4

Multiple Linear Regression

4.1 Overview

In Chapter 2 we pointed out that the *population regression function of Y on $X_1, \ldots, X_k$ is the best function for predicting a response variable Y, using the predictor variables $X_1, \ldots, X_k$*. In applied problems, the population prediction function is seldom available, and sample data are used to estimate it.

In many situations we either know or assume the *form* of the population regression function so that the regression function is known except for some unknown parameters (regression coefficients). In Chapter 3 we discussed point estimation, confidence interval estimation, and tests for unknown parameters when the population regression function of the response variable Y on the predictor variable X is a straight line function of X; i.e., $\mu_Y(x) = \beta_0 + \beta_1 x$. In particular, the number of predictor variables is one. In this chapter we consider the situation where there are several predictor (or explanatory) variables, say $X_1, \ldots, X_k$, and the population regression function Y on $X_1, \ldots, X_k$ is of the form

$$\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad (4.1.1)$$

When data for the entire population are available, the constants $\beta_0, \beta_1, \ldots, \beta_k$ can be calculated exactly. However, as we pointed out earlier, this is seldom the case in applied problems, and so the constants $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters that have to be estimated using sample data. This chapter is primarily concerned with point and confidence interval estimation of the parameters in the **multiple linear regression function** of (4.1.1). The formulas and procedures for point and confidence interval estimation, discussed in Chapter 3, are special cases (with $k = 1$) of the formulas and procedures that are discussed in this chapter.

The chapter is organized as follows. Section 4.2 includes definitions and notation that are used throughout this chapter and elsewhere in the book. Section 4.3 contains a discussion of the assumptions required for valid statistical inference procedures for multiple linear regression. Procedures for point estimation are given in Section 4.4. Section 4.5 discusses some methods for investigating the validity of the assumptions

required for multiple linear regression. Confidence interval procedures for multiple linear regression are given in Section 4.6. A discussion of statistical tests is provided in Section 4.7. Analysis of variance in the case of multiple linear regression is discussed in Section 4.8. Procedures for comparing two regression functions are given in Sections 4.9 and 4.10. Section 4.9 also discusses multiple correlation, the coefficient of determination, multiple-partial correlation, and partial coefficient of determination. Procedures for evaluating the lack-of-fit of a straight line regression model are provided in Section 4.11, and Section 4.12 contains exercises.

For a clear and accurate explanation of the concepts developed in this chapter, we need to introduce some notation that may initially appear complicated. However, if you take the time to understand the notation, it will help you to understand the concepts as well as the procedures.

4.2 Notation and Definitions

The word *multiple* in multiple linear regression means there is more than one predictor variable. Throughout this chapter, it is assumed that there are k predictor variables that are denoted by $X_1, \ldots, X_k$. Recall from Section 2.3 that the word *linear* means the regression function, denoted by $\mu_Y(x_1, \ldots, x_k)$, is *linear in the unknown parameters* $\beta_0, \beta_1, \ldots, \beta_k$ (see (4.1.1)). The word *regression* means that the study is concerned with the prediction of the response variable Y using the relationship between Y and the k predictor variables $X_1, X_2, \ldots, X_k$.

We begin with two examples illustrating how regression functions can be useful in practical problems.

EXAMPLE 4.2.1

Consider the population of high school graduates who were admitted to a particular university during the past ten years and who completed at least the first year of coursework after being admitted. Suppose the director of admissions at this university is interested in investigating how well Y, the first year grade point average (GPA) of a student, can be predicted by using the following quantities:

X_1 = the score on the mathematics part of the Scholastic Aptitude Test (SATmath)

X_2 = the score on the verbal part of the Scholastic Aptitude Test (SATverbal)

X_3 = the grade point average of all high school mathematics courses (HSmath)

X_4 = the grade point average of all high school English courses (HSenglish)

If a relationship does exist between Y and X_1, X_2, X_3, X_4, then it may be possible to predict the performance of new applicants during their first year in college based on their SAT scores and high school grades; this would provide the director of admissions with a very valuable tool for making decisions about whether or not an applicant should be recommended for financial assistance.

Suppose, in this example, that the population regression function of Y on X_1, X_2, X_3, X_4 is of the form

$$\mu_Y(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

This is a special case of the multiple linear regression function in (4.1.1) with $k = 4$. Methods discussed in this chapter can therefore be used to investigate the relationship between Y and X_1, X_2, X_3, X_4. ■

EXAMPLE 4.2.2

Consider Example 2.2.9. Suppose that the regression function of Y, the strength of plastic containers, on X_1 and X_2, the temperature and pressure, respectively, during the production process, is of the form

$$\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

This is a special case of the multiple linear regression function in (4.1.1) with $k = 2$. ■

Other situations where multiple linear regression may be applicable are described in Examples 2.2.3, 2.2.5, and 2.2.6.

Basic Observable Variables and Derived Variables

It is useful to make a distinction between *basic observable variables* and *derived variables*. If Z represents the height of an individual, then Z is a basic observable variable and $\sqrt{Z}$ or $\log Z$ are derived variables. Thus derived variables are known functions of observable variables. The quantities x_i in the multiple linear regression function in (4.1.1) may be values of basic observable variables or derived variables. For instance, if the variables Z_1 and Z_2 are basic observable variables and the regression function is

$$\mu_Y(z_1, z_2) = \beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \sqrt{z_2}$$

then we can define the variables X_1, X_2, X_3 by

$$X_1 = Z_1 \qquad X_2 = Z_1^2 \qquad X_3 = \sqrt{Z_2}$$

and their values x_1, x_2, x_3 by

$$x_1 = z_1 \qquad x_2 = z_1^2 \qquad x_3 = \sqrt{z_2}$$

so that the regression function can be written as

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

This demonstrates the versatility of the regression function in (4.1.1). More examples are given in (4.2.1).

$$\mu_Y(z_1, z_2) = \beta_0 + \beta_1 z_1 + \beta_2 z_1 z_2 + \beta_3 z_1^2 \qquad \text{(4.2.1a)}$$

$$\mu_Y(z_1, z_2) = \beta_0 + \beta_1 e^{z_1} + \beta_2(\cos z_1 + \cos z_2) + \beta_3 \sqrt{z_1} \qquad \text{(4.2.1b)}$$

$$\mu_Y(z_1, z_2, z_3) = \beta_0 + \beta_1 z_1 + \beta_2 z_1^2 + \beta_3 \frac{z_1}{z_1 + z_2 + z_3} + \beta_4 \log|z_3| \quad \textbf{(4.2.1c)}$$

Each of these regression functions may be written in the form given in (4.1.1) by suitably defining variables $X_1, X_2, \ldots$, etc. For instance, in (4.2.1a) we define $X_1 = Z_1, X_2 = Z_1 Z_2$, and $X_3 = Z_1^2$ so that the regression function can be rewritten as

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Here each X_i is a *known* function of the Z_i and involves no unknown parameters. In (4.2.1b) we use $X_1 = e^{Z_1}, X_2 = (\cos Z_1 + \cos Z_2)$, and $X_3 = \sqrt{Z_1}$, which allow us to rewrite the regression function as

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Similarly, in (4.2.1c) we use $X_1 = Z_1, X_2 = Z_1^2, X_3 = Z_1/(Z_1 + Z_2 + Z_3)$, and $X_4 = \log|Z_3|$ and rewrite the regression function as

$$\mu_Y(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Polynomial Regression

A population regression model that is quite useful in many applied problems uses the polynomial function given by

$$\mu_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_k x^k$$

which is a kth degree polynomial in the predictor X. We can write this as

$$\mu_Y(x_1, x_2, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where $X_1 = X, X_2 = X^2, \ldots, X_k = X^k$. This is a multiple regression model using the derived variables $X_1, \ldots, X_k$, where X is the basic observable variable.

Notation

The starting point of any investigation involving multiple linear regression is the precise definition of the $(k+1)$ variable population $\{(Y, X_1, \ldots, X_k)\}$. To identify the elements of this population, we use a double subscript notation for the values of the predictor variables. In this notation, $X_{I,J}$ refers to the value of the Ith population item corresponding to the Jth predictor variable; i.e., the first subscript I is the *reference number* or *label* of the population item under consideration, and the second subscript J refers to the Jth predictor variable. For example, $X_{4,8}$ is the value of the 8th predictor variable of the 4th population item; $X_{24,6}$ is the value of the 6th predictor variable of the 24th population item, etc. Of course, Y_I denotes the value of the response variable for the Ith population item.

Table 4.2.1 gives a schematic display of the five-variable population $\{(Y, X_1, X_2, X_3, X_4)\}$ discussed in Example 4.2.1. We suppose that this population has N items. The items in this population are *all high school graduates who were admitted to the university during the past ten years and completed at least the first year of*

coursework after being admitted. Of course, all of the population values Y_I and $X_{I,J}$ may not be known in a real problem, but conceptually they exist.

In Table 4.2.1, $X_{6,3}$ represents the high school math GPA of the 6th individual, $X_{32,2}$ represents the SAT verbal score of the 32nd individual, etc. The theory for deriving point estimates, confidence intervals, and tests assumes that N, the population size, is infinite, but in most practical applications N is indeed finite. However, N is generally very large (but unknown) so the theory is approximately valid for all practical purposes.

TABLE 4.2.1
Schematic Display of the Population in Example 4.2.1

Item (Individual Student) I	GPA at the End of One Year Y	SAT Math Score X_1	SAT Verbal Score X_2	High School Math GPA X_3	High School English GPA X_4
1	Y_1	$X_{1,1}$	$X_{1,2}$	$X_{1,3}$	$X_{1,4}$
2	Y_2	$X_{2,1}$	$X_{2,2}$	$X_{2,3}$	$X_{2,4}$
⋮	⋮	⋮	⋮	⋮	⋮
I	Y_I	$X_{I,1}$	$X_{I,2}$	$X_{I,3}$	$X_{I,4}$
⋮	⋮	⋮	⋮	⋮	⋮
N	Y_N	$X_{N,1}$	$X_{N,2}$	$X_{N,3}$	$X_{N,4}$

Parameters of a $(k+1)$-Variable Population

A $(k+1)$-variable population $\{(Y, X_1, X_2, \ldots, X_k)\}$ has associated with it several parameters that are useful in describing and summarizing it. The parameters

$$\mu_Y, \mu_{X_1}, \ldots, \mu_{X_k}$$

are the means of each one-variable population, and

$$\sigma_Y, \sigma_{X_1}, \ldots, \sigma_{X_k}$$

are the standard deviations of each one-variable population. They are useful for summarizing the one-variable populations $\{Y\}, \{X_1\}, \ldots, \{X_k\}$. The parameters

$$\rho_{Y,X_1}, \ldots, \rho_{Y,X_k}, \rho_{X_1,X_2}, \ldots, \rho_{X_{k-1},X_k}$$

are the coefficients of correlation of all pairs of variables, which may be useful for investigating the relationships between pairs of variables.

The regression function $\mu_Y(x_1, \ldots, x_k)$ can be used for predicting Y using the k predictor variables $X_1, \ldots, X_k$. It is in fact the best prediction function for predicting Y using $X_1, \ldots, X_k$. When the regression function $\mu_Y(x_1, \ldots, x_k)$ is of the form given in (4.1.1), the parameters $\beta_0, \beta_1, \ldots, \beta_k$ completely determine this regression function. Therefore we are interested in knowing the values of these (and possibly

other) parameters to help make decisions about the population. Of course the population values are unknown, so the parameters that summarize the population are unknown. As in simple linear regression, a sample is selected from the population, and inferences about population parameters are made by using the sample data.

The Population Regression Model

For the Ith item in the population, the true value of the response variable is Y_I, and the predicted value using the regression function in (4.1.1), which is the best prediction function, is

$$\mu_Y(X_{I,1}, X_{I,2}, \ldots, X_{I,k}) = \beta_0 + \beta_1 X_{I,1} + \beta_2 X_{I,2} + \cdots + \beta_k X_{I,k}$$

The error of prediction for the Ith element of the population is denoted by E_I; it is the difference between the true value Y_I and the predicted value $\mu_Y(X_{I,1}, X_{I,2}, \ldots, X_{I,k})$ and is given by

$$E_I = Y_I - \mu_Y(X_{I,1}, \ldots, X_{I,k})$$

or

$$E_I = Y_I - (\beta_0 + \beta_1 X_{I,1} + \cdots + \beta_k X_{I,k})$$

This equation is generally written as

$$Y_I = \beta_0 + \beta_1 X_{I,1} + \cdots + \beta_k X_{I,k} + E_I \qquad (4.2.2)$$

for $I = 1, \ldots, N$ and is called the **population regression model**. We use the symbol $Y(x_1, \ldots, x_k)$ to denote the Y value of a randomly chosen item with $X_1 = x_1, \ldots, X_k = x_k$.

Sample

To make statistical inferences about the unknown parameters in the population regression function in (4.1.1), we must use sample data to compute point estimates and interval estimates of these unknown parameters, so we select a sample of size n from the population. A schematic display of the sample data corresponding to the population in Table 4.2.1 is given in Table 4.2.2. Note that we use lower-case letters y, $x_{i,j}$, n, etc., to emphasize that these quantities refer to the sample and not to the population. For example, $(y_i, x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})$ are the five measurements on the ith sample item selected from the population. The sample size n and the sample observations are, of course, known quantities.

You are encouraged to carry out the following task to develop familiarity with the terminology as well as some of the concepts underlying multiple regression.

TABLE 4.2.2
Schematic Display of a Sample from the Population in Table 4.2.1

Item (Individual Student) i	GPA at the End of One Year Y	SAT Math Score X_1	SAT Verbal Score X_2	High School Math GPA X_3	High School English GPA X_4
1	y_1	$x_{1,1}$	$x_{1,2}$	$x_{1,3}$	$x_{1,4}$
2	y_2	$x_{2,1}$	$x_{2,2}$	$x_{2,3}$	$x_{2,4}$
⋮	⋮	⋮	⋮	⋮	⋮
i	y_i	$x_{i,1}$	$x_{i,2}$	$x_{i,3}$	$x_{i,4}$
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	$x_{n,1}$	$x_{n,2}$	$x_{n,3}$	$x_{n,4}$

Task 4.2.1

In this task we illustrate the concept of subpopulations when the number of predictor variables is greater than one.

Table D-4 in Appendix D contains a three-variable population $\{(Y, X_1, X_2)\}$ where Y is the strength of plastic containers, and X_1 and X_2 are, respectively, the temperature (in degrees Celsius) and the pressure (in pounds per square inch) used during the production process. These data are also in file **plastic.dat** on the data disk. There are 1,650 items in this (artificial) population. Column 1 contains the item numbers for the items in the population. Column 2 contains the strengths of the plastic containers, and columns 3 and 4 contain the temperatures and pressures, respectively, used during the production of each item. These population data may be thought of as having been obtained from the records, for the past two years, of the research and development division of a company that manufactures plastic containers.

The means of Y, X_1, and X_2 are

$$\mu_Y = 30 \qquad \mu_{X_1} = 250 \qquad \mu_{X_2} = 15$$

and the standard deviations are

$$\sigma_Y = 7.39 \qquad \sigma_{X_1} = 31.62 \qquad \sigma_{X_2} = 3.42$$

An examination of the population reveals that there are 66 distinct subpopulations of Y values determined by the distinct pairs of values of (X_1, X_2). Note that this is a very special population in which the standard deviation of each subpopulation is the same. We say more about this in Section 4.3. The means and the standard deviations of these 66 subpopulations of Y values are listed in Table 4.2.3. They are also stored in the file **table423.dat** on the data disk.

TABLE 4.2.3

The Sixty-Six Subpopulation Means and Standard Deviations for the Population Data in Table D-4

Subpopulation	Temperature	Pressure	*Y* Mean	*Y* Standard Deviation
1	200	10	25	1.7076
2	200	12	23	1.7076
3	200	14	21	1.7076
4	200	16	19	1.7076
5	200	18	17	1.7076
6	200	20	15	1.7076
7	210	10	27	1.7076
8	210	12	25	1.7076
9	210	14	23	1.7076
10	210	16	21	1.7076
11	210	18	19	1.7076
12	210	20	17	1.7076
13	220	10	29	1.7076
14	220	12	27	1.7076
15	220	14	25	1.7076
16	220	16	23	1.7076
17	220	18	21	1.7076
18	220	20	19	1.7076
19	230	10	31	1.7076
20	230	12	29	1.7076
21	230	14	27	1.7076
22	230	16	25	1.7076
23	230	18	23	1.7076
24	230	20	21	1.7076
25	240	10	33	1.7076
26	240	12	31	1.7076
27	240	14	29	1.7076
28	240	16	27	1.7076
29	240	18	25	1.7076
30	240	20	23	1.7076
31	250	10	35	1.7076
32	250	12	33	1.7076
33	250	14	31	1.7076
34	250	16	29	1.7076
35	250	18	27	1.7076
36	250	20	25	1.7076
37	260	10	37	1.7076
38	260	12	35	1.7076
39	260	14	33	1.7076
40	260	16	31	1.7076

(Continued)

TABLE 4.2.3
(Continued)

Subpopulation	Temperature	Pressure	Y Mean	Y Standard Deviation
41	260	18	29	1.7076
42	260	20	27	1.7076
43	270	10	39	1.7076
44	270	12	37	1.7076
45	270	14	35	1.7076
46	270	16	33	1.7076
47	270	18	31	1.7076
48	270	20	29	1.7076
49	280	10	41	1.7076
50	280	12	39	1.7076
51	280	14	37	1.7076
52	280	16	35	1.7076
53	280	18	33	1.7076
54	280	20	31	1.7076
55	290	10	43	1.7076
56	290	12	41	1.7076
57	290	14	39	1.7076
58	290	16	37	1.7076
59	290	18	35	1.7076
60	290	20	33	1.7076
61	300	10	45	1.7076
62	300	12	43	1.7076
63	300	14	41	1.7076
64	300	16	39	1.7076
65	300	18	37	1.7076
66	300	20	35	1.7076

1 Verify that the population regression function of Y on X_1 and X_2 is of the form (4.1.1) with $k = 2$, i.e., linear in X_1 and X_2 (for those values of the pair (X_1, X_2) occurring in the population). In particular, what are the values of β_0, β_1, and β_2?

Suppose that the regression function of Y on X_1 and X_2 is of the form $\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Then, in particular, the average Y value of each subpopulation must be equal to the value obtained when the corresponding values of temperature X_1 and pressure X_2 are substituted into the regression function. For instance, each one of the 66 equations given in Table 4.2.4 (obtained by examining the subpopulation values in Table 4.2.3) must be satisfied. Consider the first two equations in Table 4.2.4, viz.,

$$\beta_0 + \beta_1(200) + \beta_2(10) = 25$$
$$\beta_0 + \beta_1(200) + \beta_2(12) = 23$$

TABLE 4.2.4

Subpopulation Number	X_1	X_2	Value of the Regression Function = $\mu_Y(X_{i1}, X_{i2}) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} = Y$ mean
1	200	10	$\mu_Y(200, 10) = \beta_0 + \beta_1(200) + \beta_2(10) = 25$
2	200	12	$\mu_Y(200, 12) = \beta_0 + \beta_1(200) + \beta_2(12) = 23$
⋮	⋮	⋮	⋮
7	210	10	$\mu_Y(210, 10) = \beta_0 + \beta_1(210) + \beta_2(10) = 27$
⋮	⋮	⋮	⋮
66	300	20	$\mu_Y(300, 20) = \beta_0 + \beta_1(300) + \beta_2(20) = 35$

We subtract the second equation from the first and obtain $-2\beta_2 = 2$ *or* $\beta_2 = -1$. *Next, consider the first and the last equation in Table 4.2.4, viz.,*

$$\beta_0 + \beta_1(200) + \beta_2(10) = 25$$
$$\beta_0 + \beta_1(300) + \beta_2(20) = 35$$

Subtract the first equation from the second equation above and obtain $100\beta_1 + 10\beta_2 = 10$. *Substituting* $\beta_2 = -1$, *we get* $\beta_1 = 0.2$. *Finally, substituting the value* $\beta_1 = 0.2$ *and* $\beta_2 = -1.0$ *into the first equation in Table 4.2.4, viz., into*

$$\beta_0 + \beta_1(200) + \beta_2(10) = 25$$

we get $\beta_0 = -5$. *Hence, if the regression function of* Y *on* X_1, X_2 *is indeed of the form* $\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$, *then it must be given by*

$$\mu_Y(x_1, x_2) = -5 + 0.2x_1 - x_2 \qquad \textbf{(4.2.3)}$$

You should verify that all of the subpopulation means for Y *are obtained by substituting the appropriate temperature and pressure values into (4.2.3). Thus we obtain*

$$\beta_0 = -5 \qquad \beta_1 = 0.2 \qquad \beta_2 = -1 \qquad \textbf{(4.2.4)}$$

2 Examine the Y values in the subpopulation with $X_1 = 220$ and $X_2 = 16$.

Table 4.2.5 contains these subpopulation values, and they are also stored in the file ***table425.dat*** *on the data disk.*

3 Compute the mean and the standard deviation of the Y values in the subpopulation in (2) where $X_1 = 220$, and $X_2 = 16$.

There are three ways to obtain the Y mean $\mu_Y(220, 16)$.

i *Compute the mean of the Y values in the subpopulation in (2) by direct calculation; i.e.,* $\mu_Y(220, 16) = \overline{Y} = (1/25)\sum Y_i$.

ii *Look up the Y mean corresponding to subpopulation number 16 in Table 4.2.3 where temperature* $X_1 = 220$ *and pressure* $X_2 = 16$.

iii *Use (4.2.3) and plug in* $X_1 = 220$ *and* $X_2 = 16$.

TABLE 4.2.5

Row	Item Number I	Y	X_1	X_2
1	76	22.8	220	16
2	183	22.2	220	16
3	332	24.2	220	16
4	473	25.8	220	16
5	551	24.5	220	16
6	592	22.2	220	16
7	623	23.3	220	16
8	683	26.5	220	16
9	858	25.0	220	16
10	883	20.4	220	16
11	888	21.7	220	16
12	900	20.7	220	16
13	931	22.3	220	16
14	951	26.7	220	16
15	952	22.6	220	16
16	978	22.6	220	16
17	989	21.5	220	16
18	993	22.7	220	16
19	1165	21.7	220	16
20	1228	24.1	220	16
21	1316	23.9	220	16
22	1463	20.7	220	16
23	1526	21.2	220	16
24	1557	22.2	220	16
25	1572	23.5	220	16

To apply method (i) we compute the mean of the Y values in Table 4.2.5 and get $\mu_Y(220, 16) = \overline{Y} = 23.0$. *To apply method (ii) we look up subpopulation number 16 and observe that* $\mu_Y(220, 16) = 23.0$. *To apply method (iii) we use (4.2.3) and get*

$$\mu_Y(220, 16) = -5.0 + 0.2(220) - 16 = 23.0$$

There are two ways to get the standard deviation of a subpopulation. For instance, the standard deviation of the subpopulation with $X_1 = 220$ *and* $X_2 = 16$ *can be obtained using one of the following two ways.*

iv *By directly computing the standard deviation of the Y values in the subpopulation in (2) using the definition of standard deviation.*

v *By looking up the Y standard deviation for subpopulation 16 in Table 4.2.3.*

By both of these procedures we get $\sigma_Y(220, 16) = 1.7076$.

4 The population regression model is

$$Y_I = \beta_0 + \beta_1 X_{I,1} + \beta_2 X_{I,2} + E_I$$

For the subpopulation with $X_1 = 220$ and $X_2 = 16$ we get

$$E_I = Y_I - \mu_Y(220, 16) = Y_I - \beta_0 - \beta_1(220) - \beta_2(16)$$
$$= Y_I + 5 - 0.2(220) + 1(16) = Y_I - 23$$

The E_I's satisfy the following conditions:

i The mean of the E_I values is zero.

ii The standard deviation of the E_I values in any subpopulation is the same as the standard deviation of the Y_I values in that subpopulation.

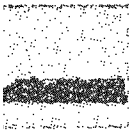
To see that this is the case in this subpopulation, calculate the values of E_I for the items in the subpopulation referred to in (2). Verify that the mean of these E_I values is zero and that their standard deviation is the same as the standard deviation of the Y values in this subpopulation, namely 1.7076.

By subtracting the subpopulation mean Y value of 23 from each of the Y values in the subpopulation in Table 4.2.5, we obtain the E_I values, which are

```
-0.20000  -0.80000   1.20000   2.80000   1.50000  -0.80000   0.30000
 3.50000   2.00000  -2.60000  -1.30000  -2.30000  -0.70000   3.70000
-0.40000  -0.40000  -1.50000  -0.30000  -1.30000   1.10000   0.90000
-2.30000  -1.80000  -0.80000   0.50000
```

You should check these and also check that the mean of these numbers is 0 (to within rounding error) and that the standard deviation is 1.7076, the same as the standard deviation of the Y values in Table 4.2.5, as it should be.

In Task 4.2.1 we examined a population and its subpopulations whose values were all *known*. However, in applied problems the entire population is seldom known and so the population parameters must be estimated using sample data. For these estimates to be valid, certain assumptions must be satisfied, as in the case of straight line regression. These assumptions are stated and discussed in the next section.



Problems 4.2

Problems 4.2.1–4.2.5 refer to the population data given in Table D-4 in Appendix D and also in the file **plastic.dat** on the data disk. Consider the subpopulation of Y values with $X_1 = 240$ and $X_2 = 18$. This subpopulation has 25 values and they are given in Table 4.2.6 along with the corresponding E_I values. These subpopulation

TABLE 4.2.6

Row	Population Item Number	Y	X_1	X_2	E_I
1	19	24.6	240	18	−0.400
2	44	24.2	240	18	−0.800
3	71	23.7	240	18	−1.300
4	174	26.5	240	18	1.500
5	175	24.3	240	18	−0.700
6	212	24.2	240	18	−0.800
7	216	23.7	240	18	−1.300
8	218	24.6	240	18	−0.400
9	235	22.7	240	18	−2.300
10	378	28.7	240	18	3.700
11	407	22.4	240	18	−2.600
12	584	26.1	240	18	1.100
13	588	27.8	240	18	2.800
14	948	26.2	240	18	1.200
15	953	27.0	240	18	2.000
16	962	25.3	240	18	*
17	1118	22.7	240	18	−2.300
18	1135	24.8	240	18	−0.200
19	1209	25.5	240	18	0.500
20	1285	24.2	240	18	−0.800
21	1302	28.5	240	18	3.500
22	1376	25.9	240	18	*
23	1406	24.7	240	18	−0.300
24	1447	23.5	240	18	−1.500
25	1459	23.2	240	18	−1.800

* These values were omitted and you will be asked to supply them.

values are also stored in **table426.dat** on the data disk. For this subpopulation, use the following:

$$\sum Y_I = 625.0 \qquad \sum Y_I^2 = 15697.90$$

4.2.1 **a** Show that the mean of this subpopulation is 25.0, i.e. $\mu_Y(240, 18) = 25.0$.

b The subpopulation of E_I values can be obtained by computing

$$E_I = Y_I - \mu_Y(240, 18) = Y_I - \beta_0 - \beta_1(240) - \beta_2(18)$$

Substituting the values for $\beta_0, \beta_1, \beta_2$ from (4.2.4) we get

$$E_I = Y_I - [-5 + 0.2X_{I1} - X_{I2}] = Y_I - [-5 + 0.2(240) - 18] = Y_I - 25$$

In Table 4.2.6, the 25 values of E_I have been computed except for population items 962 and 1376 (where the asterisk * appears) . Compute the E_I values for items 962 and 1376.

c Verify that the sum of the E_I values in Table 4.2.6 (including items 962 and 1376) is zero.

d Find the standard deviation of the Y_I values in this subpopulation.

e Find the standard deviation of all 25 E_I values in this subpopulation.

f In Problem 4.2.1(e), is this standard deviation what you should get? (Examine Table 4.2.3.)

4.2.2 Predict the strength of a single plastic container that was produced with a temperature of 280° and a pressure of 10 units.

4.2.3 Predict the strength of a single plastic container if it is produced with a temperature of 205° and a pressure of 11 units.

4.2.4 What is the *average* strength of all plastic containers in the study population produced using a temperature of 280° and a pressure of 18 units? Do you know this exactly or do you have to estimate it?

4.2.5 For the subpopulation with $X_1 = 280$ and $X_2 = 10$, find a number L such that 80% of the plastic containers produced would have strength greater than or equal to L.

