REGRESSION

9.1 INTRODUCTION

Many engineering and scientific problems are concerned with determining a relationship between a set of variables. For instance, in a chemical process, we might be interested in the relationship between the output of the process, the temperature at which it occurs, and the amount of catalyst employed. Knowledge of such a relationship would enable us to predict the output for various values of temperature and amount of catalyst.

In many situations, there is a single *response* variable Y, also called the *dependent* variable, which depends on the value of a set of *input*, also called *independent*, variables $x_1, \ldots, x_r$. The simplest type of relationship between the dependent variable Y and the input variables $x_1, \ldots, x_r$ is a linear relationship. That is, for some constants $\beta_0, \beta_1, \ldots, \beta_r$ the equation

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r \tag{9.1.1}$$

would hold. If this was the relationship between Y and the x_i, $i = 1, \ldots, r$, then it would be possible (once the β_i were learned) to exactly predict the response for any set of input values. However, in practice, such precision is almost never attainable, and the most that one can expect is that Equation 9.1.1 would be valid *subject to random error*. By this we mean that the explicit relationship is

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r + e \tag{9.1.2}$$

where e, representing the random error, is assumed to be a random variable having mean 0. Indeed, another way of expressing Equation 9.1.2 is as follows:

$$E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_r x_r$$

where $\mathbf{x} = (x_1, \ldots, x_r)$ is the set of independent variables, and $E[Y|\mathbf{x}]$ is the expected response given the inputs $\mathbf{x}$.

Equation 9.1.2 is called a *linear regression equation*. We say that it describes the regression of Y on the set of independent variables $x_1, \ldots, x_r$. The quantities $\beta_0, \beta_1, \ldots, \beta_r$ are

called the *regression coefficients*, and must usually be estimated from a set of data. A regression equation containing a single independent variable — that is, one in which $r = 1$ — is called a *simple regression equation*, whereas one containing many independent variables is called a *multiple regression equation*.

Thus, a simple linear regression model supposes a linear relationship between the mean response and the value of a single independent variable. It can be expressed as

$$Y = \alpha + \beta x + e$$

where x is the value of the independent variable, also called the input level, Y is the response, and e, representing the random error, is a random variable having mean 0.

EXAMPLE 9.1a Consider the following 10 data pairs (x_i, y_i), $i = 1, \ldots, 10$, relating y, the percent yield of a laboratory experiment, to x, the temperature at which the experiment was run.

i	x_i	y_i	i	x_i	y_i
1	100	45	6	150	68
2	110	52	7	160	75
3	120	54	8	170	76
4	130	63	9	180	92
5	140	62	10	190	88

A plot of y_i versus x_i — called a *scatter diagram* — is given in Figure 9.1. As this scatter diagram appears to reflect, subject to random error, a linear relation between y and x, it seems that a simple linear regression model would be appropriate. ■

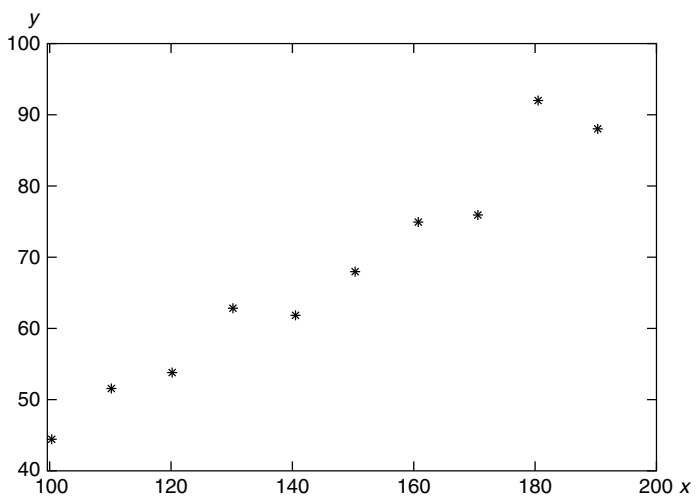

FIGURE 9.1 *Scatter plot.*

9.2 LEAST SQUARES ESTIMATORS OF THE REGRESSION PARAMETERS

Suppose that the responses Y_i corresponding to the input values x_i, $i = 1, \ldots, n$ are to be observed and used to estimate α and β in a simple linear regression model. To determine estimators of α and β we reason as follows: If A is the estimator of α and B of β, then the estimator of the response corresponding to the input variable x_i would be $A + Bx_i$. Since the actual response is Y_i, the squared difference is $(Y_i - A - Bx_i)^2$, and so if A and B are the estimators of α and β, then the sum of the squared differences between the estimated responses and the actual response values — call it SS — is given by

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

The method of least squares chooses as estimators of α and β the values of A and B that minimize SS. To determine these estimators, we differentiate SS first with respect to A and then to B as follows:

$$\frac{\partial SS}{\partial A} = -2 \sum_{i=1}^{n} (Y_i - A - Bx_i)$$

$$\frac{\partial SS}{\partial B} = -2 \sum_{i=1}^{n} x_i (Y_i - A - Bx_i)$$

Setting these partial derivatives equal to zero yields the following equations for the minimizing values A and B:

$$\sum_{i=1}^{n} Y_i = nA + B \sum_{i=1}^{n} x_i \tag{9.2.1}$$

$$\sum_{i=1}^{n} x_i Y_i = A \sum_{i=1}^{n} x_i + B \sum_{i=1}^{n} x_i^2$$

The Equations 9.2.1 are known as the *normal equations*. If we let

$$\overline{Y} = \sum_i Y_i / n, \qquad \overline{x} = \sum_i x_i / n$$

then we can write the first normal equation as

$$A = \overline{Y} - B\overline{x} \tag{9.2.2}$$

Substituting this value of A into the second normal equation yields

$$\sum_i x_i Y_i = (\overline{Y} - B\overline{x})n\overline{x} + B\sum_i x_i^2$$

or

$$B\left(\sum_i x_i^2 - n\overline{x}^2\right) = \sum_i x_i Y_i - n\overline{x}\,\overline{Y}$$

or

$$B = \frac{\sum_i x_i Y_i - n\overline{x}\,\overline{Y}}{\sum_i x_i^2 - n\overline{x}^2}$$

Hence, using Equation 9.2.2 and the fact that $n\overline{Y} = \sum_{i=1}^n Y_i$, we have proven the following proposition.

PROPOSITION 9.2.1 The least squares estimators of β and α corresponding to the data set $x_i, Y_i, i = 1, \ldots, n$ are, respectively,

$$B = \frac{\sum_{i=1}^n x_i Y_i - \overline{x}\sum_{i=1}^n Y_i}{\sum_{i=1}^n x_i^2 - n\overline{x}^2}$$

$$A = \overline{Y} - B\overline{x}$$

The straight line $A + Bx$ is called the estimated regression line.

Program 9.2 computes the least squares estimators A and B. It also gives the user the option of computing some other statistics whose values will be needed in the following sections.

EXAMPLE 9.2a The raw material used in the production of a certain synthetic fiber is stored in a location without a humidity control. Measurements of the relative humidity in the storage location and the moisture content of a sample of the raw material were taken over 15 days with the following data (in percentages) resulting.

Relative humidity	46	53	29	61	36	39	47	49	52	38	55	32	57	54	44
Moisture content	12	15	7	17	10	11	11	12	14	9	16	8	18	14	12

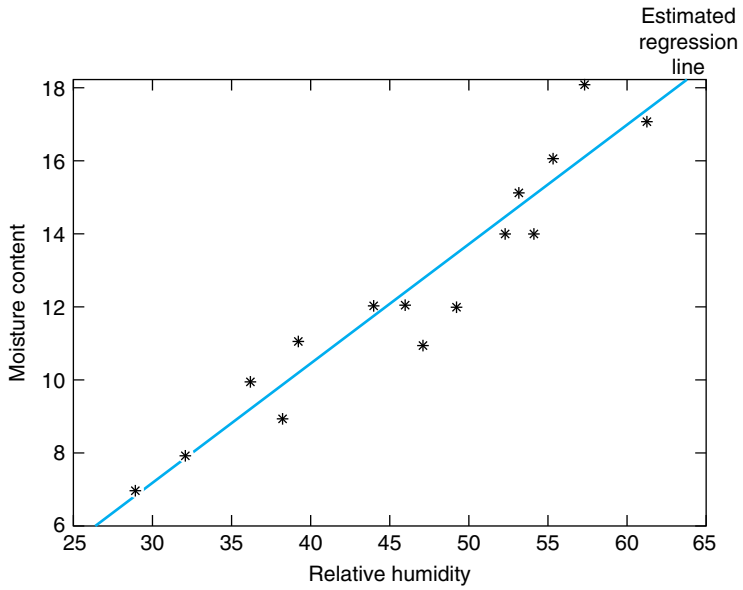

FIGURE 9.2 *Example 9.2a.*

These data are plotted in Figure 9.2. To compute the least squares estimator and the estimated regression line, we run Program 9.2; results are shown in Figure 9.3. ■

9.3 DISTRIBUTION OF THE ESTIMATORS

To specify the distribution of the estimators A and B, it is necessary to make additional assumptions about the random errors aside from just assuming that their mean is 0. The usual approach is to assume that the random errors are independent normal random variables having mean 0 and variance σ^2. That is, we suppose that if Y_i is the response corresponding to the input value x_i, then $Y_1, \ldots, Y_n$ are independent and

$$Y_i \sim \mathcal{N}(\alpha + \beta x_i, \sigma^2)$$

Note that the foregoing supposes that the variance of the random error does not depend on the input value but rather is a constant. This value σ^2 is not assumed to be known but rather must be estimated from the data.

Since the least squares estimator B of β can be expressed as

$$B = \frac{\sum_i (x_i - \bar{x}) Y_i}{\sum_i x_i^2 - n\bar{x}^2} \tag{9.3.1}$$

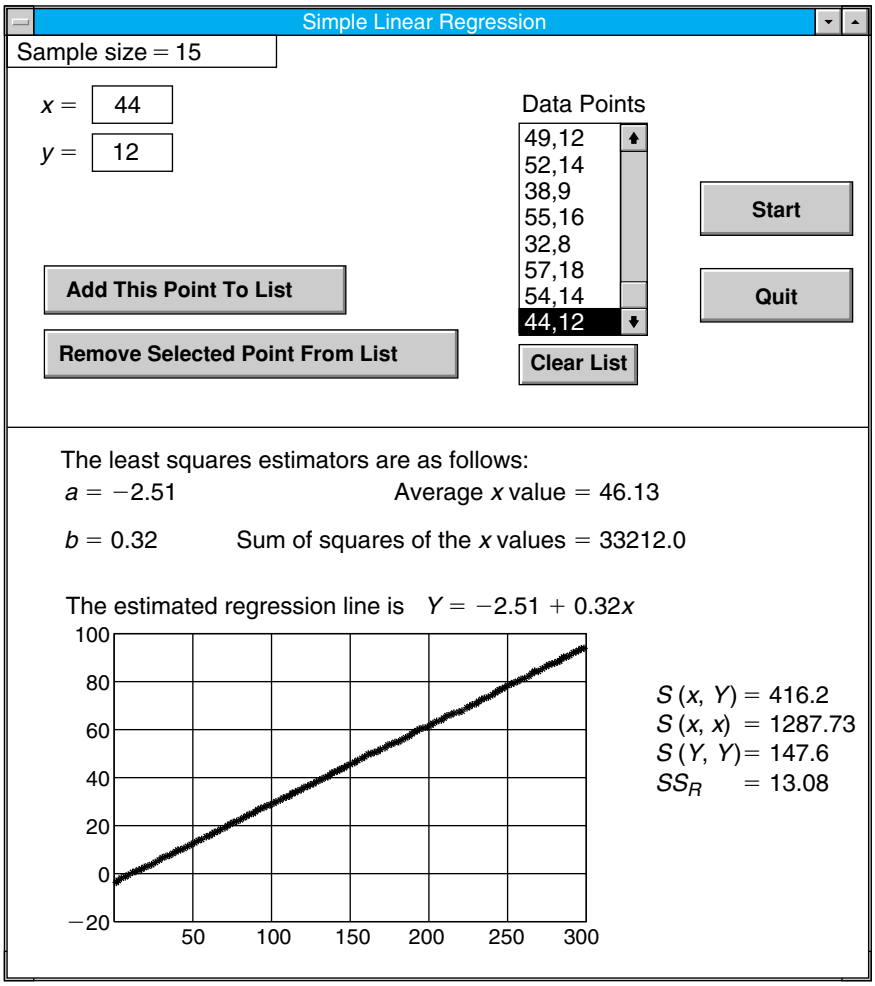

FIGURE 9.3

we see that it is a linear combination of the independent normal random variables Y_i, $i = 1, \ldots, n$ and so is itself normally distributed. Using Equation 9.3.1, the mean and variance of B are computed as follows:

$$E[B] = \frac{\sum_i (x_i - \bar{x}) E[Y_i]}{\sum_i x_i^2 - n\bar{x}^2}$$

$$= \frac{\sum_i (x_i - \bar{x})(\alpha + \beta x_i)}{\sum_i x_i^2 - n\bar{x}^2}$$

$$= \frac{\alpha \sum_i (x_i - \bar{x}) + \beta \sum_i x_i (x_i - \bar{x})}{\sum_i x_i^2 - n\bar{x}^2}$$

$$= \beta \frac{\left[\sum_i x_i^2 - \bar{x} \sum_i x_i\right]}{\sum_i x_i^2 - n\bar{x}^2} \qquad \text{since } \sum_i (x_i - \bar{x}) = 0$$

$$= \beta$$

Thus $E[B] = \beta$ and so B is an unbiased estimator of β. We will now compute the variance of B.

$$\text{Var}(B) = \frac{\text{Var}\left(\sum_{i=1}^{n} (x_i - \bar{x}) Y_i\right)}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2}$$

$$= \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2 \,\text{Var}(Y_i)}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2} \text{ by independence}$$

$$= \frac{\sigma^2 \sum_{i=1}^{n} (x_i - \bar{x})^2}{\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)^2}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \tag{9.3.2}$$

where the final equality results from the use of the identity

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2$$

Using Equation 9.3.1 along with the relationship

$$A = \sum_{i=1}^{n} \frac{Y_i}{n} - B\bar{x}$$

shows that A can also be expressed as a linear combination of the independent normal random variables $Y_i, i = 1, \ldots, n$ and is thus also normally distributed. Its mean is obtained from

$$\begin{aligned} E[A] &= \sum_{i=1}^{n} \frac{E[Y_i]}{n} - \bar{x}E[B] \\ &= \sum_{i=1}^{n} \frac{(\alpha + \beta x_i)}{n} - \bar{x}\beta \\ &= \alpha + \beta\bar{x} - \bar{x}\beta \\ &= \alpha \end{aligned}$$

Thus A is also an unbiased estimator. The variance of A is computed by first expressing A as a linear combination of the Y_i. The result (whose details are left as an exercise) is that

$$\text{Var}(A) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\left(\sum_{i=1}^{n} x_i^2 - n\bar{x}^2\right)} \tag{9.3.3}$$

The quantities $Y_i - A - Bx_i, i = 1, \ldots, n$, which represent the differences between the actual responses (that is, the Y_i) and their least squares estimators (that is, $A + Bx_i$) are called the *residuals*. The sum of squares of the residuals

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

can be utilized to estimate the unknown error variance σ^2. Indeed, it can be shown that

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

That is, SS_R/σ^2 has a chi-square distribution with $n-2$ degrees of freedom, which implies that

$$E\left[\frac{SS_R}{\sigma^2}\right] = n - 2$$

or

$$E\left[\frac{SS_R}{n-2}\right] = \sigma^2$$

Thus $SS_R/(n-2)$ is an unbiased estimator of σ^2. In addition, it can be shown that SS_R is independent of the pair A and B.

REMARKS

A plausibility argument as to why SS_R/σ^2 might have a chi-square distribution with $n-2$ degrees of freedom and be independent of A and B runs as follows. Because the Y_i are independent normal random variables, it follows that $(Y_i - E[Y_i])/\sqrt{\text{Var}(Y_i)}$, $i = 1, \ldots, n$ are independent standard normals and so

$$\sum_{i=1}^{n} \frac{(Y_i - E[Y_i])^2}{\text{Var}(Y_i)} = \sum_{i=1}^{n} \frac{(Y_i - \alpha - \beta x_i)^2}{\sigma^2} \sim \chi^2_n$$

Now if we substitute the estimators A and B for α and β, then 2 degrees of freedom are lost, and so it is not an altogether surprising result that SS_R/σ^2 has a chi-square distribution with $n-2$ degrees of freedom.

The fact that SS_R is independent of A and B is quite similar to the fundamental result that in normal sampling $\overline{X}$ and S^2 are independent. Indeed this latter result states that if $Y_1, \ldots, Y_n$ is a normal sample with population mean μ and variance σ_2, then if in the sum of squares $\sum_{i=1}^{n}(Y_i - \mu)^2/\sigma^2$, which has a chi-square distribution with n degrees of freedom, one substitutes the estimator $\overline{Y}$ for μ to obtain the new sum of squares $\sum_i (Y_i - \overline{Y})^2/\sigma^2$, then this quantity [equal to $(n-1)S^2/\sigma^2$] will be independent of $\overline{Y}$ and will have a chi-square distribution with $n-1$ degrees of freedom. Since SS_R/σ^2 is obtained by substituting the estimators A and B for α and β in the sum of squares $\sum_{i=1}^{n}(Y_i - \alpha - \beta x_i)^2/\sigma^2$, it is not unreasonable to expect that this quantity might be independent of A and B.

When the Y_i are normal random variables, the least squares estimators are also the maximum likelihood estimators. To verify this remark, note that the joint density of $Y_1, \ldots, Y_n$ is given by

$$\begin{aligned} f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) &= \prod_{i=1}^{n} f_{Y_i}(y_i) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-(y_i - \alpha - \beta x_i)^2/2\sigma^2} \\ &= \frac{1}{(2\pi)^{n/2}\sigma^n} e^{-\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2/2\sigma^2} \end{aligned}$$

Consequently, the maximum likelihood estimators of α and β are precisely the values of α and β that minimize $\sum_{i=1}^{n}(y_i - \alpha - \beta x_i)^2$. That is, they are the least squares estimators.

Notation

If we let

$$S_{xY} = \sum_{i=1}^{n}(x_i - \overline{x})(Y_i - \overline{Y}) = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}$$
$$S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n\overline{x}^2$$
$$S_{YY} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2$$

then the least squares estimators can be expressed as

$$B = \frac{S_{xY}}{S_{xx}}$$
$$A = \overline{Y} - B\overline{x}$$

The following computational identity for SS_R, the sum of squares of the residuals, can be established.

Computational Identity for SS_R

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}} \tag{9.3.4}$$

The following proposition sums up the results of this section.

PROPOSITION 9.3.I Suppose that the responses $Y_i, i = 1, \ldots, n$ are independent normal random variables with means $\alpha + \beta x_i$ and common variance σ^2. The least squares estimators of β and α

$$B = \frac{S_{xY}}{S_{xx}}, \qquad A = \overline{Y} - B\overline{x}$$

are distributed as follows:

$$A \sim \mathcal{N}\left(\alpha, \frac{\sigma^2 \sum_i x_i^2}{nS_{xx}}\right)$$
$$B \sim \mathcal{N}(\beta, \sigma^2/S_{xx})$$

In addition, if we let

$$SS_R = \sum_i (Y_i - A - Bx_i)^2$$

denote the sum of squares of the residuals, then

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

and SS_R is independent of the least squares estimators A and B. Also, SS_R can be computed from

$$SS_R = \frac{S_{xx}S_{YY} - (S_{xY})^2}{S_{xx}}$$

Program 9.2 will compute the least squares estimators A and B as well as $\bar{x}$, $\sum_i x_i^2$, S_{xx}, S_{xY}, S_{YY}, and SS_R.

EXAMPLE 9.3a The following data relate x, the moisture of a wet mix of a certain product, to Y, the density of the finished product.

x_i	y_i
5	7.4
6	9.3
7	10.6
10	15.4
12	18.1
15	22.2
18	24.1
20	24.8

Fit a linear curve to these data. Also determine SS_R.

