11

Simple Linear Regression and Correlation

CHAPTER OUTLINE

LEARNING OBJECTIVES

After careful study of this chapter, you should be able to do the following:

1. Use simple linear regression for building empirical models to engineering and scientific data
2. Understand how the method of least squares is used to estimate the parameters in a linear regression model
3. Analyze residuals to determine if the regression model is an adequate fit to the data or to see if any underlying assumptions are violated
4. Test statistical hypotheses and construct confidence intervals on regression model parameters
5. Use the regression model to make a prediction of a future observation and construct an appropriate prediction interval on the future observation
6. Use simple transformations to achieve a linear regression model
7. Apply the correlation model

CD MATERIAL

8. **Conduct a lack-of-fit test in a regression model where there are replicated observations.**

Answers for many odd numbered exercises are at the end of the book. Answers to exercises whose numbers are surrounded by a box can be accessed in the e-Text by clicking on the box. Complete worked solutions to certain exercises are also available in the e-Text. These are indicated in the Answers to Selected Exercises section by a box around the exercise number. Exercises are also available for some of the text sections that appear on CD only. These exercises may be found within the e-Text immediately following the section they accompany.

11-1 EMPIRICAL MODELS

Many problems in engineering and science involve exploring the relationships between two or more variables. **Regression analysis** is a statistical technique that is very useful for these types of problems. For example, in a chemical process, suppose that the yield of the product is related to the process-operating temperature. Regression analysis can be used to build a model to predict yield at a given temperature level. This model can also be used for process optimization, such as finding the level of temperature that maximizes yield, or for process control purposes.

As an illustration, consider the data in Table 11-1. In this table y is the purity of oxygen produced in a chemical distillation process, and x is the percentage of hydrocarbons that are present in the main condenser of the distillation unit. Figure 11-1 presents a **scatter diagram**

Table 11-1 Oxygen and Hydrocarbon Levels

Observation Number	Hydrocarbon Level x (%)	Purity y (%)
1	0.99	90.01
2	1.02	89.05
3	1.15	91.43
4	1.29	93.74
5	1.46	96.73
6	1.36	94.45
7	0.87	87.59
8	1.23	91.77
9	1.55	99.42
10	1.40	93.65
11	1.19	93.54
12	1.15	92.52
13	0.98	90.56
14	1.01	89.54
15	1.11	89.85
16	1.20	90.39
17	1.26	93.25
18	1.32	93.41
19	1.43	94.98
20	0.95	87.33

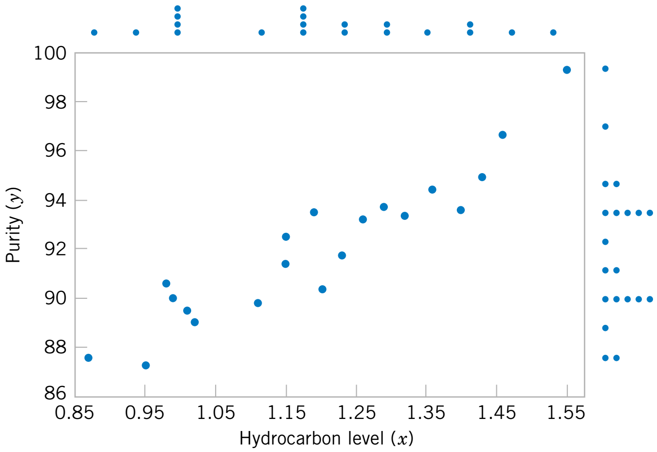

Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

of the data in Table 11-1. This is just a graph on which each (x_i, y_i) pair is represented as a point plotted in a two-dimensional coordinate system. This scatter diagram was produced by Minitab, and we selected an option that shows dot diagrams of the x and y variables along the top and right margins of the graph, respectively, making it easy to see the distributions of the individual variables (box plots or histograms could also be selected). Inspection of this scatter diagram indicates that, although no simple curve will pass exactly through all the points, there is a strong indication that the points lie scattered randomly around a straight line. Therefore, it is probably reasonable to assume that the mean of the random variable Y is related to x by the following straight-line relationship:

$$E(Y|x) = \mu_{Y|x} = \beta_0 + \beta_1 x$$

where the slope and intercept of the line are called **regression coefficients.** While the mean of Y is a linear function of x, the actual observed value y does not fall exactly on a straight line. The appropriate way to generalize this to a **probabilistic linear model** is to assume that the expected value of Y is a linear function of x, but that for a fixed value of x the actual value of Y is determined by the mean value function (the linear model) plus a random error term, say,

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{11-1}$$

where ϵ is the random error term. We will call this model the **simple linear regression model,** because it has only one independent variable or **regressor.** Sometimes a model like this will arise from a theoretical relationship. At other times, we will have no theoretical knowledge of the relationship between x and y, and the choice of the model is based on inspection of a scatter diagram, such as we did with the oxygen purity data. We then think of the regression model as an **empirical model.**

To gain more insight into this model, suppose that we can fix the value of x and observe the value of the random variable Y. Now if x is fixed, the random component ϵ on the right-hand side of the model in Equation 11-1 determines the properties of Y. Suppose that the mean and variance of ϵ are 0 and σ^2, respectively. Then

$$E(Y|x) = E(\beta_0 + \beta_1 x + \epsilon) = \beta_0 + \beta_1 x + E(\epsilon) = \beta_0 + \beta_1 x$$

Notice that this is the same relationship that we initially wrote down empirically from inspection of the scatter diagram in Fig. 11-1. The variance of Y given x is

$$V(Y|x) = V(\beta_0 + \beta_1 x + \epsilon) = V(\beta_0 + \beta_1 x) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

Thus, the true regression model $\mu_{Y|x} = \beta_0 + \beta_1 x$ is a line of mean values; that is, the height of the regression line at any value of x is just the expected value of Y for that x. The slope, β_1, can be interpreted as the change in the mean of Y for a unit change in x. Furthermore, the variability of Y at a particular value of x is determined by the error variance σ^2. This implies that there is a distribution of Y-values at each x and that the variance of this distribution is the same at each x.

For example, suppose that the true regression model relating oxygen purity to hydrocarbon level is $\mu_{Y|x} = 75 + 15x$, and suppose that the variance is $\sigma^2 = 2$. Figure 11-2 illustrates this situation. Notice that we have used a normal distribution to describe the random variation in ϵ. Since Y is the sum of a constant $\beta_0 + \beta_1 x$ (the mean) and a normally distributed random variable, Y is a normally distributed random variable. The variance σ^2 determines the

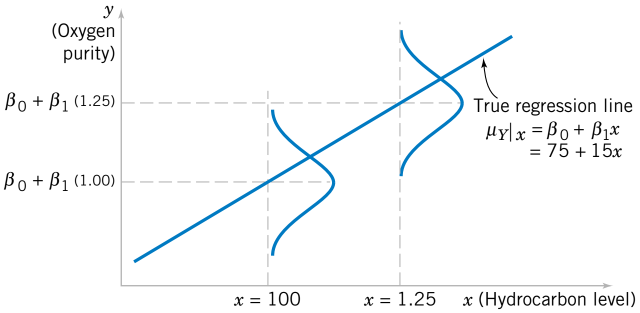

Figure 11-2 The distribution of Y for a given value of x for the oxygen purity-hydrocarbon data.

variability in the observations Y on oxygen purity. Thus, when σ^2 is small, the observed values of Y will fall close to the line, and when σ^2 is large, the observed values of Y may deviate considerably from the line. Because σ^2 is constant, the variability in Y at any value of x is the same.

The regression model describes the relationship between oxygen purity Y and hydrocarbon level x. Thus, for any value of hydrocarbon level, oxygen purity has a normal distribution with mean $75 + 15x$ and variance 2. For example, if $x = 1.25$, Y has mean value $\mu_{Y|x} = 75 + 15(1.25) = 93.75$ and variance 2.

In most real-world problems, the values of the intercept and slope (β_0, β_1) and the error variance σ^2 will not be known, and they must be estimated from sample data. Then this fitted regression equation or model is typically used in prediction of future observations of Y, or for estimating the mean response at a particular level of x. To illustrate, a chemical engineer might be interested in estimating the mean purity of oxygen produced when the hydrocarbon level is $x = 1.25\%$. This chapter discusses such procedures and applications for the simple linear regression model. Chapter 12 will discuss multiple linear regression models that involve more than one regressor.

11-2 SIMPLE LINEAR REGRESSION

The case of **simple linear regression** considers a single **regressor** or **predictor** x and a dependent or **response variable** Y. Suppose that the true relationship between Y and x is a straight line and that the observation Y at each level of x is a random variable. As noted previously, the expected value of Y for each value of x is

$$E(Y|x) = \beta_0 + \beta_1 x$$

where the intercept β_0 and the slope β_1 are unknown regression coefficients. We assume that each observation, Y, can be described by the model

$$Y = \beta_0 + \beta_1 x + \epsilon \qquad (11\text{-}2)$$

where ϵ is a random error with mean zero and (unknown) variance σ^2. The random errors corresponding to different observations are also assumed to be uncorrelated random variables.

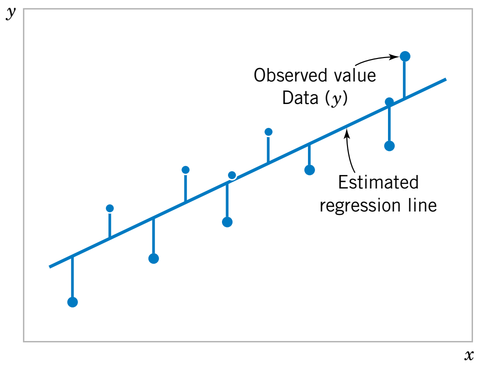

Figure 11-3 Deviations of the data from the estimated regression model.

Suppose that we have n pairs of observations (x_1, y_1), (x_2, y_2), … (x_n, y_n). Figure 11-3 shows a typical scatter plot of observed data and a candidate for the estimated regression line. The estimates of β_0 and β_1 should result in a line that is (in some sense) a “best fit” to the data. The German scientist Karl Gauss (1777–1855) proposed estimating the parameters β_0 and β_1 in Equation 11-2 to minimize the sum of the squares of the vertical deviations in Fig. 11-3.

We call this criterion for estimating the regression coefficients the **method of least squares.** Using Equation 11-2, we may express the n observations in the sample as

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad i = 1, 2, \ldots, n \tag{11-3}$$

and the sum of the squares of the deviations of the observations from the true regression line is

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2 \tag{11-4}$$

The least squares estimators of β_0 and β_1, say, $\hat{\beta}_0$ and $\hat{\beta}_1$, must satisfy

$$\begin{aligned} \left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\ \left.\frac{\partial L}{\partial \beta_1}\right|_{\hat{\beta}_0,\hat{\beta}_1} &= -2\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)x_i = 0 \end{aligned} \tag{11-5}$$

Simplifying these two equations yields

$$\begin{aligned} n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i &= \sum_{i=1}^{n} y_i \\ \hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 &= \sum_{i=1}^{n} y_i x_i \end{aligned} \tag{11-6}$$

Equations 11-6 are called the **least squares normal equations.** The solution to the normal equations results in the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Definition

The **least squares estimates** of the intercept and slope in the simple linear regression model are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{11-7}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} y_i x_i - \frac{\left(\sum_{i=1}^{n} y_i\right)\left(\sum_{i=1}^{n} x_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} \tag{11-8}$$

where $\bar{y} = (1/n)\sum_{i=1}^{n} y_i$ and $\bar{x} = (1/n)\sum_{i=1}^{n} x_i$.

The **fitted** or **estimated regression line** is therefore

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{11-9}$$

Note that each pair of observations satisfies the relationship

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \qquad i = 1, 2, \ldots, n$$

where $e_i = y_i - \hat{y}_i$ is called the **residual.** The residual describes the error in the fit of the model to the ith observation y_i. Later in this chapter we will use the residuals to provide information about the adequacy of the fitted model.

Notationally, it is occasionally convenient to give special symbols to the numerator and denominator of Equation 11-8. Given data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, let

$$S_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n} \tag{11-10}$$

and

$$S_{xy} = \sum_{i=1}^{n} y_i(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n} \tag{11-11}$$

EXAMPLE 11-1

We will fit a simple linear regression model to the oxygen purity data in Table 11-1. The following quantities may be computed:

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21 \quad \bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

$$\sum_{i=1}^{20} y_i^2 = 170{,}044.5321 \quad \sum_{i=1}^{20} x_i^2 = 29.2892 \quad \sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

$$S_{xx} = \sum_{i=1}^{20} x_i^2 - \frac{\left(\sum_{i=1}^{20} x_i\right)^2}{20} = 29.2892 - \frac{(23.92)^2}{20} = 0.68088$$

and

$$S_{xy} = \sum_{i=1}^{20} x_i y_i - \frac{\left(\sum_{i=1}^{20} x_i\right)\left(\sum_{i=1}^{20} y_i\right)}{20} = 2{,}214.6566 - \frac{(23.92)(1{,}843.21)}{20} = 10.17744$$

Therefore, the least squares estimates of the slope and intercept are

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{10.17744}{0.68088} = 14.94748$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 92.1605 - (14.94748)1.196 = 74.28331$$

The fitted simple linear regression model (with the coefficients reported to three decimal places) is

$$\hat{y} = 74.283 + 14.947x$$

This model is plotted in Fig. 11-4, along with the sample data.

Computer software programs are widely used in regression modeling. These programs typically carry more decimal places in the calculations. Table 11-2 shows a portion of the output from Minitab for this problem. The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are highlighted. In subsequent sections we will provide explanations for the information provided in this computer output.

Using the regression model of Example 11-1, we would predict oxygen purity of $\hat{y} = 89.23\%$ when the hydrocarbon level is $x = 1.00\%$. The purity 89.23% may be interpreted as

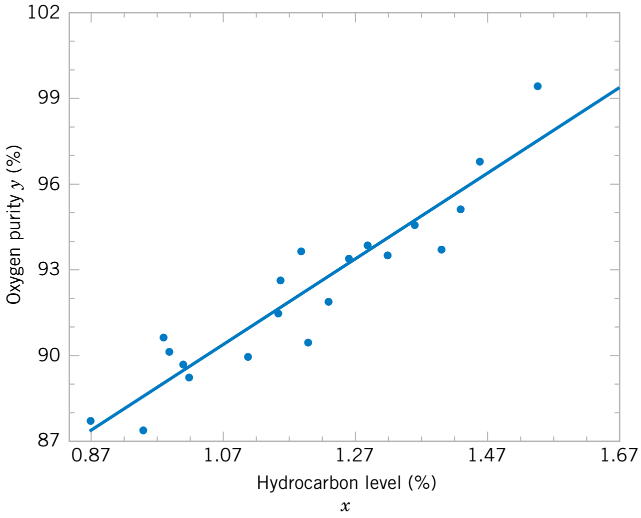

Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.20 + 14.97x$.

Table 11-2 Minitab Output for the Oxygen Purity Data in Example 11-1

```
Regression Analysis
The regression equation is

Purity = 74.3 + 14.9 HC Level

Predictor        Coef          SE Coef        T         P
Constant         74.283 ←β̂0      1.593      46.62     0.000
HC Level         14.947 ←β̂1      1.317      11.35     0.000

S = 1.087          R-Sq = 87.7%          R-Sq (adj) = 87.1%

Analysis of Variance

Source           DF       SS                MS               F          P
Regression        1     152.13            152.13           128.86      0.000
Residual Error   18      21.25 ←SSE         1.18 ←σ̂²
Total            19     173.38

Predicted Values for New Observations

New Obs           Fit         SE Fit      95.0%    CI          95.0%    PI
1               89.231         0.354    (88.486,  89.975)    (86.830,  91.632)

Values of Predictors for New Observations

New Obs       HC Level
1                 1.00
```

an estimate of the true population mean purity when $x = 1.00\%$, or as an estimate of a new observation when $x = 1.00\%$. These estimates are, of course, subject to error; that is, it is unlikely that a future observation on purity would be exactly 89.23% when the hydrocarbon level is 1.00%. In subsequent sections we will see how to use confidence intervals and prediction intervals to describe the error in estimation from a regression model.

Estimating σ^2

There is actually another unknown parameter in our regression model, σ^2 (the variance of the error term ϵ). The residuals $e_i = y_i - \hat{y}_i$ are used to obtain an estimate of σ^2. The sum of squares of the residuals, often called the **error sum of squares,** is

$$SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \qquad (11\text{-}12)$$

We can show that the expected value of the error sum of squares is $E(SS_E) = (n-2)\sigma^2$. Therefore an **unbiased estimator** of σ^2 is

$$\hat{\sigma}^2 = \frac{SS_E}{n-2} \qquad (11\text{-}13)$$

Computing SS_E using Equation 11-12 would be fairly tedious. A more convenient computing formula can be obtained by substituting $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ into Equation 11-12 and simplifying.