4.3
Assumptions for Multiple Linear Regression

To obtain valid point and confidence interval estimates for parameters in the multiple linear regression function

$$\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

we must make some assumptions about the $(k+1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$ and about the method used to obtain a sample from this population. One set of assumptions under which the theory for multiple linear regression has been extensively developed is given in Box 4.3.1 and is referred to as assumptions (A) throughout this book. Three of the assumptions concern the population and two concern the sample.

BOX 4.3.1 Assumptions (A) for Multiple Linear Regression

Notation: The $(k+1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$ is the study population.

(Population) Assumption 1. The mean $\mu_Y(x_1, \ldots, x_k)$ of the subpopulation of Y values with $X_1 = x_1, \ldots, X_k = x_k$ is

$$\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

where $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters and $x_1, \cdots, x_k$ belong to the set of allowable values (sometimes called the *domain*) of the predictor variables.

(Population) Assumption 2. The standard deviation of the Y values in the subpopulations with $X_1 = x_1, \ldots, X_k = x_k$ does not depend on the values $x_1, \ldots, x_k$ (i.e., the standard deviations are the same for each subpopulation determined by specified values of the predictor variables $X_1, \ldots, X_k$). This common standard deviation of all the subpopulations is denoted by $\sigma_{Y|X_1,\ldots,X_k}$. When there is no possibility of confusion, we use the simpler notation σ instead of the more complete notation $\sigma_{Y|X_1,\ldots,X_k}$.

(Population) Assumption 3. Each subpopulation of Y values, determined by specified values of $X_1, \ldots, X_k$, is a Gaussian population.

(Sample) Assumption 4. The sample data are obtained by simple random sampling or by sampling with preselected values of $X_1, \ldots, X_k$, discussed in Section 2.3. The number of items in the sample is n.

(Sample) Assumption 5. All sample values $y_i, x_{i,1}, \ldots, x_{i,k}$ for $i = 1, \ldots, n$ are observed without error (but read Section 3.10).

Several quantities of interest are associated with the $(k+1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$. They include

a Parameters and randomly chosen Y values associated with subpopulations determined by $X_1, \ldots, X_k$, viz.,

$$\beta_0, \beta_1, \ldots, \beta_k, Y(x_1, x_2, \ldots, x_k), \mu_Y(x_1, \ldots, x_k), \sigma = \sigma_{Y|X_1,\ldots,X_k}$$

b Parameters of individual populations $\{Y\}, \{X_1\}, \ldots, \{X_k\}$, viz.,

$$\mu_Y, \sigma_Y, \mu_{X_1}, \sigma_{X_1}, \mu_{X_2}, \sigma_{X_2}, \ldots, \mu_{X_k}, \sigma_{X_k}$$

c Correlation coefficients of all two-variable populations, viz.,

$$\rho_{Y,X_1}, \cdots, \rho_{Y,X_k}, \rho_{X_1,X_2}, \cdots, \rho_{X_{k-1},X_k}$$

The assumptions in Box 4.3.1 are sufficient for making valid inferences about the quantities in (a), but in some situations the investigator may also be interested in making inferences about the quantities in (b) and (c). If data are obtained by sampling with preselected values of $X_1, \ldots, X_k$, we cannot make valid inferences about the parameters in (b) and (c) unless every subpopulation is sampled and the relative subpopulation sizes are known, which is almost never the case in real applications. If data are obtained by simple random sampling, then valid point estimates of the quantities in (b) and (c) are available. See (1.6.1), (1.6.2), and (1.6.3). A more restrictive set of assumptions, referred to as assumptions (B) and given in Box 4.3.2, is sufficient for making inferences about *all* of the quantities in (a), (b), and (c).

BOX 4.3.2 Assumptions (B) for Multiple Linear Regression

(Population) Assumption 1. The study population $\{(Y, X_1, \ldots, X_k)\}$ is a $(k+1)$-variable Gaussian population.

(Sample) Assumption 2. The sample data are obtained by simple random sampling described in Section 2.3; i.e., a simple random sample of n items is selected from the population and the values of the variables $Y, X_1, \ldots, X_k$, are observed.

(Sample) Assumption 3. The sample values $y_i, x_{i,1}, \ldots, x_{i,k}$, for $i = 1, \ldots, n$ are measured without error.

We make several comments about the assumptions.

1 For assumptions (B) to be met, sample data must be obtained by simple random sampling.

2 If $\{(Y, X_1, \ldots, X_k)\}$ is a $(k+1)$-variable Gaussian population as in (population) assumption 1 of Box 4.3.2, then (population) assumptions 1, 2, 3 in Box 4.3.1 are automatically satisfied, and $\mu_Y(x_1, \ldots, x_k)$ is indeed of the form

$$\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Thus, *assumptions* (B) *for multiple linear regression imply assumptions* (A); *however, the converse is true only when the k-variable population* $\{(X_1, \ldots, X_k)\}$ *is Gaussian* and data are obtained by simple random sampling.

3 Strictly speaking, in (population) assumption 1 of Box 4.3.1, we should clearly state what the allowable values of the predictor variables are. This can be done by stating a lower limit and an upper limit for each predictor variable: for instance, $a_1 \le X_1 \le b_1, \ldots, a_k \le X_k \le b_k$. For simplicity of presentation, we often omit specification of these limits, but investigators and statisticians should know what these limits are for each given problem.

4 If we are interested only in point estimation and not in confidence intervals or tests, then (population) assumptions 1 and 2 and (sample) assumptions 4 and 5 of regression assumptions (A) in Box 4.3.1 are sufficient, and the Gaussian assumption is not needed.

Some Comments on Measurement Errors

In many applied problems, the values of $X_1, \ldots, X_k$ and Y of the sample items are not known exactly but are subject to measurement errors. For many applications of regression, measurement errors can be ignored provided that they are small. For such situations the inference procedures developed in Chapters 3 through 7 should be adequate. However, there are situations where measurement errors should not be ignored, and certain modifications may be required for the inference procedures to be valid. For the straight line regression model, these are discussed in Section 3.10. For the multiple linear regression model involving measurement errors in the predictor variables $X_1, \ldots, X_k$, and possibly in the response variable Y, results similar to those in Section 3.10 are applicable. You should consult more advanced textbooks [7].

Often, when discussing statistical procedures, we make no mention of measurement errors. While we want you to be aware of the additional complications that may

arise when measurement errors are not small, we want to assure you that the methods developed in Chapters 3 through 7, under the assumption of no measurement errors, are adequate for most applications.

Unless specifically stated otherwise, we assume that the values of $X_1, \ldots, X_k$ and Y of the sample items are measured without appreciable error and that (sample) assumption 5 in Box 4.3.1 and/or (sample) assumption 3 in Box 4.3.2 are satisfied.

Note that the (artificial) population of Task 4.2.1 satisfies (population) assumptions 1 and 2 in Box 4.3.1. It can never be determined if all of the assumptions in Box 4.3.1 or 4.3.2 are exactly satisfied in a real problem. Investigators may not know for certain that $\mu_Y(x_1, \ldots, x_k)$ is of the form in (4.1.1), but they can generally determine the form of the regression function that fits the problem so that the theory will be approximately valid. The same can be said for all of the assumptions. The sample assumptions mainly concern collecting the data, and often the investigator is restricted by money, time, or other constraints so the data collection methods may not exactly meet the requirements of randomness, etc. Sometimes the investigator who must analyze the data and draw conclusions from them is not the one who collected the data. We may know or suspect that some errors were made in the sampling procedures or in recording the data. The view we take is that all data contain some information, and the investigator is in the best position to determine whether the assumptions are close enough to being satisfied to allow valid conclusions to be drawn about the populations under study. Investigators should always be aware of abnormalities in the data and deal with them. In this chapter, we give inference procedures that are valid if assumptions (A) or (B) for multiple linear regression hold. However, the results are generally reliable and useful when the assumptions are approximately satisfied or if the sample size is large.

The next section discusses point estimation for the unknown parameters in the multiple linear regression model. In Section 4.5, we consider some methods for examining the validity of assumptions (A) and (B) and, if they appear not to be satisfied, alternative procedures are sometimes available. These are discussed in Chapter 8.

4.4 Point Estimation

One of the main objectives of multiple regression analysis is to use sample data to obtain point and confidence interval estimates for the unknown quantities $\beta_0, \beta_1, \ldots, \beta_k$, $\mu_Y(x_1, \ldots, x_k)$, $Y(x_1, \ldots, x_k)$, and σ, and also for selected functions of these quantities. In Section 3.4 we discussed point estimation for straight line regression. In this section we discuss point estimation for multiple linear regression with k predictor factors when assumptions (A) and (B) are satisfied. Consequently, the population regression function is of the form given in (4.1.1). A sample of size n is selected using simple random sampling or by sampling with preselected values of $X_1, \ldots, X_k$ from the population $\{(Y, X_1, \ldots, X_k)\}$ of N items. The sample data may be organized as shown in Table 4.4.1.

The *method of least squares* is used to obtain point estimates of $\beta_0, \beta_1, \ldots, \beta_k$, and we discuss this next.

TABLE 4.4.1
Schematic Representation of Sample Data Values

Y	X_1	X_2	$\cdots$	X_k
y_1	$x_{1,1}$	$x_{1,2}$	$\cdots$	$x_{1,k}$
y_2	$x_{2,1}$	$x_{2,2}$	$\cdots$	$x_{2,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_i	$x_{i,1}$	$x_{i,2}$	$\cdots$	$x_{i,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{n,1}$	$x_{n,2}$	$\cdots$	$x_{n,k}$

Least Squares Estimates of $\beta_0, \beta_1, \ldots, \beta_k$

Consider the population regression function

$$\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{4.4.1}$$

which we would like to use to predict Y with $X_1, \ldots, X_k$ as predictor factors. However, since the β_i are not known, we cannot use this function. Thus we must use sample data given in Table 4.4.1 to obtain estimates of the β_i and an estimate of $\mu_Y(x_1, \ldots, x_k)$ in (4.4.1). To obtain estimates of β_i, we use the **principle of least squares**. The resulting estimates are denoted by $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, respectively. The corresponding estimate of the regression function is denoted by

$$\hat{\mu}_Y(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \tag{4.4.2}$$

Consider the n sample observations in Table 4.4.1. The predicted Y value of the ith sample item, with $x_{i,1}, \ldots, x_{i,k}$ as the values of the predictor variables, is given by $\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k}) = \hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k}$. The corresponding prediction error is denoted by $\hat{e}_i$ and is given by

$$\hat{e}_i = y_i - [\hat{\beta}_0 + \hat{\beta}_1 x_{i,1} + \cdots + \hat{\beta}_k x_{i,k}] \tag{4.4.3}$$

The quantities $\hat{e}_i$, $i = 1, \ldots, n$ are called *residuals*. They are useful in examining the validity of the assumptions given in Box 4.3.1 as well as those given in Box 4.3.2. This is discussed in Section 4.5.

The least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ of $\beta_0, \beta_1, \ldots, \beta_k$ are chosen in such a way that the quantity $SSE(X_1, \ldots, X_k)$, which is defined by

$$SSE(X_1, \ldots, X_k) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_k x_{i,k})^2 = \sum_{i=1}^{n} \hat{e}_i^2 \tag{4.4.4}$$

attains its smallest possible value among all the possible choices we could make for estimating $\beta_0, \beta_1, \ldots, \beta_k$. The corresponding minimum value of $SSE(X_1, \ldots, X_k)$ is called the **sum of squared errors** for predicting Y using the estimated regression

function

$$\hat{\mu}_Y(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

Note that we are really not interested in predicting the Y values of the sample items because we already know their true values, namely $y_1, \ldots, y_n$. But if the estimated regression function

$$\hat{\mu}_Y(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k$$

is a good predictor of the *known sample values* y_i corresponding to the n sample items, then we have reason to expect that it will be a good prediction function for *all* values of Y in the population. Thus we use the y_i and the $x_{i,j}$, $j = 1, \ldots, k$, values of the items in the sample to assess the performance of the estimated regression function.

For the sake of completeness, we now enunciate the **principle of least squares.**

The principle of least squares states that the best estimate of the population regression function $\mu_Y(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$, using sample data, is obtained by choosing $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ in (4.4.2) in such a way that the sum of squares of the prediction errors

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_k x_{i,k})^2 \qquad (4.4.5)$$

is the minimum possible.

Normal Equations

It is convenient to use matrices to compute the least squares estimates of $\beta_0, \beta_1, \ldots, \beta_k$. It can be shown mathematically that the least squares estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ of the parameters $\beta_0, \beta_1, \ldots, \beta_k$ are obtained by solving the following matrix equation for $\hat{\boldsymbol{\beta}}$:

$$\boldsymbol{X}^T \boldsymbol{X} \hat{\boldsymbol{\beta}} = \boldsymbol{X}^T \mathbf{y} \qquad (4.4.6)$$

where $\mathbf{y}$ is an $n \times 1$ vector, $\boldsymbol{X}$ is an $n \times (k+1)$ matrix, and $\hat{\boldsymbol{\beta}}$ is a $(k+1) \times 1$ vector, given by

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,k} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{i,1} & x_{i,2} & \cdots & x_{i,k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,k} \end{bmatrix} \quad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_i \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \qquad (4.4.7)$$

respectively. Note that the $\mathbf{y}$ vector and $\mathbf{X}$ matrix are known since the entries are the data values, except that $\mathbf{X}$ has a column of 1's as its first column. The equations in (4.4.6) are known as the **normal equations**. The solution of these normal equations is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{4.4.8}$$

provided that the inverse of $\mathbf{X}^T\mathbf{X}$ exists (which is always the case in this book). The estimate of the regression function $\mu_Y(x_1, \ldots, x_k)$ is

$$\hat{\mu}_Y(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \tag{4.4.9}$$

and the *best* predicted Y value for a randomly chosen item from the subpopulation with $X_1 = x_1, \ldots, X_k = x_k$ is

$$\hat{Y}(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \tag{4.4.10}$$

Note that, as in the case of straight line regression, the regression function is the best function for predicting Y using $X_1, \ldots, X_k$.

EXAMPLE 4.4.1

In this example we use a small (artificial) data set with three predictors (i.e., $k = 3$) to illustrate how to form the $\mathbf{y}$ vector and the $\mathbf{X}$ matrix from the data. We also show how to use matrix calculations to solve for $\hat{\boldsymbol{\beta}}$. The data are given below in Table 4.4.2 and are also in the file **table442.dat** on the data disk. The regression function of Y on $X_1, \ldots, X_k$ is assumed to be of the form

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

TABLE 4.4.2

Y	X_1	X_2	X_3
6	3	9	16
9	6	13	13
12	4	3	17
5	8	2	10
13	3	4	9
2	2	4	7

and assumptions (A) are presumed to hold. Thus the y vector, the X matrix, and the $\hat{\beta}$ vector are

$$y = \begin{bmatrix} 6 \\ 9 \\ 12 \\ 5 \\ 13 \\ 2 \end{bmatrix} \quad X = \begin{bmatrix} 1 & 3 & 9 & 16 \\ 1 & 6 & 13 & 13 \\ 1 & 4 & 3 & 17 \\ 1 & 8 & 2 & 10 \\ 1 & 3 & 4 & 9 \\ 1 & 2 & 4 & 7 \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix}$$

From these we get

$$X^T = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 3 & 6 & 4 & 8 & 3 & 2 \\ 9 & 13 & 3 & 2 & 4 & 4 \\ 16 & 13 & 17 & 10 & 9 & 7 \end{bmatrix} \quad X^T X = \begin{bmatrix} 6 & 26 & 35 & 72 \\ 26 & 138 & 153 & 315 \\ 35 & 153 & 295 & 448 \\ 72 & 315 & 448 & 944 \end{bmatrix} \quad X^T y = \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

So

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} 2.59578 & -0.15375 & -0.01962 & -0.13737 \\ -0.15375 & 0.03965 & -0.00014 & -0.00144 \\ -0.01962 & -0.00014 & 0.01234 & -0.00431 \\ -0.13737 & -0.00144 & -0.00431 & 0.01406 \end{bmatrix} \begin{bmatrix} 47 \\ 203 \\ 277 \\ 598 \end{bmatrix}$$

$$= \begin{bmatrix} 3.20975 \\ -0.07573 \\ -0.11162 \\ 0.46691 \end{bmatrix}$$

Thus

$$\hat{\beta}_0 = 3.20975 \quad \hat{\beta}_1 = -0.07573 \quad \hat{\beta}_2 = -0.11162 \quad \hat{\beta}_3 = 0.46691$$

$$\hat{\mu}_Y(x_1, x_2, x_3) = 3.20975 - 0.07573x_1 - 0.11162x_2 + 0.46691x_3$$

and

$$\hat{Y}(x_1, x_2, x_3) = \hat{\mu}_Y(x_1, x_2, x_3) = 3.20975 - 0.07573x_1 - 0.11162x_2 + 0.46691x_3$$

■

Point Estimates for Linear Functions of $\beta_0, \beta_1, \ldots, \beta_k$

A question of interest to an investigator can often be formulated in terms of a question involving a linear combination of the parameters $\beta_0, \beta_1, \ldots, \beta_k$, given by

$$\theta = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k = a^T\beta$$

where the components a_i in the vector $a^T = [a_0, a_1, \ldots, a_k]$ are specified by the investigator. For example, we may want to estimate $\beta_1 - \beta_2$, in which case $a_1 = 1$, $a_2 = -1$, and all other $a_i = 0$; or we may want to estimate $2\beta_0 - \beta_1 + 3\beta_k$, in which case $a_0 = 2, a_1 = -1, a_k = 3$, and all other a_i equal zero.

The point estimate of $\theta = \boldsymbol{a}^T\boldsymbol{\beta}$ is

$$\hat{\theta} = \boldsymbol{a}^T\hat{\boldsymbol{\beta}} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 + \cdots + a_k\hat{\beta}_k \tag{4.4.11}$$

where the $\hat{\beta}_i$ are computed by the formula in (4.4.8).

Observe that $\mu_Y(x_1, \ldots, x_k)$ itself is a quantity of the form $\boldsymbol{a}^T\boldsymbol{\beta}$ with $a_0 = 1$, $a_1 = x_1, \ldots, a_k = x_k$. Likewise, every β_i is a special case of $\boldsymbol{a}^T\boldsymbol{\beta}$ with $a_i = 1$ and all remaining elements of $\boldsymbol{a}$ equal to zero.

Residuals

If we use the estimated regression function $\hat{\mu}(x_1, \ldots, x_k)$ to predict the Y value of sample item i, which has $x_{i,1}, \ldots, x_{i,k}$ as the values for the predictor variables, the predicted Y value is $\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})$, and the error of prediction is $\hat{e}_i = y_i - \hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})$. See (4.4.3). Thus

$$\hat{e}_i = y_i - \hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k}) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_k x_{i,k} \quad \text{for} \quad i = 1, \ldots, n \tag{4.4.12}$$

The quantities $\hat{e}_i$ are called **residuals** and they can be computed from sample data because $y_i, x_{i,1}, \ldots, x_{i,k}$ and $\hat{\beta}_i$ are all known. As discussed in Chapter 3, residuals are useful in examining the validity of assumptions (A) and (B) given in Boxes 4.3.1 and 4.3.2, respectively. This is discussed in Section 4.5.

Point Estimate of σ (i.e., $\sigma_{Y|X_1,\ldots,X_k}$)

The estimate of σ is given by

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^n \hat{e}_i^2}{(n-k-1)}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_k x_{i,k})^2}{(n-k-1)}} \tag{4.4.13}$$

The quantity

$$SSE(X_1, \ldots, X_k) = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_k x_{i,k})^2 = \sum_{i=1}^n \hat{e}_i^2 \tag{4.4.14}$$

is referred to as the **sum of squared errors** (sometimes also called **sum of squares due to error** or **error sum of squares** or **residual sum of squares**) for the estimated regression of Y on $X_1, \ldots, X_k$. The quantity

$$MSE(X_1, \ldots, X_k) = \frac{SSE(X_1, \ldots, X_k)}{(n-k-1)} \tag{4.4.15}$$

is called the **mean squared error** (or **error mean square**, or **residual mean square**) for the estimated regression of Y on $X_1, \ldots, X_k$. When there is no possibility of confusion, we write *SSE* for $SSE(X_1, \ldots, X_k)$ and *MSE* for $MSE(X_1, \ldots, X_k)$. With this notation, the estimate of σ can be written as

$$\hat{\sigma} = \sqrt{MSE} \tag{4.4.16}$$

Equivalently, the estimate of σ^2 is

$$\hat{\sigma}^2 = MSE = \frac{SSE}{n-k-1} \tag{4.4.17}$$

If we write

$$\hat{\boldsymbol{e}} = \begin{bmatrix} \hat{e}_1 \\ \hat{e}_2 \\ \vdots \\ \hat{e}_n \end{bmatrix} \tag{4.4.18}$$

for the vector of residuals, then by using (4.4.12) for $\hat{e}_i$, we get

$$\hat{\boldsymbol{e}} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

The sum of squared errors *SSE* can be expressed in various equivalent ways using matrix notation. For instance, we have

$$SSE = \hat{\boldsymbol{e}}^T\hat{\boldsymbol{e}} \tag{4.4.19}$$

$$SSE = (\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) \tag{4.4.20}$$

$$SSE = \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{X}^T\boldsymbol{y} \tag{4.4.21}$$

$$SSE = \boldsymbol{y}^T[\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T]\boldsymbol{y} \tag{4.4.22}$$

Degrees of Freedom Associated with $\hat{\sigma}^2$

Note that to obtain $\hat{\sigma}^2$ we divide the sum of squared errors *SSE* by $(n-k-1)$ (see (4.4.17)). This number $(n-k-1)$ is called the **degrees of freedom for error** or the **degrees of freedom associated with** $\hat{\sigma}^2$. Observe that *the degrees of freedom for error equals n, the sample size, minus the number of β's in the model.* This is a general rule. In Chapter 1 we saw that the estimated variance for a one-variable population $\{Y\}$ is

$$\hat{\sigma}_Y^2 = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_Y)^2}{(n-1)}$$

the divisor is $(n-1)$ because the model is $y_i = \mu_Y + e_i$, and there is one β (say β_0, represented by μ_Y) in the model. In Chapter 3 the estimated variance $\hat{\sigma}^2$ was seen to be

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{(n-2)}$$

The divisor is $(n-2)$ because the straight line regression function $\mu_Y(x) = \beta_0 + \beta_1 x$ contains two β's (β_0 and β_1).

Straight Line Regression

Observe that the formula in (4.4.8) for $\hat{\boldsymbol{\beta}}$ is also valid for straight line regression. In that case it is a reexpression of formulas (3.4.8) and (3.4.9) using matrix notation. Note that for straight line regression, the $\boldsymbol{X}$ matrix is of size n by 2, the vectors $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}$ are 2 by 1, and the vector $\boldsymbol{y}$ is n by 1. We display them here.

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \qquad \hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \qquad \boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_n \end{bmatrix} \tag{4.4.23}$$

Whereas matrix notation is not really necessary for expressing the formulas for parameter estimates in straight line regression, it is invaluable for multiple linear regression; otherwise the formulas for parameter estimates would be too cumbersome, if not impossible, to write down.

Estimates of population parameters in a regression problem may be calculated using formulas (4.4.8), (4.4.9), (4.4.10), (4.4.14), (4.4.15), (4.4.16), etc., but the computations are tedious. It is often convenient to use any one of several readily available statistical packages such as SAS, SPSS, SPLUS, BMDP, MINITAB, and so forth, to obtain the estimates. The printouts from these packages are quite similar in content, but differences do exist in the style of presentation of the results. We will present computer output from MINITAB or SAS whenever appropriate. The following example illustrates a typical computer output (obtained using MINITAB) for regression analysis that can be used for obtaining estimates of the unknown parameters in the multiple linear regression model.

EXAMPLE 4.4.2

Consider the situation described in Example 4.2.1. Suppose that the population regression function of Y (GPA at the end of one year) on X_1 (SATmath), X_2 (SATverbal), X_3 (HSmath), and X_4 (HSenglish) is of the form

$$\mu_Y(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{4.4.24}$$

and that (population) assumptions (B) for multiple linear regression hold. The director of admissions decides to obtain an estimate of this regression function based on a sample of 20 students, obtained using simple random sampling, from all students who were admitted over the past four years and who completed their first year. The data for the 20 students in the sample are given in Table 4.4.3. For convenience these data are also stored in the file **gpa.dat** on the data disk. The MINITAB output from a regression analysis of Y on X_1, X_2, X_3, X_4 for the data in Table 4.4.3 is shown in Exhibit 4.4.1. As mentioned earlier, other computer packages produce similar output but may differ slightly in the style of presentation. Only portions of the MINITAB output that are relevant to the present discussion are shown in Exhibit 4.4.1. Other aspects of the MINITAB output will be discussed as the need arises.

TABLE 4.4.3
GPA Data

Student i	First-Year GPA Y	SATmath X_1	SATverbal X_2	HSmath X_3	HSenglish X_4
1	1.97	321	247	2.30	2.63
2	2.74	718	436	3.80	3.57
3	2.19	358	578	2.98	2.57
4	2.60	403	447	3.58	2.21
5	2.98	640	563	3.38	3.48
6	1.65	237	342	1.48	2.14
7	1.89	270	472	1.67	2.64
8	2.38	418	356	3.73	2.52
9	2.66	443	327	3.09	3.20
10	1.96	359	385	1.54	3.46
11	3.14	669	664	3.21	3.37
12	1.96	409	518	2.77	2.60
13	2.20	582	364	1.47	2.90
14	3.90	750	632	3.14	3.49
15	2.02	451	435	1.54	3.20
16	3.61	645	704	3.50	3.74
17	3.07	791	341	3.20	2.93
18	2.63	521	483	3.59	3.32
19	3.11	594	665	3.42	2.70
20	3.20	653	606	3.69	3.52

EXHIBIT 4.4.1
MINITAB Output for Regression Analysis of Data in Table 4.4.3

```
The regression equation is
GPA = 0.162 + 0.00201 SATmath + 0.00125 SATverb +
      0.189 HSmath + 0.088 HSengl                          (4.4.25)

Predictor       Coef        Stdev    t-ratio       p
Constant      0.1615       0.4375       0.37   0.717   (4.4.26)
SATmath   0.0020102   0.0005844       3.44   0.004   (4.4.27)
SATverb   0.0012522   0.0005515       2.27   0.038   (4.4.28)
HSmath      0.18944      0.09187       2.06   0.057   (4.4.29)
HSengl       0.0876       0.1765       0.50   0.627   (4.4.30)

s = 0.2685      R-sq = 85.3%     R-sq(adj) = 81.4%      (4.4.31)
```

The estimated regression equation is in (4.4.25), and the values of $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_3$, and $\hat{\beta}_4$ can be obtained from this. They are also listed under the heading `Coef` in (4.4.26), (4.4.27), (4.4.28), (4.4.29), and (4.4.30), respectively. The value of $\hat{\sigma}$ is the quantity labeled `s` in (4.4.31). Thus

$$\hat{\beta}_0 = 0.1615 \qquad \hat{\beta}_1 = 0.0020102 \qquad \hat{\beta}_2 = 0.0012522 \qquad \hat{\beta}_3 = 0.18944$$

$$\hat{\beta}_4 = 0.0876 \qquad \hat{\sigma} = 0.2685$$

Note that the $\hat{\beta}_i$ given in the estimated regression equation are often *rounded* values of the $\hat{\beta}_i$ given under `Coef`. The estimated regression function of Y on X_1, X_2, X_3, X_4 is

$$\hat{\mu}_Y(x_1, x_2, x_3, x_4) = 0.162 + 0.00201x_1 + 0.00125x_2 + 0.189x_3 + 0.088x_4 \tag{4.4.32}$$

and the best prediction $\hat{Y}(x_1, x_2, x_3, x_4)$ of the Y value of an item from the subpopulation with $X_1 = x_1, X_2 = x_2, X_3 = x_3, X_4 = x_4$ is

$$\hat{Y}(x_1, x_2, x_3, x_4) = 0.162 + 0.00201x_1 + 0.00125x_2 + 0.189x_3 + 0.088x_4 \tag{4.4.33}$$

For example, the predicted first-year GPA for an applicant with SATmath = 730, SATverbal = 570, HSmath = 3.2, and HSenglish = 2.7, is

$$\begin{aligned}\hat{Y}(730, 570, 3.2, 2.7) &= 0.162 + 0.00201(730) + 0.00125(570) \\ &\quad +0.189(3.2) + 0.088(2.7) \\ &= 3.2 \text{ (rounded to one decimal)}\end{aligned}$$

The director of admissions can use this information to make a decision regarding whether or not this student should be recommended for financial assistance.