SOLUTION A plot of the data and the estimated regression line is shown in Figure 9.4.

To solve the foregoing, run Program 9.2; results are shown in Figure 9.5. ■

9.4 STATISTICAL INFERENCES ABOUT THE REGRESSION PARAMETERS

Using Proposition 9.3.1, it is a simple matter to devise hypothesis tests and confidence intervals for the regression parameters.

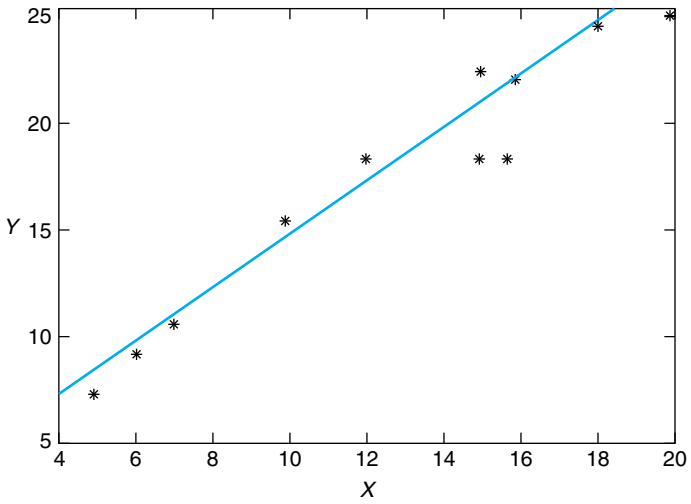

FIGURE 9.4 *Example 9.3a.*

9.4.1 Inferences Concerning β

An important hypothesis to consider regarding the simple linear regression model

$$Y = \alpha + \beta x + e$$

is the hypothesis that $\beta = 0$. Its importance derives from the fact that it is equivalent to stating that the mean response does not depend on the input, or, equivalently, that there is no regression on the input variable. To test

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_1 : \beta \neq 0$$

note that, from Proposition 9.3.1,

$$\frac{B - \beta}{\sqrt{\sigma^2 / S_{xx}}} = \sqrt{S_{xx}} \frac{(B - \beta)}{\sigma} \sim \mathcal{N}(0, 1) \tag{9.4.1}$$

and is independent of

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

Hence, from the definition of a t-random variable it follows that

$$\frac{\sqrt{S_{xx}}(B - \beta)/\sigma}{\sqrt{\frac{SS_R}{\sigma^2(n-2)}}} = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B - \beta) \sim t_{n-2} \tag{9.4.2}$$

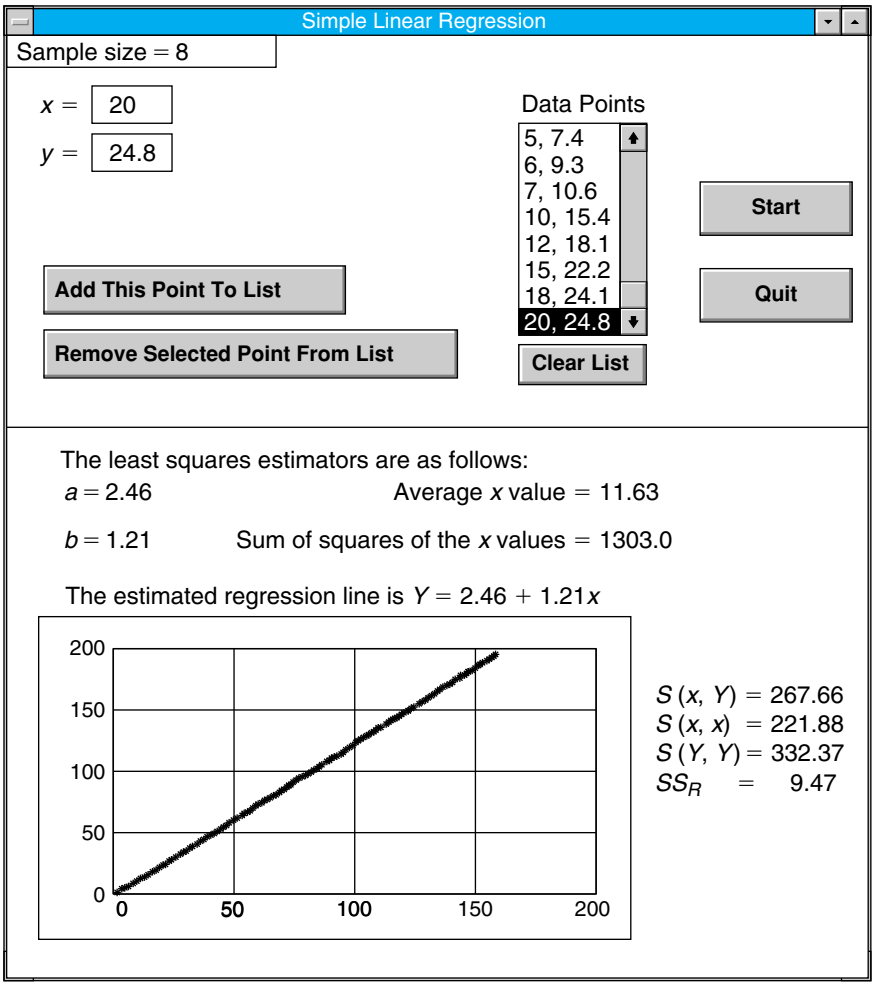

FIGURE 9.5

That is, $\sqrt{(n-2)S_{xx}/SS_R}(B-\beta)$ has a t-distribution with $n-2$ degrees of freedom. Therefore, if H_0 is true (and so $\beta = 0$), then

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}B \sim t_{n-2}$$

which gives rise to the following test of H_0.

Hypothesis Test of H_0: $\beta = 0$

A significance level γ test of H_0 is to

$$\begin{array}{lll} \text{reject} & H_0 & \text{if } \sqrt{\dfrac{(n-2)S_{xx}}{SS_R}}|B| > t_{\gamma/2,n-2} \\ \text{accept} & H_0 & \text{otherwise} \end{array}$$

This test can be performed by first computing the value of the test statistic $\sqrt{(n-2)S_{xx}/SS_R}|B|$ — call its value v — and then rejecting H_0 if the desired significance level is at least as large as

$$\begin{aligned} p\text{-value} &= P\{|T_{n-2}| > v\} \\ &= 2P\{T_{n-2} > v\} \end{aligned}$$

where T_{n-2} is a t-random variable with $n-2$ degrees of freedom. This latter probability can be obtained by using Program 5.8.2a.

EXAMPLE 9.4a An individual claims that the fuel consumption of his automobile does not depend on how fast the car is driven. To test the plausibility of this hypothesis, the car was tested at various speeds between 45 and 70 miles per hour. The miles per gallon attained at each of these speeds was determined, with the following data resulting:

Speed	Miles per Gallon
45	24.2
50	25.0
55	23.3
60	22.0
65	21.5
70	20.6
75	19.8

Do these data refute the claim that the mileage per gallon of gas is unaffected by the speed at which the car is being driven?

SOLUTION Suppose that a simple linear regression model

$$Y = \alpha + \beta x + e$$

relates Y, the miles per gallon of the car, to x, the speed at which it is being driven. Now, the claim being made is that the regression coefficient β is equal to 0. To see if the data are strong enough to refute this claim, we need to see if it leads to a rejection of the null hypothesis when testing

$$H_0 : \beta = 0 \qquad \text{versus} \qquad H_1 : \beta \neq 0$$

To compute the value of the test statistic, we first compute the values of S_{xx}, S_{YY}, and S_{xY}. A hand calculation yields that

$$S_{xx} = 700, \qquad S_{YY} = 21.757, \qquad S_{xY} = -119$$

Using Equation 9.3.4 gives

$$\begin{aligned} SS_R &= [S_{xx}S_{YY} - S_{xY}^2]/S_{xx} \\ &= [700(21.757) - (119)^2]/700 = 1.527 \end{aligned}$$

Because

$$B = S_{xY}/S_{xx} = -119/700 = -.17$$

the value of the test statistic is

$$TS = \sqrt{5(700)/1.527}(.17) = 8.139$$

Since, from Table A2 of the Appendix, $t_{.005,5} = 4.032$, it follows that the hypothesis $\beta = 0$ is rejected at the 1 percent level of significance. Thus, the claim that the mileage does not depend on the speed at which the car is driven is rejected; there is strong evidence that increased speeds lead to decreased mileages. ■

A confidence interval estimator for β is easily obtained from Equation 9.4.2. Indeed, it follows from Equation 9.4.2 that for any a, $0 < a < 1$,

$$P\left\{-t_{a/2,n-2} < \sqrt{\frac{(n-2)S_{xx}}{SS_R}}(B-\beta) < t_{a/2,n-2}\right\} = 1 - a$$

or, equivalently,

$$P\left\{B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2} < \beta < B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right\} = 1 - a$$

which yields the following.

Confidence Interval for β

A $100(1-a)$ percent confidence interval estimator of β is

$$\left(B - \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}, B + \sqrt{\frac{SS_R}{(n-2)S_{xx}}}t_{a/2,n-2}\right)$$

REMARK

The result that

$$\frac{B-\beta}{\sqrt{\sigma^2/S_{xx}}} \sim \mathcal{N}(0,1)$$

cannot be immediately applied to make inferences about β since it involves the unknown parameter σ^2. Instead, what we do is use the preceding statistic with σ^2 replaced by its estimator $SS_R/(n-2)$, which has the effect of changing the distribution of the statistic from the standard normal to the t-distribution with $n-2$ degrees of freedom.

EXAMPLE 9.4b Derive a 95 percent confidence interval estimate of β in Example 9.4a.

SOLUTION Since $t_{.025,5} = 2.571$, it follows from the computations of this example that the 95 percent confidence interval is

$$-.170 \pm 2.571\sqrt{\frac{1.527}{3{,}500}} = -.170 \pm .054$$

That is, we can be 95 percent confident that β lies between $-.224$ and $-.116$. ■

9.4.1.1 REGRESSION TO THE MEAN

The term *regression* was originally employed by Francis Galton while describing the laws of inheritance. Galton believed that these laws caused population extremes to "regress toward the mean." By this he meant that children of individuals having extreme values of a certain characteristic would tend to have less extreme values of this characteristic than their parent.

If we assume a linear regression relationship between the characteristic of the offspring (Y) and that of the parent (x), then a regression to the mean will occur when the regression parameter β is between 0 and 1. That is, if

$$E[Y] = \alpha + \beta x$$

and $0 < \beta < 1$, then $E[Y]$ will be smaller than x when x is large and greater than x when x is small. That this statement is true can be easily checked either algebraically or by plotting the two straight lines

$$y = \alpha + \beta x$$

and

$$y = x$$

A plot indicates that, when $0 < \beta < 1$, the line $y = \alpha + \beta x$ is above the line $y = x$ for small values of x and is below it for large values of x.

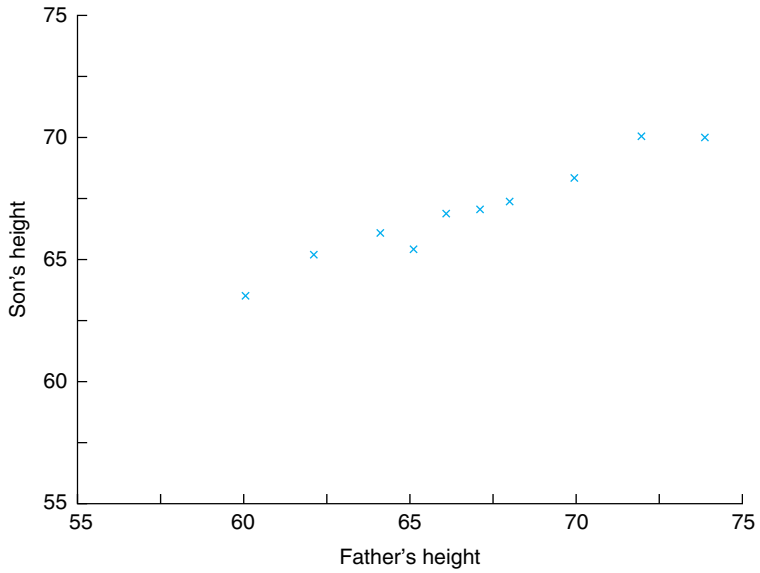

FIGURE 9.6 *Scatter diagram of son's height versus father's height.*

EXAMPLE 9.4c To illustrate Galton's thesis of regression to the mean, the British statistician Karl Pearson plotted the heights of 10 randomly chosen sons versus that of their fathers. The resulting data (in inches) were as follows.

Fathers' height	60	62	64	65	66	67	68	70	72	74
Sons' height	63.6	65.2	66	65.5	66.9	67.1	67.4	68.3	70.1	70

A scatter diagram representing these data is presented in Figure 9.6.

Note that whereas the data appear to indicate that taller fathers tend to have taller sons, it also appears to indicate that the sons of fathers who are either extremely short or extremely tall tend to be more "average" than their fathers — that is, there is a "regression toward the mean."

We will determine whether the preceding data are strong enough to prove that there is a regression toward the mean by taking this statement as the alternative hypothesis. That is, we will use the above data to test

$$H_0 : \beta \geq 1 \qquad \text{versus} \qquad H_1 : \beta < 1$$

which is equivalent to a test of

$$H_0 : \beta = 1 \qquad \text{versus} \qquad H_1 : \beta < 1$$

It now follows from Equation 9.4.2 that when $\beta = 1$, the test statistic

$$TS = \sqrt{8S_{xx}/SS_R}(B - 1)$$

has a t-distribution with 8 degrees of freedom. The significance level α test will reject H_0 when the value of TS is sufficiently small (since this will occur when B, the estimator of β, is sufficiently smaller than 1). Specifically, the test is to

$$\text{reject} \quad H_0 \quad \text{if} \quad \sqrt{8S_{xx}/SS_R}(B - 1) < -t_{\alpha,8}$$

Program 9.2 gives that

$$\sqrt{8S_{xx}/SS_R}(B - 1) = 30.2794(.4646 - 1) = -16.21$$

Since $t_{.01,8} = 2.896$, we see that

$$TS < -t_{.01,8}$$

and so the null hypothesis that $\beta \geq 1$ is rejected at the 1 percent level of significance. In fact, the p-value is

$$p\text{-value} = P\{T_8 \leq -16.213\} \approx 0$$

and so the null hypothesis that $\beta \geq 1$ is rejected at almost any significance level, thus establishing a regression toward the mean (see Figure 9.7).

A modern biological explanation for the regression to the mean phenomenon would roughly go along the lines of noting that as an offspring obtains a random selection of

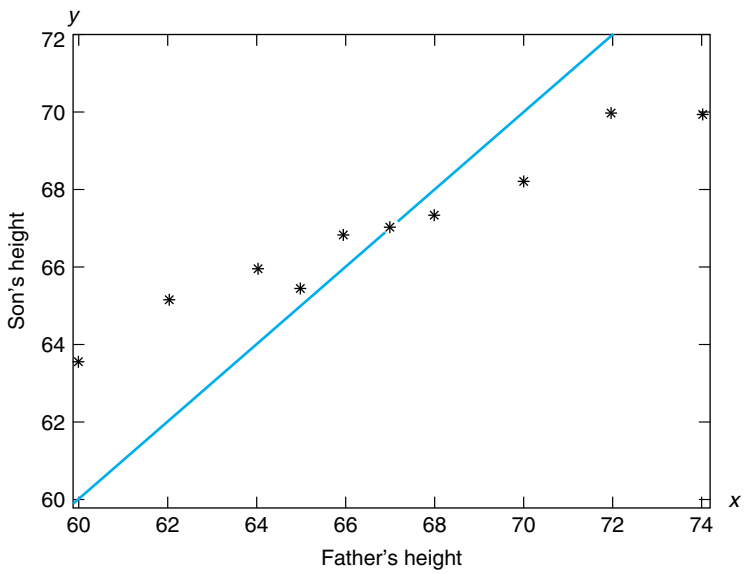

FIGURE 9.7 *Example 9.4c for x small, y > x. For x large, y < x.*

one-half of its parents' genes, it follows that the offspring of, say, a very tall parent would, by chance, tend to have fewer "tall" genes than its parent.

While the most important applications of the regression to the mean phenomenon concern the relationship between the biological characteristics of an offspring and that of its parents, this phenomenon also arises in situations where we have two sets of data referring to the same variables. ■

EXAMPLE 9.4d The data of Table 9.1 relate the number of motor vehicle deaths occurring in 12 counties in the northwestern United States in the years 1988 and 1989.

A glance at Figure 9.8 indicates that in 1989 there was, for the most part, a reduction in the number of deaths in those counties that had a large number of motor deaths in 1988. Similarly, there appears to have been an increase in those counties that had a low value in 1988. Thus, we would expect that a regression to the mean is in effect. In fact, running Program 9.2 yields that the estimated regression equation is

$$y = 74.589 + .276x$$

showing that the estimated value of β indeed appears to be less than 1.

One must be careful when considering the reason behind the regression to the mean phenomenon in the preceding data. For instance, it might be natural to suppose that those counties that had a large number of deaths caused by motor vehicles in 1988 would have made a large effort — perhaps by improving the safety of their roads or by making people more aware of the potential dangers of unsafe driving — to reduce this number. In addition, we might suppose that those counties that had the fewest number of deaths in 1988 might have "rested on their laurels" and not made much of an effort to further improve their numbers — and as a result had an increase in the number of casualties the following year.

TABLE 9.1 *Motor Vehicle Deaths, Northwestern United States, 1988 and 1989*

County	Deaths in 1988	Deaths in 1989
1	121	104
2	96	91
3	85	101
4	113	110
5	102	117
6	118	108
7	90	96
8	84	102
9	107	114
10	112	96
11	95	88
12	101	106

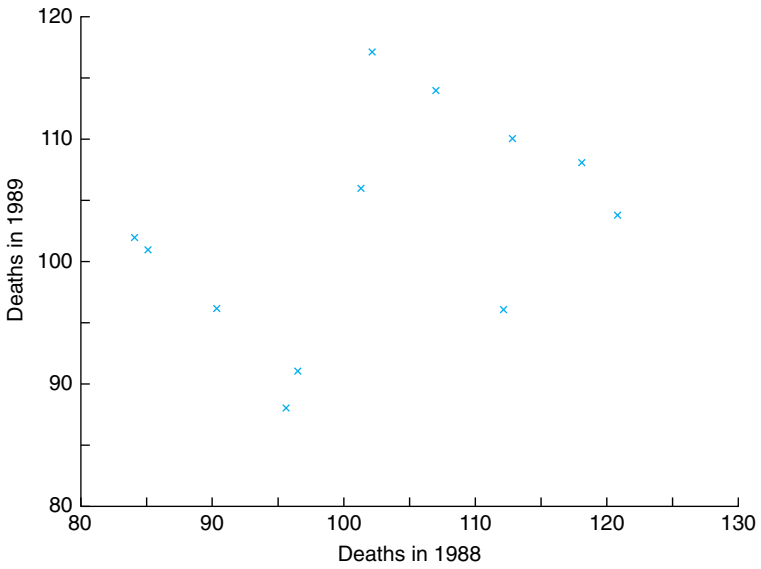

FIGURE 9.8 *Scatter diagram of 1989 deaths versus 1988 deaths.*

While this supposition might be correct, it is important to realize that a regression to the mean would probably have occurred even if none of the counties had done anything out of the ordinary. Indeed, it could very well be the case that those counties having large numbers of casualties in 1988 were just very unlucky in that year and thus a decrease in the next year was just a return to a more normal result for them. (For an analogy, if 9 heads result when 10 fair coins are flipped then it is quite likely that another flip of these 10 coins will result in fewer than 9 heads.) Similarly, those counties having few deaths in 1988 might have been "lucky" that year and a more normal result in 1989 would thus lead to an increase.