The resulting computing formula is

$$SS_E = SS_T - \hat{\beta}_1 S_{xy} \tag{11-14}$$

where $SS_T = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ is the **total sum of squares of the response variable** y. The error sum of squares and the estimate of σ^2 for the oxygen purity data, $\hat{\sigma}^2 = 1.18$, are highlighted in the Minitab output in Table 11-2.

EXERCISES FOR SECTION 11-2

11-1. An article in *Concrete Research* ("Near Surface Characteristics of Concrete: Intrinsic Permeability," Vol. 41, 1989), presented data on compressive strength x and intrinsic permeability y of various concrete mixes and cures. Summary quantities are $n = 14$, $\sum y_i = 572$, $\sum y_i^2 = 23{,}530$, $\sum x_i = 43$, $\sum x_i^2 = 157.42$, and $\sum x_i y_i = 1697.80$. Assume that the two variables are related according to the simple linear regression model.

(a) Calculate the least squares estimates of the slope and intercept.

(b) Use the equation of the fitted line to predict what permeability would be observed when the compressive strength is $x = 4.3$.

(c) Give a point estimate of the mean permeability when compressive strength is $x = 3.7$.

(d) Suppose that the observed value of permeability at $x = 3.7$ is $y = 46.1$. Calculate the value of the corresponding residual.

11-2. Regression methods were used to analyze the data from a study investigating the relationship between roadway surface temperature (x) and pavement deflection (y). Summary quantities were $n = 20$, $\sum y_i = 12.75$, $\sum y_i^2 = 8.86$, $\sum x_i = 1478$, $\sum x_i^2 = 143{,}215.8$, and $\sum x_i y_i = 1083.67$.

(a) Calculate the least squares estimates of the slope and intercept. Graph the regression line.

(b) Use the equation of the fitted line to predict what pavement deflection would be observed when the surface temperature is 85°F.

(c) What is the mean pavement deflection when the surface temperature is 90°F?

(d) What change in mean pavement deflection would be expected for a 1°F change in surface temperature?

11-3. Consider the regression model developed in Exercise 11-2.

(a) Suppose that temperature is measured in °C rather than °F. Write the new regression model that results.

(b) What change in expected pavement deflection is associated with a 1°C change in surface temperature?

11-4. Montgomery, Peck, and Vining (2001) present data concerning the performance of the 28 National Football League teams in 1976. It is suspected that the number of games won (y) is related to the number of yards gained rushing by an opponent (x). The data are shown in the following table.

Teams	Games Won (y)	Yards Rushing by Opponent (x)	Teams	Games Won (y)	Yards Rushing by Opponent (x)
Washington	10	2205	Detroit	6	1901
Minnesota	11	2096	Green Bay	5	2288
New England	11	1847	Houston	5	2072
Oakland	13	1903	Kansas City	5	2861
Pittsburgh	10	1457	Miami	6	2411
Baltimore	11	1848	New Orleans	4	2289
Los Angeles	10	1564	New York Giants	3	2203
Dallas	11	1821	New York Jets	3	2592
Atlanta	4	2577	Philadelphia	4	2053
Buffalo	2	2476	St. Louis	10	1979
Chicago	7	1984	San Diego	6	2048
Cincinnati	10	1917	San Francisco	8	1786
Cleveland	9	1761	Seattle	2	2876
Denver	9	1709	Tampa Bay	0	2560

(a) Calculate the least squares estimates of the slope and intercept. What is the estimate of σ^2? Graph the regression model.

(b) Find an estimate of the mean number of games won if the opponents can be limited to 1800 yards rushing.

(c) What change in the expected number of games won is associated with a decrease of 100 yards rushing by an opponent?

(d) To increase by 1 the mean number of games won, how much decrease in rushing yards must be generated by the defense?

(e) Given that $x = 1917$ yards (Cincinnati), find the fitted value of y and the corresponding residual.

11-5. An article in *Technometrics* by S. C. Narula and J. F. Wellington ("Prediction, Linear Regression, and a Minimum Sum of Relative Errors," Vol. 19, 1977) presents data on the selling price and annual taxes for 24 houses. The data are shown in the following table.

(a) Assuming that a simple linear regression model is appropriate, obtain the least squares fit relating selling price to taxes paid. What is the estimate of σ^2?

Sale Price/1000	Taxes (Local, School), County)/1000	Sale Price/1000	Taxes (Local, School), County)/1000
25.9	4.9176	30.0	5.0500
29.5	5.0208	36.9	8.2464
27.9	4.5429	41.9	6.6969
25.9	4.5573	40.5	7.7841
29.9	5.0597	43.9	9.0384
29.9	3.8910	37.5	5.9894
30.9	5.8980	37.9	7.5422
28.9	5.6039	44.5	8.7951
35.9	5.8282	37.9	6.0831
31.5	5.3003	38.9	8.3607
31.0	6.2712	36.9	8.1400
30.9	5.9592	45.8	9.1416

(b) Find the mean selling price given that the taxes paid are $x = 7.50$.

(c) Calculate the fitted value of y corresponding to $x = 5.8980$. Find the corresponding residual.

(d) Calculate the fitted $\hat{y}_i$ for each value of x_i used to fit the model. Then construct a graph of $\hat{y}_i$ versus the corresponding observed value y_i and comment on what this plot would look like if the relationship between y and x was a deterministic (no random error) straight line. Does the plot actually obtained indicate that taxes paid is an effective regressor variable in predicting selling price?

11-6. The number of pounds of steam used per month by a chemical plant is thought to be related to the average ambient temperature (in° F) for that month. The past year's usage and temperature are shown in the following table:

Month	Temp.	Usage/1000	Month	Temp.	Usage/1000
Jan.	21	185.79	July	68	621.55
Feb.	24	214.47	Aug.	74	675.06
Mar.	32	288.03	Sept.	62	562.03
Apr.	47	424.84	Oct.	50	452.93
May	50	454.58	Nov.	41	369.95
June	59	539.03	Dec.	30	273.98

(a) Assuming that a simple linear regression model is appropriate, fit the regression model relating steam usage (y) to the average temperature (x). What is the estimate of σ^2?

(b) What is the estimate of expected steam usage when the average temperature is 55°F?

(c) What change in mean steam usage is expected when the monthly average temperature changes by 1°F?

(d) Suppose the monthly average temperature is 47°F. Calculate the fitted value of y and the corresponding residual.

11-7. The data shown in the following table are highway gasoline mileage performance and engine displacement for a sample of 20 cars.

Make	Model	MPG (highway)	Engine Displacement (in^3)	Make	Model	MPG (highway)	Engine Displacement (in^3)
Acura	Legend	30	97	Ford	Taurus	27	153
BMW	735i	19	209	Ford	Tempo	33	90
Buick	Regal	29	173	Honda	Accord	30	119
Chevrolet	Cavalier	32	121	Mazda	RX-7	23	80
Chevrolet	Celebrity	30	151	Mercedes	260E	24	159
Chrysler	Conquest	24	156	Mercury	Tracer	29	97
Dodge	Aries	30	135	Nissan	Maxima	26	181
Dodge	Dynasty	28	181	Oldsmobile	Cutlass	29	173
Ford	Escort	31	114	Plymouth	Laser	37	122
Ford	Mustang	25	302	Pontiac	Grand Prix	29	173

(a) Fit a simple linear model relating highway miles per gallon (y) to engine displacement (x) using least squares.

(b) Find an estimate of the mean highway gasoline mileage performance for a car with 150 cubic inches engine displacement.

(c) Obtain the fitted value of y and the corresponding residual for a car, the Ford Escort, with engine displacement of 114 cubic inches.

11-8. An article in the *Tappi Journal* (March, 1986) presented data on green liquor Na_2S concentration (in grams per liter) and paper machine production (in tons per day). The data (read from a graph) are shown as follows:

y	40	42	49	46	44	48	46	43	53	52	54	57	58
x	825	830	890	895	890	910	915	960	990	1010	1012	1030	1050

(a) Fit a simple linear regression model with y = green liquor Na_2S concentration and x = production. Find an estimate of σ^2. Draw a scatter diagram of the data and the resulting least squares fitted model.

(b) Find the fitted value of y corresponding to $x = 910$ and the associated residual.

(c) Find the mean green liquor Na_2S concentration when the production rate is 950 tons per day.

11-9. An article in the *Journal of Sound and Vibration* (Vol. 151, 1991, pp. 383–394) described a study investigating the relationship between noise exposure and hypertension. The following data are representative of those reported in the article.

y	1	0	1	2	5	1	4	6	2	3	5	4	6	8	4	5	7	9	7	6
x	60	63	65	70	70	70	80	90	80	80	85	89	90	90	90	90	94	100	100	100

(a) Draw a scatter diagram of y (blood pressure rise in millimeters of mercury) versus x (sound pressure level in decibels). Does a simple linear regression model seem reasonable in this situation?

(b) Fit the simple linear regression model using least squares. Find an estimate of σ^2.

(c) Find the predicted mean rise in blood pressure level associated with a sound pressure level of 85 decibals.

11-10. An article in *Wear* (Vol. 152, 1992, pp. 171–181) presents data on the fretting wear of mild steel and oil viscosity. Representative data follow, with x = oil viscosity and y = wear volume (10^{-4} cubic millimeters).

y	240	181	193	155	172	110	113	75	94
x	1.6	9.4	15.5	20.0	22.0	35.5	43.0	40.5	33.0

(a) Construct a scatter plot of the data. Does a simple linear regression model appear to be plausible?

(b) Fit the simple linear regression model using least squares. Find an estimate of σ^2.

(c) Predict fretting wear when viscosity $x = 30$.

(d) Obtain the fitted value of y when $x = 22.0$ and calculate the corresponding residual.

11-11. An article in the *Journal of Environmental Engineering* (Vol. 115, No. 3, 1989, pp. 608–619) reported the results of a study on the occurrence of sodium and chloride in surface streams in central Rhode Island. The following data are chloride concentration y (in milligrams per liter) and roadway area in the watershed x (in percentage).

y	4.4	6.6	9.7	10.6	10.8	10.9	11.8	12.1	14.3	14.7	15.0	17.3	19.2	23.1	27.4	27.7	31.8	39.5
x	0.19	0.15	0.57	0.70	0.67	0.63	0.47	0.70	0.60	0.78	0.81	0.78	0.69	1.30	1.05	1.06	1.74	1.62

(a) Draw a scatter diagram of the data. Does a simple linear regression model seem appropriate here?

(b) Fit the simple linear regression model using the method of least squares. Find an estimate of σ^2.

(c) Estimate the mean chloride concentration for a watershed that has 1% roadway area.

(d) Find the fitted value corresponding to $x = 0.47$ and the associated residual.

11-12. A rocket motor is manufactured by bonding together two types of propellants, an igniter and a sustainer. The shear strength of the bond y is thought to be a linear function of the age of the propellant x when the motor is cast. Twenty observations are shown in the table on the next page.

(a) Draw a scatter diagram of the data. Does the straight-line regression model seem to be plausible?

(b) Find the least squares estimates of the slope and intercept in the simple linear regression model. Find an estimate of σ^2.

(c) Estimate the mean shear strength of a motor made from propellant that is 20 weeks old.

(d) Obtain the fitted values $\hat{y}_i$ that correspond to each observed value y_i. Plot $\hat{y}_i$ versus y_i, and comment on what this plot would look like if the linear relationship between

Observation Number	Strength y (psi)	Age x (weeks)	Observation Number	Strength y (psi)	Age x (weeks)
1	2158.70	15.50	11	2165.20	13.00
2	1678.15	23.75	12	2399.55	3.75
3	2316.00	8.00	13	1779.80	25.00
4	2061.30	17.00	14	2336.75	9.75
5	2207.50	5.00	15	1765.30	22.00
6	1708.30	19.00	16	2053.50	18.00
7	1784.70	24.00	17	2414.40	6.00
8	2575.00	2.50	18	2200.50	12.50
9	2357.90	7.50	19	2654.20	2.00
10	2277.70	11.00	20	1753.70	21.50

shear strength and age were perfectly deterministic (no error). Does this plot indicate that age is a reasonable choice of regressor variable in this model?

11-13. Show that in a simple linear regression model the point $(\bar{x}, \bar{y})$ lies exactly on the least squares regression line.

11-14. Consider the simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$. Suppose that the analyst wants to use $z = x - \bar{x}$ as the regressor variable.

(a) Using the data in Exercise 11-12, construct one scatter plot of the (x_i, y_i) points and then another of the $(z_i = x_i - \bar{x}, y_i)$ points. Use the two plots to intuitively explain how the two models, $Y = \beta_0 + \beta_1 x + \epsilon$ and $Y = \beta_0^* + \beta_1^* z + \epsilon$, are related.

(b) Find the least squares estimates of β_0^* and β_1^* in the model $Y = \beta_0^* + \beta_1^* z + \epsilon$. How do they relate to the least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$?

11-15. Suppose we wish to fit the model $y_i^* = \beta_0^* + \beta_1^*(x_i - \bar{x}) + \epsilon_i$, where $y_i^* = y_i - \bar{y}$ $(i = 1, 2, \ldots, n)$. Find the least squares estimates of β_0^* and β_1^*. How do they relate to $\hat{\beta}_0$ and $\hat{\beta}_1$?

11-16. Suppose we wish to fit a regression model for which the true regression line passes through the point (0, 0). The appropriate model is $Y = \beta x + \epsilon$. Assume that we have n pairs of data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Find the least squares estimate of β.

11-17. Using the results of Exercise 11-16, fit the model $Y = \beta x + \epsilon$ to the chloride concentration-roadway area data in Exercise 11-11. Plot the fitted model on a scatter diagram of the data and comment on the appropriateness of the model.

11-3 PROPERTIES OF THE LEAST SQUARES ESTIMATORS

The statistical properties of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ may be easily described. Recall that we have assumed that the error term ϵ in the model $Y = \beta_0 + \beta_1 x + \epsilon$ is a random variable with mean zero and variance σ^2. Since the values of x are fixed, Y is a random variable with mean $\mu_{Y|x} = \beta_0 + \beta_1 x$ and variance σ^2. Therefore, the values of $\hat{\beta}_0$ and $\hat{\beta}_1$ depend on the observed y's; thus, the least squares estimators of the regression coefficients may be viewed as random variables. We will investigate the bias and variance properties of the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

Consider first $\hat{\beta}_1$. Because $\hat{\beta}_1$ is a linear combination of the observations Y_i, we can use properties of expectation to show that expected value of $\hat{\beta}_1$ is

$$E(\hat{\beta}_1) = \beta_1 \tag{11-15}$$

Thus, $\hat{\beta}_1$ is an **unbiased estimator** of the true slope β_1.

Now consider the variance of $\hat{\beta}_1$. Since we have assumed that $V(\epsilon_i) = \sigma^2$, it follows that $V(Y_i) = \sigma^2$, and it can be shown that

$$V(\hat{\beta}_1) = \frac{\sigma^2}{S_{xx}} \tag{11-16}$$

For the intercept, we can show that

$$E(\hat{\beta}_0) = \beta_0 \quad \text{and} \quad V(\hat{\beta}_0) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right] \tag{11-17}$$

Thus, $\hat{\beta}_0$ is an unbiased estimator of the intercept β_0. The covariance of the random variables $\hat{\beta}_0$ and $\hat{\beta}_1$ is not zero. It can be shown (see Exercise 11-69) that $\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\sigma^2 \bar{x}/S_{xx}$.

The estimate of σ^2 could be used in Equations 11-16 and 11-17 to provide estimates of the variance of the slope and the intercept. We call the square roots of the resulting variance estimators the **estimated standard errors** of the slope and intercept, respectively.

Definition

In simple linear regression the **estimated standard error of the slope** and the **estimated standard error of the intercept** are

$$se(\hat{\beta}_1) = \sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \quad \text{and} \quad se(\hat{\beta}_0) = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right]}$$

respectively, where $\hat{\sigma}^2$ is computed from Equation 11-13.

The Minitab computer output in Table 11-2 reports the estimated standard errors of the slope and intercept under the column heading "*SE* coeff."

11-4 SOME COMMENTS ON USES OF REGRESSION (CD ONLY)

11-5 HYPOTHESIS TESTS IN SIMPLE LINEAR REGRESSION

An important part of assessing the adequacy of a linear regression model is testing statistical hypotheses about the model parameters and constructing certain confidence intervals. Hypothesis testing in simple linear regression is discussed in this section, and Section 11-6 presents methods for constructing confidence intervals. To test hypotheses about the slope and intercept of the regression model, we must make the additional assumption that the error component in the model, ϵ, is normally distributed. Thus, the complete assumptions are that the errors are normally and independently distributed with mean zero and variance σ^2, abbreviated NID(0, σ^2).