We must not lose sight of the fact that such predictions may not be reliable. For instance, it is hard to believe that the applicant just referred to would get a GPA of exactly 3.2 at the end of the first year. To get an idea of how good the prediction is, we must also compute appropriate confidence intervals. This is discussed in Section 4.6.

Suppose for a moment that the true subpopulation standard deviation σ is indeed equal to 0.2685. Then, based on the assumption that the subpopulations of Y values are Gaussian, a proportion $p = 0.8$ of the first-year GPAs are within $z_{0.9} \times (0.2685) = (1.28) \times (0.2685) = 0.3437$ unit of $\mu_Y(x_1, x_2, x_3, x_4)$, the mean of the subpopulation to which the individual belongs. These calculations suggest that, if $\mu_Y(x_1, x_2, x_3, x_4)$ is used to predict the Y values of applicants, then 80% of the GPAs will be within 0.3437 unit of $\mu_Y(x_1, x_2, x_3, x_4)$. However, to account for the fact that the number 0.2685 is the *estimated value and not the true value* of σ, we must use a confidence interval for σ. This is discussed in Section 4.6.

In Exhibit 4.4.2 we give a SAS output for the GPA data of Example 4.4.2 analogous to the MINITAB output in Exhibit 4.4.1. Compare the SAS output with the MINITAB output and note the similarities and differences. Only those portions

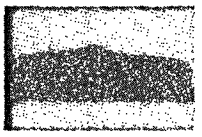
EXHIBIT 4.4.2
SAS Output for Regression Analysis of Data in Table 4.4.3

```
                    The SAS System   0:00 Saturday,   Jan 1, 1994

Model: MODEL1
Dependent Variable: GPA

                          Analysis of Variance

                          Sum of          Mean
Source          DF        Squares        Square      F Value      Prob>F

Model            4        6.26432       1.56608       21.721      0.0001
Error           15        1.08150       0.07210
C Total         19        7.34582

    Root MSE       0.26851     R-square      0.8528                    (4.4.34)
    Dep Mean       2.59300     Adj R-sq      0.8135
    C.V.          10.35535

                       Parameter Estimates

                 Parameter      Standard   T for H0:
Variable   DF    Estimate         Error  Parameter=0  Prob > |T|

INTERCEP    1    0.161550    0.43753205        0.369      0.7171    (4.4.35)
SATMATH     1    0.002010    0.00058444        3.439      0.0036    (4.4.36)
SATVERB     1    0.001252    0.00055152        2.270      0.0383    (4.4.37)
HSMATH      1    0.189440    0.09186804        2.062      0.0570    (4.4.38)
HSENGL      1    0.087564    0.17649628        0.496      0.6270    (4.4.39)
```

of the SAS output that are relevant to the present discussion are shown in this exhibit. Other parts of the SAS output will be discussed as the need arises.

The estimates of $\beta_0, \beta_1, \beta_2, \beta_3$, and β_4 are given in (4.4.35)–(4.4.39), respectively, under the column labeled `Parameter Estimate`. The estimate of σ is in (4.4.34) corresponding to the label `Root MSE`. Compare these results with the results obtained from MINITAB. ■

In Task 4.4.1 we show how regression can be useful for solving practical problems. Use the computer output and perform all the subsequent calculations needed to solve the problems.



Task 4.4.1

The questions in this task refer to the (artificial) population given in Table D-4 in Appendix D and also in the file **plastic.dat** on the data disk. Refer to Task 4.2.1 at this point.

Suppose we undertake a study to determine the relationship of Y, the strength of plastic containers with temperature X_1 and pressure X_2 for the population in the file **plastic.dat**, but the population data are unavailable. However, a simple random sample of size 16 from this population is available. They are given in Table 4.4.4 and are also stored in the file **table444.dat** on the data disk.

Even though we have the complete population in this example, we use the sample data to illustrate how they can be used to make inferences about population quantities. This gives us an opportunity to compare the estimated parameter values with the true population parameter values.

TABLE 4.4.4

A Simple Random Sample of Size 16 from the Plastic Data Population

Sample Item Number	Population Item Number	Strength Y	Temperature (°C) X_1	Pressure (psi) X_2
1	1150	36.6	260	10
2	1186	20.7	230	18
3	200	36.5	290	18
4	1305	16.4	200	16
5	783	23.2	200	10
6	1066	26.6	230	14
7	1023	22.5	210	16
8	448	17.0	200	20
9	945	32.7	290	18
10	508	34.4	260	10
11	704	32.4	260	12
12	1135	24.8	240	18
13	107	26.8	220	12
14	742	37.7	280	12
15	749	26.7	260	20
16	1585	24.6	250	20

Suppose that assumptions (A) for multiple linear regression are satisfied and that the regression function is $\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. For questions (1) through (7), we express the answer in three parts. They are

a In terms of the *symbols* for the **population parameters**, namely β_0, β_1, etc.

b In terms of the *values* of the **population parameters** from Task 4.2.1.

c In terms of *estimates* of these **population parameters** based on the sample data in Table 4.4.4.

Even though part (c) is the answer we seek, we deliberately provide parts (a) and (b) so you can get accustomed to thinking about the population, the population parameters, and finally the estimates of these parameters.

A SAS output from a regression analysis of the sample data in Table 4.4.4 appears in Exhibit 4.4.3.

EXHIBIT 4.4.3

SAS Output for Regression Analysis of Data in Table 4.4.4

The SAS System 0:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: STRENGTH

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	678.56980	339.28490	153.361	0.0001
Error	13	28.76020	2.21232		
C Total	15	707.33000			

Root MSE	1.48739	R-square	0.9593	(4.4.40)
Dep Mean	27.47500	Adj R-sq	0.9531	
C.V.	5.41361			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|	
INTERCEP	1	-6.522361	3.36888546	-1.936	.0749	(4.4.41)
TEMP	1	0.192693	0.01235387	15.598	0.0001	(4.4.42)
PRESSURE	1	-0.834794	0.10145367	-8.228	0.0001	(4.4.43)

1 Estimate the regression function $\mu_Y(x_1, x_2)$, the mean strength of the plastic containers corresponding to a production temperature of x_1 (degrees Celsius), and pressure of x_2 (pounds per square inch); i.e., find $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\mu}_Y(x_1, x_2)$.

a *The population regression function has the form*

$$\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

b *In Task 4.2.1 we found the population regression function to be*

$$\mu_Y(x_1, x_2) = -5 + 0.2x_1 - x_2$$

c *From (4.4.41)–(4.4.43) the sample estimate of the regression function is*

$$\hat{\mu}_Y(x_1, x_2) = -6.52 + 0.193x_1 - 0.835x_2$$

2 Estimate the difference between the average strengths of two batches of plastic containers if they were produced using the same value for pressure, but their production temperatures differed by 1°C.

Suppose the production temperature and pressure for one batch of plastic containers are x_1 (°C) and x_2 (psi), whereas for the second batch they are $x_1 + 1$ (°C) and x_2 (psi), respectively.

a *The difference between the average strengths of these two batches is* $\mu_Y(x_1 + 1, x_2) - \mu_Y(x_1, x_2) = [\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2] = \beta_1$.

b *The population value of β_1 is* 0.2.

c *The estimate of β_1 based on the sample is $\hat{\beta}_1 = 0.193$ from (4.4.42).*

3 Estimate the difference between the average strengths of two batches of products if they were produced using the same value for temperature but their production pressures differed by 1 psi.

Suppose the production temperature and pressure for one batch of plastic containers are x_1 (°C) and x_2 (psi), whereas for the second batch they are x_1 (°C) and $x_2 + 1$ (psi), respectively.

a *The difference between the average strengths of these two batches is* $\mu_Y(x_1, x_2 + 1) - \mu_Y(x_1, x_2) = [\beta_0 + \beta_1 x_1 + \beta_2(x_2 + 1)] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2] = \beta_2$.

b *The population value of β_2 is* -1.0.

c *The estimated value of β_2 is $\hat{\beta}_2 = -0.835$ from (4.4.43).*

4 Estimate the difference between the average strengths that would result if a temperature of 260°C is used during production instead of 250°C, while the pressure is kept fixed at some value, say x_2.

a *The required quantity is* $\mu_Y(260, x_2) - \mu_Y(250, x_2) = [\beta_0 + \beta_1(260) + \beta_2 x_2] - [\beta_0 + \beta_1(250) + \beta_2 x_2] = 10\beta_1$.

b *The population value of this quantity is* $10\beta_1 = 10(0.2) = 2.0$.

c *The estimated value of this quantity is* $10\hat{\beta}_1 = 10(0.193) = 1.93$.

Thus, based on the sample, we estimate that on the average the strength of the containers will increase by 1.93 units if a temperature of 260°C is used during production instead of a temperature of 250°C, while the pressure is kept fixed at some value.

5 If a pressure of 16 psi is used during production instead of 14 psi, while the temperature is kept fixed at some value, say x_1, estimate the resulting difference in the average strength of the plastic containers.

a *The quantity of interest is* $\mu_Y(x_1, 16) - \mu_Y(x_1, 14) = [\beta_0 + \beta_1 x_1 + \beta_2(16)] - [\beta_0 + \beta_1 x_1 + \beta_2(14)] = 2\beta_2$.

b *The population value of this quantity is* $2\beta_2 = 2(-1) = -2.0$.

c *The estimated value of this quantity is* $2\hat{\beta}_2 = 2(-0.835) = -1.67$.

Thus the sample data suggest that using a pressure of 16 psi instead of 14 psi, while keeping the temperature fixed at some value will decrease the average strength of the plastic containers by 1.67 units.

6 Estimate the difference in the average strengths between containers manufactured at a temperature of 260°C and pressure of 16 psi and those manufactured at a temperature of 240°C and pressure of 14 psi.

a *The quantity of interest is* $\mu_Y(260, 16) - \mu_Y(240, 14) = [\beta_0 + 260\beta_1 + 16\beta_2] - [\beta_0 + 240\beta_1 + 14\beta_2] = 20\beta_1 + 2\beta_2$.

b *The population value of this quantity is* $20(0.2) + 2(-1) = 2$.

c *The estimated value of this quantity is* $20(0.193) + 2(-0.835) = 2.19$.

7 Estimate the standard deviation of the strength of plastic containers manufactured under identical temperature and pressure conditions.

First note that plastic containers, manufactured under identical temperature and pressure conditions, all belong to the same subpopulation.

a *Recall from the results of Task 4.2.1 that the subpopulation standard deviations for strength are all equal. This common subpopulation standard deviation is denoted by* σ.

b *The population value of the standard deviation is* $\sigma = 1.7076$ *(from Task 4.2.1).*

c *The sample estimate of this standard deviation is* $\hat{\sigma} = 1.487$ *from (4.4.40).*

8 What is the interpretation of the parameter β_0 in this problem?

The quantity β_0 *is the value of the regression function when* X_1 *is zero and* X_2 *is zero. Do not, however, be tempted to conclude that this is the average strength of the plastic containers when the process temperature is set at* 0°C *and pressure at 0 psi. We know the regression function to be valid only for the range of values 200–300°C for temperature, and 10–20 psi for pressure. The pair of values, 0°C and 0 psi, are far removed from the allowable range of values, and no meaningful conclusions can be derived from the value of* β_0 *(or* $\hat{\beta}_0$*) regarding the average strength of the containers manufactured at* 0°C *and 0 psi. In all likelihood, it may not even make any sense to run the process at these values of temperature and pressure.*

Problems 4.4

Use the following information for Problems 4.4.1–4.4.12. A simple random sample of size 10 is selected from the population discussed in Task 4.2.1, where an investigator wants to study the relationship of strength (Y) of plastic containers to the predictor factors temperature (X_1) and pressure (X_2). The sample data are given in Table 4.4.5, and are also stored in the file **table445.dat** on the data disk. We suppose that assumptions (A) for multiple regression are satisfied. A MINITAB output from a regression analysis of Y on X_1 and X_2 is given in Exhibit 4.4.4.

TABLE 4.4.5

Sample Item Number	Population Item Number	Strength Y	Temperature X_1	Pressure X_2
1	1001	40.2	290	12
2	260	38.2	270	12
3	1085	30.2	240	12
4	1267	18.5	210	20
5	733	28.2	250	16
6	1173	35.3	260	12
7	438	27.3	220	12
8	129	28.1	210	10
9	1072	35.0	250	12
10	1381	16.7	210	20

EXHIBIT 4.4.4

MINITAB Output for Regression of Plastic Data in Table 4.4.5

```
The regression equation is
 strength = - 0.65 + 0.187 temp - 1.07 pressure

 Predictor       Coef      Stdev     t-ratio        p
 Constant      -0.646      4.509       -0.14    0.890
 temp          0.18721    0.01477       12.68    0.000
 pressure      -1.0653     0.1157       -9.21    0.000

 s = 1.117       R-sq = 98.4%     R-sq(adj) = 97.9%
```

4.4.1 Exhibit the X matrix and the y vector for these data (see (4.4.7)).

4.4.2 Compute X^TX and X^Ty.

4.4.3 The matrix $(X^TX)^{-1}$ is as follows:

$$(X^TX)^{-1} = \begin{bmatrix} 16.300423 & -0.050442 & -0.293043 \\ -0.050442 & 0.000175 & 0.000602 \\ -0.293043 & 0.000602 & 0.010723 \end{bmatrix}$$

Calculate $\hat{\beta}$ using (4.4.8) and compare the results with the $\hat{\beta}_i$ given in the computer output in Exhibit 4.4.4.

4.4.4 Compute SSY, $SSX(1)$, and $SSX(2)$ where $SSX(i)$ is SSX for X_i.

4.4.5 Are valid estimates available for the following population quantities?

$$\mu_Y \quad \mu_{X_1} \quad \mu_{X_2} \quad \sigma_Y \quad \sigma_{X_1} \quad \sigma_{X_2}$$

Explain.

4.4.6 What is the value of $\hat{\sigma}$ (i.e., $\hat{\sigma}_{Y|X_1,X_2}$)? Compare this with the population value.

4.4.7 Estimate $\mu_Y(300, 16)$ using the sample data in Table 4.4.5. Compare this estimate with the population value.

4.4.8 Write the symbol for the population quantity that represents the best predicted value for the strength of a plastic container that was produced using a temperature of 280°C and a pressure of 19 psi.

4.4.9 In Problem 4.4.8, is that value available from the population? If not, why not?

4.4.10 What is the estimate of $\mu_Y(280, 18)$ using the sample data in Table 4.4.5?

4.4.11 Use the *population values of the parameters* and determine what proportion of the plastic containers has a strength greater than 31 units if they are produced with a temperature of 250°C and a pressure of 16 psi.

4.4.12 Estimate the quantity of interest in Problem 4.4.11 using the data in Table 4.4.5.

4.5 Residual Analysis

Residuals in multiple linear regression were defined in (4.4.12). They are useful for checking the validity of assumptions (A) and (B). For reasons similar to those discussed in Section 3.5, it is customary to examine the residuals after appropriately *standardizing* them.

Standardized Residuals

The **standardized residuals** for multiple linear regression are defined by

DEFINITION

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}} \tag{4.5.1}$$

where $h_{i,i}$ is the ith diagonal element of the matrix

$$H = X(X^T X)^{-1} X^T \quad \blacksquare \tag{4.5.2}$$

The matrix H is sometimes called the **hat matrix** and the quantities $h_{i,i}$ are called the **hat values**. If assumptions (A) or (B) for multiple linear regression are satisfied, then the standardized residuals $r_1, \ldots, r_n$ will be (approximately) a simple random sample of n observations from a Gaussian population with mean zero and unit standard deviation. This fact is useful in examining the validity of some of the model assumptions.

Plotting Standardized Residuals Against the Predictors

In some instances a plot of the standardized residuals r_i against the sample values $x_{i,j}$ of the predictor variable X_j $(j = 1, \ldots, k)$ can be useful in detecting inadequacies in the assumed regression function. When assumptions (A) or (B) are satisfied, the points on this plot should be scattered about the X_j axis in a random fashion, showing no obvious trends or other patterns. If this is found not to be the case, then one or more of assumptions (A) or (B) are likely to be violated.

Plotting Standardized Residuals Against Fitted Values

The estimated subpopulation means $\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})$ corresponding to the sample values $(x_{i,1}, \ldots, x_{i,k})$ are often referred to as **fitted values** or simply **fits**. When assumptions (A) or (B) are satisfied, the fitted values are *independent* of the standardized residuals r_i. Hence the points in the plot of r_i versus $\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})$ for $i = 1, \ldots, n$ should be randomly scattered about the horizontal axis (axis of fitted values), with no specific pattern. If any systematic pattern is observed, then this may indicate violation of one or more of assumptions (A) in Box 4.3.1 or assumptions (B) in Box 4.3.2. In some instances where a pattern is observed instead of a random scatter, it may be possible to diagnose the cause of the observed pattern. Refer to Section 3.5 for details.

Rankit-Plots of Standardized Residuals

If assumptions (A) or (B) are satisfied, then the standardized residuals r_i will be (approximately) a simple random sample from a *standard* Gaussian population (Gaussian with mean zero and standard deviation one). This can be checked graphically

by examining a (Gaussian) rankit-plot of the standardized residuals (i.e., a plot of the standardized residuals against the Gaussian scores) as discussed in Section 3.5.

Rankit-Plots of Linear Combinations of $X_1, \ldots, X_k$ and Y

As discussed in Section 4.3, for inferences on certain parameters we need assumptions (B) in Box 4.3.2 and, in particular, we need the assumption that the population $\{(Y, X_1, \ldots, X_k)\}$ is Gaussian. To check this assumption, we investigate whether or not the sample data $(y_1, x_{1,1}, \ldots, x_{1,k}), \ldots, (y_n, x_{n,1}, \ldots, x_{n,k})$ appear to be a simple random sample from a $(k+1)$-variable Gaussian population. We can do this by examining the rankit-plots of linear combinations of $x_{i,1}, \ldots, x_{i,k}$ and y_i. Recall that a theoretical population $\{(Y, X_1, \ldots, X_k)\}$ is Gaussian if and only if every linear combination $b_0Y + b_1X_1 + b_2X_2 + \cdots + b_kX_k$ is Gaussian. Since only sample data are available, we examine linear combinations of y_i and $x_{i,1}, \ldots, x_{i,k}$ in the sample. In practice, it is impossible to examine *every* linear combination of y_i and $x_{i,1}, \ldots, x_{i,k}$, but we can consider several linear combinations and examine the corresponding rankit-plots. We illustrate in the following example.

EXAMPLE 4.5.1

A utility company is interested in investigating how Y, the electricity consumption by each household, is related to monthly income (X_1), number of persons (X_2) in the household, and the living area (X_3) of the house (or apartment). A simple random sample of 34 households served by this utility company is surveyed, and the following information is obtained for each household:

Y = (total) electric bill (in dollars) for the past year

X_1 = monthly income for the household (in dollars)

X_2 = number of persons in the household

X_3 = living area (in square feet) of the house or apartment

The data are displayed in Table 4.5.1 and are also stored in the file **electric.dat** on the data disk. The investigator assumes that the regression function of Y on X_1, X_2, and X_3 is of the form (at least approximately)

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{4.5.3}$$

for $2{,}000 \le x_1 \le 6{,}000$, $x_2 = 1, 2, 3, 4, 5, 6$, and $500 \le x_3 \le 4{,}000$. We are interested in checking the validity of assumptions (A) for multiple linear regression. For this purpose we carry out a residual analysis of the data. The computations, plots, etc., were done using MINITAB although any regression package can be used. The computer output containing these results is given in Exhibit 4.5.1 where we have also printed the fitted values $\hat{\mu}_Y(x_{i,1}, x_{i,2}, x_{i,3})$ (labeled `fits`), the residuals $\hat{e}_i$ (labeled `residual`), the standardized residuals r_i (labeled `stdresid`), and the Gaussian scores $z_i^{(n)}$ (labeled `nscores`) of the standardized residuals.

TABLE 4.5.1
Electric Bill Data

Household	Bill Y	Income X_1	Persons X_2	Area X_3
1	228	3220	2	1160
2	156	2750	1	1080
3	648	3620	2	1720
4	528	3940	1	1840
5	552	4510	3	2240
6	636	3990	4	2190
7	444	2430	1	830
8	144	3070	1	1150
9	744	3750	2	1570
10	1104	4790	5	2660
11	204	2490	1	900
12	420	3600	3	1680
13	876	5370	1	2550
14	840	3180	7	1770
15	876	5910	2	2960
16	276	3020	2	1190
17	1236	5920	3	3130
18	372	3520	2	1560
19	276	3720	1	1510
20	540	4840	1	2190
21	1044	4700	6	2620
22	552	3270	2	1350
23	756	4420	2	1990
24	636	4480	2	2070
25	708	3820	4	1850
26	960	5740	2	2700
27	1080	5600	3	3030
28	480	3950	2	1700
29	96	2290	3	890
30	1272	5580	5	3270
31	1056	5820	2	2660
32	156	3160	2	1330
33	396	2880	4	1280
34	768	3780	3	1950

EXHIBIT 4.5.1

MINITAB Output for Regression Analysis of Electric Bill Data

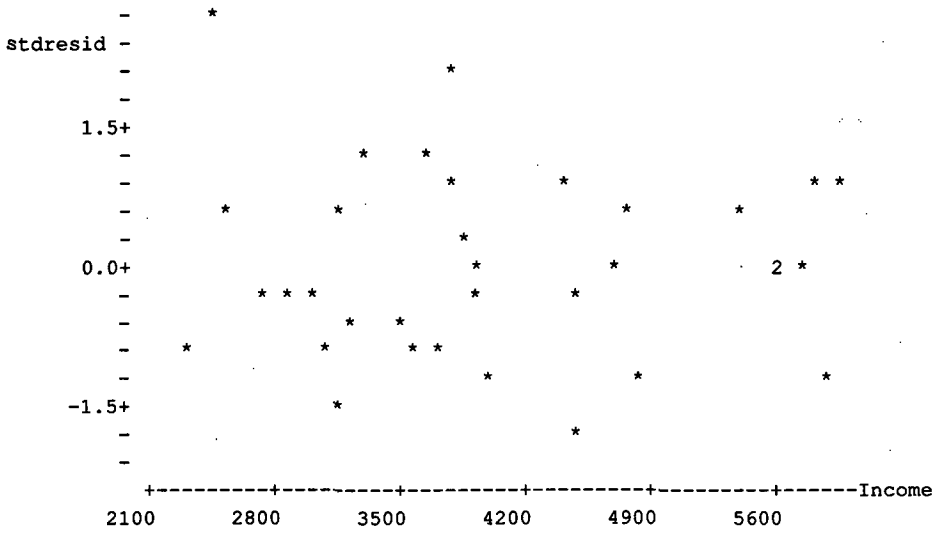
The regression equation is
Bill = - 358 + 0.075 Income + 55.1 Persons + 0.281 Area

Predictor	Coef	Stdev	t-ratio	p
Constant	-358.4	198.7	-1.80	0.081
Income	0.0751	0.1361	0.55	0.585
Persons	55.09	29.05	1.90	0.068
Area	0.2811	0.2261	1.24	0.223

s = 135.4 R-sq = 85.1% R-sq(adj) = 83.7%

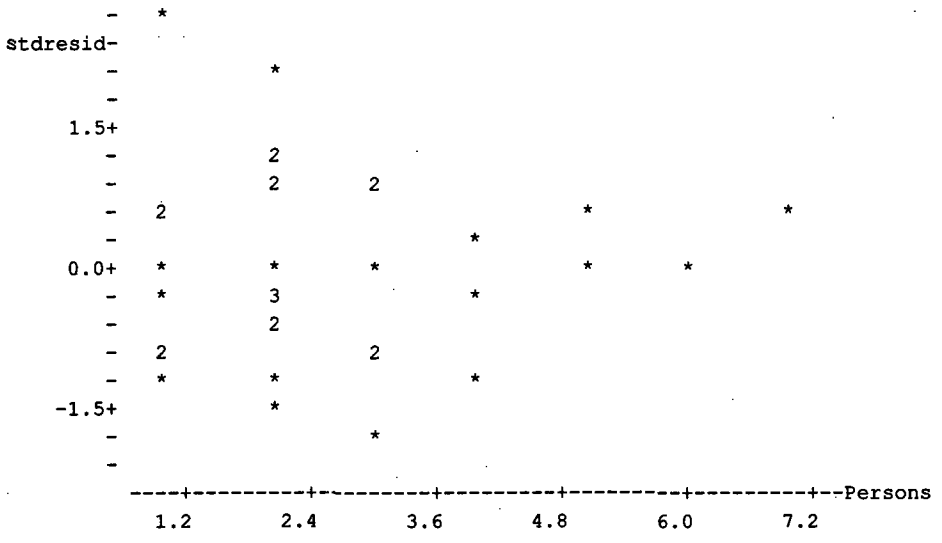
ROW	Bill	Income	Persons	Area	fits	residual	stdresid	nscores
1	228	3220	2	1160	319.75	-91.755	-0.73740	-0.57777
2	156	2750	1	1080	206.86	-50.865	-0.40260	-0.41240
3	648	3620	2	1720	507.23	140.772	1.08468	1.24834
4	528	3940	1	1840	509.92	18.084	0.14376	0.18319
5	552	4510	3	2240	775.36	-223.361	-1.68098	-2.09417
6	636	3990	4	2190	777.32	-141.322	-1.11948	-1.24834
7	444	2430	1	830	112.54	331.455	2.61553	2.09417
8	144	3070	1	1150	250.59	-106.586	-0.82029	-0.66662
9	744	3750	2	1570	474.83	269.170	2.05367	1.67229
10	1104	4790	5	2660	1024.64	79.362	0.62575	0.57777
11	204	2490	1	900	136.73	67.270	0.53344	0.49336
12	420	3600	3	1680	549.57	-129.568	-0.97566	-0.86326
13	876	5370	1	2550	816.95	59.054	0.46565	0.41240
14	840	3180	7	1770	763.66	76.339	0.70903	0.66662
15	876	5910	2	2960	1027.86	-151.860	-1.20401	-1.42855
16	276	3020	2	1190	313.16	-37.161	-0.28362	-0.18319
17	1236	5920	3	3130	1131.49	104.514	0.82864	0.76119
18	372	3520	2	1560	454.74	-82.737	-0.62388	-0.49336
19	276	3720	1	1510	400.62	-124.622	-0.95231	-0.76119
20	540	4840	1	2190	675.93	-135.926	-1.05102	-1.10154
21	1044	4700	6	2620	1061.72	-17.719	-0.14601	-0.03644
22	552	3270	2	1350	376.92	175.079	1.32665	1.42855
23	756	4420	2	1990	643.24	112.765	0.85709	0.86326
24	636	4480	2	2070	670.23	-34.232	-0.25848	-0.10951
25	708	3820	4	1850	668.97	39.026	0.30005	0.33413
26	960	5740	2	2700	942.00	18.000	0.14687	0.25790
27	1080	5600	3	3030	1079.33	0.668	0.00538	0.03644
28	480	3950	2	1700	526.40	-46.401	-0.35265	-0.33413
29	96	2290	3	890	229.07	-133.067	-1.04615	-0.97539
30	1272	5580	5	3270	1255.47	16.530	0.14132	0.10951
31	1056	5820	2	2660	936.77	119.234	1.04082	1.10154
32	156	3160	2	1330	363.03	-207.034	-1.56974	-1.67229
33	396	2880	4	1280	438.12	-42.116	-0.33233	-0.25790
34	768	3780	3	1950	638.99	129.009	1.00438	0.97539

First we plot the standardized residuals r_i against monthly income X_1.



The standardized residuals exhibit only a random scatter about the horizontal line through the origin with `stdresid` = 0 (which is not shown in the plot, but you should draw this line for ease of interpretation) and no definite pattern is seen. So this plot does not point to any violations of assumptions (A).

