11

(15 July 2009)—Space Shuttle Endeavour and its seven-member STS-127 crew head toward Earth orbit and rendezvous with the International Space Station
Courtesy NASA

Simple Linear Regression and Correlation

The space shuttle *Challenger* accident in January 1986 was the result of the failure of O-rings used to seal field joints in the solid rocket motor due to the extremely low ambient temperatures at the time of launch. Prior to the launch there were data on the occurrence of O-ring failure and the corresponding temperature on 24 prior launches or static firings of the motor. In this chapter we will see how to build a statistical model relating the probability of O-ring failure to temperature. This model provides a measure of the risk associated with launching the shuttle at the low temperature occurring when *Challenger* was launched.

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. **Apply the correlation model**
7. **Use simple transformations to achieve a linear regression model**

11-1 EMPIRICAL MODELS

Many problems in engineering and the sciences involve a study or analysis of the relationship between two or more variables. For example, the pressure of a gas in a container is related to the temperature, the velocity of water in an open channel is related to the width of the channel, and the displacement of a particle at a certain time is related to its velocity. In this last example, if we let d_0 be the displacement of the particle from the origin at time $t = 0$ and v be the velocity, then the displacement at time t is $d_t = d_0 + vt$. This is an example of a **deterministic** linear relationship, because (apart from measurement errors) the model predicts displacement perfectly.

However, there are many situations where the relationship between variables is not deterministic. For example, the electrical energy consumption of a house (y) is related to the size of the house (x, in square feet), but it is unlikely to be a deterministic relationship. Similarly, the fuel usage of an automobile (y) is related to the vehicle weight x, but the relationship is not a deterministic one. In both of these examples the value of the response of interest y (energy consumption, fuel usage) cannot be predicted perfectly from knowledge of the corresponding x. It is possible for different automobiles to have different fuel usage even if they weigh the same, and it is possible for different houses to use different amounts of electricity even if they are the same size.

The collection of statistical tools that are used to model and explore relationships between variables that are related in a nondeterministic manner is called **regression analysis**. Because problems of this type occur so frequently in many branches of engineering and science, regression analysis is one of the most widely used statistical tools. In this chapter we present the situation where there is only one independent or predictor variable x and the relationship with the response y is assumed to be linear. While this seems to be a simple scenario, there are many practical problems that fall into this framework.

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** 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,

$$Y = \beta_0 + \beta_1 x + \epsilon \qquad (11\text{-}1)$$

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

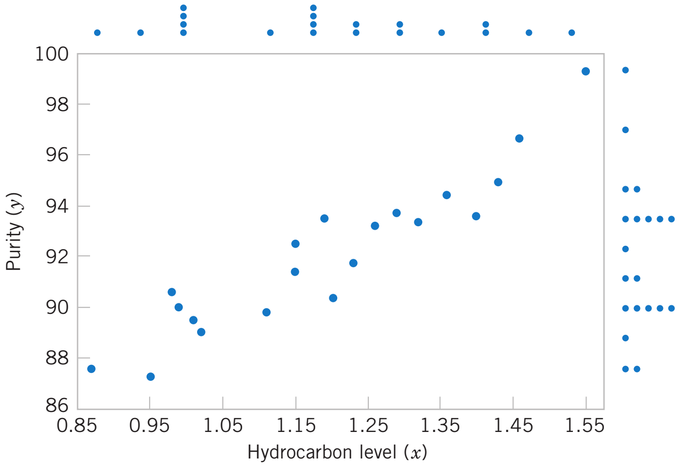


Figure 11-1 Scatter diagram of oxygen purity versus hydrocarbon level from Table 11-1.

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 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

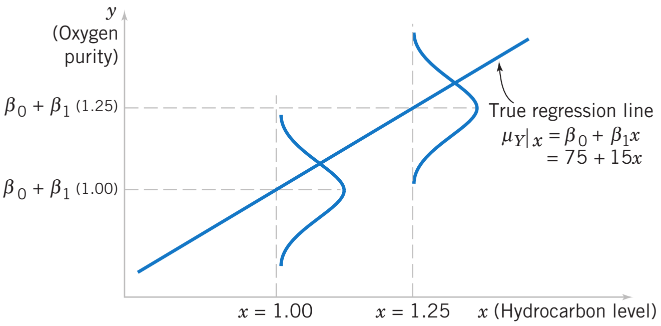

y
(Oxygen purity)
$\beta_0 + \beta_1$ (1.25)
$\beta_0 + \beta_1$ (1.00)
True regression line
$\mu_{Y|x} = \beta_0 + \beta_1 x$
$= 75 + 15x$
x = 1.00
x = 1.25
x (Hydrocarbon level)

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

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.

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 father's 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 below average 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.

11-2 SIMPLE LINEAR REGRESSION

The case of **simple linear regression** considers a single **regressor variable** or **predictor variable** 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

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{11-2}$$

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.

Suppose that we have n pairs of observations $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. 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

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

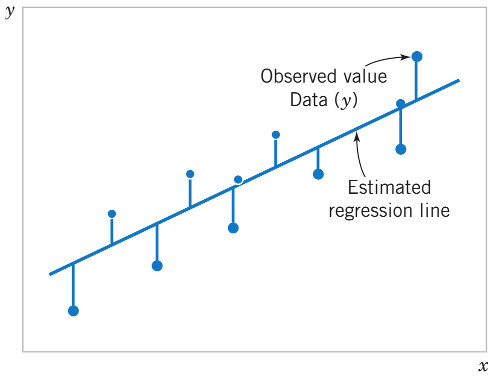


Figure 11-3 Deviations of the data from the estimated regression model.

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 $\hat{\beta}_1$.

Least Squares Estimates

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

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

EXAMPLE 11-1 Oxygen Purity

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

$$n = 20 \quad \sum_{i=1}^{20} x_i = 23.92 \quad \sum_{i=1}^{20} y_i = 1{,}843.21$$

$$\bar{x} = 1.1960 \quad \bar{y} = 92.1605$$

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

$$\sum_{i=1}^{20} x_i y_i = 2{,}214.6566$$

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

$$= 0.68088$$

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.

Practical Interpretation: Using the regression model, 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 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.

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.

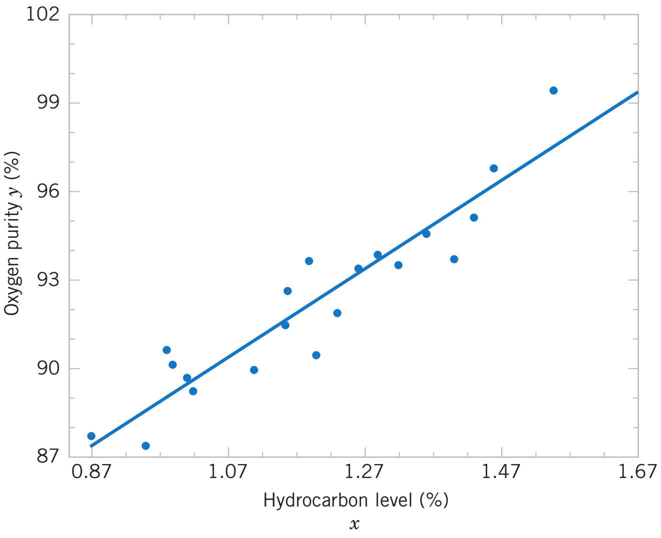


Figure 11-4 Scatter plot of oxygen purity y versus hydrocarbon level x and regression model $\hat{y} = 74.283 + 14.947x$.

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 ← $\hat{\beta}_0$     1.593      46.62      0.000
HC Level         14.947 ← $\hat{\beta}_1$     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 ← $SS_E$       1.18 ← $\hat{\sigma}^2$
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
```

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

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

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

Estimator of Variance

$$\hat{\sigma}^2 = \frac{SS_E}{n - 2} \tag{11-13}$$

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 $SS_T = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2$ is the total sum of squares of the response variable y. Formulas such as this are presented in Section 11-4. 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. Estimate σ^2. Graph the regression line.

(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. Estimate σ^2.

(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. The following table presents data on the ratings of quarterbacks for the 2008 National Football League season (source: *The Sports Network*). It is suspected that the rating (y) is related to the average number of yards gained per pass attempt (x).

(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 rating if a quarterback averages 7.5 yards per attempt.

(c) What change in the mean rating is associated with a decrease of one yard per attempt?

(d) To increase the mean rating by 10 points, how much increase in the average yards per attempt must be generated?

(e) Given that $x = 7.21$ yards, find the fitted value of y and the corresponding residual.

Player		Team	Yards per Attempt	Rating Points
Philip	Rivers	SD	8.39	105.5
Chad	Pennington	MIA	7.67	97.4
Kurt	Warner	ARI	7.66	96.9
Drew	Brees	NO	7.98	96.2
Peyton	Manning	IND	7.21	95
Aaron	Rodgers	GB	7.53	93.8
Matt	Schaub	HOU	8.01	92.7
Tony	Romo	DAL	7.66	91.4
Jeff	Garcia	TB	7.21	90.2
Matt	Cassel	NE	7.16	89.4
Matt	Ryan	ATL	7.93	87.7
Shaun	Hill	SF	7.10	87.5
Seneca	Wallace	SEA	6.33	87
Eli	Manning	NYG	6.76	86.4
Donovan	McNabb	PHI	6.86	86.4
Jay	Cutler	DEN	7.35	86
Trent	Edwards	BUF	7.22	85.4
Jake	Delhomme	CAR	7.94	84.7
Jason	Campbell	WAS	6.41	84.3
David	Garrard	JAC	6.77	81.7
Brett	Favre	NYJ	6.65	81
Joe	Flacco	BAL	6.94	80.3
Kerry	Collins	TEN	6.45	80.2
Ben	Roethlisberger	PIT	7.04	80.1
Kyle	Orton	CHI	6.39	79.6
JaMarcus	Russell	OAK	6.58	77.1
Tyler	Thigpen	KC	6.21	76
Gus	Freotte	MIN	7.17	73.7
Dan	Orlovsky	DET	6.34	72.6
Marc	Bulger	STL	6.18	71.4
Ryan	Fitzpatrick	CIN	5.12	70
Derek	Anderson	CLE	5.71	66.5

11-4. 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.

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

(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?

(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?

11-5. 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? Graph the regression line.

(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.

11-6. The following table presents the highway gasoline mileage performance and engine displacement for DaimlerChrysler vehicles for model year 2005 (source: U.S. Environmental Protection Agency).

(a) Fit a simple linear model relating highway miles per gallon (y) to engine displacement (x) in cubic inches 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 Neon, with an engine displacement of 122 cubic inches.

Carline	Engine Displacement (in^3)	MPG (highway)
300C/SRT-8	215	30.8
CARAVAN 2WD	201	32.5
CROSSFIRE ROADSTER	196	35.4
DAKOTA PICKUP 2WD	226	28.1
DAKOTA PICKUP 4WD	226	24.4
DURANGO 2WD	348	24.1
GRAND CHEROKEE 2WD	226	28.5
GRAND CHEROKEE 4WD	348	24.2
LIBERTY/CHEROKEE 2WD	148	32.8
LIBERTY/CHEROKEE 4WD	226	28
NEON/SRT-4/SX 2.0	122	41.3
PACIFICA 2WD	215	30.0
PACIFICA AWD	215	28.2
PT CRUISER	148	34.1
RAM 1500 PICKUP 2WD	500	18.7
RAM 1500 PICKUP 4WD	348	20.3
SEBRING 4-DR	165	35.1
STRATUS 4-DR	148	37.9
TOWN & COUNTRY 2WD	148	33.8
VIPER CONVERTIBLE	500	25.9
WRANGLER/TJ 4WD	148	26.4

11-7. 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
x	825	830	890	895	890	910

y	46	43	53	52	54	57	58
x	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.

11-8. 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
x	60	63	65	70	70	70	80	90	80	80

y	5	4	6	8	4	5	7	9	7	6
x	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 decibels.

11-9. 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
x	1.6	9.4	15.5	20.0	22.0

y	110	113	75	94
x	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.

11-10. 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
x	0.19	0.15	0.57	0.70	0.67	0.63

y	11.8	12.1	14.3	14.7	15.0	17.3
x	0.47	0.70	0.60	0.78	0.81	0.78

y	19.2	23.1	27.4	27.7	31.8	39.5
x	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.

11-11. 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 following table.
(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 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?

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

11-12. An article in the *Journal of the American Ceramic Society* ["Rapid Hot-Pressing of Ultrafine PSZ Powders" (1991, Vol. 74, pp. 1547–1553)] considered the microstructure of the ultrafine powder of partially stabilized zirconia as a function of temperature. The data are shown below:

x = Temperature (°C): 1100 1200 1300 1100 1500
1200 1300

y = Porosity (%): 30.8 19.2 6.0 13.5 11.4
7.7 3.6

(a) Fit the simple linear regression model using the method of least squares. Find an estimate of σ^2.
(b) Estimate the mean porosity for a temperature of 1400°C.
(c) Find the fitted value corresponding to $y = 11.4$ and the associated residual.
(d) Draw a scatter diagram of the data. Does a simple linear regression model seem appropriate here? Explain.

11-13. An article in the *Journal of the Environmental Engineering Division* ["Least Squares Estimates of BOD Parameters" (1980, Vol. 106, pp. 1197–1202)] took a sample from the Holston River below Kingport, Tennessee, during August 1977. The biochemical oxygen demand (BOD) test is conducted over a period of time in days. The resulting data are shown below:

Time (days): 1 2 4 6 8 10 12 14 16
18 20

BOD (mg/liter): 0.6 0.7 1.5 1.9 2.1 2.6 2.9 3.7 3.5
3.7 3.8

(a) Assuming that a simple linear regression model is appropriate, fit the regression model relating BOD (y) to the time (x). What is the estimate of σ^2?
(b) What is the estimate of expected BOD level when the time is 15 days?
(c) What change in mean BOD is expected when the time changes by three days?
(d) Suppose the time used is six days. Calculate the fitted value of y and the corresponding residual.
(e) 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 values 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 time is an effective regressor variable in predicting BOD?

11-14. An article in *Wood Science and Technology* ["Creep in Chipboard, Part 3: Initial Assessment of the Influence of Moisture Content and Level of Stressing on Rate of Creep and Time to Failure" (1981, Vol. 15, pp. 125–144)] studied the deflection (mm) of particleboard from stress levels of relative humidity. Assume that the two variables are related according to the simple linear regression model. The data are shown below:

x = Stress level (%):	54	54	61	61	68
y = Deflection (mm):	16.473	18.693	14.305	15.121	13.505

x = Stress level (%):	68	75	75	75
y = Deflection (mm):	11.640	11.168	12.534	11.224

(a) Calculate the least square estimates of the slope and intercept. What is the estimate of σ^2? Graph the regression model and the data.
(b) Find the estimate of the mean deflection if the stress level can be limited to 65%.
(c) Estimate the change in the mean deflection associated with a 5% increment in stress level.
(d) To decrease the mean deflection by one millimeter, how much increase in stress level must be generated?
(e) Given that the stress level is 68%, find the fitted value of deflection and the corresponding residual.

11-15. In an article in *Statistics and Computing* ["An Iterative Monte Carlo Method for Nonconjugate Bayesian Analysis" (1991, pp. 119–128)] Carlin and Gelfand investigated the age (x) and length (y) of 27 captured dugongs (sea cows).

x = 1.0, 1.5, 1.5, 1.5, 2.5, 4.0, 5.0, 5.0, 7.0, 8.0, 8.5, 9.0, 9.5, 9.5, 10.0, 12.0, 12.0, 13.0, 13.0, 14.5, 15.5, 15.5, 16.5, 17.0, 22.5, 29.0, 31.5

y = 1.80, 1.85, 1.87, 1.77, 2.02, 2.27, 2.15, 2.26, 2.47, 2.19, 2.26, 2.40, 2.39, 2.41, 2.50, 2.32, 2.32, 2.43, 2.47, 2.56, 2.65, 2.47, 2.64, 2.56, 2.70, 2.72, 2.57

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

(b) Estimate the mean length of dugongs at age 11.

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

11-16. 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.

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

11-17. Consider the regression model developed in Exercise 11-6. Suppose that engine displacement is measured in cubic centimeters instead of cubic inches.

(a) Write the new regression model.

(b) What change in gasoline mileage is associated with a 1 cm^3 change is engine displacement?

11-18. Show that in a simple linear regression model the point $(\bar{x}, \bar{y})$ lies exactly on the least squares regression line.

11-19. 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 11-11, 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 $\hat{\beta}_1$?

11-20. 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 $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

(a) Find the least squares estimate of β.

(b) Fit the model $Y = \beta x + \epsilon$ to the chloride concentration-roadway area data in Exercise 11-10. 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 the 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 $V(Y_i) = \sigma^2$. Because $\hat{\beta}_1$ is a linear combination of the observations Y_i, the results in Section 5-5 can be applied to show that

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

For the intercept, we can show in a similar manner 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 11-98) 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.

Estimated Standard Errors

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 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 11-5 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$(0, \sigma^2)$.

11-4.1 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

$$H_0\colon \beta_1 = \beta_{1,0}$$
$$H_1\colon \beta_1 \neq \beta_{1,0} \tag{11-18}$$

where we have assumed a two-sided alternative. Since the errors ϵ_i are NID$(0, \sigma^2)$, it follows directly that the observations Y_i are NID$(\beta_0 + \beta_1 x_i, \sigma^2)$. 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

Test 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

Test 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, 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*)].

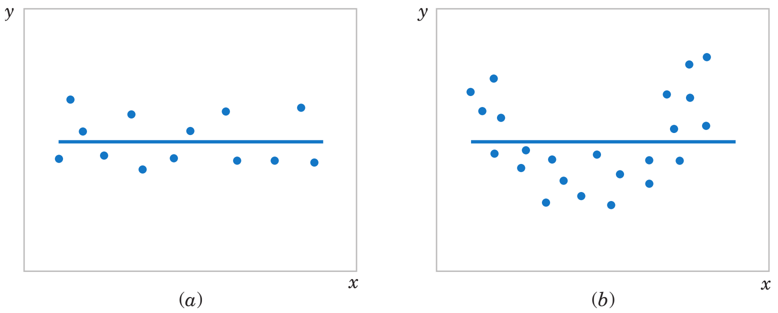


Figure 11-5 The hypothesis H_0: $\beta_1 = 0$ is not rejected.

EXAMPLE 11-2 Oxygen Purity Tests of Coefficients

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

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

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

$$\hat{\beta}_1 = 14.947 \quad n = 20, \quad S_{xx} = 0.68088, \quad \hat{\sigma}^2 = 1.18$$

so the t-statistic in Equation 10-20 becomes

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

Practical Interpretation: 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. There is strong evidence to support this claim. 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.

11-4.2 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:

Analysis of Variance Identity

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

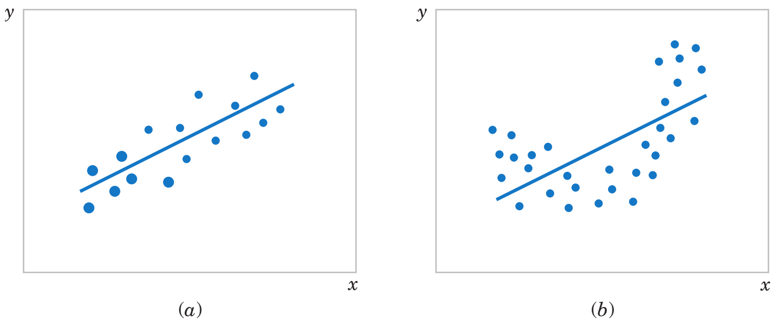


Figure 11-6 The hypothesis H_0: $\beta_1 = 0$ is rejected.

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

$$SS_T = SS_R + SS_E \qquad (11\text{-}25)$$

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 11-25 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

Test for Significance of Regression

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

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.

EXAMPLE 11-3 Oxygen Purity ANOVA

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: $\beta_1 = 0$ 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.

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

Note that the analysis of variance procedure for testing for significance of regression is equivalent to the *t*-test in Section 11-4.1. 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

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

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

$$T_0^2 = \frac{\hat{\beta}_1^2 S_{xx}}{MS_E} = \frac{\hat{\beta}_1 S_{xy}}{MS_E} = \frac{MS_R}{MS_E} \tag{11-28}$$

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 11-4

11-21. Consider the computer output below.

```
The regression equation is
Y = 12.9 + 2.34 x

Predictor        Coef      SE Coef       T        P
Constant       12.857        1.032       ?        ?
X              2.3445       0.1150       ?        ?

S = 1.48111     R−Sq = 98.1%     R−Sq(adj) = 97.9%

Analysis of Variance

Source            DF        SS         MS       F       P
Regression         1    912.43     912.43       ?       ?
Residual Error     8     17.55          ?
Total              9    929.98
```

(a) Fill in the missing information. You may use bounds for the *P*-values.
(b) Can you conclude that the model defines a useful linear relationship?
(c) What is your estimate of σ^2?

11-22. Consider the computer output below.

```
The regression equation is
Y = 26.8 + 1.48 x

Predictor        Coef      SE Coef       T        P
Constant       26.753        2.373       ?        ?
X              1.4756       0.1063       ?        ?

S = 2.70040     R−Sq = 93.7%     R-Sq (adj) = 93.2%

Analysis of Variance

Source            DF         SS       MS       F       P
Regression         1          ?        ?       ?       ?
Residual Error     ?       94.8      7.3
Total             15     1500.0
```

(a) Fill in the missing information. You may use bounds for the *P*-values.
(b) Can you conclude that the model defines a useful linear relationship?
(c) What is your estimate of σ^2?

11-23. 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?

11-24. 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.

11-25. Consider the National Football League data in Exercise 11-3.
(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 H_0: $\beta_1 = 10$ versus H_1: $\beta_1 \neq 10$ with $\alpha = 0.01$. Would you agree with the statement that this is a test of the hypothesis that a one-yard increase in the average yards per attempt results in a mean increase of 10 rating points?

11-26. Consider the data from Exercise 11-4 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 $\beta_0 = 0$.

11-27. Consider the data from Exercise 11-5 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.

11-28. Consider the data from Exercise 11-6 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?

11-29. Consider the data from Exercise 11-7 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?

11-30. Consider the data from Exercise 11-8 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.

11-31. Consider the data from Exercise 11-11, 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?

11-32. Consider the data from Exercise 11-10 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?

11-33. Consider the data in Exercise 11-13 on y = oxygen demand and x = time.
(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 the hypothesis that $\beta_0 = 0$.

11-34. Consider the data in Exercise 11-14 on y = deflection and x = stress level.
(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) Does this model appear to be adequate?
(c) Estimate the standard errors of the slope and intercept.

11-35. An article in *The Journal of Clinical Endocrinology and Metabolism* ["Simultaneous and Continuous 24-Hour Plasma and Cerebrospinal Fluid Leptin Measurements: Dissociation of Concentrations in Central and Peripheral Compartments" (2004, Vol. 89, pp. 258–265)] studied the demographics of simultaneous and continuous 24-hour plasma and cerebrospinal fluid leptin measurements. The data follow:

y = BMI (kg/m^2):	19.92	20.59	29.02	20.78	25.97	20.39	23.29	17.27	35.24
x = Age (yr):	45.5	34.6	40.6	32.9	28.2	30.1	52.1	33.3	47.0

(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?

11-36. 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.

11-37. 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 VII*e* through VII*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 11-3, using $\sigma = 5.5$ and $\beta_1' = 12.5$, where the hypotheses are H_0: $\beta_1 = 10$ versus H_1: $\beta_1 \neq 10$.

11-38. 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 11-20.

11-5 CONFIDENCE INTERVALS

11-5.1 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.

Confidence Intervals on Parameters

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 Oxygen Purity Confidence Interval on the Slope

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 11-29 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

$$12.181 \le \beta_1 \le 17.713$$

Practical Interpretation: This CI does not include zero, so there is strong evidence (at $\alpha = 0.05$) that the slope is not zero. The CI is reasonably narrow (±2.766) because the error variance is fairly small.

11-5.2 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 $\hat{\mu}_{Y|x_0} = \bar{y} + \hat{\beta}_1(x_0 - \bar{x})$ and cov $(\bar{Y}, \hat{\beta}_1) = 0$. The zero covariance result is left as a mind-expanding exercise. 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.

Confidence Interval on the Mean Response

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 CI 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 Oxygen Purity Confidence Interval on the Mean Response

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

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

or

$$89.23 \pm 0.75$$

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

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

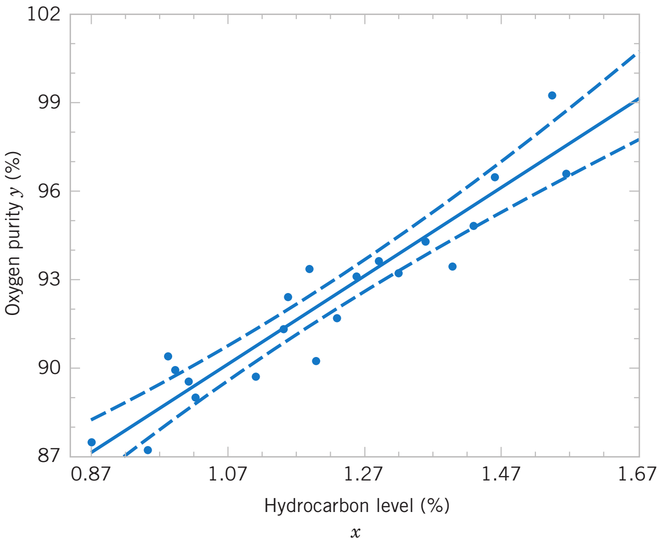


Figure 11-7 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line and 95 percent confidence limits on $\mu_{Y|x_0}$.

This is a reasonable narrow CI.

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.

11-6 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.

Prediction Interval

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 Oxygen Purity Prediction Interval

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

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

which simplifies to

$$86.83 \leq y_0 \leq 91.63$$

This is a reasonably narrow prediction interval.

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.

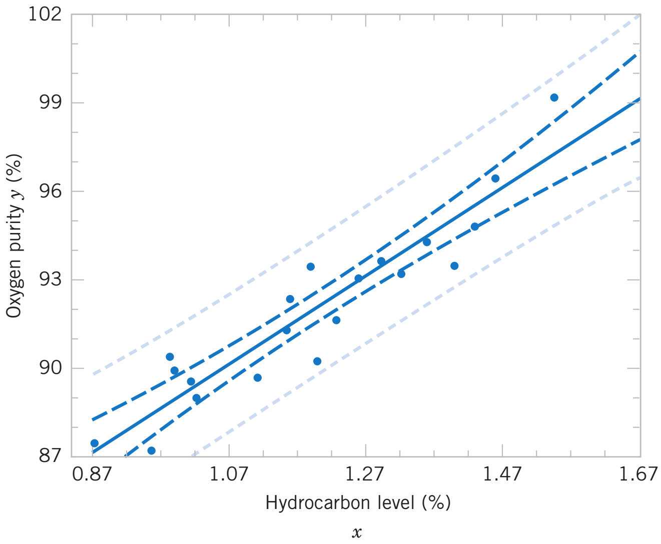


Figure 11-8 Scatter diagram of oxygen purity data from Example 11-1 with fitted regression line, 95% prediction limits (outer lines) and 95% confidence limits on $\mu_{Y|x_0}$.

EXERCISES FOR SECTIONS 11-5 AND 11-6

11-39. 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).

11-40. 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.

11-41. Refer to the NFL quarterback ratings data in Exercise 11-3. Find a 95% confidence interval on each of the following:

(a) Slope

(b) Intercept

(c) Mean rating when the average yards per attempt is 8.0

(d) Find a 95% prediction interval on the rating when the average yards per attempt is 8.0.

11-42. Refer to the data on y = house selling price and x = taxes paid in Exercise 11-4. 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 $x = 7.50$.

11-43. Exercise 11-5 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).

11-44. Exercise 11-6 presented gasoline mileage performance for 21 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.

11-45. Consider the data in Exercise 11-7 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.

11-46. Exercise 11-8 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 decibels

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

11-47. Refer to the data in Exercise 11-9 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 $x = 30$

11-48. Exercise 11-10 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 $x = 1.0\%$.

11-49. Refer to the data in Exercise 11-11 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.

11-50. Refer to the data in Exercise 11-12 on the microstructure of zirconia. Find a 95% confidence interval on each of the following:

(a) Slope (b) Intercept

(c) Mean length when $x = 1500$

(d) Find a 95% prediction interval on length when $x = 1500$. Explain why this interval is wider than the interval in part (c).

11-51. Refer to the data in Exercise 11-13 on oxygen demand. Find a 99% confidence interval on each of the following:

(a) β_1

(b) β_0

(c) Find a 95% confidence interval on mean BOD when the time is 8 days.

11-7 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.

11-7.1 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 is preferred. It requires judgment to assess the abnormality of such plots. (Refer to the discussion of the "fat pencil" method in Section 6-6).

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

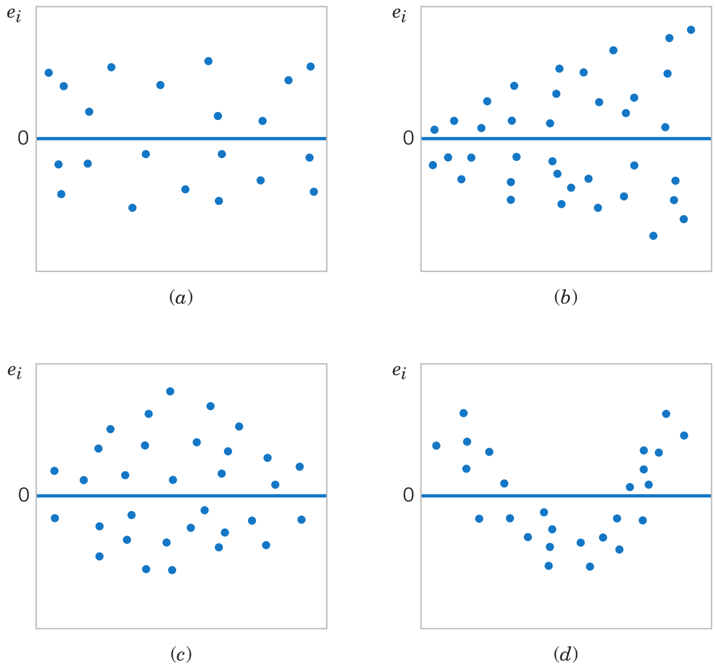


Figure 11-9 Patterns for residual plots. (a) Satisfactory, (b) Funnel, (c) Double bow, (d) Nonlinear. [Adapted from Montgomery, Peck, and Vining (2006).]

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 (2006).

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 (2006) for more details regarding methods for selecting an appropriate transformation. 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 Oxygen Purity Residuals

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 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.

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	89.081	0.929	11	1.19	93.54	92.071	1.469
2	1.02	89.05	89.530	−0.480	12	1.15	92.52	91.473	1.047
3	1.15	91.43	91.473	−0.043	13	0.98	90.56	88.932	1.628
4	1.29	93.74	93.566	0.174	14	1.01	89.54	89.380	0.160
5	1.46	96.73	96.107	0.623	15	1.11	89.85	90.875	−1.025
6	1.36	94.45	94.612	−0.162	16	1.20	90.39	92.220	−1.830
7	0.87	87.59	87.288	0.302	17	1.26	93.25	93.117	0.133
8	1.23	91.77	92.669	−0.899	18	1.32	93.41	94.014	−0.604
9	1.55	99.42	97.452	1.968	19	1.43	94.98	95.658	−0.678
10	1.40	93.65	95.210	−1.560	20	0.95	87.33	88.483	−1.153

11-7.2 Coefficient of Determination (R^2)

A widely used measure for a regression model is the following ratio of sum of squares.

R^2

The **coefficient of determination** is

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

The coefficient 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 the analysis of variance identity in Equations 11-24 and 11-25, $0 \leq R^2 \leq 1$. 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.

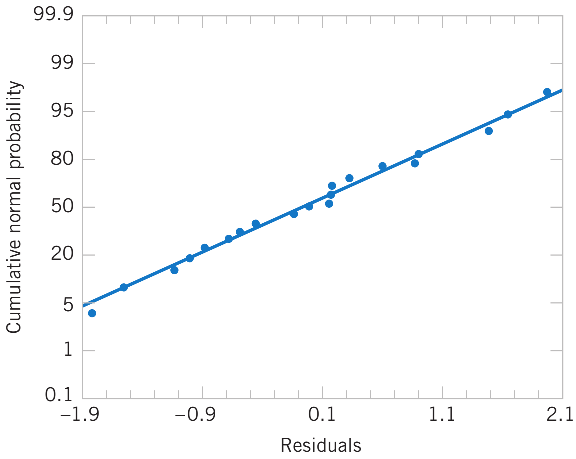


Figure 11-10 Normal probability plot of residuals, Example 11-7.

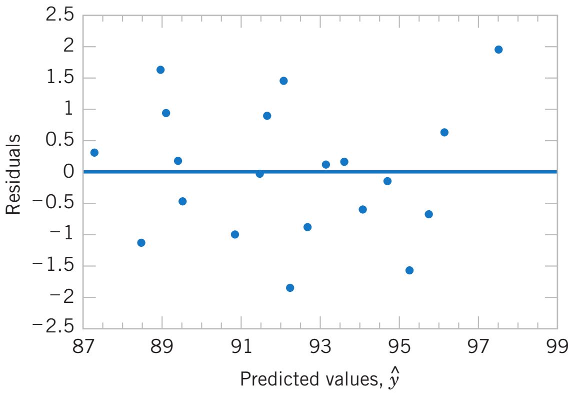


Figure 11-11 Plot of residuals versus predicted oxygen purity $\hat{y}$, Example 11-7.

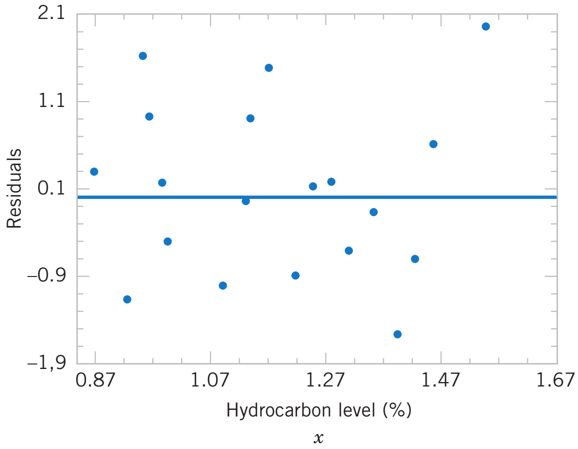


Figure 11-12 Plot of residuals versus hydrocarbon level x, Example 11-8.

The 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.

EXERCISES FOR SECTION 11-7

11-52. Refer to the compressive strength data in Exercise 11-1. Use the summary statistics provided to calculate R^2 and provide a practical interpretation of this quantity.

11-53. Refer to the NFL quarterback ratings data in Exercise 11-3.

(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 the fitted values and against x. Interpret these graphs.

11-54. Refer to the data in Exercise 11-4 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?

11-55. Refer to the data in Exercise 11-5 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?

11-56. Refer to the gasoline mileage data in Exercise 11-6.

(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?

11-57. Exercise 11-9 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?

11-58. Refer to Exercise 11-8, 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.

11-59. Refer to Exercise 11-10, 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?

11-60. An article in the *Journal of the American Statistical Association* ["Markov Chain Monte Carlo Methods for Computing Bayes Factors: A Comparative Review" (2001, Vol. 96, pp. 1122–1132)] analyzed the tabulated data on compressive strength parallel to the grain versus resin-adjusted density for specimens of radiata pine.

Compressive Strength	Density	Compressive Strength	Density
3040	29.2	3840	30.7
2470	24.7	3800	32.7
3610	32.3	4600	32.6
3480	31.3	1900	22.1
3810	31.5	2530	25.3
2330	24.5	2920	30.8
1800	19.9	4990	38.9
3110	27.3	1670	22.1
3160	27.1	3310	29.2
2310	24.0	3450	30.1
4360	33.8	3600	31.4
1880	21.5	2850	26.7
3670	32.2	1590	22.1
1740	22.5	3770	30.3
2250	27.5	3850	32.0
2650	25.6	2480	23.2
4970	34.5	3570	30.3
2620	26.2	2620	29.9
2900	26.7	1890	20.8
1670	21.1	3030	33.2
2540	24.1	3030	28.2

(a) Fit a regression model relating compressive strength to density.

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

(c) Estimate σ^2 for this model.

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

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

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

11-61. Consider the rocket propellant data in Exercise 11-11.

(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?

11-62. Consider the data in Exercise 11-7 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.

11-63. Consider the rocket propellant data in Exercise 11-11. Calculate the standardized residuals for these data. Does this provide any helpful information about the magnitude of the residuals?

11-64. **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.

11-65. 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 $f_0 > f_{\alpha,1,n-2}$. 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 $\alpha = 0.05$.
(b) What is the smallest value of R^2 that would lead to the conclusion of a significant regression if $\alpha = 0.05$?

11-8 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 $\sigma^2_{Y|x}$. 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,

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

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

$$H_0\colon \rho = 0$$
$$H_1\colon \rho \neq 0 \qquad (11\text{-}45)$$

The appropriate test statistic for these hypotheses is

Test Statistic for Zero Correlation

$$T_0 = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \qquad (11\text{-}46)$$

which has the t distribution with $n - 2$ degrees of freedom if $H_0\colon \rho = 0$ 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 $H_0\colon \beta_1 = 0$ given in Section 11-5.1. This equivalence follows directly from Equation 11-46.

The test procedure for the hypotheses

$$H_0\colon \rho = \rho_0$$
$$H_1\colon \rho \neq \rho_0 \qquad (11\text{-}47)$$

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

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

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 $H_0\colon \rho = \rho_0$, we may use the test statistic

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

and reject $H_0\colon \rho = \rho_0$ 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 11-48. The approximate $100(1-\alpha)\%$ confidence interval is

Confidence Interval for a Correlation Coefficient

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

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

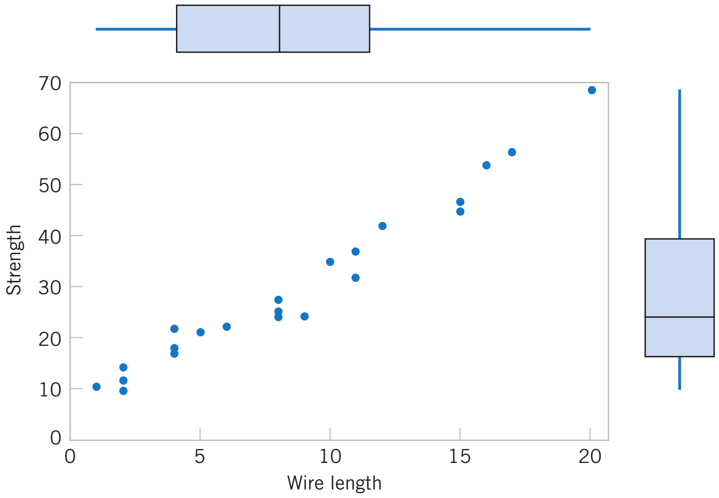


Figure 11-13 Scatter plot of wire bond strength versus wire length, Example 11-8.

EXAMPLE 11-8 Wire Bond Pull Strength

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 below.

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 hypotheses

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

