval for the slope β_1 using the output from Example 11.4.

Solution In the output, $\hat{\beta}_1 = -7.859$ and the estimated standard error of $\hat{\beta}_1$ is shown in the column labelled **SE Coef** as 1.090. Because n is 20, there are $20 - 2 = 18$ df for error. The required table value for $\alpha/2 = .05/2 = .025$ is 2.101. The corresponding confidence interval for the true value of β_1 is then

$$-7.859 \pm 2.101(1.090) \quad \text{or} \quad -10.149 \text{ to } -5.569$$

The predicted decrease in growth retardation for a unit increase in soil pH ranges from −10.149 to −5.569. The large width of this interval is mainly due to the small sample size.

There is an alternative test, an F test, for the null hypothesis of no predictive value. It was designed to test the null hypothesis that *all* predictors have no value in predicting y. This test gives the same result as a two-sided t test of H_0: $\beta_1 = 0$ in simple linear regression; to say that all predictors have no value is to say that the (only) slope is 0. The F test is summarized next.

F Test for H_0: $\beta_1 = 0$

H_0: $\beta_1 = 0$

H_a: $\beta_1 \neq 0$

T.S.: $$F = \frac{\text{SS(Regression)}/1}{\text{SS(Residual)}/(n-2)} = \frac{\text{MS(Regression)}}{\text{MS(Residual)}}$$

R.R.: With $df_1 = 1$ and $df_2 = n - 2$, reject H_0 if $F > F_\alpha$.

Check assumptions and draw conclusions.

SS(Regression) is the sum of squared deviations of predicted y values from the y mean. SS(Regression) $= \Sigma(\hat{y}_i - \bar{y})^2$. SS(Residual) is the sum of squared deviations of actual y values from predicted y values. SS(Residual) $= \Sigma(\hat{y}_i - \bar{y}_i)^2$.

Virtually all computer packages calculate this F statistic. In Example 11.3, the output shows $F = 54.03$ with a p-value given by 0.000 (in fact, p-value = .00008). Again, the hypothesis of no predictive value can be rejected. It is always true for simple linear regression problems that $F = t^2$; in the example, $54.03 = (7.35)^2$, to within round-off error. The F and two-sided t tests are equivalent in simple linear regression; they serve different purposes in multiple regression.

EXAMPLE 11.8

For the output of Example 11.4, reproduced here, use the F test for testing H_0: $\beta_1 = 0$. Show that $t^2 = F$ for this data set.

```
Predictor      Coef    SE Coef       T       P
Constant     47.475      4.428   10.72   0.000
SoilpH       -7.859      1.090   -7.21   0.000


S = 2.72162   R-Sq = 74.3%   R-Sq(adj) = 72.9%


Analysis of Variance

Source           DF        SS      MS      F       P
Regression        1    385.28  385.28  52.01   0.000
Residual Error   18    133.33    7.41
Total            19    518.61
```

Solution The F statistic is shown in the output as 52.01, with a p-value of .000 (indicating the actual p-value is something less than .0005). Using a computer program, the actual p-value is .00000104. Note that the t statistic is -7.21, and $t^2 = (-7.21)^2 = 51.984$, which equals the F value, to within round-off error.

A confidence interval for β_0 can be computed using the estimated standard error of $\hat{\beta}_0$ as

$$\hat{\sigma}_{\hat{\beta}_0} = s_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

Confidence Interval for Intercept β_0

$$\hat{\beta}_0 \pm t_{\alpha/2} s_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

The required degrees of freedom for the table value of $t_{\alpha/2}$ is $n - 2$, the error df.

In practice, this parameter is of less interest than the slope. In particular, there is often no reason to hypothesize that the true intercept is zero (or any other particular value). Computer packages almost always test the null hypothesis of zero slope, but some don't bother with a test on the intercept term.

11.4 Predicting New *y* Values Using Regression

In all the regression analyses we have done so far, we have been summarizing and making inferences about relations in data that have already been observed. Thus, we have been predicting the past. One of the most important uses of regression is trying to forecast the future. In the road resurfacing example, the county highway director wants to predict the cost of a new contract that is up for bids. In a regression relating the change in systolic blood pressure for a specified dose of a drug, the doctor will want to predict the change in systolic blood pressure for a dose level not used in the study. In this section, we discuss how to make such regression predictions and how to determine prediction intervals which will convey our uncertainty in these predictions.