11-5.1 Use of *t*-Tests

Suppose we wish to test the hypothesis that the slope equals a constant, say, $\beta_{1,0}$. The appropriate hypotheses are

$$\begin{aligned} H_0&: \beta_1 = \beta_{1,0} \\ H_1&: \beta_1 \neq \beta_{1,0} \end{aligned} \tag{11-18}$$

where we have assumed a two-sided alternative. Since the errors ϵ_i are NID(0, σ^2), it follows directly that the observations Y_i are NID($\beta_0 + \beta_1 x_i$, σ^2). Now $\hat{\beta}_1$ is a linear combination of independent normal random variables, and consequently, $\hat{\beta}_1$ is $N(\beta_1, \sigma^2/S_{xx})$, using the bias and variance properties of the slope discussed in Section 11-3. In addition, $(n-2)\hat{\sigma}^2/\sigma^2$ has a chi-square distribution with $n-2$ degrees of freedom, and $\hat{\beta}_1$ is independent of $\hat{\sigma}^2$. As a result of those properties, the statistic

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\hat{\sigma}^2/S_{xx}}} \tag{11-19}$$

follows the t distribution with $n-2$ degrees of freedom under H_0: $\beta_1 = \beta_{1,0}$. We would reject H_0: $\beta_1 = \beta_{1,0}$ if

$$|t_0| > t_{\alpha/2,n-2} \tag{11-20}$$

where t_0 is computed from Equation 11-19. The denominator of Equation 11-19 is the standard error of the slope, so we could write the test statistic as

$$T_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

A similar procedure can be used to test hypotheses about the intercept. To test

$$\begin{aligned} H_0&: \beta_0 = \beta_{0,0} \\ H_1&: \beta_0 \neq \beta_{0,0} \end{aligned} \tag{11-21}$$

we would use the statistic

$$T_0 = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}} = \frac{\hat{\beta}_0 - \beta_{0,0}}{se(\hat{\beta}_0)} \tag{11-22}$$

and reject the null hypothesis if the computed value of this test statistic, t_0, is such that $|t_0| > t_{\alpha/2,n-2}$. Note that the denominator of the test statistic in Equation 11-22 is just the standard error of the intercept.

A very important special case of the hypotheses of Equation 11-18 is

$$\begin{aligned} H_0&: \beta_1 = 0 \\ H_1&: \beta_1 \neq 0 \end{aligned} \tag{11-23}$$

These hypotheses relate to the **significance of regression.** Failure to reject H_0: $\beta_1 = 0$ is equivalent to concluding that there is no linear relationship between x and Y. This situation is illustrated in Fig. 11-5. Note that this may imply either that x is of little value in explaining the variation in Y and that the best estimator of Y for any x is $\hat{y} = \overline{Y}$ (Fig. 11-5*a*) or that the true relationship between x and Y is not linear (Fig. 11-5*b*). Alternatively, if H_0: $\beta_1 = 0$ is rejected, this implies that x is of value in explaining the variability in Y (see Fig. 11-6). Rejecting H_0: $\beta_1 = 0$ could mean either that the straight-line model is adequate (Fig. 11-6*a*) or that,

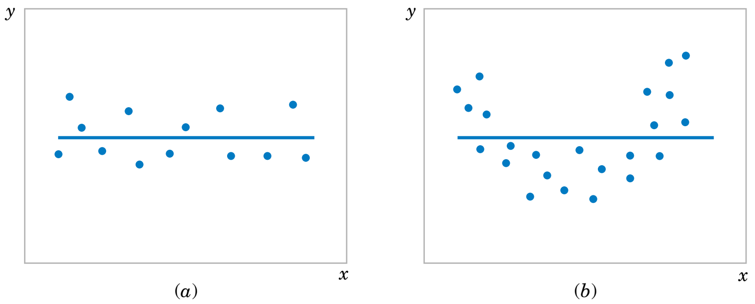

Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

although there is a linear effect of x, better results could be obtained with the addition of higher order polynomial terms in x (Fig. 11-6*b*).

EXAMPLE 11-2

We will test for significance of regression using the model for the oxygen purity data from Example 11-1. The hypotheses are

$$H_0\text{: } \beta_1 = 0$$
$$H_1\text{: } \beta_1 \neq 0$$

and we will use $\alpha = 0.01$. From Example 11-1 and Table 11-2 we have

$$\hat{\beta}_1 = 14.97 \quad n = 20, \quad S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$$

so the t-statistic in Equation 10-20 becomes

$$t_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{14.947}{\sqrt{1.18/0.68088}} = 11.35$$

Since the reference value of t is $t_{0.005,18} = 2.88$, the value of the test statistic is very far into the critical region, implying that H_0: $\beta_1 = 0$ should be rejected. The P-value for this test is $P \simeq 1.23 \times 10^{-9}$. This was obtained manually with a calculator.

Table 11-2 presents the Minitab output for this problem. Notice that the t-statistic value for the slope is computed as 11.35 and that the reported P-value is $P = 0.000$. Minitab also reports the t-statistic for testing the hypothesis H_0: $\beta_0 = 0$. This statistic is computed from Equation 11-22, with $\beta_{0,0} = 0$, as $t_0 = 46.62$. Clearly, then, the hypothesis that the intercept is zero is rejected.

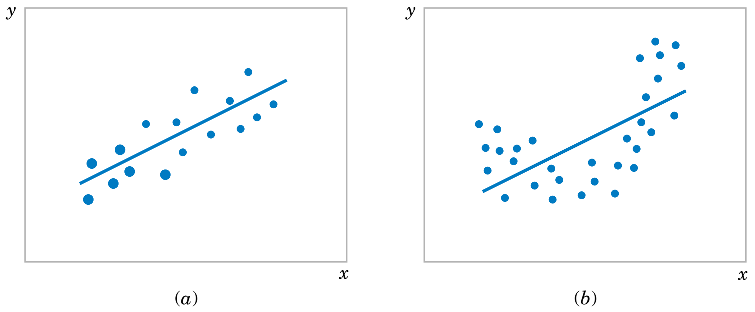

Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

11-5.2 Analysis of Variance Approach to Test Significance of Regression

A method called the **analysis of variance** can be used to test for significance of regression. The procedure partitions the total variability in the response variable into meaningful components as the basis for the test. The **analysis of variance identity** is as follows:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{11-24}$$

The two components on the right-hand-side of Equation 11-24 measure, respectively, the amount of variability in y_i accounted for by the regression line and the residual variation left unexplained by the regression line. We usually call $SS_E = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ the **error sum of squares** and $SS_R = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ the **regression sum of squares.** Symbolically, Equation 11-24 may be written as

$$SS_T = SS_R + SS_E \tag{11-25}$$

where $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the **total corrected sum of squares** of y. In Section 11-2 we noted that $SS_E = SS_T - \hat{\beta}_1 S_{xy}$ (see Equation 11-14), so since $SS_T = \hat{\beta}_1 S_{xy} + SS_E$, we note that the regression sum of squares in Equation 10-26 is $SS_R = \hat{\beta}_1 S_{xy}$. The total sum of squares SS_T has $n - 1$ degrees of freedom, and SS_R and SS_E have 1 and $n - 2$ degrees of freedom, respectively.

We may show that $E[SS_E/(n-2)] = \sigma^2$, $E(SS_R) = \sigma^2 + \beta_1^2 S_{xx}$ and that SS_E/σ^2 and SS_R/σ^2 are independent chi-square random variables with $n - 2$ and 1 degrees of freedom, respectively. Thus, if the null hypothesis H_0: $\beta_1 = 0$ is true, the statistic

$$F_0 = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E} \tag{11-26}$$

follows the $F_{1,n-2}$ distribution, and we would reject H_0 if $f_0 > f_{\alpha,1,n-2}$. The quantities $MS_R = SS_R/1$ and $MS_E = SS_E/(n-2)$ are called **mean squares.** In general, a mean square is always computed by dividing a sum of squares by its number of degrees of freedom. The test procedure is usually arranged in an **analysis of variance table,** such as Table 11-3.

Table 11-3 Analysis of Variance for Testing Significance of Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n - 2$	MS_E	
Total	SS_T	$n - 1$		

Note that $MS_E = \hat{\sigma}^2$.

EXAMPLE 11-3 We will use the analysis of variance approach to test for significance of regression using the oxygen purity data model from Example 11-1. Recall that $SS_T = 173.38$, $\hat{\beta}_1 = 14.947$, $S_{xy} = 10.17744$, and $n = 20$. The regression sum of squares is

$$SS_R = \hat{\beta}_1 S_{xy} = (14.947)10.17744 = 152.13$$

and the error sum of squares is

$$SS_E = SS_T - SS_R = 173.38 - 152.13 = 21.25$$

The analysis of variance for testing $H_0\colon \beta_1 = 0$ is summarized in the Minitab output in Table 11-2. The test statistic is $f_0 = MS_R/MS_E = 152.13/1.18 = 128.86$, for which we find that the P-value is $P \simeq 1.23 \times 10^{-9}$, so we conclude that β_1 is not zero.

There are frequently minor differences in terminology among computer packages. For example, sometimes the regression sum of squares is called the "model" sum of squares, and the error sum of squares is called the "residual" sum of squares.

Note that the analysis of variance procedure for testing for significance of regression is equivalent to the t-test in Section 11-5.1. That is, either procedure will lead to the same conclusions. This is easy to demonstrate by starting with the t-test statistic in Equation 11-19 with $\beta_{1,0} = 0$, say

$$T_0 = \frac{\hat{\beta}_1}{\sqrt{\hat{\sigma}^2/S_{xx}}} \qquad (11\text{-}27)$$

Squaring both sides of Equation 11-27 and using the fact that $\hat{\sigma}^2 = MS_E$ results in

$$T_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_E} = \frac{\hat{\beta}_1 S_{xY}}{MS_E} = \frac{MS_R}{MS_E} \qquad (11\text{-}28)$$

Note that T_0^2 in Equation 11-28 is identical to F_0 in Equation 11-26 It is true, in general, that the square of a t random variable with v degrees of freedom is an F random variable, with one and v degrees of freedom in the numerator and denominator, respectively. Thus, the test using T_0 is equivalent to the test based on F_0. Note, however, that the t-test is somewhat more flexible in that it would allow testing against a one-sided alternative hypothesis, while the F-test is restricted to a two-sided alternative.

EXERCISES FOR SECTION 11-5

11-18. Consider the data from Exercise 11-1 on x = compressive strength and y = intrinsic permeability of concrete.

(a) Test for significance of regression using $\alpha = 0.05$. Find the P-value for this test. Can you conclude that the model specifies a useful linear relationship between these two variables?

(b) Estimate σ^2 and the standard deviation of $\hat{\beta}_1$.

(c) What is the standard error of the intercept in this model?

11-19. Consider the data from Exercise 11-2 on x = roadway surface temperature and y = pavement deflection.

(a) Test for significance of regression using $\alpha = 0.05$. Find the P-value for this test. What conclusions can you draw?

(b) Estimate the standard errors of the slope and intercept.

11-20. Consider the National Football League data in Exercise 11-4.

(a) Test for significance of regression using $\alpha = 0.01$. Find the P-value for this test. What conclusions can you draw?

(b) Estimate the standard errors of the slope and intercept.

(c) Test (using $\alpha = 0.01$) $H_0\colon \beta_1 = -0.01$ versus $H_1\colon \beta_1 \neq -0.01$. Would you agree with the statement that this is a test of the claim that if you can decrease the opponent's rushing yardage by 100 yards the team will win one more game?

11-21. Consider the data from Exercise 11-5 on y = sales price and x = taxes paid.
(a) Test H_0: $\beta_1 = 0$ using the t-test; use $\alpha = 0.05$.
(b) Test H_0: $\beta_1 = 0$ using the analysis of variance with $\alpha = 0.05$. Discuss the relationship of this test to the test from part (a).
(c) Estimate the standard errors of the slope and intercept.
(d) Test the hypothesis that $\beta_0 = 0$.

11-22. Consider the data from Exercise 11-6 on y = steam usage and x = average temperature.
(a) Test for significance of regression using $\alpha = 0.01$. What is the P-value for this test? State the conclusions that result from this test.
(b) Estimate the standard errors of the slope and intercept.
(c) Test the hypothesis H_0: $\beta_1 = 10$ versus H_1: $\beta_1 \neq 10$ using $\alpha = 0.01$. Find the P-value for this test.
(d) Test H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ using $\alpha = 0.01$. Find the P-value for this test and draw conclusions.

11-23. Exercise 11-7 gave 20 observations on y = highway gasoline mileage and x = engine displacement.
(a) Test for significance of regression using $\alpha = 0.01$. Find the P-value for this test. What conclusions can you reach?
(b) Estimate the standard errors of the slope and intercept.
(c) Test H_0: $\beta_1 = -0.05$ versus H_1: $\beta_1 < -0.05$ using $\alpha = 0.01$ and draw conclusions. What is the P-value for this test?
(d) Test the hypothesis H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ using $\alpha = 0.01$. What is the P-value for this test?

11-24. Exercise 11-8 gave 13 observations on y = green liquor Na_2S concentration and x = production in a paper mill.
(a) Test for significance of regression using $\alpha = 0.05$. Find the P-value for this test.
(b) Estimate the standard errors of the slope and intercept.
(c) Test H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ using $\alpha = 0.05$. What is the P-value for this test?

11-25. Exercise 11-9 presented data on y = blood pressure rise and x = sound pressure level.
(a) Test for significance of regression using $\alpha = 0.05$. What is the P-value for this test?
(b) Estimate the standard errors of the slope and intercept.
(c) Test H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ using $\alpha = 0.05$. Find the P-value for this test.

11-26. Exercise 11-11 presented data on y = chloride concentration in surface streams and x = roadway area.
(a) Test the hypothesis H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ using the analysis of variance procedure with $\alpha = 0.01$.
(b) Find the P-value for the test in part (a).
(c) Estimate the standard errors of $\hat\beta_1$ and $\hat\beta_0$.
(d) Test H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ using $\alpha = 0.01$. What conclusions can you draw? Does it seem that the model might be a better fit to the data if the intercept were removed?

11-27. Refer to Exercise 11-12, which gives 20 observations on y = shear strength of a propellant and x = propellant age.
(a) Test for significance of regression with $\alpha = 0.01$. Find the P-value for this test.
(b) Estimate the standard errors of $\hat\beta_0$ and $\hat\beta_1$.
(c) Test H_0: $\beta_1 = -30$ versus H_1: $\beta_1 \neq -30$ using $\alpha = 0.01$. What is the P-value for this test?
(d) Test H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ using $\alpha = 0.01$. What is the P-value for this test?
(e) Test H_0: $\beta_0 = 2500$ versus H_1: $\beta_0 > 2500$ using $\alpha = 0.01$. What is the P-value for this test?

11-28. Suppose that each value of x_i is multiplied by a positive constant a, and each value of y_i is multiplied by another positive constant b. Show that the t-statistic for testing H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ is unchanged in value.

11-29. Consider the no-intercept model $Y = \beta x + \epsilon$ with the ϵ's NID$(0, \sigma^2)$. The estimate of σ^2 is $s^2 = \sum_{i=1}^{n}(y_i - \hat\beta x_i)^2/(n-1)$ and $V(\hat\beta) = \sigma^2/\sum_{i=1}^{n} x_i^2$.
(a) Devise a test statistic for H_0: $\beta = 0$ versus H_1: $\beta \neq 0$.
(b) Apply the test in (a) to the model from Exercise 11-17.

11-30. The type II error probability for the t-test for H_0: $\beta_1 = \beta_{1,0}$ can be computed in a similar manner to the t-tests of Chapter 9. If the true value of β_1 is β_1', the value $d = |\beta_{1,0} - \beta_1'|/(\sigma\sqrt{(n-1)/S_{xx}})$ is calculated and used as the horizontal scale factor on the operating characteristic curves for the t-test, (Appendix Charts VI*e* through VI*h*) and the type II error probability is read from the vertical scale using the curve for $n - 2$ degrees of freedom. Apply this procedure to the football data of Exercise 11-4, using $\sigma = 2.4$ and $\beta_1' = -0.005$, where the hypotheses are H_0: $\beta_1 = -0.01$ versus H_1: $\beta_1 \neq -0.01$.

11-6 CONFIDENCE INTERVALS

11-6.1 Confidence Intervals on the Slope and Intercept

In addition to point estimates of the slope and intercept, it is possible to obtain **confidence interval** estimates of these parameters. The width of these confidence intervals is a measure of the overall quality of the regression line. If the error terms, ϵ_i, in the regression model are normally and independently distributed,

$$(\hat{\beta}_1 - \beta_1)/\sqrt{\hat{\sigma}^2/S_{xx}} \quad \text{and} \quad (\hat{\beta}_0 - \beta_0)/\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}$$

are both distributed as t random variables with $n - 2$ degrees of freedom. This leads to the following definition of $100(1 - \alpha)\%$ confidence intervals on the slope and intercept.