The mistaken belief that regression to the mean is due to some outside influence when it is in reality just due to "chance" occurs frequently enough that it is often referred to as the *regression fallacy*. ■

9.4.2 Inferences Concerning α

The determination of confidence intervals and hypothesis tests for α is accomplished in exactly the same manner as was done for β. Specifically, Proposition 9.3.1 can be used to show that

$$\sqrt{\frac{n(n-2)S_{xx}}{SS_R \sum_i x_i^2}}\,(A-\alpha) \tag{9.4.3}$$

which leads to the following confidence interval estimator of α.

Confidence Interval Estimator of α

The $100(1-a)$ percent confidence interval for α is the interval

$$A \pm t_{\alpha/2,n-2}\sqrt{\frac{SS_R \sum_i x_i^2}{n(n-2)S_{xx}}}$$

Hypothesis tests concerning α are easily obtained from Equation 9.4.3, and their development is left as an exercise.

9.4.3 Inferences Concerning the Mean Response $\alpha+\beta x_0$

It is often of interest to use the data pairs $(x_i, Y_i), i = 1, \ldots, n$, to estimate $\alpha + \beta x_0$, the mean response for a given input level x_0. If it is a point estimator that is desired, then the natural estimator is $A + Bx_0$, which is an unbiased estimator since

$$E[A + Bx_0] = E[A] + x_0 E[B] = \alpha + \beta x_0$$

However, if we desire a confidence interval, or are interested in testing some hypothesis about this mean response, then it is necessary to first determine the probability distribution of the estimator $A + Bx_0$. We now do so.

Using the expression for B given by Equation 9.3.1 yields that

$$B = c\sum_{i=1}^{n}(x_i - \bar{x})Y_i$$

where

$$c = \frac{1}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{1}{S_{xx}}$$

Since

$$A = \bar{Y} - B\bar{x}$$

we see that

$$\begin{aligned} A + Bx_0 &= \frac{\sum_{i=1}^{n} Y_i}{n} - B(\bar{x} - x_0) \\ &= \sum_{i=1}^{n} Y_i \left[\frac{1}{n} - c(x_i - \bar{x})(\bar{x} - x_0)\right] \end{aligned}$$

Since the Y_i are independent normal random variables, the foregoing equation shows that $A + Bx_0$ can be expressed as a linear combination of independent normal random

variables and is thus itself normally distributed. Because we already know its mean, we need only compute its variance, which is accomplished as follows:

$$\begin{aligned}
\text{Var}(A + Bx_0) &= \sum_{i=1}^{n} \left[\frac{1}{n} - c(x_i - \bar{x})(\bar{x} - x_0) \right]^2 \text{Var}(Y_i) \\
&= \sigma^2 \sum_{i=1}^{n} \left[\frac{1}{n^2} + c^2(\bar{x} - x_0)^2(x_i - \bar{x})^2 - 2c(x_i - \bar{x}) \frac{(\bar{x} - x_0)}{n} \right] \\
&= \sigma^2 \left[\frac{1}{n} + c^2(\bar{x} - x_0)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 - 2c(\bar{x} - x_0) \sum_{i=1}^{n} \frac{(x_i - \bar{x})}{n} \right] \\
&= \sigma^2 \left[\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{S_{xx}} \right]
\end{aligned}$$

where the last equality followed from

$$\sum_{i=1}^{n} (x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 = 1/c = S_{xx}, \qquad \sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Hence, we have shown that

$$A + Bx_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}} \right] \right) \tag{9.4.4}$$

In addition, because $A + Bx_0$ is independent of

$$SS_R/\sigma^2 \sim \chi^2_{n-2}$$

it follows that

$$\frac{A + Bx_0 - (\alpha + \beta x_0)}{\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \sqrt{\frac{SS_R}{n-2}}} \sim t_{n-2} \tag{9.4.5}$$

Equation 9.4.5 can now be used to obtain the following confidence interval estimator of $\alpha + \beta x_0$.

Confidence Interval Estimator of $\alpha + \beta x_0$

With $100(1 - a)$ percent confidence, $\alpha + \beta x_0$ will lie within

$$A + Bx_0 \pm \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}} \sqrt{\frac{SS_R}{n-2}} t_{a/2,\, n-2}$$

EXAMPLE 9.4e Using the data of Example 9.4c, determine a 95 percent confidence interval for the average height of all males whose fathers are 68 inches tall.

SOLUTION Since the observed values are

$$n = 10, \qquad x_0 = 68, \qquad \bar{x} = 66.8, \qquad S_{xx} = 171.6, \qquad SS_R = 1.49721$$

we see that

$$\sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}}\sqrt{\frac{SS_R}{n-2}} = .1424276$$

Also, because

$$t_{.025,8} = 2.306, \qquad A + B x_0 = 67.56751$$

we obtain the following 95 percent confidence interval:

$$\alpha + \beta x_0 \in (67.239, 67.896) \quad \blacksquare$$

9.4.4 Prediction Interval of a Future Response

It is often the case that it is more important to estimate the actual value of a future response rather than its mean value. For instance, if an experiment is to be performed at temperature level x_0, then we would probably be more interested in predicting $Y(x_0)$, the yield from this experiment, than we would be in estimating the expected yield — $E[Y(x_0)] = \alpha + \beta x_0$. (On the other hand, if a series of experiments were to be performed at input level x_0, then we would probably want to estimate $\alpha + \beta x_0$, the mean yield.)

Suppose first that we are interested in a single value (as opposed to an interval) to use as a predictor of $Y(x_0)$, the response at level x_0. Now, it is clear that the best predictor of $Y(x_0)$ is its mean value $\alpha + \beta x_0$. [Actually, this is not so immediately obvious since one could argue that the best predictor of a random variable is (1) its mean — which minimizes the expected square of the difference between the predictor and the actual value; or (2) its median — which minimizes the expected absolute difference between the predictor and the actual value; or (3) its mode — which is the most likely value to occur. However, as the mean, median, and mode of a normal random variable are all equal — and the response is, by assumption, normally distributed — there is no doubt in this situation.] Since α and β are not known, it seems reasonable to use their estimators A and B and thus use $A + B x_0$ as the predictor of a new response at input level x_0.

Let us now suppose that rather than being concerned with determining a single value to predict a response, we are interested in finding a prediction interval that, with a given degree of confidence, will contain the response. To obtain such an interval, let Y denote

the future response whose input level is x_0 and consider the probability distribution of the response minus its predicted value — that is, the distribution of $Y - A - Bx_0$. Now,

$$Y \sim \mathcal{N}(\alpha + \beta x_0, \sigma^2)$$

and, as was shown in Section 9.4.3,

$$A + Bx_0 \sim \mathcal{N}\left(\alpha + \beta x_0, \sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

Hence, because Y is independent of the earlier data values $Y_1, Y_2, \ldots, Y_n$ that were used to determine A and B, it follows that Y is independent of $A + Bx_0$ and so

$$Y - A - Bx_0 \sim \mathcal{N}\left(0, \sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\right)$$

or, equivalently,

$$\frac{Y - A - Bx_0}{\sigma\sqrt{\dfrac{n+1}{n} + \dfrac{(x_0 - \bar{x})^2}{S_{xx}}}} \sim \mathcal{N}(0, 1) \tag{9.4.6}$$

Now, using once again the result that SS_R is independent of A and B (and also of Y) and

$$\frac{SS_R}{\sigma^2} \sim \chi^2_{n-2}$$

we obtain, by the usual argument, upon replacing σ^2 in Equation 9.4.6 by its estimator $SS_R/(n-2)$ that

$$\frac{Y - A - Bx_0}{\sqrt{\dfrac{n+1}{n} + \dfrac{(x_0 - \bar{x})^2}{S_{xx}}}\sqrt{\dfrac{SS_R}{n-2}}} \sim t_{n-2}$$

and so, for any value a, $0 < a < 1$,

$$p\left\{-t_{a/2,n-2} < \frac{Y - A - Bx_0}{\sqrt{\dfrac{n+1}{n} + \dfrac{(x_0 - \bar{x})^2}{S_{xx}}}\sqrt{\dfrac{SS_R}{n-2}}} < t_{a/2,n-2}\right\} = 1 - a$$

That is, we have just established the following.

Prediction Interval for a Response at the Input Level x_0

Based on the response values Y_i corresponding to the input values $x_i, i = 1, 2, \ldots, n$: With $100(1-a)$ percent confidence, the response Y at the input level x_0 will be contained in the interval

$$A + B x_0 \pm t_{a/2,n-2}\sqrt{\left[\frac{n+1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]\frac{SS_R}{n-2}}$$

EXAMPLE 9.4f In Example 9.4c, suppose we want an interval that we can "be 95 percent certain" will contain the height of a given male whose father is 68 inches tall. A simple computation now yields the prediction interval

$$Y(68) \in 67.568 \pm 1.050$$

or, with 95 percent confidence, the person's height will be between 66.518 and 68.618. ■

REMARKS

(a) There is often some confusion about the difference between a confidence and a prediction interval. A confidence interval is an interval that does contain, with a given degree of confidence, a fixed parameter of interest. A prediction interval, on the other hand, is an interval that will contain, again with a given degree of confidence, a random variable of interest.

(b) One should not make predictions about responses at input levels that are far from those used to obtain the estimated regression line. For instance, the data of Example 9.4c should not be used to predict the height of a male whose father is 42 inches tall.

9.4.5 Summary of Distributional Results

We now summarize the distributional results of this section.

$$\textit{Model: } Y = \alpha + \beta x + e, \quad e \sim \mathcal{N}(0, \sigma^2)$$
$$\textit{Data: } (x_i, Y_i), \quad i = 1, 2, \ldots, n$$

Inferences About	Use the Distributional Result
β	$\sqrt{\frac{(n-2)S_{xx}}{SS_r}}(B - \beta) \sim t_{n-2}$
α	$\sqrt{\frac{n(n-2)S_{xx}}{\sum_i x_i^2 SS_R}}(A - \alpha) \sim t_{n-2}$

Inferences About	Use the Distributional Result
$\alpha + \beta x_0$	$\dfrac{A + Bx_0 - \alpha - \beta x_0}{\sqrt{\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\left(\frac{SS_R}{n-2}\right)}} \sim t_{n-2}$
$Y(x_0)$	$\dfrac{Y(x_0) - A - Bx_0}{\sqrt{\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)\left(\frac{SS_R}{n-2}\right)}} \sim t_{n-2}$

9.5 THE COEFFICIENT OF DETERMINATION AND THE SAMPLE CORRELATION COEFFICIENT

Suppose we wanted to measure the amount of variation in the set of response values $Y_1, \ldots, Y_n$ corresponding to the set of input values $x_1, \ldots, x_n$. A standard measure in statistics of the amount of variation in a set of values $Y_1, \ldots, Y_n$ is given by the quantity

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

For instance, if all the Y_i are equal — and thus are all equal to $\overline{Y}$ — then S_{YY} would equal 0.

The variation in the values of the Y_i arises from two factors. First, because the input values x_i are different, the response variables Y_i all have different mean values, which will result in some variation in their values. Second, the variation also arises from the fact that even when the differences in the input values are taken into account, each of the response variables Y_i has variance σ^2 and thus will not exactly equal the predicted value at its input x_i.

Let us consider now the question as to how much of the variation in the values of the response variables is due to the different input values, and how much is due to the inherent variance of the responses even when the input values are taken into account. To answer this question, note that the quantity

$$SS_R = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

measures the remaining amount of variation in the response values after the different input values have been taken into account.

Thus,

$$S_{YY} - SS_R$$

represents the amount of variation in the response variables that is *explained* by the different input values, and so the quantity R^2 defined by

$$R^2 = \frac{S_{YY} - SS_R}{S_{YY}} = 1 - \frac{SS_R}{S_{YY}}$$

represents the proportion of the variation in the response variables that is explained by the different input values. R^2 is called the *coefficient of determination.*

The coefficient of determination R^2 will have a value between 0 and 1. A value of R^2 near 1 indicates that most of the variation of the response data is explained by the different input values, whereas a value of R^2 near 0 indicates that little of the variation is explained by the different input values.

EXAMPLE 9.5a In Example 9.4c, which relates the height of a son to that of his father, the output from Program 9.2 yielded that

$$S_{YY} = 38.521, \qquad SS_R = 1.497$$

Thus,

$$R^2 = 1 - \frac{1.497}{38.531} = .961$$

In other words, 96 percent of the variation of the heights of the 10 individuals is explained by the heights of their fathers. The remaining (unexplained) 4 percent of the variation is due to the variance of a son's height even when the father's height is taken into account. (That is, it is due to σ^2, the variance of the error random variable.) ■

The value of R^2 is often used as an indicator of how well the regression model fits the data, with a value near 1 indicating a good fit, and one near 0 indicating a poor fit. In other words, if the regression model is able to explain most of the variation in the response data, then it is considered to fit the data well.

Recall that in Section 2.6 we defined the sample correlation coefficient r of the set of data pairs (x_i, Y_i), $i = 1, \ldots, n$, by

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

It was noted that r provided a measure of the degree to which high values of x are paired with high values of Y and low values of x with low values of Y. A value of r

near +1 indicated that large x values were strongly associated with large Y values and small x values were strongly associated with small Y values, whereas a value near −1 indicated that large x values were strongly associated with small Y values and small x values with large Y values.

In the notation of this chapter,

$$r = \frac{S_{xY}}{\sqrt{S_{xx}S_{YY}}}$$

Upon using identity (9.3.4):

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

we see that

$$\begin{aligned} r^2 &= \frac{S_{xY}^2}{S_{xx}S_{YY}} \\ &= \frac{S_{xx}S_{YY} - SS_R S_{xx}}{S_{xx}S_{YY}} \\ &= 1 - \frac{SS_R}{S_{YY}} \\ &= R^2 \end{aligned}$$

That is,

$$|r| = \sqrt{R^2}$$

and so, except for its sign indicating whether it is positive or negative, the sample correlation coefficient is equal to the square root of the coefficient of determination. The sign of r is the same as that of B.

The above gives additional meaning to the sample correlation coefficient. For instance, if a data set has its sample correlation coefficient r equal to .9, then this implies that a simple linear regression model for these data explains 81 percent (since $R^2 = .9^2 = .81$) of the variation in the response values. That is, 81 percent of the variation in the response values is explained by the different input values.

9.6 ANALYSIS OF RESIDUALS: ASSESSING THE MODEL

The initial step for ascertaining whether or not the simple linear regression model

$$Y = \alpha + \beta x + e, \qquad e \sim \mathcal{N}(0, \sigma^2)$$

is appropriate in a given situation is to investigate the scatter diagram. Indeed, this is often sufficient to convince one that the regression model is or is not correct. When the scatter diagram does not by itself rule out the preceding model, then the least squares estimators A and B should be computed and the residual $Y_i - (A + Bx_i), i = 1, \ldots, n$ analyzed. The analysis begins by normalizing, or standardizing, the residuals by dividing them by $\sqrt{SS_R/(n-2)}$, the estimate of the standard deviation of the Y_i. The resulting quantities

$$\frac{Y_i - (A + Bx_i)}{\sqrt{SS_R/(n-2)}}, \qquad i = 1, \ldots, n$$

are called the *standardized residuals*.

When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables, and thus should be randomly distributed about 0 with about 95 percent of their values being between -2 and $+2$ (since $P\{-1.96 < Z < 1.96\} = .95$). In addition, a plot of the standardized residuals should not indicate any distinct pattern. Indeed, any indication of a distinct pattern should make one suspicious about the validity of the assumed simple linear regression model.

Figure 9.9 presents three different scatter diagrams and their associated standardized residuals. The first of these, as indicated both by its scatter diagram and the random nature of its standardized residuals, appears to fit the straight-line model quite well. The second residual plot shows a discernible pattern, in that the residuals appear to be first

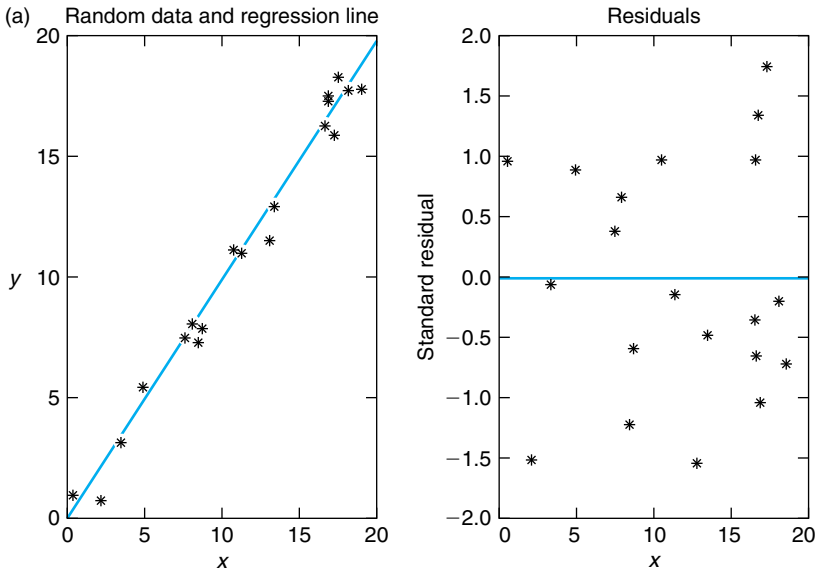

FIGURE 9.9

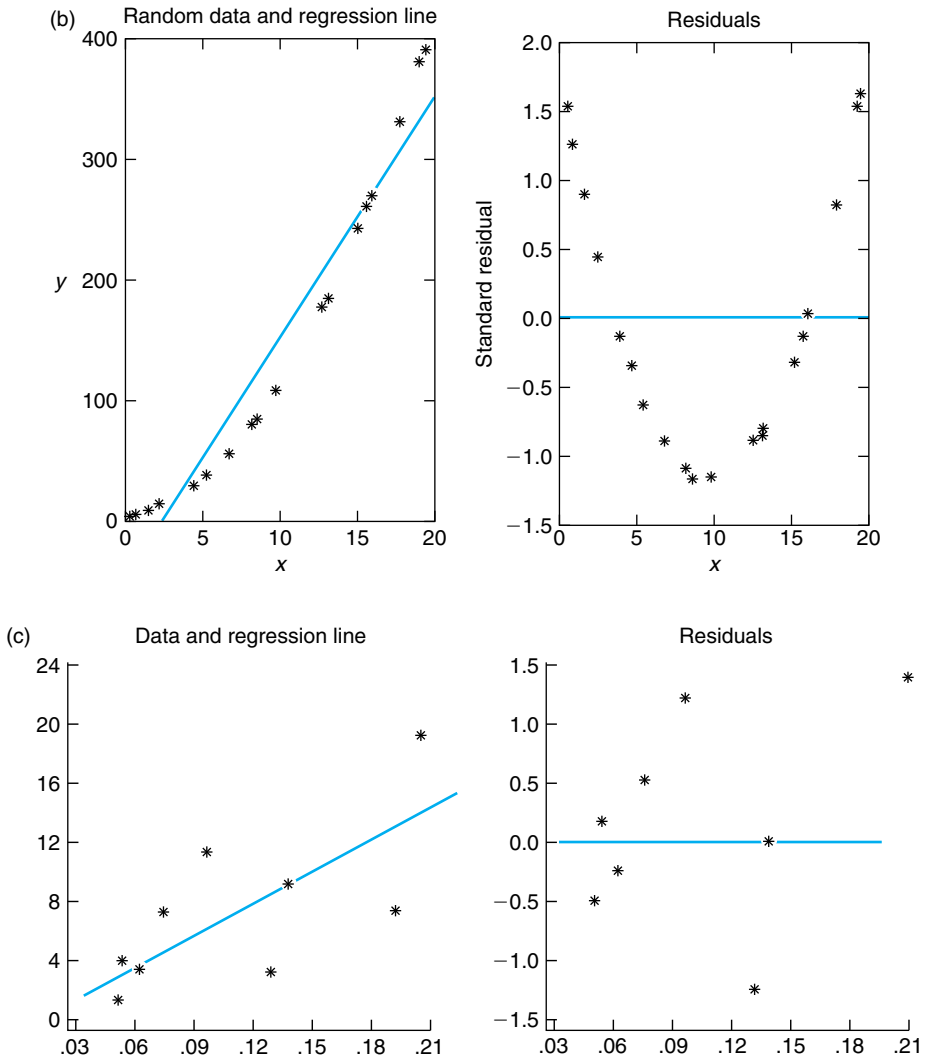

FIGURE 9.9 *(continued)*

decreasing and then increasing as the input level increases. This often means that higher-order (than just linear) terms are needed to describe the relationship between the input and response. Indeed, this is also indicated by the scatter diagram in this case. The third standardized residual plot also shows a pattern, in that the absolute value of the residuals, and thus their squares, appear to be increasing, as the input level increases. This often indicates that the variance of the response is not constant but, rather, increases with the input level.