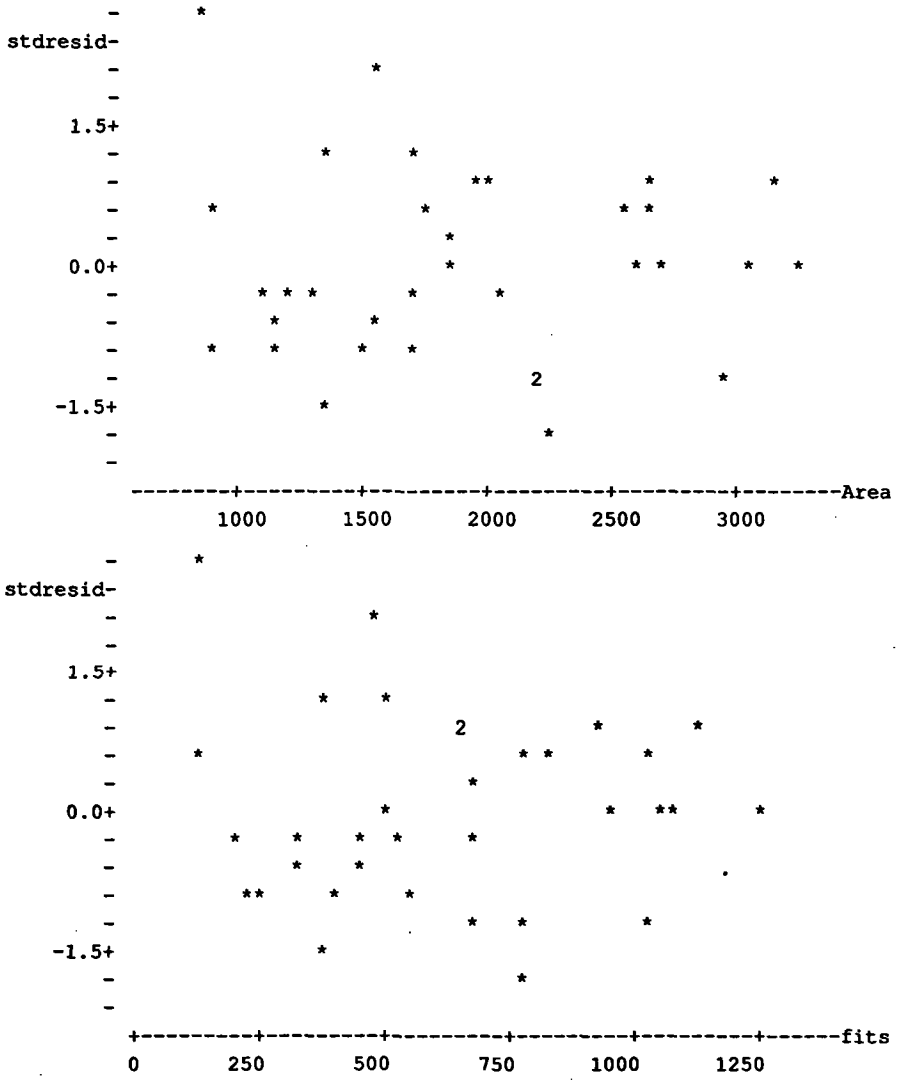
Next we plot the standardized residuals r_i against X_2, the number of persons in the household.



This plot suggests that perhaps the assumption of homogeneity of standard deviations may *not* be satisfied. The standardized residuals seem to vary over a wider range for small values of X_2 than for large values of X_2. However, this apparent

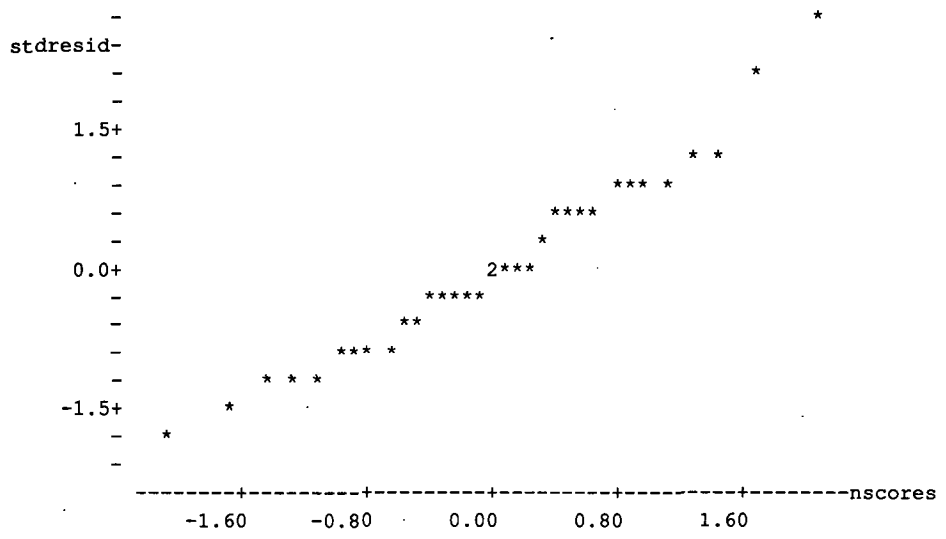
pattern may be due to the fact that there are only seven observations with X_2 values equal to 4 or more. Nevertheless, further examination may be warranted.

Next we plot the standardized residuals r_i against living area X_3 and also against fitted values $\hat{\mu}_Y(x_{i,1}, x_{i,2}, x_{i,3})$.



These plots might suggest the possibility that the standard deviations of subpopulations may not all be the same. There seems to be a greater variability in the values of the standardized residuals corresponding to small values of `fits` (and area) than for large values of `fits` (and area). It is possible that this apparent pattern arises due to the two data values giving rise to the two largest positive standardized residuals (check the data values for sample items 7 and 9). Further investigation may be warranted.

To examine whether or not the subpopulations determined by X_1, X_2, and X_3 are Gaussian, we obtain a rankit-plot of the standardized residuals.



This plot appears to be consistent with the assumption that the residuals are a simple random sample from a standard Gaussian population and there is no indication of violations of the assumption that subpopulations determined by X_1, X_2, X_3 are Gaussian.

To summarize, the only part of assumptions (A) whose validity is in some doubt is the homogeneity of standard deviations. In a real study, the investigator should probably consult a professional statistician for advice. ■

In Section 4.5 of the laboratory manuals we show how to carry out the computations and plots to examine the plausibility of assumptions (A) or (B) using MINITAB and/or SAS.

There may exist other diagnostic procedures which might indicate that the assumptions for multiple linear regression may be violated. Even when these procedures do not suggest the failure of any of the assumptions, there is no guarantee that the assumptions are actually met. This is necessarily a subjective exercise, and at some point the investigator must either conclude that the assumptions required for a valid analysis are satisfied (for all practical purposes) and proceed with the analyses and inferences, or he/she must conclude that one or more of the assumptions is seriously violated and look for alternative procedures. For certain types of violations of assumptions, we discuss alternative procedures in Chapter 8.

Authors' Recommendation

When performing a multiple linear regression analysis of a set of data $(y_1, x_{1,1}, \ldots, x_{1,k}), \ldots, (y_n, x_{n,1}, \ldots, x_{n,k})$, we suggest that you include the following steps.

1 Obtain the standardized residuals r_i and the fitted values $\hat{\mu}_Y(x_{i,1}, \ldots, x_{i,k})$, denoted here by $\hat{\mu}_i$ for ease of notation.

2 Plot r_i against $\hat{\mu}_i$ and also r_i against $x_{i,j}$ for $j = 1, \cdots, k$. Examine these plots for evidence of unequal subpopulation variances or an incorrect model.

3 Obtain a rankit-plot of r_i to evaluate the validity of the assumption that each subpopulation of Y values is a Gaussian population.

4 If you wish to examine the validity of assumptions (B) and the data are obtained by simple random sampling, then examine the Gaussian rankit-plots of y_i, $x_{i,1}, \ldots,$ and $x_{i,k}$, and several linear combinations of these, to assess whether or not the data appear to be a simple random sample from a $(k+1)$-variable Gaussian population.

5 Make an overall evaluation of the validity (at least approximately) of assumptions (A) or (B) within the context of the particular application in question.

Problems 4.5

4.5.1 Consider Task 4.4.1 where an investigator wants to study the relationship of the strength (Y) of plastic containers to the predictor factors, temperature (X_1) and pressure (X_2). A simple random sample of size 16 is selected from the population and assumptions (A) are presumed to be satisfied. The data are given in Table 4.4.4 and are also stored in the file **table444.dat** on the data disk. We want to perform a residual analysis to determine whether assumptions (A) seem to be valid for this problem. A SAS output for regression appears in Exhibit 4.5.2, along with a printout of y_i (STRENGTH), $x_{i,1}$ (TEMP), $x_{i,2}$ (PRESSURE), fitted values $\hat{\mu}_Y(x_{i,1}, x_{i,2})$ (FITS), residuals $\hat{e}_i$ (RESIDUAL), standardized residuals r_i (STDRESID), and Gaussian scores $z_i^{(n)}$ (NSCORES) with $n = 16$.

a Plot the standardized residuals r_i against fits $\hat{\mu}_Y(x_{i,1}, x_{i,2})$ (for convenience round all numbers to one decimal).

b Plot the standardized residuals r_i against X_1 = temperature.

c Plot the standardized residuals r_i against X_2 = pressure.

d Plot the standardized residuals r_i against the Gaussian scores (i.e., rankits) $z_i^{(n)}$.

What do you conclude regarding the plausibility of assumptions (A) for this problem?

EXHIBIT 4.5.2

SAS Output for Problem 4.5.1

The SAS System 0:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: STRENGTH

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	678.56980	339.28490	153.361	0.0001
Error	13	28.76020	2.21232		
C Total	15	707.33000			

Root MSE	1.48739	R-square	0.9593
Dep Mean	27.47500	Adj R-sq	0.9531
C.V.	5.41361		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-6.522361	3.36888546	-1.936	0.0749
TEMP	1	0.192693	0.01235387	15.598	0.0001
PRESSURE	1	-0.834794	0.10145367	-8.228	0.0001

SAMPLE	PROPPANT	TEMP	TIME	RESULT	FIT	RESIDUAL	STUDENT	COOKD
1	1150	36.6	260	10	35.2298	1.37021	1.03884	0.76184
2	1186	20.7	230	18	22.7707	-2.07066	-1.47494	-1.28155
3	200	36.5	290	18	34.3322	2.16778	1.68383	1.76883
4	1305	16.4	200	16	18.6595	-2.25947	-1.68822	-1.76883
5	783	23.2	200	10	23.6682	-0.46823	-0.37926	-0.39573
6	1066	26.6	230	14	26.1098	0.49017	0.34362	0.39573
7	1023	22.5	210	16	20.5864	1.91361	1.38608	1.28155
8	448	17.0	200	20	15.3203	1.67971	1.34590	0.98815
9	945	32.7	290	18	34.3322	-1.63222	-1.26783	-0.98815
10	508	34.4	260	10	35.2298	-0.82979	-0.62912	-0.56918
11	704	32.4	260	12	33.5602	-1.16020	-0.83814	-0.76184
12	1135	24.8	240	18	24.6976	0.10242	0.07251	0.07720
13	107	26.8	220	12	25.8525	0.94750	0.68899	0.56918
14	742	37.7	280	12	37.4141	0.28594	0.21643	0.23349
15	749	26.7	260	20	26.8818	-0.18185	-0.13559	-0.07720
16	1585	24.6	250	20	24.9549	-0.35492	-0.26203	-0.23349

4.5.2 Consider the GPA data of Example 4.4.2, which is also stored in the file **gpa.dat** on the data disk. We want to perform a residual analysis to determine if assumptions (B) seem to be valid for this problem. A MINITAB output for regression appears in Exhibit 4.5.3, along with a printout of the data, the residuals $\hat{e}_i$, fits $\hat{\mu}_Y(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})$, standardized residuals r_i, and Gaussian scores $z_i^{(n)}$ with $n = 20$.

- **a** Plot the standardized residuals r_i against the fits $\hat{\mu}_Y(x_{i,1}, x_{i,2}, x_{i,3}, x_{i,4})$ (for convenience round all numbers to one decimal).
- **b** Plot the standardized residuals r_i against X_1 = SATmath.
- **c** Plot the standardized residuals r_i against X_2 = SATverbal.
- **d** Plot the standardized residuals r_i against X_3 = HSmath.
- **e** Plot the standardized residuals r_i against X_4 = HSenglish.
- **f** Obtain Gaussian rankit-plots for several linear combinations of Y, X_1, X_2, X_3, and X_4.

Based on parts (a)–(f), judge whether or not the five-variable population $\{(Y, X_1, X_2, X_3, X_4)\}$ may be assumed to be multivariate Gaussian.

4.5.3 In Problem 4.5.2, plot the standardized residuals r_i against the Gaussian scores (i.e., rankits) $z_i^{(n)}$. Assess whether or not the standardized residuals appear to be a simple random sample from a standard Gaussian population.

4.5.4 Based on the results of Problems 4.5.2 and 4.5.3, decide whether or not assumptions (B) appear to hold for this problem.

EXHIBIT 4.5.3

MINITAB Output for Problem 4.5.3

```
The regression equation is
GPA = 0.162 + 0.00201 SATmath + 0.00125 SATverb +
      0.189 HSmath + 0.088 HSengl

Predictor       Coef       Stdev    t-ratio       p
Constant      0.1615      0.4375       0.37   0.717
SATmath    0.0020102   0.0005844       3.44   0.004
SATverb    0.0012522   0.0005515       2.27   0.038
HSmath       0.18944     0.09187       2.06   0.057
HSengl        0.0876      0.1765       0.50   0.627

s = 0.2685      R-sq = 85.3%     R-sq(adj) = 81.4%
```

EXHIBIT 4.5.3
(Continued)

student	gpa	SAT math	SAT verb	HS math	HS engl	fits	residual	stdresid	nscores
1	1.97	321	247	2.30	2.63	1.7821	0.1879	0.7966	0.7420
2	2.74	718	436	3.80	3.57	3.1833	-0.4433	-1.9039	-1.8713
3	2.19	358	578	2.98	2.57	2.3945	-0.2045	-0.8595	-1.1269
4	2.60	403	447	3.58	2.21	2.4031	0.1969	0.8708	0.9172
5	2.98	640	563	3.38	3.48	3.0981	-0.1181	-0.4655	-0.4460
6	1.65	237	342	1.48	2.14	1.5340	0.1160	0.5137	0.5874
7	1.89	270	472	1.67	2.64	1.8429	0.0471	0.1998	0.4460
8	2.38	418	356	3.73	2.52	2.3749	0.0051	0.0226	0.0617
9	2.66	443	327	3.09	3.20	2.3271	0.3329	1.4335	1.4038
10	1.96	359	385	1.54	3.46	1.9600	-0.0000	-0.0000	-0.0617
11	3.14	669	664	3.21	3.37	3.2410	-0.1010	-0.4127	-0.3132
12	1.96	409	518	2.77	2.60	2.3848	-0.4248	-1.6851	-1.4038
13	2.20	582	364	1.47	2.90	2.3197	-0.1197	-0.5717	-0.5874
14	3.90	750	632	3.14	3.49	3.3610	0.5390	2.2510	1.8713
15	2.02	451	435	1.54	3.20	2.1848	-0.1648	-0.6919	-0.9172
16	3.61	645	704	3.50	3.74	3.3302	0.2798	1.2155	1.1269
17	3.07	791	341	3.20	2.93	3.0414	0.0286	0.1549	0.1859
18	2.63	521	483	3.59	3.32	2.7845	-0.1545	-0.6295	-0.7420
19	3.11	594	665	3.42	2.70	3.0726	0.0374	0.1645	0.3132
20	3.20	653	606	3.69	3.52	3.2403	-0.0403	-0.1622	-0.1859

4.6
Confidence Intervals

The practical importance of confidence intervals cannot be overemphasized, and you should recall the discussions of Section 1.6 in this regard. In this section we give formulas for computing confidence intervals for the unknown parameters in multiple linear regression.

Confidence Intervals for β_i, $\mu_Y(x_1, \ldots, x_k)$, $Y(x_1, \ldots, x_k)$, and $\sum_{i=0}^{k} a_i\beta_i = a^T\beta$

The general form for $1 - \alpha$ two-sided confidence intervals for

$$\beta_0, \beta_1, \ldots, \beta_k, \mu_Y(x_1, \ldots, x_k), Y(x_1, \ldots, x_k)$$

and linear functions

$$a^T\beta = \sum_{i=0}^{k} a_i\beta_i = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k$$

for a specified vector a is

$$\hat{\theta} - \text{table-value} \times SE(\hat{\theta}) \le \theta \le \hat{\theta} + \text{table-value} \times SE(\hat{\theta}) \qquad (4.6.1)$$

where θ refers to any of the quantities

$$\beta_0, \beta_1, \ldots, \beta_k, \mu_Y(x_1, \ldots, x_k), Y(x_1, \ldots, x_k), \text{ or } a^T\beta$$

and $\hat{\theta}$ refers to the corresponding estimate. The quantity to be used as the table-value is $t_{1-\alpha/2;df}$ obtained from a student's t-table (Table T-2 in Appendix T) with df = degrees of freedom $= n - k - 1 = n - (k+1)$, where $k+1$ is the number of parameters β_i in the regression function in (4.1.1). As usual, the quantity $SE(\hat{\theta})$ is the standard error of $\hat{\theta}$ and is an estimate of the *precision* of $\hat{\theta}$. These confidence intervals are valid under assumptions (A) or (B).

From the two-sided $1-\alpha$ confidence interval in (4.6.1) a $1-\alpha/2$ lower confidence bound or a $1-\alpha/2$ upper confidence bound can be obtained as explained in Section 1.6. (4.6.2)

Point estimates for $\beta = [\beta_0, \beta_1, \ldots, \beta_k]^T$, $\mu_Y(x_1, \ldots, x_k)$, $Y(x_1, \ldots, x_k)$, and $a^T\beta$ are given in (4.4.8), (4.4.9), (4.4.10), and (4.4.11), respectively. The formulas for standard errors require elements from the matrix C where C is defined by

$$C = (X^TX)^{-1} \qquad (4.6.3)$$

and the matrix X is defined in (4.4.7). We write $c_{i,j}$ for the (i,j) element of the matrix C. Formulas for standard errors of various estimated quantities of interest are listed in (4.6.4)–(4.6.8).

$$SE(\hat{\beta}_{i-1}) = \hat{\sigma}\sqrt{c_{i,i}} \quad \text{for } i = 1, \ldots, (k+1) \qquad (4.6.4)$$

$$SE(\hat{\mu}_Y(x_1, \ldots, x_k)) = \hat{\sigma}\sqrt{x^TCx} \qquad (4.6.5)$$

$$SE(\hat{Y}(x_1, \ldots, x_k)) = \hat{\sigma}\sqrt{1 + (x^TCx)} \qquad (4.6.6)$$

$$= \sqrt{\hat{\sigma}^2 + [SE(\hat{\mu}_Y(x_1, \ldots, x_k))]^2} \qquad (4.6.7)$$

$$SE(a^T\hat{\beta}) = \hat{\sigma}\sqrt{a^TCa} \qquad (4.6.8)$$

The vector x in (4.6.5) and (4.6.6) is given by

$$x = [1, x_1, x_2, \ldots, x_k]^T \qquad (4.6.9)$$

and the vector a in (4.6.8) is given by

$$a = [a_0, a_1, \ldots, a_k]^T \qquad (4.6.10)$$

where the linear combination $a^T\beta$ of interest is $a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k$. For instance, in Example 4.4.2 if we wish to compute a confidence interval for $\mu_Y(720, 570, 3.2, 2.7)$, then we would compute $SE(\hat{\mu}_Y(720, 570, 3.2, 2.7))$ using (4.6.5) with $x = [1, 720, 570, 3.2, 2.7]^T$.

Remarks

1 Note that $SE(\hat{\beta}_i)$ for $i = 0, \ldots, k$ and $SE(\hat{\mu}_Y(x_1, \ldots, x_k))$ can be obtained from (4.6.8) by using appropriate vectors a. For instance, $SE(\hat{\beta}_0)$ is obtained from (4.6.8) by taking $a = [1, 0, \ldots, 0]^T$, $SE(\hat{\beta}_1)$ is obtained by taking $a = [0, 1, 0, \ldots, 0]^T$, $SE(\mu_Y(x_1, \ldots, x_k))$ is obtained by taking $a = [1, x_1, \ldots, x_k]^T$, etc.

2 In any particular application, it is important for the user to determine whether a confidence interval for $\mu_Y(x_1, \ldots, x_k)$ or a confidence interval for $Y(x_1, \ldots, x_k)$ is required. Some authors call the confidence interval for $Y(x_1, \ldots, x_k)$ a **prediction interval** because the term confidence interval is traditionally reserved for *parameters*, but $Y(x_1, \ldots, x_k)$ is a *random variable*. (It is the Y value of a randomly chosen item with $X_1 = x_1, \ldots, X_k = x_k$.)

3 Note that even though

$$\hat{\mu}_Y(x_1, \ldots, x_k) = \hat{Y}(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \qquad \textbf{(4.6.11)}$$

their standard errors are not the same. This is so because there is greater uncertainty in predicting $Y(x_1, \ldots, x_k)$, a single randomly chosen value from a subpopulation with $X_1 = x_1, \ldots, X_k = x_k$ than in estimating the mean $\mu_Y(x_1, \ldots, x_k)$ of the entire subpopulation.

Confidence Interval for σ

When regression assumptions (A) or (B) hold, a two-sided $1 - \alpha$ confidence interval for σ is given by

$$\sqrt{\frac{(n-k-1)\hat{\sigma}^2}{\chi^2_{1-\alpha/2:n-k-1}}} \le \sigma \le \sqrt{\frac{(n-k-1)\hat{\sigma}^2}{\chi^2_{\alpha/2:n-k-1}}} \qquad \textbf{(4.6.12)}$$

where $\chi^2_{\alpha/2:n-k-1}$ and $\chi^2_{1-\alpha/2:n-k-1}$ may be obtained from Table T-3 in Appendix T. It is worth noting that the quantity $(n-k-1)\hat{\sigma}^2$ in (4.6.12) is in fact equal to $SSE = SSE(X_1, \ldots, X_k)$ (refer to (4.4.17)), and so we can rewrite (4.6.12) in terms of SSE. Thus the $(1-\alpha)$ two-sided confidence interval for σ in (4.6.12) may be reexpressed in the form

$$\sqrt{\frac{SSE}{\chi^2_{1-\alpha/2:n-k-1}}} \le \sigma \le \sqrt{\frac{SSE}{\chi^2_{\alpha/2:n-k-1}}} \qquad \textbf{(4.6.13)}$$

The formula for a confidence interval for σ in (4.6.12) is not of the general form given in (4.6.1). In fact, the general form for a two-sided $(1-\alpha)$ confidence

interval for a parameter that is a standard deviation (when a valid confidence interval is available) is

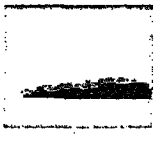
$$\sqrt{\frac{(df)\ (\text{estimated standard deviation})^2}{\chi^2_{1-\alpha/2:df}}} \le \sigma \le \sqrt{\frac{(df)\ (\text{estimated standard deviation})^2}{\chi^2_{\alpha/2:df}}} \tag{4.6.14}$$

where df represents the number of degrees of freedom associated with the estimate $\hat{\sigma}$. Note that the number of degrees of freedom associated with $\hat{\sigma}$ is $n-(k+1)$ when the regression function of Y on $X_1, \ldots, X_k$ is the one given in (4.1.1); i.e., it has $(k+1)$ regression coefficients $\beta_0, \beta_1, \ldots, \beta_k$. It is also worth observing that while the confidence intervals for the quantities, $\beta_0, \beta_1, \ldots, \beta_k$, $\mu_Y(x_1, \ldots, x_k)$, $Y(x_1, \ldots, x_k)$, and $a^T\beta$ are symmetric about the corresponding point estimates, this is not so in the case of σ. However, the confidence interval for σ is equal-tailed and contains the point estimate $\hat{\sigma}$, and hence it gives us some indication of how close $\hat{\sigma}$ might be to the population value σ.

One-Sided Confidence Bounds

In this section, most of the discussion has been about two-sided confidence intervals for a parameter of interest. However, in some applications an investigator may be interested in only the lower bound or the upper bound for a parameter in a decision-making situation, and hence one-sided confidence bounds are useful. As discussed in Section 1.6, we can obtain one-sided confidence bounds with confidence coefficient $1-\alpha/2$ by first constructing a two-sided confidence interval with confidence coefficient $1-\alpha$ and reading off either the lower or the upper endpoint as appropriate. This is valid for all of the quantities, $\beta_0, \beta_1, \ldots, \beta_k$, $\mu_Y(x_1, \ldots, x_k)$, $Y(x_1, \ldots, x_k)$, and $a^T\beta$, as well as σ, because all of these confidence intervals discussed so far are *equal-tailed*.

We illustrate the use of the preceeding formulas in the following task.



Task 4.6.1

Here we consider some practical questions that may arise in the context of Example 4.4.2, where a director of admissions is studying how $X_1 = \text{SATmath}$, $X_2 = \text{SATverbal}$, $X_3 = \text{HSmath}$, and $X_4 = \text{HSenglish}$ can be used to predict $Y = \text{GPA}$ at the end of the first year after admission to a certain university. The population

regression function is assumed to be of the form

$$\mu_Y(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

and population assumptions (B) are presumed valid. A sample of 20 students was selected using simple random sampling from all students who were admitted over the past four years and completed the first year. The data are given in Table 4.4.3 and are also in the file **gpa.dat** on the data disk.

In Exhibit 4.6.1 we give a MINITAB output from a regression analysis of GPA on SATmath, SATverbal, HSmath, and HSenglish along with the $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ matrix. The matrix $\boldsymbol{C}$ is needed to obtain standard errors and hence confidence intervals for $\boldsymbol{a}^T\boldsymbol{\beta}$. A similar output can be obtained from most statistical computing packages.

EXHIBIT 4.6.1

MINITAB Output for Task 4.6.1

```
The regression equation is
GPA = 0.162 + 0.00201 SATmath + 0.00125 SATverb +
      0.189 HSmath + 0.088 HSengl                                    (4.6.15)

Predictor       Coef       Stdev    t-ratio       p
Constant      0.1615      0.4375       0.37   0.717                 (4.6.16)
SATmath    0.0020102   0.0005844       3.44   0.004                 (4.6.17)
SATverb    0.0012522   0.0005515       2.27   0.038                 (4.6.18)
HSmath       0.18944     0.09187       2.06   0.057                 (4.6.19)
HSengl        0.0876      0.1765       0.50   0.627                 (4.6.20)

s = 0.2685      R-sq = 85.3%     R-sq(adj) = 81.4%                   (4.6.21)
```

The C matrix is

$$\boldsymbol{C} = \begin{bmatrix} 2.6551249569 & 0.0013784337 & -0.0003619435 & -0.2006841873 & -0.8521280877 \\ 0.0013784337 & 0.0000047375 & -0.0000004016 & -0.0003559377 & -0.0008620172 \\ -0.0003619435 & -0.0000004016 & 0.0000042188 & -0.0001953018 & -0.0002966849 \\ -0.2006841873 & -0.0003559377 & -0.0001953018 & 0.1170561208 & 0.0472194123 \\ -0.8521280877 & -0.0008620172 & -0.0002966849 & 0.0472194123 & 0.4320523096 \end{bmatrix} \quad (4.6.22)$$

In the computer output in Exhibit 4.6.1, the quantities in (4.6.16)–(4.6.20) under the heading `coef` are the $\hat{\beta}_i$, the point estimates of the corresponding β's. Also in (4.6.16)–(4.6.20) under `Stdev` are the $SE(\hat{\beta}_i)$, the standard errors of the $\hat{\beta}_i$'s. The value of $\hat{\sigma}$ is the quantity labeled `s` in (4.6.21), and the matrix $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is given in (4.6.22).

Regression calculations are notoriously prone to rounding errors. To minimize this problem, it is advisable to carry as many significant digits as possible for all intermediate calculations. For this reason, we have used ten-decimal accuracy in the matrix $\boldsymbol{C}$ given in (4.6.22). Some computer programs will round results to four or five decimals unless you specifically ask for more. Final results may, of course, be appropriately rounded.

We now consider several practical questions concerning the GPA data.

1 Suppose that the current requirement to qualify as a candidate for financial assistance at this university is that applicants must have a grade point average of 1.5 or better in high school mathematics courses. The director of admissions is considering a recommendation by a faculty committee to increase this requirement to 2.5 or better, so it is of interest to know what the difference will be between the *average* first-year GPA of students who received a grade point average of 2.5 in high school mathematics courses and that of students who received a grade point average of 1.5 in high school mathematics courses, other variables being equal. The director of admissions would like to have an estimate of this difference along with an indication of how good this estimate is.