Definition

Under the assumption that the observations are normally and independently distributed, a $100(1 - \alpha)\%$ **confidence interval on the slope** β_1 in simple linear regression is

$$\hat{\beta}_1 - t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{\alpha/2,n-2}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \tag{11-29}$$

Similarly, a $100(1 - \alpha)\%$ **confidence interval on the intercept** β_0 is

$$\hat{\beta}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \le \beta_0 \le \hat{\beta}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]} \tag{11-30}$$

EXAMPLE 11-4

We will find a 95% confidence interval on the slope of the regression line using the data in Example 11-1. Recall that $\hat{\beta}_1 = 14.947$, $S_{xx} = 0.68088$, and $\hat{\sigma}^2 = 1.18$ (see Table 11-2). Then, from Equation 10-31 we find

$$\hat{\beta}_1 - t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}} \le \beta_1 \le \hat{\beta}_1 + t_{0.025,18}\sqrt{\frac{\hat{\sigma}^2}{S_{xx}}}$$

or

$$14.947 - 2.101\sqrt{\frac{1.18}{0.68088}} \le \beta_1 \le 14.947 + 2.101\sqrt{\frac{1.18}{0.68088}}$$

This simplifies to

$$12.197 \le \beta_1 \le 17.697$$

11-6.2 Confidence Interval on the Mean Response

A confidence interval may be constructed on the mean response at a specified value of x, say, x_0. This is a confidence interval about $E(Y|x_0) = \mu_{Y|x_0}$ and is often called a confidence interval about the regression line. Since $E(Y|x_0) = \mu_{Y|x_0} = \beta_0 + \beta_1 x_0$, we may obtain a point estimate of the mean of Y at $x = x_0(\mu_{Y|x_0})$ from the fitted model as

$$\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

Now $\hat{\mu}_{Y|x_0}$ is an unbiased point estimator of $\mu_{Y|x_0}$, since $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators of β_0 and β_1. The variance of $\hat{\mu}_{Y|x_0}$ is

$$V(\hat{\mu}_{Y|x_0}) = \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

This last result follows from the fact that cov $(\bar{Y}, \hat{\beta}_1) = 0$ (Refer to Exercise 11-71). Also, $\hat{\mu}_{Y|x_0}$ is normally distributed, because $\hat{\beta}_1$ and $\hat{\beta}_0$ are normally distributed, and if we $\hat{\sigma}^2$ use as an estimate of σ^2, it is easy to show that

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}}$$

has a t distribution with $n - 2$ degrees of freedom. This leads to the following confidence interval definition.

Definition

A $100(1 - \alpha)\%$ **confidence interval about the mean response** at the value of $x = x_0$, say $\mu_{Y|x_0}$, is given by

$$\hat{\mu}_{Y|x_0} - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$
$$\leq \mu_{Y|x_0} \leq \hat{\mu}_{Y|x_0} + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \qquad (11\text{-}31)$$

where $\hat{\mu}_{Y|x_0} = \hat{\beta}_0 + \hat{\beta}_1 x_0$ is computed from the fitted regression model.

Note that the width of the confidence interval for $\mu_{Y|x_0}$ is a function of the value specified for x_0. The interval width is a minimum for $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases.

EXAMPLE 11-5

We will construct a 95% confidence interval about the mean response for the data in Example 11-1. The fitted model is $\hat{\mu}_{Y|x_0} = 74.283 + 14.947x_0$, and the 95% confidence interval on $\mu_{Y|x_0}$ is found from Equation 11-31 as

$$\hat{\mu}_{Y|x_0} \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(x_0 - 1.1960)^2}{0.68088}\right]}$$

Suppose that we are interested in predicting mean oxygen purity when $x_0 = 1.00\%$. Then

$$\hat{\mu}_{Y|x_{1.00}} = 74.283 + 14.947(1.00) = 89.23$$

and the 95% confidence interval is

$$\left\{89.23 \pm 2.101\sqrt{1.18\left[\frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}\right\}$$

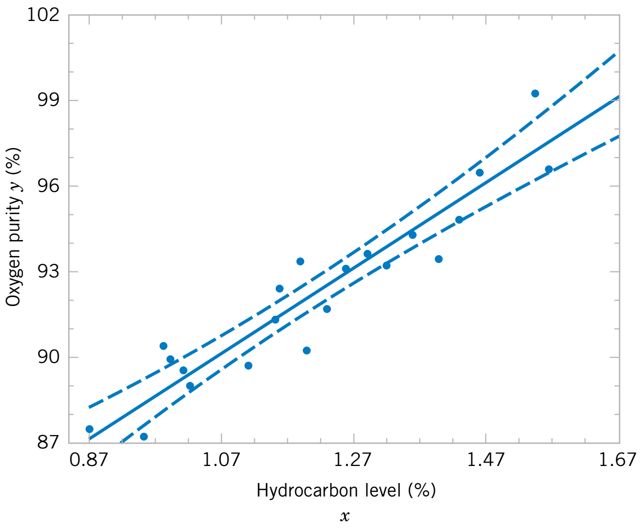

Figure 11-7 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line and 95 percent confidence limits on $\mu_{Y|x_0}$.

or

$$89.23 \pm 0.75$$

Therefore, the 95% confidence interval on $\mu_{Y|1.00}$ is

$$88.48 \leq \mu_{Y|1.00} \leq 89.98$$

Minitab will also perform these calculations. Refer to Table 11-2. The predicted value of y at $x = 1.00$ is shown along with the 95% CI on the mean of y at this level of x.

By repeating these calculations for several different values for x_0 we can obtain confidence limits for each corresponding value of $\mu_{Y|x_0}$. Figure 11-7 displays the scatter diagram with the fitted model and the corresponding 95% confidence limits plotted as the upper and lower lines. The 95% confidence level applies only to the interval obtained at one value of x and not to the entire set of x-levels. Notice that the width of the confidence interval on $\mu_{Y|x_0}$ increases as $|x_0 - \bar{x}|$ increases.

11-7 PREDICTION OF NEW OBSERVATIONS

An important application of a regression model is predicting new or future observations Y corresponding to a specified level of the regressor variable x. If x_0 is the value of the regressor variable of interest,

$$\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{11-32}$$

is the point estimator of the new or future value of the response Y_0.

Now consider obtaining an interval estimate for this future observation Y_0. This new observation is independent of the observations used to develop the regression model. Therefore, the confidence interval for $\mu_{Y|x_0}$ in Equation 11-31 is inappropriate, since it is based only on the data used to fit the regression model. The confidence interval about $\mu_{Y|x_0}$ refers to the true mean response at $x = x_0$ (that is, a population parameter), not to future observations.

Let Y_0 be the future observation at $x = x_0$, and let $\hat{Y}_0$ given by Equation 11-32 be the estimator of Y_0. Note that the error in prediction

$$e_{\hat{p}} = Y_0 - \hat{Y}_0$$

is a normally distributed random variable with mean zero and variance

$$V(e_{\hat{p}}) = V(Y_0 - \hat{Y}_0) = \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

because Y_0 is independent of $\hat{Y}_0$. If we use $\hat{\sigma}^2$ to estimate σ^2, we can show that

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}}$$

has a t distribution with $n - 2$ degrees of freedom. From this we can develop the following **prediction interval** definition.

Definition

A $100(1 - \alpha)$ % **prediction interval on a future observation** Y_0 at the value x_0 is given by

$$\hat{y}_0 - t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]}$$
$$\leq Y_0 \leq \hat{y}_0 + t_{\alpha/2,n-2}\sqrt{\hat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]} \qquad (11\text{-}33)$$

The value $\hat{y}_0$ is computed from the regression model $\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Notice that the prediction interval is of minimum width at $x_0 = \bar{x}$ and widens as $|x_0 - \bar{x}|$ increases. By comparing Equation 11-33 with Equation 11-31, we observe that the prediction interval at the point x_0 is always wider than the confidence interval at x_0. This results because the prediction interval depends on both the error from the fitted model and the error associated with future observations.

EXAMPLE 11-6 To illustrate the construction of a prediction interval, suppose we use the data in Example 11-1 and find a 95% prediction interval on the next observation of oxygen purity at $x_0 = 1.00\%$. Using Equation 11-33 and recalling from Example 11-5 that $\hat{y}_0 = 89.23$, we find that the prediction interval is

$$89.23 - 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$
$$\leq Y_0 \leq 89.23 + 2.101\sqrt{1.18\left[1 + \frac{1}{20} + \frac{(1.00 - 1.1960)^2}{0.68088}\right]}$$

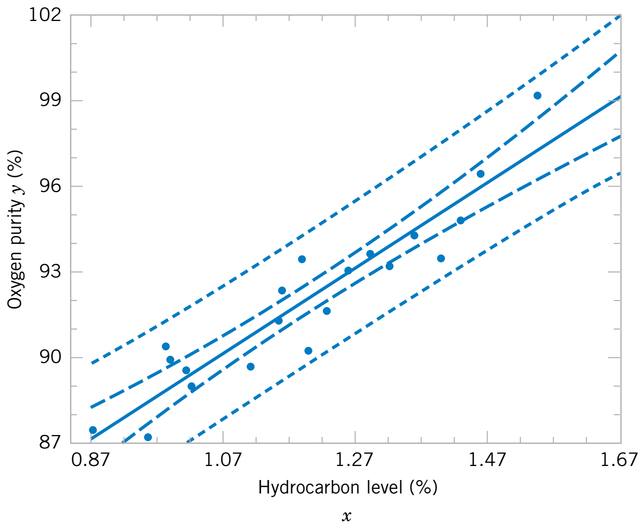

Figure 11-8 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line, 95% prediction limits (outer lines) and 95% confidence limits on $\mu_{Y|x_0}$.

which simplifies to

$$86.83 \leq y_0 \leq 91.63$$

Minitab will also calculate prediction intervals. Refer to the output in Table 11-2. The 95% PI on the future observation at $x_0 = 1.00$ is shown in the display.

By repeating the foregoing calculations at different levels of x_0, we may obtain the 95% prediction intervals shown graphically as the lower and upper lines about the fitted regression model in Fig. 11-8. Notice that this graph also shows the 95% confidence limits on $\mu_{Y|x_0}$ calculated in Example 11-5. It illustrates that the prediction limits are always wider than the confidence limits.

EXERCISES FOR SECTIONS 11-6 AND 11-7

11-31. Refer to the data in Exercise 11-1 on y = intrinsic permeability of concrete and x = compressive strength. Find a 95% confidence interval on each of the following:
(a) Slope (b) Intercept
(c) Mean permeability when $x = 2.5$
(d) Find a 95% prediction interval on permeability when $x = 2.5$. Explain why this interval is wider than the interval in part (c).

11-32. Exercise 11-2 presented data on roadway surface temperature x and pavement deflection y. Find a 99% confidence interval on each of the following:
(a) Slope (b) Intercept
(c) Mean deflection when temperature $x = 85$°F
(d) Find a 99% prediction interval on pavement deflection when the temperature is 90°F.

11-33. Exercise 11-4 presented data on the number of games won by NFL teams in 1976. Find a 95% confidence interval on each of the following:
(a) Slope (b) Intercept
(c) Mean number of games won when opponents rushing yardage is limited to $x = 1800$
(d) Find a 95% prediction interval on the number of games won when opponents rushing yards is 1800.

11-34. Refer to the data on y = house selling price and x = taxes paid in Exercise 11-5. Find a 95% confidence interval on each of the following:
(a) β_1 (b) β_0
(c) Mean selling price when the taxes paid are $x = 7.50$
(d) Compute the 95% prediction interval for selling price when the taxes paid are $x = 7.50$.

11-35. Exercise 11-6 presented data on y = steam usage and x = monthly average temperature.
(a) Find a 99% confidence interval for β_1.
(b) Find a 99% confidence interval for β_0.
(c) Find a 95% confidence interval on mean steam usage when the average temperature is 55°F.

(d) Find a 95% prediction interval on steam usage when temperature is 55°F. Explain why this interval is wider than the interval in part (c).

11-36. Exercise 11-7 presented gasoline mileage performance for 20 cars, along with information about the engine displacement. Find a 95% confidence interval on each of the following:

(a) Slope (b) Intercept

(c) Mean highway gasoline mileage when the engine displacement is $x = 150$ in^3

(d) Construct a 95% prediction interval on highway gasoline mileage when the engine displacement is $x = 150$ in^3.

11-37. Consider the data in Exercise 11-8 on y = green liquor Na_2S concentration and x = production in a paper mill. Find a 99% confidence interval on each of the following:

(a) β_1 (b) β_0

(c) Mean Na_2S concentration when production $x = 910$ tons/day

(d) Find a 99% prediction interval on Na_2S concentration when $x = 910$ tons/day.

11-38. Exercise 11-9 presented data on y = blood pressure rise and x = sound pressure level. Find a 95% confidence interval on each of the following:

(a) β_1 (b) β_0

(c) Mean blood pressure rise when the sound pressure level is 85 decibals

(d) Find a 95% prediction interval on blood pressure rise when the sound pressure level is 85 decibals.

11-39. Refer to the data in Exercise 11-10 on y = wear volume of mild steel and x = oil viscosity. Find a 95% confidence interval on each of the following:

(a) Intercept (b) Slope

(c) Mean wear when oil viscosity $x = 30$

11-40. Exercise 11-11 presented data on chloride concentration y and roadway area x on watersheds in central Rhode Island. Find a 99% confidence interval on each of the following:

(a) β_1 (b) β_0

(c) Mean chloride concentration when roadway area $x = 1.0\%$

(d) Find a 99% prediction interval on chloride concentration when roadway area $x = 1.0\%$.

11-41. Refer to the data in Exercise 11-12 on rocket motor shear strength y and propellant age x. Find a 95% confidence interval on each of the following:

(a) Slope β_1 (b) Intercept β_0

(c) Mean shear strength when age $x = 20$ weeks

(d) Find a 95% prediction interval on shear strength when age $x = 20$ weeks.

11-8 ADEQUACY OF THE REGRESSION MODEL

Fitting a regression model requires several **assumptions.** Estimation of the model parameters requires the assumption that the errors are uncorrelated random variables with mean zero and constant variance. Tests of hypotheses and interval estimation require that the errors be normally distributed. In addition, we assume that the order of the model is correct; that is, if we fit a simple linear regression model, we are assuming that the phenomenon actually behaves in a linear or first-order manner.

The analyst should always consider the validity of these assumptions to be doubtful and conduct analyses to examine the adequacy of the model that has been tentatively entertained. In this section we discuss methods useful in this respect.

11-8.1 Residual Analysis

The **residuals** from a regression model are $e_i = y_i - \hat{y}_i$, $i = 1, 2, \ldots, n$, where y_i is an actual observation and $\hat{y}_i$ is the corresponding fitted value from the regression model. Analysis of the residuals is frequently helpful in checking the assumption that the errors are approximately normally distributed with constant variance, and in determining whether additional terms in the model would be useful.

As an approximate check of normality, the experimenter can construct a frequency histogram of the residuals or a **normal probability plot of residuals.** Many computer programs will produce a normal probability plot of residuals, and since the sample sizes in regression are often too small for a histogram to be meaningful, the normal probability plotting method

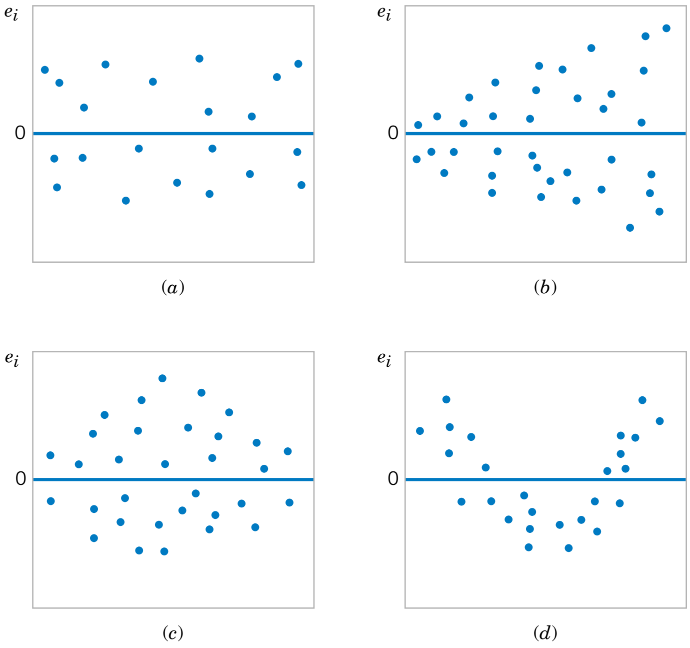

Figure 11-9 Patterns for residual plots. (a) satisfactory, (b) funnel, (c) double bow, (d) nonlinear. [Adapted from Montgomery, Peck, and Vining (2001).]

is preferred. It requires judgment to assess the abnormality of such plots. (Refer to the discussion of the "fat pencil" method in Section 6-7).