9.7 TRANSFORMING TO LINEARITY

In many situations, it is clear that the mean response is not a linear function of the input level. In such cases, if the form of the relationship can be determined it is sometimes possible, by a change of variables, to transform it into a linear form. For instance, in certain applications it is known that $W(t)$, the amplitude of a signal a time t after its origination, is approximately related to t by the functional form

$$W(t) \approx ce^{-dt}$$

On taking logarithms, this can be expressed as

$$\log W(t) \approx \log c - dt$$

If we now let

$$\begin{aligned} Y &= \log W(t) \\ \alpha &= \log c \\ \beta &= -d \end{aligned}$$

then the foregoing can be modeled as a regression of the form

$$Y = \alpha + \beta t + e$$

The regression parameters α and β would then be estimated by the usual least squares approach and the original functional relationships can be predicted from

$$W(t) \approx e^{A+Bt}$$

EXAMPLE 9.7a The following table gives the percentages of a chemical that were used up when an experiment was run at various temperatures (in degrees Celsius). Use it to estimate the percentage of the chemical that would be used up if the experiment were to be run at 350 degrees.

Temperature	Percentage
5°	.061
10°	.113
20°	.192
30°	.259
40°	.339
50°	.401
60°	.461
80°	.551

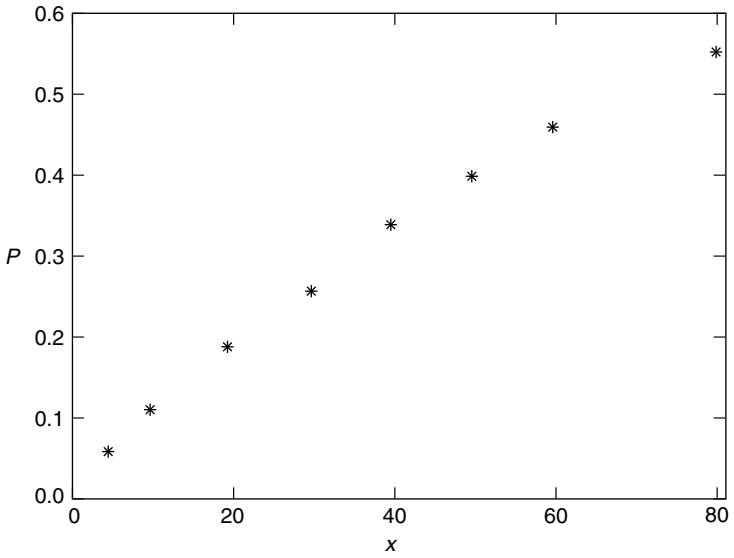

FIGURE 9.10 *Example 9.7a.*

SOLUTION Let $P(x)$ be the percentage of the chemical that is used up when the experiment is run at $10x$ degrees. Even though a plot of $P(x)$ looks roughly linear (see Figure 9.10), we can improve upon the fit by considering a nonlinear relationship between x and $P(x)$. Specifically, let us consider a relationship of the form

$$1 - P(x) \approx c(1 - d)^x$$

That is, let us suppose that the percentage of the chemical that survives an experiment run at temperature x approximately decreases at an exponential rate when x increases. Taking logs, the preceding can be written as

$$\log(1 - P(x)) \approx \log(c) + x \log(1 - d)$$

Thus, setting

$$\begin{aligned} Y &= -\log(1 - P) \\ \alpha &= -\log c \\ \beta &= -\log(1 - d) \end{aligned}$$

we obtain the usual regression equation

$$Y = \alpha + \beta x + e$$

TABLE 9.2

Temperature	$-\log(1 - P)$
5°	.063
10°	.120
20°	.213
30°	.300
40°	.414
50°	.512
60°	.618
80°	.801

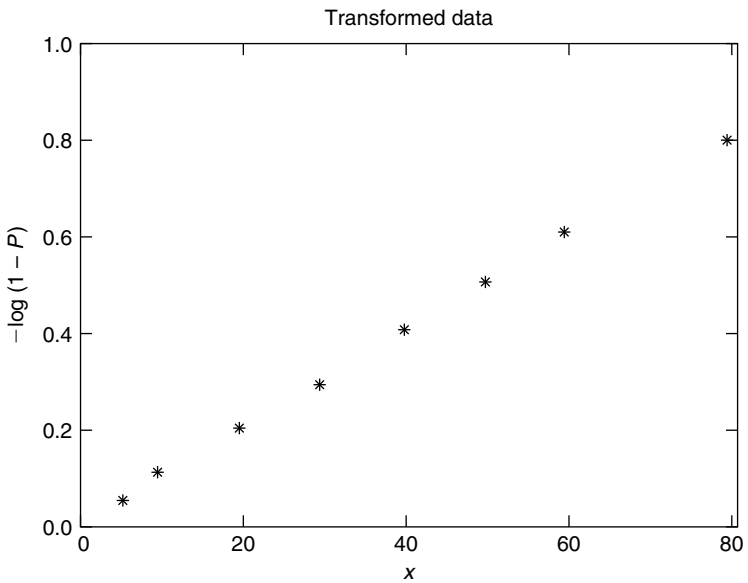

FIGURE 9.11

To see whether the data support this model, we can plot $-\log(1 - P)$ versus x. The transformed data are presented in Table 9.2 and the graph in Figure 9.11.

Running Program 9.2 yields that the least square estimates of α and β are

$$A = .0154$$
$$B = .0099$$

TABLE 9.3

x	P	$\hat{P}$	$P-\hat{P}$
5	.061	.063	−.002
10	.113	.109	.040
20	.192	.193	−.001
30	.259	.269	−.010
40	.339	.339	.000
50	.401	.401	.000
60	.461	.458	.003
80	.551	.556	−.005

Transforming this back into the original variable gives that the estimates of c and d are

$$\hat{c} = e^{-A} = .9847$$
$$1 - \hat{d} = e^{-B} = .9901$$

and so the estimated functional relationship is

$$\hat{P} = 1 - .9847(.9901)^x$$

The residuals $P - \hat{P}$ are presented in Table 9.3. ■

9.8 WEIGHTED LEAST SQUARES

In the regression model

$$Y = \alpha + \beta x + e$$

it often turns out that the variance of a response is not constant but rather depends on its input level. If these variances are known — at least up to a proportionality constant — then the regression parameters α and β should be estimated by minimizing a weighted sum of squares. Specifically, if

$$\text{Var}(Y_i) = \frac{\sigma^2}{w_i}$$

then the estimators A and B should be chosen to minimize

$$\sum_i \frac{[Y_i - (A + Bx_i)]^2}{\text{Var}(Y_i)} = \frac{1}{\sigma^2} \sum_i w_i (Y_i - A - Bx_i)^2$$

On taking partial derivatives with respect to A and B and setting them equal to 0, we obtain the following equations for the minimizing A and B.

$$\sum_i w_i Y_i = A \sum_i w_i + B \sum_i w_i x_i \tag{9.8.1}$$
$$\sum_i w_i x_i Y_i = A \sum_i w_i x_i + B \sum_i w_i x_i^2$$

These equations are easily solved to yield the least squares estimators.

EXAMPLE 9.8a To develop a feel as to why the estimators should be obtained by minimizing the weighted sum of squares rather than the ordinary sum of squares, consider the following situation. Suppose that $X_1, \ldots, X_n$ are independent normal random variables each having mean μ and variance σ^2. Suppose further that the X_i are not directly observable but rather only Y_1 and Y_2, defined by

$$Y_1 = X_1 + \cdots + X_k, \qquad Y_2 = X_{k+1} + \cdots + X_n, \qquad k < n$$

are directly observable. Based on Y_1 and Y_2, how should we estimate μ?

Whereas the best estimator of μ is clearly $\overline{X} = \sum_{i=1}^n X_i/n = (Y_1 + Y_2)/n$, let us see what the ordinary least squares estimator would be. Since

$$E[Y_1] = k\mu, \qquad E[Y_2] = (n-k)\mu$$

the least squares estimator of μ would be that value of μ that minimizes

$$(Y_1 - k\mu)^2 + (Y_2 - [n-k]\mu)^2$$

On differentiating and setting equal to zero, we see that the least squares estimator of μ — call it $\hat{\mu}$ — is such that

$$-2k(Y_1 - k\hat{\mu}) - 2(n-k)[Y_2 - (n-k)\hat{\mu}] = 0$$

or

$$[k^2 + (n-k)^2]\hat{\mu} = kY_1 + (n-k)Y_2$$

or

$$\hat{\mu} = \frac{kY_1 + (n-k)Y_2}{k^2 + (n-k)^2}$$

Thus we see that while the ordinary least squares estimator is an unbiased estimator of μ — since

$$E[\hat{\mu}] = \frac{kE[Y_1] + (n-k)E[Y_2]}{k^2 + (n-k)^2} = \frac{k^2\mu + (n-k)^2\mu}{k^2 + (n-k)^2} = \mu,$$

it is not the best estimator $\overline{X}$.

Now let us determine the estimator produced by minimizing the weighted sum of squares. That is, let us determine the value of μ — call it μ_w — that minimizes

$$\frac{(Y_1 - k\mu)^2}{\text{Var}(Y_1)} + \frac{[Y_2 - (n-k)\mu]^2}{\text{Var}(Y_2)}$$

Since

$$\text{Var}(Y_1) = k\sigma^2, \qquad \text{Var}(Y_2) = (n-k)\sigma^2$$

this is equivalent to choosing μ to minimize

$$\frac{(Y_1 - k\mu)^2}{k} + \frac{[Y_2 - (n-k)\mu]^2}{n-k}$$

Upon differentiating and then equating to 0, we see that μ_w, the minimizing value, satisfies

$$\frac{-2k(Y_1 - k\mu_w)}{k} - \frac{2(n-k)[Y_2 - (n-k)\mu_w]}{n-k} = 0$$

or

$$Y_1 + Y_2 = n\mu_w$$

or

$$\mu_w = \frac{Y_1 + Y_2}{n}$$

That is, the weighted least squares estimator is indeed the preferred estimator $(Y_1 + Y_2)/n = \overline{X}$. ■

REMARKS

(a) Assuming normally distributed data, the weighted least squares estimators are precisely the maximum likelihood estimators. This follows because the joint density of the data $Y_1, \ldots, Y_n$ is

$$\begin{aligned} f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}(\sigma/\sqrt{w_i})} e^{-(y_i - \alpha - \beta x_i)^2/(2\sigma^2/w_i)} \\ &= \frac{\sqrt{w_1 \cdots w_n}}{(2\pi)^{n/2}\sigma^n} e^{-\sum_{i=1}^{n} w_i (y_i - \alpha - \beta x_i)^2/2\sigma^2} \end{aligned}$$

Consequently, the maximum likelihood estimators of α and β are precisely the values of α and β that minimize the weighted sum of squares $\sum_{i=1}^{n} w_i (y_i - \alpha - \beta x_i)^2$.

(b) The weighted sum of squares can also be seen as the relevant quantity to be minimized by multiplying the regression equation

$$Y = \alpha + \beta x + e$$

by $\sqrt{w}$. This results in the equation

$$Y\sqrt{w} = \alpha\sqrt{w} + \beta x\sqrt{w} + e\sqrt{w}$$

Now, in this latter equation the error term $e\sqrt{w}$ has mean 0 and constant variance. Hence, the natural least squares estimators of α and β would be the values of A and B that minimize

$$\sum_i (Y_i\sqrt{w_i} - A\sqrt{w_i} - Bx_i\sqrt{w_i})^2 = \sum_i w_i(Y_i - A - Bx_i)^2$$

(c) The weighted least squares approach puts the greatest emphasis on those data pairs having the greatest weights (and thus the smallest variance in their error term). ■

At this point it might appear that the weighted least squares approach is not particularly useful since it requires a knowledge, up to a constant, of the variance of a response at an arbitrary input level. However, by analyzing the model that generates the data, it is often possible to determine these values. This will be indicated by the following two examples.

EXAMPLE 9.8b The following data represent travel times in a downtown area of a certain city. The independent, or input, variable is the distance to be traveled.

Distance (miles)	.5	1	1.5	2	3	4	5	6	8	10
Travel time (minutes)	15.0	15.1	16.5	19.9	27.7	29.7	26.7	35.9	42	49.4

Assuming a linear relationship of the form

$$Y = \alpha + \beta x + e$$

between Y, the travel time, and x, the distance, how should we estimate α and β? To utilize the weighted least squares approach we need to know, up to a multiplicative constant, the variance of Y as a function of x. We will now present an argument that $\text{Var}(Y)$ should be proportional to x.

SOLUTION Let d denote the length of a city block. Thus a trip of distance x will consist of x/d blocks. If we let $Y_i, i = 1, \ldots, x/d$, denote the time it takes to traverse block i, then the total travel time can be expressed as

$$Y = Y_1 + Y_2 + \cdots + Y_{x/d}$$

Now in many applications it is probably reasonable to suppose that the Y_i are independent random variables with a common variance, and thus,

$$\begin{aligned}\text{Var}(Y) &= \text{Var}(Y_1) + \cdots + \text{Var}(Y_{x/d}) \\ &= (x/d)\text{Var}(Y_1) \qquad \text{since Var}(Y_i) = \text{Var}(Y_1) \\ &= x\sigma^2, \qquad \text{where } \sigma^2 = \text{Var}(Y_1)/d\end{aligned}$$

Thus, it would seem that the estimators A and B should be chosen so as to minimize

$$\sum_i \frac{(Y_i - A - Bx_i)^2}{x_i}$$

Using the preceding data with the weights $w_i = 1/x_i$, the least squares Equations 9.8.1 are

$$\begin{aligned}104.22 &= 5.34A + 10B \\ 277.9 &= 10A + 41B\end{aligned}$$

which yield the solution

$$A = 12.561, \qquad B = 3.714$$

A graph of the estimated regression line $12.561 + 3.714x$ along with the data points is presented in Figure 9.12. As a qualitative check of our solution, note that the regression line fits the data pairs best when the input levels are small, which is as it should be since the weights are inversely proportional to the inputs. ■

EXAMPLE 9.8c Consider the relationship between Y, the number of accidents on a heavily traveled highway, and x, the number of cars traveling on the highway. After a little thought it would probably seem to most that the linear model

$$Y = \alpha + \beta x + e$$

would be appropriate. However, as there does not appear to be any *a priori* reason why $\text{Var}(Y)$ should not depend on the input level x, it is not clear that we would be justified in using the ordinary least squares approach to estimate α and β. Indeed, we will now argue that a weighted least squares approach with weights $1/x$ should be employed — that is, we should choose A and B to minimize

$$\sum_i \frac{(Y_i - A - Bx_i)^2}{x_i}$$

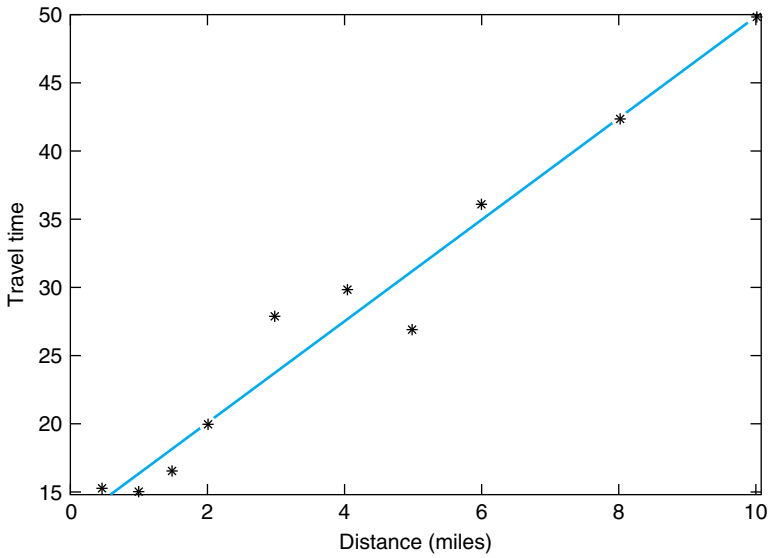

FIGURE 9.12 *Example 9.8b.*

The rationale behind this claim is that it seems reasonable to suppose that Y has approximately a Poisson distribution. This is so since we can imagine that each of the x cars will have a small probability of causing an accident and so, for large x, the number of accidents should be approximately a Poisson random variable. Since the variance of a Poisson random variable is equal to its mean, we see that

$$\begin{aligned} \text{Var}(Y) &\simeq E[Y] \qquad \text{since } Y \text{ is approximately Poisson} \\ &= \alpha + \beta x \\ &\simeq \beta x \qquad \text{for large } x \end{aligned}$$ ■

REMARKS

(a) Another technique that is often employed when the variance of the response depends on the input level is to attempt to stabilize the variance by an appropriate transformation. For example, if Y is a Poisson random variable with mean λ, then it can be shown [see Remark (b)] that $\sqrt{Y}$ has approximate variance .25 no matter what the value of λ. Based on this fact, one might try to model $E[\sqrt{Y}]$ as a linear function of the input. That is, one might consider the model

$$\sqrt{Y} = \alpha + \beta x + e$$

(b) Proof that $\text{Var}(\sqrt{Y}) \approx .25$ when Y is Poisson with mean λ. Consider the Taylor series expansion of $g(y) = \sqrt{y}$ about the value λ. By ignoring all terms beyond the second derivative term, we obtain that

$$g(y) \approx g(\lambda) + g'(\lambda)(y - \lambda) + \frac{g''(\lambda)(y - \lambda)^2}{2} \tag{9.8.2}$$

Since

$$g'(\lambda) = \tfrac{1}{2}\lambda^{-1/2}, \qquad g''(\lambda) = -\tfrac{1}{4}\lambda^{-3/2}$$

we obtain, on evaluating Equation 9.8.2 at $y = Y$, that

$$\sqrt{Y} \approx \sqrt{\lambda} + \tfrac{1}{2}\lambda^{-1/2}(Y - \lambda) - \tfrac{1}{8}\lambda^{-3/2}(Y - \lambda)^2$$

Taking expectations, and using the results that

$$E[Y - \lambda] = 0, \quad E[(Y - \lambda)^2] = \text{Var}(Y) = \lambda$$

yields that

$$E[\sqrt{Y}] \approx \sqrt{\lambda} - \frac{1}{8\sqrt{\lambda}}$$

Hence

$$\begin{aligned}(E[\sqrt{Y}])^2 &\approx \lambda + \frac{1}{64\lambda} - \frac{1}{4} \\ &\approx \lambda - \frac{1}{4}\end{aligned}$$

and so

$$\begin{aligned}\text{Var}(\sqrt{Y}) &= E[Y] - (E[\sqrt{Y}])^2 \\ &\approx \lambda - \left(\lambda - \frac{1}{4}\right) \\ &= \frac{1}{4}\end{aligned}$$

9.9 POLYNOMIAL REGRESSION

In situations where the functional relationship between the response Y and the independent variable x cannot be adequately approximated by a linear relationship, it is sometimes possible to obtain a reasonable fit by considering a polynomial relationship. That is, we might try to fit to the data set a functional relationship of the form

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_r x^r + e$$

where $\beta_0, \beta_1, \ldots, \beta_r$ are regression coefficients that would have to be estimated. If the data set consists of the n pairs $(x_i, Y_i), i = 1, \ldots, n$, then the least squares estimators of $\beta_0, \ldots, \beta_r$ — call them $B_0, \ldots, B_r$ — are those values that minimize

$$\sum_{i=1}^{n} (Y_i - B_0 - B_1 x_i - B_2 x_i^2 - \cdots - B_r x_i^r)^2$$

To determine these estimators, we take partial derivatives with respect to $B_0 \ldots B_r$ of the foregoing sum of squares, and then set these equal to 0 so as to determine the minimizing values. On doing so, and then rearranging the resulting equations, we obtain that the least squares estimators $B_0, B_1, \ldots, B_r$ satisfy the following set of $r + 1$ linear equations called the *normal equations*.

$$\begin{aligned}
\sum_{i=1}^{n} Y_i &= B_0 n + B_1 \sum_{i=1}^{n} x_i + B_2 \sum_{i=1}^{n} x_i^2 + \cdots + B_r \sum_{i=1}^{n} x_i^r \\
\sum_{i=1}^{n} x_i Y_i &= B_0 \sum_{i=1}^{n} x_i + B_1 \sum_{i=1}^{n} x_i^2 + B_2 \sum_{i=1}^{n} x_i^3 + \cdots + B_r \sum_{i=1}^{n} x_i^{r+1} \\
\sum_{i=1}^{n} x_i^2 Y_i &= B_0 \sum_{i=1}^{n} x_i^2 + B_1 \sum_{i=1}^{n} x_i^3 + \cdots + B_r \sum_{i=1}^{n} x_i^{r+2} \\
&\vdots \qquad\qquad \vdots \qquad\qquad\qquad\qquad \vdots \\
\sum_{i=1}^{n} x_i^r Y_i &= B_0 \sum_{i=1}^{n} x_i^r + B_1 \sum_{i=1}^{n} x_i^{r+1} + \cdots + B_r \sum_{i=1}^{n} x_i^{2r}
\end{aligned}$$

In fitting a polynomial to a set of data pairs, it is often possible to determine the necessary degree of the polynomial by a study of the scatter diagram. We emphasize that one should always use the lowest possible degree that appears to adequately describe the data. [Thus, for instance, whereas it is usually possible to find a polynomial of degree n that passes through all the n pairs $(x_i, Y_i), i = 1, \ldots, n$, it would be hard to ascribe much confidence to such a fit.]