```
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
```

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 11-50. Since arctanh $r =$ arctanh $0.9818 = 2.3452$, Equation 11-50 becomes

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

which reduces to

$$0.9585 \le \rho \le 0.9921$$

EXERCISES FOR SECTION 11–8

11-66. Suppose data is obtained from 20 pairs of (x, y) and the sample correlation coefficient is 0.8.

(a) Test the hypothesis that H_0: $\rho = 0$ against H_1: $\rho \neq 0$ with $\alpha = 0.05$. Calculate the P-value.

(b) Test the hypothesis that H_1: $\rho = 0.5$ against H_1: $\rho \neq 0.5$ with $\alpha = 0.05$. Calculate the P-value.

(c) Construct a 95% two-sided confidence interval for the correlation coefficient. Explain how the questions in parts (a) and (b) could be answered with a confidence interval.

11-67. Suppose data are obtained from 20 pairs of (x, y) and the sample correlation coefficient is 0.75.

(a) Test the hypothesis that H_0: $\rho = 0$ against H_1: $\rho > 0$ with $\alpha = 0.05$. Calculate the P-value.

(b) Test the hypothesis that H_1: $\rho = 0.5$ against H_1: $\rho > 0.5$ with $\alpha = 0.05$. Calculate the P-value.

(c) Construct a 95% one-sided confidence interval for the correlation coefficient. Explain how the questions in parts (a) and (b) could be answered with a confidence interval.

11-68. 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.

11-69. 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?

11-70. 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) Analyze the residuals and comment on model adequacy.

11-71. 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. Graph the data.

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

(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.

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

11-72. 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.

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

(a) Find a regression line relating systolic blood pressure to weight.
(b) Test for significance of regression using $\alpha = 0.05$.
(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.

11-73. In an article in *IEEE Transactions on Instrumentation and Measurement* (2001, Vol. 50, pp. 986–990), researchers studied the effects of reducing current draw in a magnetic core by electronic means. They measured the current in a magnetic winding with and without the electronics in a paired experiment. Data for the case without electronics are provided in the following table.

Supply Voltage	Current Without Electronics (mA)
0.66	7.32
1.32	12.22
1.98	16.34
2.64	23.66
3.3	28.06
3.96	33.39
4.62	34.12
3.28	39.21
5.94	44.21
6.6	47.48

(a) Graph the data and fit a regression line to predict current without electronics to supply voltage. Is there a significant regression at $\alpha = 0.05$? What is the P-value?
(b) Estimate the correlation coefficient.
(c) Test the hypothesis that $\rho = 0$ against the alternative $\rho \neq 0$ with $\alpha = 0.05$. What is the P-value?
(d) Compute a 95% confidence interval for the correlation coefficient.

11-74. The monthly absolute estimate of global (land and ocean combined) temperature indexes (degrees C) in 2000 and 2001 are (source: http://www.ncdc.noaa.gov/oa/climate/):

2000: 12.28, 12.63, 13.22, 14.21, 15.13, 15.82, 16.05, 16.02, 15.29, 14.29, 13.16, 12.47

2001: 12.44, 12.55, 13.35, 14.22, 15.28, 15.99, 16.23, 16.17, 15.44, 14.52, 13.52, 12.61

(a) Graph the data and fit a regression line to predict 2001 temperatures from those in 2000. Is there a significant regression at $\alpha = 0.05$? What is the P-value?
(b) Estimate the correlation coefficient.
(c) Test the hypothesis that $\rho = 0.9$ against the alternative $\rho \neq 0.9$ with $\alpha = 0.05$. What is the P-value?
(d) Compute a 95% confidence interval for the correlation coefficient.

11-75 Refer to the NFL quarterback ratings data in Exercise 11-3.
(a) Estimate the correlation coefficient between the ratings and the average yards per attempt.
(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.

11-76. Consider the following (x, y) data. Calculate the correlation coefficient. Graph the data and comment on the relationship between x and y. Explain why the correlation coefficient does not detect the relationship between x and y.

x	y	x	y
−4	0	0	−4
−3	−2.65	1	3.87
−3	2.65	1	−3.87
−2	−3.46	2	3.46
−2	3.46	2	−3.46
−1	−3.87	3	2.65
−1	3.87	3	−2.65
0	4	4	0

11-9 REGRESSION ON TRANSFORMED VARIABLES

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 (2006) or Myers (1990).

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 11-9 Windmill Power

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 11-14 and listed in Table 11-5 (p.439).

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

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

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).

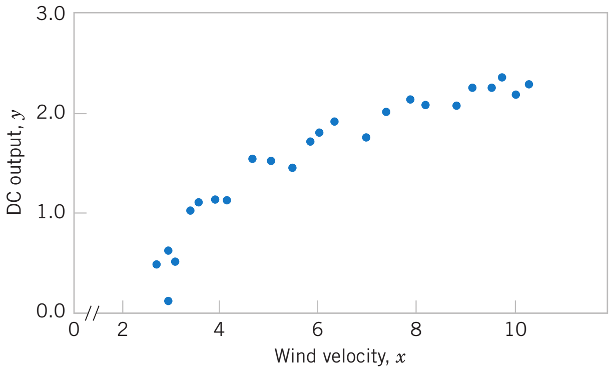


Figure 11-14 Plot of DC output y versus wind velocity x for the windmill data.

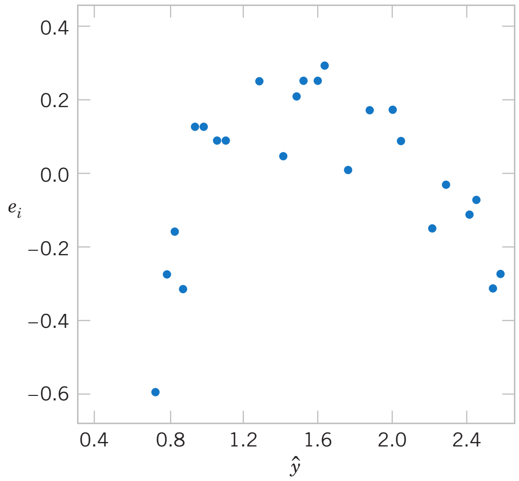


Figure 11-15 Plot of residuals e_i versus fitted values $\hat{y}_i$ for the windmill data.

A plot of the residuals versus $\hat{y}_i$ is shown in Figure 11-15. 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 11-14 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 of Figure 11-14 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

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

Figure 11-16 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 11-17. This plot does not reveal any serious problem with inequality of variance. The normal probability plot, shown in Figure 11-18, 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.

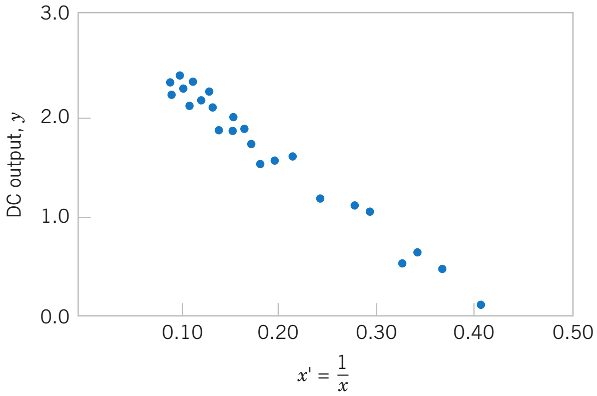


Figure 11-16 Plot of DC output versus $x' = 1/x$ for the windmill data.

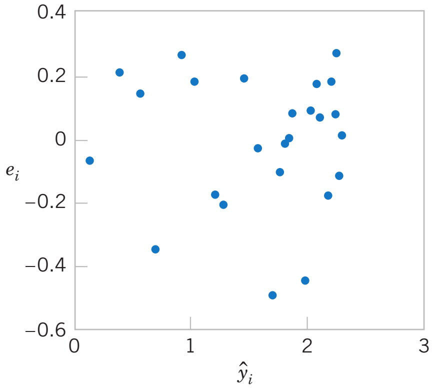


Figure 11-17 Plot of residuals versus fitted values $\hat{y}_i$ for the transformed model for the windmill data.

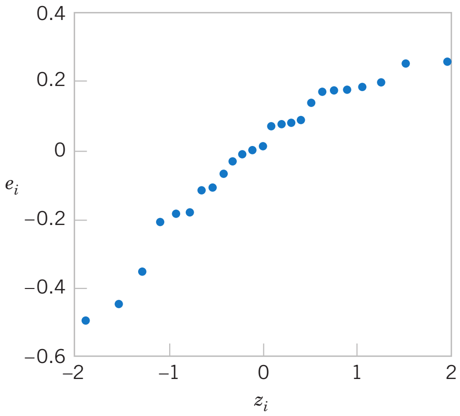


Figure 11-18 Normal probability plot of the residuals for the transformed model for the windmill data.

Table 11-5 Observed Values y_i and Regressor Variable x_i for Example 11-9

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

EXERCISES FOR SECTION 11–9

11-77. Determine if the following models are intrinsically linear. If yes, determine the appropriate transformation to generate the linear model.

(a) $Y = \beta_0 x^{\beta_1}\epsilon$ (b) $Y = \dfrac{3 + 5x}{x} + \epsilon$

(c) $Y = \beta_0 \beta_1^{x}\epsilon$ (d) $Y = \dfrac{x}{\beta_0 x + \beta_1 + x\epsilon}$

11-78. 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 Clausis–Clapeyron relationship 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.

11-79. 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

continued

Customer	x	y	Customer	x	y
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?

11-10 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 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

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{11-51}$$

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 11-51. First, note that if the response is binary, the error terms ϵ_i can only take on two values, namely,

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

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 11-51. 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 11-19, is usually employed. This function is called the **logit response function,** and has the form

$$E(Y) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} \tag{11-52}$$

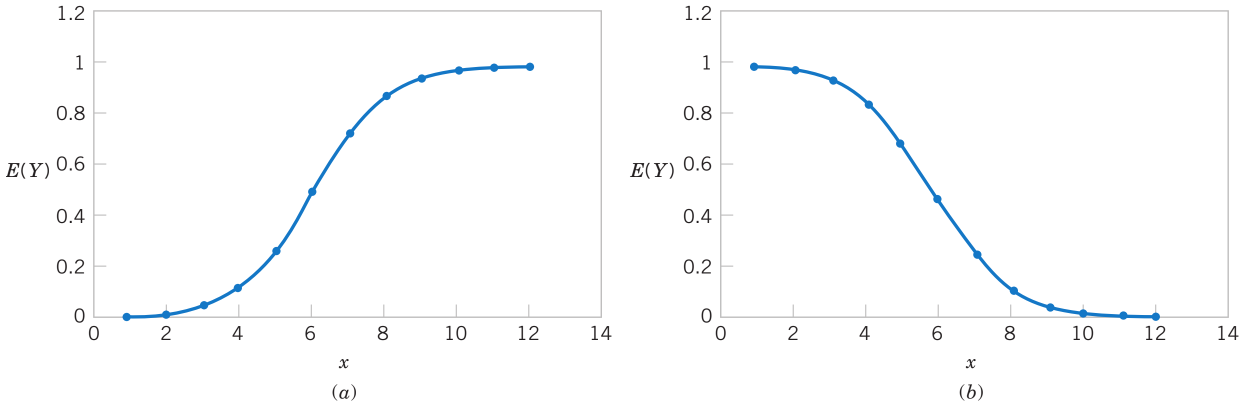


Figure 11-19 Examples of the logistic response function. (a) $E(Y) = 1/(1 + e^{-6.0-1.0x})$, (b) $E(Y) = 1/(1 + e^{-6.0+1.0x})$.

or equivalently,

$$E(Y) = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x)]} \tag{11-53}$$

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

$$\frac{E(Y)}{1 - E(Y)} = \exp(\beta_0 + \beta_1 x) \tag{11-54}$$

The quantity $\exp(\beta_0 + \beta_1 x)$ on the right-hand side of Equation 11-54 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 (2006). Minitab will fit logistic regression models and provide useful information on the quality of the fit.

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 to seal field joints on the rocket motor assembly. 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 11-20 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.

The fitted logistic regression model is

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

```
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 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 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 11-21 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

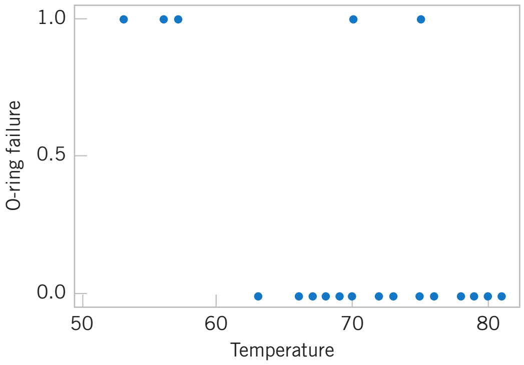


Figure 11-20 Scatter plot of O-ring failures versus launch temperature for 24 space shuttle flights.

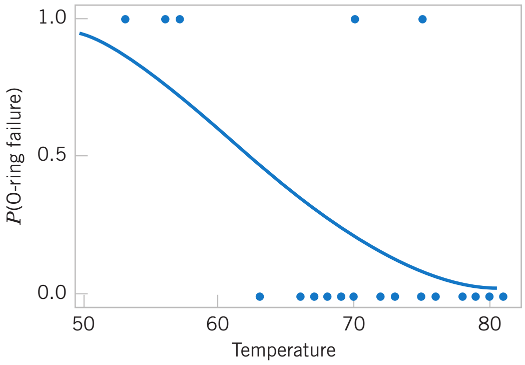


Figure 11-21 Probability of O-ring failure versus launch temperature (based on a logistic regression model).

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.**

EXERCISES FOR SECTION 11–10

11-80 A study was conducted attempting to relate home ownership to family income. Twenty households were selected and family income was estimated, along with information concerning home ownership ($y = 1$ indicates *yes* and $y = 0$ indicates *no*). The data are shown below.

Household	Income	Home Ownership Status
1	38,000	0
2	51,200	1
3	39,600	0
4	43,400	1
5	47,700	0
6	53,000	0
7	41,500	1
8	40,800	0
9	45,400	1
10	52,400	1
11	38,700	1
12	40,100	0
13	49,500	1
14	38,000	0
15	42,000	1
16	54,000	1
17	51,700	1
18	39,400	0
19	40,900	0
20	52,800	1

(a) Fit a logistic regression model to the response variable y. Use a simple linear regression model as the structure for the linear predictor.

(b) Is the logistic regression model in part (a) adequate?

(c) Provide an interpretation of the parameter β_1 in this model.

11-81 The compressive strength of an alloy fastener used in aircraft construction is being studied. Ten loads were selected over the range 2500– 4300 psi and a number of fasteners were tested at those loads. The numbers of fasteners failing at each load were recorded. The complete test data follow.

Load, x (psi)	Sample Size, n	Number Failing, r
2500	50	10
2700	70	17
2900	100	30
3100	60	21
3300	40	18
3500	85	43
3700	90	54
3900	50	33
4100	80	60
4300	65	51

(a) Fit a logistic regression model to the data. Use a simple linear regression model as the structure for the linear predictor.

(b) Is the logistic regression model in part (a) adequate?

11-82 The market research department of a soft drink manufacturer is investigating the effectiveness of a price discount coupon on the purchase of a two-liter beverage product. A sample of 5500 customers was given coupons for varying price discounts between 5 and 25 cents. The response variable was the number of coupons in each price discount category redeemed after one month. The data are shown below.

Discount, x	Sample Size, n	Number Redeemed, r
5	500	100
7	500	122
9	500	147
11	500	176
13	500	211
15	500	244
17	500	277
19	500	310
21	500	343
23	500	372
25	500	391

(a) Fit a logistic regression model to the data. Use a simple linear regression model as the structure for the linear predictor.

(b) Is the logistic regression model in part (a) adequate?

(c) Draw a graph of the data and the fitted logistic regression model.

(d) Expand the linear predictor to include a quadratic term. Is there any evidence that this quadratic term is required in the model?

(e) Draw a graph of this new model on the same plot that you prepared in part (c). Does the expanded model visually provide a better fit to the data than the original model from part (a)?

11-83 A study was performed to investigate new automobile purchases. A sample of 20 families was selected. Each family was surveyed to determine the age of their oldest vehicle and their total family income. A follow-up survey was conducted six months later to determine if they had actually purchased a new vehicle during that time period ($y = 1$ indicates *yes* and $y = 0$ indicates *no*). The data from this study are shown in the following table.

Income, x_1	Age, x_2	y	Income, x_1	Age, x_2	y
45,000	2	0	37,000	5	1
40,000	4	0	31,000	7	1
60,000	3	1	40,000	4	1
50,000	2	1	75,000	2	0
55,000	2	0	43,000	9	1
50,000	5	1	49,000	2	0
35,000	7	1	37,500	4	1
65,000	2	1	71,000	1	0
53,000	2	0	34,000	5	0
48,000	1	0	27,000	6	0

(a) Fit a logistic regression model to the data.

(b) Is the logistic regression model in part (a) adequate?

(c) Interpret the model coefficients β_1 and β_2.

(d) What is the estimated probability that a family with an income of $45,000 and a car that is five years old will purchase a new vehicle in the next six months?

(e) Expand the linear predictor to include an interaction term. Is there any evidence that this term is required in the model?

Supplemental Exercises

11-84. Show that, for the simple linear regression model, the following statements are true:

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

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

11-85. 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" (1991, Vol. 40, 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?

11-86. 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:

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

(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.

11-87. 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	9	8	11	6
y	0.1	0.13	0.09	0.15	0.20	0.21	0.18	0.24

11-88. The following data, adapted from Montgomery, Peck, and Vining (2006), 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

11-89. Consider the weight and blood pressure data in Exercise 11-72. Fit a no-intercept model to the data, and compare it to the model obtained in Exercise 11-70. Which model is superior?

11-90. 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.

11-91. 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 model.
(c) Test for significance of regression. Calculate R^2 for the model.
(d) Analyze the residuals and comment on model adequacy.

11-92. Two different methods can be used for measuring the temperature of the solution in a Hall cell used in aluminum smelting, a thermocouple implanted in the cell and an indirect measurement produced from an IR device. The indirect method is preferable because 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.

11-93. 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.

11-94. Cesium atoms cooled by laser light could be used to build inexpensive atomic clocks. In a study in *IEEE Transactions on Instrumentation and Measurement* (2001, Vol. 50, pp. 1224–1228), the number of atoms cooled by lasers of various powers were counted.

Power (mW)	Number of Atoms (×10E9)
11	0
12	0.02
18	0.08
21	0.13
22	0.15
24	0.18
28	0.31
32	0.4
37	0.49
39	0.57
41	0.64
46	0.71
48	0.79
50	0.82
51	0.83

(a) Graph the data and fit a regression line to predict the number of atoms from laser power. Comment on the adequacy of a linear model.

(b) Is there a significant regression at $\alpha = 0.05$? What is the *P*-value?

(c) Estimate the correlation coefficient.

(d) Test the hypothesis that $\rho = 0$ against the alternative $\rho \neq 0$ with $\alpha = 0.05$. What is the *P*-value?

(e) Compute a 95% confidence interval for the slope coefficient.

11-95. The following data related diamond carats to purchase prices. It appeared in Singapore's *Business Times*, February 18, 2000.

Carat	Price	Carat	Price
0.3	1302	0.33	1327
0.3	1510	0.33	1098
0.3	1510	0.34	1693
0.3	1260	0.34	1551
0.31	1641	0.34	1410
0.31	1555	0.34	1269
0.31	1427	0.34	1316
0.31	1427	0.34	1222
0.31	1126	0.35	1738
0.31	1126	0.35	1593
0.32	1468	0.35	1447
0.32	1202	0.35	1255
0.36	1635	0.45	1572
0.36	1485	0.46	2942
0.37	1420	0.48	2532
0.37	1420	0.5	3501
0.4	1911	0.5	3501
0.4	1525	0.5	3501
0.41	1956	0.5	3293
0.43	1747	0.5	3016

(a) Graph the data. What is the relation between carat and price? Is there an outlier?

(b) What would you say to the person who purchased the diamond that was an outlier?

(c) Fit two regression models, one with all the data and the other with unusual data omitted. Estimate the slope coefficient with a 95% confidence interval in both cases. Comment on any difference.

11-96. The following table shows the population and the average count of wood storks sighted per sample period for South Carolina from 1991 to 2004. Fit a regression line with population as the response and the count of wood storks as the predictor. Such an analysis might be used to evaluate the relationship between storks and babies. Is regression significant at $\alpha = 0.05$? What do you conclude about the role of regression analysis to establish a cause-and-effect relationship?

Year	Population	Stork Count
1991	3,559,470	0.342
1992	3,600,576	0.291
1993	3,634,507	0.291
1994	3,666,456	0.291
1995	3,699,943	0.291
1996	3,738,974	0.509
1997	3,790,066	0.294
1998	3,839,578	0.799
1999	3,885,736	0.542
2000	4,012,012	0.495
2001	4,061,209	0.859
2002	4,105,848	0.364
2003	4,148,744	0.501
2004	4,198,068	0.656

MIND-EXPANDING EXERCISES

11-97. 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.

11-98. 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 $\text{cov}(\bar{Y}, \hat{\beta}_1) = 0$.

11-99. 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 $E(MS_R) = \sigma^2 + \beta_1^2 S_{xx}$.

11-100. 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?

11-101. 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.

11-102. **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 which 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.

11-103. 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

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

11-104. 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

Analysis of variance test in regression
Confidence interval on mean response
Correlation coefficient
Empirical model
Confidence intervals on model parameters
Intrinsically linear model
Least squares estimation of regression model parameters
Logistic regression
Model adequacy checking
Odds ratio
Prediction interval on a future observation
Regression analysis
Residual plots
Residuals
Scatter diagram
Significance of regression
Simple linear regression model standard errors
Statistical tests on model parameters
Transformations

12

© David Lewis/
iStockphoto

Multiple Linear Regression

This chapter generalizes the simple linear regression to a situation where there is more than one predictor or regressor variable. This situation occurs frequently in science and engineering; for example, in Chapter 1 we provided data on the pull strength of a wire bond on a semiconductor package and illustrated its relationship to the wire length and the die height. Understanding the relationship between strength and the other two variables may provide important insight to the engineer when the package is designed, or to the manufacturing personnel who assemble the die into the package. We used a multiple **linear regression** model to relate strength to wire length and die height. There are many examples of such relationships: The life of a cutting tool is related to the cutting speed and the tool angle; patient satisfaction in a hospital is related to patient age, type of procedure performed, and length of stay; and the fuel economy of a vehicle is related to the type of vehicle (car versus truck), engine displacement, horsepower, type of transmission, and vehicle weight. Multiple regression models give insight into the relationships between these variables that can have important practical implications.

This chapter shows how to fit multiple linear regression models, perform the statistical tests and confidence procedures that are analogous to those for simple linear regression, and check for model adequacy. We also show how models that have polynomial terms in the regressor variables are just multiple linear regression models. We also discuss some aspects of building a good regression model from a collection of candidate regressors.

CHAPTER OUTLINE

LEARNING OBJECTIVES

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

1. Use multiple regression techniques to build empirical models to engineering and scientific data
2. Understand how the method of least squares extends to fitting multiple regression models
3. Assess regression model adequacy
4. Test hypotheses and construct confidence intervals on the regression coefficients
5. Use the regression model to estimate the mean response and to make predictions and to construct confidence intervals and prediction intervals
6. Build regression models with polynomial terms
7. Use indicator variables to model categorical regressors
8. Use stepwise regression and other model building techniques to select the appropriate set of variables for a regression model

12-1 MULTIPLE LINEAR REGRESSION MODEL

12-1.1 Introduction

Many applications of regression analysis involve situations in which there are more than one regressor or predictor variable. A regression model that contains more than one regressor variable is called a **multiple regression model.**

As an example, suppose that the effective life of a cutting tool depends on the cutting speed and the tool angle. A multiple regression model that might describe this relationship is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \qquad (12\text{-}1)$$

where Y represents the tool life, x_1 represents the cutting speed, x_2 represents the tool angle, and ϵ is a random error term. This is a multiple linear regression model with two regressors. The term **linear** is used because Equation 12-1 is a linear function of the unknown parameters β_0, β_1, and β_2.

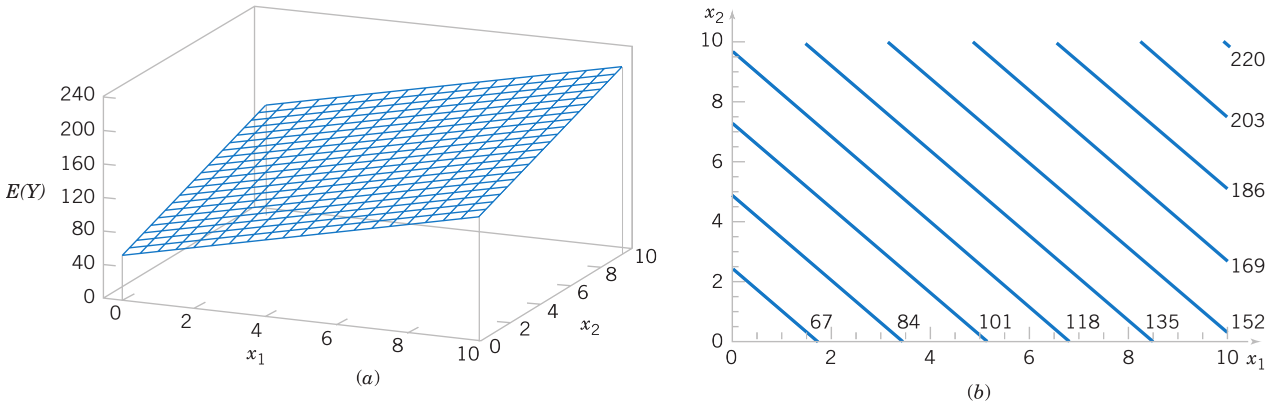


Figure 12-1 (a) The regression plane for the model $E(Y) = 50 + 10x_1 + 7x_2$. (b) The contour plot.

The regression model in Equation 12-1 describes a plane in the three-dimensional space of Y, x_1, and x_2. Figure 12-1(a) shows this plane for the regression model

$$E(Y) = 50 + 10x_1 + 7x_2$$

where we have assumed that the expected value of the error term is zero; that is $E(\epsilon) = 0$. The parameter β_0 is the **intercept** of the plane. We sometimes call β_1 and β_2 **partial regression coefficients,** because β_1 measures the expected change in Y per unit change in x_1 when x_2 is held constant, and β_2 measures the expected change in Y per unit change in x_2 when x_1 is held constant. Figure 12-1(b) shows a **contour plot** of the regression model—that is, lines of constant $E(Y)$ as a function of x_1 and x_2. Notice that the contour lines in this plot are straight lines.

In general, the **dependent variable** or **response** Y may be related to k **independent** or **regressor variables.** The model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon \tag{12-2}$$

is called a multiple linear regression model with k regressor variables. The parameters β_j, $j = 0, 1, \ldots, k$, are called the regression coefficients. This model describes a hyperplane in the k-dimensional space of the regressor variables $\{x_j\}$. The parameter β_j represents the expected change in response Y per unit change in x_j when all the remaining regressors x_i $(i \neq j)$ are held constant.

Multiple linear regression models are often used as approximating functions. That is, the true functional relationship between Y and $x_1, x_2, \ldots, x_k$ is unknown, but over certain ranges of the independent variables the linear regression model is an adequate approximation.

Models that are more complex in structure than Equation 12-2 may often still be analyzed by multiple linear regression techniques. For example, consider the **cubic polynomial** model in one regressor variable.

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon \tag{12-3}$$

If we let $x_1 = x$, $x_2 = x^2$, $x_3 = x^3$, Equation 12-3 can be written as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon \tag{12-4}$$

which is a multiple linear regression model with three regressor variables.