We may also **standardize** the residuals by computing $d_i = e_i/\sqrt{\hat{\sigma}^2}$, $i = 1, 2 \ldots, n$. If the errors are normally distributed, approximately 95% of the standardized residuals should fall in the interval $(-2, +2)$. Residuals that are far outside this interval may indicate the presence of an **outlier,** that is, an observation that is not typical of the rest of the data. Various rules have been proposed for discarding outliers. However, outliers sometimes provide important information about unusual circumstances of interest to experimenters and should not be automatically discarded. For further discussion of outliers, see Montgomery, Peck and Vining (2001).

It is frequently helpful to plot the residuals (1) in time sequence (if known), (2), against the $\hat{y}_i$, and (3) against the independent variable x. These graphs will usually look like one of the four general patterns shown in Fig. 11-9. Pattern (*a*) in Fig. 11-9 represents the ideal situation, while patterns (*b*), (*c*), and (*d*) represent anomalies. If the residuals appear as in (*b*), the variance of the observations may be increasing with time or with the magnitude of y_i or x_i. Data transformation on the response y is often used to eliminate this problem. Widely used variance-stabilizing transformations include the use of $\sqrt{y}$, ln y, or $1/y$ as the response. See Montgomery, Peck, and Vining (2001) for more details regarding methods for selecting an appropriate transformation. If a plot of the residuals against time has the appearance of (*b*), the variance of the observations is increasing with time. Plots of residuals against $\hat{y}_i$ and x_i that look like (*c*) also indicate inequality of variance. Residual plots that look like (*d*) indicate model inadequacy; that is, higher order terms should be added to the model, a transformation on the x-variable or the y-variable (or both) should be considered, or other regressors should be considered.

EXAMPLE 11-7

The regression model for the oxygen purity data in Example 11-1 is $\hat{y} = 74.283 + 14.947x$. Table 11-4 presents the observed and predicted values of y at each value of x from this data set, along with the corresponding residual. These values were computed using Minitab and show

Table 11-4 Oxygen Purity Data from Example 11-1, Predicted Values, and Residuals

	Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, $\hat{y}$	Residual $e = y - \hat{y}$		Hydrocarbon Level, x	Oxygen Purity, y	Predicted Value, $\hat{y}$	Residual $e = y - \hat{y}$
1	0.99	90.01	89.069009	0.940991	11	1.19	93.54	92.063189	1.476811
2	1.02	89.05	89.518136	−0.468136	12	1.15	92.52	91.614062	0.905938
3	1.15	91.43	91.464353	−0.034353	13	0.98	90.56	88.919300	1.640700
4	1.29	93.74	93.560279	0.179721	14	1.01	89.54	89.368427	0.171573
5	1.46	96.73	96.105332	0.624668	15	1.11	89.85	90.865517	−1.015517
6	1.36	94.45	94.608242	−0.158242	16	1.20	90.39	92.212898	−1.822898
7	0.87	87.59	87.272501	0.317499	17	1.26	93.25	93.111152	0.138848
8	1.23	91.77	92.662025	−0.892025	18	1.32	93.41	94.009406	−0.599406
9	1.55	99.42	97.452713	1.967287	19	1.43	94.98	95.656205	−0.676205
10	1.40	93.65	95.207078	−1.557078	20	0.95	87.33	88.470173	−1.140173

the number of decimal places typical of computer output. A normal probability plot of the residuals is shown in Fig. 11-10. Since the residuals fall approximately along a straight line in the figure, we conclude that there is no severe departure from normality. The residuals are also plotted against the predicted value $\hat{y}_i$ in Fig. 11-11 and against the hydrocarbon levels x_i in Fig. 11-12. These plots do not indicate any serious model inadequacies.

11-8.2 Coefficient of Determination (R^2)

The quantity

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \tag{11-34}$$

is called the **coefficient of determination** and is often used to judge the adequacy of a regression model. Subsequently, we will see that in the case where X and Y are jointly distributed random variables, R^2 is the square of the correlation coefficient between X and Y. From

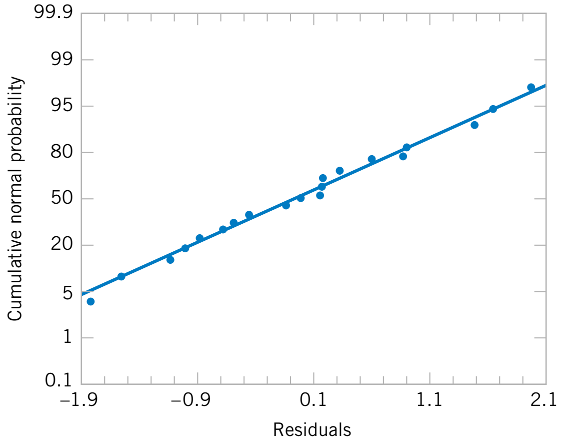

Figure 11-10 Normal probability plot of residuals, Example 11-7.

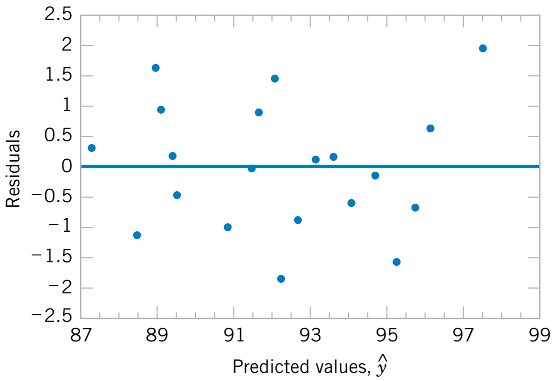

Figure 11-11 Plot of residuals versus predicted oxygen purity $\hat{y}$, Example 11-7.

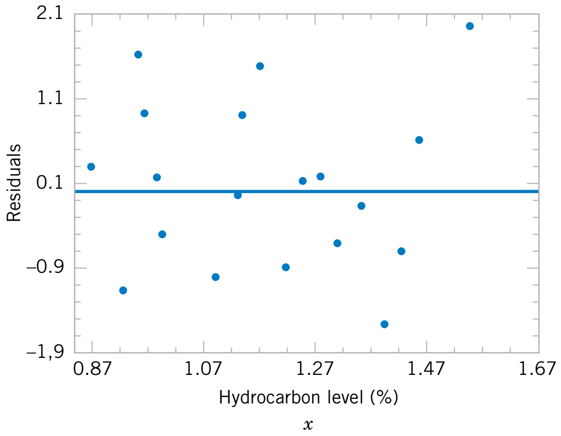

Figure 11-12 Plot of residuals versus hydrocarbon level x, Example 11-8.

the analysis of variance identity in Equations 11-24 and 11-25, $0 \le R^2 \le 1$. We often refer loosely to R^2 as the amount of variability in the data explained or accounted for by the regression model. For the oxygen purity regression model, we have $R^2 = SS_R/SS_T = 152.13/173.38 = 0.877$; that is, the model accounts for 87.7% of the variability in the data.

The statistic R^2 should be used with caution, because it is always possible to make R^2 unity by simply adding enough terms to the model. For example, we can obtain a "perfect" fit to n data points with a polynomial of degree $n - 1$. In addition, R^2 will always increase if we add a variable to the model, but this does not necessarily imply that the new model is superior to the old one. Unless the error sum of squares in the new model is reduced by an amount equal to the original error mean square, the new model will have a larger error mean square than the old one, because of the loss of one error degree of freedom. Thus, the new model will actually be worse than the old one.

There are several misconceptions about R^2. In general, R^2 does not measure the magnitude of the slope of the regression line. A large value of R^2 does not imply a steep slope. Furthermore, R^2 does not measure the appropriateness of the model, since it can be artificially inflated by adding higher order polynomial terms in x to the model. Even if y and x are related in a nonlinear fashion, R^2 will often be large. For example, R^2 for the regression equation in Fig. 11-6(*b*) will be relatively large, even though the linear approximation is poor. Finally, even though R^2 is large, this does not necessarily imply that the regression model will provide accurate predictions of future observations.

11-8.3 Lack-of-Fit Test (CD Only)

EXERCISES FOR SECTION 11-8

11-42. Refer to the NFL team performance data in Exercise 11-4.

(a) Calculate R^2 for this model and provide a practical interpretation of this quantity.

(b) Prepare a normal probability plot of the residuals from the least squares model. Does the normality assumption seem to be satisfied?

(c) Plot the residuals versus $\hat{y}$ and against x. Interpret these graphs.

11-43. Refer to the data in Exercise 11-5 on house selling price y and taxes paid x.

(a) Find the residuals for the least squares model.

(b) Prepare a normal probability plot of the residuals and interpret this display.

(c) Plot the residuals versus $\hat{y}$ and versus x. Does the assumption of constant variance seem to be satisfied?

(d) What proportion of total variability is explained by the regression model?

11-44. Exercise 11-6 presents data on y = steam usage and x = average monthly temperature.

(a) What proportion of total variability is accounted for by the simple linear regression model?

(b) Prepare a normal probability plot of the residuals and interpret this graph.

(c) Plot residuals versus $\hat{y}$ and x. Do the regression assumptions appear to be satisfied?

11-45. Refer to the gasoline mileage data in Exercise 11-7.

(a) What proportion of total variability in highway gasoline mileage performance is accounted for by engine displacement?

(b) Plot the residuals versus $\hat{y}$ and x, and comment on the graphs.

(c) Prepare a normal probability plot of the residuals. Does the normality assumption appear to be satisfied?

11-46. Consider the data in Exercise 11-8 on y = green liquor Na_2S concentration and x = paper machine production. Suppose that a 14th sample point is added to the original data, where $y_{14} = 59$ and $x_{14} = 855$.

(a) Prepare a scatter diagram of y versus x. Fit the simple linear regression model to all 14 observations.

(b) Test for significance of regression with $\alpha = 0.05$.

(c) Estimate σ^2 for this model.

(d) Compare the estimate of σ^2 obtained in part (c) above with the estimate of σ^2 obtained from the original 13 points. Which estimate is larger and why?

(e) Compute the residuals for this model. Does the value of e_{14} appear unusual?

(f) Prepare and interpret a normal probability plot of the residuals.

(g) Plot the residuals versus $\hat{y}$ and versus x. Comment on these graphs.

11-47. Refer to Exercise 11-9, which presented data on blood pressure rise y and sound pressure level x.

(a) What proportion of total variability in blood pressure rise is accounted for by sound pressure level?

(b) Prepare a normal probability plot of the residuals from this least squares model. Interpret this plot.

(c) Plot residuals versus $\hat{y}$ and versus x. Comment on these plots.

11-48. Exercise 11-10 presents data on wear volume y and oil viscosity x.

(a) Calculate R^2 for this model. Provide an interpretation of this quantity.

(b) Plot the residuals from this model versus $\hat{y}$ and versus x. Interpret these plots.

(c) Prepare a normal probability plot of the residuals. Does the normality assumption appear to be satisfied?

11-49. Refer to Exercise 11-11, which presented data on chloride concentration y and roadway area x.

(a) What proportion of the total variability in chloride concentration is accounted for by the regression model?

(b) Plot the residuals versus $\hat{y}$ and versus x. Interpret these plots.

(c) Prepare a normal probability plot of the residuals. Does the normality assumption appear to be satisfied?

11-50. Consider the rocket propellant data in Exercise 11-12.

(a) Calculate R^2 for this model. Provide an interpretation of this quantity.

(b) Plot the residuals on a normal probability scale. Do any points seem unusual on this plot?

(c) Delete the two points identified in part (b) from the sample and fit the simple linear regression model to the remaining 18 points. Calculate the value of R^2 for the new model. Is it larger or smaller than the value of R^2 computed in part (a)? Why?

(d) Did the value of $\hat{\sigma}^2$ change dramatically when the two points identified above were deleted and the model fit to the remaining points? Why?

11-51. Show that an equivalent way to define the test for significance of regression in simple linear regression is to base the test on R^2 as follows: to test H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$, calculate

$$F_0 = \frac{R^2(n-2)}{1-R^2}$$

and to reject H_0: $\beta_1 = 0$ if the computed value $f_0 > f_{\alpha,1,n-2}$.

11-52. Suppose that a simple linear regression model has been fit to $n = 25$ observations and $R^2 = 0.90$.

(a) Test for significance of regression at $\alpha = 0.05$. Use the results of Exercise 11-51.

(b) What is the smallest value of R^2 that would lead to the conclusion of a significant regression if $\alpha = 0.05$?

11-53. Consider the rocket propellant data in Exercise 11-12. Calculate the standardized residuals for these data. Does this provide any helpful information about the magnitude of the residuals?

11-54. Studentized Residuals. Show that the variance of the ith residual is

$$V(e_i) = \sigma^2\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]$$

Hint:

$$\text{cov}(Y_i, \hat{Y}_i) = \sigma^2\left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right].$$

The ith studentized residual is defined as

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2\left[1 - \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]}}$$

(a) Explain why r_i has unit standard deviation.

(b) Do the **standardized residuals** have unit standard deviation?

(c) Discuss the behavior of the studentized residual when the sample value x_i is very close to the middle of the range of x.

(d) Discuss the behavior of the studentized residual when the sample value x_i is very near one end of the range of x.

11-9 TRANSFORMATIONS TO A STRAIGHT LINE

We occasionally find that the straight-line regression model $Y = \beta_0 + \beta_1 x + \epsilon$ is inappropriate because the true regression function is nonlinear. Sometimes nonlinearity is visually determined from the scatter diagram, and sometimes, because of prior experience or underlying theory, we know in advance that the model is nonlinear. Occasionally, a scatter diagram will exhibit an apparent nonlinear relationship between Y and x. In some of these situations, a nonlinear function can be expressed as a straight line by using a suitable transformation. Such nonlinear models are called **intrinsically linear.**

As an example of a nonlinear model that is intrinsically linear, consider the exponential function

$$Y = \beta_0 e^{\beta_1 x}\epsilon$$

This function is intrinsically linear, since it can be transformed to a straight line by a logarithmic transformation

$$\ln Y = \ln \beta_0 + \beta_1 x + \ln \epsilon$$

This transformation requires that the transformed error terms $\ln \epsilon$ are normally and independently distributed with mean 0 and variance σ^2.

Another intrinsically linear function is

$$Y = \beta_0 + \beta_1\left(\frac{1}{x}\right) + \epsilon$$

By using the reciprocal transformation $z = 1/x$, the model is linearized to

$$Y = \beta_0 + \beta_1 z + \epsilon$$

Sometimes several transformations can be employed jointly to linearize a function. For example, consider the function

$$Y = \frac{1}{\exp(\beta_0 + \beta_1 x + \epsilon)}$$

letting $Y^* = 1/Y$, we have the linearized form

$$\ln Y^* = \beta_0 + \beta_1 x + \epsilon$$

For examples of fitting these models, refer to Montgomery, Peck, and Vining (2001) or Myers (1990).

11-10 MORE ABOUT TRANSFORMATIONS (CD ONLY)

11-11 CORRELATION

Our development of regression analysis has assumed that x is a mathematical variable, measured with negligible error, and that Y is a random variable. Many applications of regression analysis involve situations in which both X and Y are random variables. In these situations, it is usually assumed that the observations (X_i, Y_i), $i = 1, 2, \ldots, n$ are jointly distributed random variables obtained from the distribution $f(x, y)$.

For example, suppose we wish to develop a regression model relating the shear strength of spot welds to the weld diameter. In this example, weld diameter cannot be controlled. We would randomly select n spot welds and observe a diameter (X_i) and a shear strength (Y_i) for each. Therefore (X_i, Y_i) are jointly distributed random variables.

We assume that the joint distribution of X_i and Y_i is the bivariate normal distribution presented in Chapter 5, and μ_Y and σ_Y^2 are the mean and variance of Y, μ_X and σ_X^2 are the mean and variance of X, and ρ is the **correlation coefficient** between Y and X. Recall that the correlation coefficient is defined as

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{11-35}$$

where σ_{XY} is the covariance between Y and X.

The conditional distribution of Y for a given value of $X = x$ is

$$f_{Y|x}(y) = \frac{1}{\sqrt{2\pi}\sigma_{Y|x}} \exp\left[-\frac{1}{2}\left(\frac{y - \beta_0 - \beta_1 x}{\sigma_{Y|x}}\right)^2\right] \tag{11-36}$$

where

$$\beta_0 = \mu_Y - \mu_X \rho \frac{\sigma_Y}{\sigma_X} \tag{11-37}$$

$$\beta_1 = \frac{\sigma_Y}{\sigma_X}\rho \tag{11-38}$$

and the variance of the conditional distribution of Y given $X = x$ is

$$\sigma_{Y|x}^2 = \sigma_Y^2(1 - \rho^2) \tag{11-39}$$

That is, the conditional distribution of Y given $X = x$ is normal with mean

$$E(Y|x) = \beta_0 + \beta_1 x \tag{11-40}$$

and variance $\sigma_{Y|x}^2$. Thus, the mean of the conditional distribution of Y given $X = x$ is a simple linear regression model. Furthermore, there is a relationship between the correlation coefficient ρ and the slope β_1. From Equation 11-38 we see that if $\rho = 0$, then $\beta_1 = 0$, which implies that there is no regression of Y on X. That is, knowledge of X does not assist us in predicting Y.