Let the values of X_1, X_2, and X_4 for the two groups of students being compared, be x_1, x_2, and x_4, respectively. Then the required difference is in fact equal to

$$\begin{aligned}&\mu_Y(x_1, x_2, 2.5, x_4) - \mu_Y(x_1, x_2, 1.5, x_4)\\ &= [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(2.5) + \beta_4 x_4] - [\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(1.5) + \beta_4 x_4]\\ &= \beta_3[2.5 - 1.5] = (1)\beta_3 = \beta_3.\end{aligned}$$

From (4.6.19) above under `coef` *we obtain $\hat{\beta}_3 = 0.18944$. A confidence interval for β_3 tells us how well we can estimate β_3 using the sample data. For illustration, we compute a 90% two-sided confidence interval for β_3 using the formula in (4.6.1). Here $1 - \alpha = 0.90$, so we have $\alpha = 0.10$. Also $n - k - 1 = 20 - 4 - 1 = 15$. From Table T-2 in Appendix T we obtain $t_{1-\alpha/2:n-k-1} = t_{.95:15} = 1.753$. From (4.6.21) we get $\hat{\sigma} = 0.2685$. To compute the standard error of $\hat{\beta}_3$ we use the formula in (4.6.4) with $i = 4$. For this we need $c_{4,4}$, the (4, 4) element of the matrix $\boldsymbol{C}$. This is equal to 0.11706 (rounded to five decimals) and is obtained from (4.6.22). So $SE(\hat{\beta}_3) = 0.2685\sqrt{0.11706}$, which equals 0.09186. We can also obtain $SE(\hat{\beta}_3)$ from (4.6.19) under* `Stdev` *in the MINITAB output in Exhibit 4.6.1. Putting together the various quantities, we get*

$$C[0.18944 - (1.753)(0.09186) \le \beta_3 \le 0.18944 + (1.753)(0.09186)] = 0.90$$

i.e.,

$$C[0.0284 \le \beta_3 \le 0.3505] = 0.90$$

This means that if X_1, X_2, and X_4 remain fixed, we can state with 90% confidence that an increase in HSmath of 1 grade point unit (from 1.5 units to 2.5 units) is associated with an increase in the average first-year GPA of at least 0.0284 grade point unit and not more than 0.3504 grade point unit. The director of admissions has to decide whether the information provided by this confidence statement is adequate for the purpose at hand, or if the interval is too wide for decision-making purposes; in the latter case a larger sample would be necessary to obtain the required information.

2 One of the applicants has furnished the following information in support of her application for financial aid. She has a score of 594 in SATmath, a score of 665 in SATverbal, a grade point average of 3.42 in high school mathematics courses, and a grade point average of 2.70 in high school English courses. The financial aid committee will grant her financial aid if there is evidence that she will obtain a GPA of 2.5 or higher at the end of the first year. The committee will use a 95% lower confidence bound to help make this decision.

So we are interested in a lower bound for $Y(594, 665, 3.42, 2.70)$. We compute a 90% two-sided confidence interval for $Y(594, 665, 3.42, 2.70)$, the lower endpoint of which gives us the required 95% lower confidence bound. The values of $\hat{\beta}_i$ can be obtained from (4.6.16)–(4.6.20), and using these we obtain $\hat{Y}(594, 665, 3.42, 2.70) = 3.0726$. Next we need the standard error of $\hat{Y}(594, 665, 3.42, 2.70)$. This is computed using the formula in (4.6.6) or (4.6.7). The vector $\boldsymbol{x}$ is equal to $[1, 594, 665, 3.42, 2.70]^T$, and the matrix $\boldsymbol{C}$ is in (4.6.22). So we get $\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} = 0.2831$. From (4.6.21) we obtain $\hat{\sigma} = 0.2685$. So $SE(\hat{Y}(x_1, x_2, x_3, x_4)) = (0.2685)\sqrt{1 + 0.2831} = 0.3041$. From Table T-2 we obtain $t_{1-\alpha/2:n-k-1} = t_{0.95:15} = 1.753$. Using (4.6.1) we get

$$C[3.0726 - (1.753)(0.3041) \leq Y(594, 665, 3.42, 2.7) \leq 3.0726 + (1.753)(0.3041)] = 0.90$$

i.e.,

$$C[2.54 \leq Y(594, 665, 3.42, 2.7) \leq 3.61] = 0.90$$

In particular we have 95% confidence that this student will have a GPA no smaller than 2.54 at the end of the first year. The financial aid committee will use this information to help decide whether or not to grant assistance to this student.

3 The director of admissions wants to determine how good the population regression function $\mu_Y(x_1, x_2, x_3, x_4)$ is for predicting Y = first-year GPA (it is assumed that the model $\mu_Y(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$ holds). For this reason the director wants a point and an interval estimate of the subpopulation standard deviation σ.

The quantity σ is a measure of how good the four predictor variables X_1, X_2, X_3, X_4 together are for predicting the value of Y. The point estimate of σ is given in (4.6.21) and is equal to 0.2685. We now compute a 90% two-sided confidence interval for σ using (4.6.12). We have $n - k - 1 = 15$,

$1-\alpha/2=.95$, $\alpha/2=0.05$, *and, from Table T-3,* $\chi^2_{0.05:15}=7.261$, *and* $\chi^2_{0.95:15}=24.996$. *Hence the confidence statement is*

$$C\left[\sqrt{\frac{15(0.2685)^2}{24.996}}\le\sigma\le\sqrt{\frac{15(0.2685)^2}{7.261}}\right]=0.90$$

i.e.,

$$C[0.208\le\sigma\le 0.386]=0.90$$

So we have 90% confidence that the subpopulation standard deviation σ is between 0.208 grade point unit and 0.386 grade point unit.

We can directly read $\hat{\sigma}$, $\hat{\beta}_i$, and the standard errors of $\hat{\beta}_i$ from the regression output of most statistical computer packages. These are the ingredients required to compute confidence intervals for each β_i. However, the point estimate and the standard error of $a^T\beta$ are not given directly in all statistical packages but must be computed by using the C matrix in (4.6.3). In Section 4.6 of the laboratory manuals we discuss how confidence intervals can be obtained using computer commands.

Conversation 4.6.1

Investigator: I want to ask you some questions about confidence intervals and in particular about confidence coefficients. Can you discuss these with me now?

Statistician: Certainly.

Investigator: One of the scientists with whom I work says that 90%, 95%, and 99% confidence coefficients are part of statistical theory and that we should always use $1-\alpha=$ 0.90, 0.95, or 0.99. Is that correct?

Statistician: No, it is not correct. You can use any value between zero and one for the confidence coefficient $1-\alpha$.

Investigator: That's what I thought. But most statistics books almost always use a confidence coefficient of 90%, 95%, or 99%. Why is that?

Statistician: I guess this is partly due to tradition or habit. In the 1930s, when modern statistical inference was beginning to include confidence intervals, some books used confidence intervals with 90%, 95%, or 99% confidence coefficients, and it appears that other books continued to use these values. As people read and studied these and subsequent books, these values became quite standard although there is nothing sacred about these or any other values for confidence coefficients.

Investigator: What does it really mean to have a confidence of, say, 95% that an interval contains a parameter θ?

Statistician: I will try to explain it this way. Suppose that you are handed a box containing 95 white balls and 5 red balls, and the balls are indistinguishable except for color. You are going to select a ball at random from this box, and we say the probability is 95% that the ball you choose will be white, or we have 95% confidence that the ball drawn will be white. Now, theory and experience tell us that if you draw a ball repeatedly (with replacement) from the box many times, say 100,000 times, the percentage of white balls drawn will be very close to 95%.

Investigator: But in practice I'll draw a ball only once.

Statistician: That's correct, and it isn't completely clear what the 95% probability means for this one draw. But suppose you are handed two boxes. Box 1 contains 99 white balls and 1 red ball, and Box 2 contains 50 white balls and 50 red balls. You can randomly choose one ball from either box and, if the ball is white, you get a valuable prize. From which box would you choose this one ball?

Investigator: I'd choose a ball from Box 1—the box that has 99 white balls and one red ball.

Statistician: Why would you select Box 1?

Investigator: Because there are more white balls in Box 1 than in Box 2.

Statistician: Well, let's change the problem slightly. Suppose Box 1 still contains 99 white balls and 1 red ball, but Box 2 now contains 1,000 white balls and 1,000 red balls. Box 2 now has more white balls in it than Box 1 has. Which box would you choose a ball from now?

Investigator: I'd still choose Box 1.

Statistician: Why?

Investigator: I just *feel* that if I select a ball from Box 1, I have a better chance of getting a white ball than if I select a ball from Box 2. In fact, if I draw a ball from Box 1 100,000 times (with replacement), and if you draw a ball from Box 2 100,000 times (with replacement), I'll get a white ball about 95% of the time, but you'll get a white ball only about 50% of the time.

Statistician: That's right, but you are going to draw only one ball. From which box will you draw it?

Investigator: Box 1.

Statistician: Why?

Investigator: Since the ratio of the number of white balls to the number of red balls in Box 1 is much greater than their ratio in Box 2, I feel that I'm more apt to get a white ball if I draw it from Box 1 rather than Box 2.

Statistician: More apt? What does that mean?

Investigator: I have a better chance, or a higher probability. In fact, I'll have 95% probability of getting a white ball if I draw it from Box 1, but I'll have only a 50% probability of getting a white ball if I draw it from Box 2.

Statistician: I guess you've answered your original question: "What does 95% confidence mean?" When an investigator selects a simple random sample of size n from a population of size N and computes a confidence interval for a parameter θ, he/she is using one of $H = \binom{N}{n}$ possible intervals. We can view this process as follows. Suppose that a box contains H balls, and each ball has written on it a confidence interval for θ. Now it is known (from theory) that for 95% of these balls, the confidence intervals written on them actually cover the unknown parameter θ. Color these balls white. Color the balls red if the confidence intervals written on them do not cover θ. Thus selecting a sample of size n and computing a confidence interval for θ with confidence coefficient 95% is equivalent to choosing a ball at random and noting its color. Since 95% of the balls are white, we have 95% confidence that the chosen ball will be white and the confidence interval will cover θ.

Investigator: I think I see your point, but since scientists want to be certain that their confidence interval covers the unknown parameter θ, why should they use a 95% or even a 99% confidence interval? Why shouldn't I advise them to use a 99.9999% confidence interval? Then they can be practically certain that their interval includes θ.

Statistician: Because, in general, the larger the confidence coefficient is, the wider the confidence interval will be, and so to reach a decision a scientist must also consider the width of the interval. Let me give you a simple illustration. Suppose a company is considering moving to city A, and it needs to know μ, the average annual income of wage earners in the city. Suppose a sample is selected and a 99.9999% confidence interval for μ, is computed as

$$\$6.59 \leq \mu \leq \$750{,}000$$

The investigator is quite certain that the confidence interval is correct, but it is so wide that it is useless for making a decision. Now suppose that a 90% confidence interval for μ is

$$\$21{,}000 \leq \mu \leq \$23{,}981$$

(Of course these numbers were chosen to make our point in a dramatic manner.) The investigator is less certain that μ is in this interval, but the interval is useful. So you can see why a very large confidence coefficient may not be the thing to use.

Investigator: Are you saying that a confidence interval with confidence coefficient of either 90%, 95%, or 99% will always have a desirable width? Is that the reason these values are generally used?

Statistician: No, I'm not saying that. That isn't the case. What I am saying is that the larger the confidence coefficient is, the wider the confidence interval will be, so using confidence coefficients very close to 1 may result in very wide confidence intervals that will be of no help in making decisions.

Investigator: So what should I do? In one situation the confidence interval is practically certain (99.9999% confidence) to cover θ, but the confidence interval is so wide that it is not useful. In the other situation the result is not so certain (90% confidence), but if it is not too wide it may be useful.

Statistician: That's correct, but in many situations we can have a desirable confidence coefficient (say 95%, 99%, etc.) and still have the width such that the result is useful.

Investigator: How do I do this?

Statistician: By designing the study carefully before any sample values are selected and then choosing the **sample size** judiciously! In general, in statistical inference problems for a fixed confidence coefficient $1 - \alpha$, one can reduce the width of the interval by increasing n, the size of the sample. So in many problems if you choose the proper value for the sample size n, you may be able to obtain a confidence interval of desired width and desired confidence coefficient $1 - \alpha$. You must remember, however, that the confidence coefficient you specify may be 0.95, but due to the fact that the assumptions are generally not exactly satisfied, the actual confidence coefficient will typically differ from the specified nominal value. It may be 0.94 or 0.96, or even 0.92 or 0.98. But if you choose a confidence coefficient equal to $1 - \alpha$, and if the assumptions are approximately satisfied, then the actual confidence coefficient will be close to $1 - \alpha$.

Investigator: That sounds good to me, but in almost all of the problems that I've been associated with, the data have already been collected. Thus the sample size n has already been fixed.

Statistician: Then you may be much more restricted in what conclusions you can draw from the data. That is why it's important to design the study carefully before samples are selected. Perhaps we can discuss this later.

Investigator: I'd like to do that. But let me ask you one final question. What value should I recommend that the scientists use for the confidence coefficient $1 - \alpha$?

Statistician: I can't answer that question definitively. You might explain to the scientists you work with about confidence coefficients as we have discussed them using balls in boxes and ask them to think about the risk they are willing to take in obtaining an incorrect conclusion from their experiment. Then let them make their own decision about what confidence coefficient they want to use.

Investigator: I'll try that. Can I come to see you again after I discuss this with them?

Statistician: Certainly.

Conversation 4.6.2

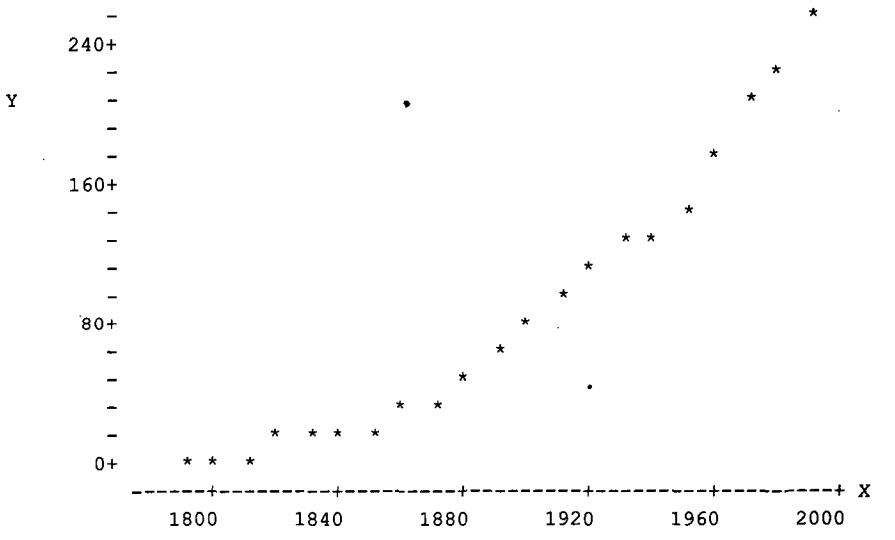
Investigator: Good morning. I know you didn't expect to see me again so soon! Do you have time to talk with me now? This won't take very long.

Statistician: Certainly. How can I help?

Investigator: We're working on a problem where we must make a census projection of the resident population of the United States in the year 2010. This will help my company determine whether or not it should start thinking about plans for expanding its operations over the next several years. I have obtained the following U.S. Bureau of Census data of the resident population for every 10 years from 1790 through 1990.

Resident Population (in millions) Y	Year X
3.929	1790
5.308	1800
7.240	1810
9.638	1820
12.866	1830
17.069	1840
23.192	1850
31.443	1860
39.818	1870
50.156	1880
62.948	1890
75.995	1900
91.972	1910
105.711	1920
122.775	1930
131.669	1940
150.697	1950
179.323	1960
203.302	1970
226.546	1980
248.710	1990

I plotted Y against X for the data, and the result is



I used least squares and fitted a second-degree polynomial

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

to the data. The fitted least squares polynomial is denoted by $\hat{f}(x)$ where

$$\hat{f}(x) = 21010.7 - 23.3819x + 0.0065068x^2$$

I substituted $x = 2{,}010$ and the result is $\hat{f}(2010) = 301.201$. So based on these calculations the resident population in the United States in the year 2010 is estimated to be 301.201 million. Is this a valid estimate? Can I use regression techniques to obtain a valid confidence interval for $f(2010)$?

Statistician: Your estimate of 301.2 million may be a *reasonable* estimate, but you don't have a regression setup that satisfies assumptions (A) or (B). You also do not have a *random* sample from a well-defined population. What you have is a set of 21 pairs of points, and I don't see how you can define a target or study population from which you collected a random sample. The curve that you computed is a way to summarize the pattern exhibited by these 21 data points. I'm not saying that the data do not contain a great deal of information—they do. But they don't seem to fit into the framework of regression assumptions.

Investigator: We want to publish these results along with others in a professional journal, and if we don't include the results of statistical tests or confidence intervals, it won't be accepted for publication. What would you advise?

Statistician: As I stated earlier, your data contain a great deal of information about the U.S. resident population for the past 21 decades (in fact, you have the entire set of available

data for these decades), and you don't need confidence intervals, tests, etc. to make useful statements for these decades. However, you want to extrapolate to the year 2010, and extrapolation is always risky unless you can define a population that includes the extrapolated points. Perhaps you can look at your model as a *process* rather than a population and use a *stochastic process model*. You might want to talk to a professional statistician who works with such models.

Investigator: That's a possibility. I'll look into it. Thanks. I'll see you again next week.

Problems 4.6

The following questions refer to the problem discussed in Example 4.5.1 and the data in Table 4.5.1, which are also in the file **electric.dat** on the data disk. Recall that

Y = (total) electric bill (in dollars) for the past year

X_1 = monthly income for the household (in dollars)

X_2 = number of persons in the household

X_3 = living area (in square feet) of the house or apartment

Suppose that assumptions (A) are met and the data are obtained by simple random sampling. In particular, the regression function of Y on X_1, X_2, and X_3 is of the form

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \qquad (4.6.23)$$

The utility company is interested in predicting the electricity usage patterns by various households for next year so the target population is a *future* population. However, the study population is assumed to be very similar to the target population, and so the sample data from the study population may be used to make inferences about the target population. A SAS output from a regression analysis of these data appears in Exhibit 4.6.2.

Note that the C matrix is obtained from the first four rows and columns of the matrix given in (4.6.24). The rows and columns of the C matrix are labeled by the names of the predictors corresponding to $\beta_0, \beta_1, \beta_2$, and β_3, respectively. Thus the (2, 2)-element of the C matrix is 0.0000010099689, etc. The first four numbers in the last row (and column) of the matrix in (4.6.24), labeled BILL, are the regression coefficients, and the final number in the last row (and column) is $SSE(X_1, X_2, X_3, X_4)$. Use this information to solve the following problems.

4.6.1 Are valid point estimates available for any of the following parameters?

$$\mu_{X_1}, \mu_{X_2}, \mu_{X_3}, \mu_Y, \sigma_{X_1}, \sigma_{X_2}, \sigma_{X_3}, \sigma_Y$$

Explain.

4.6.2 In Problem 4.6.1, find the point estimates of the parameters for which valid estimates are available.

EXHIBIT 4.6.2
SAS Output for Problem 4.6.1

The SAS System 0:00 Saturday, Jan 1, 1994

Model: MODEL1

X'X Inverse, Parameter Estimates, and SSE

	INTERCEP	INCOME	PERSONS	AREA	BILL
INTERCEP	2.153683547	-0.001377673	-0.25570104	0.0021517892	-358.4415686
INCOME	-0.001377673	1.0099689E-6	0.000175464	-1.655901E-6	0.075136905
PERSONS	-0.25570104	0.000175464	0.046002015	-0.00029998	55.087632718
AREA	0.0021517892	-1.655901E-6	-0.00029998	2.7878446E-6	0.2811036938
BILL	-358.4415686	0.075136905	55.087632718	0.2811036938	550163.42009

(4.6.24)

Dependent Variable: BILL

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	3151504.8152	1050501.6051	57.283	0.0001
Error	30	550163.42009	18338.78067		
C Total	33	3701668.2353			

Root MSE	135.42075	R-square	0.8514
Dep Mean	619.41176	Adj R-sq	0.8365
C.V.	21.86280		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-358.441569	198.73583019	-1.804	0.0813
INCOME	1	0.075137	0.13609408	0.552	0.5850
PERSONS	1	55.087633	29.04515215	1.897	0.0675
AREA	1	0.281104	0.22610987	1.243	0.2234

4.6.3 Exhibit the values of

$$\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{SE}(\hat{\beta}_0), SE(\hat{\beta}_1), SE(\hat{\beta}_2), SE(\hat{\beta}_3), \hat{\mu}_Y(x_1, x_2, x_3)$$

4.6.4 How well can Y, the annual electric bill for a randomly chosen household, be predicted using X_1, X_2, and X_3?

4.6.5 How well can Y, the annual electric bill for a randomly chosen household, be predicted if no predictor factors are used?

4.6.6 What is the difference between the average annual electric bill of households consisting of five individuals and households with four individuals if they have the same monthly income and the same living area? Obtain a point estimate and a 95% upper confidence bound for this quantity.

4.6.7 What is the difference between the average annual electric bill of households with a monthly income of $4,000.00 and households with a monthly income of $3,000.00 if they have the same number of individuals in the household and the same living area? Obtain a point estimate and a 95% upper confidence bound for this quantity.

4.6.8 What is the difference between the average annual electric bill of households with a living area of 2,000 square feet and households with a living area of 1,500 square feet if they have the same monthly income and the same number of individuals in the household? Obtain a point estimate and a 95% upper confidence bound for this quantity.

4.6.9 The monthly income of a *particular* household is $3,200. There are six individuals in the household, and the living area equals 2,800 square feet. Predict the annual monthly electric bill for this household.

4.6.10 In Problem 4.6.9, compute a 90% upper confidence bound for the annual electric bill for this household.

4.6.11 Repeat Problems 4.6.9 and 4.6.10 for the *average* annual electric bill of all households with this monthly income, number of individuals in the household, and living area.

4.6.12 What is the difference between the average annual electric bill for households consisting of seven individuals with a living area of 3,500 square feet and a monthly income of $5,400.00 and that for households consisting of four individuals with a living area of 2,400 square feet and a monthly income of $5,000.00? Estimate this difference and also compute a 95% two-sided confidence interval for this quantity.

4.6.13 The model in (4.6.23) would be considered adequate for predicting Y if the standard deviation of the prediction errors—i.e., σ—is less than $50.00. If this is not so, then the investigator will look for additional explanatory variables to include in the regression model. To assist the investigator in making a decision in this connection, compute a two-sided 95% confidence interval for σ. Do you believe that the investigator needs to look for additional explanatory variables? Or do you believe that the model in (4.6.23) is adequate? Explain.

4.7 Tests

It has been stated several times that *statistical tests alone should never be used in situations where appropriate confidence intervals for parameters of interest are available.* Refer to Sections 1.6 and 3.6 for a more thorough discussion in this regard. However, as we said earlier, statistical tests are quite popular among many practitioners. We feel that you should be familiar with some of the commonly used testing procedures because of their widespread use so that you can properly interpret published results of investigations. Formulas and procedures for some statistical tests that pertain to the multiple linear regression model in (4.1.1) are summarized in Boxes 4.7.1 and 4.7.2. Here θ denotes any one of the quantities

$$\beta_0, \beta_1, \ldots, \beta_k, \mu_Y(x_1, \ldots, x_k)$$

or

$$a^T\beta = a_0\beta_0 + a_1\beta_1 + \cdots + a_k\beta_k$$

The quantity $\hat{\theta}$ denotes the point estimate of θ.

BOX 4.7.1 Statistical Tests for $\beta_i, i = 0, \ldots, k, \mu_Y(x_1, \ldots, x_k)$ or $a^T\beta$ (θ Stands for Any One of These)

Let q be a specified number (the investigator specifies q). Compute the statistic

$$t_C = \frac{\hat{\theta} - q}{SE(\hat{\theta})}$$

1. For testing NH: $\theta = q$ versus AH: $\theta \neq q$, the P-value is the value of α such that $|t_C| = t_{1-\alpha/2:n-k-1}$.
2. For testing NH: $\theta \leq q$ versus AH: $\theta > q$, the P-value is the value of α such that $t_C = t_{1-\alpha:n-k-1}$.
3. For testing NH: $\theta \geq q$ versus AH: $\theta < q$, the P-value is the value of α such that $-t_C = t_{1-\alpha:n-k-1}$.

BOX 4.7.2 Statistical Tests for σ

Let q be a positive number specified by the investigator. Compute the statistic

$$\chi^2_C = \frac{(n-k-1)\hat{\sigma}^2}{q^2} = \frac{SSE}{q^2}$$

1. For testing NH: $\sigma = q$ versus AH: $\sigma \neq q$, the P-value is equal to α where α is a number between 0 and 1 and satisfies

$$\chi^2_C = \chi^2_{\alpha/2:n-k-1} \quad \text{or} \quad \chi^2_C = \chi^2_{1-\alpha/2:n-k-1}$$

(only one of these two equalities can be satisfied unless $\alpha = 1$).

2 For testing NH: $\sigma \leq q$ versus AH: $\sigma > q$, the P-value is the value of α such that $\chi_C^2 = \chi_{1-\alpha:n-k-1}^2$.

3 For testing NH: $\sigma \geq q$ versus AH: $\sigma < q$, the P-value is the value of α such that $\chi_C^2 = \chi_{\alpha:n-k-1}^2$.

We illustrate the use of these procedures with Example 4.7.1.

EXAMPLE 4.7.1

(a) Consider the GPA problem discussed in Task 4.6.1, and suppose the financial aid committee is interested in knowing whether the average GPA at the end of the first year of applicants with SATmath = 594, SATverbal = 665, HSmath = 3.42, and HSenglish = 2.70 would equal or exceed 2.5. We might consider the following pair of hypotheses:

$$\text{NH}: \mu_Y(594, 665, 3.42, 2.70) \leq 2.5$$

versus

$$\text{AH}: \mu_Y(594, 665, 3.42, 2.70) > 2.5$$

Here we have $\theta = \mu_Y(594, 665, 3.42, 2.70)$ and $q = 2.5$. The appropriate test procedure is in part 2 in Box 4.7.1. From (4.6.16)–(4.6.20) we know that

$$\hat{\mu}_Y(x_1, x_2, x_3, x_4) = 0.1615 + 0.0020102x_1 + 0.0012522x_2 + 0.18944x_3 + 0.0876x_4$$

so the value of $\hat{\mu}_Y(594, 665, 3.42, 2.70)$ is 3.0726. To calculate $SE(\hat{\mu}_Y(594, 665, 3.42, 2.70))$ we use the formula in (4.6.5). From (4.6.21) we have $\hat{\sigma} = 0.2685$. Also from (4.6.22) we have the matrix $\boldsymbol{C}$, and using it we get $\boldsymbol{x}^T\boldsymbol{C}\boldsymbol{x} = 0.2831$ where $\boldsymbol{x}^T = [1, 594, 665, 3.42, 2.70]$. So $SE(\hat{\mu}_Y(594, 665, 3.42, 2.70)) = 0.2685\sqrt{0.2831} = 0.1429$. Also, $n - k - 1 = 20 - 4 - 1 = 15$. Hence

$$t_C = \frac{3.0726 - 2.5}{0.1429} = 4.007$$

The value of α for which $t_C = 4.007 = t_T = t_{1-\alpha:15}$ is less than 0.005 by Table T-2 in Appendix T, so the P-value is less than 0.005. So if NH were indeed true, then the probability of obtaining a value for t_C as large as, or larger than, 4.007 is less than 0.005. Since this probability is so small, the committee will very likely conclude that the average first-year GPA of applicants with SATmath = 594, SATverbal = 665, HSmath = 3.42, and HSenglish = 2.70, is greater than 2.5 (i.e., reject NH). A confidence interval for $\theta = \mu_Y(594, 665, 3.42, 2.70)$ would provide the committee with additional information.

(b) Test procedures given in Box 4.7.2 can be used to help decide whether σ is equal to, greater than or equal to, or less than or equal to a specified value q. To illustrate these procedures, we consider the following problem.

Suppose the director of admissions wants to know how well he can predict the first-year GPAs of applicants using X_1 = SATmath, X_2 = SATverbal, X_3 = HSmath, and X_4 = HSenglish. The prediction would be considered *adequate for this problem*

if σ is less than 0.2. To help determine whether σ is less than 0.2, the director might consider the following statistical test.

$$\text{NH: } \sigma \geq 0.2 \qquad \text{versus} \qquad \text{AH: } \sigma < 0.2 \tag{4.7.1}$$

If NH is rejected, then the predictors X_1, X_2, X_3, and X_4 together will be considered to be adequate for predicting Y.