Even more so than in linear regression, it is extremely risky to use a polynomial fit to predict the value of a response at an input level x_0 that is far away from the input levels $x_i, i = 1, \ldots, n$ used in finding the polynomial fit. (For one thing, the polynomial fit may be valid only in a region around the $x_i, i = 1, \ldots, n$ and not including x_0.)

EXAMPLE 9.9a Fit a polynomial to the following data.

x	Y
1	20.6
2	30.8
3	55
4	71.4
5	97.3
6	131.8
7	156.3
8	197.3
9	238.7
10	291.7

SOLUTION A plot of these data (see Figure 9.13) indicates that a quadratic relationship

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

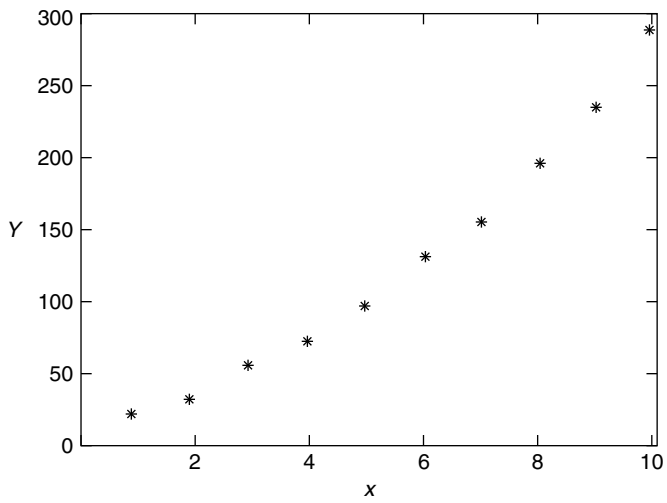

FIGURE 9.13

might hold. Since

$$\sum_i x_i = 55, \quad \sum_i x_i^2 = 385, \quad \sum_i x_i^3 = 3{,}025, \quad \sum_i x_i^4 = 25{,}333$$
$$\sum_i Y_i = 1{,}291.1, \quad \sum_i x_i Y_i = 9{,}549.3, \quad \sum_i x_i^2 Y_i = 77{,}758.9$$

the least squares estimates are the solution of the following set of equations.

$$\begin{aligned} 1{,}291.1 &= 10B_0 + 55B_1 + 385B_2 \\ 9{,}549.3 &= 55B_0 + 385B_1 + 3{,}025B_2 \\ 77{,}758.9 &= 385B_0 + 3{,}025B_1 + 25{,}333B_2 \end{aligned} \tag{9.9.1}$$

Solving these equations (see the remark following this example) yields that the least squares estimates are

$$B_0 = 12.59326, \quad B_1 = 6.326172, \quad B_2 = 2.122818$$

Thus, the estimated quadratic regression equation is

$$Y = 12.59 + 6.33x + 2.12x^2$$

This equation, along with the data, is plotted in Figure 9.14. ■

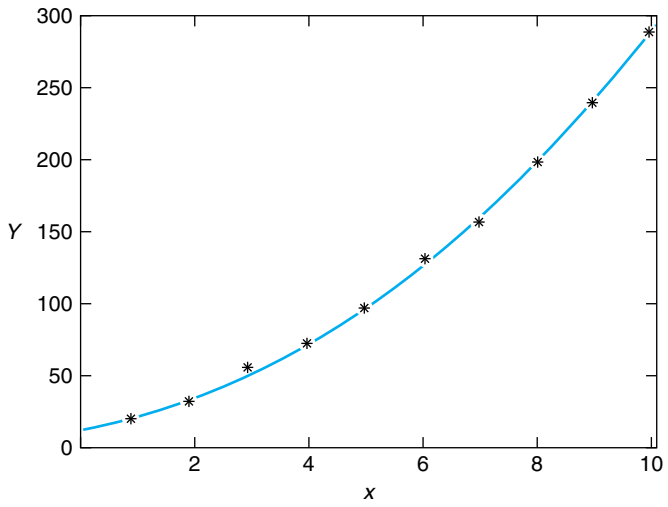

FIGURE 9.14

REMARK

In matrix notation Equation 9.9.1 can be written as

$$\begin{bmatrix} 1{,}291.1 \\ 9{,}549.3 \\ 77{,}758.9 \end{bmatrix} = \begin{bmatrix} 10 & 55 & 385 \\ 55 & 385 & 3{,}025 \\ 385 & 3{,}025 & 25{,}333 \end{bmatrix} \begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix}$$

which has the solution

$$\begin{bmatrix} B_0 \\ B_1 \\ B_2 \end{bmatrix} = \begin{bmatrix} 10 & 55 & 385 \\ 55 & 385 & 3{,}025 \\ 385 & 3{,}025 & 25{,}333 \end{bmatrix}^{-1} \begin{bmatrix} 1{,}291.1 \\ 9{,}549.3 \\ 77{,}758.9 \end{bmatrix}$$

*9.10 MULTIPLE LINEAR REGRESSION

In the majority of applications, the response of an experiment can be predicted more adequately not on the basis of a single independent input variable but on a collection of such variables. Indeed, a typical situation is one in which there are a set of, say, k input variables and the response Y is related to them by the relation

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + e$$

where $x_j, j = 1, \ldots, k$ is the level of the jth input variable and e is a random error that we shall assume is normally distributed with mean 0 and (constant) variance σ^2. The parameters $\beta_0, \beta_1, \ldots, \beta_k$ and σ^2 are assumed to be unknown and must be estimated from the data, which we shall suppose will consist of the values of $Y_1, \ldots, Y_n$ where Y_i is the response level corresponding to the k input levels $x_{i1}, x_{i2}, \ldots, x_{ik}$. That is, the Y_i are related to these input levels through

$$E[Y_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}$$

If we let $B_0, B_1, \ldots, B_k$ denote estimators of $\beta_0, \ldots, \beta_k$, then the sum of the squared differences between the Y_i and their estimated expected values is

$$\sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2$$

The least squares estimators are those values of $B_0, B_1, \ldots, B_k$ that minimize the foregoing.

---

* Optional section.

To determine the least squares estimators, we repeatedly take partial derivatives of the preceding sum of squares first with respect to B_0, then to $B_1, \ldots,$ then to B_k. On equating these $k+1$ equations to 0, we obtain the following set of equations:

$$\begin{aligned}
&\sum_{i=1}^{n}(Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik}) = 0\\
&\sum_{i=1}^{n} x_{i1}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}) = 0\\
&\sum_{i=1}^{n} x_{i2}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}) = 0\\
&\quad\vdots\\
&\sum_{i=1}^{n} x_{ik}(Y_i - B_0 - B_1 x_{i1} - \cdots - B_i x_{ik}) = 0
\end{aligned}$$

Rewriting these equations yields that the least squares estimators $B_0, B_1, \ldots, B_k$ satisfy the following set of linear equations, called the *normal equations*:

$$\sum_{i=1}^{n} Y_i = nB_0 + B_1\sum_{i=1}^{n} x_{i1} + B_2\sum_{i=1}^{n} x_{i2} + \cdots + B_k\sum_{i=1}^{n} x_{ik} \tag{9.10.1}$$

$$\sum_{i=1}^{n} x_{i1}Y_i = B_0\sum_{i=1}^{n} x_{i1} + B_1\sum_{i=1}^{n} x_{i1}^2 + B_2\sum_{i=1}^{n} x_{i1}x_{i2} + \cdots + B_k\sum_{i=1}^{n} x_{i1}x_{ik}$$

$$\vdots$$

$$\sum_{i=1}^{k} x_{ik}Y_i = B_0\sum_{i=1}^{n} x_{ik} + B_1\sum_{i=1}^{n} x_{ik}x_{i1} + B_2\sum_{i=1}^{n} x_{ik}x_{i2} + \cdots + B_k\sum_{i=1}^{n} x_{ik}^2$$

Before solving the normal equations, it is convenient to introduce matrix notation. If we let

$$\mathbf{Y} = \begin{bmatrix} Y_1\\ Y_2\\ \vdots\\ Y_n \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k}\\ 1 & x_{21} & x_{22} & \cdots & x_{2k}\\ \vdots & \vdots & \vdots & & \vdots\\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \qquad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

then $\mathbf{Y}$ is an $n \times 1$, $\mathbf{X}$ an $n \times p$, $\boldsymbol{\beta}$ a $p \times 1$, and $\mathbf{e}$ an $n \times 1$ matrix where $p \equiv k + 1$.

The multiple regression model can now be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$$

In addition, if we let

$$\mathbf{B} = \begin{bmatrix} B_0 \\ B_1 \\ \vdots \\ B_k \end{bmatrix}$$

be the matrix of least squares estimators, then the normal Equations 9.10.1 can be written as

$$\mathbf{X}'\mathbf{X}\mathbf{B} = \mathbf{X}'\mathbf{Y} \tag{9.10.2}$$

where $\mathbf{X}'$ is the transpose of $\mathbf{X}$.

To see that Equation 9.10.2 is equivalent to the normal Equations 9.10.1, note that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}$$

$$= \begin{bmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum_i x_{ik} & \sum_i x_{ik}x_{i1} & \sum_i x_{ik}x_{i2} & \cdots & \sum_i x_{ik}^2 \end{bmatrix}$$

and

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum_i Y_i \\ \sum_i x_{i1} Y_i \\ \vdots \\ \sum_i x_{ik} Y_i \end{bmatrix}$$

It is now easy to see that the matrix equation

$$\mathbf{X}'\mathbf{X}\mathbf{B} = \mathbf{X}'\mathbf{Y}$$

is equivalent to the set of normal Equations 9.10.1. Assuming that $(\mathbf{X}'\mathbf{X})^{-1}$ exists, which is usually the case, we obtain, upon multiplying it by both sides of the foregoing, that the least squares estimators are given by

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \tag{9.10.3}$$

Program 9.10 computes the least squares estimates, the inverse matrix $(\mathbf{X}'\mathbf{X})^{-1}$, and SS_R.

EXAMPLE 9.10a The data in Table 9.4 relate the suicide rate to the population size and the divorce rate at eight different locations.

TABLE 9.4

Location	Population in Thousands	Divorce Rate per 100,000	Suicide Rate per 100,000
Akron, OH	679	30.4	11.6
Anaheim, CA	1,420	34.1	16.1
Buffalo, NY	1,349	17.2	9.3
Austin, TX	296	26.8	9.1
Chicago, IL	6,975	29.1	8.4
Columbia, SC	323	18.7	7.7
Detroit, MI	4,200	32.6	11.3
Gary, IN	633	32.5	8.4

Fit a multiple linear regression model to these data. That is, fit a model of the form

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

where Y is the suicide rate, x_1 is the population, and x_2 is the divorce rate.

SOLUTION We run Program 9.10, and results are shown in Figures 9.15, 9.16, and 9.17. Thus the estimated regression line is

$$Y = 3.5073 - .0002x_1 + .2609x_2$$

The value of β_1 indicates that the population does not play a major role in predicting the suicide rate (at least when the divorce rate is also given). Perhaps the population density, rather than the actual population, would have been more useful. ■

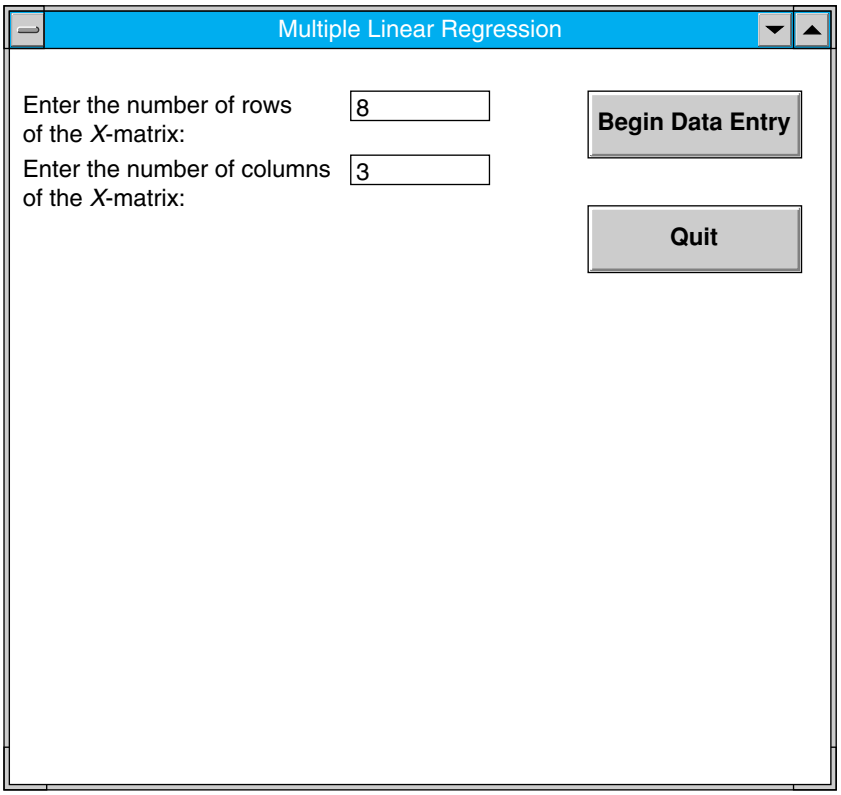

FIGURE 9.15

It follows from Equation 9.10.3 that the least squares estimators $B_0, B_1, \ldots, B_k$ — the elements of the matrix $\mathbf{B}$ — are all linear combinations of the independent normal random variables $Y_1, \ldots, Y_n$ and so will also be normally distributed. Indeed in such a situation — namely, when each member of a set of random variables can be expressed as a linear combination of independent normal random variables — we say that the set of random variables has a joint *multivariate normal distribution*.

The least squares estimators turn out to be unbiased. This can be shown as follows:

$$
\begin{aligned}
E[\mathbf{B}] &= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}] \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{e})] \quad \text{since } \mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \\
&= E[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}] \\
&= E[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}] \\
&= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}[\mathbf{e}] \\
&= \boldsymbol{\beta}
\end{aligned}
$$

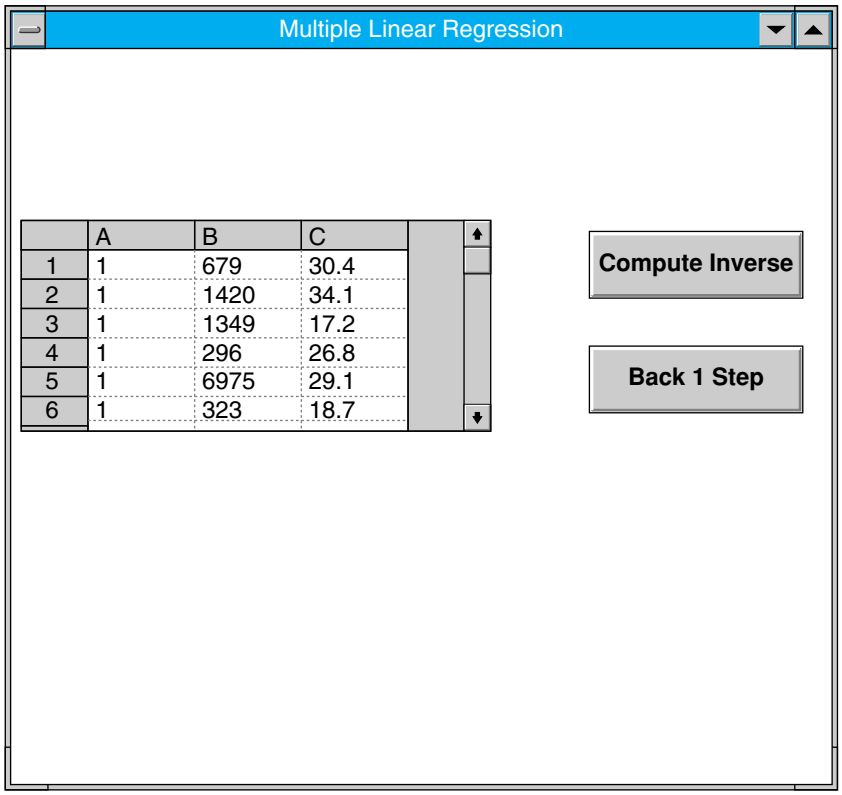

FIGURE 9.16

The variances of the least squares estimators can be obtained from the matrix $(\mathbf{X}'\mathbf{X})^{-1}$. Indeed, the values of this matrix are related to the covariances of the B_i's. Specifically, the element in the $(i+1)$st row, $(j+1)$st column of $(\mathbf{X}'\mathbf{X})^{-1}$ is equal to $\text{Cov}(B_i, B_j)/\sigma^2$.