The method of maximum likelihood may be used to estimate the parameters β_0 and β_1. It can be shown that the maximum likelihood estimators of those parameters are

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{11-41}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} Y_i(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{S_{XY}}{S_{XX}} \tag{11-42}$$

We note that the estimators of the intercept and slope in Equations 11-41 and 11-42 are identical to those given by the method of least squares in the case where X was assumed to be a mathematical variable. That is, the regression model with Y and X jointly normally distributed is equivalent to the model with X considered as a mathematical variable. This follows because the random variables Y given $X = x$ are independently and normally distributed with mean $\beta_0 + \beta_1 x$ and constant variance $\sigma^2_{Y|x}$. These results will also hold for any joint distribution of Y and X such that the conditional distribution of Y given X is normal.

It is possible to draw inferences about the correlation coefficient ρ in this model. The estimator of ρ is the **sample correlation coefficient**

$$R = \frac{\sum_{i=1}^{n} Y_i(X_i - \overline{X})}{\left[\sum_{i=1}^{n}(X_i - \overline{X})^2 \sum_{i=1}^{n}(Y_i - \overline{Y})^2\right]^{1/2}} = \frac{S_{XY}}{(S_{XX}SS_T)^{1/2}} \tag{11-43}$$

Note that

$$\hat{\beta}_1 = \left(\frac{SS_T}{S_{XX}}\right)^{1/2} R \tag{11-44}$$

so the slope $\hat{\beta}_1$ is just the sample correlation coefficient R multiplied by a scale factor that is the square root of the "spread" of the Y values divided by the "spread" of the X values.Thus, $\hat{\beta}_1$ and R are closely related, although they provide somewhat different information. The sample correlation coefficient R measures the linear association between Y and X, while $\hat{\beta}_1$ measures the predicted change in the mean of Y for a unit change in X. In the case of a mathematical variable x, R has no meaning because the magnitude of R depends on the choice of spacing of x. We may also write, from Equation 11-44,

$$R^2 = \hat{\beta}_1^2 \frac{S_{XX}}{S_{YY}} = \frac{\hat{\beta}_1 S_{XY}}{SS_T} = \frac{SS_R}{SS_T}$$

which is just the coefficient of determination. That is, the coefficient of determination R^2 is just the square of the correlation coefficient between Y and X.

It is often useful to test the hypotheses

$$\begin{aligned} H_0&: \rho = 0 \\ H_1&: \rho \neq 0 \end{aligned} \tag{11-45}$$

The appropriate test statistic for these hypotheses is

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \tag{11-46}$$

which has the t distribution with $n - 2$ degrees of freedom if H_0: $\rho = 0$ is true. Therefore, we would reject the null hypothesis if $|t_0| > t_{\alpha/2,n-2}$. This test is equivalent to the test of the hypothesis H_0: $\beta_1 = 0$ given in Section 11-6.1. This equivalence follows directly from Equation 10-51.

The test procedure for the hypothesis is

$$H_0: \rho = \rho_0$$
$$H_1: \rho \neq \rho_0 \tag{11-47}$$

where $\rho_0 \neq 0$ is somewhat more complicated. For moderately large samples (say, $n \geq 25$) the statistic

$$Z = \text{arctanh } R = \frac{1}{2} \ln \frac{1 + R}{1 - R} \tag{11-48}$$

is approximately normally distributed with mean and variance

$$\mu_Z = \text{arctanh } \rho = \frac{1}{2} \ln \frac{1 + \rho}{1 - \rho} \qquad \text{and} \qquad \sigma_Z^2 = \frac{1}{n - 3}$$

respectively. Therefore, to test the hypothesis H_0: $\rho = \rho_0$, we may use the test statistic

$$Z_0 = (\text{arctanh } R - \text{arctanh } \rho_0)(n - 3)^{1/2} \tag{11-49}$$

and reject H_0: $\rho = \rho_0$ if the value of the test statistic in Equation 11-49 is such that $|z_0| > z_{\alpha/2}$.

It is also possible to construct an approximate $100(1 - \alpha)\%$ confidence interval for ρ, using the transformation in Equation 10-55. The approximate $100(1 - \alpha)\%$ confidence interval is

$$\tanh\left(\text{arctanh } r - \frac{z_{\alpha/2}}{\sqrt{n - 3}}\right) \leq \rho \leq \tanh\left(\text{arctanh } r + \frac{z_{\alpha/2}}{\sqrt{n - 3}}\right) \tag{11-50}$$

where $\tanh u = (e^u - e^{-u})/(e^u + e^{-u})$.

EXAMPLE 11-8

In Chapter 1 (Section 1-3) an application of regression analysis is described in which an engineer at a semiconductor assembly plant is investigating the relationship between pull strength of a wire bond and two factors: wire length and die height. In this example, we will consider only one of the factors, the wire length. A random sample of 25 units is selected and tested, and the wire bond pull strength and wire length are observed for each unit. The data are shown in Table 1-2. We assume that pull strength and wire length are jointly normally distributed.

Figure 11-13 shows a scatter diagram of wire bond strength versus wire length. We have used the Minitab option of displaying box plots of each individual variable on the scatter diagram. There is evidence of a linear relationship between the two variables.

The Minitab output for fitting a simple linear regression model to the data is shown on the following page.

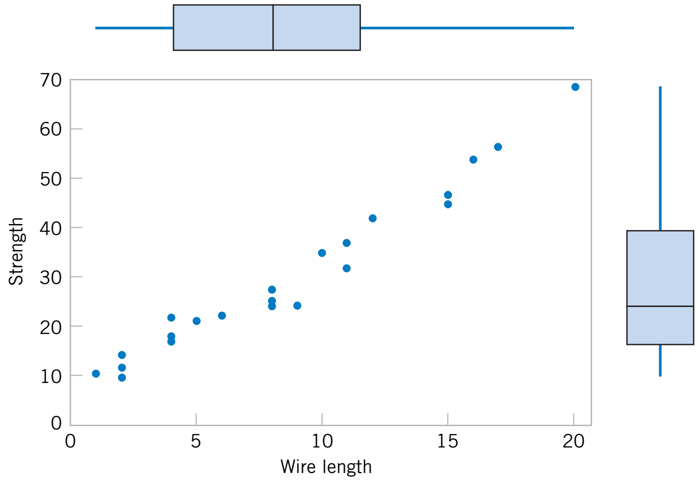

Figure 11-13 Scatter plot of wire bond strength versus wire length, Example 11-8.

```
Regression Analysis: Strength versus Length

The regression equation is
Strength = 5.11 + 2.90 Length

Predictor       Coef       SE Coef        T          P
Constant        5.115       1.146        4.46      0.000
Length          2.9027      0.1170      24.80      0.000

S = 3.093               R-Sq = 96.4%                R-Sq(adj) = 96.2%
PRESS = 272.144         R-Sq(pred) = 95.54%

Analysis of Variance

Source            DF       SS         MS          F          P
Regression         1     5885.9     5885.9     615.08      0.000
Residual Error    23      220.1        9.6
Total             24     6105.9
```

Now $S_{xx} = 698.56$ and $S_{xy} = 2027.7132$, and the sample correlation coefficient is

$$r = \frac{S_{xy}}{[S_{xx}SS_T]^{1/2}} = \frac{2027.7132}{[(698.560)(6105.9)]^{1/2}} = 0.9818$$

Note that $r^2 = (0.9818)^2 = 0.9640$ (which is reported in the Minitab output), or that approximately 96.40% of the variability in pull strength is explained by the linear relationship to wire length.

Now suppose that we wish to test the hypothesis

$$H_0: \rho = 0$$
$$H_1: \rho \neq 0$$

with $\alpha = 0.05$. We can compute the t-statistic of Equation 11-46 as

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.9818\sqrt{23}}{\sqrt{1-0.9640}} = 24.8$$

This statistic is also reported in the Minitab output as a test of H_0: $\beta_1 = 0$. Because $t_{0.025,23} = 2.069$, we reject H_0 and conclude that the correlation coefficient $\rho \neq 0$.

Finally, we may construct an approximate 95% confidence interval on ρ from Equation 10-57. Since arctanh r = arctanh 0.9818 = 2.3452, Equation 11-50 becomes

$$\tanh\left(2.3452 - \frac{1.96}{\sqrt{22}}\right) \leq \rho \leq \tanh\left(2.3452 + \frac{1.96}{\sqrt{22}}\right)$$

which reduces to

$$0.9585 \leq \rho \leq 0.9921$$

EXERCISES FOR SECTION 11–10

11-55. The final test and exam averages for 20 randomly selected students taking a course in engineering statistics and a course in operations research follow. Assume that the final averages are jointly normally distributed.

(a) Find the regression line relating the statistics final average to the OR final average.

(b) Test for significance of regression using $\alpha = 0.05$.

Statistics	86	75	69	75	90
OR	80	81	75	81	92
Statistics	94	83	86	71	65
OR	95	80	81	76	72
Statistics	84	71	62	90	83
OR	85	72	65	93	81
Statistics	75	71	76	84	97
OR	70	73	72	80	98

(c) Estimate the correlation coefficient.

(d) Test the hypothesis that $\rho = 0$, using $\alpha = 0.05$.

(e) Test the hypothesis that $\rho = 0.5$, using $\alpha = 0.05$.

(f) Construct a 95% confidence interval for the correlation coefficient.

11-56. The weight and systolic blood pressure of 26 randomly selected males in the age group 25 to 30 are shown in the following table. Assume that weight and blood pressure are jointly normally distributed.

(a) Find a regression line relating systolic blood pressure to weight.

(b) Test for significance of regression using $\alpha = 0.05$.

Subject	Weight	Systolic BP	Subject	Weight	Systolic BP
1	165	130	14	172	153
2	167	133	15	159	128
3	180	150	16	168	132
4	155	128	17	174	149
5	212	151	18	183	158
6	175	146	19	215	150
7	190	150	20	195	163
8	210	140	21	180	156
9	200	148	22	143	124
10	149	125	23	240	170
11	158	133	24	235	165
12	169	135	25	192	160
13	170	150	26	187	159

(c) Estimate the correlation coefficient.

(d) Test the hypothesis that $\rho = 0$, using $\alpha = 0.05$.

(e) Test the hypothesis that $\rho = 0.6$, using $\alpha = 0.05$.

(f) Construct a 95% confidence interval for the correlation coefficient.

11-57. Consider the NFL data introduced in Exercise 11-4.

(a) Estimate the correlation coefficient between the number of games won and the yards rushing by the opponents.

(b) Test the hypothesis H_0: $\rho = 0$ versus H_1: $\rho \neq 0$ using $\alpha = 0.05$. What is the P-value for this test?

(c) Construct a 95% confidence interval for ρ.

(d) Test the hypothesis H_0: $\rho = -0.7$ versus H_1: $\rho \neq -0.7$ using $\alpha = 0.05$. Find the P-value for this test.

11-58. Show that the t-statistic in Equation 11-46 for testing H_0: $\rho = 0$ is identical to the t-statistic for testing H_0: $\beta_1 = 0$.

11-59. A random sample of 50 observations was made on the diameter of spot welds and the corresponding weld shear strength.

(a) Given that $r = 0.62$, test the hypothesis that $\rho = 0$, using $\alpha = 0.01$. What is the P-value for this test?

(b) Find a 99% confidence interval for ρ.

(c) Based on the confidence interval in part (b), can you conclude that $\rho = 0.5$ at the 0.01 level of significance?

11-60. Suppose that a random sample of 10,000 (X, Y) pairs yielded a sample correlation coefficient of $r = 0.02$.

(a) What is the conclusion that you would reach if you tested H_0: $\rho = 0$ using $\alpha = 0.05$? What is the P-value for this test?

(b) Comment on the practical significance versus the statistical significance of your answer.

11-61. The following data gave X = the water content of snow on April 1 and Y = the yield from April to July (in inches) on the Snake River watershed in Wyoming for 1919 to 1935. (The data were taken from an article in *Research Notes,* Vol. 61, 1950, Pacific Northwest Forest Range Experiment Station, Oregon)

x	y	x	y
23.1	10.5	37.9	22.8
32.8	16.7	30.5	14.1
31.8	18.2	25.1	12.9
32.0	17.0	12.4	8.8
30.4	16.3	35.1	17.4
24.0	10.5	31.5	14.9
39.5	23.1	21.1	10.5
24.2	12.4	27.6	16.1
52.5	24.9		

(a) Estimate the correlation between Y and X.

(b) Test the hypothesis that $\rho = 0$, using $\alpha = 0.05$.

(c) Fit a simple linear regression model and test for significance of regression using $\alpha = 0.05$. What conclusions can you draw? How is the test for significance of regression related to the test on ρ in part (b)?

(d) Test the hypothesis H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ and draw conclusions. Use $\alpha = 0.05$.

(e) Analyze the residuals and comment on model adequacy.

11-62. A random sample of $n = 25$ observations was made on the time to failure of an electronic component and the temperature in the application environment in which the component was used.

(a) Given that $r = 0.83$, test the hypothesis that $\rho = 0$, using $\alpha = 0.05$. What is the P-value for this test?

(b) Find a 95% confidence interval on ρ.

(c) Test the hypothesis H_0: $\rho = 0.8$ versus H_1: $\rho \neq 0.8$, using $\alpha = 0.05$. Find the P-value for this test.

Supplemental Exercises

11-63. Show that, for the simple linear regression model, the following statements are true:

(a) $\displaystyle\sum_{i=1}^{n} (y_i - \hat{y}_i) = 0$ (b) $\displaystyle\sum_{i=1}^{n} (y_i - \hat{y}_i)x_i = 0$

(c) $\displaystyle\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i = \bar{y}$

11-64. An article in the *IEEE Transactions on Instrumentation and Measurement* ("Direct, Fast, and Accurate Measurement of V_T and K of MOS Transistor Using V_T-Sift Circuit," Vol. 40, 1991, pp. 951–955) described the use of a simple linear regression model to express drain current y (in milliamperes) as a function of ground-to-source voltage x (in volts). The data are as follows:

y	x	y	x
0.734	1.1	1.50	1.6
0.886	1.2	1.66	1.7
1.04	1.3	1.81	1.8
1.19	1.4	1.97	1.9
1.35	1.5	2.12	2.0

(a) Draw a scatter diagram of these data. Does a straight-line relationship seem plausible?

(b) Fit a simple linear regression model to these data.

(c) Test for significance of regression using $\alpha = 0.05$. What is the P-value for this test?

(d) Find a 95% confidence interval estimate on the slope.

(e) Test the hypothesis H_0: $\beta_0 = 0$ versus H_1: $\beta_0 \neq 0$ using $\alpha = 0.05$. What conclusions can you draw?

11-65. The strength of paper used in the manufacture of cardboard boxes (y) is related to the percentage of hardwood concentration in the original pulp (x). Under controlled conditions, a pilot plant manufactures 16 samples, each from a different batch of pulp, and measures the tensile strength. The data are shown in the table that follows:

(a) Fit a simple linear regression model to the data.

(b) Test for significance of regression using $\alpha = 0.05$.

(c) Construct a 90% confidence interval on the slope β_1.

(d) Construct a 90% confidence interval on the intercept β_0.

(e) Construct a 95% confidence interval on the mean strength at $x = 2.5$.

(f) Analyze the residuals and comment on model adequacy.

y	101.4	117.4	117.1	106.2
x	1.0	1.5	1.5	1.5
y	131.9	146.9	146.8	133.9
x	2.0	2.0	2.2	2.4
y	111.0	123.0	125.1	145.2
x	2.5	2.5	2.8	2.8
y	134.3	144.5	143.7	146.9
x	3.0	3.0	3.2	3.3

11-66. The vapor pressure of water at various temperatures follows:

Observation Number, i	Temperature (K)	Vapor pressure (mm Hg)
1	273	4.6
2	283	9.2
3	293	17.5
4	303	31.8
5	313	55.3
6	323	92.5
7	333	149.4
8	343	233.7
9	353	355.1
10	363	525.8
11	373	760.0

(a) Draw a scatter diagram of these data. What type of relationship seems appropriate in relating y to x?
(b) Fit a simple linear regression model to these data.
(c) Test for significance of regression using $\alpha = 0.05$. What conclusions can you draw?
(d) Plot the residuals from the simple linear regression model versus $\hat{y}_i$. What do you conclude about model adequacy?
(e) The Clausius-Clapeyron equation states that $\ln(P_v) \propto -\frac{1}{T}$, where P_v is the vapor pressure of water. Repeat parts (a)–(d). using an appropriate transformation.

11-67. An electric utility is interested in developing a model relating peak hour demand (y in kilowatts) to total monthly energy usage during the month (x, in kilowatt hours). Data for 50 residential customers are shown in the following table.