Using the procedure given in part 3 of Box 4.7.2 with $q = 0.2$, we obtain

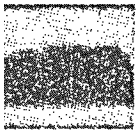
$$\chi_C^2 = \frac{(n-k-1)\hat{\sigma}^2}{(0.2)^2} = \frac{SSE}{(0.2)^2} = \frac{1.0815}{0.04} = 27.04$$

From Table T-3 in Appendix T we find that $\chi_C^2 = 27.04$ is between $\chi^2_{0.95;15} = 24.996$ and $\chi^2_{0.975,15} = 27.488$ so the value of α for which χ_C^2 is equal to $\chi^2_{\alpha;15}$ is between 0.95 and 0.975. Hence the P-value for the preceding hypothesis test is between 0.95 and 0.975. So NH will not be rejected at any of the usual α levels. The director of admissions would perhaps conclude that the predictors X_1, X_2, X_3, X_4 together may not be adequate for predicting Y. ■

Most computer programs for regression routinely output t_C values for testing whether or not the various β_i are equal to zero (i.e., $q = 0$). In Box 4.7.1 note that when $q = 0$, the expression for t_C becomes

$$t_C = \frac{\hat{\theta}}{SE(\hat{\theta})}$$

In the computer output in Exhibit 4.6.1, the quantities in (4.6.16)–(4.6.20) under `'t-ratio'` are the quantities t_C in Box 4.7.1 when $q = 0$. These quantities cannot be used for values of q other than zero.



Problems 4.7

The following problems refer to Example 4.5.1. The data are given in Table 4.5.1 and are also stored in the file **electric.dat** on the data disk. These data are repeated in Table 4.7.1 for convenience.

Recall that

Y = (total) annual electric bill (in dollars) for the past year

X_1 = monthly income for the household (in dollars)

X_2 = number of individuals in the household

X_3 = living area (in square feet) of the house or apartment

Suppose that assumptions (A) are met and the data are obtained by simple random sampling. In particular, the regression function of Y on X_1, X_2, and X_3 is of the form

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{4.7.2}$$

TABLE 4.7.1
Electric Bill Data

Sample Item Number	Bill Y	Income X_1	Persons X_2	Area X_3
1	228	3220	2	1160
2	156	2750	1	1080
3	648	3620	2	1720
4	528	3940	1	1840
5	552	4510	3	2240
6	636	3990	4	2190
7	444	2430	1	830
8	144	3070	1	1150
9	744	3750	2	1570
10	1104	4790	5	2660
11	204	2490	1	900
12	420	3600	3	1680
13	876	5370	1	2550
14	840	3180	7	1770
15	876	5910	2	2960
16	276	3020	2	1190
17	1236	5920	3	3130
18	372	3520	2	1560
19	276	3720	1	1510
20	540	4840	1	2190
21	1044	4700	6	2620
22	552	3270	2	1350
23	756	4420	2	1990
24	636	4480	2	2070
25	708	3820	4	1850
26	960	5740	2	2700
27	1080	5600	3	3030
28	480	3950	2	1700
29	96	2290	3	890
30	1272	5580	5	3270
31	1056	5820	2	2660
32	156	3160	2	1330
33	396	2880	4	1280
34	768	3780	3	1950

A SAS output containing the results of a regression analysis for this problem is given in Exhibit 4.7.1. Answer the following.

4.7.1 The utility company wants to know whether the annual electric bill for a household is dependent on the monthly income in each subpopulation of households having a specified number of individuals and a specified living area.

EXHIBIT 4.7.1
SAS Output for Regression Analysis of Electric Data

The SAS System 0:00 Saturday, Jan 1, 1994

Dependent Variable: BILL

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	3151504.8152	1050501.6051	57.283	0.0001
Error	30	550163.42009	18338.78067		
C Total	33	3701668.2353			

Root MSE	135.42075	R-square	0.8514
Dep Mean	619.41176	Adj R-sq	0.8365
C.V.	21.86280		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-358.441569	198.73583019	-1.804	0.0813
INCOME	1	0.075137	0.13609408	0.552	0.5850
PERSONS	1	55.087633	29.04515215	1.897	0.0675
AREA	1	0.281104	0.22610987	1.243	0.2234

a Show that, for the model in (4.7.2), the average annual electric bill is *not* dependent on the monthly income in each subpopulation of households having a specified number of individuals and a specified living area if and only if $\beta_1 = 0$.

b Formulate an appropriate pair of hypotheses to test this and calculate the *P*-value. Use $\alpha = 0.05$ and state your conclusion.

4.7.2 In Problem 4.7.1, suppose, for company purposes, the manager is willing to conclude that the annual electric bill does not depend on income if the difference between the average electric bills of households whose monthly incomes differ by \$1,000 (the other factors being the same) is less than \$120. Compute a 95% confidence interval for the parameter of interest and make a decision. Which is more informative, the confidence interval of this problem or the test of Problem 4.7.1?

4.7.3 The regression function in (4.7.2) is considered adequate for predicting Y if the standard deviation of the prediction errors—i.e., σ—is less than \$50.00. If this is not so, then the investigator will look for additional explanatory variables to include

in the regression model. With this in mind, the investigator wants to test (using $\alpha = 0.05$)

$$\text{NH: } \sigma \geq 50 \qquad \text{against} \qquad \text{AH: } \sigma < 50$$

Carry out this test and state your conclusions. Which is more informative, the confidence statement obtained in Problem 4.6.13 or this test?

4.8 Analysis of Variance

Recall that for the straight line regression model we used a table to summarize certain key numerical quantities that are useful for computing confidence intervals, standard errors, and tests. The process of calculating, tabulating, and examining these key numerical quantities was termed analysis of variance, which is often abbreviated to ANOVA. We now discuss analysis of variance for the multiple linear regression model.

As in the case of straight line regression, the first key quantity is

$$SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.8.1}$$

This is called the **total sum of squares** (of deviations from the sample mean) for Y. This is the sum of squares of prediction errors when $\hat{\mu}_Y = \bar{y}$ is used to predict the Y values of the sample items. (Note that we are really not interested in predicting the Y values of the sample items because we already know them, but we do this to assess how good the predictions are likely to be when predicting unknown Y values.)

We know that the best predictor of the Y value of any item in the population without using any of the predictor variables $X_1, \ldots, X_k$, is μ_Y, the mean Y value of all items in the population. We also know that σ_Y is a measure of how well μ_Y represents the entire population of Y values. If data are obtained by simple random sampling, then we can estimate μ_Y by the sample mean $\bar{y}$ and estimate σ_Y^2 by $\sum_{i=1}^{n}(y_i - \bar{y})^2/(n-1) = SSY/(n-1)$. The quantity $SSY/(n-1)$ is sometimes written as MSY and is called the *total mean square for* Y. The divisor $n-1$ of SSY is called the *degrees of freedom associated with SSY* (or with $\hat{\sigma}_Y$).

The second key quantity is the sum of squares of the prediction errors when the sample regression function $\hat{\mu}_Y(x_1, \ldots, x_k)$ is used to predict the Y values of the sample items (in order to assess the performance of the sample regression function as a prediction function for Y). This quantity is called the **sum of squared errors** and is denoted by $SSE(X_1, \ldots, X_k)$, or SSE for short. It was defined in (4.4.14) to be

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \cdots - \hat{\beta}_k x_{i,k})^2 \tag{4.8.2}$$

and it has $(n-k-1)$ degrees of freedom associated with it. The quantity $MSE(X_1, \ldots, X_k)$ (or MSE for short), called the *mean squared error*, was defined in

(4.4.15) to be

$$MSE = \frac{SSE}{(n-k-1)} \tag{4.8.3}$$

Note that $MSE(X_1, \ldots, X_k) = \hat{\sigma}^2$ is the estimate of σ^2, regardless of whether the data are obtained by simple random sampling or by sampling with preselected X values.

The third key quantity is called the **sum of squares due to regression** and is denoted by $SSR(X_1, \ldots, X_k)$ (SSR for short). This quantity is the difference

$$SSR = SSY - SSE \tag{4.8.4}$$

and it has k degrees of freedom associated with it. Note that the degrees of freedom associated with SSR is the difference between the degrees of freedom for SSY and the degrees of freedom for SSE, i.e., $k = (n-1) - (n-k-1)$. The quantity

$$MSR = \frac{SSR}{k} \tag{4.8.5}$$

is called the *mean square due to regression.* All of these quantities are generally displayed in an ANOVA table such as Table 4.8.1.

When assumptions (A) or (B) hold, the statistic F_C in the last column of the analysis of variance table can be used to test the null hypothesis that μ_Y is equal to $\mu_Y(x_1, \ldots, x_k)$ against the alternative hypothesis that $\mu_Y(x_1, \ldots, x_k)$ is not equal to μ_Y. An equivalent way of stating this hypothesis is as follows:

$$\begin{aligned} &\text{NH: } \beta_1 = \beta_2 = \cdots = \beta_k = 0 \qquad \text{against} \\ &\text{AH: at least one of } \beta_1, \beta_2, \ldots, \beta_k \text{ is not zero} \end{aligned} \tag{4.8.6}$$

A test of the NH in (4.8.6) can be carried out using the numerical quantities exhibited in the analysis of variance table. Specifically, if a statistical test is used, the test statistic can be calculated as

$$F_C = \frac{MSR}{MSE} \tag{4.8.7}$$

If NH in (4.8.6) is true, then the quantity F_C in (4.8.7) has an F-distribution with k degrees of freedom for the numerator and $n-k-1$ degrees of freedom for the denominator. The P-value for this test is the value of α such that $F_C = F_{1-\alpha;k,n-k-1}$.

Example 4.8.1 explains how to use an analysis of variance table for multiple linear regression obtained from a computer output.

TABLE 4.8.1
ANOVA for Multiple Linear Regression

Source	Degrees of Freedom (df)	Sum of Squares (SS)	Mean Square (MS)	Computed F-Value
Regression	k	SSR	MSR	$F_C = \frac{MSR}{MSE}$
Error	$n-k-1$	SSE	MSE	
Total	$n-1$	SSY	MSY	

EXAMPLE 4.8.1

Table 4.8.2 is an ANOVA table for the GPA data of Example 4.4.2 (see Table 4.4.3). The quantity *SSY* is computed by using (4.8.1), *SSE* by using (4.8.2), and *SSR* by using (4.8.4). You will notice that this computing can be quite an arduous task if it is not done on a computer. The test statistic F_C for the test of

$$\text{NH: } \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

against

$$\text{AH: at least one of } \beta_1, \beta_2, \beta_3, \beta_4 \text{ is not zero}$$

is 21.72 with 4 degrees of freedom for the numerator and 15 degrees of freedom for the denominator. If we use $\alpha = 0.01$, then we would reject NH since the *P*-value for the test is seen to be less than 0.01 using Table T-5 in Appendix T. Hence we conclude that at least one of $\beta_1, \beta_2, \beta_3, \beta_4$ is nonzero so the regression function $\mu_Y(x_1, x_2, x_3, x_4)$ is not equal to μ_Y.

An examination of $\hat{\sigma}$ and $\hat{\sigma}_Y$ gives us an idea of how well $\mu_Y(x_1, x_2, x_3, x_4)$ and μ_Y, respectively, predict the values of *Y*. From Table 4.8.2 we find $\hat{\sigma}_Y = \sqrt{0.3866} = 0.6218$. In comparison, $\hat{\sigma} = \sqrt{0.0721} = 0.2685$, which is considerably lower than 0.6218. In Section 4.9 we discuss appropriate confidence interval procedures (valid under assumptions (B)) for comparing σ_Y and σ in any multiple linear regression problem. We note that for this problem, $\hat{\sigma}_Y$ is a valid estimate of σ_Y because data are obtained by simple random sampling. Do not take it for granted that a valid estimate of σ_Y exists. Check to see what the assumptions are. ■

Exhibit 4.8.1 shows a computer output for regression analysis of the GPA data (obtained using MINITAB) that includes an ANOVA table. Of course the ANOVA in the computer output of Exhibit 4.8.1 is the same as the one we computed in Table 4.8.2, except that the computer output includes a *P*-value (rounded to three decimal places in this case) and does not include *MSY*.

Reminder The *F* test in an ANOVA table can be used to test only

$$\text{NH: } \beta_1 = \beta_2 \ldots = \beta_k = 0 \qquad \text{against} \qquad \text{AH: at least one } \beta_i \neq 0$$

For one-sided tests or tests of $\beta_1 = \beta_2 = \ldots = \beta_k = q$ where $q \neq 0$, an ANOVA table cannot be used directly.

TABLE 4.8.2

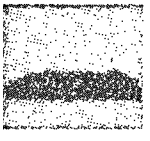
ANOVA for GPA Data

Source of Variation	Degrees of Freedom (DF)	Sum of Squares (SS)	Mean square MS	Computed *F*-Value
Regression	4	6.2643	1.5661	$F_C = \dfrac{1.5661}{0.0721} = 21.72$
Error	15	1.0815	0.0721	
Total	19	7.3458	0.3866	

EXHIBIT 4.8.1
MINITAB Output for Regression Analysis of GPA Data

```
The regression equation is
GPA = 0.162 + 0.00201 SATmath + 0.00125 SATverb +
      0.189 HSmath + 0.088 HSengl


Predictor       Coef       Stdev    t-ratio       p
Constant      0.1615      0.4375       0.37   0.717
SATmath    0.0020102   0.0005844       3.44   0.004
SATverb    0.0012522   0.0005515       2.27   0.038
HSmath       0.18944     0.09187       2.06   0.057
HSengl        0.0876      0.1765       0.50   0.627

s = 0.2685      R-sq = 85.3%     R-sq(adj) = 81.4%


Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    4      6.2643      1.5661     21.72   0.000
Error        15      1.0815      0.0721
Total        19      7.3458
```

(4.8.8)



Problems 4.8

4.8.1 A MINITAB output from a regression analysis is given in Exhibit 4.8.2 for the electric bill data of Table 4.5.1. Calculate the P-value for a test of

NH: $\beta_1 = \beta_2 = \beta_3 = 0$ against AH: at least one of β_1, β_2, and β_3 is not zero

What is your conclusion if you use $\alpha = 0.01$?

4.8.2 The manager of the marketing division of a grocery store chain wants to conduct a study in a particular city where the company wants to open a store to understand the relationship between the number of dollars Y a household spends in grocery stores each month and the following variables—monthly income X_1 for the household,

EXHIBIT 4.8.2
MINITAB Output for Regression Analysis of Electric Data

```
The regression equation is
Bill = - 358 + 0.075 Income + 55.1 Persons + 0.281 Area
```

Predictor	Coef	Stdev	t-ratio	p
Constant	-358.4	198.7	-1.80	0.081
Income	0.0751	0.1361	0.55	0.585
Persons	55.09	29.05	1.90	0.068
Area	0.2811	0.2261	1.24	0.223

```
s = 135.4       R-sq = 85.1%     R-sq(adj) = 83.7%
```

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	3151505	1050502	57.28	0.000
Error	30	550163	18339		
Total	33	3701668			

number of children X_2 in the household, and the number of adults X_3 in the household. A group of 27 grocery shoppers are selected by simple random sampling from a study population and are requested to provide the needed information. The data for these 27 shoppers are given in Table 4.8.3 and are also stored in the file **grocery.dat** on the data disk.

TABLE 4.8.3
Grocery Shoppers Data

Sample Item Number	Amount Spent in Grocery Stores Each Month Y (in dollars)	Monthly Income X_1 (in dollars)	Number of Children X_2	Number of Adults X_3
1	486	3800	2	2
2	164	1200	1	2
3	245	5000	0	1
4	714	5700	2	2
5	565	5600	1	2
6	728	4500	3	2
7	221	4400	0	1
8	209	2200	1	1
9	299	2300	3	1
10	477	4700	2	1
11	711	4300	4	2
12	379	3100	2	2
13	738	4900	4	1
14	325	3000	2	1
15	517	2000	4	2
16	441	2700	2	2
17	168	3400	1	1
18	525	2200	4	2
19	201	3800	0	1
20	358	4700	0	2
21	202	1400	3	2
22	272	1200	3	2
23	257	1700	2	2
24	376	1900	3	2
25	697	4900	3	2
26	248	2600	0	2
27	507	5100	2	1

Suppose that assumptions (A) for multiple linear regression are satisfied; thus the regression function of Y on X_1, X_2, and X_3 is of the form

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \tag{4.8.9}$$

A SAS output for a regression of Y on X_1, X_2, X_3 follows.

EXHIBIT 4.8.3
SAS Output for Regression Analysis of Grocery Data

```
                          The SAS System   0:00 Saturday,   Jan 1, 1994

Model: MODEL1 Dependent Variable: AMOUNT

                        Analysis of Variance

                          Sum of          Mean
Source            DF     Squares        Square      F Value      Prob>F

Model              3 881100.31876 293700.10625      104.399      Q.0001
Error             23  64704.42198   2813.23574
C Total           26 945804.74074

    Root MSE       53.03994     R-square       0.9316
    Dep Mean      408.51852     Adj R-sq       0.9227
    C.V.           12.98349

                      Parameter Estimates

                 Parameter       Standard     T for H0:
Variable  DF      Estimate         Error    Parameter=0    Prob > |T|

INTERCEP   1   -324.608237   51.16247889        -6.345        0.0001
INCOME     1      0.105141    0.00760990        13.816        0.0001
CHILDREN   1     96.601256    8.27176593        11.678        0.0001
ADULTS     1    110.760869   22.58900750         4.903        0.0001
```

Calculate the P-value for a test of

$$\text{NH: } \beta_1 = \beta_2 = \beta_3 = 0$$

$$\text{AH: at least one of } \beta_1, \beta_2, \text{ and } \beta_3 \text{ is not zero}$$

What is your conclusion if you use $\alpha = 0.01$?

4.8.3 An investigator is studying a population of males who have lived in mountain isolation for several generations, and she is interested in investigating the relationship between the heights Y of these males at age 18 years and the following variables.

X_1 = Length at birth
X_2 = Mother's height at age 18
X_3 = Father's height at age 18
X_4 = Maternal grandmother's height at age 18
X_5 = Maternal grandfather's height at age 18

X_6 = Paternal grandmother's height at age 18

X_7 = Paternal grandfather's height at age 18

All heights and lengths are in inches. A simple random sample of 20 males of age 18 or more was drawn from the study population, and all the preceding information was recorded. The data are given in Table 4.8.4 and are also stored in the file **age18.dat** on the data disk.

Assumptions (B) are presumed to hold. In particular, the regression function is of the form

$$\mu_Y(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7$$

A MINITAB output from a regression analysis of Y on $X_1, X_2, X_3, X_4, X_5, X_6, X_7$ follows in Exhibit 4.8.4.

Calculate the P-value for a test of

$$\text{NH:} \quad \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

against

$$\text{AH:} \quad \text{at least one of } \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \text{ and } \beta_7 \text{ is not zero}$$

What is your conclusion if you use $\alpha = 0.05$?

TABLE 4.8.4
Heights at Age 18 of a Random Sample of Mountain People

Sample Item Number	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	67.2	19.7	60.5	70.3	65.7	69.3	65.7	67.3
2	69.1	19.6	64.9	70.4	62.6	69.6	64.6	66.4
3	67.0	19.4	65.4	65.8	66.2	68.8	64.0	69.4
4	72.4	19.4	63.4	71.9	60.7	68.0	64.9	67.1
5	63.6	19.7	65.1	65.1	65.5	65.5	61.8	70.9
6	72.7	19.6	65.2	71.1	63.5	66.2	67.3	68.6
7	68.5	19.8	64.3	67.9	62.4	71.4	63.4	69.4
8	69.7	19.7	65.3	68.8	61.5	66.0	62.4	67.7
9	68.4	19.7	64.5	68.7	63.9	68.8	62.3	68.8
10	70.4	19.9	63.4	70.3	65.9	69.0	63.7	65.1
11	67.5	18.9	63.3	70.4	63.7	68.2	66.2	68.5
12	73.3	20.8	66.2	70.2	65.4	66.6	61.7	64.0
13	70.0	20.3	64.9	68.8	65.2	70.2	62.4	67.0
14	69.8	19.7	63.5	70.3	63.1	64.4	65.1	67.0
15	63.6	19.9	62.0	65.5	64.1	67.7	62.1	66.5
16	64.3	19.6	63.5	65.2	63.9	70.0	64.2	64.5
17	68.5	21.3	66.1	65.4	64.8	68.4	66.4	70.8
18	70.5	20.1	64.8	70.2	65.3	65.5	63.7	66.9
19	68.1	20.2	62.6	68.6	63.7	69.8	66.7	68.0
20	66.1	19.2	62.2	67.3	63.6	70.9	63.6	66.7

EXHIBIT 4.8.4
MINITAB Output for Regression Analysis of Data in Table 4.8.4

```
The regression equation is
Y = - 78.3 + 1.37 X1 + 0.782 X2 + 1.05 X3 - 0.120 X4 + 0.091 X5
    + 0.088 X6 - 0.102 X7

Predictor       Coef      Stdev    t-ratio       p
Constant      -78.27      26.96      -2.90   0.013
X1            1.3718     0.5207       2.63   0.022
X2            0.7824     0.1992       3.93   0.002
X3            1.0514     0.1358       7.74   0.000
X4           -0.1199     0.1717      -0.70   0.498
X5            0.0914     0.1301       0.70   0.496
X6            0.0883     0.1613       0.55   0.594
X7           -0.1017     0.1549      -0.66   0.524

s = 1.004       R-sq = 91.7%     R-sq(adj) = 86.9%

Analysis of Variance

SOURCE       DF          SS          MS        F       p
Regression    7     133.657      19.094    18.95   0.000
Error        12      12.088       1.007
Total        19     145.746
```

4.9
Comparison of Two Regression Functions (Nested Case) and Coefficients of Determination

The notation is quite heavy in this section and the next, but if you study it carefully, it will help you to learn and understand the material.

Consider a $(k+1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$. We know that the best function for predicting Y using the k predictor variables $X_1, \ldots, X_k$ is the regression function $\mu_Y(x_1, \ldots, x_k)$. However, to predict Y, investigators may want to use a smaller number of variables, which for ease of notation we take to be the first m variables $X_1, \ldots, X_m (m < k)$. They may want to know how much better the regression function of Y on $X_1, \ldots, X_k$ is for predicting Y than the regression function of Y on $X_1, \ldots, X_m$. It can be shown that when population assumptions (B) hold for the $(k+1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$, the regression function of Y on $X_1, \ldots, X_k$ cannot be worse than the regression function of Y on $X_1, \ldots, X_m$ for predicting Y.

Notation

In this section, we work with more than one regression function. It therefore becomes necessary to introduce some notational conventions to help avoid confusion. When dealing with two or more regression functions based on different sets of predictor variables, we distinguish between them by using appropriate superscripts. For instance, in the discussion in the previous paragraph, we would use the symbol $\mu_Y^{(A)}(x_1, \ldots, x_k)$ to denote the regression function of Y on $X_1, \ldots, X_k$, which we will refer to as model A, and the symbol $\mu_Y^{(B)}(x_1, \ldots, x_m)$ to denote the regression function of Y on $X_1, \ldots, X_m$, which we will refer to as model B. When discussing subpopulation standard deviations, we must again distinguish between the subpopulation standard deviation for model A and the subpopulation standard deviation for model B. We do this either by using the more complete notation $\sigma_{Y|X_1,\ldots,X_k}$ and $\sigma_{Y|X_1,\ldots,X_m}$ or, when there is no possibility of confusion, by abbreviating these to σ_A and σ_B, respectively. Recall that to judge how good a regression function is for predicting Y, the measure we use is the corresponding subpopulation standard deviation.

We illustrate with an example.

EXAMPLE 4.9.1

For an illustration, consider Example 2.2.3 and suppose an investigator wants to determine if age X_1 and weight X_2 together are better than age X_1 alone for predicting blood pressure Y and, if so, how much better. This can be stated as follows: How much better is the regression function $\mu_Y^{(A)}(x_1, x_2)$ than the regression function $\mu_Y^{(B)}(x_1)$ for predicting Y? So the investigator will be interested in comparing the subpopulation standard deviation $\sigma_{Y|X_1,X_2} = \sigma_A$ with $\sigma_{Y|X_1} = \sigma_B$. ■

The decision of whether to use the full set $X_1, \ldots, X_k$ of predictor variables, or the subset $X_1, \ldots, X_m$, usually depends, at least in part, on

1. The cost of observing the values of the additional variables $X_{m+1}, \ldots, X_k$
2. The improvement in prediction that is made possible by using the full set of predictors rather than the subset of predictors under consideration

As explained earlier, we let $\mu_Y^{(A)}(x_1, \ldots, x_k)$ denote the regression function of Y on the k predictor variables $X_1, \ldots, X_k$ and call this model A. We let $\mu_Y^{(B)}(x_1, \ldots, x_m)$ denote the regression function of Y on the subset of m predictor variables $X_1, \ldots, X_m$ and call this model B.

The investigator is likely to be interested in obtaining the answers to the following questions.

1. How well does model A predict Y?
2. Is model A an adequate predictor of Y?
3. How well does model B predict Y?
4. Is model B an adequate predictor of Y?
5. How much better is model A than model B for predicting Y?

In this section we discuss methods for answering these and other questions. An important difference between assumptions (A) and assumptions (B) should be noted, especially when discussing several regression models. *Even though population assumptions* (A) *may hold for* $\{(Y, X_1, \ldots, X_k)\}$, *this does not imply that population assumptions* (A) *will hold for* $\{(Y, X_1, \ldots, X_m)\}$, *where* $\{X_1, \ldots, X_m\}$ *is a subset of the variables in* $\{X_1, \ldots, X_k\}$. *But if population assumptions* (B) *hold for* $\{(Y, X_1, \ldots, X_k)\}$, *then population assumptions* (B) *must hold for* $\{(Y, X_1, \ldots, X_m)\}$ *for every subset* $\{X_1, \ldots, X_m\}$ *of predictors from the full set* $\{X_1, \ldots, X_k\}$. This leads us to make the following assumption.

Throughout this section, we presume that assumptions (B) are valid. This means that the $(k+1)$-variable population $\{(Y, X_1, \ldots, X_k)\}$ and the $(m+1)$-variable population $\{(Y, X_1, \ldots, X_m)\}$ (where $m < k$) both satisfy (population) assumptions (B) in Box 4.3.2. This also means that the data are obtained by simple random sampling. (4.9.1)

Remarks

a By virtue of (4.9.1), the regression functions $\mu_Y^{(A)}(x_1, \ldots, x_k)$—i.e., model A—and $\mu_Y^{(B)}(x_1, \ldots, x_m)$—i.e., model B—are of the form

$$\mu_Y^{(A)}(x_1, \ldots, x_k) = \beta_0^A + \beta_1^A x_1 + \cdots + \beta_k^A x_k \tag{4.9.2}$$

and

$$\mu_Y^{(B)}(x_1, \ldots, x_m) = \beta_0^B + \beta_1^B x_1 + \cdots + \beta_m^B x_m \tag{4.9.3}$$

respectively.

b The regression function for model B is said to be nested in the regression function for model A because the set of predictor variables in $\mu_Y^{(B)}(x_1, \ldots, x_m)$ is a *subset* of the set of predictor variables in $\mu_Y^{(A)}(x_1, \ldots, x_k)$.

c We use the superscripts A and B to distinguish between the β coefficients in the two regression functions $\mu_Y^{(A)}(x_1, \ldots, x_k)$ and $\mu_Y^{(B)}(x_1, \ldots, x_m)$, respectively. The regression coefficients β_i^A and β_i^B for $i = 0, 1, 2, \ldots, m$ are different unless $\beta_{m+1}^A = \cdots = \beta_k^A = 0$, in which case the two regression functions are identical and $\beta_i^A = \beta_i^B$ for $i = 1, \ldots, m$.

d As mentioned earlier, we sometimes use σ_A to denote the subpopulation standard deviation for model A and σ_B to denote the subpopulation standard deviation for model B.

In Example 4.9.1 the set of factors in model B (i.e., $\{X_1\}$) is a subset of the set of factors in model A (i.e., $\{X_1, X_2\}$), and so model B is nested in model A. However, if model A is $\mu_Y^{(A)}(x_1) = \beta_0^A + \beta_1^A x_1$ and model B is $\mu_Y^{(B)}(x_2) = \beta_0^B + \beta_2^B x_2$, then the

set of factors in model B (i.e., $\{X_2\}$) is not a subset of the set of factors in model A (i.e., $\{X_1\}$), so for this case the discussion in this section does not apply, but the discussion in Section 4.10 does.