To verify the preceding statement concerning $\text{Cov}(B_i, B_j)$, let

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

Since $\mathbf{X}$ is an $n \times p$ matrix and $\mathbf{X}'$ a $p \times n$ matrix, it follows that $\mathbf{X}'\mathbf{X}$ is $p \times p$, as is $(\mathbf{X}'\mathbf{X})^{-1}$, and so $\mathbf{C}$ will be a $p \times n$ matrix. Let C_{ij} denote the element in row i, column j of this matrix. Now

$$\begin{bmatrix} B_0 \\ \vdots \\ B_{i-1} \\ \vdots \\ B_k \end{bmatrix} = \mathbf{B} = \mathbf{CY} = \begin{bmatrix} C_{11} & \cdots & C_{1n} \\ \vdots & & \vdots \\ C_{i1} & \cdots & C_{in} \\ \vdots & & \vdots \\ C_{p1} & \cdots & C_{pn} \end{bmatrix} \begin{bmatrix} Y_1 \\ \\ \vdots \\ \\ Y_n \end{bmatrix}$$

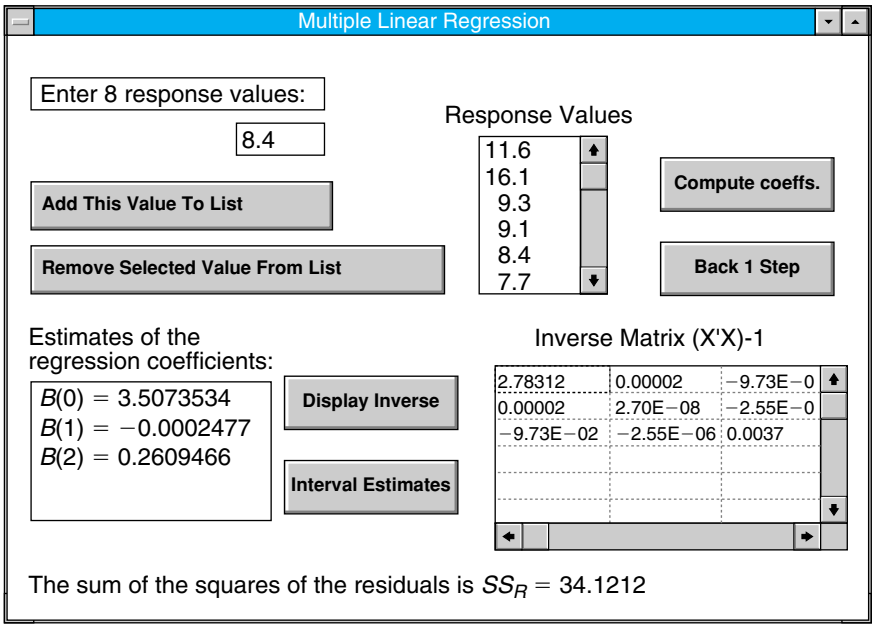

FIGURE 9.17

and so

$$B_{i-1} = \sum_{l=1}^{n} C_{il} Y_l$$

$$B_{j-1} = \sum_{r=1}^{n} C_{jr} Y_r$$

Hence

$$\begin{aligned} \text{Cov}(B_{i-1}, B_{j-1}) &= \text{Cov}\left(\sum_{l=1}^{n} C_{il} Y_l, \sum_{r=1}^{n} C_{jr} Y_r \right) \\ &= \sum_{r=1}^{n} \sum_{l=1}^{n} C_{il} C_{jr} \, \text{Cov}(Y_l, Y_r) \end{aligned}$$

Now Y_l and Y_r are independent when $l \neq r$, and so

$$\text{Cov}(Y_l, Y_r) = \begin{cases} 0 & \text{if } l \neq r \\ \text{Var}(Y_r) & \text{if } l = r \end{cases}$$

Since $\text{Var}(Y_r) = \sigma^2$, we see that

$$\begin{aligned}\text{Cov}(B_{i-1}, B_{j-1}) &= \sigma^2 \sum_{r=1}^{n} C_{ir} C_{jr} \\ &= \sigma^2 (\mathbf{CC}')_{ij}\end{aligned} \tag{9.10.4}$$

where $(\mathbf{CC}')_{ij}$ is the element in row i, column j of $\mathbf{CC}'$.

If we now let $\text{Cov}(\mathbf{B})$ denote the matrix of covariances — that is,

$$\text{Cov}(\mathbf{B}) = \begin{bmatrix} \text{Cov}(B_0, B_0) & \cdots & \text{Cov}(B_0, B_k) \\ \vdots & & \vdots \\ \text{Cov}(B_k, B_0) & \cdots & \text{Cov}(B_k, B_k) \end{bmatrix}$$

then it follows from Equation 9.10.4 that

$$\text{Cov}(\mathbf{B}) = \sigma^2 \mathbf{CC}' \tag{9.10.5}$$

Now

$$\begin{aligned}\mathbf{C}' &= \left((\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}' \right)' \\ &= \mathbf{X} \left((\mathbf{X}'\mathbf{X})^{-1} \right)' \\ &= \mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

where the last equality follows since $(\mathbf{X}'\mathbf{X})^{-1}$ is symmetric (since $\mathbf{X}'\mathbf{X}$ is) and so is equal to its transpose. Hence

$$\begin{aligned}\mathbf{CC}' &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

and so we can conclude from Equation 9.10.5 that

$$\text{Cov}(\mathbf{B}) = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1} \tag{9.10.6}$$

Since $\text{Cov}(B_i, B_i) = \text{Var}(B_i)$, it follows that the variances of the least squares estimators are given by σ^2 multiplied by the diagonal elements of $(\mathbf{X}'\mathbf{X})^{-1}$.

The quantity σ^2 can be estimated by using the sum of squares of the residuals. That is, if we let

$$SS_R = \sum_{i=1}^{n} (Y_i - B_0 - B_1 x_{i1} - B_2 x_{i2} - \cdots - B_k x_{ik})^2$$

then it can be shown that

$$\frac{SS_r}{\sigma^2} \sim \chi^2_{n-(k+1)}$$

and so

$$E\left[\frac{SS_R}{\sigma^2}\right] = n - k - 1$$

or

$$E[SS_R/(n-k-1)] = \sigma^2$$

That is, $SS_R/(n-k-1)$ is an unbiased estimator of σ^2. In addition, as in the case of simple linear regression, SS_R will be independent of the least squares estimators $B_0, B_1, \ldots, B_k$.

REMARK

If we let r_i denote the ith residual

$$r_i = Y_i - B_0 - B_1 x_{i1} - \cdots - B_k x_{ik}, \qquad i = 1, \ldots, n$$

then

$$\mathbf{r} = \mathbf{Y} - \mathbf{XB}$$

where

$$\mathbf{r} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_n \end{bmatrix}$$

Hence, we may write

$$\begin{aligned} SS_R &= \sum_{i=1}^{n} r_i^2 \\ &= \mathbf{r}'\mathbf{r} \\ &= (\mathbf{Y} - \mathbf{XB})'(\mathbf{Y} - \mathbf{XB}) \\ &= [\mathbf{Y}' - (\mathbf{XB})'](\mathbf{Y} - \mathbf{XB}) \end{aligned} \tag{9.10.7}$$

$$
\begin{aligned}
&= (\mathbf{Y}' - \mathbf{B}'\mathbf{X}')(\mathbf{Y} - \mathbf{XB}) \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{XB} - \mathbf{B}'\mathbf{X}'\mathbf{Y} + \mathbf{B}'\mathbf{X}'\mathbf{XB} \\
&= \mathbf{Y}'\mathbf{Y} - \mathbf{Y}'\mathbf{XB}
\end{aligned}
$$

where the last equality follows from the normal equations

$$\mathbf{X}'\mathbf{XB} = \mathbf{X}'\mathbf{Y}$$

Because $\mathbf{Y}'$ is $1 \times n$, $\mathbf{X}$ is $n \times p$, and $\mathbf{B}$ is $p \times 1$, it follows that $\mathbf{Y}'\mathbf{XB}$ is a 1×1 matrix. That is, $\mathbf{Y}'\mathbf{XB}$ is a scalar and thus is equal to its transpose, which shows that

$$
\begin{aligned}
\mathbf{Y}'\mathbf{XB} &= (\mathbf{Y}'\mathbf{XB})' \\
&= \mathbf{B}'\mathbf{X}'\mathbf{Y}
\end{aligned}
$$

Hence, using Equation 9.10.7 we have proven the following identity:

$$SS_R = \mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y}$$

The foregoing is a useful computational formula for SS_R (though one must be careful of possible roundoff error when using it).

EXAMPLE 9.10b For the data of Example 9.10a, we computed that $SS_R = 34.12$. Since $n = 8, k = 2$, the estimate of σ^2 is $34.12/5 = 6.824$. ■

EXAMPLE 9.10c The diameter of a tree at its breast height is influenced by many factors. The data in Table 9.5 relate the diameter of a particular type of eucalyptus tree to its age, average rainfall at its site, site's elevation, and the wood's mean specific gravity. (The data come from R. G. Skolmen, 1975, "Shrinkage and Specific Gravity Variation in Robusta Eucalyptus Wood Grown in Hawaii." USDA Forest Service PSW-298.)

Assuming a linear regression model of the form

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

where x_1 is the age, x_2 is the elevation, x_3 is the rainfall, x_4 is the specific gravity, and Y is the tree's diameter, test the hypothesis that $\beta_2 = 0$. That is, test the hypothesis that, given the other three factors, the elevation of the tree does not affect its diameter.

SOLUTION To test this hypothesis, we begin by running Program 9.10, which yields, among other things, the following:

$$(\mathbf{X}'\mathbf{X})^{-1}_{3,3} = .379, \qquad SS_R = 19.262, \qquad B_2 = .075$$

TABLE 9.5

	Age (years)	Elevation (1,000 ft)	Rainfall (inches)	Specific Gravity	Diameter at Breast Height (inches)
1	44	1.3	250	.63	18.1
2	33	2.2	115	.59	19.6
3	33	2.2	75	.56	16.6
4	32	2.6	85	.55	16.4
5	34	2.0	100	.54	16.9
6	31	1.8	75	.59	17.0
7	33	2.2	85	.56	20.0
8	30	3.6	75	.46	16.6
9	34	1.6	225	.63	16.2
10	34	1.5	250	.60	18.5
11	33	2.2	255	.63	18.7
12	36	1.7	175	.58	19.4
13	33	2.2	75	.55	17.6
14	34	1.3	85	.57	18.3
15	37	2.6	90	.62	18.8

It now follows from Equation 9.10.6 that

$$\text{Var}(B_2) = .379\sigma^2$$

Since B_2 is normal and

$$E[B_2] = \beta_2$$

we see that

$$\frac{B_2 - \beta_2}{.616\sigma} \sim N(0, 1)$$

Replacing σ by its estimator $SS_R/10$ transforms the foregoing standard normal distribution into a t-distribution with $10(= n - k - 1)$ degrees of freedom. That is,

$$\frac{B_2 - \beta_2}{.616\sqrt{SS_R/10}} \sim t_{10}$$

Hence, if $\beta_2 = 0$ then

$$\frac{\sqrt{10/SS_R}\, B_2}{.616}$$

Since the value of the preceding statistic is $(\sqrt{10/19.262})(.075)/.616 = .088$, the p-value of the test of the hypothesis that $\beta_2 = 0$ is

$$\begin{aligned} p\text{-value} &= P\{|T_{10}| > .088\} \\ &= 2P\{T_{10} > .088\} \\ &= .9316 \quad \text{by Program 5.8.2.A} \end{aligned}$$

Hence, the hypothesis is accepted (and, in fact, would be accepted at any significance level less than .9316). ■

REMARK

The quantity

$$R^2 = 1 - \frac{SS_R}{\sum_i (Y_i - \overline{Y})^2}$$

which measures the amount of reduction in the sum of squares of the residuals when using the model

$$Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + e$$

as opposed to the model

$$Y = \beta_0 + e$$

is called the *coefficient of multiple determination*.

9.10.1 Predicting Future Responses

Let us now suppose that a series of experiments is to be performed using the input levels $x_1, \ldots, x_k$. Based on our data, consisting of the prior responses $Y_1, \ldots, Y_n$, suppose we would like to estimate the mean response. Since the mean response is

$$E[Y|\mathbf{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

a point estimate of it is simply $\sum_{i=0}^{k} B_i x_i$ where $x_0 \equiv 1$.

To determine a confidence interval estimator, we need the distribution of $\sum_{i=0}^{k} B_i x_i$. Because it can be expressed as a linear combination of the independent normal random variables $Y_i, i = 1, \ldots, n$, it follows that it is also normally distributed. Its mean and variance are obtained as follows:

$$\begin{aligned} E\left[\sum_{i=0}^{k} x_i B_i\right] &= \sum_{i=0}^{k} x_i E[B_i] \qquad (9.10.8) \\ &= \sum_{i=0}^{k} x_i \beta_i \qquad \text{since } E[B_i] = \beta_i \end{aligned}$$

That is, it is an unbiased estimator. Also, using the fact that the variance of a random variable is equal to the covariance between that random variable and itself, we see that

$$\begin{aligned}\text{Var}\left(\sum_{i=0}^{k} x_i B_i\right) &= \text{Cov}\left(\sum_{i=0}^{k} x_i B_i, \sum_{j=0}^{k} x_j B_j\right) \\ &= \sum_{i=0}^{k}\sum_{j=0}^{k} x_i x_j \text{Cov}(B_i, B_j)\end{aligned} \tag{9.10.9}$$

If we let x denote the matrix

$$\mathbf{x} = \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_k \end{bmatrix}$$

then, recalling that $\text{Cov}(B_i, B_j)/\sigma^2$ is the element in the $(i+1)$st row and $(j+1)$st column of $(\mathbf{X}'\mathbf{X})^{-1}$, we can express Equation 9.10.9 as

$$\text{Var}\left(\sum_{i=0}^{k} x_i B_i\right) = \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}\sigma^2 \tag{9.10.10}$$

Using Equations 9.10.8 and 9.10.10, we see that

$$\frac{\sum_{i=0}^{k} x_i B_i - \sum_{i=0}^{k} x_i \beta_i}{\sigma\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim N(0, 1)$$

If we now replace σ by its estimator $\sqrt{SS_R/(n-k-1)}$ we obtain, by the usual argument, that

$$\frac{\sum_{i=0}^{k} x_i B_i - \sum_{i=0}^{k} x_i \beta_i}{\sqrt{\frac{SS_R}{(n-k-1)}}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim t_{n-k-1}$$

which gives rise to the following confidence interval estimator of $\sum_{i=0}^{k} x_i \beta_i$.

Confidence Interval Estimate of $E[Y \mid \mathbf{x}] = \sum_{i=0}^{k} x_i \beta_i$, $(x_0 \equiv 1)$

A $100(1-a)$ percent confidence interval estimate of $\sum_{i=0}^{k} x_i \beta_i$ is given by

$$\sum_{i=0}^{k} x_i b_i \pm \sqrt{\frac{SS_r}{(n-k-1)}}\sqrt{\mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}} \quad t_{a/2,n-k-1}$$

TABLE 9.6

Hardness	Copper Content	Annealing Temperature (units of 1,000°F)
79.2	.02	1.05
64.0	.03	1.20
55.7	.03	1.25
56.3	.04	1.30
58.6	.10	1.30
84.3	.15	1.00
70.4	.15	1.10
61.3	.09	1.20
51.3	.13	1.40
49.8	.09	1.40

where $b_0, \ldots, b_k$ are the values of the least squares estimators $B_0, B_1, \ldots, B_k$, and ss_r is the value of SS_R.

EXAMPLE 9.10d A steel company is planning to produce cold reduced sheet steel consisting of .15 percent copper at an annealing temperature of 1,150 (degrees F), and is interested in estimating the average (Rockwell 30-T) hardness of a sheet. To determine this, they have collected the data shown in Table 9.6 on 10 different specimens of sheet steel having different copper contents and annealing temperatures. Estimate the average hardness and determine an interval in which it will lie with 95 percent confidence.

SOLUTION To solve this, we first run Program 9.10, which gives the results shown in Figures 9.18, 9.19, and 9.20.

Hence, a point estimate of the expected hardness of sheets containing .15 percent copper at an annealing temperature of 1,150 is 69.862. In addition, since $t_{.025,7} = 2.365$, a 95 percent confidence interval for this value is

$$69.862 \pm 4.083 \quad \blacksquare$$

When it is only a single experiment that is going to be performed at the input levels $x_1, \ldots, x_k$, we are usually more concerned with predicting the actual response than its mean value. That is, we are interested in utilizing our data set $Y_1, \ldots, Y_n$ to predict

$$Y(\mathbf{x}) = \sum_{i=0}^{k} \beta_i x_i + e, \qquad \text{where } x_0 = 1$$

A point prediction is given by $\sum_{i=0}^{k} B_i x_i$ where B_i is the least squares estimator of β_i based on the set of prior responses $Y_1, \ldots, Y_n, i = 1, \ldots, k$.

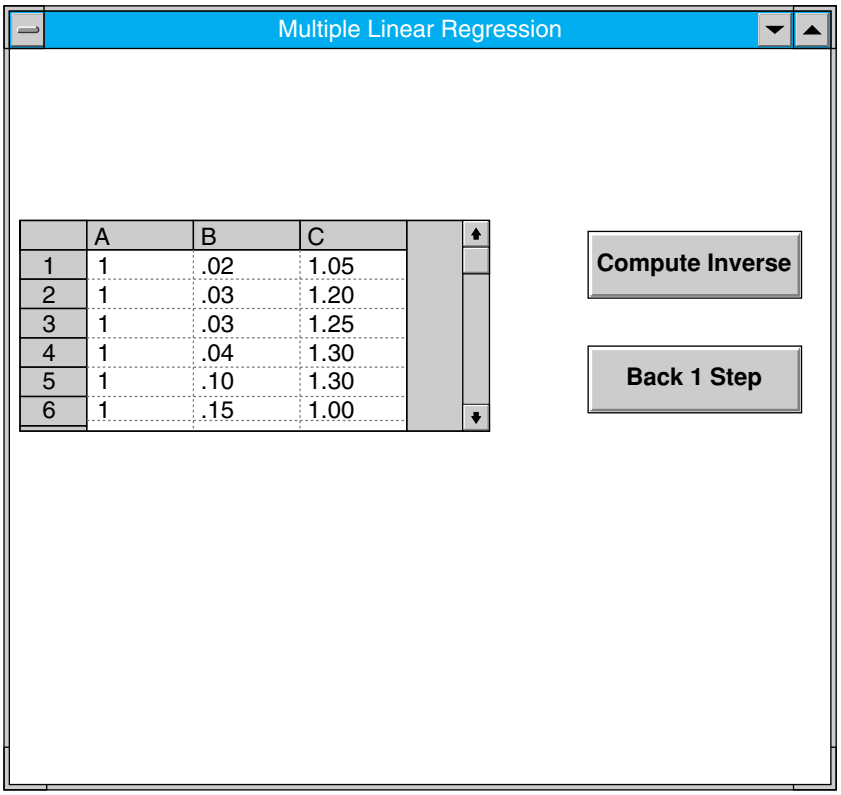

FIGURE 9.18

To determine a prediction interval for $Y(\mathbf{x})$, note first that since $B_0, \ldots, B_k$ are based on prior responses, it follows that they are independent of $Y(\mathbf{x})$. Hence, it follows that $Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i$ is normal with mean 0 and variance given by

$$\begin{aligned} \operatorname{Var}\left[Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i\right] &= \operatorname{Var}[Y(\mathbf{x})] + \operatorname{Var}\left(\sum_{i=0}^{k} B_i x_i\right) \qquad \text{by independence} \\ &= \sigma^2 + \sigma^2 \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x} \qquad \text{from Equation 9.10.10} \end{aligned}$$

and so

$$\frac{Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i}{\sigma\sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim N(0, 1)$$

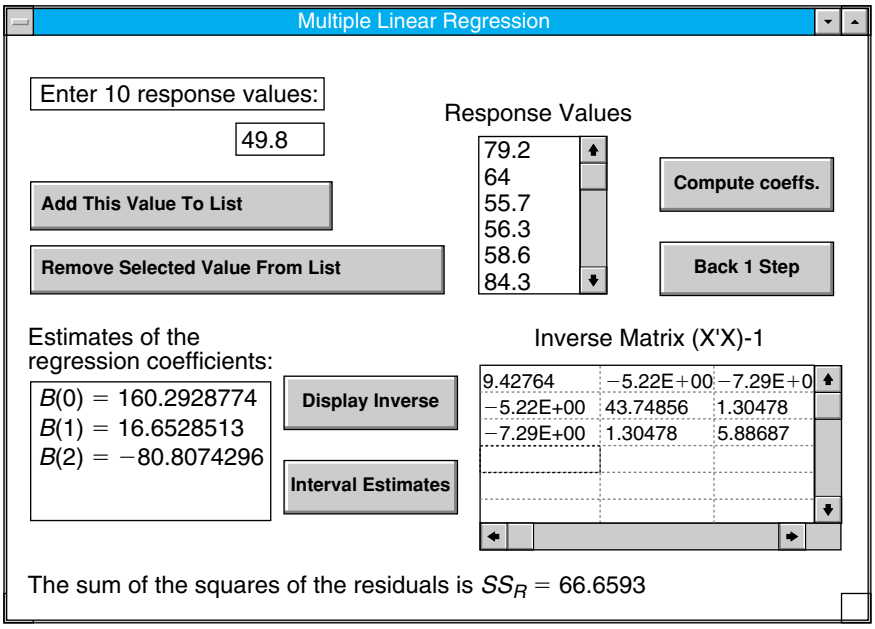

FIGURE 9.19

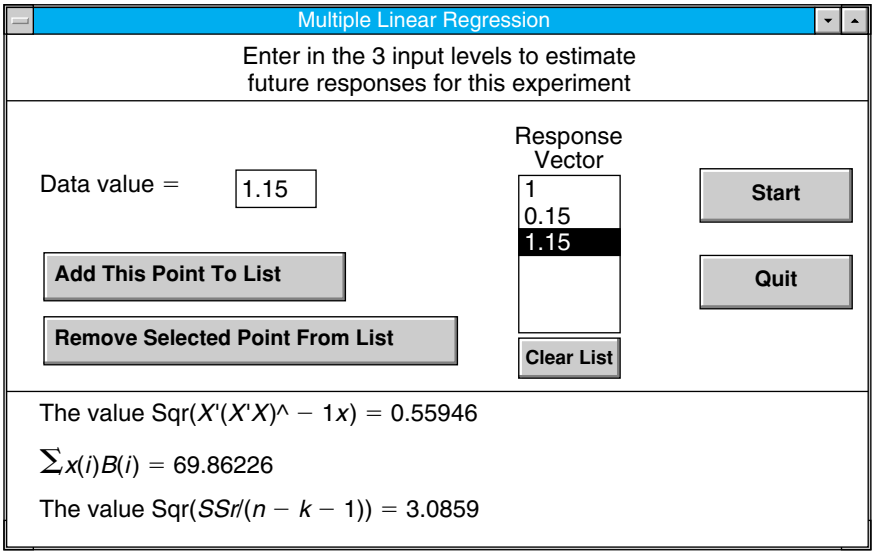

FIGURE 9.20

which yields, upon replacing σ by its estimator, that

$$\frac{Y(\mathbf{x}) - \sum_{i=0}^{k} B_i x_i}{\sqrt{\frac{SS_R}{(n-k-1)}}\sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}}} \sim t_{n-k-1}$$

We thus have:

Prediction Interval for *Y*(x)

With $100(1-a)$ percent confidence $Y(\mathbf{x})$ will lie between

$$\sum_{i=0}^{k} x_i b_i \pm \sqrt{\frac{ss_r}{(n-k-1)}}\sqrt{1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}} \quad t_{a/2,n-k-1}$$

where $b_0, \ldots, b_k$ are the values of the least squares estimators $B_0, B_1, \ldots, B_k$, and ss_r is the value of SS_R.