Models that include **interaction** effects may also be analyzed by multiple linear regression methods. An interaction between two variables can be represented by a cross-product term in the model, such as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon \tag{12-5}$$

If we let $x_3 = x_1 x_2$ and $\beta_3 = \beta_{12}$, Equation 12-5 can be written as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

which is a linear regression model.

Figure 12-2(a) and (b) shows the three-dimensional plot of the regression model

$$Y = 50 + 10x_1 + 7x_2 + 5x_1 x_2$$

and the corresponding two-dimensional contour plot. Notice that, although this model is a linear regression model, the shape of the surface that is generated by the model is not linear. In general, **any regression model that is linear in parameters** (the β's) **is a linear regression model, regardless of the shape of the surface that it generates.**

Figure 12-2 provides a nice graphical interpretation of an interaction. Generally, interaction implies that the effect produced by changing one variable (x_1, say) depends on the level of the other variable (x_2). For example, Fig. 12-2 shows that changing x_1 from 2 to 8 produces a much smaller change in $E(Y)$ when $x_2 = 2$ than when $x_2 = 10$. Interaction effects occur frequently in the study and analysis of real-world systems, and regression methods are one of the techniques that we can use to describe them.

As a final example, consider the second-order model with interaction

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \tag{12-6}$$

If we let $x_3 = x_1^2$, $x_4 = x_2^2$, $x_5 = x_1 x_2$, $\beta_3 = \beta_{11}$, $\beta_4 = \beta_{22}$, and $\beta_5 = \beta_{12}$, Equation 12-6 can be written as a multiple linear regression model as follows:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$$

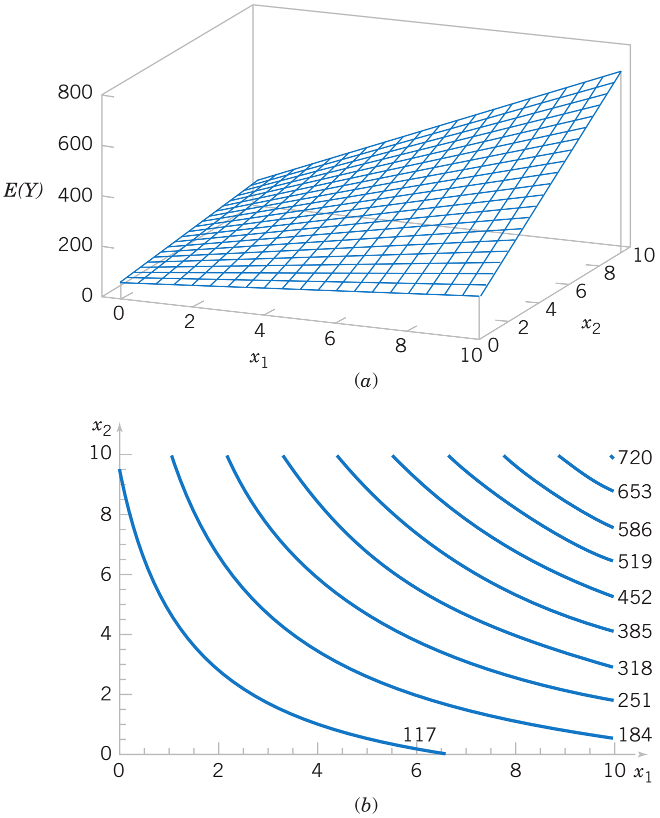
Figure 12-3(a) and (b) show the three-dimensional plot and the corresponding contour plot for

$$E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1 x_2$$

These plots indicate that the expected change in Y when x_1 is changed by one unit (say) is a function of *both* x_1 and x_2. The quadratic and interaction terms in this model produce a mound-shaped function. Depending on the values of the regression coefficients, the second-order model with interaction is capable of assuming a wide variety of shapes; thus, it is a very flexible regression model.

12-1.2 Least Squares Estimation of the Parameters

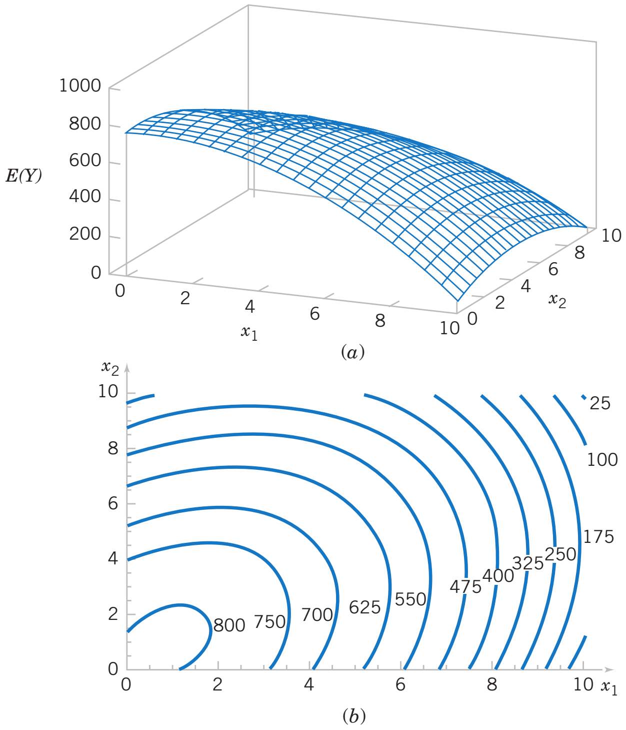
The **method of least squares** may be used to estimate the regression coefficients in the multiple regression model, Equation 12-2. Suppose that $n > k$ observations are available, and let



(*a*)

(*b*)

Figure 12-2 (a) Three-dimensional plot of the regression model $E(Y) = 50 + 10x_1 + 7x_2 + 5x_1x_2$. (b) The contour plot.



(*a*)

(*b*)

Figure 12-3 (a) Three-dimensional plot of the regression model $E(Y) = 800 + 10x_1 + 7x_2 - 8.5x_1^2 - 5x_2^2 + 4x_1x_2$. (b) The contour plot.

x_{ij} denote the ith observation or level of variable x_j. The observations are

$$(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i), \qquad i = 1, 2, \ldots, n \quad \text{and} \quad n > k$$

It is customary to present the data for multiple regression in a table such as Table 12-1. Each observation $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, satisfies the model in Equation 12-2, or

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \\ &= \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i \qquad i = 1, 2, \ldots, n \end{aligned} \tag{12-7}$$

Table 12-1 Data for Multiple Linear Regression

y	x_1	x_2	...	x_k
y_1	x_{11}	x_{12}	...	x_{1k}
y_2	x_{21}	x_{22}	...	x_{2k}
⋮	⋮	⋮		⋮
y_n	x_{n1}	x_{n2}	...	x_{nk}

The least squares function is

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2 \qquad (12\text{-}8)$$

We want to minimize L with respect to $\beta_0, \beta_1, \ldots, \beta_k$. The **least squares estimates** of $\beta_0, \beta_1, \ldots, \beta_k$ must satisfy

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat\beta_0 - \sum_{j=1}^{k} \hat\beta_j x_{ij} \right) = 0 \qquad (12\text{-}9a)$$

and

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat\beta_0 - \sum_{j=1}^{k} \hat\beta_j x_{ij} \right) x_{ij} = 0 \quad j = 1, 2, \ldots, k \qquad (12\text{-}9b)$$

Simplifying Equation 12-9, we obtain the **least squares normal equations**

$$\begin{array}{lllllll}
n\hat\beta_0 & + \hat\beta_1 \sum_{i=1}^{n} x_{i1} & + \hat\beta_2 \sum_{i=1}^{n} x_{i2} & + \cdots + & \hat\beta_k \sum_{i=1}^{n} x_{ik} & = & \sum_{i=1}^{n} y_i \\
\hat\beta_0 \sum_{i=1}^{n} x_{i1} & + \hat\beta_1 \sum_{i=1}^{n} x_{i1}^2 & + \hat\beta_2 \sum_{i=1}^{n} x_{i1} x_{i2} & + \cdots + & \hat\beta_k \sum_{i=1}^{n} x_{i1} x_{ik} & = & \sum_{i=1}^{n} x_{i1} y_i \\
\vdots & \vdots & \vdots & \vdots & \vdots & & \vdots \\
\hat\beta_0 \sum_{i=1}^{n} x_{ik} & + \hat\beta_1 \sum_{i=1}^{n} x_{ik} x_{i1} & + \hat\beta_2 \sum_{i=1}^{n} x_{ik} x_{i2} & + \cdots + & \hat\beta_k \sum_{i=1}^{n} x_{ik}^2 & = & \sum_{i=1}^{n} x_{ik} y_i
\end{array} \qquad (12\text{-}10)$$

Note that there are $p = k + 1$ normal equations, one for each of the unknown regression coefficients. The solution to the normal equations will be the **least squares estimators** of the regression coefficients, $\hat\beta_0, \hat\beta_1, \ldots, \hat\beta_k$. The normal equations can be solved by any method appropriate for solving a system of linear equations.

EXAMPLE 12-1 Wire Bond Strength

In Chapter 1, we used data on pull strength of a wire bond in a semiconductor manufacturing process, wire length, and die height to illustrate building an empirical model. We will use the same data, repeated for convenience in Table 12-2, and show the details of estimating the model parameters. A three-dimensional scatter plot of the data is presented in Fig. 1-15. Figure 12-4 shows a matrix of two-dimensional scatter plots of the data. These displays can be helpful in visualizing the relationships among variables in a multivariable data set. For example, the plot indicates that there is a strong linear relationship between strength and wire length.

Specifically, we will fit the multiple linear regression model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where Y = pull strength, x_1 = wire length, and x_2 = die height. From the data in Table 12-2 we calculate

$$n = 25, \sum_{i=1}^{25} y_i = 725.82$$

$$\sum_{i=1}^{25} x_{i1} = 206, \sum_{i=1}^{25} x_{i2} = 8{,}294$$

$$\sum_{i=1}^{25} x_{i1}^2 = 2{,}396, \sum_{i=1}^{25} x_{i2}^2 = 3{,}531{,}848$$

$$\sum_{i=1}^{25} x_{i1} x_{i2} = 77{,}177, \sum_{i=1}^{25} x_{i1} y_i = 8{,}008.47,$$

$$\sum_{i=1}^{25} x_{i2} y_i = 274{,}816.71$$

Table 12-2 Wire Bond Data for Example 12-1

Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2	Observation Number	Pull Strength y	Wire Length x_1	Die Height x_2
1	9.95	2	50	14	11.66	2	360
2	24.45	8	110	15	21.65	4	205
3	31.75	11	120	16	17.89	4	400
4	35.00	10	550	17	69.00	20	600
5	25.02	8	295	18	10.30	1	585
6	16.86	4	200	19	34.93	10	540
7	14.38	2	375	20	46.59	15	250
8	9.60	2	52	21	44.88	15	290
9	24.35	9	100	22	54.12	16	510
10	27.50	8	300	23	56.63	17	590
11	17.08	4	412	24	22.13	6	100
12	37.00	11	400	25	21.15	5	400
13	41.95	12	500				

For the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$, the normal equations 12-10 are

$$\begin{aligned}
n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} &= \sum_{i=1}^{n} y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} &= \sum_{i=1}^{n} x_{i1}y_i \\
\hat{\beta}_0 \sum_{i=1}^{n} x_{i2} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}x_{i2} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2}^2 &= \sum_{i=1}^{n} x_{i2}y_i
\end{aligned}$$

Inserting the computed summations into the normal equations, we obtain

$$\begin{aligned}
25\hat{\beta}_0 + \quad 206\hat{\beta}_1 + \quad 8294\hat{\beta}_2 &= 725.82 \\
206\hat{\beta}_0 + \quad 2396\hat{\beta}_1 + \quad 77{,}177\hat{\beta}_2 &= 8{,}008.47 \\
8294\hat{\beta}_0 + 77{,}177\hat{\beta}_1 + 3{,}531{,}848\hat{\beta}_2 &= 274{,}816.71
\end{aligned}$$

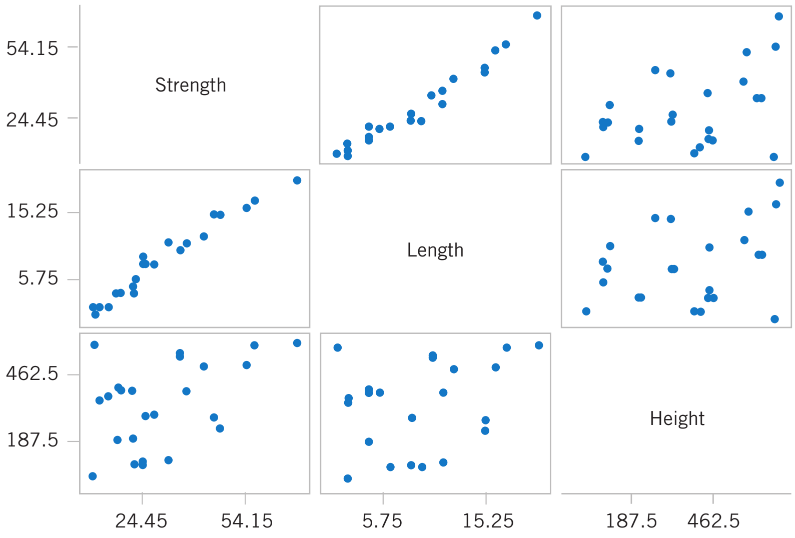


Figure 12-4 Matrix of scatter plots (from Minitab) for the wire bond pull strength data in Table 12-2.

The solution to this set of equations is

$$\hat{\beta}_0 = 2.26379, \quad \hat{\beta}_1 = 2.74427, \quad \hat{\beta}_2 = 0.01253$$

Therefore, the fitted regression equation is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

Practical Interpretation: This equation can be used to predict pull strength for pairs of values of the regressor variables wire length (x_1) and die height (x_2). This is essentially the same regression model given in Section 1-3. Figure 1-16 shows a three-dimensional plot of the plane of predicted values $\hat{y}$ generated from this equation.

12-1.3 Matrix Approach to Multiple Linear Regression

In fitting a multiple regression model, it is much more convenient to express the mathematical operations using **matrix notation.** Suppose that there are k regressor variables and n observations, $(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i)$, $i = 1, 2, \ldots, n$ and that the model relating the regressors to the response is

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad i = 1, 2, \ldots, n$$

This model is a system of n equations that can be expressed in matrix notation as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{12-11}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1k} \\ 1 & x_{21} & x_{22} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nk} \end{bmatrix} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \quad \text{and} \quad \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

In general, $\mathbf{y}$ is an $(n \times 1)$ vector of the observations, $\mathbf{X}$ is an $(n \times p)$ matrix of the levels of the independent variables (assuming that the intercept is always multiplied by a constant value—unity), $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of the regression coefficients, and $\boldsymbol{\epsilon}$ is a $(n \times 1)$ vector of random errors. The $\mathbf{X}$ matrix is often called the **model matrix.**

We wish to find the vector of least squares estimators, $\hat{\boldsymbol{\beta}}$, that minimizes

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \boldsymbol{\epsilon}'\boldsymbol{\epsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The least squares estimator $\hat{\boldsymbol{\beta}}$ is the solution for $\boldsymbol{\beta}$ in the equations

$$\frac{\partial L}{\partial \boldsymbol{\beta}} = \mathbf{0}$$

We will not give the details of taking the derivatives above; however, the resulting equations that must be solved are

Normal Equations

$$\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y} \tag{12-12}$$

Equations 12-12 are the least squares normal equations in matrix form. They are identical to the scalar form of the normal equations given earlier in Equations 12-10. To solve the normal equations, multiply both sides of Equations 12-12 by the inverse of **X′X.** Therefore, the least squares estimate of $\boldsymbol{\beta}$ is

Least Square Estimate of β

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y} \qquad (12\text{-}13)$$

Note that there are $p = k + 1$ normal equations in $p = k + 1$ unknowns (the values of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$). Furthermore, the matrix **X′X** is always nonsingular, as was assumed above, so the methods described in textbooks on determinants and matrices for inverting these matrices can be used to find $(\mathbf{X}'\mathbf{X})^{-1}$. In practice, multiple regression calculations are almost always performed using a computer.

It is easy to see that the matrix form of the normal equations is identical to the scalar form. Writing out Equation 12-12 in detail, we obtain

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \sum_{i=1}^{n} x_{i1}x_{i2} & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{ik}x_{i1} & \sum_{i=1}^{n} x_{ik}x_{i2} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{bmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_{i1}y_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}y_i \end{bmatrix}$$

If the indicated matrix multiplication is performed, the scalar form of the normal equations (that is, Equation 12-10) will result. In this form it is easy to see that **X′X** is a $(p \times p)$ symmetric matrix and **X′y** is a $(p \times 1)$ column vector. Note the special structure of the **X′X** matrix. The diagonal elements of **X′X** are the sums of squares of the elements in the columns of **X,** and the off-diagonal elements are the sums of cross-products of the elements in the columns of **X.** Furthermore, note that the elements of **X′y** are the sums of cross-products of the columns of **X** and the observations $\{y_i\}$.

The fitted regression model is

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \qquad i = 1, 2, \ldots, n \qquad (12\text{-}14)$$

In matrix notation, the fitted model is

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$$

The difference between the observation y_i and the fitted value $\hat{y}_i$ is a **residual,** say, $e_i = y_i - \hat{y}_i$. The $(n \times 1)$ vector of residuals is denoted by

$$\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} \qquad (12\text{-}15)$$

EXAMPLE 12-2 Wire Bond Strength with Matrix Notation

In Example 12-1, we illustrated fitting the multiple regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where y is the observed pull strength for a wire bond, x_1 is the wire length, and x_2 is the die height. The 25 observations are in Table 12-2. We will now use the matrix approach to fit the regression model above to these data. The model matrix $\mathbf{X}$ and $\mathbf{y}$ vector for this model are

$$\mathbf{X} = \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ 1 & 11 & 120 \\ 1 & 10 & 550 \\ 1 & 8 & 295 \\ 1 & 4 & 200 \\ 1 & 2 & 375 \\ 1 & 2 & 52 \\ 1 & 9 & 100 \\ 1 & 8 & 300 \\ 1 & 4 & 412 \\ 1 & 11 & 400 \\ 1 & 12 & 500 \\ 1 & 2 & 360 \\ 1 & 4 & 205 \\ 1 & 4 & 400 \\ 1 & 20 & 600 \\ 1 & 1 & 585 \\ 1 & 10 & 540 \\ 1 & 15 & 250 \\ 1 & 15 & 290 \\ 1 & 16 & 510 \\ 1 & 17 & 590 \\ 1 & 6 & 100 \\ 1 & 5 & 400 \end{bmatrix} \qquad \mathbf{y} = \begin{bmatrix} 9.95 \\ 24.45 \\ 31.75 \\ 35.00 \\ 25.02 \\ 16.86 \\ 14.38 \\ 9.60 \\ 24.35 \\ 27.50 \\ 17.08 \\ 37.00 \\ 41.95 \\ 11.66 \\ 21.65 \\ 17.89 \\ 69.00 \\ 10.30 \\ 34.93 \\ 46.59 \\ 44.88 \\ 54.12 \\ 56.63 \\ 22.13 \\ 21.15 \end{bmatrix}$$

The $\mathbf{X}'\mathbf{X}$ matrix is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 1 & 2 & 50 \\ 1 & 8 & 110 \\ \vdots & \vdots & \vdots \\ 1 & 5 & 400 \end{bmatrix}$$

$$= \begin{bmatrix} 25 & 206 & 8{,}294 \\ 206 & 2{,}396 & 77{,}177 \\ 8{,}294 & 77{,}177 & 3{,}531{,}848 \end{bmatrix}$$

and the $\mathbf{X}'\mathbf{y}$ vector is

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 2 & 8 & \cdots & 5 \\ 50 & 110 & \cdots & 400 \end{bmatrix} \begin{bmatrix} 9.95 \\ 24.45 \\ \vdots \\ 21.15 \end{bmatrix} = \begin{bmatrix} 725.82 \\ 8{,}008.47 \\ 274{,}816.71 \end{bmatrix}$$

The least squares estimates are found from Equation 12-13 as

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

or

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 25 & 206 & 8{,}294 \\ 206 & 2{,}396 & 77{,}177 \\ 8{,}294 & 77{,}177 & 3{,}531{,}848 \end{bmatrix}^{-1} \begin{bmatrix} 725.82 \\ 8{,}008.37 \\ 274{,}811.31 \end{bmatrix}$$

$$= \begin{bmatrix} 0.214653 & -0.007491 & -0.000340 \\ -0.007491 & 0.001671 & -0.000019 \\ -0.000340 & -0.000019 & +0.0000015 \end{bmatrix} \begin{bmatrix} 725.82 \\ 8{,}008.47 \\ 274{,}811.31 \end{bmatrix}$$

$$= \begin{bmatrix} 2.26379143 \\ 2.74426964 \\ 0.01252781 \end{bmatrix}$$

Therefore, the fitted regression model with the regression coefficients rounded to five decimal places is

$$\hat{y} = 2.26379 + 2.74427x_1 + 0.01253x_2$$

This is identical to the results obtained in Example 12-1.

This regression model can be used to predict values of pull strength for various values of wire length (x_1) and die height (x_2). We can also obtain the **fitted values** $\hat{y}_i$ by substituting each observation (x_{i1}, x_{i2}), $i = 1, 2, \ldots, n$, into the equation. For example, the first observation has $x_{11} = 2$ and $x_{12} = 50$, and the fitted value is

$$\begin{aligned} \hat{y}_1 &= 2.26379 + 2.74427x_{11} + 0.01253x_{12} \\ &= 2.26379 + 2.74427(2) + 0.01253(50) \\ &= 8.38 \end{aligned}$$

The corresponding observed value is $y_1 = 9.95$. The *residual* corresponding to the first observation is

$$\begin{aligned} e_1 &= y_1 - \hat{y}_1 \\ &= 9.95 - 8.38 \\ &= 1.57 \end{aligned}$$

Table 12-3 displays all 25 fitted values $\hat{y}_i$ and the corresponding residuals. The fitted values and residuals are calculated to the same accuracy as the original data.

Table 12-3 Observations, Fitted Values, and Residuals for Example 12-2

Observation Number	y_i	$\hat{y}_i$	$e_i = y_i - \hat{y}_i$
1	9.95	8.38	1.57
2	24.45	25.60	−1.15
3	31.75	33.95	−2.20
4	35.00	36.60	−1.60
5	25.02	27.91	−2.89
6	16.86	15.75	1.11
7	14.38	12.45	1.93
8	9.60	8.40	1.20
9	24.35	28.21	−3.86
10	27.50	27.98	−0.48
11	17.08	18.40	−1.32
12	37.00	37.46	−0.46
13	41.95	41.46	0.49
14	11.66	12.26	−0.60
15	21.65	15.81	5.84
16	17.89	18.25	−0.36
17	69.00	64.67	4.33
18	10.30	12.34	−2.04
19	34.93	36.47	−1.54
20	46.59	46.56	0.03
21	44.88	47.06	−2.18
22	54.12	52.56	1.56
23	56.63	56.31	0.32
24	22.13	19.98	2.15
25	21.15	21.00	0.15

Computers are almost always used in fitting multiple regression models. Table 12-4 presents some annotated output from Minitab for the least squares regression model for wire bond pull strength data. The upper part of the table contains the numerical estimates of the regression coefficients. The computer also calculates several other quantities that reflect important information about the regression model. In subsequent sections, we will define and explain the quantities in this output.

Estimating σ^2

Just as in simple linear regression, it is important to estimate σ^2, the variance of the error term ϵ, in a multiple regression model. Recall that in simple linear regression the estimate of σ^2 was obtained by dividing the sum of the squared residuals by $n - 2$. Now there are two parameters in the simple linear regression model, so in multiple linear regression with p parameters a logical estimator for σ^2 is

Estimator of Variance

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-p} = \frac{SS_E}{n-p} \tag{12-16}$$

This is an **unbiased estimator** of σ^2. Just as in simple linear regression, the estimate of σ^2 is usually obtained from the **analysis of variance** for the regression model. The numerator of Equation 12-16 is called the **error** or **residual sum of squares,** and the denominator $n - p$ is called the **error** or **residual degrees of freedom.**

We can find a computing formula for SS_E as follows:

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e}$$

Substituting $\mathbf{e} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ into the above, we obtain

$$\begin{aligned} SS_E &= \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \\ &= 27{,}178.5316 - 27{,}063.3581 = 115.174 \end{aligned} \tag{12-17}$$

Table 12-4 Minitab Multiple Regression Output for the Wire Bond Pull Strength Data

```
Regression Analysis: Strength versus Length, Height

 The regression equation is
 Strength = 2.26 + 2.74 Length + 0.0125 Height

 Predictor               Coef     SE Coef       T        P     VIF
 Constant     β̂0 →     2.264       1.060     2.14    0.044
 Length      β̂1 →    2.74427     0.09352    29.34    0.000     1.2
 Height     β̂2 →   0.012528    0.002798     4.48    0.000     1.2

 S = 2.288              R-Sq = 98.1%              R-Sq (adj) = 97.9%
 PRESS = 156.163        R-Sq (pred) = 97.44%

 Analysis of Variance

 Source           DF       SS        MS              F         P
 Regression        2   5990.8    2995.4         572.17     0.000
 Residual Error   22    115.2       5.2 ← σ̂2
 Total            24   6105.9

 Source    DF    Seq SS
 Length     1    5885.9
 Height     1     104.9

 Predicted Values for New Observations

 New Obs       Fit    SE Fit        95.0%  CI           95.0%  PI
     1      27.663     0.482   (26.663,  28.663)   (22.814,  32.512)

 Values of Predictors for New Observations

 New Obs    Length    Height
     1        8.00       275
```

Table 12-4 shows that the estimate of σ^2 for the wire bond pull strength regression model is $\hat{\sigma}^2 = 115.2/22 = 5.2364$. The Minitab output rounds the estimate to $\hat{\sigma}^2 = 5.2$.

12-1.4 Properties of the Least Squares Estimators

The statistical properties of the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ may be easily found, under certain assumptions on the error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$, in the regression model. Paralleling the assumptions made in Chapter 11, we assume that the errors ϵ_i are statistically independent with mean zero and variance σ^2. Under these assumptions, the least squares estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ are **unbiased estimators** of the regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$. This property may be shown as follows:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})] \\ &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}] \\ &= \boldsymbol{\beta} \end{aligned}$$

since $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} = \mathbf{I}$, the identity matrix. Thus, $\hat{\boldsymbol{\beta}}$ is an unbiased estimator of $\boldsymbol{\beta}$.

The variances of the $\hat{\boldsymbol{\beta}}$'s are expressed in terms of the elements of the inverse of the $\mathbf{X}'\mathbf{X}$ matrix. The inverse of $\mathbf{X}'\mathbf{X}$ times the constant σ^2 represents the **covariance matrix** of the regression coefficients $\hat{\boldsymbol{\beta}}$**.** The diagonal elements of $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$ are the variances of $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, and the off-diagonal elements of this matrix are the covariances. For example, if we have $k = 2$ regressors, such as in the pull-strength problem,

$$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} C_{00} & C_{01} & C_{02} \\ C_{10} & C_{11} & C_{12} \\ C_{20} & C_{21} & C_{22} \end{bmatrix}$$

which is symmetric ($C_{10} = C_{01}$, $C_{20} = C_{02}$, and $C_{21} = C_{12}$) because $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric, and we have

$$\begin{aligned} V(\hat{\beta}_j) &= \sigma^2 C_{jj}, \qquad j = 0, 1, 2 \\ \text{cov}(\hat{\beta}_i, \hat{\beta}_j) &= \sigma^2 C_{ij}, \qquad i \neq j \end{aligned}$$

In general, the covariance matrix of $\hat{\boldsymbol{\beta}}$ is a $(p \times p)$ symmetric matrix whose jjth element is the variance of $\hat{\beta}_j$ and whose i, jth element is the covariance between $\hat{\beta}_i$ and $\hat{\beta}_j$, that is,

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2\mathbf{C}$$

The estimates of the variances of these regression coefficients are obtained by replacing σ^2 with an estimate. When σ^2 is replaced by its estimate $\hat{\sigma}^2$, the square root of the estimated variance of the jth regression coefficient is called the **estimated standard error** of $\hat{\beta}_j$ or $se(\hat{\beta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$. These standard errors are a useful measure of the **precision of estimation** for the regression coefficients; small standard errors imply good precision.

Multiple regression computer programs usually display these standard errors. For example, the Minitab output in Table 12-4 reports $se(\hat{\beta}_0) = 1.060$, $se(\hat{\beta}_1) = 0.09352$, and $se(\hat{\beta}_2) = 0.002798$. The intercept estimate is about twice the magnitude of its standard error, and $\hat{\beta}_1$ and $\hat{\beta}_2$ are considerably larger than $se(\hat{\beta}_1)$ and $se(\hat{\beta}_2)$. This implies reasonable precision of estimation, although the parameters β_1 and β_2 are much more precisely estimated than the intercept (this is not unusual in multiple regression).

EXERCISES FOR SECTION 12-1

12-1. A study was performed to investigate the shear strength of soil (y) as it related to depth in feet (x_1) and % moisture content (x_2). Ten observations were collected, and the following summary quantities obtained: $n = 10$, $\sum x_{i1} = 223$, $\sum x_{i2} = 553$, $\sum y_i = 1{,}916$, $\sum x_{i1}^2 = 5{,}200.9$, $\sum x_{i2}^2 = 31{,}729$, $\sum x_{i1}x_{i2} = 12{,}352$, $\sum x_{i1}y_i = 43{,}550.8$, $\sum x_{i2}y_i = 104{,}736.8$, and $\sum y_i^2 = 371{,}595.6$.

(a) Set up the least squares normal equations for the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$.

(b) Estimate the parameters in the model in part (a).

(c) What is the predicted strength when $x_1 = 18$ feet and $x_2 = 43\%$?

12-2. A regression model is to be developed for predicting the ability of soil to absorb chemical contaminants. Ten observations have been taken on a soil absorption index (y) and two regressors: x_1 = amount of extractable iron ore and x_2 = amount of bauxite. We wish to fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$. Some necessary quantities are:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1.17991 & -7.30982\text{ E-3} & 7.3006\text{ E-4} \\ -7.30982\text{ E-3} & 7.9799\text{ E-5} & -1.23713\text{ E-4} \\ 7.3006\text{ E-4} & -1.23713\text{ E-4} & 4.6576\text{ E-4} \end{bmatrix},$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 220 \\ 36{,}768 \\ 9{,}965 \end{bmatrix}$$

(a) Estimate the regression coefficients in the model specified above.

(b) What is the predicted value of the absorption index y when $x_1 = 200$ and $x_2 = 50$?

12-3. A chemical engineer is investigating how the amount of conversion of a product from a raw material (y) depends on reaction temperature (x_1) and the reaction time (x_2). He has developed the following regression models:

1. $\hat{y} = 100 + 2x_1 + 4x_2$

2. $\hat{y} = 95 + 1.5x_1 + 3x_2 + 2x_1x_2$

Both models have been built over the range $0.5 \le x_2 \le 10$.

(a) What is the predicted value of conversion when $x_2 = 2$? Repeat this calculation for $x_2 = 8$. Draw a graph of the predicted values for both conversion models. Comment on the effect of the interaction term in model 2.

(b) Find the expected change in the mean conversion for a unit change in temperature x_1 for model 1 when $x_2 = 5$. Does this quantity depend on the specific value of reaction time selected? Why?

(c) Find the expected change in the mean conversion for a unit change in temperature x_1 for model 2 when $x_2 = 5$. Repeat this calculation for $x_2 = 2$ and $x_2 = 8$. Does the result depend on the value selected for x_2? Why?

12-4. You have fit a multiple linear regression model and the $(\mathbf{X'X})^{-1}$ matrix is:

$$(\mathbf{X'X})^{-1} = \begin{bmatrix} 0.893758 & -0.0282448 & -0.0175641 \\ -0.028245 & 0.0013329 & 0.0001547 \\ -0.017564 & 0.0001547 & 0.0009108 \end{bmatrix}$$

(a) How many regressor variables are in this model?

(b) If the error sum of squares is 307 and there are 15 observations, what is the estimate of σ^2?

(c) What is the standard error of the regression coefficient $\hat{\beta}_1$?

12-5. Data from a patient satisfaction survey in a hospital are shown in the following table:

Observation	Age	Severity	Surg-Med	Anxiety	Satisfaction
1	55	50	0	2.1	68
2	46	24	1	2.8	77
3	30	46	1	3.3	96
4	35	48	1	4.5	80
5	59	58	0	2.0	43
6	61	60	0	5.1	44
7	74	65	1	5.5	26
8	38	42	1	3.2	88
9	27	42	0	3.1	75
10	51	50	1	2.4	57
11	53	38	1	2.2	56
12	41	30	0	2.1	88
13	37	31	0	1.9	88
14	24	34	0	3.1	102
15	42	30	0	3.0	88
16	50	48	1	4.2	70
17	58	61	1	4.6	52
18	60	71	1	5.3	43
19	62	62	0	7.2	46
20	68	38	0	7.8	56
21	70	41	1	7.0	59
22	79	66	1	6.2	26
23	63	31	1	4.1	52
24	39	42	0	3.5	83
25	49	40	1	2.1	75

The regressor variables are the patient's age, an illness severity index (larger values indicate greater severity), an indicator variable denoting whether the patient is a medical patient (0) or a surgical patient (1), and an anxiety index (larger values indicate greater anxiety).