Subpopulation Standard Deviations as Measures of Goodness of Prediction

As usual, we use the quantity σ_A as a summary measure of how well model A in (4.9.2) predicts Y. Likewise, the quantity σ_B is a summary measure of how well model B in (4.9.3) predicts Y.

Relationship Between Subpopulation Standard Deviations in Nested Models

Under the assumption in (4.9.1), it can be shown that

$$\sigma_A \leq \sigma_B \tag{4.9.4}$$

i.e., the subpopulation standard deviation for the larger model (model A) is smaller than or equal to the subpopulation standard deviation for the smaller nested model (model B). To elaborate on this, consider the k predictor factors $X_1, \ldots, X_k$, and let S_1 be any subset of these k factors and S_2 be any subset of the factors in S_1. Then the following inequalities are true.

$$0 \leq \sigma_A \leq \sigma_{S_1} \leq \sigma_{S_2} \leq \sigma_Y \tag{4.9.5}$$

For instance consider the five-variable population of Example 4.2.1. Let $S_1 = \{X_1, X_2, X_4\}$ and $S_2 = \{X_1, X_2\}$. Then σ_{S_1} stands for $\sigma_{Y|X_1,X_2,X_4}$ and σ_{S_2} stands for $\sigma_{Y|X_1,X_2}$. By (4.9.5) the following is true:

$$0 \leq \sigma_{Y|X_1,X_2,X_3,X_4} \leq \sigma_{Y|X_1,X_2,X_4} \leq \sigma_{Y|X_1,X_2} \leq \sigma_Y$$

On the other hand, if $S_1 = \{X_1, X_2, X_3\}$ and $S_2 = \{X_1, X_4\}$, then S_2 is not a subset of S_1 (and S_1 is not a subset of S_2), so it is not known whether $\sigma_{Y|X_1,X_2,X_3}$ is larger or smaller than $\sigma_{Y|X_1,X_4}$. However, from (4.9.5) we get

$$0 \leq \sigma_{Y|X_1,X_2,X_3,X_4} \leq \sigma_{Y|X_1,X_2,X_3} \leq \sigma_Y$$

and

$$0 \leq \sigma_{Y|X_1,X_2,X_3,X_4} \leq \sigma_{Y|X_1,X_4} \leq \sigma_Y$$

To illustrate inequalities such as (4.9.5), we consider the simple case described in Example 2.2.2, where Y = first-year maintenance cost of a new car and X = number of miles the car is driven the first year. The target population $\{(Y, X)\}$ is the set of all cars that will be made by company A next year and driven between 5,000 and 20,000 miles. The study population is a set of similar cars made by company A last year. The quantity σ_Y is the standard deviation of the first-year maintenance costs of *all* the cars in the study population. If only the subpopulation of cars that were driven 10,000 miles is considered, then the standard deviation of the first-year maintenance costs of all cars driven 10,000 miles the first-year is equal to $\sigma_{Y|X}$.

We would typically expect the first-year maintenance costs of these cars to be more similar than for the entire population of cars, and hence we would expect $\sigma_{Y|X} \leq \sigma_Y$.

Adequacy of Prediction Functions

Whether or not a prediction function is adequate for predicting Y depends on the particular problem.

We consider a prediction function to be adequate for predicting values of Y if a proportion p (p is taken to be greater than 0.5) of the Y values are within d units of the corresponding predicted values. The values of p and d are specified by the investigator.

Let us examine the conditions under which the regression function $\mu_Y(x_1, \ldots, x_k)$ may be regarded as an adequate prediction function for predicting values of Y. When population assumptions (A) or (B) hold for $\{(Y, X_1, \ldots X_k)\}$, a proportion p of the Y values with $X_1 = x_1, \ldots, X_k = x_k$ will lie in the interval

$$[\mu_Y(x_1, \ldots, x_k) - z_{(1+p)/2}\sigma_{Y|X_1,\ldots,X_k}, \mu_Y(x_1, \ldots, x_k) + z_{(1+p)/2}\sigma_{Y|X_1,\ldots,X_k}]$$

So if $\mu_Y(x_1, \ldots, x_k)$ is to be within d units of a proportion p of the Y values in the corresponding subpopulation, we must have

$$z_{(1+p)/2}\sigma_{Y|X_1,\ldots,X_k} \leq d \qquad \text{i.e.,} \qquad \sigma_{Y|X_1,\ldots,X_k} \leq d/z_{(1+p)/2}$$

Sometimes an investigator may say "$\mu_Y(x_1, \ldots, x_k)$ is adequate for predicting Y values if $\sigma_{Y|X_1,\ldots,X_k}$ is less than a specified value d^*." In this case d and d^* are related by the equation

$$d^* = d/z_{(1+p)/2} \qquad \text{i.e.,} \qquad d = d^* z_{(1+p)/2}$$

For example, in Problem 4.7.3, the investigator considers the regression function in (4.7.2) to be adequate for predicting the annual electric bill if $\sigma_{Y|X_1,X_2,X_3} \leq \$50$. So here $d^* = 50$. The numbers d and p (or d^* and p) specified by an investigator are based on various practical considerations in the context of the particular problem.

We now apply the preceding criterion of adequacy to the regression functions $\mu_Y^{(A)}(x_1, \ldots, x_k)$ and $\mu_Y^{(B)}(x_1, \ldots, x_m)$ given in (4.9.2) and (4.9.3), respectively. By (4.9.1), assumptions (B) for multiple regression hold for both models. So we can conclude that $\mu_Y^{(A)}(x_1, \ldots, x_k)$ is an adequate predictor of Y if $z_{(1+p)/2}\sigma_A \leq d$, i.e., if $\sigma_A \leq d/z_{(1+p)/2}$. Similarly, $\mu_Y^{(B)}(x_1, \ldots, x_m)$ is an adequate predictor of Y if $\sigma_B \leq d/z_{(1+p)/2}$. It is possible that $\mu_Y^{(A)}(x_1, \ldots, x_k)$ and $\mu_Y^{(B)}(x_1, \ldots, x_m)$ are *both* adequate for predicting Y or that *neither* $\mu_Y^{(A)}(x_1, \ldots, x_k)$ nor $\mu_Y^{(B)}(x_1, \ldots, x_m)$ is adequate for predicting Y. Of course, in a real problem the regression functions $\mu_Y^{(A)}(x_1, \ldots, x_k)$ and $\mu_Y^{(B)}(x_1, \ldots, x_m)$ are not known and must be estimated. This

adds another source of uncertainty to the problem. Nevertheless, $\hat{\sigma}_A$ and $\hat{\sigma}_B$, the estimates of σ_A and σ_B, respectively, and the confidence intervals for these, are useful for determining the adequacy of prediction functions.

Performance of Model A Relative to Model B

There are instances where it is difficult to specify a number d to determine whether a prediction function is adequate. In these instances and in other situations, we may be interested in the relative performances of competing prediction functions. In the present situation, two prediction functions, $\mu_Y^{(A)}(x_1, \ldots, x_k)$ and $\mu_Y^{(B)}(x_1, \ldots, x_m)$ (i.e., two models A and B), are being considered, and hence we may want to compare σ_A and σ_B. We can either examine σ_A and σ_B individually or examine some function of σ_A and σ_B that may be particularly meaningful in a given problem. For instance, we may want to examine the following functions of σ_A and σ_B to determine how much better $\mu_Y^{(A)}(x_1, \ldots, x_k)$ is than $\mu_Y^{(B)}(x_1, \ldots, x_m)$ for predicting Y.

1 $\sigma_B - \sigma_A$
2 σ_B/σ_A
3 $\sigma_B^2 - \sigma_A^2$
4 σ_B^2/σ_A^2

In this book we use the ratio σ_B/σ_A for a comparison of σ_A relative to σ_B. A function of σ_A and σ_B that is related to the ratio σ_B/σ_A and has found widespread use in the literature is called the *multiple coefficient of determination*, and we discuss this next.

Multiple Coefficient of Determination

A commonly used measure that summarizes the performance of $\mu_Y^{(A)}(x_1, \ldots, x_k)$ (model A) as a predictor of Y relative to $\mu_Y^{(B)}(x_1, \ldots, x_m)$ (model B) is the **multiple-partial coefficient of determination** of Y with $X_{m+1}, \ldots, X_k$ when $X_1, \ldots, X_m$ are held fixed. It is denoted by $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ and is defined by

$$\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} = \frac{\sigma^2_{Y|X_1,\ldots,X_m} - \sigma^2_{Y|X_1,\ldots,X_k}}{\sigma^2_{Y|X_1,\ldots,X_m}} = \frac{\sigma_B^2 - \sigma_A^2}{\sigma_B^2} = 1 - \frac{1}{(\sigma_B/\sigma_A)^2} \tag{4.9.6}$$

Thus $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ is the *proportional reduction in the variance of prediction errors by using* $\mu_Y^{(A)}(x_1, \ldots, x_k)$ *to predict* Y, *relative to using* $\mu_Y^{(B)}(x_1, \ldots, x_m)$ *to predict* Y.

The positive square root of $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ is generally called the *multiple-partial correlation coefficient* of Y with $X_{m+1}, X_{m+2}, \ldots, X_k$ when $X_1, \ldots, X_m$ are held fixed. It is denoted by $\rho_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$. The word *multiple* in multiple-partial coefficient of determination means there is more than one predictor variable; i.e., $X_1, \ldots, X_k$, $k > 1$. The word *partial* means that some of the predictor variables are held fixed.

To examine (4.9.6) in more detail, consider only the numerator

$$\sigma_B^2 - \sigma_A^2 \tag{4.9.7}$$

There are two equivalent ways of looking at (4.9.7).

1 As a measure of how much better the k factors $X_1, \ldots, X_k$ together are than the m factors $X_1, \ldots, X_m$ for predicting values of Y.

2 As a measure of how much factors $X_{m+1}, X_{m+2}, \ldots, X_k$ contribute to predicting Y in addition to what factors $X_1, X_2, \ldots, X_m$ contribute.

Consider the special case in Example 4.9.1 where the relationship of blood pressure Y with age X_1 and weight X_2 is studied. Let $A = \{X_1, X_2\} = \{\text{age, weight}\}$, $B = \{X_1\} = \{\text{age}\}$, and $C = \{X_2\} = \{\text{weight}\}$.

1 σ_A is a measure of how good age and weight together are for predicting blood pressure Y.

2 σ_B is a measure of how good age alone is for predicting blood pressure Y.

3 σ_C is a measure of how good weight alone is for predicting blood pressure Y.

4 The quantity $\sigma_B^2 - \sigma_A^2$ is often used in the following two equivalent ways.
   a As a measure of how much better age and weight, X_1 and X_2, together are for predicting blood pressure Y than age X_1 alone is.
   b As a measure of how much weight contributes to predicting Y in addition to what age contributes.

When model B in (4.9.3) uses no predictors, i.e., when $m = 0$, then $\mu_Y^{(B)} = \beta_0^B = \mu_Y$ is the quantity used to predict Y under model B and so σ_B is simply σ_Y. In this case there are no predictor variables to the right of the '|' symbol in $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ (since $m = 0$), and therefore it is simply written as $\rho^2_{Y(X_1,\ldots,X_k)}$ and is called the **multiple coefficient of determination** of Y with $X_1, \ldots, X_k$. The word *partial* is omitted since no predictors are held fixed. Thus $\rho^2_{Y(X_1,\ldots,X_k)}$ measures relatively how much better $\mu_Y^{(A)}(x_1, \ldots, x_k)$ is for predicting Y than μ_Y is. For completeness, we give the definition of $\rho^2_{Y(X_1,\ldots,X_k)}$ in (4.9.8).

D E F I N I T I O N

$$\rho^2_{Y(X_1,\ldots,X_k)} = \frac{\sigma_Y^2 - \sigma^2_{Y|X_1,\ldots,X_k}}{\sigma_Y^2} = \frac{\sigma_Y^2 - \sigma_A^2}{\sigma_Y^2} = 1 - \frac{1}{(\sigma_Y/\sigma_A)^2} \quad \blacksquare \tag{4.9.8}$$

The case when $k = 1$ and $m = 0$ was discussed in Section 3.9, and we write ρ^2_{Y,X_1} instead of $\rho^2_{Y(X_1)}$. Also when $m = k - 1$, the word *multiple* is often omitted; for instance, the quantity $\rho^2_{Y,X_1|X_2}$ is simply called the *partial coefficient of determination* of Y with X_1 where X_2 is held fixed.

To summarize:

1 $\rho^2_{Y(X_{m+1},X_{m+2},\ldots,X_k)}$ is the multiple coefficient of determination of Y with the predictor variables in parentheses, namely $X_{m+1}, X_{m+2}, \ldots, X_k$.

2 $\rho^2_{Y(X_{m+1},X_{m+2},\ldots,X_k)|X_1,X_2,\ldots,X_m}$ is the multiple-partial coefficient of determination of Y with $X_{m+1}, X_{m+2}, \ldots, X_k$ (the variables in parentheses) when the variables $X_1, X_2, \ldots, X_m$ (the variables following the vertical line '|') are held fixed.

Properties of $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$

Under population assumptions (B), the statements in Box 4.9.1 hold.

BOX 4.9.1

1 $0 \le \rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} \le 1$.

2 $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} = 0$ if and only if $\beta^A_{m+1} = \ldots = \beta^A_k = 0$ in (4.9.2). In this case the two regression functions, $\mu^{(A)}_Y(x_1, \ldots, x_k)$ in (4.9.2) and $\mu^{(B)}_Y(x_1, \ldots, x_m)$ in (4.9.3), are identical. Moreover, $\sigma_A = \sigma_B$.

3 If $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} > 0$, then at least one of the parameters $\beta^A_{m+1}, \ldots, \beta^A_k$ in (4.9.2) is nonzero. Also $\sigma_A < \sigma_B$.

4 $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} = 1$ if and only if $\sigma_A = 0$. In this case $\mu^{(A)}_Y(x_1, \ldots, x_k)$ is a *perfect predictor* of Y.

For the case when $m = 0$ (i.e., model B is $\mu^{(B)}_Y = \beta^B_0 = \mu_Y$ and it contains none of the predictors $X_1, \ldots, X_k$), the statements in Box 4.9.1 specialize to those in Box 4.9.2.

BOX 4.9.2

1 $0 \le \rho^2_{Y(X_1,\ldots,X_k)} \le 1$.

2 $\rho^2_{Y(X_1,\ldots,X_k)} = 0$ if and only if $\beta^A_1 = \ldots = \beta^A_k = 0$ in (4.9.2). In this case the regression function $\mu^{(A)}_Y(x_1, \ldots, x_k)$ is equal to $\mu^{(B)}_Y = \mu_Y$. Moreover, $\sigma_A = \sigma_Y$.

3 If $\rho^2_{Y(X_1,\ldots,X_k)} > 0$, then at least one of the parameters $\beta^A_1, \ldots, \beta^A_k$ in (4.9.2) is nonzero. Also $\sigma_A < \sigma_Y$.

4 $\rho^2_{Y(X_1,\ldots,X_k)} = 1$ if and only if $\sigma_A = 0$. In this case $\mu^{(A)}_Y(x_1, \ldots, x_k)$ is a *perfect predictor* of Y.

EXAMPLE 4.9.2

To illustrate the preceding concepts, let us consider Example 2.2.3 where Y is (systolic) blood pressure, X_1 is age, and X_2 is weight. Suppose we want to examine the relationship among Y, X_1, and X_2 for the population of all females in California

between the ages of 25 and 75. Suppose the three-variable population $\{(Y, X_1, X_2)\}$ is Gaussian.

- The two-variable populations $\{(Y, X_1)\}$ and $\{(Y, X_2)\}$ and the one-variable population $\{Y\}$ are also Gaussian.
- The average blood pressure of the entire population $\{Y\}$ is μ_Y, and the standard deviation is σ_Y.
- The average blood pressure for the subpopulation of all females who are x_1 years old is

$$\mu_Y^{(B)}(x_1) = \beta_0^B + \beta_1^B x_1 \tag{4.9.9}$$

  with standard deviation σ_B.
- The average blood pressure for the subpopulation of all females who weigh x_2 pounds is

$$\mu_Y^{(C)}(x_2) = \beta_0^C + \beta_2^C x_2 \tag{4.9.10}$$

  with standard deviation σ_C.
- The average blood pressure for the subpopulation of all females whose age is x_1 and weight is x_2 is

$$\mu_Y^{(A)}(x_1, x_2) = \beta_0^A + \beta_1^A x_1 + \beta_2^A x_2 \tag{4.9.11}$$

  with standard deviation σ_A.

Also

$$\sigma_Y, \sigma_B, \sigma_C, \sigma_A$$

are the quantities that are used as summary measures of how good μ_Y, $\mu_Y^{(B)}(x_1)$, $\mu_Y^{(C)}(x_2)$, and $\mu_Y^{(A)}(x_1, x_2)$, respectively, are as predictors of Y. From (4.9.5) we have

$$0 \le \sigma_A \le \sigma_B \le \sigma_Y \qquad \text{and} \qquad 0 \le \sigma_A \le \sigma_C \le \sigma_Y \tag{4.9.12}$$

We may want to determine the answer to several questions about X_1 and X_2 as predictors of Y. Some of these questions follow.

1 How good are age and weight together as predictors of blood pressure Y?
2 How well can Y be predicted if no predictor variables are used?
3 How good is age alone as a predictor of Y (if weight is ignored)?
4 How good is weight alone as a predictor of Y (if age is ignored)?
5 How good is weight as a predictor of Y in the subpopulation of all females who are 35 years old?
6 How much better are age and weight together for predicting blood pressure than is age alone?
7 How good is weight as a predictor of blood pressure after age has been accounted for?

8 *Relatively*, how much better are age and weight together for predicting blood pressure than age alone?

9 *Relatively*, how good is weight for predicting blood pressure after age has been taken into account?

The answers to these questions are given next in terms of the population parameters.

1 A measure of how good age and weight together are for predicting blood pressure—i.e., how good $\mu_Y^{(A)}(x_1, x_2)$ is for predicting Y—is σ_A because this is the standard deviation of Y in the subpopulation determined by fixed values of X_1 and X_2.

2 When no predictor variables are used to predict Y, the best value to use to predict Y is μ_Y, the mean of the Y values in the entire population. A measure of how good μ_Y is for predicting Y is provided by σ_Y, the standard deviation of the population $\{Y\}$.

3 The best prediction function for predicting Y using age alone is $\mu_Y^{(B)}(x_1)$, the regression function of Y on X_1. A measure of how good $\mu_Y^{(B)}(x_1)$ is for predicting Y is provided by σ_B, the standard deviation of the Y values in the subpopulation determined by X_1 (age) when X_2 (weight) is not considered.

4 The best prediction function for predicting Y using weight alone is $\mu_Y^{(C)}(x_2)$, the regression function of Y on X_2. A measure of how good $\mu_Y^{(C)}(x_2)$ is for predicting Y is provided by σ_C, the standard deviation of the Y values in the subpopulation determined by X_2 (weight) when X_1 (age) is not considered.

5 σ_A, the same as the answer to question (1), because assumptions (B) in Box 4.3.2 imply that the standard deviations of the subpopulations determined by X_1 and X_2 are the same for all values of X_1 and X_2.

6 $\sigma_B - \sigma_A$, or σ_B/σ_A, or $\sigma_B^2 - \sigma_A^2$, or σ_B^2/σ_A^2, depending on which measure you want to use.

7 Same answer as (6).

8 $\dfrac{\sigma_B^2 - \sigma_A^2}{\sigma_B^2} = \rho^2_{Y,X_2|X_1}$ if we use variances as the summary measures.

9 Same answer as (8).

We illustrate other important points using this example.

1 In the multiple regression model in (4.1.1), the coefficient β_i (for $i = 1, \ldots, k$) is the change in $\mu_Y(x_1, \ldots, x_k)$ per unit change in X_i when the other factors are held fixed. For example, in the model in (4.9.11), β_2^A is the change in the average blood pressure per unit change in weight for all females of the same age (say, 35 years old), i.e., when X_1 (age) is held fixed.

Note *Sometimes it is not meaningful to change one variable, say X_i, and hold other variables fixed. For example, consider the regression model in (4.2.1a). If we let $X_1 = Z_1$, $X_2 = Z_1Z_2$, and $X_3 = Z_1^2$, we get the multiple linear regression*

model

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

and clearly we cannot hold X_1 and X_2 fixed and let X_3 change.

2 Some predictor factors are controllable factors and some are noncontrollable factors. In the preceding example, suppose β_1^A and β_2^A in (4.9.11) are positive, which means that $\mu_Y^{(A)}(x_1, x_2)$ decreases as x_1 and/or x_2 decreases. If a physician recognizes that a patient's blood pressure should be reduced, the model indicates that it *may be possible* to reduce it by lowering the patient's age or weight. Of course it is not possible to reduce the patient's age (X_1 is a *noncontrollable* factor). But weight is a controllable factor (at least controllable to some extent), and it *may be possible* to reduce the patient's weight by diet and hence possibly reduce the blood pressure. Whether changing the value of X_2 will actually result in a change in the value of Y cannot be known based on observational studies, and it has to be studied using controlled experiments. ■

The quantity $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ is a population parameter, and so it is unknown in any real problem. A valid point estimate of $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ can be calculated from sample data according to the formula given in (4.9.13) if assumptions (B) are satisfied; in particular, *the data must be obtained by simple random sampling. If data are obtained by sampling with preselected values of $X_1, \ldots, X_k$, then no valid estimate of $\rho^2_{(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ is available.*

Point Estimate of $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$

The point estimate of $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$, the multiple-partial coefficient of determination of Y with $X_{m+1}, \ldots, X_k$ when $X_1, \ldots, X_m$ are held fixed, is given by

$$\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} = \frac{SSE(X_1, \ldots, X_m) - SSE(X_1, \ldots, X_k)}{SSE(X_1, \ldots, X_m)} \tag{4.9.13}$$

The sum of squares $SSE(X_1, \ldots, X_k)$ was defined in (4.4.14), and the sum of squares $SSE(X_1, \ldots, X_m)$ is defined similarly but using only the variables $X_1, \ldots, X_m$. For notational convenience we write $SSE(A)$ for $SSE(X_1, \ldots, X_k)$ and $SSE(B)$ for $SSE(X_1, \ldots, X_m)$ because these correspond to models A and B in (4.9.2) and (4.9.3), respectively. Thus we have

$$\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} = \frac{SSE(B) - SSE(A)}{SSE(B)} \tag{4.9.14}$$

An Alternate Point Estimate of $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$

The definition of $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ in (4.9.6) might suggest that the point estimate of $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ would be given by

$$\frac{\hat{\sigma}^2_{Y|X_1,\ldots,X_m} - \hat{\sigma}^2_{Y|X_1,\ldots,X_k}}{\hat{\sigma}^2_{Y|X_1,\ldots,X_m}}$$

and indeed this estimate is sometimes used. It is called *the estimate of the multiple-partial coefficient of determination (adjusted for degrees of freedom) of* Y *with* $X_{m+1},\ldots,X_k$ *when* $X_1,\ldots,X_m$ *are held fixed* and is written as $Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}]$. Thus,

$$Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}] = \frac{\hat{\sigma}^2_{Y|X_1,\ldots,X_m} - \hat{\sigma}^2_{Y|X_1,\ldots,X_k}}{\hat{\sigma}^2_{YX_1,\ldots,X_m}} \tag{4.9.15}$$

$$= \frac{MSE(X_1,\ldots,X_m) - MSE(X_1,\ldots,X_k)}{MSE(X_1,\ldots,X_m)} \tag{4.9.16}$$

The mean square $MSE(X_1,\ldots,X_k)$ was defined in (4.4.15), and the mean square $MSE(X_1,\ldots,X_m)$ is defined similarly but using only the variables $X_1,\ldots,X_m$. For notational convenience we write $MSE(A)$ for $MSE(X_1,\ldots,X_k)$ and $MSE(B)$ for $MSE(X_1,\ldots,X_m)$ because these correspond to models A and B in (4.9.2) and (4.9.3), respectively. Thus we have

$$Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}] = \frac{MSE(B) - MSE(A)}{MSE(B)} \tag{4.9.17}$$

Relationship Between $Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}]$ and $\hat{\rho}^2_{Y(X_{m+1}\ldots,X_k)|X_1,\ldots,X_m}$

The following equations relate $Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}]$ and $\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$.

$$1 - Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}] = \frac{n-m-1}{n-k-1}(1 - \hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}) \tag{4.9.18}$$

$$1 - \hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} = \frac{n-k-1}{n-m-1}(1 - Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}]) \tag{4.9.19}$$

It follows that $Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}]$ is always less than or equal to $\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$. Note that $Adj[\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}]$ can take negative values, and if it does, we replace the negative value with zero because it is estimating $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$, which is always greater than or equal to zero. On the other hand, the estimate $\hat{\rho}^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ is never negative. Unless we specifically state otherwise, we always use the estimate in (4.9.13).

Point Estimates for $\rho^2_{Y(X_1,\ldots,X_k)}$

When $m = 0$, i.e., when model B uses none of the predictors $X_1, \ldots, X_k$, the formulas in (4.9.13)–(4.9.16) reduce to

$$\hat{\rho}^2_{Y(X_1,\ldots,X_k)} = \frac{SSY - SSE(X_1,\ldots,X_k)}{SSY} \tag{4.9.20}$$

and

$$Adj[\hat{\rho}^2_{Y(X_1,\ldots,X_k)}] = \frac{\hat{\sigma}^2_Y - \hat{\sigma}^2_{Y|X_1,\ldots,X_k}}{\hat{\sigma}^2_Y} \tag{4.9.21}$$

$$= \frac{MSY - MSE(X_1,\ldots,X_k)}{MSY} \tag{4.9.22}$$

respectively.

Investigators often use $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ to decide whether $\mu_Y^{(A)}(x_1,\ldots,x_k)$ is a better predictor of Y than $\mu_Y^{(B)}(x_1,\ldots,x_m)$. Technically, if $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} > 0$, then we can conclude that σ_A (i.e., $\sigma_{Y|X_1,\ldots,X_k}$) is smaller than σ_B (i.e., $\sigma_{Y|X_1,\ldots,X_m}$) and hence $\mu_Y^{(A)}(x_1,\ldots,x_k)$ is better than $\mu_Y^{(B)}(x_1,\ldots,x_m)$ for predicting Y. However, in practice, we generally want to know *how much smaller σ_A is than σ_B*. (Recall that, as stated in (4.9.4), σ_A is never greater than σ_B.) An obvious procedure is to examine the estimated values and confidence intervals for σ_A and σ_B. Additionally, we could use a confidence interval for σ_B/σ_A for this purpose.

Confidence Interval for $\sigma_{Y|X_1,\ldots,X_m}/\sigma_{Y|X_1,\ldots,X_k}$ (i.e., σ_B/σ_A)

A two-sided confidence interval for $\sigma_{Y|X_1,\ldots,X_m}/\sigma_{Y|X_1,\ldots,X_k}$ with confidence coefficient equal to $1-\alpha$, can be obtained by first obtaining a confidence interval for $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_2,\ldots,X_m}$. Confidence intervals for $\sigma_Y/\sigma_{Y|X_1,\ldots,X_k}$ can be obtained as a special case of this procedure with $m = 0$. We do not discuss this here and you are referred to [10]. However, a procedure using the Bonferroni method, given in Box 4.10.2 in the next section, can be used here to obtain confidence intervals for σ_B/σ_A with confidence coefficient greater than or equal to $1-\alpha$, but the resulting confidence intervals are wide and thus cannot be recommended for routine use in applications. We discuss them in Box 4.10.2 for illustrative purposes only.

We illustrate the procedures discussed in this section in Examples 4.9.3 and 4.9.4.