EXAMPLE 9.10e If in Example 9.10d we were interested in determining an interval in which a single steel sheet, produced with a carbon content of .15 percent and at an annealing temperature of 1,150°F, would lie, then the midpoint of the prediction interval would be as given before. However, the half-length of this prediction interval would differ from the confidence interval for the mean value by the factor $\sqrt{1.313}/\sqrt{.313} \approx 2.048$. Consequently, the 95 percent prediction interval is

$$69.862 \pm (4.083)(2.048) = 69.862 \pm 8.363 \quad \blacksquare$$

9.10.2 Dummy Variables for Categorical Data

Suppose that in determining an appropriate multiple regression model for predicting a person's blood cholesterol level a researcher has decided on the following five independent variables:

1. number of pounds overweight
2. number of pounds underweight
3. average number of hours of exercise per week
4. average number of calories due to saturated fats eaten daily
5. whether a smoker or not

Whereas each of the first four variables takes on values in some interval, the final variable is a categorical variable that indicates whether the person under consideration has or

does not have a certain characteristic (which, in this case, is the characteristic of being a smoker). To determine which category the person belongs to, let

$$x_5 = \begin{cases} 1, & \text{if person is a smoker} \\ 0, & \text{if person is not a smoker} \end{cases}$$

The researcher can now try to fit the multiple regression model

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

where x_1 is the number of pounds the individual is overweight, x_2 is the number of pounds the individual is underweight, x_3 is the average number of hours the individual exercises per week, x_4 is the average number of calories due to saturated fats that is eaten daily, x_5 is as above, and Y is the individual's cholesterol level. The variable x_5 is called a *dummy variable*, as its only purpose is to indicate whether or not the Y value is determined from data having a particular characteristic.

One might wonder at this point why a dummy variable is used rather than just running separate multiple regressions for smokers and for nonsmokers. The main reason for using a dummy variable is that we can use all the data in a single regression, thus yielding more precise estimates than if we broke the data into two parts (one for smokers and the other for nonsmokers) and then used the divided data to run separate regressions. However, it is important to understand what is being assumed when dummy variables are being used. Namely, we are assuming that if Y_s stands for the cholesterol level of a smoker, and Y_n, the cholesterol level of a nonsmoker, then for specified values of x_1, x_2, x_3, and x_4

$$E[Y_s] = \beta_0 + \beta_5 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

and

$$E[Y_n] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

In other words, in using the model with a dummy variable we are assuming that if a smoker and nonsmoker had the same values for the four quantitative variables x_1, x_2, x_3, x_4 then the difference between their mean cholesterol levels would always be a constant, no matter what are the values of x_1, x_2, x_3, x_4. Thus, for instance, the dummy variable model assumes that the amount that one is overweight has the same effect on raising the expected cholesterol level on a smoker as it does on a nonsmoker. Because this might seem like a questionable assumption, it is typically preferable when the data set is large enough to use two regression models rather than combining into one model by the use of a dummy variable.

In situations where there are multiple qualitative characteristics that the researcher feels are relevant it might be necessary to utilize dummy variables, for otherwise the data set may become too fragmented to yield reliable estimates of the regression parameters. So, for instance, if the researcher felt that the sex of the person was also a relevant factor, then the

researcher could utilize a multiple regression model having two dummy variables: namely x_5 and

$$x_6 = \begin{cases} 1, & \text{if person is a male} \\ 0, & \text{if person is a female} \end{cases}$$

9.11 LOGISTIC REGRESSION MODELS FOR BINARY OUTPUT DATA

In this section we consider experiments that result in either a success or a failure. We will suppose that these experiments can be performed at various levels, and that an experiment performed at level x will result in a success with probability $p(x)$, $-\infty < x < \infty$. If $p(x)$ is of the form

$$p(x) = \frac{e^{a+bx}}{1 + e^{a+bx}}$$

then the experiments are said to come from a *logistic regression* model and $p(x)$ is called the *logistics regression function*. If $b > 0$, then $p(x) = 1/[e^{-(a+bx)} + 1]$ is an increasing function that converges to 1 as $x \to \infty$; if $b < 0$, then $p(x)$ is a decreasing function that converges to 0 as $x \to \infty$. (When $b = 0$, $p(x)$ is constant.) Plots of logistics regression functions are given in Figure 9.21. Notice the **s**-shape of these curves.

Writing $p(x) = 1 - [1/(1 + e^{a+bx})]$ and differentiating give that

$$\frac{\partial}{\partial x} p(x) = \frac{be^{a+bx}}{(1 + e^{a+bx})^2} = b\,p(x)[1 - p(x)]$$

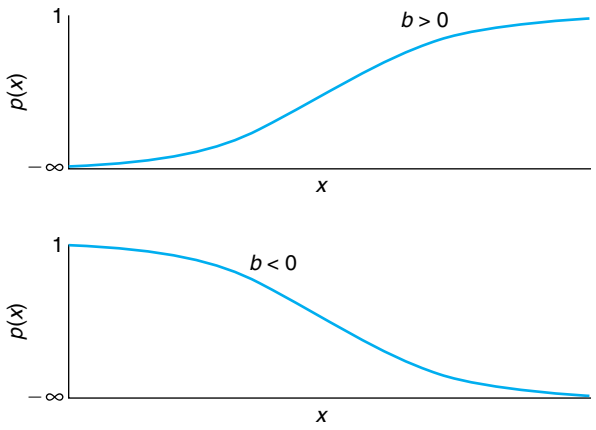

FIGURE 9.21 *Logistic regression functions.*

Thus the rate of change of $p(x)$ depends on x and is largest at those values of x for which $p(x)$ is near .5. For instance, at the value x such that $p(x) = .5$, the rate of change is $\frac{\partial}{\partial x}p(x) = .25b$, whereas at that value x for which $p(x) = .8$ the rate of change is $.16b$.

If we let $o(x)$ be the odds for success when the experiment is run at level x, then

$$o(x) = \frac{p(x)}{1 - p(x)} = e^{a+bx}$$

Thus, when $b > 0$, the odds increase exponentially in the input level x; when $b < 0$, the odds decrease exponentially in the input level x. Taking logs of the preceding shows that the log odds, called the *logit*, is a linear function:

$$\log[o(x)] = a + bx$$

The parameters a and b of the logistic regression function are assumed to be unknown and need to be estimated. This can be accomplished by using the maximum likelihood approach. That is, suppose that the experiment is to be performed at levels $x_1, \ldots, x_k$. Let Y_i be the result (either 1 if a success or 0 if a failure) of the experiment when performed at level x_i. Then, using the Bernoulli density function (that is, the binomial density for a single trial), gives

$$P\{Y_i = y_i\} = [p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i} = \left(\frac{e^{a+bx_i}}{1 + e^{a+bx_i}}\right)^{y_i} \left(\frac{1}{1 + e^{a+bx_i}}\right)^{1-y_i}, \quad y_i = 0, 1$$

Thus, the probability that the experiment at level x_i results in outcome y_i, for all $i = 1, \ldots, k$, is

$$\begin{aligned} P\{Y_i = y_i, i = 1, \ldots, k\} &= \prod_i \left(\frac{e^{a+bx_i}}{1 + e^{a+bx_i}}\right)^{y_i} \left(\frac{1}{1 + e^{a+bx_i}}\right)^{1-y_i} \\ &= \prod_i \frac{\left(e^{a+bx_i}\right)^{y_i}}{1 + e^{a+bx_i}} \end{aligned}$$

Taking logarithms gives that

$$\log\left(P\{Y_i = y_i, i = 1, \ldots, k\}\right) = \sum_{i=1}^{k} y_i(a + bx_i) - \sum_{i=1}^{k} \log\left(1 + e^{a+bx_i}\right)$$

The maximum likelihood estimates can now be obtained by numerically finding the values of a and b that maximize the preceding likelihood. However, because the likelihood

is nonlinear this requires an iterative approach; consequently, one typically resorts to specialized software to obtain the estimates.

Whereas the logistic regression model is the most frequently used model when the response data are binary, other models are often employed. For instance in situations where it is reasonable to suppose that $p(x)$, the probability of a positive response when the input level is x, is an increasing function of x, it is often supposed that $p(x)$ has the form of a specified probability distribution function. Indeed, when $b > 0$, the logistic regression model is of this form because $p(x)$ is equal to the distribution function of a logistic random variable (Section 5.9) with parameters $\mu = -a/b$, $\nu = 1/b$. Another model of this type is the *probit* model, which supposes that for some constants, $\alpha, \beta > 0$

$$p(x) = \Phi(\alpha + \beta x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\alpha+\beta x} e^{-y^2/2}\, dy$$

In other words $p(x)$ is equal to the probability that a standard normal random variable is less than $\alpha + \beta x$.

EXAMPLE 9.11a A common assumption for whether an animal becomes sick when exposed to a chemical at dosage level x is to assume a threshold model, which supposes that each animal has a random threshold and will become ill if the dosage level exceeds that threshold. The exponential distribution has sometimes been used as the threshold distribution. For instance, a model considered in Freedman and Zeisel ("From Mouse to Man: The Quantitative Assessment of Cancer Risks," *Statistical Science*, 1988, **3**, 1, 3–56) supposes that a mouse exposed to x units of DDT (measured in ppm) will contract cancer of the liver with probability

$$p(x) = 1 - e^{-ax}, \quad x > 0$$

Because of the lack of memory of the exponential distribution, this is equivalent to assuming that if the mouse who is still healthy after receiving a (partial) dosage of level x is as good as it was before receiving any dosage.

It was reported in Freedman and Zeisel that 84 of 111 mice exposed to DDT at a level of 250 ppm developed cancer. Therefore, α can be estimated from

$$1 - e^{-250\hat{\alpha}} = \frac{84}{111}$$

or

$$\hat{\alpha} = -\frac{\log(27/111)}{250} = .005655 \quad \blacksquare$$

Problems

1. The following data relate x, the moisture of a wet mix of a certain product, to Y, the density of the finished product.

x_i	Y_i
5	7.4
6	9.3
7	10.6
10	15.4
12	18.1
15	22.2
18	24.1
20	24.8

(a) Draw a scatter diagram.
(b) Fit a linear curve to the data.

2. The following data relate the number of units of a good that were ordered as a function of the price of the good at six different locations.

Number ordered	88	112	123	136	158	172
Price	50	40	35	30	20	15

How many units do you think would be ordered if the price were 25?

3. The corrosion of a certain metallic substance has been studied in dry oxygen at 500 degrees centigrade. In this experiment, the gain in weight after various periods of exposure was used as a measure of the amount of oxygen that had reacted with the sample. Here are the data:

Hours	Percent Gain
1.0	.02
2.0	.03
2.5	.035
3.0	.042
3.5	.05
4.0	.054

(a) Plot a scatter diagram.
(b) Fit a linear relation.
(c) Predict the percent weight gain when the metal is exposed for 3.2 hours.

4. The following data indicate the relationship between x, the specific gravity of a wood sample, and Y, its maximum crushing strength in compression parallel to the grain.

x_i	y_i(psi)	x_i	y_i(psi)
.41	1,850	.39	1,760
.46	2,620	.41	2,500
.44	2,340	.44	2,750
.47	2,690	.43	2,730
.42	2,160	.44	3,120

(a) Plot a scatter diagram. Does a linear relationship seem reasonable?
(b) Estimate the regression coefficients.
(c) Predict the maximum crushing strength of a wood sample whose specific gravity is .43.

5. The following data indicate the gain in reading speed versus the number of weeks in the program of 10 students in a speed-reading program.

Number of Weeks	Speed Gain (wds/min)
2	21
3	42
8	102
11	130
4	52
5	57
9	105
7	85
5	62
7	90

(a) Plot a scatter diagram to see if a linear relationship is indicated.
(b) Find the least squares estimates of the regression coefficients.
(c) Estimate the expected gain of a student who plans to take the program for 7 weeks.

6. Infrared spectroscopy is often used to determine the natural rubber content of mixtures of natural and synthetic rubber. For mixtures of known percentages, the infrared spectroscopy gave the following readings:

Percentage	0	20	40	60	80	100
Reading	.734	.885	1.050	1.191	1.314	1.432

If a new mixture gives an infrared spectroscopy reading of 1.15, estimate its percentage of natural rubber.

7. The following table gives the 1996 SAT mean math and verbal scores in each state and the District of Columbia, along with the percentage of the states' graduating high school students that took the examination. Use data relating to the first 20 locations listed (Alabama to Maine) to develop a prediction of the mean student mathematics score in terms of the percentage of students that take the examination. Then compare your predicted values for the next 5 states (based on the percentage taking the exam in these states) with the actual mean math scores.

SAT Mean Scores by State, 1996 (recentered scale)

	1996		% Graduates Taking SAT
	Verbal	Math	
Alabama	565	558	8
Alaska	521	513	47
Arizona	525	521	28
Arkansas	566	550	6
California	495	511	45
Colorado	536	538	30
Connecticut	507	504	79
Delaware	508	495	66
Dist. of Columbia	489	473	50
Florida	498	496	48
Georgia	484	477	63
Hawaii	485	510	54
Idaho	543	536	15
Illinois	564	575	14
Indiana	494	494	57
Iowa	590	600	5

(*continued*)

	1996		% Graduates Taking SAT
	Verbal	Math	
Kansas	579	571	9
Kentucky	549	544	12
Louisiana	559	550	9
Maine	504	498	68
Maryland	507	504	64
Massachusetts	507	504	80
Michigan	557	565	11
Minnesota	582	593	9
Mississippi	569	557	4
Missouri	570	569	9
Montana	546	547	21
Nebraska	567	568	9
Nevada	508	507	31
New Hampshire	520	514	70
New Jersey	498	505	69
New Mexico	554	548	12
New York	497	499	73
North Carolina	490	486	59
North Dakota	596	599	5
Ohio	536	535	24
Oklahoma	566	557	8
Oregon	523	521	50
Pennsylvania	498	492	71
Rhode Island	501	491	69
South Carolina	480	474	57
South Dakota	574	566	5
Tennessee	563	552	14
Texas	495	500	48
Utah	583	575	4
Vermont	506	500	70
Virginia	507	496	68
Washington	519	519	47
West Virginia	526	506	17
Wisconsin	577	586	8
Wyoming	544	544	11
National Average	505	508	41

Source: The College Board.

8. Verify Equation 9.3.3, which states that

$$\text{Var}(A) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n} (x_i - \bar{x})^2}$$

9. In Problem 4,

(a) Estimate the variance of an individual response.
(b) Determine a 90 percent confidence interval for the variance.

10. Verify that

$$SS_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

11. The following table relates the number of sunspots that appeared each year from 1970 to 1983 to the number of auto accident deaths during that year. Test the hypothesis that the number of auto deaths is not affected by the number of sunspots. (The sunspot data are from Jastrow and Thompson, *Fundamentals and Frontiers of Astronomy,* and the auto death data are from *General Statistics of the U.S. 1985.*)

Year	Sunspots	Auto Accident Deaths (1,000s)
70	165	54.6
71	89	53.3
72	55	56.3
73	34	49.6
74	9	47.1
75	30	45.9
76	59	48.5
77	83	50.1
78	109	52.4
79	127	52.5
80	153	53.2
81	112	51.4
82	80	46
83	45	44.6

12. The following data set presents the heights of 12 male law school classmates whose law school examination scores were roughly equal. It also gives their first year

salaries. Each of them went into corporate law. The height is in inches and the salary in units of $1,000.

Height	Salary
64	91
65	94
66	88
67	103
69	77
70	96
72	105
72	88
74	122
74	102
75	90
76	114

(a) Do the above data establish the hypothesis that a lawyer's salary is related to his height? Use the 5 percent level of significance.

(b) What was the null hypothesis in part (a)?

13. Suppose in the simple linear regression model

$$Y = \alpha + \beta x + e$$

that $0 < \beta < 1$.

(a) Show that if $x < \alpha/(1 - \beta)$, then

$$x < E[Y] < \frac{\alpha}{1 - \beta}$$

(b) Show that if $x > \alpha/(1 - \beta)$, then

$$x > E[Y] > \frac{\alpha}{1 - \beta}$$

and conclude that $E[Y]$ is always between x and $\alpha/(1 - \beta)$.

14. A study has shown that a good model for the relationship between X and Y, the first and second year batting averages of a randomly chosen major league baseball player, is given by the equation

$$Y = .159 + .4X + e$$

where e is a normal random variable with mean 0. That is, the model is a simple linear regression with a regression toward the mean.

(a) If a player's batting average is .200 in his first year, what would you predict for the second year?
(b) If a player's batting average is .265 in his first year, what would you predict for the second year?
(c) If a player's batting average is .310 in his first year, what would you predict for the second year?

15. Experienced flight instructors have claimed that praise for an exceptionally fine landing is typically followed by a poorer landing on the next attempt, whereas criticism of a faulty landing is typically followed by an improved landing. Should we thus conclude that verbal praise tends to lower performance levels, whereas verbal criticism tends to raise them? Or is some other explanation possible?

16. Verify Equation 9.4.3.

17. The following data represent the relationship between the number of alignment errors and the number of missing rivets for 10 different aircraft.

Number of Missing Rivets $= x$	Number of Alignment Errors $= y$
13	7
15	7
10	5
22	12
30	15
7	2
25	13
16	9
20	11
15	8

(a) Plot a scatter diagram.
(b) Estimate the regression coefficients.
(c) Test the hypothesis that $\alpha = 1$.
(d) Estimate the expected number of alignment errors of a plane having 24 missing rivets.
(e) Compute a 90 percent confidence interval estimate for the quantity in (d).

18. The following are the average scores on the mathematics part of the Scholastic Aptitude Test (SAT) for some of the years from 1994 to 2009.