(a) Fit a multiple linear regression model to the satisfaction response using age, illness severity, and the anxiety index as the regressors.

(b) Estimate σ^2.

(c) Find the standard errors of the regression coefficients.

(d) Are all of the model parameters estimated with nearly the same precision? Why or why not?

12-6. The electric power consumed each month by a chemical plant is thought to be related to the average ambient temperature (x_1), the number of days in the month (x_2), the average product purity (x_3), and the tons of product produced (x_4). The past year's historical data are available and are presented in the following table:

y	x_1	x_2	x_3	x_4
240	25	24	91	100
236	31	21	90	95
270	45	24	88	110
274	60	25	87	88
301	65	25	91	94
316	72	26	94	99
300	80	25	87	97
296	84	25	86	96
267	75	24	88	110
276	60	25	91	105
288	50	25	90	100
261	38	23	89	98

(a) Fit a multiple linear regression model to these data.

(b) Estimate σ^2.

Table 12-5 DaimlerChrysler Fuel Economy and Emissions

mfr	carline	car/truck	cid	rhp	trns	drv	od	etw	cmp	axle	n/v	a/c	hc	co	co2	mpg
20	300C/SRT-8	C	215	253	L5	4	2	4500	9.9	3.07	30.9	Y	0.011	0.09	288	30.8
20	CARAVAN 2WD	T	201	180	L4	F	2	4500	9.3	2.49	32.3	Y	0.014	0.11	274	32.5
20	CROSSFIRE ROADSTER	C	196	168	L5	R	2	3375	10	3.27	37.1	Y	0.001	0.02	250	35.4
20	DAKOTA PICKUP 2WD	T	226	210	L4	R	2	4500	9.2	3.55	29.6	Y	0.012	0.04	316	28.1
20	DAKOTA PICKUP 4WD	T	226	210	L4	4	2	5000	9.2	3.55	29.6	Y	0.011	0.05	365	24.4
20	DURANGO 2WD	T	348	345	L5	R	2	5250	8.6	3.55	27.2	Y	0.023	0.15	367	24.1
20	GRAND CHEROKEE 2WD	T	226	210	L4	R	2	4500	9.2	3.07	30.4	Y	0.006	0.09	312	28.5
20	GRAND CHEROKEE 4WD	T	348	230	L5	4	2	5000	9	3.07	24.7	Y	0.008	0.11	369	24.2
20	LIBERTY/CHEROKEE 2WD	T	148	150	M6	R	2	4000	9.5	4.1	41	Y	0.004	0.41	270	32.8
20	LIBERTY/CHEROKEE 4WD	T	226	210	L4	4	2	4250	9.2	3.73	31.2	Y	0.003	0.04	317	28
20	NEON/SRT-4/SX 2.0	C	122	132	L4	F	2	3000	9.8	2.69	39.2	Y	0.003	0.16	214	41.3
20	PACIFICA 2WD	T	215	249	L4	F	2	4750	9.9	2.95	35.3	Y	0.022	0.01	295	30
20	PACIFICA AWD	T	215	249	L4	4	2	5000	9.9	2.95	35.3	Y	0.024	0.05	314	28.2
20	PT CRUISER	T	148	220	L4	F	2	3625	9.5	2.69	37.3	Y	0.002	0.03	260	34.1
20	RAM 1500 PICKUP 2WD	T	500	500	M6	R	2	5250	9.6	4.1	22.3	Y	0.01	0.1	474	18.7
20	RAM 1500 PICKUP 4WD	T	348	345	L5	4	2	6000	8.6	3.92	29	Y	0	0	0	20.3
20	SEBRING 4-DR	C	165	200	L4	F	2	3625	9.7	2.69	36.8	Y	0.011	0.12	252	35.1
20	STRATUS 4-DR	C	148	167	L4	F	2	3500	9.5	2.69	36.8	Y	0.002	0.06	233	37.9
20	TOWN & COUNTRY 2WD	T	148	150	L4	F	2	4250	9.4	2.69	34.9	Y	0	0.09	262	33.8
20	VIPER CONVERTIBLE	C	500	501	M6	R	2	3750	9.6	3.07	19.4	Y	0.007	0.05	342	25.9
20	WRANGLER/TJ 4WD	T	148	150	M6	4	2	3625	9.5	3.73	40.1	Y	0.004	0.43	337	26.4

mfr-mfr code
carline-car line name (test vehicle model name)
car/truck-‘C’ for passenger vehicle and ‘T’ for truck
cid-cubic inch displacement of test vehicle
rhp-rated horsepower
trns-transmission code
drv-drive system code
od-overdrive code
etw-equivalent test weight

cmp-compression ratio
axle-axle ratio
n/v-n/v ratio (engine speed versus vehicle speed at 50 mph)
a/c-indicates air conditioning simulation
hc-HC(hydrocarbon emissions) Test level composite results
co-CO(carbon monoxide emissions) Test level composite results
co2-CO2(carbon dioxide emissions) Test level composite results
mpg-mpg(fuel economy, miles per gallon)

(c) Compute the standard errors of the regression coefficients. Are all of the model parameters estimated with the same precision? Why or why not?

(d) Predict power consumption for a month in which $x_1 = 75°\text{F}$, $x_2 = 24$ days, $x_3 = 90\%$, and $x_4 = 98$ tons.

12-7. Table 12-5 provides the highway gasoline mileage test results for 2005 model year vehicles from DaimlerChrysler. The full table of data (available on the book’s Web site) contains the same data for 2005 models from over 250 vehicles from many manufacturers (source: Environmental Protection Agency Web site www.epa.gov/ otaq/cert/mpg/testcars/database).

(a) Fit a multiple linear regression model to these data to estimate gasoline mileage that uses the following regressors: *cid, rhp, etw, cmp, axle, n/v.*

(b) Estimate σ^2 and the standard errors of the regression coefficients.

(c) Predict the gasoline mileage for the first vehicle in the table.

12-8. The pull strength of a wire bond is an important characteristic. The following table gives information on pull strength (y), die height (x_1), post height (x_2), loop height (x_3), wire length (x_4), bond width on the die (x_5), and bond width on the post (x_6).

(a) Fit a multiple linear regression model using x_2, x_3, x_4, and x_5 as the regressors.

(b) Estimate σ^2.

(c) Find the $se(\hat{\beta}_j)$. How precisely are the regression coefficients estimated, in your opinion?

y	x_1	x_2	x_3	x_4	x_5	x_6
8.0	5.2	19.6	29.6	94.9	2.1	2.3
8.3	5.2	19.8	32.4	89.7	2.1	1.8
8.5	5.8	19.6	31.0	96.2	2.0	2.0
8.8	6.4	19.4	32.4	95.6	2.2	2.1
9.0	5.8	18.6	28.6	86.5	2.0	1.8
9.3	5.2	18.8	30.6	84.5	2.1	2.1
9.3	5.6	20.4	32.4	88.8	2.2	1.9
9.5	6.0	19.0	32.6	85.7	2.1	1.9
9.8	5.2	20.8	32.2	93.6	2.3	2.1
10.0	5.8	19.9	31.8	86.0	2.1	1.8
10.3	6.4	18.0	32.6	87.1	2.0	1.6
10.5	6.0	20.6	33.4	93.1	2.1	2.1
10.8	6.2	20.2	31.8	83.4	2.2	2.1
11.0	6.2	20.2	32.4	94.5	2.1	1.9
11.3	6.2	19.2	31.4	83.4	1.9	1.8
11.5	5.6	17.0	33.2	85.2	2.1	2.1
11.8	6.0	19.8	35.4	84.1	2.0	1.8
12.3	5.8	18.8	34.0	86.9	2.1	1.8
12.5	5.6	18.6	34.2	83.0	1.9	2.0

(d) Use the model from part (a) to predict pull strength when $x_2 = 20$, $x_3 = 30$, $x_4 = 90$, and $x_5 = 2.0$.

12-9. An engineer at a semiconductor company wants to model the relationship between the device HFE (y) and three parameters: Emitter-RS (x_1), Base-RS (x_2), and Emitter-to-Base RS (x_3). The data are shown in the following table.

x_1 Emitter-RS	x_2 Base-RS	x_3 E-B-RS	y HFE-1M-5V
14.620	226.00	7.000	128.40
15.630	220.00	3.375	52.62
14.620	217.40	6.375	113.90
15.000	220.00	6.000	98.01
14.500	226.50	7.625	139.90
15.250	224.10	6.000	102.60
16.120	220.50	3.375	48.14
15.130	223.50	6.125	109.60
15.500	217.60	5.000	82.68
15.130	228.50	6.625	112.60
15.500	230.20	5.750	97.52
16.120	226.50	3.750	59.06
15.130	226.60	6.125	111.80
15.630	225.60	5.375	89.09
15.380	229.70	5.875	101.00
14.380	234.00	8.875	171.90
15.500	230.00	4.000	66.80
14.250	224.30	8.000	157.10
14.500	240.50	10.870	208.40
14.620	223.70	7.375	133.40

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

(b) Estimate σ^2.

(c) Find the standard errors $se(\hat{\beta}_j)$. Are all of the model parameters estimated with the same precision? Justify your answer.

(d) Predict HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$.

12-10. Heat treating is often used to carburize metal parts, such as gears. The thickness of the carburized layer is considered a crucial feature of the gear and contributes to the overall reliability of the part. Because of the critical nature of this feature, two different lab tests are performed on each furnace load. One test is run on a sample pin that accompanies each load. The other test is a destructive test, where an actual part is cross-sectioned. This test involves running a carbon analysis on the surface of both the gear pitch (top of the gear tooth) and the gear root (between the gear teeth). Table 12-6 shows the results of the pitch carbon analysis test for 32 parts.

The regressors are furnace temperature (TEMP), carbon concentration and duration of the carburizing cycle (SOAKPCT, SOAKTIME), and carbon concentration and duration of the diffuse cycle (DIFFPCT, DIFFTIME).

(a) Fit a linear regression model relating the results of the pitch carbon analysis test (PITCH) to the five regressor variables.

(b) Estimate σ^2.

(c) Find the standard errors $se(\hat{\beta}_j)$.

(d) Use the model in part (a) to predict PITCH when TEMP = 1650, SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.

12-11. An article in *Electronic Packaging and Production* (2002, Vol. 42) considered the effect of X-ray inspection of integrated circuits. The rads (radiation dose) were studied as a function of current (in milliamps) and exposure time (in minutes).

Rads	mAmps	Exposure Time
7.4	10	0.25
14.8	10	0.5
29.6	10	1
59.2	10	2
88.8	10	3
296	10	10
444	10	15
592	10	20
11.1	15	0.25
22.2	15	0.5
44.4	15	1

continued

Rads	mAmps	Exposure Time
88.8	15	2
133.2	15	3
444	15	10
666	15	15
888	15	20
14.8	20	0.25
29.6	20	0.5
59.2	20	1
118.4	20	2
177.6	20	3
592	20	10
888	20	15
1184	20	20
22.2	30	0.25
44.4	30	0.5
88.8	30	1
177.6	30	2
266.4	30	3
888	30	10
1332	30	15
1776	30	20
29.6	40	0.25
59.2	40	0.5
118.4	40	1
236.8	40	2
355.2	40	3
1184	40	10
1776	40	15
2368	40	20

(a) Fit a multiple linear regression model to these data with rads as the response.

(b) Estimate σ^2 and the standard errors of the regression coefficients.

(c) Use the model to predict rads when the current is 15 milliamps and the exposure time is 5 seconds.

12-12. An article in *Cancer Epidemiology, Biomarkers and Prevention* (1996, Vol. 5, pp. 849–852) conducted a pilot study to assess the use of toenail arsenic concentrations as an indicator of ingestion of arsenic-containing water. Twenty-one participants were interviewed regarding use of their private (unregulated) wells for drinking and cooking, and each provided a sample of water and toenail clippings. The table below showed the data of age (years), sex of person (1 = male, 2 = female), proportion of times household well used for drinking ($1 \leq 1/4$, $2 = 1/4$, $3 = 1/2$, $4 = 3/4$, $5 \geq 3/4$), proportion of times household well used for cooking ($1 \leq 1/4$, $2 = 1/4$, $3 = 1/2$, $4 = 3/4$, $5 \geq 3/4$), arsenic in water (ppm), and arsenic in toenails (ppm) respectively.

Age	Sex	Drink Use	Cook Use	Arsenic Water	Arsenic Nails
44	2	5	5	0.00087	0.119
45	2	4	5	0.00021	0.118
44	1	5	5	0	0.099
66	2	3	5	0.00115	0.118
37	1	2	5	0	0.277
45	2	5	5	0	0.358
47	1	5	5	0.00013	0.08
38	2	4	5	0.00069	0.158
41	2	3	2	0.00039	0.31
49	2	4	5	0	0.105
72	2	5	5	0	0.073
45	2	1	5	0.046	0.832
53	1	5	5	0.0194	0.517
86	2	5	5	0.137	2.252
8	2	5	5	0.0214	0.851
32	2	5	5	0.0175	0.269
44	1	5	5	0.0764	0.433
63	2	5	5	0	0.141
42	1	5	5	0.0165	0.275
62	1	5	5	0.00012	0.135
36	1	5	5	0.0041	0.175

(a) Fit a multiple linear regression model using arsenic concentration in nails as the response and age, drink use, cook use, and arsenic in the water as the regressors.

(b) Estimate σ^2 and the standard errors of the regression coefficients.

(c) Use the model to predict the arsenic in nails when the age is 30, the drink use is category 5, the cook use is category 5, and arsenic in the water is 0.135 ppm.

12-13. In an article in *IEEE Transactions on Instrumentation and Measurement* (2001, Vol. 50, pp. 2033–2040) powdered mixtures of coal and limestone were analyzed for permittivity. The errors in the density measurement was the response.

Density	Dielectric Constant	Loss Factor
0.749	2.05	0.016
0.798	2.15	0.02
0.849	2.25	0.022
0.877	2.3	0.023
0.929	2.4	0.026
0.963	2.47	0.028
0.997	2.54	0.031
1.046	2.64	0.034
1.133	2.85	0.039
1.17	2.94	0.042
1.215	3.05	0.045

Table 12-6

TEMP	SOAKTIME	SOAKPCT	DIFFTIME	DIFFPCT	PITCH
1650	0.58	1.10	0.25	0.90	0.013
1650	0.66	1.10	0.33	0.90	0.016
1650	0.66	1.10	0.33	0.90	0.015
1650	0.66	1.10	0.33	0.95	0.016
1600	0.66	1.15	0.33	1.00	0.015
1600	0.66	1.15	0.33	1.00	0.016
1650	1.00	1.10	0.50	0.80	0.014
1650	1.17	1.10	0.58	0.80	0.021
1650	1.17	1.10	0.58	0.80	0.018
1650	1.17	1.10	0.58	0.80	0.019
1650	1.17	1.10	0.58	0.90	0.021
1650	1.17	1.10	0.58	0.90	0.019
1650	1.17	1.15	0.58	0.90	0.021
1650	1.20	1.15	1.10	0.80	0.025
1650	2.00	1.15	1.00	0.80	0.025
1650	2.00	1.10	1.10	0.80	0.026
1650	2.20	1.10	1.10	0.80	0.024
1650	2.20	1.10	1.10	0.80	0.025
1650	2.20	1.15	1.10	0.80	0.024
1650	2.20	1.10	1.10	0.90	0.025
1650	2.20	1.10	1.10	0.90	0.027
1650	2.20	1.10	1.50	0.90	0.026
1650	3.00	1.15	1.50	0.80	0.029
1650	3.00	1.10	1.50	0.70	0.030
1650	3.00	1.10	1.50	0.75	0.028
1650	3.00	1.15	1.66	0.85	0.032
1650	3.33	1.10	1.50	0.80	0.033
1700	4.00	1.10	1.50	0.70	0.039
1650	4.00	1.10	1.50	0.70	0.040
1650	4.00	1.15	1.50	0.85	0.035
1700	12.50	1.00	1.50	0.70	0.056
1700	18.50	1.00	1.50	0.70	0.068

(a) Fit a multiple linear regression model to these data with the density as the response.

(b) Estimate σ^2 and the standard errors of the regression coefficients.

(c) Use the model to predict the density when the dielectric constant is 2.5 and the loss factor is 0.03.

12-14. An article in *Biotechnology Progress* (2001, Vol. 17, pp. 366–368) reported on an experiment to investigate and optimize nisin extraction in aqueous two-phase systems (ATPS). The nisin recovery was the dependent variable (y). The two regressor variables were concentration (%) of PEG 4000 (denoted as x_1) and concentration (%) of Na_2SO_4 (denoted as x_2).

(a) Fit a multiple linear regression model to these data.

(b) Estimate σ^2 and the standard errors of the regression coefficients.

x_1	x_2	y
13	11	62.8739
15	11	76.1328
13	13	87.4667
15	13	102.3236
14	12	76.1872
14	12	77.5287
14	12	76.7824
14	12	77.4381
14	12	78.7417

(c) Use the model to predict the nisin recovery when $x_1 = 14.5$ and $x_2 = 12.5$.

12-15. An article in *Optical Engineering* ["Operating Curve Extraction of a Correlator's Filter" (2004, Vol. 43, pp. 2775–2779)] reported on use of an optical correlator to perform an experiment by varying brightness and contrast. The resulting modulation is characterized by the useful range of gray levels. The data are shown below:

Brightness (%):	54	61	65	100	100	100	50	57	54
Contrast (%):	56	80	70	50	65	80	25	35	26
Useful range (ng):	96	50	50	112	96	80	155	144	255

(a) Fit a multiple linear regression model to these data.
(b) Estimate σ^2.
(c) Compute the standard errors of the regression coefficients.
(d) Predict the useful range when brightness = 80 and contrast = 75.

12-16. An article in *Technometrics* (1974, Vol. 16, pp. 523–531) considered the following stack-loss data from a plant oxidizing ammonia to nitric acid. Twenty-one daily responses of stack loss y (the amount of ammonia escaping) were measured with air flow x_1, temperature x_2, and acid concentration x_3.

Stack loss y = 42, 37, 37, 28, 18, 18, 19, 20, 15, 14, 14, 13, 11, 12, 8, 7, 8, 8, 9, 15, 15
x_1 = 80, 80, 75, 62, 62, 62, 62, 62, 58, 58, 58, 58, 58, 58, 50, 50, 50, 50, 50, 56, 70
x_2 = 27, 27, 25, 24, 22, 23, 24, 24, 23, 18, 18, 17, 18, 19, 18, 18, 19, 19, 20, 20, 20
x_3 = 89, 88, 90, 87, 87, 87, 93, 93, 87, 80, 89, 88, 82, 93, 89, 86, 72, 79, 80, 82, 91

(a) Fit a linear regression model relating the results of the stack loss to the three regressor varilables.
(b) Estimate σ^2.
(c) Find the standard error $se(\hat{\beta}_j)$.
(d) Use the model in part (a) to predict stack loss when $x_1 = 60$, $x_2 = 26$, and $x_3 = 85$.

12-17. Table 12-7 presents quarterback ratings for the 2008 National Football League season (source: *The Sports Network*).
(a) Fit a multiple regression model to relate the quarterback rating to the percentage of completions, the percentage of TDs, and the percentage of interceptions.
(b) Estimate σ^2.
(c) What are the standard errors of the regression coefficients?
(d) Use the model to predict the rating when the percentage of completions is 60%, the percentage of TDs is 4%, and the percentage of interceptions is 3%.

12-18. Table 12-8 presents statistics for the National Hockey League teams from the 2008–2009 season (source: *The Sports Network*). Fit a multiple linear regression model that relates *Wins* to the variables *GF* through *FG*. Because teams play 82 games $W = 82 - L - T - OTL$, but such a model does not help build a better team. Estimate σ^2 and find the standard errors of the regression coefficients for your model.

12-19. A study was performed on wear of a bearing y and its relationship to x_1 = oil viscosity and x_2 = load. The following data were obtained.

y	x_1	x_2
293	1.6	851
230	15.5	816
172	22.0	1058
91	43.0	1201
113	33.0	1357
125	40.0	1115

(a) Fit a multiple linear regression model to these data.
(b) Estimate σ^2 and the standard errors of the regression coefficients.
(c) Use the model to predict wear when $x_1 = 25$ and $x_2 = 1000$.
(d) Fit a multiple linear regression model with an interaction term to these data.
(e) Estimate σ^2 and $se(\hat{\beta}_j)$ for this new model. How did these quantities change? Does this tell you anything about the value of adding the interaction term to the model?
(f) Use the model in (d) to predict when $x_1 = 25$ and $x_2 = 1000$. Compare this prediction with the predicted value from part (c) above.

12-20. Consider the linear regression model

$$Y_i = \beta_0' + \beta_1(x_{i1} - \bar{x}_1) + \beta_2(x_{i2} - \bar{x}_2) + \epsilon_i$$

where $\bar{x}_1 = \sum x_{i1}/n$ and $\bar{x}_2 = \sum x_{i2}/n$.
(a) Write out the least squares normal equations for this model.
(b) Verify that the least squares estimate of the intercept in this model is $\hat{\beta}_0' = \sum y_i/n = \bar{y}$.
(c) Suppose that we use $y_i - \bar{y}$ as the response variable in the model above. What effect will this have on the least squares estimate of the intercept?

Table 12-7 Quarterback Ratings for the 2008 National Football League Season

Player		Team	Att	Comp	Pct Comp	Yds	Yds per Att	TD	Pct TD	Lng	Int	Pct Int	Rating Pts
Philip	Rivers	SD	478	312	65.3	4,009	8.39	34	7.1	67	11	2.3	105.5
Chad	Pennington	MIA	476	321	67.4	3,653	7.67	19	4.0	80	7	1.5	97.4
Kurt	Warner	ARI	598	401	67.1	4,583	7.66	30	5.0	79	14	2.3	96.9
Drew	Brees	NO	635	413	65	5,069	7.98	34	5.4	84	17	2.7	96.2
Peyton	Manning	IND	555	371	66.8	4,002	7.21	27	4.9	75	12	2.2	95
Aaron	Rodgers	GB	536	341	63.6	4,038	7.53	28	5.2	71	13	2.4	93.8
Matt	Schaub	HOU	380	251	66.1	3,043	8.01	15	3.9	65	10	2.6	92.7
Tony	Romo	DAL	450	276	61.3	3,448	7.66	26	5.8	75	14	3.1	91.4
Jeff	Garcia	TB	376	244	64.9	2,712	7.21	12	3.2	71	6	1.6	90.2
Matt	Cassel	NE	516	327	63.4	3,693	7.16	21	4.1	76	11	2.1	89.4
Matt	Ryan	ATL	434	265	61.1	3,440	7.93	16	3.7	70	11	2.5	87.7
Shaun	Hill	SF	288	181	62.8	2,046	7.10	13	4.5	48	8	2.8	87.5
Seneca	Wallace	SEA	242	141	58.3	1,532	6.33	11	4.5	90	3	1.2	87
Eli	Manning	NYG	479	289	60.3	3,238	6.76	21	4.4	48	10	2.1	86.4
Donovan	McNabb	PHI	571	345	60.4	3,916	6.86	23	4.0	90	11	1.9	86.4
Jay	Cutler	DEN	616	384	62.3	4,526	7.35	25	4.1	93	18	2.9	86
Trent	Edwards	BUF	374	245	65.5	2,699	7.22	11	2.9	65	10	2.7	85.4
Jake	Delhomme	CAR	414	246	59.4	3,288	7.94	15	3.6	65	12	2.9	84.7
Jason	Campbell	WAS	506	315	62.3	3,245	6.41	13	2.6	67	6	1.2	84.3
David	Garrard	JAC	535	335	62.6	3,620	6.77	15	2.8	41	13	2.4	81.7
Brett	Favre	NYJ	522	343	65.7	3,472	6.65	22	4.2	56	22	4.2	81
Joe	Flacco	BAL	428	257	60	2,971	6.94	14	3.3	70	12	2.8	80.3
Kerry	Collins	TEN	415	242	58.3	2,676	6.45	12	2.9	56	7	1.7	80.2
Ben	Roethlisberger	PIT	469	281	59.9	3,301	7.04	17	3.6	65	15	3.2	80.1
Kyle	Orton	CHI	465	272	58.5	2,972	6.39	18	3.9	65	12	2.6	79.6
JaMarcus	Russell	OAK	368	198	53.8	2,423	6.58	13	3.5	84	8	2.2	77.1
Tyler	Thigpen	KC	420	230	54.8	2,608	6.21	18	4.3	75	12	2.9	76
Gus	Frerotte	MIN	301	178	59.1	2,157	7.17	12	4.0	99	15	5.0	73.7
Dan	Orlovsky	DET	255	143	56.1	1,616	6.34	8	3.1	96	8	3.1	72.6
Marc	Bulger	STL	440	251	57	2,720	6.18	11	2.5	80	13	3.0	71.4
Ryan	Fitzpatrick	CIN	372	221	59.4	1,905	5.12	8	2.2	79	9	2.4	70
Derek	Anderson	CLE	283	142	50.2	1,615	5.71	9	3.2	70	8	2.8	66.5

Att	Attempts (number of pass attempts)
Comp	Completed passes
Pct Comp	Percentage of completed passes
Yds	Yards gained passing
Yds per Att	Yards gained per pass attempt
TD	Number of touchdown passes
Pct TD	Percentage of attempts that are touchdowns
Long	Longest pass completion
Int	Number of interceptions
Pct Int	Percentage of attempts that are interceptions
Rating Pts	Rating points

Table 12-8 Team Statistics for the 2008–2009 National Hockey League Season

Team	W	L	OTL	PTS	GF	GA	ADV	PPGF	PCTG	PEN	BMI	AVG	SHT	PPGA	PKPCT	SHGF	SHGA	FG
Anaheim	42	33	7	91	238	235	309	73	23.6	1418	8	17.4	385	78	79.7	6	6	43
Atlanta	35	41	6	76	250	279	357	69	19.3	1244	12	15.3	366	88	76	13	9	39
Boston	53	19	10	116	270	190	313	74	23.6	1016	12	12.5	306	54	82.4	8	7	47
Buffalo	41	32	9	91	242	229	358	75	21	1105	16	13.7	336	61	81.8	7	4	44
Carolina	45	30	7	97	236	221	374	70	18.7	786	16	9.8	301	59	80.4	8	7	39
Columbus	41	31	10	92	220	223	322	41	12.7	1207	20	15	346	62	82.1	8	9	41
Calgary	46	30	6	98	251	246	358	61	17	1281	18	15.8	349	58	83.4	6	13	37
Chicago	46	24	12	104	260	209	363	70	19.3	1129	28	14.1	330	64	80.6	10	5	43
Colorado	32	45	5	69	190	253	318	50	15.7	1044	18	13	318	64	79.9	4	5	31
Dallas	36	35	11	83	224	251	351	54	15.4	1134	10	14	327	70	78.6	2	2	38
Detroit	51	21	10	112	289	240	353	90	25.5	810	14	10	327	71	78.3	6	4	46
Edmonton	38	35	9	85	228	244	354	60	17	1227	20	15.2	338	76	77.5	3	8	39
Florida	41	30	11	93	231	223	308	51	16.6	884	16	11	311	54	82.6	7	6	39
Los Angeles	34	37	11	79	202	226	360	69	19.2	1191	16	14.7	362	62	82.9	4	7	39
Minnesota	40	33	9	89	214	197	328	66	20.1	869	20	10.8	291	36	87.6	9	6	39
Montreal	41	30	11	93	242	240	374	72	19.2	1223	6	15	370	65	82.4	10	10	38
New Jersey	51	27	4	106	238	207	307	58	18.9	1038	20	12.9	324	65	79.9	12	3	44
Nashville	40	34	8	88	207	228	318	50	15.7	982	12	12.1	338	59	82.5	9	8	41
NI Islanders	26	47	9	61	198	274	320	54	16.9	1198	18	14.8	361	73	79.8	12	5	37
NY Rangers	43	30	9	95	200	212	346	48	13.9	1175	24	14.6	329	40	87.8	9	13	42
Ottawa	36	35	11	83	213	231	339	66	19.5	1084	14	13.4	346	64	81.5	8	5	46
Philadelphia	44	27	11	99	260	232	316	71	22.5	1408	26	17.5	393	67	83	16	1	43
Phoenix	36	39	7	79	205	249	344	50	14.5	1074	18	13.3	293	68	76.8	5	4	36
Pittsburgh	45	28	9	99	258	233	360	62	17.2	1106	8	13.6	347	60	82.7	7	11	46
San Jose	53	18	11	117	251	199	360	87	24.2	1037	16	12.8	306	51	83.3	12	10	46
St. Louis	41	31	10	92	227	227	351	72	20.5	1226	22	15.2	357	58	83.8	10	8	35
Tampa Bay	24	40	18	66	207	269	343	61	17.8	1280	26	15.9	405	89	78	4	8	34
Toronto	34	35	13	81	244	286	330	62	18.8	1113	12	13.7	308	78	74.7	6	7	40
Vancouver	45	27	10	100	243	213	357	67	18.8	1323	28	16.5	371	69	81.4	7	5	47
Washington	50	24	8	108	268	240	337	85	25.2	1021	20	12.7	387	75	80.6	7	9	45

W	Wins
L	Losses during regular time
OTL	Overtime losses
PTS	Points. Two points for winning a game, one point for a tie or losing in overtime, zero points for losing in regular time.
GF	Goals for
GA	Goals against
ADV	Total advantages. Power play opportunities.
PPGF	Power-play goals for. Goals scored while on power play.
PCTG	Power play percentage. Power-play goals divided by total advantages.
PEN	Total penalty minutes including bench minutes
BMI	Total bench minor minutes
AVG	Average penalty minutes per game
SHT	Total times short-handed. Measures opponent opportunities.
PPGA	Power-play goals against
PKPCT	Penalty killing percentage. Measures a team's ability to prevent goals while its opponent is on a power play. Opponent opportunities minus power play goals divided by opponent's opportunities.
SHGF	Short-handed goals for
SHGA	Short-handed goals against
FG	Games scored first

12-2 HYPOTHESIS TESTS IN MULTIPLE LINEAR REGRESSION

In multiple linear regression problems, certain tests of hypotheses about the model parameters are useful in measuring model adequacy. In this section, we describe several important hypothesis-testing procedures. As in the simple linear regression case, hypothesis testing requires that the error terms ϵ_i in the regression model are normally and independently distributed with mean zero and variance σ^2.

12-2.1 Test for Significance of Regression

The test for significance of regression is a test to determine whether a linear relationship exists between the response variable y and a subset of the regressor variables $x_1, x_2, \ldots, x_k$. The appropriate hypotheses are

Hypotheses for ANOVA Test

$$H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$$
$$H_1\colon \beta_j \neq 0 \quad \text{for at least one } j \tag{12-18}$$

Rejection of $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$ implies that at least one of the regressor variables $x_1, x_2, \ldots, x_k$ contributes significantly to the model.

The test for significance of regression is a generalization of the procedure used in simple linear regression. The total sum of squares SS_T is partitioned into a sum of squares due to the model or to regression and a sum of squares due to error, say,

$$SS_T = SS_R + SS_E$$

Now if $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is true, SS_R/σ^2 is a chi-square random variable with k degrees of freedom. Note that the number of degrees of freedom for this chi-square random variable is equal to the number of regressor variables in the model. We can also show that the SS_E/σ^2 is a chi-square random variable with $n - p$ degrees of freedom, and that SS_E and SS_R are independent. The test statistic for $H_0\colon \beta_1 = \beta_2 = \cdots = \beta_k = 0$ is

Test Statistic for ANOVA

$$F_0 = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MS_R}{MS_E} \tag{12-19}$$

We should reject H_0 if the computed value of the test statistic in Equation 12-19, f_0, is greater than $f_{\alpha,k,n-p}$. The procedure is usually summarized in an analysis of variance table such as Table 12-9.

A computational formula for SS_R may be found easily. Now since $SS_T = \Sigma_{i=1}^n y_i^2 - (\Sigma_{i=1}^n y_i)^2/n = \mathbf{y}'\mathbf{y} - (\Sigma_{i=1}^n y_i)^2/n$, we may rewrite Equation 12-19 as

$$SS_E = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n} - \left[\hat{\boldsymbol\beta}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^n y_i\right)^2}{n}\right]$$

Table 12-9 Analysis of Variance for Testing Significance of Regression in Multiple Regression

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error or residual	SS_E	$n - p$	MS_E	
Total	SS_T	$n - 1$		

or

$$SS_E = SS_T - SS_R$$

Therefore, the regression sum of squares is

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} \tag{12-21}$$

EXAMPLE 12-3 Wire Bond Strength ANOVA

We will test for significance of regression (with $\alpha = 0.05$) using the wire bond pull strength data from Example 12-1. The total sum of squares is

$$SS_T = \mathbf{y}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 27{,}178.5316 - \frac{(725.82)^2}{25}$$
$$= 6105.9447$$

The regression or model sum of squares is computed from Equation 12-20 as follows:

$$SS_R = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 27{,}063.3581 - \frac{(725.82)^2}{25}$$
$$= 5990.7712$$

and by subtraction

$$SS_E = SS_T - SS_R = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} = 115.1716$$

The analysis of variance is shown in Table 12-10. To test $H_0\colon \beta_1 = \beta_2 = 0$, we calculate the statistic

$$f_0 = \frac{MS_R}{MS_E} = \frac{2995.3856}{5.2352} = 572.17$$

Since $f_0 > f_{0.05,2,22} = 3.44$ (or since the P-value is considerably smaller than $\alpha = 0.05$), we reject the null hypothesis and conclude that pull strength is linearly related to either wire length or die height, or both.