Customer	x	y	Customer	x	y
1	679	0.79	26	1434	0.31
2	292	0.44	27	837	4.20
3	1012	0.56	28	1748	4.88
4	493	0.79	29	1381	3.48
5	582	2.70	30	1428	7.58
6	1156	3.64	31	1255	2.63
7	997	4.73	32	1777	4.99
8	2189	9.50	33	370	0.59
9	1097	5.34	34	2316	8.19
10	2078	6.85	35	1130	4.79
11	1818	5.84	36	463	0.51
12	1700	5.21	37	770	1.74
13	747	3.25	38	724	4.10
14	2030	4.43	39	808	3.94
15	1643	3.16	40	790	0.96
16	414	0.50	41	783	3.29
17	354	0.17	42	406	0.44
18	1276	1.88	43	1242	3.24
19	745	0.77	44	658	2.14
20	795	3.70	45	1746	5.71
21	540	0.56	46	895	4.12
22	874	1.56	47	1114	1.90
23	1543	5.28	48	413	0.51
24	1029	0.64	49	1787	8.33
25	710	4.00	50	3560	14.94

(a) Draw a scatter diagram of y versus x.
(b) Fit the simple linear regression model.
(c) Test for significance of regression using $\alpha = 0.05$.
(d) Plot the residuals versus $\hat{y}_i$ and comment on the underlying regression assumptions. Specifically, does it seem that the equality of variance assumption is satisfied?
(e) Find a simple linear regression model using $\sqrt{y}$ as the response. Does this transformation on y stabilize the inequality of variance problem noted in part (d) above?

11-68. Consider the following data. Suppose that the relationship between Y and x is hypothesized to be $Y = (\beta_0 + \beta_1 x + \epsilon)^{-1}$. Fit an appropriate model to the data. Does the assumed model form seem reasonable?

x	10	15	18	12
y	0.1	0.13	0.09	0.15
x	9	8	11	6
y	0.20	0.21	0.18	0.24

11-69. Consider the weight and blood pressure data in Exercise 11-56. Fit a no-intercept model to the data, and compare it to the model obtained in Exercise 11-56. Which model is superior?

11-70. The following data, adapted from Montgomery, Peck, and Vining (2001), present the number of certified mental defectives per 10,000 of estimated population in the United Kingdom (y) and the number of radio receiver licenses issued (x) by the BBC (in millions) for the years 1924 through 1937. Fit a regression model relating y and x. Comment on the model. Specifically, does the existence of a strong correlation imply a cause-and-effect relationship?

Year	y	x	Year	y	x
1924	8	1.350	1931	16	4.620
1925	8	1.960	1932	18	5.497
1926	9	2.270	1933	19	6.260
1927	10	2.483	1934	20	7.012
1928	11	2.730	1935	21	7.618
1929	11	3.091	1936	22	8.131
1930	12	3.674	1937	23	8.593

11-71. An article in *Air and Waste* ("Update on Ozone Trends in California's South Coast Air Basin," Vol. 43, 1993) studied the ozone levels on the South Coast air basin of California for the years 1976–1991. The author believes that the number of days that the ozone level exceeds 0.20 parts per million depends on the seasonal meteorological index (the seasonal average 850 millibar temperature). The data follow:

Year	Days	Index	Year	Days	Index
1976	91	16.7	1984	81	18.0
1977	105	17.1	1985	65	17.2
1978	106	18.2	1986	61	16.9
1979	108	18.1	1987	48	17.1
1980	88	17.2	1988	61	18.2
1981	91	18.2	1989	43	17.3
1982	58	16.0	1990	33	17.5
1983	82	17.2	1991	36	16.6

(a) Construct a scatter diagram of the data.

(b) Fit a simple linear regression model to the data. Test for significance of regression.

(c) Find a 95% CI on the slope β_1.

(d) Analyze the residuals and comment on model adequacy.

11-72. An article in the *Journal of Applied Polymer Science* (Vol. 56, pp. 471–476, 1995) studied the effect of the mole ratio of sebacic acid on the intrinsic viscosity of copolyesters. The data follow:

Mole ratio x	1.0	0.9	0.8	0.7	0.6	0.5	0.4	0.3
Viscosity y	0.45	0.20	0.34	0.58	0.70	0.57	0.55	0.44

(a) Construct a scatter diagram of the data.

(b) Fit a simple linear repression module.

(c) Test for significance of regression. Calculate R^2 for the model.

(d) Analyze the residuals and comment on model adequacy.

11-73. Suppose that we have n pairs of observations (x_i, y_i) such that the sample correlation coefficient r is unity (approximately). Now let $z_i = y_i^2$ and consider the sample correlation coefficient for the n-pairs of data (x_i, z_i). Will this sample correlation coefficient be approximately unity? Explain why or why not.

11-74. The grams of solids removed from a material (y) is thought to be related to the drying time. Ten observations obtained from an experimental study follow:

y	4.3	1.5	1.8	4.9	4.2	4.8	5.8	6.2	7.0	7.9
x	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	6.5	7.0

(a) Construct a scatter diagram for these data.

(b) Fit a simple linear regression model.

(c) Test for significance of regression.

(d) Based on these data, what is your estimate of the mean grams of solids removed at 4.25 hours? Find a 95% confidence interval on the mean.

(e) Analyze the residuals and comment on model adequacy.

11-75. Two different methods can be used for measuring the temperature of the solution in a Hall cell used in aluminum smelting, a thermocouples implanted in the cell and an indirect measurement produced from an IR device. The indirect method is preferable became the thermocouples are eventually destroyed by the solution. Consider the following 10 measurements:

Thermocouple	921	935	916	920	940
IR	918	934	924	921	945
Thermocouple	936	925	940	933	927
IR	930	919	943	932	935

(a) Construct a scatter diagram for these data, letting x = thermocouple measurement and y = IR measurement.

(b) Fit a simple linear regression model.

(c) Test for significance a regression and calculate R^2. What conclusions can you draw?

(d) Is there evidence to support a claim that both devices produce equivalent temperature measurements? Formulate and test an appropriate hypothesis to support this claim.

(e) Analyze the residuals and comment on model adequacy.

MIND-EXPANDING EXERCISES

11-76. Consider the simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$, with $E(\epsilon) = 0$, $V(\epsilon) = \sigma^2$, and the errors ϵ uncorrelated.

(a) Show that $\text{cov}(\hat{\beta}_0, \hat{\beta}_1) = -\bar{x}\sigma^2/S_{xx}$.

(b) Show that $\text{cov}(\overline{Y}, \hat{\beta}_1) = 0$.

11-77. Consider the simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$, with $E(\epsilon) = 0$, $V(\epsilon) = \sigma^2$, and the errors ϵ uncorrelated.

(a) Show that $E(\hat{\sigma}^2) = E(MS_E) = \sigma^2$.

(b) Show that $E(MS_R) = \sigma^2 + \beta_1^2 S_{xx}$.

11-78. Suppose that we have assumed the straight-line regression model

$$Y = \beta_0 + \beta_1 x_1 + \epsilon$$

but the response is affected by a second variable x_2 such that the true regression function is

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Is the estimator of the slope in the simple linear regression model unbiased?

11-79. Suppose that we are fitting a line and we wish to make the variance of the regression coefficient $\hat{\beta}_1$ as small as possible. Where should the observations x_i, $i = 1, 2, \ldots, n$, be taken so as to minimize $V(\hat{\beta}_1)$? Discuss the practical implications of this allocation of the x_i.

11-80. **Weighted Least Squares.** Suppose that we are fitting the line $Y = \beta_0 + \beta_1 x + \epsilon$, but the variance of Y depends on the level of x; that is,

$$V(Y_i \mid x_i) = \sigma_i^2 = \frac{\sigma^2}{w_i} \qquad i = 1, 2, \ldots, n$$

where the w_i are constants, often called *weights*. Show that for an objective function in whole each squared residual is multiplied by the reciprocal of the variance of the corresponding observation, the resulting **weighted** least squares normal equations are

$$\hat{\beta}_0 \sum_{i=1}^{n} w_i + \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i = \sum_{i=1}^{n} w_i y_i$$
$$\hat{\beta}_0 \sum_{i=1}^{n} w_i x_i + \hat{\beta}_1 \sum_{i=1}^{n} w_i x_i^2 = \sum_{i=1}^{n} w_i x_i y_i$$

Find the solution to these normal equations. The solutions are weighted least squares estimators of β_0 and β_1.

11-81. Consider a situation where both Y and X are random variables. Let s_x and s_y be the sample standard deviations of the observed x's and y's, respectively. Show that an alternative expression for the fitted simple linear regression model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ is

$$\hat{y} = \bar{y} + r\frac{s_y}{s_x}(x - \bar{x})$$

11-82. Suppose that we are interested in fitting a simple linear regression model $Y = \beta_0 + \beta_1 x + \epsilon$, where the intercept, β_0, is known.

(a) Find the least squares estimator of β_1.

(b) What is the variance of the estimator of the slope in part (a)?

(c) Find an expression for a $100(1 - \alpha)\%$ confidence interval for the slope β_1. Is this interval longer than the corresponding interval for the case where both the intercept and slope are unknown? Justify your answer.

IMPORTANT TERMS AND CONCEPTS

In the E-book, click on any term or concept below to go to that subject.

Analysis of variance test in regression

Confidence interval on mean response

Correlation coefficient

Empirical models

Confidence intervals on model parameters

Least squares estimation of regression model parameters

Model adequacy checking

Prediction interval on a future observation

Residual plots

Residuals

Scatter diagram

Significance of regression

Statistical tests on model parameters

Transformations

CD MATERIAL

Lack of fit test

Logistic regression

11-4 SOME COMMENTS ON USES OF REGRESSION (CD ONLY)

Historical Note

Sir Francis Galton first used the term **regression analysis** in a study of the heights of fathers (x) and sons (y). Galton fit a least squares line and used it to predict the son's height from the fathers height. He found that if a father's height was above average, the son's height would also be above average, but not by as much as the father's height was. A similar effect was observed for short heights. That is, the son's height "regressed" toward the average. Consequently, Galton referred to the least squares line as a **regression line.**

Abuses of Regression.

Regression is widely used and frequently misused; several common abuses of regression are briefly mentioned here. Care should be taken in selecting variables with which to construct regression equations and in determining the form of the model. It is possible to develop statistically significant relationships among variables that are completely unrelated in a **causal** sense. For example, we might attempt to relate the shear strength of spot welds with the number of empty parking spaces in the visitor parking lot. A straight line may even appear to provide a good fit to the data, but the relationship is an unreasonable one on which to rely. You can't increase the weld strength by blocking off parking spaces. A strong observed association between variables does not necessarily imply that a causal relationship exists between those variables. This type of effect is encountered fairly often in retrospective data analysis, and even in observational studies. **Designed experiments** are the only way to determine cause-and-effect relationships.

Regression relationships are valid only for values of the regressor variable within the range of the original data. The linear relationship that we have tentatively assumed may be valid over the original range of x, but it may be unlikely to remain so as we extrapolate—that is, if we use values of x beyond that range. In other words, as we move beyond the range of values of x for which data were collected, we become less certain about the validity of the assumed model. Regression models are not necessarily valid for extrapolation purposes.

Now this does not mean *don't ever extrapolate*. There are many problem situations in science and engineering where extrapolation of a regression model is the only way to even approach the problem. However, there is a strong warning to **be careful.** A modest extrapolation may be perfectly all right in many cases, but a large extrapolation will almost never produce acceptable results.

11-8.3 Lack-of-Fit Test (CD Only)

Regression models are often fit to data to provide an empirical model when the true relationship between the variables Y and x is unknown. Naturally, we would like to know whether the order of the model tentatively assumed is correct. This section describes a test for the validity of this assumption.

The danger of using a regression model that is a poor approximation of the true functional relationship is illustrated in Fig. S11-1. Obviously, a polynomial of degree two or greater in x should have been used in this situation.

We present a test for the "goodness of fit" of the regression model. Specifically, the hypotheses we wish to test are

H_0: The simple linear regression model is correct.

H_1: The simple linear regression model is not correct.

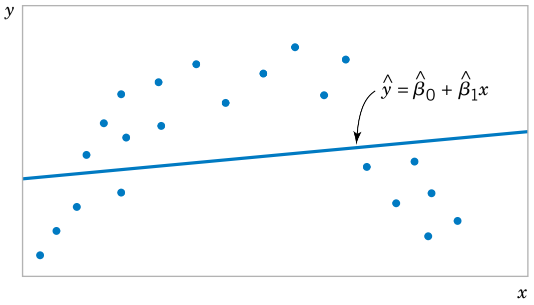

Figure S11-1 A regression model displaying lack of fit.

The test involves partitioning the error or residual sum of squares into the following components:

$$SS_E = SS_{PE} + SS_{LOF} \tag{S11-1}$$

where SS_{PE} is the sum of squares attributable to **pure error,** and SS_{LOF} is the sum of squares attributable to the **lack of fit** of the model. To compute SS_{PE}, we must have repeated observations on the response Y for at least one level of x. Suppose we have n total observations such that

$$\begin{array}{rl} y_{11}, y_{12}, \dots, y_{1n_1} & \text{repeated observations at } x_1 \\ y_{21}, y_{22}, \dots, y_{2n_2} & \text{repeated observations at } x_2 \\ \vdots & \\ y_{m1}, y_{m2}, \dots, y_{mn_m} & \text{repeated observations at } x_m \end{array}$$

Note that there are m distinct levels of x. The contribution to the pure-error sum of squares at x_1 (say) would be

$$\sum_{u=1}^{n_1} (y_{1u} - \bar{y}_1)^2 \tag{S11-2}$$

where $\bar{y}_1$ represents the average of all n_1 repeat observations on the response y at x_1. The total sum of squares for pure error would be obtained by summing Equation S11-2 over all levels of x as

$$SS_{PE} = \sum_{i=1}^{m} \sum_{u=1}^{n_i} (y_{iu} - \bar{y}_i)^2 \tag{S11-3}$$

There are $n_{pe} = \sum_{i=1}^{m} (n_i - 1) = n - m$ degrees of freedom associated with the pure-error sum of squares. The sum of squares for lack of fit is simply

$$SS_{LOF} = SS_E - SS_{PE} \tag{S11-4}$$

with $n - 2 - n_{pe} = m - 2$ degrees of freedom. The test statistic for lack of fit would then be

$$F_0 = \frac{SS_{LOF}/(m-2)}{SS_{PE}/(n-m)} = \frac{MS_{LOF}}{MS_{PE}} \tag{S11-5}$$

and we would reject the hypothesis that the model adequately fits the data if $f_0 > f_{\alpha, m-2, n-m}$.

This test procedure may be easily introduced into the analysis of variance conducted for the significance of regression. If the null hypothesis of model adequacy is rejected, the model must be abandoned and attempts must be made to find a more appropriate model. If H_0 is not rejected, there is no apparent reason to doubt the adequacy of the model, and MS_{PE} and MS_{LOF} are often combined to estimate σ^2.

EXAMPLE S11-1 Consider the data on two variables y and x shown below. Fit a simple linear regression model and test for lack of fit, using $\alpha = 0.05$.

x	y	x	y
1.0	2.3, 1.8	5.6	3.5, 2.8, 2.1
2.0	2.8	6.0	3.4, 3.2
3.3	1.8, 3.7	6.5	3.4
4.0	2.6, 2.6, 2.2	6.9	5.0
5.0	2.0		

The regression model is $\hat{y} = 1.697 + 0.259x$, and the regression sum of squares is $SS_R = 3.4930$. The pure-error sum of squares is computed as follows:

Level of x	$\sum_{u=1}^{n_i} (y_{iu} - \bar{y}_i)^2$	Degrees of Freedom
1.0	0.1250	1
3.3	1.8050	1
4.0	0.0166	2
5.6	0.9800	2
6.0	0.0200	1
Total	3.0366	7

The analysis of variance is summarized in Table S11-1. Since the lack-of-fit F-statistic is $f_0 = 1.42$, which has a P-value of $P = 0.3276$, we cannot reject the hypothesis that the tentative model adequately describes the data. We will pool lack-of-fit and pure-error mean squares to form the residual mean square that is the denominator mean square in the test for significance of regression. In addition, since the P-value for the statistic $f_0 = 6.66$ with 1 and 14 degrees of freedom associated with significance of regression is $P = 0.0216$, we conclude that $\beta_1 \neq 0$.