Year	SAT Score
1994	504
1996	508
1998	512
2000	514
2002	516
2004	518
2005	520
2007	515
2009	515

Assuming a simple linear regression model, predict the average scores in the years 1997, 2006, 2008, and 2010.

19. **(a)** Draw a scatter diagram of cigarette consumption versus death rate from bladder cancer.
(b) Does the diagram indicate the possibility of a linear relationship?
(c) Find the best linear fit.
(d) If next year's average cigarette consumption is 2,500, what is your prediction of the death rate from bladder cancer?

20. **(a)** Draw a scatter diagram relating cigarette use and death rates from lung cancer.
(b) Estimate the regression parameters α and β.
(c) Test at the .05 level of significance the hypothesis that cigarette consumption does not affect the death rate from lung cancer.
(d) What is the p-value of the test in part (c)?

21. **(a)** Draw a scatter diagram of cigarette use versus death rate from kidney cancer.
(b) Estimate the regression line.
(c) What is the p-value in the test that the slope of the regression line is 0?
(d) Determine a 90 percent confidence interval for the mean death rate from kidney cancer in a state whose citizens smoke an average of 3,400 cigarettes per year.

22. **(a)** Draw a scatter diagram of cigarettes smoked versus death rate from leukemia.
(b) Estimate the regression coefficients.
(c) Test the hypothesis that there is no regression of the death rate from leukemia on the number of cigarettes used. That is, test that $\beta = 0$.
(d) Determine a 90 percent prediction interval for the leukemia death rate in a state whose citizens smoke an average of 2,500 cigarettes.

23. **(a)** Estimate the variances in Problems 19 through 22.
(b) Determine a 95 percent confidence interval for the variance in the data relating to lung cancer.

(c) Break up the lung cancer data into two parts — the first corresponding to states whose average cigarette consumption is less than 2,300, and the second greater. Assume a linear regression model for both sets of data. How would you test the hypothesis that the variance of a response is the same for both sets?

(d) Do the test in part (c) at the .05 level of significance.

24. Plot the standardized residuals from the data of Problem 1. What does the plot indicate about the assumptions of the linear regression model?

25. It is difficult and time consuming to measure directly the amount of protein in a liver sample. As a result, medical laboratories often make use of the fact that the amount of protein is related to the amount of light that would be absorbed by the sample. As a result, a spectrometer that emits light is shined on a solution that contains the liver sample and the amount of light absorbed is then used to estimate the amount of protein.

The above procedure was tried on five samples having known amounts of protein, with the following data resulting.

Light Absorbed	Amount of Protein (mg)
.44	2
.82	16
1.20	30
1.61	46
1.83	55

(a) Determine the coefficient of determination.

(b) Does this appear to be a reasonable way of estimating the amount of protein in a liver sample?

(c) What is the estimate of the amount of protein when the light absorbed is 1.5?

(d) Determine a prediction interval, in which we can have 90 percent confidence, for the quantity in part (c).

26. The determination of the shear strength of spot welds is relatively difficult, whereas measuring the weld diameter of spot welds is relatively simple. As a result, it would be advantageous if shear strength could be predicted from a measurement of weld diameter. The data are as follows:

(a) Draw a scatter diagram.

(b) Find the least squares estimates of the regression coefficients.

(c) Test the hypothesis that the slope of the regression line is equal to 1 at the .05 level of significance.

Shear Strength (psi)	Weld Diameter (.0001 in.)
370	400
780	800
1,210	1,250
1,560	1,600
1,980	2,000
2,450	2,500
3,070	3,100
3,550	3,600
3,940	4,000
3,950	4,000

(d) Estimate the expected value of shear strength when the weld diameter is .2500.

(e) Find a prediction interval such that, with 95 percent confidence, the value of shear strength corresponding to a weld diameter of .2250 inch will be contained in it.

(f) Plot the standardized residuals.

(g) Does the plot in part (f) support the assumptions of the model?

27. The following are the ages and weights of a random sample of 10 high school male students

Age	Weight
14	129
16	173
18	188
15	121
17	190
16	166
14	133
16	155
15	152
14	115

Assuming a simple linear regression model, give an interval that, with 95 percent confidence, will contain the average weight of all 17 year old male high school students.

28. Glass plays a key role in criminal investigations, because criminal activity often results in the breakage of windows and other glass objects. Since glass fragments often lodge in the clothing of the criminal, it is of great importance to be able

to identify such fragments as originating at the scene of the crime. Two physical properties of glass that are useful for identification purposes are its refractive index, which is relatively easy to measure, and its density, which is much more difficult to measure. The exact measurement of density is, however, greatly facilitated if one has a good estimate of this value before setting up the laboratory experiment needed to determine it exactly. Thus, it would be quite useful if one could use the refractive index of a glass fragment to estimate its density.

The following data relate the refractive index to the density for 18 pieces of glass.

Refractive Index	Density	Refractive Index	Density
1.5139	2.4801	1.5161	2.4843
1.5153	2.4819	1.5165	2.4858
1.5155	2.4791	1.5178	2.4950
1.5155	2.4796	1.5181	2.4922
1.5156	2.4773	1.5191	2.5035
1.5157	2.4811	1.5227	2.5086
1.5158	2.4765	1.5227	2.5117
1.5159	2.4781	1.5232	2.5146
1.5160	2.4909	1.5253	2.5187

(a) Predict the density of a piece of glass with a refractive index 1.52.
(b) Determine an interval that, with 95 percent confidence, will contain the density of the glass in part (a).

29. The regression model

$$Y = \beta x + e, \qquad e \sim N(0, \sigma^2)$$

is called regression through the origin since it presupposes that the expected response corresponding to the input level $x = 0$ is equal to 0. Suppose that $(x_i, Y_i), i = 1, \ldots, n$ is a data set from this model.

(a) Determine the least squares estimator B of β.
(b) What is the distribution of B?
(c) Define SS_R and give its distribution.
(d) Derive a test of $H_0 : \beta = \beta_0$ versus $H_1 : \beta \neq \beta_0$.
(e) Determine a $100(1 - \alpha)$ percent prediction interval for $Y(x_0)$, the response at input level x_0.

30. The following are the body mass index (BMI) and the systolic blood pressure of eight randomly chosen men who do not take any blood pressure medication.

BMI	Systolic Blood Pressure
20.3	116
22.0	110
26.4	131
28.2	136
31.0	144
32.6	138
17.6	122
19.4	115

Give an interval that, with 95 percent confidence, will include the systolic blood pressure of a man whose BMI is 26.0.

31. The weight and systolic blood pressure of randomly selected males in age group 25 to 30 are shown in the following table.

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

(a) Estimate the regression coefficients.
(b) Do the data support the claim that systolic blood pressure does not depend on an individual's weight?
(c) If a large number of males weighing 182 pounds have their blood pressures taken, determine an interval that, with 95 percent confidence, will contain their average blood pressure.
(d) Analyze the standardized residuals.
(e) Determine the sample correlation coefficient.

32. It has been determined that the relation between stress (S) and the number of cycles to failure (N) for a particular type of alloy is given by

$$S = \frac{A}{N^m}$$

where A and m are unknown constants. An experiment is run yielding the following data.

Stress (thousand psi)	N (million cycles to failure)
55.0	.223
50.5	.925
43.5	6.75
42.5	18.1
42.0	29.1
41.0	50.5
35.7	126
34.5	215
33.0	445
32.0	420

Estimate A and m.

33. In 1957 the Dutch industrial engineer J. R. DeJong proposed the following model for the time it takes to perform a simple manual task as a function of the number of times the task has been practiced:

$$T \approx ts^{-n}$$

where T is the time, n is the number of times the task has been practiced, and t and s are parameters depending on the task and individual. Estimate t and s for the following data set.

T	22.4	21.3	19.7	15.6	15.2	13.9	13.7
n	0	1	2	3	4	5	6

34. The chlorine residual in a swimming pool at various times after being cleaned is as given:

Time (hr)	Chlorine Residual (pt/million)
2	1.8
4	1.5
6	1.45
8	1.42
10	1.38
12	1.36

Fit a curve of the form

$$Y \approx ae^{-bx}$$

What would you predict for the chlorine residual 15 hours after a cleaning?

35. The proportion of a given heat rise that has dissipated a time t after the source is cut off is of the form

$$P = 1 - e^{-\alpha t}$$

for some unknown constant α. Given the data

P	.07	.21	.32	.38	.40	.45	.51
t	.1	.2	.3	.4	.5	.6	.7

estimate the value of α. Estimate the value of t at which half of the heat rise is dissipated.

36. The following data represent the bacterial count of five individuals at different times after being inoculated by a vaccine consisting of the bacteria.

Days Since Inoculation	Bacterial Count
3	121,000
6	134,000
7	147,000
8	210,000
9	330,000

(a) Fit a curve.
(b) Estimate the bacteria count of a new patient after 8 days.

37. The following data yield the amount of hydrogen present (in parts per million) in core drillings of fixed size at the following distances (in feet) from the base of a vacuum-cast ingot.

Distance	1	2	3	4	5	6	7	8	9	10
Amount	1.28	1.50	1.12	.94	.82	.75	.60	.72	.95	1.20

(a) Draw a scatter diagram.
(b) Fit a curve of the form

$$Y = \alpha + \beta x + \gamma x^2 + e$$

to the data.

38. A new drug was tested on mice to determine its effectiveness in reducing cancerous tumors. Tests were run on 10 mice, each having a tumor of size 4 grams, by varying the amount of the drug used and then determining the resulting reduction

in the weight of the tumor. The data were as follows:

Coded Amount of Drug	Tumor Weight Reduction
1	.50
2	.90
3	1.20
4	1.35
5	1.50
6	1.60
7	1.53
8	1.38
9	1.21
10	.65

Estimate the maximum expected tumor reduction and the amount of the drug that attains it by fitting a quadratic regression equation of the form

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

39. The following data represent the relation between the number of cans damaged in a boxcar shipment of cans and the speed of the boxcar at impact.

Speed	Number of Cans Damaged
3	54
3	62
3	65
5	94
5	122
5	84
6	142
7	139
7	184
8	254

(a) Analyze as a simple linear regression model.
(b) Plot the standardized residuals.
(c) Do the results of part (b) indicate any flaw in the model?
(d) If the answer to part (c) is yes, suggest a better model and estimate all resulting parameters.

40. Redo Problem 5 under the assumption that the variance of the gain in reading speed is proportional to the number of weeks in the program.

41. The following data relate the proportions of coal miners who exhibit symptoms of pneumoconiosis to the number of years of working in coal mines.

Years Working	Proportion Having Penumoconiosis
5	0
10	.0090
15	.0185
20	.0672
25	.1542
30	.1720
35	.1840
40	.2105
45	.3570
50	.4545

Estimate the probability that a coal miner who has worked for 42 years will have pneumoconiosis.

42. The following data set refers to Example 9.8c.

Number of Cars (Daily)	Number of Accidents (Monthly)
2,000	15
2,300	27
2,500	20
2,600	21
2,800	31
3,000	16
3,100	22
3,400	23
3,700	40
3,800	39
4,000	27
4,600	43
4,800	53

(a) Estimate the number of accidents in a month when the number of cars using the highway is 3,500.

(b) Use the model

$$\sqrt{Y} = \alpha + \beta x + e$$

and redo part (a).

43. The peak discharge of a river is an important parameter for many engineering design problems. Estimates of this parameter can be obtained by relating it to the watershed area (x_1) and watershed slope (x_2). Estimate the relationship based on the following data.

x_1 (m^2)	x_2 (ft/ft)	Peak Discharge (ft^3/sec)
36	.005	50
37	.040	40
45	.004	45
87	.002	110
450	.004	490
550	.001	400
1,200	.002	650
4,000	.0005	1,550

44. The sediment load in a stream is related to the size of the contributing drainage area (x_1) and the average stream discharge (x_2). Estimate this relationship using the following data.

Area ($\times 10^3$ mi^2)	Discharge (ft^3/sec)	Sediment Yield (Millions of tons/yr)
8	65	1.8
19	625	6.4
31	1,450	3.3
16	2,400	1.4
41	6,700	10.8
24	8,500	15.0
3	1,550	1.7
3	3,500	.8
3	4,300	.4
7	12,100	1.6

45. Fit a multiple linear regression equation to the following data set.

x_1	x_2	x_3	x_4	y
1	11	16	4	275
2	10	9	3	183
3	9	4	2	140
4	8	1	1	82
5	7	2	1	97
6	6	1	−1	122
7	5	4	−2	146
8	4	9	−3	246
9	3	16	−4	359
10	2	25	−5	482

46. The following data refer to Stanford heart transplants. It relates the survival time of patients that have received heart transplants to their age when the transplant occurred and to a so-called mismatch score that is supposed to be an indicator of how well the transplanted heart should fit the recipient.

Survival Time (in days)	Mismatch Score	Age
624	1.32	51.0
46	.61	42.5
64	1.89	54.6
1,350	.87	54.1
280	1.12	49.5
10	2.76	55.3
1,024	1.13	43.4
39	1.38	42.8
730	.96	58.4
136	1.62	52.0
836	1.58	45.0
60	.69	64.5

(a) Letting the dependent variable be the logarithm of the survival time, fit a regression on the independent variable's mismatch score and age.
(b) Estimate the variance of the error term.

47. **(a)** Fit a multiple linear regression equation to the following data set.
(b) Test the hypothesis that $\beta_0 = 0$.
(c) Test the hypothesis that $\beta_3 = 0$.
(d) Test the hypothesis that the mean response at the input levels $x_1 = x_2 = x_3 = 1$ is 8.5.

x_1	x_2	x_3	y
7.1	.68	4	41.53
9.9	.64	1	63.75
3.6	.58	1	16.38
9.3	.21	3	45.54
2.3	.89	5	15.52
4.6	.00	8	28.55
.2	.37	5	5.65
5.4	.11	3	25.02
8.2	.87	4	52.49
7.1	.00	6	38.05
4.7	.76	0	30.76
5.4	.87	8	39.69
1.7	.52	1	17.59
1.9	.31	3	13.22
9.2	.19	5	50.98

48. The tensile strength of a certain synthetic fiber is thought to be related to x_1, the percentage of cotton in the fiber, and x_2, the drying time of the fiber. A test of 10 pieces of fiber produced under different conditions yielded the following results.

$Y =$ Tensile Strength	$x_1 =$ Percentage of Cotton	$x_2 =$ Drying Time
213	13	2.1
220	15	2.3
216	14	2.2
225	18	2.5
235	19	3.2
218	20	2.4
239	22	3.4
243	17	4.1
233	16	4.0
240	18	4.3

(a) Fit a multiple regression equation.
(b) Determine a 90 percent confidence interval for the mean tensile strength of a synthetic fiber having 21 percent cotton whose drying time is 3.6.

49. The time to failure of a machine component is related to the operating voltage (x_1), the motor speed in revolutions per minute (x_2), and the operating temperature (x_3).

A designed experiment is run in the research and development laboratory, and the following data, where y is the time to failure in minutes, are obtained.

y	x_1	x_2	x_3
2,145	110	750	140
2,155	110	850	180
2,220	110	1,000	140
2,225	110	1,100	180
2,260	120	750	140
2,266	120	850	180
2,334	120	1,000	140
2,340	130	1,000	180
2,212	115	840	150
2,180	115	880	150

(a) Fit a multiple regression model to these data.
(b) Estimate the error variance.
(c) Determine a 95 percent confidence interval for the mean time to failure when the operating voltage is 125, the motor speed is 900, and the operating temperature is 160.

50. Explain why, for the same data, a prediction interval for a future response always contains the corresponding confidence interval for the mean response.

51. Consider the following data set:

x_1	x_2	y
5.1	2	55.42
5.4	8	100.21
5.9	−2	27.07
6.6	12	169.95
7.5	−6	−17.93
8.6	16	197.77
9.9	−10	−25.66
11.4	20	264.18
13.1	−14	−53.88
15	24	317.84
17.1	−18	−72.53
19.4	28	385.53

(a) Fit a linear relationship between y and x_1, x_2.
(b) Determine the variance of the error term.

(c) Determine an interval that, with 95 percent confidence, will contain the response when the inputs are $x_1 = 10.2$ and $x_2 = 17$.

52. The cost of producing power per kilowatt hour is a function of the load factor and the cost of coal in cents per million Btu. The following data were obtained from 12 mills.

Load Factor (in percent)	Cost of Coal	Power Cost
84	14	4.1
81	16	4.4
73	22	5.6
74	24	5.1
67	20	5.0
87	29	5.3
77	26	5.4
76	15	4.8
69	29	6.1
82	24	5.5
90	25	4.7
88	13	3.9

(a) Estimate the relationship.
(b) Test the hypothesis that the coefficient of the load factor is equal to 0.
(c) Determine a 95 percent prediction interval for the power cost when the load factor is 85 and the coal cost is 20.

53. The following data relate the systolic blood pressure to the age (x_1) and weight (x_2) of a set of individuals of similar body type and lifestyle.

Age	Weight	Blood Pressure
25	162	112
25	184	144
42	166	138
55	150	145
30	192	152
40	155	110
66	184	118
60	202	160
38	174	108

(a) Test the hypothesis that, when an individual's weight is known, age gives no additional information in predicting blood pressure.

(b) Determine an interval that, with 95 percent confidence, will contain the average blood pressure of all individuals of the preceding type who are 45 years old and weigh 180 pounds.

(c) Determine an interval that, with 95 percent confidence, will contain the blood pressure of a given individual of the preceding type who is 45 years old and weighs 180 pounds.

54. A recently completed study attempted to relate job satisfaction to income (in 1,000s) and seniority for a random sample of 9 municipal workers. The job satisfaction value given for each worker is his or her own assessment of such, with a score of 1 being the lowest and 10 being the highest. The following data resulted.

Yearly Income	Years on the Job	Job Satisfaction
52	8	5.6
47	4	6.3
59	12	6.8
53	9	6.7
61	16	7.0
64	14	7.7
58	10	7.0
67	15	8.0
71	22	7.8

(a) Estimate the regression parameters.

(b) What qualitative conclusions can you draw about how job satisfaction changes when income remains fixed and the number of years of service increases?

(c) Predict the job satisfaction of an employee who has spent 5 years on the job and earns a yearly salary of $56,000.

55. Suppose in Problem 54 that job satisfaction was related solely to years on the job, with the following data resulting.

Years on the Job	Job Satisfaction
8	5.6
4	6.3
12	6.8
9	6.7
16	7.0
14	7.7
10	7.0
15	8.0
22	7.8

(a) Estimate the regression parameters α and β.
(b) What is the qualitative relationship between years of service and job satisfaction? That is, what appears to happen to job satisfaction as service increases?
(c) Compare your answer to part (b) with the answer you obtained in part (b) of Problem 54.
(d) What conclusion, if any, can you draw from your answer in part (c)?

56. For the logistics regression model, find the value x such that $p(x) = .5$.

57. A study of 64 prematurely born infants was interested in the relation between the gestational age (in weeks) of the infant at birth and whether the infant was breast-feeding at the time of release from the birthing hospital. The following data resulted:

Gestational Age	Frequency	Number Breast-Feeding
28	6	2
29	5	2
30	9	7
31	9	7
32	20	16
33	15	14

In the preceding, the frequency column refers to the number of babies born after the specified gestational number of weeks.

(a) Explain how the relationship between gestational age and whether the infant was breast-feeding can be analyzed via a logistics regression model.
(b) Use appropriate software to estimate the parameters for this model.
(c) Estimate the probability that a newborn with a gestational age of 29 weeks will be breast-feeding.

58. Twelve first-time heart attack victims were given a test that measures internal anger. The following data relates their scores and whether they had a second heart attack within 5 years.

Anger Score	Second Heart Attack
80	yes
77	yes
70	no
68	yes
64	no

(*continued*)

Anger Score	Second Heart Attack
60	yes
50	yes
46	no
40	yes
35	no
30	no
25	yes

(a) Explain how the relationship between a second heart attack and one's anger score can be analyzed via a logistics regression model.
(b) Use appropriate software to estimate the parameters for this model.
(c) Estimate the probability that a heart attack victim with an anger score of 55 will have a second attack within 5 years.