Practical Interpretation: Rejection of H_0 does not necessarily imply that the relationship found is an appropriate model for predicting pull strength as a function of wire length and die height. Further tests of model adequacy are required before we can be comfortable using this model in practice.

Most multiple regression computer programs provide the test for significance of regression in their output display. The middle portion of Table 12-4 is the Minitab output for this example. Compare Tables 12-4 and 12-10 and note their equivalence apart from rounding. The P-value is rounded to zero in the computer output.

Table 12-10 Test for Significance of Regression for Example 12-3

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	5990.7712	2	2995.3856	572.17	1.08E-19
Error or residual	115.1735	22	5.2352		
Total	6105.9447	24			

R^2 and Adjusted R^2

We may also use the **coefficient of multiple determination** R^2 as a global statistic to assess the fit of the model. Computationally,

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T} \qquad (12\text{-}22)$$

For the wire bond pull strength data, we find that $R^2 = SS_R/SS_T = 5990.7712/6105.9447 = 0.9811$. Thus the model accounts for about 98% of the variability in the pull strength response (refer to the Minitab output in Table 12-4). The R^2 statistic is somewhat problematic as a measure of the quality of the fit for a multiple regression model because it never decreases when a variable is added to a model.

To illustrate, consider the model fit to the wire bond pull strength data in Example 11-8. This was a simple linear regression model with x_1 = wire length as the regressor. The value of R^2 for this model is $R^2 = 0.9640$. Therefore, adding x_2 = die height to the model increases R^2 by $0.9811 - 0.9640 = 0.0171$, a very small amount. Since R^2 can never decrease when a regressor is added, it can be difficult to judge whether the increase is telling us anything useful about the new regressor. It is particularly hard to interpret a small increase, such as observed in the pull strength data.

Many regression users prefer to use an **adjusted** R^2 statistic:

Adjusted R^2

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)} \qquad (12\text{-}23)$$

Because $SS_E/(n-p)$ is the error or residual mean square and $SS_T/(n-1)$ is a constant, R^2_{adj} will only increase when a variable is added to the model if the new variable reduces the error mean square. Note that for the multiple regression model for the pull strength data $R^2_{\text{adj}} = 0.979$ (see the Minitab output in Table 12-4), whereas in Example 11-8 the adjusted R^2 for the one-variable model is $R^2_{\text{adj}} = 0.962$. Therefore, we would conclude that adding x_2 = die height to the model does result in a meaningful reduction in unexplained variability in the response.

The **adjusted** R^2 statistic essentially penalizes the analyst for adding terms to the model. It is an easy way to guard against **overfitting,** that is, including regressors that are not really useful. Consequently, it is very useful in comparing and evaluating competing regression models. We will use R^2_{adj} for this when we discuss **variable selection** in regression in Section 12-6.3.

12-2.2 Tests on Individual Regression Coefficients and Subsets of Coefficients

We are frequently interested in testing hypotheses on the individual regression coefficients. Such tests would be useful in determining the potential value of each of the regressor variables in the regression model. For example, the model might be more effective with the inclusion of additional variables or perhaps with the deletion of one or more of the regressors presently in the model.

The hypothesis to test if an individual regression coefficient, say β_j equals a value β_{j0} is

$$H_0\colon \beta_j = \beta_{j0}$$
$$H_1\colon \beta_j \neq \beta_{j0} \qquad (12\text{-}24)$$

The test statistic for this hypothesis is

$$T_0 = \frac{\hat{\beta}_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\beta}_j - \beta_{j0}}{se(\hat{\beta}_j)} \qquad (12\text{-}25)$$

where C_{jj} is the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to $\hat{\beta}_j$. Notice that the denominator of Equation 12-24 is the standard error of the regression coefficient $\hat{\beta}_j$. The null hypothesis H_0: $\beta_j = \beta_{j0}$ is rejected if $|t_0| > t_{\alpha/2,n-p}$. This is called a **partial** or **marginal test** because the regression coefficient $\hat{\beta}_j$ depends on all the other regressor variables $x_i (i \neq j)$ that are in the model. More will be said about this in the following example.

An important special case of the previous hypothesis occurs for $\beta_{j0} = 0$. If H_0: $\beta_j = 0$ is not rejected, this indicates that the regressor x_j can be deleted from the model. Adding a variable to a regression model always causes the sum of squares for regression to increase and the error sum of squares to decrease (this is why R^2 always increases when a variable is added). We must decide whether the increase in the regression sum of squares is large enough to justify using the additional variable in the model. Furthermore, adding an unimportant variable to the model can actually increase the error mean square, indicating that adding such a variable has actually made the model a poorer fit to the data (this is why R^2_{adj} is a better measure of global model fit then the ordinary R^2).

EXAMPLE 12-4 Wire Bond Strength Coefficient Test

Consider the wire bond pull strength data, and suppose that we want to test the hypothesis that the regression coefficient for x_2 (die height) is zero. The hypotheses are

$$H_0\colon \beta_2 = 0$$
$$H_1\colon \beta_2 \neq 0$$

The main diagonal element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix corresponding to $\hat{\beta}_2$ is $C_{22} = 0.0000015$, so the t-statistic in Equation 12-25 is

$$t_0 = \frac{\hat{\beta}_2}{\sqrt{\hat{\sigma}^2 C_{22}}} = \frac{0.01253}{\sqrt{(5.2352)(0.0000015)}} = 4.477$$

Note that we have used the estimate of σ^2 reported to four decimal places in Table 12-10. Since $t_{0.025,22} = 2.074$, we reject H_0: $\beta_2 = 0$ and conclude that the variable x_2 (die height) contributes significantly to the model. We could also have used a P-value to draw conclusions. The P-value for $t_0 = 4.477$ is $P = 0.0002$, so with $\alpha = 0.05$ we would reject the null hypothesis.

Practical Interpretation: Note that this test measures the marginal or partial contribution of x_2 given that x_1 is in the model. That is, the t-test measures the contribution of adding the variable x_2 = die height to a model that already contains x_1 = wire length. Table 12-4 shows the value of the t-test computed by Minitab. The Minitab t-test statistic is reported to two decimal places. Note that the computer produces a t-test for each regression coefficient in the model. These t-tests indicate that both regressors contribute to the model.

EXAMPLE 12-5 Wire Bond Strength One-Sided Coefficient Test

There is an interest in the effect of die height on strength. This can be evaluated by the magnitude of the coefficient for die height. To conclude that the coefficient for die height exceeds 0.01 the hypotheses become

$$H_0\colon \beta_2 = 0.01 \qquad H_1\colon \beta_2 > 0.01$$

For such a test, computer software can complete much of the hard work. We only need to assemble the pieces. From the Minitab output in Table 12-4, $\hat{\beta}_2 = 0.012528$ and the standard error of $\hat{\beta}_2 = 0.002798$. Therefore the t-statistic is

$$t_0 = \frac{0.012528 - 0.01}{0.002798} = 0.9035$$

with 22 degrees of freedom (error degrees of freedom). From Table IV in Appendix A, $t_{0.25,\,22} = 0.686$ and $t_{0.1,\,22} = 1.321$. Therefore, the P-value can be bounded as $0.1 < P$-value < 0.25. One cannot conclude that the coefficient exceeds 0.01 at common levels of significance.

There is another way to test the contribution of an individual regressor variable to the model. This approach determines the increase in the regression sum of squares obtained by adding a variable x_j (say) to the model, given that other variables $x_i (i \neq j)$ are already included in the regression equation.

The procedure used to do this is called the **general regression significance test**, or the **extra sum of squares method.** This procedure can also be used to investigate the contribution of a *subset* of the regressor variables to the model. Consider the regression model with k regressor variables

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{12-26}$$

where $\mathbf{y}$ is $(n \times 1)$, $\mathbf{X}$ is $(n \times p)$, $\boldsymbol{\beta}$ is $(p \times 1)$, $\boldsymbol{\epsilon}$ is $(n \times 1)$, and $p = k + 1$. We would like to determine if the subset of regressor variables $x_1, x_2, \ldots, x_r\ (r < k)$ as a whole contributes significantly to the regression model. Let the vector of regression coefficients be partitioned as follows:

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} \tag{12-27}$$

where $\boldsymbol{\beta}_1$ is $(r \times 1)$ and $\boldsymbol{\beta}_2$ is $[(p - r) \times 1]$. We wish to test the hypotheses

Hypotheses for General Regression Test

$$H_0\colon \boldsymbol{\beta}_1 = \mathbf{0}$$
$$H_1\colon \boldsymbol{\beta}_1 \neq \mathbf{0} \tag{12-28}$$

where $\mathbf{0}$ denotes a vector of zeroes. The model may be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \tag{12-29}$$

where $\mathbf{X}_1$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_1$ and $\mathbf{X}_2$ represents the columns of $\mathbf{X}$ associated with $\boldsymbol{\beta}_2$.

For the **full model** (including both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$), we know that $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\mathbf{y}.$ In addition, the regression sum of squares for all variables including the intercept is

$$SS_R(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y} \qquad (p = k + 1 \text{ degrees of freedom})$$

and

$$MS_E = \frac{\mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}\mathbf{X}'\mathbf{y}}{n - p}$$

$SS_R(\boldsymbol{\beta})$ is called the regression sum of squares due to $\boldsymbol{\beta}$. To find the contribution of the terms in $\boldsymbol{\beta}_1$ to the regression, fit the model assuming the null hypothesis H_0: $\boldsymbol{\beta}_1 = \mathbf{0}$ to be true. The **reduced model** is found from Equation 12-29 as

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon} \qquad (12\text{-}30)$$

The least squares estimate of $\boldsymbol{\beta}_2$ is $\hat{\boldsymbol{\beta}}_2 = (\mathbf{X}_2'\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{y}$, and

$$SS_R(\boldsymbol{\beta}_2) = \hat{\boldsymbol{\beta}}_2'\mathbf{X}_2'\mathbf{y} \qquad (p - r \text{ degrees of freedom}) \qquad (12\text{-}31)$$

The regression sum of squares due to $\boldsymbol{\beta}_1$ given that $\boldsymbol{\beta}_2$ is already in the model is

$$SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = SS_R(\boldsymbol{\beta}) - SS_R(\boldsymbol{\beta}_2) \qquad (12\text{-}32)$$

This sum of squares has r degrees of freedom. It is sometimes called the extra sum of squares due to $\boldsymbol{\beta}_1$. Note that $SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is the increase in the regression sum of squares due to including the variables $x_1, x_2, \ldots, x_r$ in the model. Now $SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)$ is independent of MS_E, and the null hypothesis $\boldsymbol{\beta}_1 = \mathbf{0}$ may be tested by the statistic.

***F* Statistic for General Regression Test**

$$F_0 = \frac{SS_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2)/r}{MS_E} \qquad (12\text{-}33)$$

If the computed value of the test statistic $f_0 > f_{\alpha,r,n-p}$, we reject H_0, concluding that at least one of the parameters in $\boldsymbol{\beta}_1$ is not zero and, consequently, at least one of the variables $x_1, x_2, \ldots, x_r$ in $\mathbf{X}_1$ contributes significantly to the regression model. Some authors call the test in Equation 12-33 a **partial *F*-test.**

The partial F-test is very useful. We can use it to measure the contribution of each individual regressor x_j as if it were the last variable added to the model by computing

$$SS_R(\beta_j|\beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_k), \qquad j = 1, 2, \ldots, k$$

This is the increase in the regression sum of squares due to adding x_j to a model that already includes $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_k$. The partial F-test is a more general procedure in that we can measure the effect of sets of variables. In Section 12-6.3 we show how the partial F-test plays a major role in *model building*—that is, in searching for the best set of regressor variables to use in the model.

EXAMPLE 12-6 Wire Bond Strength General Regression Test

Consider the wire bond pull-strength data in Example 12-1. We will investigate the contribution of two new variables, x_3 and x_4, to the model using the partial F-test approach. The new variables are explained at the end of this example. That is, we wish to test

$$H_0: \beta_3 = \beta_4 = 0 \qquad H_1: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0$$

To test this hypothesis, we need the extra sum of squares due to β_3 and β_4 or

$$\begin{aligned} SS_R(\beta_4, \beta_3|\beta_2, \beta_1, \beta_0) &= SS_R(\beta_4, \beta_3, \beta_2, \beta_1, \beta_0) - SS_R(\beta_2, \beta_1, \beta_0) \\ &= SS_R(\beta_4, \beta_3, \beta_2, \beta_1|\beta_0) - SS_R(\beta_2, \beta_1|\beta_0) \end{aligned}$$

In Example 12-3 we calculated

$$SS_R(\beta_2, \beta_1|\beta_0) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{y} - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n} = 5990.7712 \text{ (two degrees of freedom)}$$

Also, Table 12-4 shows the Minitab output for the model with only x_1 and x_2 as predictors. In the analysis of variance table, we can see that $SS_R = 5990.8$ and this agrees with our calculation. In practice, the computer output would be used to obtain this sum of squares.

If we fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$, we can use the same matrix formula. Alternatively, we can look at SS_R from computer output for this model. The analysis of variance table for this model is shown in Table 12-11 and we see that

$$SS_R(\beta_4, \beta_3, \beta_2, \beta_1|\beta_0) = 6024.0 \text{ (four degrees of freedom)}$$

Therefore,

$$SS_R(\beta_4, \beta_3|\beta_2, \beta_1, \beta_0) = 6024.0 - 5990.8 = 33.2 \text{ (two degrees of freedom)}$$

This is the increase in the regression sum of squares due to adding x_3 and x_4 to a model already containing x_1 and x_2. To test H_0, calculate the test statistic

$$f_0 = \frac{SS_R(\beta_4, \beta_3|\beta_2, \beta_1, \beta_0)/2}{MS_E} = \frac{33.2/2}{4.1} = 4.05$$

Note that MS_E from the full model using x_1, x_2, x_3 and x_4 is used in the denominator of the test statistic. Because $f_{0.05,2,20} = 3.49$, we reject H_0 and conclude that at least one of the new variables contributes significantly to the model. Further analysis and tests will be needed to refine the model and determine if one or both of x_3 and x_4 are important.

The mystery of the new variables can now be explained. These are quadratic powers of the original predictors wire length and wire height. That is, $x_3 = x_1^2$ and $x_4 = x_2^2$. A test for quadratic terms is a common use of partial F-tests. With this information and the original data for x_1 and x_2, you can use computer software to reproduce these calculations. Multiple regression allows models to be extended in such a simple manner that the real meaning of x_3 and x_4 did not even enter into the test procedure. Polynomial models such as this are discussed further in Section 12-6.

If a partial F-test is applied to a single variable, it is equivalent to a t-test. To see this, consider the Minitab regression output for the wire bond pull strength in Table 12-4. Just below the analysis of variance summary in this table, the quantity labeled ” ‘SeqSS”’ shows the sum

Table 12-11 Regression Analysis: y versus x1, x2, x3, x4

The regression equation is y = 5.00 + 1.90 x1 + 0.0151 x2 + 0.0460 x3 − 0.000008 x4

Predictor	Coef	SE Coef	T	P
Constant	4.996	1.655	3.02	0.007
x1	1.9049	0.3126	6.09	0.000
x2	0.01513	0.01051	1.44	0.165
x3	0.04595	0.01666	2.76	0.012
x4	−0.00000766	0.00001641	−0.47	0.646

S = 2.02474 R−Sq = 98.7% R−Sq (adj) = 98.4%

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	4	6024.0	1506.0	367.35	0.000
Residual Error	20	82.0	4.1		
Total	24	6105.9			

Source	DF	Seq SS
x1	1	5885.9
x2	1	104.9
x3	1	32.3
x4	1	0.9

of squares obtained by fitting x_1 alone (5885.9) and the sum of squares obtained by fitting x_2 after x_1 (104.9). In out notation, these are referred to as $SS_R(\beta_1|\beta_0)$ and $SS_R(\beta_2, \beta_1|\beta_0)$, respectively. Therefore, to test $H_0: \beta_2 = 0$, $H_1: \beta_2 \neq 0$ the partial F-test is

$$f_0 = \frac{SS_R(\beta_2|\beta_1, \beta_0)/1}{MS_E} = \frac{104.92}{5.24} = 20.2$$

where MS_E is the mean square for residual in the computer output in Table 12.4. This statistic should be compared to an F-distribution with 1 and 22 degrees of freedom in the numerator and denominator, respectively. From Table 12-4, the t-test for the same hypothesis is $t_0 = 4.48$. Note that $t_0^2 = 4.48^2 = 20.07 = f_0$, except for round-off error. Furthermore, the square of a t-random variable with ν degrees of freedom is an F-random variable with one and ν degrees of freedom. Consequently, the t-test provides an equivalent method to test a single variable for contribution to a model. Because the t-test is typically provided by computer output, it is the preferred method to test a single variable.

EXERCISES FOR SECTION 12-2

12-21. Consider the computer output below.

```
The regression equation is
Y = 254 + 2.77 x1 − 3.58 x2

Predictor        Coef     SE Coef        T      P
Constant      253.810       4.781        ?      ?
x1             2.7738      0.1846    15.02      ?
x2            −3.5753      0.1526        ?      ?

S = 5.05756    R-Sq = ?      R-Sq (adj) = 98.4%

Analysis of Variance

Source            DF       SS        MS      F     P
Regression         2    22784     11392      ?     ?
Residual Error     ?        ?         ?
Total             14    23091
```

(a) Fill in the missing quantities. You may use bounds for the P-values.

(b) What conclusions can you draw about the significance of regression?

(c) What conclusions can you draw about the contributions of the individual regressors to the model?

12-22. You have fit a regression model with two regressors to a data set that has 20 observations. The total sum of squares is 1000 and the model sum of squares is 750.

(a) What is the value of R^2 for this model?

(b) What is the adjusted R^2 for this model?

(c) What is the value of the F-statistic for testing the significance of regression? What conclusions would you draw about this model if $\alpha = 0.05$? What if $\alpha = 0.01$?

(d) Suppose that you add a third regressor to the model and as a result the model sum of squares is now 785. Does it seem to you that adding this factor has improved the model?

12-23. Consider the regression model fit to the soil shear strength data in Exercise 12-1.

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

(b) Construct the t-test on each regression coefficient. What are your conclusions, using $\alpha = 0.05$? Calculate P-values.

12-24. Consider the absorption index data in Exercise 12-2. The total sum of squares for y is $SS_T = 742.00$.

(a) Test for significance of regression using $\alpha = 0.01$. What is the P-value for this test?

(b) Test the hypothesis H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$ using $\alpha = 0.01$. What is the P-value for this test?

(c) What conclusion can you draw about the usefulness of x_1 as a regressor in this model?

12-25. A regression model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ has been fit to a sample of $n = 25$ observations. The calculated t-ratios $\hat{\beta}_j / se(\hat{\beta}_j)$, $j = 1, 2, 3$ are as follows: for β_1, $t_0 = 4.82$, for β_2, $t_0 = 8.21$ and for β_3, $t_0 = 0.98$.

(a) Find P-values for each of the t-statistics.

(b) Using $\alpha = 0.05$, what conclusions can you draw about the regressor x_3? Does it seem likely that this regressor contributes significantly to the model?

12-26. Consider the electric power consumption data in Exercise 12-6.

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

(b) Use the t-test to assess the contribution of each regressor to the model. Using $\alpha = 0.05$, what conclusions can you draw?

12-27. Consider the gasoline mileage data in Exercise 12-7.

(a) Test for significance of regression using $\alpha = 0.05$. What conclusions can you draw?

(b) Find the t-test statistic for each regressor. Using $\alpha = 0.05$, what conclusions can you draw? Does each regressor contribute to the model?

12-28. Consider the wire bond pull strength data in Exercise 12-8.

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

(b) Calculate the *t*-test statistic for each regression coefficient. Using $\alpha = 0.05$, what conclusions can you draw? Do all variables contribute to the model?

12-29. Reconsider the semiconductor data in Exercise 12-9.

(a) Test for significance of regression using $\alpha = 0.05$. What conclusions can you draw?

(b) Calculate the *t*-test statistic and *P*-value for each regression coefficient. Using $\alpha = 0.05$, what conclusions can you draw?

12-30. Consider the regression model fit to the arsenic data in Exercise 12-12. Use arsenic in nails as the response and age, drink use, and cook use as the regressors.

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

(b) Construct a *t*-test on each regression coefficient. What conclusions can you draw about the variables in this model? Use $\alpha = 0.05$.

12-31. Consider the regression model fit to the X-ray inspection data in Exercise 12-11. Use rads as the response.

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

(b) Construct a *t*-test on each regression coefficient. What conclusions can you draw about the variables in this model? Use $\alpha = 0.05$.

12-32. Consider the regression model fit to the nisin extraction data in Exercise 12-14. Use nisin extraction as the response.

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

(b) Construct a *t*-test on each regression coefficient. What conclusions can you draw about the variables in this model? Use $\alpha = 0.05$.

(c) Comment on the effect of a small sample size to the tests in the previous parts.

12-33. Consider the regression model fit to the grey range modulation data in Exercise 12-15. Use the useful range as the response.

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

(b) Construct a *t*-test on each regression coefficient. What conclusions can you draw about the variables in this model? Use $\alpha = 0.05$.

12-34. Consider the regression model fit to the stack loss data in Exercise 12-16. Use stack loss as the response.

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

(b) Construct a *t*-test on each regression coefficient. What conclusions can you draw about the variables in this model? Use $\alpha = 0.05$.

12-35. Consider the NFL data in Exercise 12-17.

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

(b) Conduct the *t*-test for each regression coefficient. Using $\alpha = 0.05$, what conclusions can you draw about the variables in this model?

(c) Find the amount by which the regressor x_2 (TD percentage) increases the regression sum of squares, and conduct an *F*-test for H_0: $\beta_2 = 0$ versus H_1: $\beta_2 \neq 0$ using $\alpha = 0.05$. What is the *P*-value for this test? What conclusions can you draw?

12-36. Exercise 12-10 presents data on heat treating gears.

(a) Test the regression model for significance of regression. Using $\alpha = 0.05$, find the *P*-value for the test and draw conclusions.

(b) Evaluate the contribution of each regressor to the model using the *t*-test with $\alpha = 0.05$.

(c) Fit a new model to the response PITCH using new regressors x_1 = SOAKTIME × SOAKPCT and x_2 = DIFFTIME × DIFFPCT.

(d) Test the model in part (c) for significance of regression using $\alpha = 0.05$. Also calculate the *t*-test for each regressor and draw conclusions.

(e) Estimate σ^2 for the model from part (c) and compare this to the estimate of σ^2 for the model in part (a). Which estimate is smaller? Does this offer any insight regarding which model might be preferable?

12-37. Consider the bearing wear data in Exercise 12-19.

(a) For the model with no interaction, test for significance of regression using $\alpha = 0.05$. What is the *P*-value for this test? What are your conclusions?

(b) For the model with no interaction, compute the *t*-statistics for each regression coefficient. Using $\alpha = 0.05$, what conclusions can you draw?

(c) For the model with no interaction, use the extra sum of squares method to investigate the usefulness of adding x_2 = load to a model that already contains x_1 = oil viscosity. Use $\alpha = 0.05$.

(d) Refit the model with an interaction term. Test for significance of regression using $\alpha = 0.05$.

(e) Use the extra sum of squares method to determine whether the interaction term contributes significantly to the model. Use $\alpha = 0.05$.

(f) Estimate σ^2 for the interaction model. Compare this to the estimate of σ^2 from the model in part (a).

12-38. Data on National Hockey League team performance was presented in Exercise 12-18.

(a) Test the model from this exercise for significance of regression using $\alpha = 0.05$. What conclusions can you draw?

(b) Use the *t*-test to evaluate the contribution of each regressor to the model. Does it seem that all regressors are necessary? Use $\alpha = 0.05$.

(c) Fit a regression model relating the number of games won to the number of goals for and the number of power play goals for. Does this seem to be a logical choice of regressors, considering your answer to part (b)? Test this new model for significance of regression and evaluate the contribution of each regressor to the model using the *t*-test. Use $\alpha = 0.05$.

12-39. Data from a hospital patient satisfaction survey were presented in Exercise 12-5.

(a) Test the model from this exercise for significance of regression. What conclusions can you draw if $\alpha = 0.05$? What if $\alpha = 0.01$?

(b) Test the contribution of the individual regressors using the *t*-test. Does it seem that all regressors used in the model are really necessary?

12-40. Data from a hospital patient satisfaction survey were presented in Exercise 12-5.

(a) Fit a regression model using only the patient age and severity regressors. Test the model from this exercise for significance of regression. What conclusions can you draw if $\alpha = 0.05$? What if $\alpha = 0.01$?

(b) Test the contribution of the individual regressors using the *t*-test. Does it seem that all regressors used in the model are really necessary?

(c) Find an estimate of the error variance σ^2. Compare this estimate of σ^2 with the estimate obtained from the model containing the third regressor, anxiety. Which estimate is smaller? Does this tell you anything about which model might be preferred?

12-3 CONFIDENCE INTERVALS IN MULTIPLE LINEAR REGRESSION

12-3.1 Confidence Intervals on Individual Regression Coefficients

In multiple regression models, it is often useful to construct confidence interval estimates for the regression coefficients $\{\beta_j\}$. The development of a procedure for obtaining these confidence intervals requires that the errors $\{\epsilon_i\}$ are normally and independently distributed with mean zero and variance σ^2. This is the same assumption required in hypothesis testing. Therefore, the observations $\{Y_i\}$ are normally and independently distributed with mean $\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}$ and variance σ^2. Since the least squares estimator $\hat{\boldsymbol{\beta}}$ is a linear combination of the observations, it follows that $\hat{\boldsymbol{\beta}}$ is normally distributed with mean vector $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Then each of the statistics

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2 C_{jj}}} \qquad j = 0, 1, \ldots, k \tag{12-34}$$

has a *t* distribution with $n - p$ degrees of freedom, where C_{jj} is the *jj*th element of the $(\mathbf{X}'\mathbf{X})^{-1}$ matrix, and $\hat{\sigma}^2$ is the estimate of the error variance, obtained from Equation 12-16. This leads to the following $100(1 - \alpha)\%$ confidence interval for the regression coefficient $\beta_j, j = 0, 1, \ldots, k$.

Confidence Interval on a Regression Coefficient

A $100(1 - \alpha)$ % **confidence interval on the regression coefficient** β_j, $j = 0, 1, \ldots,$ k in the multiple linear regression model is given by

$$\hat{\beta}_j - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{jj}} \tag{12-35}$$

Because $\sqrt{\hat{\sigma}^2 C_{jj}}$ is the standard error of the regression coefficient $\hat{\beta}_j$, we would also write the CI formula as $\hat{\beta}_j - t_{\alpha/2,n-p}\, se(\hat{\beta}_j) \le \beta_j \le \hat{\beta}_j + t_{\alpha/2,n-p}\, se(\hat{\beta}_j)$.

EXAMPLE 12-7 Wire Bond Strength Confidence Interval

We will construct a 95% confidence interval on the parameter β_1 in the wire bond pull strength problem. The point estimate of β_1 is $\hat{\beta}_1 = 2.74427$ and the diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$ corresponding to β_1 is $C_{11} = 0.001671$. The estimate of σ^2 is $\hat{\sigma}^2 = 5.2352$, and $t_{0.025,22} = 2.074$. Therefore, the 95% CI on β_1 is computed from Equation 12-35 as

$$2.74427 - (2.074)\sqrt{(5.2352)(.001671)} \le \beta_1 \le 2.74427 + (2.074)\sqrt{(5.2352)(.001671)}$$

which reduces to

$$2.55029 \le \beta_1 \le 2.93825$$

Also, computer software such as Minitab can be used to help calculate this confidence interval. From the regression output in Table 10-4, $\hat{\beta}_1 = 2.74427$ and the standard error of $\hat{\beta}_1 = 0.0935$. This standard error is the multiplier of the t-table constant in the confidence interval. That is, $0.0935 = \sqrt{(5.2352)(0.001671)}$. Consequently, all the numbers are available from the computer output to construct the interval and this is the typical method used in practice.

12-3.2 Confidence Interval on the Mean Response

We may also obtain a confidence interval on the mean response at a particular point, say, $x_{01}, x_{02}, \ldots, x_{0k}$. To estimate the mean response at this point, define the vector

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

The mean response at this point is $E(Y|\mathbf{x}_0) = \mu_{Y|\mathbf{x}_0} = \mathbf{x}_0'\boldsymbol{\beta}$, which is estimated by

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0'\hat{\boldsymbol{\beta}} \tag{12-36}$$

This estimator is unbiased, since $E(\mathbf{x}_0'\hat{\boldsymbol{\beta}}) = \mathbf{x}_0'\boldsymbol{\beta} = E(Y|\mathbf{x}_0) = \mu_{\mathrm{Y}|\mathbf{x}_0}$ and the variance of $\hat{\mu}_{Y|\mathbf{x}_0}$ is

$$V(\hat{\mu}_{Y|\mathbf{x}_0}) = \sigma^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 \tag{12-37}$$

A $100(1-\alpha)$ % CI on $\mu_{Y|\mathbf{x}_0}$ can be constructed from the statistic

$$\frac{\hat{\mu}_{Y|\mathbf{x}_0} - \mu_{Y|\mathbf{x}_0}}{\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}} \tag{12-38}$$

Confidence Interval on the Mean Response

For the multiple linear regression model, a $100(1-\alpha)$% **confidence interval on the mean response** at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\hat{\mu}_{Y|\mathbf{x}_0} - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \le \mu_{Y|\mathbf{x}_0} \le \hat{\mu}_{Y|\mathbf{x}_0} + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0} \tag{12-39}$$

Equation 12-39 is a CI about the regression plane (or hyperplane). It is the multiple regression generalization of Equation 11-32.

EXAMPLE 12-8 Wire Bond Strength Confidence Interval on the Mean Response

The engineer in Example 12-1 would like to construct a 95% CI on the mean pull strength for a wire bond with wire length $x_1 = 8$ and die height $x_2 = 275$. Therefore,

$$\mathbf{x}_0 = \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$

The estimated mean response at this point is found from Equation 12-36 as

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0'\hat{\beta} = \begin{bmatrix} 1 & 8 & 275 \end{bmatrix} \begin{bmatrix} 2.26379 \\ 2.74427 \\ 0.01253 \end{bmatrix} = 27.66$$

The variance of $\hat{\mu}_{Y|\mathbf{x}_0}$ is estimated by

$$\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 5.2352\,[1\ 8\ 275]$$
$$\times \begin{bmatrix} .214653 & -.007491 & -.000340 \\ -.007491 & .001671 & -.000019 \\ -.000340 & -.000019 & .0000015 \end{bmatrix} \begin{bmatrix} 1 \\ 8 \\ 275 \end{bmatrix}$$
$$= 5.2352\,(0.0444) = 0.23244$$

Therefore, a 95% CI on the mean pull strength at this point is found from Equation 12-39 as

$$27.66 - 2.074\sqrt{0.23244} \le \mu_{Y|\mathbf{x}_0} \le 27.66 + 2.074\sqrt{0.23244}$$

which reduces to

$$26.66 \le \mu_{Y|\mathbf{x}_0} \le 28.66$$

Some computer software packages will provide estimates of the mean for a point of interest $\mathbf{x}_0$ and the associated CI. Table 12-4 shows the Minitab output for Example 12-8. Both the estimate of the mean and the 95% CI are provided.

12-4 PREDICTION OF NEW OBSERVATIONS

A regression model can be used to predict new or **future observations** on the response variable Y corresponding to particular values of the independent variables, say, $x_{01}, x_{02}, \ldots, x_{0k}$. If $\mathbf{x}_0' = [1, x_{01}, x_{02}, \ldots, x_{0k}]$, a point estimate of the future observation Y_0 at the point $x_{01}, x_{02}, \ldots, x_{0k}$ is

$$\hat{y}_0 = \mathbf{x}_0'\hat{\boldsymbol{\beta}} \qquad (12\text{-}40)$$

Prediction Interval

A $100(1-\alpha)\%$ **prediction interval for this future observation** is

$$\hat{y}_0 - t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \le Y_0 \le \hat{y}_0 + t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0)} \qquad (12\text{-}41)$$

This prediction interval is a generalization of the prediction interval given in Equation 11-33 for a future observation in simple linear regression. If you compare the prediction interval Equation 12-41 with the expression for the confidence interval on the mean, Equation 12-39,

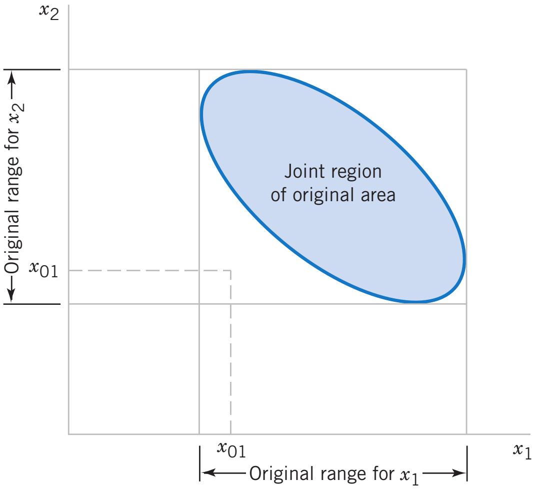


Figure 12-5 An example of extrapolation in multiple regression.

you will observe that the prediction interval is always wider than the confidence interval. The confidence interval expresses the error in estimating the mean of a distribution, while the prediction interval expresses the error in predicting a future observation from the distribution at the point $\mathbf{x}_0$. This must include the error in estimating the mean at that point, as well as the inherent variability in the random variable Y at the same value $\mathbf{x} = \mathbf{x}_0$.