In fitting a regression model to experimental data, a good practice is to use the lowest degree model that adequately describes the data. The lack-of-fit test may be useful in this

Table S11-1 Analysis of Variance for Example S11-1

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	3.4930	1	3.4930	6.66	0.0218
Residual	7.3372	14	0.5241		
(Lack of fit)	4.3005	7	0.6144	1.42	0.3276
(Pure error)	3.0366	7	0.4338		
Total	10.8300	15			

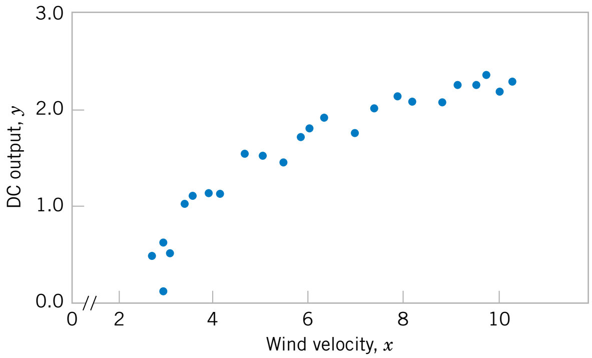

Figure S11-2 Plot of DC output y versus wind velocity x for the windmill data.

respect. However, it is always possible to fit a polynomial of degree $n - 1$ to n data points, and the experimenter should not consider using a model that is "saturated"—that is, that has very nearly as many independent variables as observations on y.

11-10 MORE ABOUT TRANSFORMATIONS (CD ONLY)

An Example

As noted earlier in Section 11-9, transformations can be very useful in many situations where the true relationship between the response Y and the regressor x is not well approximated by a straight line. The utility of a transformation is illustrated in the following example.

EXAMPLE S11-2

A research engineer is investigating the use of a windmill to generate electricity and has collected data on the DC output from this windmill and the corresponding wind velocity. The data are plotted in Figure S11-2 and listed in Table S11-2.

Inspection of the scatter diagram indicates that the relationship between DC output Y and wind velocity (x) may be nonlinear. However, we initially fit a straight-line model to the data. The regression model is

$$\hat{y} = 0.1309 + 0.2411x$$

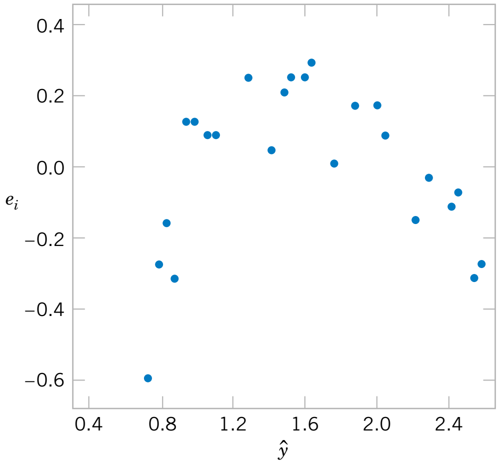

Figure S11-3 Plot of residuals e_i versus fitted values $\hat{y}_i$ for the windmill data.

The summary statistics for this model are $R^2 = 0.8745$, $MS_E = \hat{\sigma}^2 = 0.0557$ and $F_0 = 160.26$ (the P value is <0.0001).

A plot of the residuals versus $\hat{y}_i$ is shown in Figure S11-3. This residual plot indicates model inadequacy and implies that the linear relationship has not captured all of the information in the wind speed variable. Note that the curvature that was apparent in the scatter diagram of Figure S11-2 is greatly amplified in the residual plots. Clearly some other model form must be considered.

We might initially consider using a quadratic model such as

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$$

to account for the apparent curvature. However, the scatter diagram Figure S11-2 suggests that as wind speed increases, DC output approaches an upper limit of approximately 2.5. This is also consistent with the theory of windmill operation. Since the quadratic model will eventually bend downward as wind speed increases, it would not be appropriate for these data. A more reasonable model for the windmill data that incorporates an upper asymptote would be

$$y = \beta_0 + \beta_1 \left(\frac{1}{x}\right) + \epsilon$$

Table S11-2 Observed Values y_i and Regressor Variable x_i for Example S11-2

Observation Number, i	Wind Velocity (mph), x_i	DC Output, y_i
1	5.00	1.582
2	6.00	1.822
3	3.40	1.057
4	2.70	0.500
5	10.00	2.236
6	9.70	2.386
7	9.55	2.294
8	3.05	0.558
9	8.15	2.166
10	6.20	1.866
11	2.90	0.653
12	6.35	1.930
13	4.60	1.562
14	5.80	1.737
15	7.40	2.088
16	3.60	1.137
17	7.85	2.179
18	8.80	2.112
19	7.00	1.800
20	5.45	1.501
21	9.10	2.303
22	10.20	2.310
23	4.10	1.194
24	3.95	1.144
25	2.45	0.123

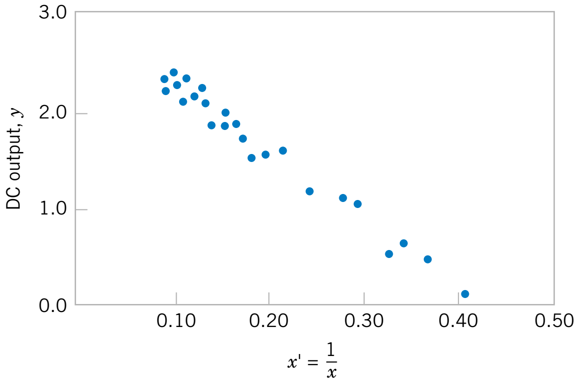

Figure S11-4 Plot of DC output versus $x' = 1/x$ for the windmill data.

Figure S11-4 is a scatter diagram with the transformed variable $x' = 1/x$. This plot appears linear, indicating that the reciprocal transformation is appropriate. The fitted regression model is

$$\hat{y} = 2.9789 - 6.9345x'$$

The summary statistics for this model are $R^2 = 0.9800$, $MS_E = \hat{\sigma}^2 = 0.0089$, and $F_0 = 1128.43$ (the P value is <0.0001).

A plot of the residuals from the transformed model versus $\hat{y}$ is shown in Figure S11-5. This plot does not reveal any serious problem with inequality of variance. The normal probability plot, shown in Figure S11-6, gives a mild indication that the errors come from a distribution with heavier tails than the normal (notice the slight upward and downward curve at the extremes). This normal probability plot has the z-score value plotted on the horizontal axis. Since there is no strong signal of model inadequacy, we conclude that the transformed model is satisfactory.

Logistic Regression

Linear regression often works very well when the response variable is **quantitative.** We now consider the situation where the response variable takes on only two possible values, 0 and 1. These could be arbitrary assignments resulting from observing a **qualitative** response. For

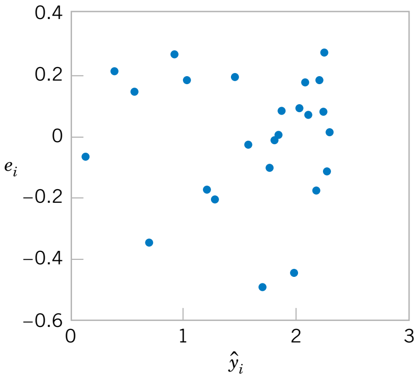

Figure S11-5 Plot of residuals versus fitted values $\hat{y}_i$ for the transformed model for the windmill data.

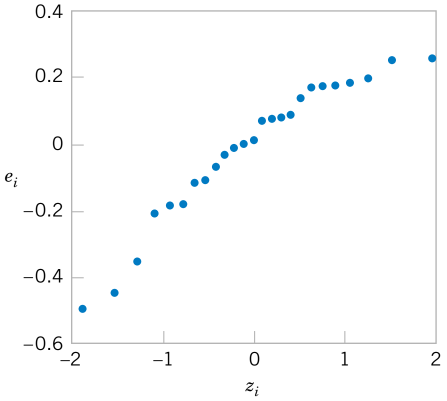

Figure S11-6 Normal probability plot of the residuals for the transformed model for the windmill data.

example, the response could be the outcome of a functional electrical test on a semiconductor device for which the results are either a "success," which means the device works properly, or a "failure," which could be due to a short, an open, or some other functional problem.

Suppose that the model has the form

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{S11-6}$$

and the response variable Y_i takes on the values either 0 or 1. We will assume that the response variable Y_i is a **Bernoulli random variable** with probability distribution as follows:

Y_i	Probability
1	$P(y_i = 1) = \pi_i$
0	$P(y_i = 0) = 1 - \pi_i$

Now since $E(\epsilon_i) = 0$, the expected value of the response variable is

$$\begin{aligned} E(Y_i) &= 1(\pi_i) + 0(1 - \pi_i) \\ &= \pi_i \end{aligned}$$

This implies that

$$E(Y_i) = \beta_0 + \beta_1 x_i = \pi_i$$

This means that the expected response given by the response function $E(Y_i) = \beta_0 + \beta_1 x_i$ is just the probability that the response variable takes on the value 1.

There are some substantive problems with the regression model in Equation S11-6. First, note that if the response is binary, the error terms ϵ_i can only take on two values, namely,

$$\begin{aligned} \epsilon_i &= 1 - (\beta_0 + \beta_1 x_i) \qquad &&\text{when } Y_i = 1 \\ \epsilon_i &= -(\beta_0 + \beta_1 x_i) \qquad &&\text{when } Y_i = 0 \end{aligned}$$

Consequently, the errors in this model cannot possibly be normal. Second, the error variance is not constant, since

$$\begin{aligned} \sigma^2_{y_i} &= E\{Y_i - E(Y_i)\}^2 \\ &= (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i) \\ &= \pi_i (1 - \pi_i) \end{aligned}$$

Notice that this last expression is just

$$\sigma^2_{y_i} = E(Y_i)[1 - E(Y_i)]$$

since $E(Y_i) = \beta_0 + \beta_1 x_i = \pi_i$. This indicates that the variance of the observations (which is the same as the variance of the errors because $\epsilon_i = Y_i - \pi_i$, and π_i is a constant) is a function of the mean. Finally, there is a constraint on the response function, because

$$0 \le E(Y_i) = \pi_i \le 1$$

This restriction can cause serious problems with the choice of a **linear response function,** as we have initially assumed in Equation S11-6. It would be possible to fit a model to the data for which the predicted values of the response lie outside the 0, 1 interval.

Generally, when the response variable is binary, there is considerable empirical evidence indicating that the shape of the response function should be nonlinear. A monotonically increasing (or decreasing) *S*-shaped (or reverse *S*-shaped) function, such as shown in Figure S11-7, is usually employed. This function is called the **logit response function,** and has the form

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{S11-7}$$

or equivalently,

$$E(Y) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x)]} \tag{S11-8}$$

In **logistic regression** we assume that $E(Y)$ is related to x by the logit function. It is easy to show that

$$\frac{E(Y)}{1 - E(Y)} = \exp^{\beta_0 + \beta_1 x} \tag{S11-9}$$

The quantity exp($\beta_0 + \beta_1 x$) on the right-hand side of Equation S11-9 is called the **odds ratio.** It has a straightforward interpretation: If the odds ratio is 2 for a particular value of x, it means that a success is twice as likely as a failure at that value of the regressor x. Notice that the natural logarithm of the odds ratio is a linear function of the regressor variable. Therefore the slope β_1 is the change in the log odds that results from a one-unit increase in x. This means that the odds ratio changes by e^{β_1} when x increases by one unit.

The parameters in this logistic regression model are usually estimated by the method of maximum likelihood. For details of the procedure, see Montgomery, Peck, and Vining (2001). Minitab will fit logistic regression models and provide useful information on the quality of the fit.

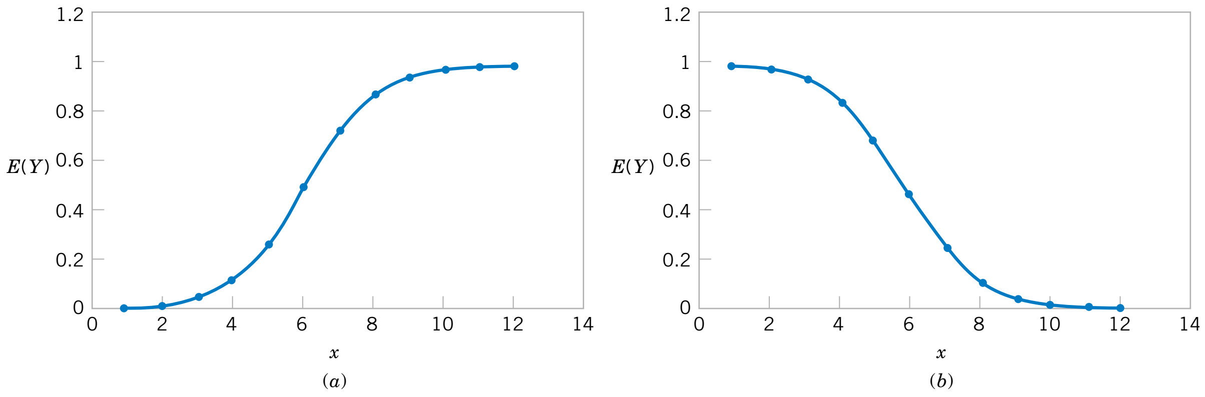

Figure S11-7 Examples of the logistic response function. (a) $E(Y) = 1/(1 + e^{-6.0-1.0x})$, (b) $E(Y) = 1/(1 + e^{-6.0+1.0x})$,

We will illustrate logistic regression using the data on launch temperature and O-ring failure for the 24 space shuttle launches prior to the *Challenger* disaster of January 1986. There are six O-rings used on the rocket motor assembly to seal field joints. The table below presents the launch temperatures. A 1 in the "O-Ring Failure" column indicates that at least one O-ring failure had occurred on that launch.

Temperature	O-Ring Failure	Temperature	O-Ring Failure	Temperature	O-Ring Failure
53	1	68	0	75	0
56	1	69	0	75	1
57	1	70	0	76	0
63	0	70	1	76	0
66	0	70	1	78	0
67	0	70	1	79	0
67	0	72	0	80	0
67	0	73	0	81	0

Figure S11-8 is a scatter plot of the data. Note that failures tend to occur at lower temperatures. The logistic regression model fit to this data from Minitab is shown in the following boxed display.

```
Binary Logistic Regression: O-Ring Failure versus Temperature

Link Function:        Logit
Response Information

Variable       Value       Count
O-Ring F       1               7       (Event)
               0              17
               Total          24

Logistic Regression Table

                                                          Odds     95%       CI
Predictor          Coef     SE Coef        Z          P   Ratio    Lower     Upper
Constant         10.875       5.703     1.91      0.057
Temperat       -0.17132     0.08344    -2.05      0.040    0.84     0.72      0.99

Log-Likelihood = -11.515
Test that all slopes are zero: G = 5.944, DF = 1, P-Value = 0.015
```

The fitted logistic regression model is

$$\hat{y} = \frac{1}{1 + \exp[-(10.875 - 0.17132x)]}$$

The standard error of the slope $\hat{\beta}_1$ is $se(\hat{\beta}_1) = 0.08344$. For large samples, $\hat{\beta}_1$ has an approximate normal distribution, and so $\hat{\beta}_1 / se(\hat{\beta}_1)$ can be compared to the standard normal distribution to

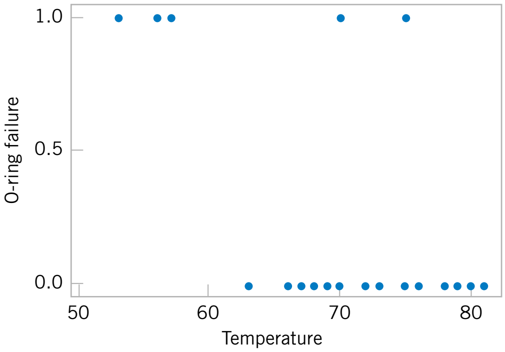

Figure S11-8 Scatter plot of O-ring failures versus launch temperature for 24 space shuttle flights.

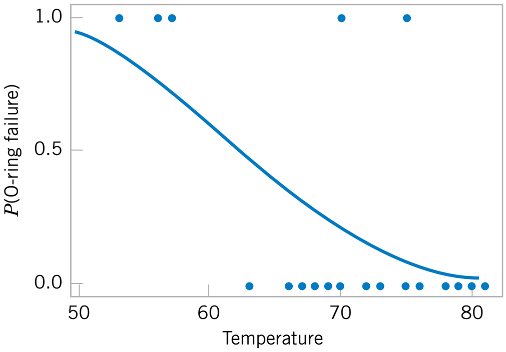

Figure S11-9 Probability of O-ring failure versus launch temperature (based on a logistic regression model).

test H_0: $\beta_1 = 0$. Minitab performs this test. The *P*-value is 0.04, indicating that temperature has a significant effect on the probability of O-ring failure. The odds ratio is 0.84, so every one degree increase in temperature reduces the odds of failure by 0.84. Figure S11-9 shows the fitted logistic regression model. The sharp increase in the probability of O-ring failure is very evident in this graph. The actual temperature at the *Challenger* launch was 31°F. This is well outside the range of other launch temperatures, so our logistic regression model is not likely to provide highly accurate predictions at that temperature, but it is clear that a launch at 31°F is almost certainly going to result in O-ring failure.

It is interesting to note that all of these data were available **prior** to launch. However, engineers were unable to effectively analyze the data and use them to provide a convincing argument against launching *Challenger* to NASA managers. Yet a simple regression analysis of the data would have provided a strong quantitative basis for this argument. This is one of the more dramatic instances that points out **why engineers and scientists need a strong background in basic statistical techniques.**