Also, one might want to predict the mean of several values of Y, say m, all at the same value $\mathbf{x} = \mathbf{x}_0$. Because the variance of a sample mean is σ^2/m, Equation 12-41 is modified as follows. Replace the constant 1 under the square root with $1/m$ to reflect the lower variability in the mean of m observations. This results in a narrower interval.

In predicting new observations and in estimating the mean response at a given point $x_{01}, x_{02}, \ldots, x_{0k}$, we must be careful about **extrapolating** beyond the region containing the original observations. It is very possible that a model that fits well in the region of the original data will no longer fit well outside of that region. In multiple regression it is often easy to inadvertently extrapolate, since the levels of the variables $(x_{i1}, x_{i2}, \ldots, x_{ik})$, $i = 1, 2, \ldots, n$, jointly define the region containing the data. As an example, consider Fig. 12-5, which illustrates the region containing the observations for a two-variable regression model. Note that the point (x_{01}, x_{02}) lies within the ranges of both regressor variables x_1 and x_2, but it is outside the region that is actually spanned by the original observations. This is sometimes called a **hidden extrapolation.** Either predicting the value of a new observation or estimating the mean response at this point is an extrapolation of the original regression model.

EXAMPLE 12-9 Wire Bond Strength Confidence Interval

Suppose that the engineer in Example 12-1 wishes to construct a 95% prediction interval on the wire bond pull strength when the wire length is $x_1 = 8$ and the die height is $x_2 = 275$. Note that $\mathbf{x}'_0 = [1 \quad 8 \quad 275]$, and the point estimate of the pull strength is $\hat{y}_0 = \mathbf{x}'_0\hat{\beta} = 27.66$. Also, in Example 12-8 we calculated $\mathbf{x}'_0(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0 = 0.04444$. Therefore, from Equation 12-41 we have

$$27.66 - 2.074\sqrt{5.2352(1 + 0.0444)} \le Y_0 \le 27.66 + 2.074\sqrt{5.2352(1 + 0.0444)}$$

and the 95% prediction interval is

$$22.81 \le Y_0 \le 32.51$$

Notice that the prediction interval is wider than the confidence interval on the mean response at the same point, calculated in Example 12-8. The Minitab output in Table 12-4 also displays this prediction interval.

EXERCISES FOR SECTIONS 12-3 AND 12-4

12-41. Consider the regression model fit to the shear strength of soil in Exercise 12-1.
(a) Calculate 95% confidence intervals on each regression coefficient.
(b) Calculate a 95% confidence interval on mean strength when $x_1 = 18$ feet and $x_2 = 43\%$.
(c) Calculate 95% prediction interval on strength for the same values of the regressors used in the previous part.

12-42. Consider the soil absorption data in Exercise 12-2.
(a) Find 95% confidence intervals on the regression coefficients.
(b) Find a 95% confidence interval on mean soil absorption index when $x_1 = 200$ and $x_2 = 50$.
(c) Find a 95% prediction interval on the soil absorption index when $x_1 = 200$ and $x_2 = 50$.

12-43. Consider the semiconductor data in Exercise 12-9.
(a) Find 99% confidence intervals on the regression coefficients.
(b) Find a 99% prediction interval on HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$.
(c) Find a 99% confidence interval on mean HFE when $x_1 = 14.5$, $x_2 = 220$, and $x_3 = 5.0$.

12-44. Consider the electric power consumption data in Exercise 12-6.
(a) Find 95% confidence intervals on β_1, β_2, β_3, and β_4.
(b) Find a 95% confidence interval on the mean of Y when $x_1 = 75$, $x_2 = 24$, $x_3 = 90$, and $x_4 = 98$.
(c) Find a 95% prediction interval on the power consumption when $x_1 = 75$, $x_2 = 24$, $x_3 = 90$, and $x_4 = 98$.

12-45. Consider the bearing wear data in Exercise 12-19.
(a) Find 99% confidence intervals on β_1 and β_2.
(b) Recompute the confidence intervals in part (a) after the interaction term x_1x_2 is added to the model. Compare the lengths of these confidence intervals with those computed in part (a). Do the lengths of these intervals provide any information about the contribution of the interaction term in the model?

12-46. Consider the wire bond pull strength data in Exercise 12-8.
(a) Find 95% confidence interval on the regression coefficients.
(b) Find a 95% confidence interval on mean pull strength when $x_2 = 20$, $x_3 = 30$, $x_4 = 90$, and $x_5 = 2.0$.
(c) Find a 95% prediction interval on pull strength when $x_2 = 20$, $x_3 = 30$, $x_4 = 90$, and $x_5 = 2.0$.

12-47. Consider the regression model fit to the X-ray inspection data in Exercise 12-11. Use rads as the response.
(a) Calculate 95% confidence intervals on each regression coefficient.
(b) Calculate a 99% confidence interval on mean rads at 15 milliamps and 1 second on exposure time.
(c) Calculate a 99% prediction interval on rads for the same values of the regressors used in the previous part.

12-48. Consider the regression model fit to the arsenic data in Exercise 12-12. Use arsenic in nails as the response and age, drink use, and cook use as the regressors.
(a) Calculate 99% confidence intervals on each regression coefficient.
(b) Calculate a 99% confidence interval on mean arsenic concentration in nails when age $= 30$, drink use $= 4$, and cook use $= 4$.
(c) Calculate a prediction interval on arsenic concentration in nails for the same values of the regressors used in the previous part.

12-49. Consider the regression model fit to the coal and limestone mixture data in Exercise 12-13. Use density as the response.
(a) Calculate 90% confidence intervals on each regression coefficient.
(b) Calculate a 90% confidence interval on mean density when the dielectric constant $= 2.3$ and the loss factor $= 0.025$.
(c) Calculate a prediction interval on density for the same values of the regressors used in the previous part.

12-50. Consider the regression model fit to the nisin extraction data in Exercise 12-14.
(a) Calculate 95% confidence intervals on each regression coefficient.
(b) Calculate a 95% confidence interval on mean nisin extraction when $x_1 = 15.5$ and $x_2 = 16$.
(c) Calculate a prediction interval on nisin extraction for the same values of the regressors used in the previous part.
(d) Comment on the effect of a small sample size to the widths of these intervals.

12-51. Consider the regression model fit to the grey range modulation data in Exercise 12-15. Use the useful range as the response.
(a) Calculate 99% confidence intervals on each regression coefficient.
(b) Calculate a 99% confidence interval on mean useful range when brightness $= 70$ and contrast $= 80$.
(c) Calculate a prediction interval on useful range for the same values of the regressors used in the previous part.
(d) Calculate a 99% confidence interval and a 99% a prediction interval on useful range when brightness $= 50$ and contrast $= 25$. Compare the widths of these intervals to those calculated in parts (b) and (c). Explain any differences in widths.

12-52. Consider the stack loss data in Exercise 12-16.
(a) Calculate 95% confidence intervals on each regression coefficient.
(b) Calculate a 95% confidence interval on mean stack loss when $x_1 = 80$, $x_2 = 25$ and $x_3 = 90$.
(c) Calculate a prediction interval on stack loss for the same values of the regressors used in the previous part.

(d) Calculate a 95% confidence interval and a 95% prediction interval on stack loss when $x_1 = 80$, $x_2 = 19$, and $x_3 = 93$. Compare the widths of these intervals to those calculated in parts (b) and (c). Explain any differences in widths.

12-53. Consider the NFL data in Exercise 12-17.

(a) Find 95% confidence intervals on the regression coefficients.

(b) What is the estimated standard error of $\hat{\mu}_{Y|\mathbf{x}_0}$ when the percentage of completions is 60%, the percentage of TDs is 4%, and the percentage of interceptions is 3%.

(c) Find a 95% confidence interval on the mean rating when the percentage of completions is 60%, the percentage of TDs is 4%, and the percentage of interceptions is 3%.

12-54. Consider the heat treating data from Exercise 12-10.

(a) Find 95% confidence intervals on the regression coefficients.

(b) Find a 95% confidence interval on mean PITCH when TEMP = 1650, SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.

(c) Fit a model to PITCH using regressors x_1 = SOAKTIME × SOAKPCT and x_2 = DIFFTIME × DIFFPCT. Using the model with regressors x_1 and x_2, find a 95% confidence interval on mean PITCH when SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.

(d) Compare the length of this confidence interval with the length of the confidence interval on mean PITCH at the same point from part (b), where an additive model in SOAKTIME, SOAKPCT, DIFFTIME, and DIFFPCT was used. Which confidence interval is shorter? Does this tell you anything about which model is preferable?

12-55. Consider the gasoline mileage data in Exercise 12-7.

(a) Find 99% confidence intervals on the regression coefficients.

(b) Find a 99% confidence interval on the mean of Y for the regressor values in the first row of data.

(c) Fit a new regression model to these data using *cid, etw,* and *axle* as the regressors. Find 99% confidence intervals on the regression coefficients in this new model.

(d) Compare the lengths of the confidence intervals from part (c) with those found in part (a). Which intervals are longer? Does this offer any insight about which model is preferable?

12-56. Consider the NHL data in Exercise 12-18.

(a) Find a 95% confidence interval on the regression coefficient for the variable *GF.*

(b) Fit a simple linear regression model relating the response variable W to the regressor *GF*.

(c) Find a 95% confidence interval on the slope for the simple linear regression model from part (b).

(d) Compare the lengths of the two confidence intervals computed in parts (a) and (c). Which interval is shorter? Does this tell you anything about which model is preferable?

12-5 MODEL ADEQUACY CHECKING

12-5.1 Residual Analysis

The **residuals** from the multiple regression model, defined by $e_i = y_i - \hat{y}_i$, play an important role in judging model adequacy just as they do in simple linear regression. As noted in Section 11-7.1, several residual plots are often useful; these are illustrated in Example 12-10. It is also helpful to plot the residuals against variables not presently in the model that are possible candidates for inclusion. Patterns in these plots may indicate that the model may be improved by adding the candidate variable.

EXAMPLE 12-10 Wire Bond Strength Residuals

The residuals for the model from Example 12-1 are shown in Table 12-3. A normal probability plot of these residuals is shown in Fig. 12-6. No severe deviations from normality are obviously apparent, although the two largest residuals ($e_{15} = 5.84$ and $e_{17} = 4.33$) do not fall extremely close to a straight line drawn through the remaining residuals.

The **standardized residuals**

Standardized Residual

$$d_i = \frac{e_i}{\sqrt{MS_E}} = \frac{e_i}{\sqrt{\hat{\sigma}^2}} \qquad (12\text{-}42)$$

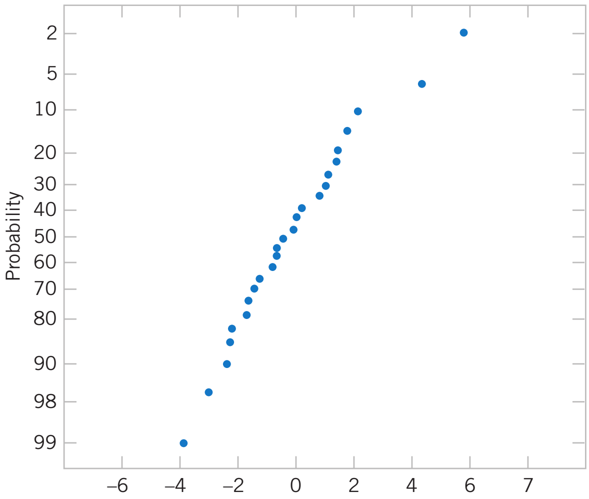


Figure 12-6 Normal probability plot of residuals.

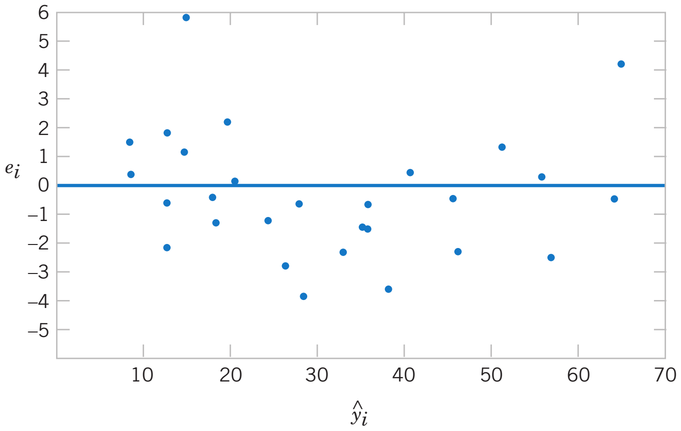


Figure 12-7 Plot of residuals against $\hat{y}$.

are often more useful than the ordinary residuals when assessing residual magnitude. For the wire bond strength example, the standardized residuals corresponding to e_{15} and e_{17} are $d_{15} = 5.84/\sqrt{5.2352} = 2.55$ and $d_{17} = 4.33/\sqrt{5.2352} = 1.89$, and they do not seem unusually large. Inspection of the data does not reveal any error in collecting observations 15 and 17, nor does it produce any other reason to discard or modify these two points.

The residuals are plotted against $\hat{y}$ in Fig. 12-7, and against x_1 and x_2 in Figs. 12-8 and 12-9, respectively.* The two largest residuals, e_{15} and e_{17}, are apparent. Figure 12-8 gives some indication that the model underpredicts the pull strength for assemblies with short wire length ($x_1 \leq 6$) and long wire length ($x_1 \geq 15$) and overpredicts the strength for assemblies with intermediate wire length ($7 \leq x_1 \leq 14$). The same impression is obtained from Fig. 12-7. Either the relationship between strength and wire length is not linear (requiring that a term involving x_1^2, say, be added to the model), or other regressor variables not presently in the model affected the response.

In the wire bond strength example we used the standardized residuals $d_i = e_i/\sqrt{\hat{\sigma}^2}$ as a measure of residual magnitude. Some analysts prefer to plot standardized residuals instead of ordinary residuals, because the standardized residuals are scaled so that their standard

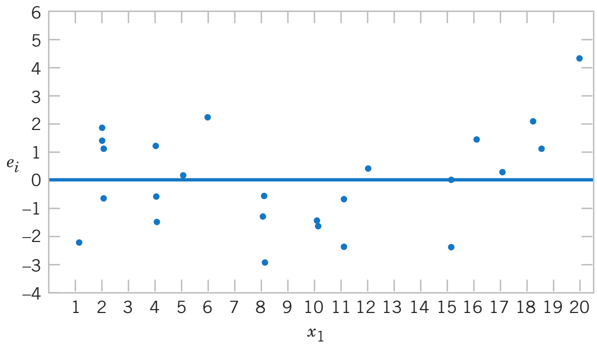


Figure 12-8 Plot of residuals against x_1.

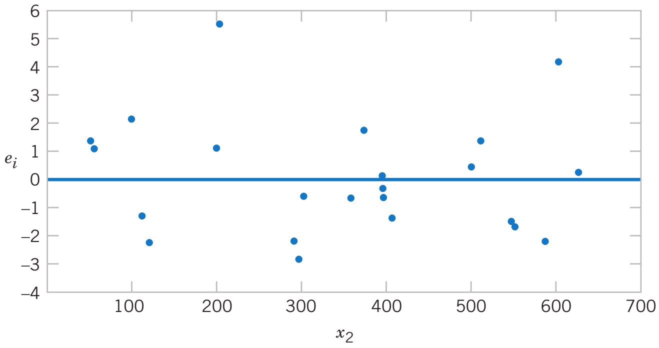


Figure 12-9 Plot of residuals against x_2.

*There are other methods, described in Montgomery, Peck, and Vining (2006) and Myers (1990), that plot a modified version of the residual, called a **partial residual,** against each regressor. These partial residual plots are useful in displaying the relationship between the response y and each individual regressor.

deviation is approximately unity. Consequently, large residuals (that may indicate possible outliers or unusual observations) will be more obvious from inspection of the residual plots.

Many regression computer programs compute other types of scaled residuals. One of the most popular are the **studentized residuals**

Studentized Residual

$$r_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1 - h_{ii})}} \qquad i = 1, 2, \ldots, n \qquad \text{(12-43)}$$

where h_{ii} is the ith diagonal element of the matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

The **H** matrix is sometimes called the **"hat" matrix,** since

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

Thus **H** transforms the observed values of **y** into a vector of fitted values $\hat{\mathbf{y}}$.

Since each row of the matrix **X** corresponds to a vector, say $\mathbf{x}_i' = [1, x_{i1}, x_{i2}, \ldots, x_{ik}]$, another way to write the diagonal elements of the hat matrix is

Diagonal Elements of Hat Matrix

$$h_{ii} = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i \qquad \text{(12-44)}$$

Note that apart from σ^2, h_{ii} is the variance of the fitted value $\hat{y}_i$. The quantities h_{ii} were used in the computation of the confidence interval on the mean response in Section 12-3.2.

Under the usual assumptions that the model errors are independently distributed with mean zero and variance σ^2, we can show that the variance of the ith residual e_i is

$$V(e_i) = \sigma^2(1 - h_{ii}), \qquad i = 1, 2, \ldots, n$$

Furthermore, the h_{ii} elements must fall in the interval $0 < h_{ii} \le 1$. This implies that the standardized residuals understate the true residual magnitude; thus, the studentized residuals would be a better statistic to examine in evaluating potential **outliers.**

To illustrate, consider the two observations identified in the wire bond strength data (Example 12-10) as having residuals that might be unusually large, observations 15 and 17. The standardized residuals are

$$d_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2}} = \frac{5.84}{\sqrt{5.2352}} = 2.55 \qquad \text{and} \qquad d_{17} = \frac{e_{17}}{\sqrt{MS_E}} = \frac{4.33}{\sqrt{5.2352}} = 1.89$$

Now $h_{15,15} = 0.0737$ and $h_{17,17} = 0.2593$, so the studentized residuals are

$$r_{15} = \frac{e_{15}}{\sqrt{\hat{\sigma}^2(1 - h_{15,15})}} = \frac{5.84}{\sqrt{5.2352(1 - 0.0737)}} = 2.65$$

and

$$r_{17} = \frac{e_{17}}{\sqrt{\hat{\sigma}^2(1 - h_{17,17})}} = \frac{4.33}{\sqrt{5.2352(1 - 0.2593)}} = 2.20$$

Notice that the studentized residuals are larger than the corresponding standardized residuals. However, the studentized residuals are still not so large as to cause us serious concern about possible outliers.

12-5.2 Influential Observations

When using multiple regression, we occasionally find that some subset of the observations is unusually influential. Sometimes these influential observations are relatively far away from the vicinity where the rest of the data were collected. A hypothetical situation for two variables is depicted in Fig. 12-10, where one observation in x-space is remote from the rest of the data. The disposition of points in the x-space is important in determining the properties of the model. For example, point (x_{i1}, x_{i2}) in Fig. 12-10 may be very influential in determining R^2, the estimates of the regression coefficients, and the magnitude of the error mean square.

We would like to examine the influential points to determine whether they control many model properties. If these influential points are "bad" points, or erroneous in any way, they should be eliminated. On the other hand, there may be nothing wrong with these points, but at least we would like to determine whether or not they produce results consistent with the rest of the data. In any event, even if an influential point is a valid one, if it controls important model properties, we would like to know this, since it could have an impact on the use of the model.

Montgomery, Peck, and Vining (2006) and Myers (1990) describe several methods for detecting influential observations. An excellent diagnostic is the **distance measure** developed by Dennis R. Cook. This is a measure of the squared distance between the usual least squares estimate of $\boldsymbol{\beta}$ based on all n observations and the estimate obtained when the ith point is removed, say, $\hat{\boldsymbol{\beta}}_{(i)}$. The **Cook's distance measure** is

Cook's Distance

$$D_i = \frac{(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\hat{\boldsymbol{\beta}}_{(i)} - \hat{\boldsymbol{\beta}})}{p\hat{\sigma}^2} \qquad i = 1, 2, \ldots, n$$

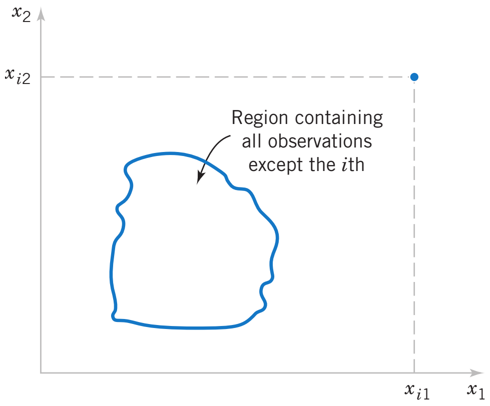


Figure 12-10 A point that is remote in x-space.

Clearly, if the ith point is influential, its removal will result in $\hat{\boldsymbol{\beta}}_{(i)}$ changing considerably from the value $\hat{\boldsymbol{\beta}}$. Thus, a large value of D_i implies that the ith point is influential. The statistic D_i is actually computed using

Cook's Distance Formula

$$D_i = \frac{r_i^2}{p}\,\frac{h_{ii}}{(1 - h_{ii})} \qquad i = 1, 2, \ldots, n \tag{12-45}$$

From Equation 12-44 we see that D_i consists of the squared studentized residual, which reflects how well the model fits the ith observation y_i [recall that $r_i = e_i/\sqrt{\hat{\sigma}^2(1 - h_{ii})}\,$] and a component that measures how far that point is from the rest of the data $[h_{ii}/(1 - h_{ii})$ is a measure of the distance of the ith point from the centroid of the remaining $n - 1$ points]. A value of $D_i > 1$ would indicate that the point is influential. Either component of D_i (or both) may contribute to a large value.

EXAMPLE 12-11 Wire Bond Strength Cook's Distances

Table 12-12 lists the values of the hat matrix diagonals h_{ii} and Cook's distance measure D_i for the wire bond pull strength data in Example 12-1. To illustrate the calculations, consider the first observation:

$$\begin{aligned} D_1 &= \frac{r_1^2}{p}\cdot\frac{h_{11}}{(1 - h_{11})} \\ &= \frac{[e_1/\sqrt{MS_E(1 - h_{11})}]^2}{p}\cdot\frac{h_{11}}{(1 - h_{11})} \\ &= \frac{[1.57/\sqrt{5.2352(1 - 0.1573)}]^2}{3}\cdot\frac{0.1573}{(1 - 0.1573)} \\ &= 0.035 \end{aligned}$$

The Cook distance measure D_i does not identify any potentially influential observations in the data, for no value of D_i exceeds unity.

Table 12-12 Influence Diagnostics for the Wire Bond Pull Strength Data 2

Observations i	h_{ii}	Cook's Distance Measure D_i	Observations i	h_{ii}	Cook's Distance Measure D_i
1	0.1573	0.035	14	0.1129	0.003
2	0.1116	0.012	15	0.0737	0.187
3	0.1419	0.060	16	0.0879	0.001
4	0.1019	0.021	17	0.2593	0.565
5	0.0418	0.024	18	0.2929	0.155
6	0.0749	0.007	19	0.0962	0.018
7	0.1181	0.036	20	0.1473	0.000
8	0.1561	0.020	21	0.1296	0.052
9	0.1280	0.160	22	0.1358	0.028
10	0.0413	0.001	23	0.1824	0.002
11	0.0925	0.013	24	0.1091	0.040
12	0.0526	0.001	25	0.0729	0.000
13	0.0820	0.001			

EXERCISES FOR SECTION 12-5

12-57. Consider the gasoline mileage data in Exercise 12-7.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals and comment on the normality assumption.

(c) Plot residuals versus $\hat{y}$ and versus each regressor. Discuss these residual plots.

(d) Calculate Cook's distance for the observations in this data set. Are any observations influential?

12-58. Consider the electric power consumption data in Exercise 12-6.

(a) Calculate R^2 for this model. Interpret this quantity.

(b) Plot the residuals versus $\hat{y}$ and versus each regressor. Interpret this plot.

(c) Construct a normal probability plot of the residuals and comment on the normality assumption.

12-59. Consider the regression model for the NFL data in Exercise 12-17.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals. What conclusion can you draw from this plot?

(c) Plot the residuals versus $\hat{y}$ and versus each regressor, and comment on model adequacy.

(d) Are there any influential points in these data?

12-60. Consider the regression model for the heat treating data in Exercise 12-10.

(a) Calculate the percent of variability explained by this model.

(b) Construct a normal probability plot for the residuals. Comment on the normality assumption.

(c) Plot the residuals versus $\hat{y}$ and interpret the display.

(d) Calculate Cook's distance for each observation and provide an interpretation of this statistic.

12-61. Consider the regression model fit to the X-ray inspection data in Exercise 12-11. Use rads as the response.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals. What conclusion can you draw from this plot?

(c) Plot the residuals versus $\hat{y}$ and versus each regressor, and comment on model adequacy.

(d) Calculate Cook's distance for the observations in this data set. Are there any influential points in these data?

12-62. Consider the regression model fit to the arsenic data in Exercise 12-12. Use arsenic in nails as the response and age, drink use, and cook use as the regressors.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals. What conclusion can you draw from this plot?

(c) Plot the residuals versus $\hat{y}$ and versus each regressor, and comment on model adequacy.

(d) Calculate Cook's distance for the observations in this data set. Are there any influential points in these data?

12-63. Consider the regression model fit to the coal and limestone mixture data in Exercise 12-13. Use density as the response.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals. What conclusion can you draw from this plot?

(c) Plot the residuals versus $\hat{y}$ and versus each regressor, and comment on model adequacy.

(d) Calculate Cook's distance for the observations in this data set. Are there any influential points in these data?

12-64. Consider the regression model fit to the nisin extraction data in Exercise 12-14.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals. What conclusion can you draw from this plot?

(c) Plot the residuals versus $\hat{y}$ and versus each regressor, and comment on model adequacy.

(d) Calculate Cook's distance for the observations in this data set. Are there any influential points in these data?

12-65. Consider the regression model fit to the grey range modulation data in Exercise 12-15. Use the useful range as the response.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals. What conclusion can you draw from this plot?

(c) Plot the residuals versus $\hat{y}$ and versus each regressor, and comment on model adequacy.

(d) Calculate Cook's distance for the observations in this data set. Are there any influential points in these data?

12-66. Consider the stack loss data in Exercise 12-16.

(a) What proportion of total variability is explained by this model?

(b) Construct a normal probability plot of the residuals. What conclusion can you draw from this plot?

(c) Plot the residuals versus $\hat{y}$ and versus each regressor, and comment on model adequacy.

(d) Calculate Cook's distance for the observations in this data set. Are there any influential points in these data?

12-67. Consider the bearing wear data in Exercise 12-19.

(a) Find the value of R^2 when the model uses the regressors x_1 and x_2.

(b) What happens to the value of R^2 when an interaction term x_1x_2 is added to the model? Does this necessarily imply that adding the interaction term is a good idea?

12-68. Fit a model to the response PITCH in the heat treating data of Exercise 12-10 using new regressors x_1 = SOAKTIME × SOAKPCT and x_2 = DIFFTIME × DIFFPCT.

(a) Calculate the R^2 for this model and compare it to the value of R^2 from the original model in Exercise 12-10. Does this provide some information about which model is preferable?

(b) Plot the residuals from this model versus $\hat{y}$ and on a normal probability scale. Comment on model adequacy.

(c) Find the values of Cook's distance measure. Are any observations unusually influential?

12-69. Consider the semiconductor HFE data in Exercise 12-9.

(a) Plot the residuals from this model versus $\hat{y}$. Comment on the information in this plot.

(b) What is the value of R^2 for this model?

(c) Refit the model using log HFE as the response variable.

(d) Plot the residuals versus predicted log HFE for the model in part (c). Does this give any information about which model is preferable?

(e) Plot the residuals from the model in part (d) versus the regressor x_3. Comment on this plot.

(f) Refit the model to log HFE using x_1, x_2, and $1/x_3$, as the regressors. Comment on the effect of this change in the model.

12-70. Consider the regression model for the NHL data from Exercise 12-18.

(a) Fit a model using *GF* as the only regressor.

(b) How much variability is explained by this model?

(c) Plot the residuals versus $\hat{y}$ and comment on model adequacy.

(d) Plot the residuals from part (a) versus *PPGF*, the points scored while in power play. Does this indicate that the model would be better if this variable were included?

12-71. The diagonal elements of the hat matrix are often used to denote **leverage**—that is, a point that is unusual in its location in the x-space and that may be influential. Generally, the ith point is called a **leverage point** if its hat diagonal h_{ii} exceeds $2p/n$, which is twice the average size of all the hat diagonals. Recall that $p = k + 1$.

(a) Table 12-12 contains the hat diagonal for the wire bond pull strength data used in Example 12-1. Find the average size of these elements.

(b) Based on the criterion above, are there any observations that are leverage points in the data set?

12-6 ASPECTS OF MULTIPLE REGRESSION MODELING

In this section we briefly discuss several other aspects of building multiple regression models. For more extensive presentations of these topics and additional examples refer to Montgomery, Peck, and Vining (2006) and Myers (1990).

12-6.1 Polynomial Regression Models

The linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ is a general model that can be used to fit any relationship that is **linear in the unknown parameters $\boldsymbol{\beta}$.** This includes the important class of **polynomial regression models.** For example, the second-degree polynomial in one variable

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon \tag{12-46}$$

and the second-degree polynomial in two variables

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 + \epsilon \tag{12-47}$$

are linear regression models.

Polynomial regression models are widely used when the response is curvilinear, because the general principles of multiple regression can be applied. The following example illustrates some of the types of analyses that can be performed.

EXAMPLE 12-12 Airplane Sidewall Panels

Sidewall panels for the interior of an airplane are formed in a 1500-ton press. The unit manufacturing cost varies with the production lot size. The data shown below give the average cost per unit (in hundreds of dollars) for this product (y) and the production lot size (x). The scatter diagram, shown in Fig. 12-11, indicates that a second-order polynomial may be appropriate.

y	1.81	1.70	1.65	1.55	1.48	1.40
x	20	25	30	35	40	50
y	1.30	1.26	1.24	1.21	1.20	1.18
x	60	65	70	75	80	90

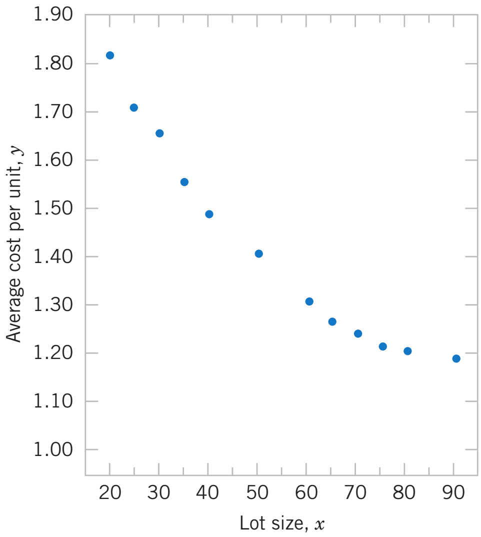


Figure 12-11 Data for Example 12-11.

We will fit the model

$$Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$$

The **y** vector, the model matrix **X** and the **β** vector are as follows:

$$\mathbf{y} = \begin{bmatrix} 1.81 \\ 1.70 \\ 1.65 \\ 1.55 \\ 1.48 \\ 1.40 \\ 1.30 \\ 1.26 \\ 1.24 \\ 1.21 \\ 1.20 \\ 1.18 \end{bmatrix} \qquad \mathbf{X} = \begin{bmatrix} 1 & 20 & 400 \\ 1 & 25 & 625 \\ 1 & 30 & 900 \\ 1 & 35 & 1225 \\ 1 & 40 & 1600 \\ 1 & 50 & 2500 \\ 1 & 60 & 3600 \\ 1 & 65 & 4225 \\ 1 & 70 & 4900 \\ 1 & 75 & 5625 \\ 1 & 80 & 6400 \\ 1 & 90 & 8100 \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_{11} \end{bmatrix}$$

Solving the normal equations $\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{y}$ gives the fitted model

$$\hat{y} = 2.19826629 - 0.02252236x + 0.00012507x^2$$

Conclusions: The test for significance of regression is shown in Table 12-13. Since $f_0 = 1762.3$ is significant at 1%, we conclude that at least one of the parameters β_1 and β_{11} is not zero. Furthermore, the standard tests for model adequacy do not reveal any unusual behavior, and we would conclude that this is a reasonable model for the sidewall panel cost data.

In fitting polynomials, we generally like to use the **lowest-degree model** consistent with the data. In this example, it would seem logical to investigate the possibility of dropping the quadratic term from the model. That is, we would like to test

$$H_0\colon \beta_{11} = 0$$
$$H_1\colon \beta_{11} \neq 0$$

Table 12-13 Test for Significance of Regression for the Second-Order Model in Example 12-12

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	*P*-value
Regression	0.52516	2	0.26258	1762.28	2.12E-12
Error	0.00134	9	0.00015		
Total	0.5265	11			

Table 12-14 Analysis of Variance for Example 12-12, Showing the Test for H_0: $\beta_{11} = 0$

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	*P*-value
Regression	$SS_R(\beta_1,\beta_{11}\mid\beta_0) = 0.52516$	2	0.26258	1767.40	2.09E-12
Linear	$SS_R(\beta_1\mid\beta_0) = 0.49416$	1	0.49416	2236.12	7.13E-13
Quadratic	$SS_R(\beta_{11}\mid\beta_0,\beta_1) = 0.03100$	1	0.03100	208.67	1.56E-7
Error	0.00133	9	0.00015		
Total	0.5265	11			

The general regression significance test can be used to test this hypothesis. We need to determine the "extra sum of squares" due to β_{11}, or

$$SS_R(\beta_{11}\mid\beta_1,\beta_0) = SS_R(\beta_1,\beta_{11}\mid\beta_0) - SS_R(\beta_1\mid\beta_0)$$

The sum of squares $SS_R(\beta_1,\beta_{11}\mid\beta_0) = 0.52516$ from Table 12-13. To find $SS_R(\beta_1\mid\beta_0)$, we fit a simple linear regression model to the original data, yielding

$$\hat{y} = 1.90036313 - 0.00910056x$$

It can be easily verified that the regression sum of squares for this model is

$$SS_R(\beta_1\mid\beta_0) = 0.4942$$

Therefore, the extra sum of the squares due to β_{11}, given that β_1 and β_0 are in the model, is

$$\begin{aligned} SS_R(\beta_{11}\mid\beta_1,\beta_0) &= SS_R(\beta_1,\beta_{11}\mid\beta_0) - SS_R(\beta_1\mid\beta_0) \\ &= 0.5252 - 0.4942 \\ &= 0.031 \end{aligned}$$

The analysis of variance, with the test of H_0: $\beta_{11} = 0$ incorporated into the procedure, is displayed in Table 12-14. Note that the quadratic term contributes significantly to the model.

12-6.2 Categorical Regressors and Indicator Variables

The regression models presented in previous sections have been based on **quantitative** variables, that is, variables that are measured on a numerical scale. For example, variables such as temperature, pressure, distance, and voltage are quantitative variables. Occasionally, we need to incorporate **categorical,** or **qualitative,** variables in a regression model. For example, suppose that one of the variables in a regression model is the operator who is associated with each observation y_i. Assume that only two operators are involved. We may wish to assign different levels to the two operators to account for the possibility that each operator may have a different effect on the response.

The usual method of accounting for the different levels of a qualitative variable is to use **indicator variables.** For example, to introduce the effect of two different operators into a regression model, we could define an indicator variable as follows:

$$x = \begin{cases} 0 \text{ if the observation is from operator 1} \\ 1 \text{ if the observation is from operator 2} \end{cases}$$

In general, a qualitative variable with r-levels can be modeled by $r - 1$ indicator variables, which are assigned the value of either zero or one. Thus, if there are *three* operators, the different levels will be accounted for by the *two* indicator variables defined as follows:

x_1	x_2	
0	0	if the observation is from operator 1
1	0	if the observation is from operator 2
0	1	if the observation is from operator 3

Indicator variables are also referred to as **dummy** variables. The following example [from Montgomery, Peck, and Vining (2006)] illustrates some of the uses of indicator variables; for other applications, see Montgomery, Peck, and Vining (2006).

EXAMPLE 12-13 Surface Finish

A mechanical engineer is investigating the surface finish of metal parts produced on a lathe and its relationship to the speed (in revolutions per minute) of the lathe. The data are shown in Table 12-15. Note that the data have been collected using two different types of cutting tools. Since the type of cutting tool likely affects the surface finish, we will fit the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

where Y is the surface finish, x_1 is the lathe speed in revolutions per minute, and x_2 is an indicator variable denoting the type of cutting tool used; that is,

$$x_2 = \begin{cases} 0, \text{ for tool type 302} \\ 1, \text{ for tool type 416} \end{cases}$$

The parameters in this model may be easily interpreted. If $x_2 = 0$, the model becomes

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

which is a straight-line model with slope β_1 and intercept β_0. However, if $x_2 = 1$, the model becomes

$$Y = \beta_0 + \beta_1 x_1 + \beta_2(1) + \epsilon = (\beta_0 + \beta_2) + \beta_1 x_1 + \epsilon$$

which is a straight-line model with slope β_1 and intercept $\beta_0 + \beta_2$. Thus, the model $Y = \beta_0 + \beta_1 x + \beta_2 x_2 + \epsilon$ implies that surface finish is linearly related to lathe speed and that the slope β_1 does not depend on the type of cutting tool used. However, the type of cutting tool does affect the intercept, and β_2 indicates the change in the intercept associated with a change in tool type from 302 to 416.

The model matrix **X** and **y** vector for this problem are as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & 225 & 0 \\ 1 & 200 & 0 \\ 1 & 250 & 0 \\ 1 & 245 & 0 \\ 1 & 235 & 0 \\ 1 & 237 & 0 \\ 1 & 265 & 0 \\ 1 & 259 & 0 \\ 1 & 221 & 0 \\ 1 & 218 & 0 \\ 1 & 224 & 1 \\ 1 & 212 & 1 \\ 1 & 248 & 1 \\ 1 & 260 & 1 \\ 1 & 243 & 1 \\ 1 & 238 & 1 \\ 1 & 224 & 1 \\ 1 & 251 & 1 \\ 1 & 232 & 1 \\ 1 & 216 & 1 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} 45.44 \\ 42.03 \\ 50.10 \\ 48.75 \\ 47.92 \\ 47.79 \\ 52.26 \\ 50.52 \\ 45.58 \\ 44.78 \\ 33.50 \\ 31.23 \\ 37.52 \\ 37.13 \\ 34.70 \\ 33.92 \\ 32.13 \\ 35.47 \\ 33.49 \\ 32.29 \end{bmatrix}$$

The fitted model is

$$\hat{y} = 14.27620 + 0.14115 x_1 - 13.28020 x_2$$

Conclusions: The analysis of variance for this model is shown in Table 12-16. Note that the hypothesis $H_0: \beta_1 = \beta_2 = 0$ (significance of regression) would be rejected at any reasonable level of significance because the P-value is very small. This table also contains the sums of squares

$$\begin{aligned} SS_R &= SS_R(\beta_1, \beta_2 | \beta_0) \\ &= SS_R(\beta_1 | \beta_0) + SS_R(\beta_2 | \beta_1, \beta_0) \end{aligned}$$

so a test of the hypothesis H_0: $\beta_2 = 0$ can be made. Since this hypothesis is also rejected, we conclude that tool type has an effect on surface finish.

Table 12-15 Surface Finish Data for Example 12-13

Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool	Observation Number, i	Surface Finish y_i	RPM	Type of Cutting Tool
1	45.44	225	302	11	33.50	224	416
2	42.03	200	302	12	31.23	212	416
3	50.10	250	302	13	37.52	248	416
4	48.75	245	302	14	37.13	260	416
5	47.92	235	302	15	34.70	243	416
6	47.79	237	302	16	33.92	238	416
7	52.26	265	302	17	32.13	224	416
8	50.52	259	302	18	35.47	251	416
9	45.58	221	302	19	33.49	232	416
10	44.78	218	302	20	32.29	216	416

It is also possible to use indicator variables to investigate whether tool type affects both the slope and intercept. Let the model be

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon$$

where x_2 is the indicator variable. Now if tool type 302 is used, $x_2 = 0$, and the model is

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

If tool type 416 is used, $x_2 = 1$, and the model becomes

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 + \beta_3 x_1 + \epsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \epsilon \end{aligned}$$

Note that β_2 is the change in the intercept and that β_3 is the change in slope produced by a change in tool type.

Another method of analyzing these data is to fit separate regression models to the data for each tool type. However, the indicator variable approach has several advantages. First, only one regression model must be fit. Second, by pooling the data on both tool types, more degrees of freedom for error are obtained. Third, tests of both hypotheses on the parameters β_2 and β_3 are just special cases of the extra sum of squares method.

12-6.3 Selection of Variables and Model Building

An important problem in many applications of regression analysis involves selecting the set of regressor variables to be used in the model. Sometimes previous experience or underlying theoretical considerations can help the analyst specify the set of regressor variables to use in a particular situation. Usually, however, the problem consists of selecting an appropriate set of

Table 12-16 Analysis of Variance for Example 12-13

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Square	f_0	P-value
Regression	1012.0595	2	506.0297	1103.69	1.02E-18
$SS_R(\beta_1 \mid \beta_0)$	130.6091	1	130.6091	284.87	4.70E-12
$SS_R(\beta_2 \mid \beta_1, \beta_0)$	881.4504	1	881.4504	1922.52	6.24E-19
Error	7.7943	17	0.4585		
Total	1019.8538	19			

regressors from a set that quite likely includes all the important variables, but we are sure that not all these candidate regressors are necessary to adequately model the response Y.

In such a situation, we are interested in **variable selection;** that is, screening the candidate variables to obtain a regression model that contains the "best" subset of regressor variables. We would like the final model to contain enough regressor variables so that in the intended use of the model (prediction, for example) it will perform satisfactorily. On the other hand, to keep model maintenance costs to a minimum and to make the model easy to use, we would like the model to use as few regressor variables as possible. The compromise between these conflicting objectives is often called finding the "best" regression equation. However, in most problems, no single regression model is "best" in terms of the various evaluation criteria that have been proposed. A great deal of judgment and experience with the system being modeled is usually necessary to select an appropriate set of regressor variables for a regression equation.

No single algorithm will always produce a good solution to the variable selection problem. Most of the currently available procedures are search techniques, and to perform satisfactorily, they require interaction with judgment by the analyst. We now briefly discuss some of the more popular variable selection techniques. We assume that there are K candidate regressors, $x_1, x_2, \ldots, x_K$, and a single response variable y. All models will include an intercept term β_0, so the model with *all* variables included would have $K + 1$ terms. Furthermore, the functional form of each candidate variable (for example, $x_1 = 1/x$, $x_2 = \ln x$, etc.) is assumed to be correct.

All Possible Regressions

This approach requires that the analyst fit all the regression equations involving one candidate variable, all regression equations involving two candidate variables, and so on. Then these equations are evaluated according to some suitable criteria to select the "best" regression model. If there are K candidate regressors, there are 2^K total equations to be examined. For example, if $K = 4$, there are $2^4 = 16$ possible regression equations; while if $K = 10$, there are $2^{10} = 1024$ possible regression equations. Hence, the number of equations to be examined increases rapidly as the number of candidate variables increases. However, there are some very efficient computing algorithms for all possible regressions available and they are widely implemented in statistical software, so it is a very practical procedure unless the number of candidate regressors is fairly large. Look for a menu choice such as "Best Subsets" regression.

Several criteria may be used for evaluating and comparing the different regression models obtained. A commonly used criterion is based on the value of R^2 or the value of the adjusted R^2, R^2_{adj}. Basically, the analyst continues to increase the number of variables in the model until the increase in R^2 or the adjusted R^2_{adj} is small. Often, we will find that the R^2_{adj} will stabilize and actually begin to decrease as the number of variables in the model increases. Usually, the model that maximizes R^2_{adj} is considered to be a good candidate for the best regression equation. Because we can write $R^2_{\text{adj}} = 1 - \{MS_E/[SS_T/(n-1)]\}$ and $SS_T/(n-1)$ is a constant, the model that maximizes the R^2_{adj} value also minimizes the mean square error, so this is a very attractive criterion.

Another criterion used to evaluate regression models is the **C_p statistic,** which is a measure of the total mean square error for the regression model. We define the total standardized mean square error for the regression model as

$$\begin{aligned}\Gamma_p &= \frac{1}{\sigma^2}\sum_{i=1}^{n} E[\hat{Y}_i - E(Y_i)]^2 \\ &= \frac{1}{\sigma^2}\left\{\sum_{i=1}^{n}[E(Y_i) - E(\hat{Y}_i)]^2 + \sum_{i=1}^{n} V(\hat{Y}_i)\right\} \\ &= \frac{1}{\sigma^2}[(\text{bias})^2 + \text{variance}]\end{aligned}$$

We use the mean square error from the *full* $K + 1$ term model as an estimate of σ^2; that is, $\hat{\sigma}^2 = MS_E(K + 1)$. Then an estimator of Γ_p is [see Montgomery, Peck, and Vining (2006) or Myers (1990) for the details]:

C_p Statistic

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p \qquad (12\text{-}48)$$

If the p-term model has negligible bias, it can be shown that

$$E(C_p \mid \text{zero bias}) = p$$

Therefore, the values of C_p for each regression model under consideration should be evaluated relative to p. The regression equations that have negligible bias will have values of C_p that are close to p, while those with significant bias will have values of C_p that are significantly greater than p. We then choose as the "best" regression equation either a model with *minimum* C_p or a model with a slightly larger C_p, that does not contain as much bias (i.e., $C_p \cong p$).

The **PRESS** statistic can also be used to evaluate competing regression models. PRESS is an acronym for **Prediction Error Sum of Squares,** and it is defined as the sum of the squares of the differences between each observation y_i and the corresponding predicted value based on a model fit to the *remaining* $n - 1$ points, say $\hat{y}_{(i)}$. So PRESS provides a measure of how well the model is likely to perform when predicting *new* data, or data that was not used to fit the regression model. The computing formula for PRESS is

Prediction Error Sum of Squares (PRESS)

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{y}_{(i)})^2 = \sum_{i=1}^{n} \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

where $e_i = y_i - \hat{y}_i$ is the usual residual. Thus PRESS is easy to calculate from the standard least squares regression results. Models that have small values of PRESS are preferred.

EXAMPLE 12-14 Wine Quality

Table 12-17 presents data on taste-testing 38 brands of pinot noir wine (the data were first reported in an article by Kwan, Kowalski, and Skogenboe in an article in the *Journal of Agricultural and Food Chemistry,* Vol. 27, 1979, and it also appears as one of the default data sets in Minitab). The response variable is y = quality, and we wish to find the "best" regression equation that relates quality to the other five parameters.

Figure 12-12 is the matrix of scatter plots for the wine quality data, as constructed by Minitab. We notice that there are some indications of possible linear relationships between quality and the regressors, but there is no obvious visual impression of which regressors would be appropriate. Table 12-18 lists the all possible regressions output from Minitab. In this analysis, we asked Minitab to present the best three equations for each subset size. Note that Minitab reports the values of R^2, R^2_{adj}, C_p, and $S = \sqrt{MS_E}$ for each model. From Table 12-18 we see that the three-variable equation with x_2 = aroma, x_4 = flavor, and x_5 = oakiness produces the minimum C_p equation, whereas the four-variable model, which adds x_1 = clarity to the previous three regressors, results in maximum R^2_{adj} (or minimum MS_E). The three-variable model is

$$\hat{y} = 6.47 + 0.580x_2 + 1.20x_4 - 0.602x_5$$

and the four-variable model is

$$\hat{y} = 4.99 + 1.79x_1 + 0.530x_2 + 1.26x_4 - 0.659x_5$$

Table 12-17 Wine Quality Data

	x_1 Clarity	x_2 Aroma	x_3 Body	x_4 Flavor	x_5 Oakiness	y Quality
1	1.0	3.3	2.8	3.1	4.1	9.8
2	1.0	4.4	4.9	3.5	3.9	12.6
3	1.0	3.9	5.3	4.8	4.7	11.9
4	1.0	3.9	2.6	3.1	3.6	11.1
5	1.0	5.6	5.1	5.5	5.1	13.3
6	1.0	4.6	4.7	5.0	4.1	12.8
7	1.0	4.8	4.8	4.8	3.3	12.8
8	1.0	5.3	4.5	4.3	5.2	12.0
9	1.0	4.3	4.3	3.9	2.9	13.6
10	1.0	4.3	3.9	4.7	3.9	13.9
11	1.0	5.1	4.3	4.5	3.6	14.4
12	0.5	3.3	5.4	4.3	3.6	12.3
13	0.8	5.9	5.7	7.0	4.1	16.1
14	0.7	7.7	6.6	6.7	3.7	16.1
15	1.0	7.1	4.4	5.8	4.1	15.5
16	0.9	5.5	5.6	5.6	4.4	15.5
17	1.0	6.3	5.4	4.8	4.6	13.8
18	1.0	5.0	5.5	5.5	4.1	13.8
19	1.0	4.6	4.1	4.3	3.1	11.3
20	0.9	3.4	5.0	3.4	3.4	7.9
21	0.9	6.4	5.4	6.6	4.8	15.1
22	1.0	5.5	5.3	5.3	3.8	13.5
23	0.7	4.7	4.1	5.0	3.7	10.8
24	0.7	4.1	4.0	4.1	4.0	9.5
25	1.0	6.0	5.4	5.7	4.7	12.7
26	1.0	4.3	4.6	4.7	4.9	11.6
27	1.0	3.9	4.0	5.1	5.1	11.7
28	1.0	5.1	4.9	5.0	5.1	11.9
29	1.0	3.9	4.4	5.0	4.4	10.8
30	1.0	4.5	3.7	2.9	3.9	8.5
31	1.0	5.2	4.3	5.0	6.0	10.7
32	0.8	4.2	3.8	3.0	4.7	9.1
33	1.0	3.3	3.5	4.3	4.5	12.1
34	1.0	6.8	5.0	6.0	5.2	14.9
35	0.8	5.0	5.7	5.5	4.8	13.5
36	0.8	3.5	4.7	4.2	3.3	12.2
37	0.8	4.3	5.5	3.5	5.8	10.3
38	0.8	5.2	4.8	5.7	3.5	13.2

These models should now be evaluated further using residuals plots and the other techniques discussed earlier in the chapter, to see if either model is satisfactory with respect to the underlying assumptions and to determine if one of them is preferable. It turns out that the residual plots do not reveal any major problems with either model. The value of PRESS for the three-variable model is 56.0524 and for the four-variable model it is 60.3327. Since PRESS is smaller in the model with three regressors, and since it is the model with the smallest number of predictors, it would likely be the preferred choice.

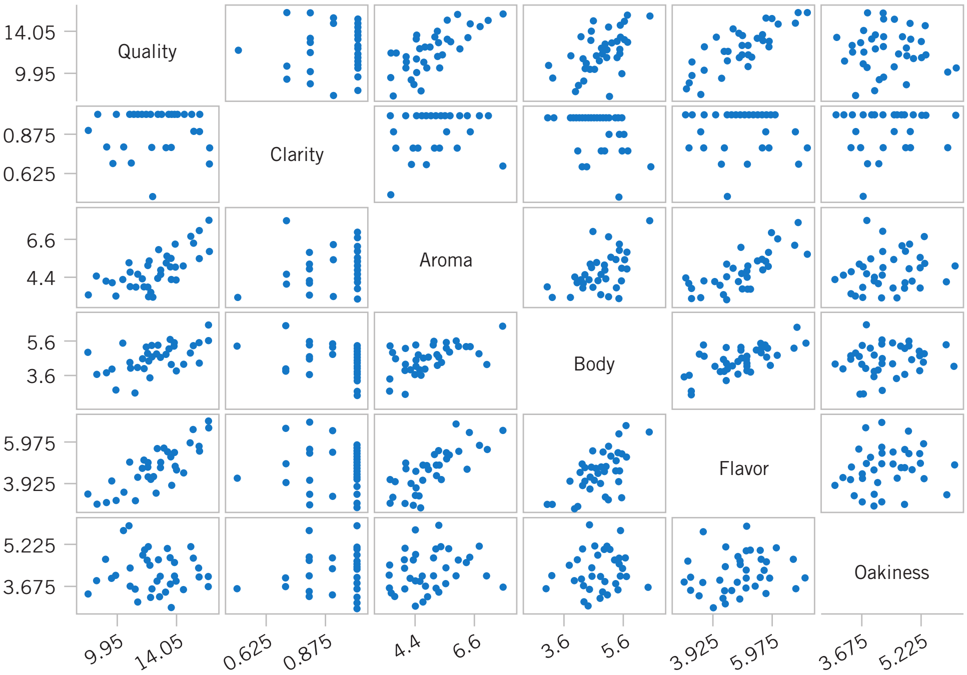


Figure 12-12 A matrix of scatter plots from Minitab for the wine quality data.

Table 12-18 Minitab All Possible Regressions Output for the Wine Quality Data

Best Subsets Regression: Quality versus Clarity, Aroma, . . .

Response is Quality

Vars	R-Sq	R-Sq (adj)	C–p	S	Clarity	Aroma	Body	Flavor	Oakiness
1	62.4	61.4	9.0	1.2712				X	
1	50.0	48.6	23.2	1.4658		X			
1	30.1	28.2	46.0	1.7335			X		
2	66.1	64.2	6.8	1.2242				X	X
2	65.9	63.9	7.1	1.2288		X		X	
2	63.3	61.2	10.0	1.2733	X			X	
3	70.4	67.8	3.9	1.1613		X		X	X
3	68.0	65.2	6.6	1.2068	X			X	X
3	66.5	63.5	8.4	1.2357			X	X	X
4	71.5	68.0	4.7	1.1568	X	X		X	X
4	70.5	66.9	5.8	1.1769		X	X	X	X
4	69.3	65.6	7.1	1.1996	X		X	X	X
5	72.1	67.7	6.0	1.1625	X	X	X	X	X

Stepwise Regression

Stepwise regression is probably the most widely used variable selection technique. The procedure iteratively constructs a sequence of regression models by adding or removing variables at each step. The criterion for adding or removing a variable at any step is usually expressed in terms of a partial F-test. Let f_{in} be the value of the F-random variable for adding a variable to the model, and let f_{out} be the value of the F-random variable for removing a variable from the model. We must have $f_{\text{in}} \geq f_{\text{out}}$, and usually $f_{\text{in}} = f_{\text{out}}$.

Stepwise regression begins by forming a one-variable model using the regressor variable that has the highest correlation with the response variable Y. This will also be the regressor producing the largest F-statistic. For example, suppose that at this step, x_1 is selected. At the second step, the remaining $K - 1$ candidate variables are examined, and the variable for which the partial F-statistic

$$F_j = \frac{SS_R(\beta_j|\beta_1,\beta_0)}{MS_E(x_j, x_1)} \tag{12-49}$$

is a maximum is added to the equation, provided that $f_j > f_{\text{in}}$. In equation 12-49, $MS_E\,(x_j, x_1)$ denotes the mean square for error for the model containing both x_1 and x_j. Suppose that this procedure indicates that x_2 should be added to the model. Now the stepwise regression algorithm determines whether the variable x_1 added at the first step should be removed. This is done by calculating the F-statistic

$$F_1 = \frac{SS_R(\beta_1|\beta_2,\beta_0)}{MS_E(x_1, x_2)} \tag{12-50}$$

If the calculated value $f_1 < f_{\text{out}}$, the variable x_1 is removed; otherwise it is retained, and we would attempt to add a regressor to the model containing both x_1 and x_2.

In general, at each step the set of remaining candidate regressors is examined, and the regressor with the largest partial F-statistic is entered, provided that the observed value of f exceeds f_{in}. Then the partial F-statistic for each regressor in the model is calculated, and the regressor with the smallest observed value of F is deleted if the observed $f < f_{\text{out}}$. The procedure continues until no other regressors can be added to or removed from the model.

Stepwise regression is almost always performed using a computer program. The analyst exercises control over the procedure by the choice of f_{in} and f_{out}. Some stepwise regression computer programs require that numerical values be specified for f_{in} and f_{out}. Since the number of degrees of freedom on MS_E depends on the number of variables in the model, which changes from step to step, a fixed value of f_{in} and f_{out} causes the type I and type II error rates to vary. Some computer programs allow the analyst to specify the type I error levels for f_{in} and f_{out}. However, the "advertised" significance level is not the true level, because the variable selected is the one that maximizes (or minimizes) the partial F-statistic at that stage. Sometimes it is useful to experiment with different values of f_{in} and f_{out} (or different advertised type I error rates) in several different runs to see if this substantially affects the choice of the final model.

EXAMPLE 12-15 Wine Quality Stepwise Regression

Table 12-19 gives the Minitab stepwise regression output for the wine quality data. Minitab uses fixed values of α for entering and removing variables. The default level is $\alpha = 0.15$ for both decisions. The output in Table 12-19 uses the default value. Notice that the variables were entered in the order Flavor (step 1), Oakiness (step 2), and Aroma (step 3) and that no variables were removed. No other variable could be entered, so the algorithm terminated. This is the three-variable model found by all possible regressions that results in a minimum value of C_p.

Table 12-19 Minitab Stepwise Regression Output for the Wine Quality Data

```
Stepwise Regression: Quality versus Clarity, Aroma, . . .
Alpha-to-Enter: 0.15 Alpha-to-Remove: 0.15

Response is Quality on 5 predictors, with N = 38

    Step            1          2          3
Constant        4.941      6.912      6.467

Flavor           1.57       1.64       1.20
T-Value          7.73       8.25       4.36
P-Value         0.000      0.000      0.000

Oakiness                   -0.54      -0.60
T-Value                    -1.95      -2.28
P-Value                    0.059      0.029

Aroma                                  0.58
T-Value                                2.21
P-Value                               0.034

S                1.27       1.22       1.16
R-Sq            62.42      66.11      70.38
R-Sq(adj)       61.37      64.17      67.76
C–p               9.0        6.8        3.9
```

Forward Selection

The **forward selection** procedure is a variation of stepwise regression and is based on the principle that regressors should be added to the model one at a time until there are no remaining candidate regressors that produce a significant increase in the regression sum of squares. That is, variables are added one at a time as long as their partial F-value exceeds f_{in}. Forward selection is a simplification of stepwise regression that omits the partial F-test for deleting variables from the model that have been added at previous steps. This is a potential weakness of forward selection; that is, the procedure does not explore the effect that adding a regressor at the current step has on regressor variables added at earlier steps. Notice that if we were to apply forward selection to the wine quality data, we would obtain exactly the same results as we did with stepwise regression in Example 12-15, since stepwise regression terminated without deleting a variable.

Backward Elimination

The **backward elimination** algorithm begins with all K candidate regressors in the model. Then the regressor with the smallest partial F-statistic is deleted if this F-statistic is insignificant, that is, if $f < f_{\text{out}}$. Next, the model with $K - 1$ regressors is fit, and the next regressor for potential elimination is found. The algorithm terminates when no further regressor can be deleted.

Table 12-20 shows the Minitab output for backward elimination applied to the wine quality data. The α value for removing a variable is $\alpha = 0.10$. Notice that this procedure removes Body at step 1 and then Clarity at step 2, terminating with the three-variable model found previously.

Some Comments on Final Model Selection

We have illustrated several different approaches to the selection of variables in multiple linear regression. The final model obtained from any model-building procedure should be subjected

Table 12-20 Minitab Backward Elimination Output for the Wine Quality Data

Stepwise Regression: Quality versus Clarity, Aroma, . . .

Backward elimination. Alpha-to-Remove: 0.1

Response is Quality on 5 predictors, with N = 38

Step	1	2	3
Constant	3.997	4.986	6.467
Clarity	2.3	1.8	
T-Value	1.35	1.12	
P-Value	0.187	0.269	
Aroma	0.48	0.53	0.58
T-Value	1.77	2.00	2.21
P-Value	0.086	0.054	0.034
Body	0.27		
T-Value	0.82		
P-Value	0.418		
Flavor	1.17	1.26	1.20
T-Value	3.84	4.52	4.36
P-Value	0.001	0.000	0.000
Oakiness	−0.68	−0.66	−0.60
T-Value	−2.52	−2.46	−2.28
P-Value	0.017	0.019	0.029
S	1.16	1.16	1.16
R-Sq	72.06	71.47	70.38
R-Sq(adj)	67.69	68.01	67.76
C–p	6.0	4.7	3.9

to the usual adequacy checks, such as residual analysis, lack-of-fit testing, and examination of the effects of influential points. The analyst may also consider augmenting the original set of candidate variables with cross-products, polynomial terms, or other transformations of the original variables that might improve the model. A major criticism of variable selection methods such as stepwise regression is that the analyst may conclude there is one "best" regression equation. Generally, this is not the case, because several equally good regression models can often be used. One way to avoid this problem is to use several different model-building techniques and see if different models result. For example, we have found the same model for the wine quality data using stepwise regression, forward selection, and backward elimination. The same model was also one of the two best found from all possible regressions. The results from variable selection methods frequently do not agree, so this is a good indication that the three-variable model is the best regression equation.

If the number of candidate regressors is not too large, the all-possible regressions method is recommended. We usually recommend using the minimum MS_E and C_p evaluation criteria in conjunction with this procedure. The all-possible regressions approach can find the "best" regression equation with respect to these criteria, while stepwise-type methods offer no such assurance. Furthermore, the all-possible regressions procedure is not distorted by dependencies among the regressors, as stepwise-type methods are.

12-6.4 Multicollinearity

In multiple regression problems, we expect to find dependencies between the response variable Y and the regressors x_j. In most regression problems, however, we find that there are also dependencies among the regressor variables x_j. In situations where these dependencies are strong, we say that **multicollinearity** exists. Multicollinearity can have serious effects on the estimates of the regression coefficients and on the general applicability of the estimated model.

The effects of multicollinearity may be easily demonstrated. The diagonal elements of the matrix $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-1}$ can be written as

$$C_{jj} = \frac{1}{(1 - R_j^2)} \qquad j = 1, 2, \ldots, k$$

where R_j^2 is the coefficient of multiple determination resulting from regressing x_j on the other $k - 1$ regressor variables. We can think of R_j^2 as a measure of the correlation between x_j and the other regressors. Clearly, the stronger the linear dependency of x_j on the remaining regressor variables, and hence the stronger the multicollinearity, the larger the value of R_j^2 will be. Recall that $V(\hat{\beta}_j) = \sigma^2 C_{jj}$. Therefore, we say that the variance of $\hat{\beta}_j$ is "inflated" by the quantity $(1 - R_j^2)^{-1}$. Consequently, we define the **variance inflation factor** for β_j as

Variance Inflation Factor (VIF)

$$VIF(\beta_j) = \frac{1}{(1 - R_j^2)} \qquad j = 1, 2, \ldots, k \qquad (12\text{-}51)$$

These factors are an important measure of the extent to which multicollinearity is present. If the columns of the model matrix **X** are **orthogonal,** then the regressors are completely uncorrelated, and the variance inflation factors will all be unity. So any VIF that exceeds one indicates some level of multicollinearity in the data.

Although the estimates of the regression coefficients are very imprecise when multicollinearity is present, the fitted model equation may still be useful. For example, suppose we wish to predict new observations on the response. If these predictions are interpolations in the original region of the x-space where the multicollinearity is in effect, satisfactory predictions will often be obtained, because while individual β_j may be poorly estimated, the function $\sum_{j=1}^{k} \beta_j x_{ij}$ may be estimated quite well. On the other hand, if the prediction of new observations requires extrapolation beyond the original region of the x-space where the data were collected, generally we would expect to obtain poor results. Extrapolation usually requires good estimates of the individual model parameters.

Multicollinearity arises for several reasons. It will occur when the analyst collects data such that a linear constraint holds approximately among the columns of the **X** matrix. For example, if four regressor variables are the components of a mixture, such a constraint will always exist because the sum of the components is always constant. Usually, these constraints do not hold exactly, and the analyst might not know that they exist.

The presence of multicollinearity can be detected in several ways. Two of the more easily understood of these will be discussed briefly.

1. The **variance inflation factors,** defined in Equation 12-51, are very useful measures of multicollinearity. The larger the variance inflation factor, the more severe the multicollinearity. Some authors have suggested that if any variance inflation factor exceeds 10, multicollinearity is a problem. Other authors consider this value too liberal and suggest that the variance inflation factors should not exceed 4 or 5. Minitab will calculate the variance inflation factors. Table 12-4 presents the Minitab

multiple regression output for the wire bond pull strength data. Since both VIF_1 and VIF_2 are small, there is no problem with multicollinearity.

2. If the F-test for significance of regression is significant, but tests on the individual regression coefficients are not significant, multicollinearity may be present.

Several remedial measures have been proposed for solving the problem of multicollinearity. Augmenting the data with new observations specifically designed to break up the approximate linear dependencies that currently exist is often suggested. However, this is sometimes impossible because of economic reasons or because of the physical constraints that relate the x_j. Another possibility is to delete certain variables from the model, but this approach has the disadvantage of discarding the information contained in the deleted variables.

Since multicollinearity primarily affects the stability of the regression coefficients, it would seem that estimating these parameters by some method that is less sensitive to multicollinearity than ordinary least squares would be helpful. Several methods have been suggested. One alternative to ordinary least squares, **ridge regression,** can be useful in combating multicollinearity. For more details on ridge regression, there are more extensive presentations in Montgomery, Peck, and Vining (2006) and Myers (1990).

EXERCISES FOR SECTION 12-6

12-72. An article entitled "A Method for Improving the Accuracy of Polynomial Regression Analysis" in the *Journal of Quality Technology* (1971, pp. 149–155) reported the following data on y = ultimate shear strength of a rubber compound (psi) and x = cure temperature (°F).

y	770	800	840	810
x	280	284	292	295
y	735	640	590	560
x	298	305	308	315

(a) Fit a second-order polynomial to these data.
(b) Test for significance of regression using $\alpha = 0.05$.
(c) Test the hypothesis that $\beta_{11} = 0$ using $\alpha = 0.05$.
(d) Compute the residuals from part (a) and use them to evaluate model adequacy.

12-73. Consider the following data, which result from an experiment to determine the effect of x = test time in hours at a particular temperature on y = change in oil viscosity:
(a) Fit a second-order polynomial to the data.

y	-1.42	-1.39	-1.55	-1.89	-2.43
x	.25	.50	.75	1.00	1.25
y	-3.15	-4.05	-5.15	-6.43	-7.89
x	1.50	1.75	2.00	2.25	2.50

(b) Test for significance of regression using $\alpha = 0.05$.
(c) Test the hypothesis that $\beta_{11} = 0$ using $\alpha = 0.05$.
(d) Compute the residuals from part (a) and use them to evaluate model adequacy.

12-74. The following data were collected during an experiment to determine the change in thrust efficiency (y, in percent) as the divergence angle of a rocket nozzle (x) changes:

y	24.60	24.71	23.90	39.50	39.60	57.12
x	4.0	4.0	4.0	5.0	5.0	6.0
y	67.11	67.24	67.15	77.87	80.11	84.67
x	6.5	6.5	6.75	7.0	7.1	7.3

(a) Fit a second-order model to the data.
(b) Test for significance of regression and lack of fit using $\alpha = 0.05$.
(c) Test the hypothesis that $\beta_{11} = 0$, using $\alpha = 0.05$.
(d) Plot the residuals and comment on model adequacy.
(e) Fit a cubic model, and test for the significance of the cubic term using $\alpha = 0.05$.

12-75. An article in the *Journal of Pharmaceuticals Sciences* (Vol. 80, 1991, pp. 971–977) presents data on the observed mole fraction solubility of a solute at a constant temperature and the dispersion, dipolar, and hydrogen bonding Hansen partial solubility parameters. The data are as shown in the following table, where y is the negative logarithm of the mole fraction solubility, x_1 is the dispersion partial solubility, x_2 is the dipolar partial solubility, and x_3 is the hydrogen bonding partial solubility.
(a) Fit the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12}x_1x_2 + \beta_{13}x_1x_3 + \beta_{23}x_2x_3 + \beta_{11}x_1^2 + \beta_{22}x_2^2 + \beta_{33}x_3^2 + \epsilon$.
(b) Test for significance of regression using $\alpha = 0.05$.
(c) Plot the residuals and comment on model adequacy.
(d) Use the extra sum of squares method to test the contribution of the second-order terms using $\alpha = 0.05$.

Observation Number	y	x_1	x_2	x_3
1	0.22200	7.3	0.0	0.0
2	0.39500	8.7	0.0	0.3
3	0.42200	8.8	0.7	1.0
4	0.43700	8.1	4.0	0.2
5	0.42800	9.0	0.5	1.0
6	0.46700	8.7	1.5	2.8
7	0.44400	9.3	2.1	1.0
8	0.37800	7.6	5.1	3.4
9	0.49400	10.0	0.0	0.3
10	0.45600	8.4	3.7	4.1
11	0.45200	9.3	3.6	2.0
12	0.11200	7.7	2.8	7.1
13	0.43200	9.8	4.2	2.0
14	0.10100	7.3	2.5	6.8
15	0.23200	8.5	2.0	6.6
16	0.30600	9.5	2.5	5.0
17	0.09230	7.4	2.8	7.8
18	0.11600	7.8	2.8	7.7
19	0.07640	7.7	3.0	8.0
20	0.43900	10.3	1.7	4.2
21	0.09440	7.8	3.3	8.5
22	0.11700	7.1	3.9	6.6
23	0.07260	7.7	4.3	9.5
24	0.04120	7.4	6.0	10.9
25	0.25100	7.3	2.0	5.2
26	0.00002	7.6	7.8	20.7

12-76. Consider the arsenic concentration data in Exercise 12-10.

(a) Discuss how you would model the information about the person's sex.

(b) Fit a regression model to the arsenic in nails using age, drink use, cook use, and the person's sex as the regressors.

(c) Is there evidence that the person's sex affects arsenic in the nails? Why?

12-77. Consider the gasoline mileage data in Exercise 12-7.

(a) Discuss how you would model the information about the type of transmission in the car.

(b) Fit a regression model to the gasoline mileage using *cid*, *etw*, and the type of transmission in the car as the regressors.

(c) Is there evidence that the type of transmission (L4, L5, or M6) affects gasoline mileage performance?

12-78. Consider the surface finish data in Example 12-15. Test the hypothesis that two different regression models (with different slopes and intercepts) are required to adequately model the data. Use indicator variables in answering this question.

12-79. Consider the X-ray inspection data in Exercise 12-11. Use rads as the response. Build regression models for the data using the following techniques:

(a) All possible regressions.

(b) Stepwise regression.

(c) Forward selection.

(d) Backward elimination.

(e) Comment on the models obtained. Which model would you prefer? Why?

12-80. Consider the electric power data in Exercise 12-6. Build regression models for the data using the following techniques:

(a) All possible regressions. Find the minimum C_p and minimum MS_E equations.

(b) Stepwise regression.

(c) Forward selection.

(d) Backward elimination.

(e) Comment on the models obtained. Which model would you prefer?

12-81. Consider the regression model fit to the coal and limestone mixture data in Exercise 12-13. Use density as the response. Build regression models for the data using the following techniques:

(a) All possible regressions.

(b) Stepwise regression.

(c) Forward selection.

(d) Backward elimination.

(e) Comment on the models obtained. Which model would you prefer? Why?

12-82. Consider the wire bond pull strength data in Exercise 12-8. Build regression models for the data using the following methods:

(a) All possible regressions. Find the minimum C_p and minimum MS_E equations.

(b) Stepwise regression.

(c) Forward selection.

(d) Backward elimination.

(e) Comment on the models obtained. Which model would you prefer?

12-83. Consider the grey range modulation data in Exercise 12-15. Use the useful range as the response. Build regression models for the data using the following techniques:

(a) All possible regressions.

(b) Stepwise regression.

(c) Forward selection.

(d) Backward elimination.

(e) Comment on the models obtained. Which model would you prefer? Why?

12-84. Consider the nisin extraction data in Exercise 12-14. Build regression models for the data using the following techniques:

(a) All possible regressions.

(b) Stepwise regression.

(c) Forward selection.

(d) Backward elimination.
(e) Comment on the models obtained. Which model would you prefer? Why?

12.85. Consider the stack loss data in Exercise 12-16. Build regression models for the data using the following techniques:
(a) All possible regressions.
(b) Stepwise regression.
(c) Forward selection.
(d) Backward elimination.
(e) Comment on the models obtained. Which model would you prefer? Why?
(f) Remove any influential data points and repeat the model building in the previous parts? Does your conclusion in part (e) change?

12-86. Consider the NHL data in Exercise 12-18. Build regression models for these data with regressors *GF* through *FG* using the following methods:
(a) All possible regressions. Find the minimum C_p and minimum MS_E equations.
(b) Stepwise regression.
(c) Forward selection.
(d) Backward elimination.
(e) Which model would you prefer?

12-87. Use the football data in Exercise 12-17 to build regression models using the following techniques:
(a) All possible regressions. Find the equations that minimize MS_E and that minimize C_p.
(b) Stepwise regression.
(c) Forward selection.
(d) Backward elimination.
(e) Comment on the various models obtained. Which model seems "best," and why?

12-88. Consider the arsenic data in Exercise 12-12. Use arsenic in nails as the response and age, drink use, and cook use as the regressors. Build regression models for the data using the following techniques:
(a) All possible regressions.
(b) Stepwise regression.
(c) Forward selection.
(d) Backward elimination.
(e) Comment on the models obtained. Which model would you prefer? Why?
(f) Now construct an indicator variable and add the person's sex to the list of regressors. Repeat the model building in the previous parts. Does your conclusion in part (e) change?

12-89. Consider the gas mileage data in Exercise 12-7. Build regression models for the data from the numerical regressors using the following techniques:
(a) All possible regressions.
(b) Stepwise regression.
(c) Forward selection.
(d) Backward elimination.
(e) Comment on the models obtained. Which model would you prefer? Why?
(f) Now construct indicator variable for *trns* and *drv* and add these to the list of regressors. Repeat the model building in the previous parts. Does your conclusion in part (e) change?

12-90. When fitting polynomial regression models, we often subtract $\bar{x}$ from each x value to produce a "centered" regressor $x' = x - \bar{x}$. This reduces the effects of dependencies among the model terms and often leads to more accurate estimates of the regression coefficients. Using the data from Exercise 12-72, fit the model $Y = \beta_0^* + \beta_1^* x' + \beta_{11}^*(x')^2 + \epsilon$.
(a) Use the results to estimate the coefficients in the uncentered model $Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$. Predict y when $x = 285$°F. Suppose that we use a standardized variable $x' = (x - \bar{x})/s_x$, where s_x is the standard deviation of x, in constructing a polynomial regression model. Fit the model $Y = \beta_0^* + \beta_1^* x' + \beta_{11}^*(x')^2 + \epsilon$.
(b) What value of y do you predict when $x = 285$°F?
(c) Estimate the regression coefficients in the unstandardized model $Y = \beta_0 + \beta_1 x + \beta_{11} x^2 + \epsilon$.
(d) What can you say about the relationship between SS_E and R^2 for the standardized and unstandardized models?
(e) Suppose that $y' = (y - \bar{y})/s_y$ is used in the model along with x'. Fit the model and comment on the relationship between SS_E and R^2 in the standardized model and the unstandardized model.

12-91. Consider the data in Exercise 12-75. Use all the terms in the full quadratic model as the candidate regressors.
(a) Use forward selection to identify a model.
(b) Use backward elimination to identify a model.
(c) Compare the two models obtained in parts (a) and (b). Which model would you prefer and why?

12-92. We have used a sample of 30 observations to fit a regression model. The full model has nine regressors, the variance estimate is $\hat{\sigma}^2 = MS_E = 100$, and $R^2 = 0.92$.
(a) Calculate the F-statistic for testing significance of regression. Using $\alpha = 0.05$, what would you conclude?
(b) Suppose that we fit another model using only four of the original regressors and that the error sum of squares for this new model is 2200. Find the estimate of σ^2 for this new reduced model. Would you conclude that the reduced model is superior to the old one? Why?
(c) Find the value of C_p for the reduced model in part (b). Would you conclude that the reduced model is better than the old model?

12-93. A sample of 25 observations is used to fit a regression model in seven variables. The estimate of σ^2 for this full model is $MS_E = 10$.
(a) A forward selection algorithm has put three of the original seven regressors in the model. The error sum of squares for the three-variable model is $SS_E = 300$. Based on C_p, would you conclude that the three-variable model has any remaining bias?

(b) After looking at the forward selection model in part (a), suppose you could add one more regressor to the model. This regressor will reduce the error sum of squares to 275. Will the addition of this variable improve the model? Why?

Supplemental Exercises

12-94. Consider the computer output below.

```
The regression equation is
Y = 517 + 11.5 x1 − 8.14 x2 + 10.9 x3

Predictor        Coef     SE Coef        T      P
Constant       517.46       11.76        ?      ?
x1            11.4720           ?    36.50      ?
x2            −8.1378      0.1969        ?      ?
x3            10.8565      0.6652        ?      ?

S = 10.2560     R–Sq = ?     R–Sq (adj) = ?

Analysis of Variance

Source            DF        SS        MS      F      P
Regression         ?    347300    115767      ?      ?
Residual Error    16         ?       105
Total             19    348983
```

(a) Fill in the missing values. Use bounds for the *P*-values.

(b) Is the overall model significant at $\alpha = 0.05$? Is it significant at $\alpha = 0.01$?

(c) Discuss the contribution of the individual regressors to the model.

12-95. Consider the following inverse of the model matrix:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.893758 & -0.028245 & -0.0175641 \\ -0.028245 & 0.0013329 & 0.0001547 \\ -0.017564 & 0.0001547 & 0.0009108 \end{bmatrix}$$

(a) How many variables are in the regression model?

(b) If the estimate of σ^2 is 50, what is the estimate of the variance of each regression coefficient?

(c) What is the standard error of the intercept?

12-96. The data shown in Table 12-22 represent the thrust of a jet-turbine engine (y) and six candidate regressors: x_1 = primary speed of rotation, x_2 = secondary speed of rotation, x_3 = fuel flow rate, x_4 = pressure, x_5 = exhaust temperature, and x_6 = ambient temperature at time of test.

(a) Fit a multiple linear regression model using x_3 = fuel flow rate, x_4 = pressure, and x_5 = exhaust temperature as the regressors.

(b) Test for significance of regression using $\alpha = 0.01$. Find the *P*-value for this test. What are your conclusions?

(c) Find the *t*-test statistic for each regressor. Using $\alpha = 0.01$, explain carefully the conclusion you can draw from these statistics.

(d) Find R^2 and the adjusted statistic for this model.

(e) Construct a normal probability plot of the residuals and interpret this graph.

(f) Plot the residuals versus $\hat{y}$. Are there any indications of inequality of variance or nonlinearity?

(g) Plot the residuals versus x_3. Is there any indication of nonlinearity?

(h) Predict the thrust for an engine for which $x_3 = 28900$, $x_4 = 170$, and $x_5 = 1589$.

12-97. Consider the engine thrust data in Exercise 12-96. Refit the model using $y^* = \ln y$ as the response variable and $x_3^* = \ln x_3$ as the regressor (along with x_4 and x_5).

(a) Test for significance of regression using $\alpha = 0.01$. Find the *P*-value for this test and state your conclusions.

(b) Use the *t*-statistic to test H_0: $\beta_j = 0$ versus H_1: $\beta_j \neq 0$ for each variable in the model. If $\alpha = 0.01$, what conclusions can you draw?

(c) Plot the residuals versus $\hat{y}^*$ and versus x_3^*. Comment on these plots. How do they compare with their counterparts obtained in Exercise 12-96 parts (f) and (g)?

12-98. Transient points of an electronic inverter are influenced by many factors. Table 12-21 gives data on the transient point (y, in volts) of PMOS-NMOS inverters and five candidate regressors: x_1 = width of the NMOS device, x_2 = length

Table 12-21 Transient Point of an Electronic Inverter

Observation Number	x_1	x_2	x_3	x_4	x_5	y
1	3	3	3	3	0	0.787
2	8	30	8	8	0	0.293
3	3	6	6	6	0	1.710
4	4	4	4	12	0	0.203
5	8	7	6	5	0	0.806
6	10	20	5	5	0	4.713
7	8	6	3	3	25	0.607
8	6	24	4	4	25	9.107
9	4	10	12	4	25	9.210
10	16	12	8	4	25	1.365
11	3	10	8	8	25	4.554
12	8	3	3	3	25	0.293
13	3	6	3	3	50	2.252
14	3	8	8	3	50	9.167
15	4	8	4	8	50	0.694
16	5	2	2	2	50	0.379
17	2	2	2	3	50	0.485
18	10	15	3	3	50	3.345
19	15	6	2	3	50	0.208
20	15	6	2	3	75	0.201
21	10	4	3	3	75	0.329
22	3	8	2	2	75	4.966
23	6	6	6	4	75	1.362
24	2	3	8	6	75	1.515
25	3	3	8	8	75	0.751

Table 12-22 Thrust of a Jet-Turbine Engine

Observation Number	y	x_1	x_2	x_3	x_4	x_5	x_6
1	4540	2140	20640	30250	205	1732	99
2	4315	2016	20280	30010	195	1697	100
3	4095	1905	19860	29780	184	1662	97
4	3650	1675	18980	29330	164	1598	97
5	3200	1474	18100	28960	144	1541	97
6	4833	2239	20740	30083	216	1709	87
7	4617	2120	20305	29831	206	1669	87
8	4340	1990	19961	29604	196	1640	87
9	3820	1702	18916	29088	171	1572	85
10	3368	1487	18012	28675	149	1522	85
11	4445	2107	20520	30120	195	1740	101
12	4188	1973	20130	29920	190	1711	100
13	3981	1864	19780	29720	180	1682	100
14	3622	1674	19020	29370	161	1630	100
15	3125	1440	18030	28940	139	1572	101
16	4560	2165	20680	30160	208	1704	98
17	4340	2048	20340	29960	199	1679	96
18	4115	1916	19860	29710	187	1642	94
19	3630	1658	18950	29250	164	1576	94
20	3210	1489	18700	28890	145	1528	94
21	4330	2062	20500	30190	193	1748	101
22	4119	1929	20050	29960	183	1713	100
23	3891	1815	19680	29770	173	1684	100
24	3467	1595	18890	29360	153	1624	99
25	3045	1400	17870	28960	134	1569	100
26	4411	2047	20540	30160	193	1746	99
27	4203	1935	20160	29940	184	1714	99
28	3968	1807	19750	29760	173	1679	99
29	3531	1591	18890	29350	153	1621	99
30	3074	1388	17870	28910	133	1561	99
31	4350	2071	20460	30180	198	1729	102
32	4128	1944	20010	29940	186	1692	101
33	3940	1831	19640	29750	178	1667	101
34	3480	1612	18710	29360	156	1609	101
35	3064	1410	17780	28900	136	1552	101
36	4402	2066	20520	30170	197	1758	100
37	4180	1954	20150	29950	188	1729	99
38	3973	1835	19750	29740	178	1690	99
39	3530	1616	18850	29320	156	1616	99
40	3080	1407	17910	28910	137	1569	100

of the NMOS device, x_3 = width of the PMOS device, x_4 = length of the PMOS device, and x_5 = temperature (°C).

(a) Fit a multiple linear regression model that uses all regressors to these data. Test for significance of regression using $\alpha = 0.01$. Find the *P*-value for this test and use it to draw your conclusions.

(b) Test the contribution of each variable to the model using the *t*-test with $\alpha = 0.05$. What are your conclusions?

(c) Delete x_5 from the model. Test the new model for significance of regression. Also test the relative contribution of each regressor to the new model with the *t*-test. Using $\alpha = 0.05$, what are your conclusions?

(d) Notice that the MS_E for the model in part (c) is smaller than the MS_E for the full model in part (a). Explain why this has occurred.

(e) Calculate the studentized residuals. Do any of these seem unusually large?

(f) Suppose that you learn that the second observation was recorded incorrectly. Delete this observation and refit the model using x_1, x_2, x_3, and x_4 as the regressors. Notice that the R^2 for this model is considerably higher than the R^2 for either of the models fitted previously. Explain why the R^2 for this model has increased.

(g) Test the model from part (f) for significance of regression using $\alpha = 0.05$. Also investigate the contribution of each regressor to the model using the *t*-test with $\alpha = 0.05$. What conclusions can you draw?

(h) Plot the residuals from the model in part (f) versus $\hat{y}$ and versus each of the regressors x_1, x_2, x_3, and x_4. Comment on the plots.

12-99. Consider the inverter data in Exercise 12-98. Delete observation 2 from the original data. Define new variables as follows: $y^* = \ln y$, $x_1^* = 1/\sqrt{x_1}$, $x_2^* = \sqrt{x_2}$, $x_3^* = 1/\sqrt{x_3}$, and $x_4^* = \sqrt{x_4}$.

(a) Fit a regression model using these transformed regressors (do not use x_5).

(b) Test the model for significance of regression using $\alpha = 0.05$. Use the *t*-test to investigate the contribution of each variable to the model ($\alpha = 0.05$). What are your conclusions?

(c) Plot the residuals versus $\hat{y}^*$ and versus each of the transformed regressors. Comment on the plots.

12-100. Following are data on y = green liquor (g/l) and x = paper machine speed (feet per minute) from a Kraft paper machine. (The data were read from a graph in an article in the *Tappi Journal,* March 1986.)

y	16.0	15.8	15.6	15.5	14.8
x	1700	1720	1730	1740	1750
y	14.0	13.5	13.0	12.0	11.0
x	1760	1770	1780	1790	1795

(a) Fit the model $Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$ using least squares.

(b) Test for significance of regression using $\alpha = 0.05$. What are your conclusions?

(c) Test the contribution of the quadratic term to the model, over the contribution of the linear term, using an *F*-statistic. If $\alpha = 0.05$, what conclusion can you draw?

(d) Plot the residuals from the model in part (a) versus $\hat{y}$. Does the plot reveal any inadequacies?

(e) Construct a normal probability plot of the residuals. Comment on the normality assumption.

12-101. Consider the jet engine thrust data in Exercise 12-96 and 12-97. Define the response and regressors as in Exercise 12-97.

(a) Use all possible regressions to select the best regression equation, where the model with the minimum value of MS_E is to be selected as "best."

(b) Repeat part (a) using the C_P criterion to identify the best equation.

(c) Use stepwise regression to select a subset regression model.

(d) Compare the models obtained in parts (a), (b), and (c) above.

(e) Consider the three-variable regression model. Calculate the variance inflation factors for this model. Would you conclude that multicollinearity is a problem in this model?

12-102. Consider the electronic inverter data in Exercise 12-98 and 12-99. Define the response and regressors variables as in Exercise 12-99, and delete the second observation in the sample.

(a) Use all possible regressions to find the equation that minimizes C_p.

(b) Use all possible regressions to find the equation that minimizes MS_E.

(c) Use stepwise regression to select a subset regression model.

(d) Compare the models you have obtained.

12-103. A multiple regression model was used to relate y = viscosity of a chemical product to x_1 = temperature and x_2 = reaction time. The data set consisted of $n = 15$ observations.

(a) The estimated regression coefficients were $\hat{\beta}_0 = 300.00$, $\hat{\beta}_1 = 0.85$, and $\hat{\beta}_2 = 10.40$. Calculate an estimate of mean viscosity when $x_1 = 100$°F and $x_2 = 2$ hours.

(b) The sums of squares were $SS_T = 1230.50$ and $SS_E = 120.30$. Test for significance of regression using $\alpha = 0.05$. What conclusion can you draw?

(c) What proportion of total variability in viscosity is accounted for by the variables in this model?

(d) Suppose that another regressor, x_3 = stirring rate, is added to the model. The new value of the error sum of squares is $SS_E = 117.20$. Has adding the new variable resulted in a smaller value of MS_E? Discuss the significance of this result.

(e) Calculate an *F*-statistic to assess the contribution of x_3 to the model. Using $\alpha = 0.05$, what conclusions do you reach?

12-104. Tables 12-23 and 12-24 present statistics for the Major League Baseball 2005 season (source: *The Sports Network*).

(a) Consider the batting data. Use model-building methods to predict *Wins* from the other variables. Check that the assumptions for your model are valid.

Table 12-23 Major League Baseball 2005 Season

	American League Batting													
Team	W	AVG	R	H	2B	3B	HR	RBI	BB	SO	SB	GIDP	LOB	OBP
Chicago	99	0.262	741	1450	253	23	200	713	435	1002	137	122	1032	0.322
Boston	95	0.281	910	1579	339	21	199	863	653	1044	45	135	1249	0.357
LA Angels	95	0.27	761	1520	278	30	147	726	447	848	161	125	1086	0.325
New York	95	0.276	886	1552	259	16	229	847	637	989	84	125	1264	0.355
Cleveland	93	0.271	790	1522	337	30	207	760	503	1093	62	128	1148	0.334
Oakland	88	0.262	772	1476	310	20	155	739	537	819	31	148	1170	0.33
Minnesota	83	0.259	688	1441	269	32	134	644	485	978	102	155	1109	0.323
Toronto	80	0.265	775	1480	307	39	136	735	486	955	72	126	1118	0.331
Texas	79	0.267	865	1528	311	29	260	834	495	1112	67	123	1104	0.329
Baltimore	74	0.269	729	1492	296	27	189	700	447	902	83	145	1103	0.327
Detroit	71	0.272	723	1521	283	45	168	678	384	1038	66	137	1077	0.321
Seattle	69	0.256	699	1408	289	34	130	657	466	986	102	115	1076	0.317
Tampa Bay	67	0.274	750	1519	289	40	157	717	412	990	151	133	1065	0.329
Kansas City	56	0.263	701	1445	289	34	126	653	424	1008	53	139	1062	0.32

	National League Batting													
Team	W	AVG	R	H	2B	3B	HR	RBI	BB	SO	SB	GIDP	LOB	OBP
St. Louis	100	0.27	805	1494	287	26	170	757	534	947	83	127	1152	0.339
Atlanta	90	0.265	769	1453	308	37	184	733	534	1084	92	146	1114	0.333
Houston	89	0.256	693	1400	281	32	161	654	481	1037	115	116	1136	0.322
Philadelphia	88	0.27	807	1494	282	35	167	760	639	1083	116	107	1251	0.348
Florida	83	0.272	717	1499	306	32	128	678	512	918	96	144	1181	0.339
New York	83	0.258	722	1421	279	32	175	683	486	1075	153	103	1122	0.322
San Diego	82	0.257	684	1416	269	39	130	655	600	977	99	122	1220	0.333
Milwaukee	81	0.259	726	1413	327	19	175	689	531	1162	79	137	1120	0.331
Washington	81	0.252	639	1367	311	32	117	615	491	1090	45	130	1137	0.322
Chicago	79	0.27	703	1506	323	23	194	674	419	920	65	131	1133	0.324
Arizona	77	0.256	696	1419	291	27	191	670	606	1094	67	132	1247	0.332
San Francisco	75	0.261	649	1427	299	26	128	617	431	901	71	147	1093	0.319
Cincinnati	73	0.261	820	1453	335	15	222	784	611	1303	72	116	1176	0.339
Los Angeles	71	0.253	685	1374	284	21	149	653	541	1094	58	139	1135	0.326
Colorado	67	0.267	740	1477	280	34	150	704	509	1103	65	125	1197	0.333
Pittsburgh	67	0.259	680	1445	292	38	139	656	471	1092	73	130	1193	0.322

Batting	
W	Wins
AVG	Batting average
R	Runs
H	Hits
2B	Doubles
3B	Triples
HR	Home runs
RBI	Runs batted in
BB	Walks
SO	Strikeouts
SB	Stolen bases
GIDP	Grounded into double play
LOB	Left on base
OBP	On-base percentage

Pitching	
ERA	Earned run average
SV	Saves
H	Hits
R	Runs
ER	Earned runs
HR	Home runs
BB	Walks
SO	Strikeouts
AVG	Opponent batting average

Table 12-24 Major League Baseball 2005 Season

	American League Pitching									
Team	W	ERA	SV	H	R	ER	HR	BB	SO	AVG
Chicago	99	3.61	54	1392	645	592	167	459	1040	0.249
Boston	95	4.74	38	1550	805	752	164	440	959	0.276
LA Angels	95	3.68	54	1419	643	598	158	443	1126	0.254
New York	95	4.52	46	1495	789	718	164	463	985	0.269
Cleveland	93	3.61	51	1363	642	582	157	413	1050	0.247
Oakland	88	3.69	38	1315	658	594	154	504	1075	0.241
Minnesota	83	3.71	44	1458	662	604	169	348	965	0.261
Toronto	80	4.06	35	1475	705	653	185	444	958	0.264
Texas	79	4.96	46	1589	858	794	159	522	932	0.279
Baltimore	74	4.56	38	1458	800	724	180	580	1052	0263
Detroit	71	4.51	37	1504	787	719	193	461	907	0.272
Seattle	69	4.49	39	1483	751	712	179	496	892	0.268
Tampa Bay	67	5.39	43	1570	936	851	194	615	949	0.28
Kansas City	56	5.49	25	1640	935	862	178	580	924	0.291

	National League Pitching									
Team	W	ERA	SV	H	R	ER	HR	BB	SO	AVG
St. Louis	100	3.49	48	1399	634	560	153	443	974	0.257
Atlanta	90	3.98	38	1487	674	639	145	520	929	0.268
Houston	89	3.51	45	1336	609	563	155	440	1164	0.246
Philadelphia	88	4.21	40	1379	726	672	189	487	1159	0.253
Florida	83	4.16	42	1459	732	666	116	563	1125	0.266
New York	83	3.76	38	1390	648	599	135	491	1012	0.255
San Diego	82	4.13	45	1452	726	668	146	503	1133	0.259
Milwaukee	81	3.97	46	1382	697	635	169	569	1173	0.251
Washington	81	3.87	51	1456	673	627	140	539	997	0.262
Chicago	79	4.19	39	1357	714	671	186	576	1256	0.25
Arizona	77	4.84	45	1580	856	783	193	537	1038	0.278
San Francisco	75	4.33	46	1456	745	695	151	592	972	0.263
Cincinnati	73	5.15	31	1657	889	820	219	492	955	0.29
Los Angeles	71	4.38	40	1434	755	695	182	471	1004	0.263
Colorado	67	5.13	37	1600	862	808	175	604	981	0.287
Pittsburgh	67	4.42	35	1456	769	706	162	612	958	0.267

Batting

W	Wins
AVG	Batting average
R	Runs
H	Hits
2B	Doubles
3B	Triples
HR	Home runs
RBI	Runs batted in
BB	Walks
SO	Strikeouts
SB	Stolen bases
GID	Grounded into double play
LOB	Left on base
OBP	On-base percentage

Pitching

ERA	Earned run average
SV	Saves
H	Hits
R	Runs
ER	Earned runs
HR	Home runs
BB	Walks
SO	Strikeouts
AVG	Opponent batting average

(b) Repeat part (a) for the pitching data.

(c) Use both the batting and pitching data to build a model to predict *Wins*. What variables are most important? Check that the assumptions for your model are valid.

12-105. An article in the *Journal of the American Ceramics Society* (1992, Vol. 75, pp. 112–116) describes a process for immobilizing chemical or nuclear wastes in soil by dissolving the contaminated soil into a glass block. The authors mix CaO and Na_2O with soil and model viscosity and electrical conductivity. The electrical conductivity model involves six regressors, and the sample consists of $n = 14$ observations.

(a) For the six-regressor model, suppose that $SS_T = 0.50$ and $R^2 = 0.94$. Find SS_E and SS_R, and use this information to test for significance of regression with $\alpha = 0.05$. What are your conclusions?

(b) Suppose that one of the original regressors is deleted from the model, resulting in $R^2 = 0.92$. What can you conclude about the contribution of the variable that was removed? Answer this question by calculating an F-statistic.

(c) Does deletion of the regressor variable in part (b) result in a smaller value of MS_E for the five-variable model, in comparison to the original six-variable model? Comment on the significance of your answer.

12-106. Exercise 12-5 introduced the hospital patient satisfaction survey data. One of the variables in that data set is a categorical variable indicating whether the patient is a medical patient or a surgical patient. Fit a model including this indicator variable to the data, using all three of the other regressors. Is there any evidence that the service the patient is on (medical versus surgical) has an impact on the reported satisfaction?

12-107. Consider the inverse model matrix shown below.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.125 & 0 & 0 & 0 \\ 0 & 0.125 & 0 & 0 \\ 0 & 0 & 0.125 & 0 \\ 0 & 0 & 0 & 0.125 \end{bmatrix}$$

(a) How many regressors are in this model?

(b) What was the sample size?

(c) Notice the special diagonal structure of the matrix. What does that tell you about the columns in the original $\mathbf{X}$ matrix?

MIND-EXPANDING EXERCISES

12-108. Consider a multiple regression model with k regressors. Show that the test statistic for significance of regression can be written as

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

Suppose that $n = 20$, $k = 4$, and $R^2 = 0.90$. If $\alpha = 0.05$, what conclusion would you draw about the relationship between y and the four regressors?

12-109. A regression model is used to relate a response y to $k = 4$ regressors with $n = 20$. What is the smallest value of R^2 that will result in a significant regression if $\alpha = 0.05$? Use the results of the previous exercise. Are you surprised by how small the value of R^2 is?

12-110. Show that we can express the residuals from a multiple regression model as $\mathbf{e} = (\mathbf{I} - \mathbf{H})\mathbf{y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

12-111. Show that the variance of the ith residual e_i in a multiple regression model is $\sigma^2(1 - h_{ii})$ and that the covariance between e_i and e_j is $-\sigma^2 h_{ij}$, where the h's are the elements of $\mathbf{H} = \mathbf{X}(\mathbf{X\,X})^{-1}\mathbf{X}'$.

12-112. Consider the multiple linear regression model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. If $\hat{\boldsymbol{\beta}}$ denotes the least squares estimator of $\boldsymbol{\beta}$, show that $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{R}\boldsymbol{\epsilon}$, where $\mathbf{R} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

12-113. Constrained Least Squares. Suppose we wish to find the least squares estimator of $\boldsymbol{\beta}$ in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ subject to a set of equality constraints, say, $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$.

(a) Show that the estimator is

$$\hat{\boldsymbol{\beta}}_{\mathbf{c}} = \hat{\boldsymbol{\beta}} + (\mathbf{X}'\mathbf{X})^{-1} \times \mathbf{T}'[\mathbf{T}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}']^{-1}(\mathbf{c} - \mathbf{T}\hat{\boldsymbol{\beta}})$$

where $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

(b) Discuss situations where this model might be appropriate.

12-114. Piecewise Linear Regression. Suppose that y is piecewise linearly related to x. That is, different linear relationships are appropriate over the intervals $-\infty < x \le x^*$ and $x^* < x < \infty$.

(a) Show how indicator variables can be used to fit such a piecewise linear regression model, assuming that the point x^* is known.

(b) Suppose that at the point x^* a discontinuity occurs in the regression function. Show how indicator variables can be used to incorporate the discontinuity into the model.

(c) Suppose that the point x^* is not known with certainty and must be estimated. Suggest an approach that could be used to fit the piecewise linear regression model.

IMPORTANT TERMS AND CONCEPTS

All possible regressions
Analysis of variance test in multiple regression
Categorical variables
Confidence interval on the mean response
Cp statistic
Extra sum of squares method
Hidden extrapolation
Indicator variables
Inference (test and intervals) on individual model parameters
Influential observations
Model parameters and their interpretation in multiple regression
Multicollinearity
Multiple Regression
Outliers
Polynomial regression model
Prediction interval on a future observation
PRESS statistic
Residual analysis and model adequacy checking
Significance of regression
Stepwise regression and related methods
Variance Inflation Factor (VIF)