EXAMPLE 4.9.3

For the GPA data in Example 4.4.2, we show how to compute point estimates for $\rho^2_{Y(X_1,X_2,X_3,X_4)}$ and for σ_B/σ_A. We suppose that assumptions (B) in Box 4.3.2 are satisfied. The quantities needed in the formulas in (4.9.20) and (4.9.21) can be obtained from the ANOVA in Table 4.8.2. These quantities are

$$SSY = 7.3458 \qquad SSE(X_1,X_2,X_3,X_4) = 1.0815 \qquad MSY = 0.3866$$

$$MSE(X_1, X_2, X_3, X_4) = 0.0721$$

The point estimate of $\rho^2_{Y(X_1,X_2,X_3,X_4)}$, computed using (4.9.20), is

$$\hat{\rho}^2_{Y(X_1,X_2,X_3,X_4)} = \frac{7.3458 - 1.0815}{7.3458} = \frac{6.2643}{7.3458} = 0.853, \text{ i.e., } 85.3\%$$

This is the quantity labeled `R-sq` in the computer output in Exhibit 4.8.1. It is an estimate of relatively how much better the regression function

$$\mu_Y(x_1, x_2, x_3, x_4) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \qquad (4.9.23)$$

is for predicting GPA than μ_Y is. Note that the quantity labeled `R-sq(adj)` in Exhibit 4.8.1 is the alternate estimate $adj[\hat{\rho}^2_{Y(X_1,X_2,X_3,X_4)}]$ of $\rho^2_{Y(X_1,X_2,X_3,X_4)}$ given in (4.9.21), and it is equal to

$$adj[\hat{\rho}^2_{Y(X_1,X_2,X_3,X_4)}] = \frac{MSY - MSE(X_1, X_2, X_3, X_4)}{MSY} = \frac{0.3866 - 0.0721}{0.3866}$$

$$= 0.814 \text{ (i.e., } 81.4\%)$$

A point estimate of σ_B/σ_A is given by $\sigma_B/\sigma_A = \sqrt{0.3866/0.0721} = 2.32$. ■

EXAMPLE 4.9.4

In Example 4.9.3, suppose the director of admissions wants to compare the performance of the model

$$\mu_Y^{(B)}(x_3, x_4) = \beta_0^B + \beta_3^B x_3 + \beta_4^B x_4 \qquad (4.9.24)$$

which uses only X_3 (HSmath) and X_4 (HSenglish), with the model

$$\mu_Y^{(A)}(x_1, x_2, x_3, x_4) = \beta_0^A + \beta_1^A x_1 + \beta_2^A x_2 + \beta_3^A x_3 + \beta_4^A x_4 \qquad (4.9.25)$$

for predicting Y. If model A is not much better than model B, then she may decide that applicants need not take the SAT.

Also, suppose the director decides that the prediction function $\mu_Y^{(A)}(x_1, x_2, x_3, x_4)$ would be considered adequate for predicting values of Y if a proportion $p = 0.95$ of the Y values being predicted are within 0.8 grade point unit of $\mu_Y^{(A)}(x_1, x_2, x_3, x_4)$, i.e., if $z_{0.975}\sigma_A \leq 0.8$, or equivalently, $\sigma_A \leq 0.8/1.96 = 0.41$. Likewise, the prediction function $\mu_Y^{(B)}(x_3, x_4)$ would be adequate for predicting values of Y if a proportion $p = 0.95$ of the Y values being predicted is within 0.8 grade point unit of $\mu_Y^{(B)}(x_3, x_4)$, i.e., if $\sigma_B \leq 0.41$. We now examine the adequacy of model A and model B and also compare them relative to one another. Assumptions (B) for regression are presumed valid.

To obtain the required point estimates and confidence intervals we need $SSE(X_3, X_4)$ and $SSE(X_1, X_2, X_3, X_4)$. We can get these from the ANOVA tables for the models in (4.9.24) and (4.9.25) given in Exhibit 4.9.1.

The standard deviation σ_B for model B is estimated to be 0.3771 (see (4.9.26)). A 90% confidence interval for σ_B is given by the statement

$$C\left[0.296 \leq \sigma_B \leq 0.528\right] = 0.90$$

EXHIBIT 4.9.1

MINITAB Output for Example 4.9.4

Regression Analysis for Model B

The regression equation is
gpa = - 0.340 + 0.417 hsmath + 0.579 hsengl

Predictor	Coef	Stdev	t-ratio	p
Constant	-0.3400	0.5644	-0.60	0.555
hsmath	0.4171	0.1054	3.96	0.001
hsengl	0.5790	0.1857	3.12	0.006

s = 0.3771 R-sq = 67.1% R-sq(adj) = 63.2% (4.9.26)

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	4.9278	2.4639	17.32	0.000
Error	17	2.4180	0.1422		
Total	19	7.3458			

Regression Analysis for Model A

The regression equation is
gpa = 0.162 + 0.00201 satmath + 0.00125 satverb +
0.189 hsmath + 0.088 hsengl

Predictor	Coef	Stdev	t-ratio	p
Constant	0.1615	0.4375	0.37	0.717
satmath	0.0020102	0.0005844	3.44	0.004
satverb	0.0012522	0.0005515	2.27	0.038
hsmath	0.18944	0.09187	2.06	0.057
hsengl	0.0876	0.1765	0.50	0.627

s = 0.2685 R-sq = 85.3% R-sq(adj) = 81.4% (4.9.27)

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	4	6.2643	1.5661	21.72	0.000
Error	15	1.0815	0.0721		
Total	19	7.3458			

From this the director of admissions can conclude, with 90% confidence, that σ_B is between 0.296 and 0.528 grade point unit. Thus she is led to conclude that more information is necessary to decide whether or not model B is *adequate* for predicting Y.

The standard deviation σ_A for model A is estimated to be 0.2685 (see (4.9.27)), and a two-sided 90% confidence interval for σ_A is given by

$$C[0.208 \leq \sigma_A \leq 0.386] = 0.90$$

From this interval the director can conclude, with 90% confidence, that σ_A is between 0.208 and 0.386 grade point unit. Thus, according to the criterion of adequacy stated earlier, she might conclude that model A is *adequate* for predicting Y.

A point estimate of σ_B/σ_A is $\hat{\sigma}_B/\hat{\sigma}_A = 0.3771/0.2685 = 1.404$, i.e., σ_B is estimated to be about 1.404 times as large as σ_A. ■

Tests for $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$ and $\rho^2_{Y(X_1,\ldots,X_k)}$

When comparing the performances of the two regression functions, $\mu_Y^{(A)}(x_1, \ldots, x_k)$ in (4.9.2) and $\mu_Y^{(B)}(x_1, \ldots, x_m)$ in (4.9.3), it is common to formulate this as a hypothesis testing problem. The null and the alternative hypotheses considered are

$$\begin{aligned} \text{NH: } & \rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} = 0 \\ & \text{or, equivalently, } \sigma_{Y|X_1,\ldots,X_k} = \sigma_{Y|X_1,\ldots,X_m} \\ & \text{or, equivalently, } \beta_{m+1} = \cdots = \beta_k = 0 \end{aligned}$$

versus (4.9.28)

$$\begin{aligned} \text{AH: } & \rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m} > 0 \\ & \text{or, equivalently, } \sigma_{Y|X_1,\ldots,X_k} < \sigma_{Y|X_1,\ldots,X_m} \\ & \text{or, equivalently, at least one of } \beta_{m+1}, \ldots, \beta_k \text{ is nonzero} \end{aligned}$$

This test procedure is given in Box 4.9.3 and is valid if assumptions (A) or (B) hold.

BOX 4.9.3 **Hypothesis Test for $\rho^2_{Y(X_{m+1},\ldots,X_k)|X_1,\ldots,X_m}$**

For a size α test of the NH versus AH in (4.9.28), compute

$$F_C = \frac{[SSE(X_1, \ldots, X_m) - SSE(X_1, \ldots, X_k)]/(k-m)}{MSE(X_1, \ldots, X_k)}$$

The P-value for the test is the value of α for which $F_C = F_{1-\alpha:k-m,n-k-1}$. The quantities $SSE(X_1, \ldots, X_m)$, $SSE(X_1, \ldots, X_k)$, and $MSE(X_1, \ldots, X_k)$ can be obtained by regressing Y on $X_1, \ldots, X_m$ and Y on $X_1, \ldots, X_k$.

Authors' Recommendation

To determine whether the regression function $\mu_Y(x_1, \ldots, x_k)$ is adequate for predicting Y, we recommend that a confidence interval for $\sigma_{Y|X_1,\ldots,X_k}$ be examined. To decide whether the regression function $\mu_Y^{(A)}(x_1, \ldots, x_k)$ or the function $\mu_Y^{(B)}(x_1, \ldots, x_m)$ should be used to predict Y, we recommend that confidence intervals for each of the two standard deviations, $\sigma_{Y|X_1,\ldots,X_k}$ and $\sigma_{Y|X_1,\ldots,X_m}$, be examined. Additionally, the investigator may want to examine a confidence interval for $\sigma_{Y|X_1,\ldots,X_m}/\sigma_{Y|X_1,\ldots,X_k}$. Equipped with such information, the investigator is better able to make a practical decision, taking into account such factors as cost of obtaining the observations for the variables under consideration, the desired level of accuracy of the predictions, etc. Do not settle for hypothesis tests only and thus waste valuable information contained in the data.

Problems 4.9

4.9.1 Consider the data of Problem 4.8.3 given in Table 4.8.4. An investigator is studying a population of males who have lived in mountain isolation for several generations and wants to investigate the relationship between the heights Y of these males at age 18 years and the following variables.

X_1 = length at birth
X_2 = mother's height at age 18
X_3 = father's height at age 18
X_4 = maternal grandmother's height at age 18
X_5 = maternal grandfather's height at age 18
X_6 = paternal grandmother's height at age 18
X_7 = paternal grandfather's height at age 18

All heights and lengths are in inches. A simple random sample of 20 males of age 18 or more was drawn from the study population, and all the preceding information was recorded. The data are also stored in the file **age18.dat** on the data disk. Assumptions (B) are presumed to hold. The investigator will consider a prediction function to be adequate for predicting values of Y if a proportion $p = 0.90$ of the Y values in the population are within $d = 2.0$ inches of the corresponding predicted value $\mu_Y(x_1, \ldots, x_7)$. You can use the computer output (obtained using SAS) in Exhibit 4.9.2 to answer questions (a)–(d). Model 1 is the regression of Y on $X_1, X_2, X_3, X_4, X_5, X_6, X_7$. Model 2 is the regression of Y on X_1, X_2, X_3.

EXHIBIT 4.9.2

SAS Output for Problem 4.9.1

The SAS System 0:00 Saturday, Jan 1, 1994

Model: MODEL1 Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	7	133.65740	19.09391	18.955	0.0001
Error	12	12.08810	1.00734		
C Total	19	145.74550			

Root MSE	1.00366	R-square	0.9171
Dep Mean	68.53500	Adj R-sq	0.8687
C.V.	1.46445		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-78.268378	26.96236510	-2.903	0.0133
X1	1	1.371816	0.52067159	2.635	0.0218
X2	1	0.782423	0.19923906	3.927	0.0020
X3	1	1.051413	0.13581316	7.742	0.0001
X4	1	-0.119914	0.17173269	-0.698	0.4983
X5	1	0.091436	0.13011662	0.703	0.4956
X6	1	0.088343	0.16132682	0.548	0.5940
X7	1	-0.101743	0.15489810	-0.657	0.5237

Model: MODEL2
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	131.89367	43.96456	50.783	0.0001
Error	16	13.85183	0.86574		
C Total	19	145.74550			

Root MSE	0.93045	R-square	0.9050
Dep Mean	68.53500	Adj R-sq	0.8871
C.V.	1.35763		

EXHIBIT 4.9.2
(Continued)

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-78.232762	13.23928437	-5.909	0.0001
X1	1	1.350302	0.44744522	3.018	0.0082
X2	1	0.692465	0.16017457	4.323	0.0005
X3	1	1.102495	0.09907801	11.128	0.0001

Consider the following two regression functions:

$$\text{model A: } \mu_Y^{(A)}(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \beta_0^A + \beta_1^A x_1 + \beta_2^A x_2 + \beta_3^A x_3 + \beta_4^A x_4 + \beta_5^A x_5 + \beta_6^A x_6 + \beta_7^A x_7$$

$$\text{model B: } \mu_Y^{(B)}(x_1, x_2, x_3) = \beta_0^B + \beta_1^B x_1 + \beta_2^B x_2 + \beta_3^B x_3$$

a Compute an appropriate 80% confidence statement to help the investigator decide whether model A is adequate for predicting Y.

b Compute an appropriate 80% confidence statement to help the investigator decide whether model B is adequate for predicting Y.

c Explain the meaning of

$$\rho^2_{Y(X_4,X_5,X_6,X_7)} \quad \rho^2_{Y(X_1,X_2,X_3)} \quad \rho^2_{Y(X_4,X_5,X_6,X_7)|X_1,X_2,X_3} \quad \rho^2_{Y(X_1,X_2,X_3,X_4,X_5,X_6,X_7)}$$

Obtain point estimates for

$$\rho^2_{Y(X_1,X_2,X_3)} \quad \rho^2_{Y(X_4,X_5,X_6,X_7)|X_1,X_2,X_3} \quad \rho^2_{Y(X_1,X_2,X_3,X_4,X_5,X_6,X_7)}$$

d Estimate the ratio σ_B/σ_A.

4.10
Comparing Two Multiple Regression Models (Nonnested Case)

The problem of comparing the performances of two regression functions for predicting Y was considered in Section 4.9 for the situation where the set of predictor variables $X_1, \ldots, X_m$ used in one regression function $\mu_Y^{(B)}(x_1, \ldots, x_m)$ is a subset of the predictor variables $X_1, \ldots, X_k$ used in the other regression function $\mu_Y^{(A)}(x_1, \ldots, x_k)$. For brevity we referred to these two regression functions as model B and model A, respectively, and the fact that the set of predictor variables in model B is a subset of the set of predictor variables in model A was expressed by saying that *model B is nested in model A*.

However, in practice, situations arise where an investigator is faced with choosing between two models when neither model is nested in the other; i.e., there are predictor variables in one set that are not in the other set and vice versa, although some of the predictor variables may belong to both sets. To avoid any possible confusion in the notation, we use superscripts A and B to represent the predictor variables in the two sets, respectively. In this section, model A may not contain the full set of k predictors but contains only $r(r < k)$ of the k predictors $X_1, \ldots, X_k$. Thus the r predictor variables in model A will be denoted by $X_1^A, \ldots, X_r^A$ and the m predictor variables in model B will be denoted by $X_1^B, \ldots, X_m^B$. To reiterate, *there are predictor variables in model A that are not in model B, and there are predictor variables in model B that are not in model A. Additionally, there may be some predictor variables that occur in both model A and model B.* We denote the union of the two collections of predictor variables, $X_1^A, \ldots, X_r^A$ and $X_1^B, \ldots, X_m^B$, by the collection $X_1, \ldots, X_k$; i.e., $X_1, \ldots, X_k$ include all the predictor variables that are in model A or model B. The choice of a model (model A or model B) for predicting Y depends not only on how good one prediction function is relative to the other but also on many other things, including costs involved in using one set of predictor variables relative to the other set, etc. We discuss this problem of comparing two nonnested models under the assumption given in Box 4.10.1 below.

BOX 4.10.1 **Assumptions for Comparing Two Nonnested Models**

Throughout this section we suppose that assumptions (B) for regression hold for $\{(Y, X_1, \ldots, X_k)\}$. Consequently, assumptions (B) also hold for $\{(Y, X_1^A, \ldots, X_r^A)\}$ and $\{(Y, X_1^B, \ldots, X_m^B)\}$. In particular, the sample data are obtained by simple random sampling.

As a consequence of this assumption, the regression function corresponding to model A is of the form

$$\mu_Y^{(A)}(x_1^A, \ldots, x_r^A) = \beta_0^A + \beta_1^A x_1^A + \cdots + \beta_r^A x_r^A \tag{4.10.1}$$

while the regression function corresponding to model B is of the form

$$\mu_Y^{(B)}(x_1^B, \ldots, x_m^B) = \beta_0^B + \beta_1^B x_1^B + \cdots + \beta_m^B x_m^B \tag{4.10.2}$$

EXAMPLE 4.10.1

To illustrate the notation of this section, consider the setup in Example 4.4.2. Suppose the director of admissions wishes to compare the performances of the regression function of Y on X_3 (HSmath) and X_4 (HSenglish) (call this model A) and the regression function of Y on X_1 (SATmath) and X_2 (SATverbal) (call this model B) as predictors of Y. Since the assumption in Box 4.10.1 implies that both three-variable populations, $\{(Y, X_3, X_4)\}$ and $\{(Y, X_1, X_2)\}$, satisfy (population) assumptions (B), the regression function of Y on X_3, X_4 is of the form

$$\mu_Y^{(A)}(x_3, x_4) = \beta_0^A + \beta_3^A x_3 + \beta_4^A x_4$$

and the regression function of Y on X_1, X_2 is of the form

$$\mu_Y^{(B)}(x_1, x_2) = \beta_0^B + \beta_1^B x_1 + \beta_2^B x_2$$

Here the value of r is 2 and the value of m is also 2. The predictor variables in model A are $\{X_1^A, \ldots, X_r^A\} = \{X_3, X_4\}$, while the predictor variables in model B are $\{X_1^B, \ldots, X_m^B\} = \{X_1, X_2\}$. Since neither model is *nested* in the other, the discussions of Section 4.9 do not apply. You must use the discussions in this section. Observe that the union of the two sets of predictor variables has $k = 4$ predictors and is the set $\{X_1, X_2, X_3, X_4\}$.

If on the other hand the director of admissions were interested in comparing the model

$$\mu_Y^{(A)}(x_1, x_3, x_4) = \beta_0^A + \beta_1^A x_1 + \beta_3^A x_3 + \beta_4^A x_4$$

with the model

$$\mu_Y^{(B)}(x_1, x_2) = \beta_0^B + \beta_1^B x_1 + \beta_2^B x_2$$

then we would have $r = 3$ and $m = 2$. The predictor variables in model A would be $\{X_1^A, \ldots, X_r^A\} = \{X_1, X_3, X_4\}$, while those in model B would be $\{X_1^B, \ldots, X_m^B\} = \{X_1, X_2\}$. The union of the two sets of predictor variables is the set $\{X_1, X_2, X_3, X_4\}$ so that $k = 4$. Note that X_1 appears in both sets of variables. ■

For notational convenience, let σ_A denote $\sigma_{Y|X_1^A,\ldots,X_r^A}$, the subpopulation standard deviation for model A, and let σ_B denote $\sigma_{Y|X_1^B,\ldots,X_m^B}$, the subpopulation standard deviation for model B. If $\sigma_A < \sigma_B$, then the regression function $\mu_Y^{(A)}(x_1^A, \ldots, x_r^A)$ is a better predictor of Y than the regression function $\mu_Y^{(B)}(x_1^B, \ldots, x_m^B)$. On the other hand, if $\sigma_A > \sigma_B$, then the regression function $\mu_Y^{(B)}(x_1^B, \ldots, X_m^B)$ is a better predictor of Y than the regression function $\mu_Y^{(A)}(x_1^A, \ldots, x_r^A)$. In practice it is highly unlikely that the two standard deviations will be exactly equal. So the investigator is actually interested in knowing *how much bigger or smaller* σ_A is than σ_B. The obvious approach is to examine the point estimates and confidence intervals for both σ_A and σ_B. However, one could additionally calculate a confidence interval for the *ratio* σ_B/σ_A and make a practical decision based on these results.

A procedure for computing a two-sided confidence interval for σ_B/σ_A with confidence coefficient greater than or equal to $1 - \alpha$ is given in Box 4.10.2. This procedure uses the Bonferroni method.

BOX 4.10.2 Two-Sided Confidence Intervals for σ_B/σ_A Using the Bonferroni Method

1 Let r and m denote the number of predictors in models A and B, respectively, and let n be the number of sample observations.

2 Regress Y on the predictors in model A and obtain the sum of squared errors $SSE(A)$.

3 Regress Y on the predictors in model B and obtain the sum of squared errors $SSE(B)$.

4 Compute a $1-\alpha/2$ two-sided confidence interval for σ_A, which is given by

$$C[L_A \le \sigma_A \le U_A] = 1-\alpha/2$$

where

$$L_A = \sqrt{\frac{SSE(A)}{\chi^2_{1-\alpha/4:n-r-1}}} \quad \text{and} \quad U_A = \sqrt{\frac{SSE(A)}{\chi^2_{\alpha/4:n-r-1}}}$$

5 Compute a $1-\alpha/2$ two-sided confidence interval for σ_B, which is given by

$$C[L_B \le \sigma_B \le U_B] = 1-\alpha/2$$

where

$$L_B = \sqrt{\frac{SSE(B)}{\chi^2_{1-\alpha/4:n-m-1}}} \quad \text{and} \quad U_B = \sqrt{\frac{SSE(B)}{\chi^2_{\alpha/4:n-m-1}}}$$

6 We have the following confidence statement for σ_B/σ_A:

$$C[L_B/U_A \le \sigma_B/\sigma_A \le U_B/L_A] \ge 1-\alpha$$

7 Equivalently, a confidence statement for σ_A/σ_B is given by

$$C[L_A/U_B \le \sigma_A/\sigma_B \le U_A/L_B] \ge 1-\alpha$$

8 If only a one-sided confidence bound is needed, then either

$$C[L_B/U_A \le \sigma_B/\sigma_A] \ge 1-\alpha/2$$

or

$$C[\sigma_B/\sigma_A \le U_B/L_A] \ge 1-\alpha/2$$

may be used.

We illustrate the procedure described in Box 4.10.2 in Example 4.10.2.

EXAMPLE 4.10.2

In Example 4.4.2, suppose the director of admissions wants to determine how much better, or worse, the set of predictors $\{X_3, X_4\}$ is for predicting Y = GPA than the set of predictors $\{X_1, X_2\}$. As required in Box 4.10.1, we suppose that (population) assumptions (B) hold for the five-variable population $\{(Y, X_1, X_2, X_3, X_4)\}$. Hence the two regression functions being compared are of the form

$$\mu_Y^{(A)}(x_3, x_4) = \beta_0^A + \beta_3^A x_3 + \beta_4^A x_4 \qquad \textbf{(4.10.3)}$$

and

$$\mu_Y^{(B)}(x_1, x_2) = \beta_0^B + \beta_1^B x_1 + \beta_2^B x_2 \qquad \textbf{(4.10.4)}$$

For this purpose we compute a confidence interval for the ratio σ_B/σ_A. If model A is judged to be better than model B, then the director of admissions will *consider* the

possibility of not requiring applicants to submit their SAT scores (although such a decision would perhaps be based on more elaborate studies).

By examining Box 4.10.2, we see that to compute confidence intervals for σ_B/σ_A we need $SSE(X_3, X_4)$ and $SSE(X_1, X_2)$, the sum of squared errors for the models in (4.10.3) and (4.10.4). These sums of squares can be obtained from the ANOVA tables for these models. We have obtained them using SAS and they are given in Exhibit 4.10.1 and Exhibit 4.10.2, respectively.

We obtain $SSE(A) = SSE(X_3, X_4) = 2.41803$, $SSE(B) = SSE(X_1, X_2) = 1.38838$, $\hat{\sigma}_A = 0.37714$, and $\hat{\sigma}_B = 0.28578$. The table-values needed to calculate L_A, U_A, L_B, and U_B in Box 4.10.2 can be obtained from Table T-3 in Appendix T, although interpolation may be required for some values of α and/or degrees of freedom.

In this example, for a 95% two-sided confidence interval for σ_B/σ_A, we need the values of $\chi^2_{\alpha/4:n-r-1} = \chi^2_{0.0125:17}$ and $\chi^2_{1-\alpha/4:n-r-1} = \chi^2_{0.9875:17}$ for computing L_A and U_A, respectively. These are also the table-values for computing L_B and U_B in this example because $n - m - 1$ is also equal to 17. These table-values (obtained from SAS) are

$$\chi^2_{0.9875:17} = 32.644 \qquad \text{and} \qquad \chi^2_{0.0125:17} = 6.664$$

so the value of L_A is 0.272, and the value of U_A is 0.602. We also have $L_B = 0.206$ and $U_B = 0.456$. From these we obtain the following confidence statement:

$$C[0.342 \leq \sigma_B/\sigma_A \leq 1.678] \geq 0.95 \tag{4.10.5}$$

Thus there is no clear-cut evidence indicating the superiority of one model over the other. The director of admissions will have to decide, based on the confidence interval in (4.10.5), either

- that the ratio σ_B/σ_A is close enough to 1 that for this problem the two models can be considered to be equally good for predicting GPA, or
- that the sample size is not large enough to determine, sufficiently precisely, the amount by which σ_A is larger or smaller than σ_B, and additional data are required for making a decision. ■

E X H I B I T 4.10.1
SAS Output for Model A in Example 4.10.2

The SAS System 0:00 Saturday, Jan 1, 1994

Dependent Variable: GPA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	4.92779	2.46389	17.322	0.0001
Error	17	2.41803	0.14224		
C Total	19	7.34582			

Root MSE	0.37714	R-square	0.6708
Dep Mean	2.59300	Adj R-sq	0.6321
C.V.	14.54467		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-0.340013	0.56441815	-0.602	0.5548
HSMATH	1	0.417116	0.10544213	3.956	0.0010
HSENGL	1	0.579021	0.18573410	3.117	0.0063

EXHIBIT 4.10.2
SAS Output for Model B in Example 4.10.2

The SAS System 0:00 Saturday, Jan 1, 1994

Dependent Variable: GPA

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	5.95744	2.97872	36.473	0.0001
Error	17	1.38838	0.08167		
C Total	19	7.34582			

Root MSE	0.28578	R-square	0.8110
Dep Mean	2.59300	Adj R-sq	0.7888
C.V.	11.02117		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.507142	0.26672665	1.901	0.0743
SATMATH	1	0.002606	0.00044323	5.879	0.0001
SATVERB	1	0.001574	0.00055547	2.834	0.0115

Problems 4.10

4.10.1 Consider the GPA data of Example 4.4.2 where we presume that assumptions (B) are valid. Exhibits 4.10.3–4.10.6 give MINITAB computer outputs for the following regression functions.

1 Model A: $\mu_Y^{(A)}(x_1) = \beta_0^A + \beta_1^A x_1$.

2 Model B: $\mu_Y^{(B)}(x_2) = \beta_0^B + \beta_2^B x_2$.

3 Model C: $\mu_Y^{(C)}(x_3) = \beta_0^C + \beta_3^C x_3$.

4 Model D: $\mu_Y^{(D)}(x_4) = \beta_0^D + \beta_4^D x_4$.

a The director of admissions asks you to determine how much better (or worse) X_1 (SATmath) is as a predictor of Y than X_3 (HSmath) is as a predictor of Y.

i Obtain 95% confidence intervals for $\sigma_{Y|X_1}$ and $\sigma_{Y|X_3}$.

ii Obtain a two-sided 90% confidence interval for $\sigma_{Y|X_1}/\sigma_{Y|X_3}$.

b The director of admissions asks you to determine how much better (or worse) X_2 (SATverbal) is as a predictor of Y than X_4 (HSenglish) is as a predictor of Y.

i Obtain 95% confidence intervals for $\sigma_{Y|X_2}$ and $\sigma_{Y|X_4}$.

ii Obtain a two-sided 90% confidence interval for $\sigma_{Y|X_2}/\sigma_{Y|X_4}$.

c Write a short report to the director of admissions summarizing the results in (a) and (b).

EXHIBIT 4.10.3

MINITAB Output for Problem 4.10.1

Regression of Y = GPA on X_1 = SATmath

```
The regression equation is
GPA = 0.967 + 0.00318 SATmath


Predictor       Coef       Stdev    t-ratio        p
Constant      0.9670      0.2496       3.87    0.001
SATmath    0.0031783   0.0004652       6.83    0.000

s = 0.3370      R-sq = 72.2%     R-sq(adj) = 70.6%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    1      5.3015      5.3015     46.68    0.000
Error        18      2.0443      0.1136
Total        19      7.3458
```

EXHIBIT 4.10.4

MINITAB Output for Problem 4.10.1

Regression of Y = GPA on X_2 = SATverb

```
The regression equation is
GPA = 1.13 + 0.00306 SATverb

Predictor       Coef       Stdev    t-ratio        p
Constant      1.1281      0.4145       2.72    0.014
SATverb    0.0030630   0.0008367       3.66    0.002

s = 0.4837      R-sq = 42.7%     R-sq(adj) = 39.5%

Analysis of Variance

SOURCE       DF          SS          MS          F        p
Regression    1      3.1350      3.1350      13.40    0.002
Error        18      4.2108      0.2339
Total        19      7.3458
```

EXHIBIT 4.10.5

MINITAB Output for Problem 4.10.1

Regression of Y = GPA on X_3 = HSmath

```
The regression equation is
GPA = 1.15 + 0.507 HSmath

Predictor       Coef       Stdev    t-ratio        p
Constant      1.1473      0.3675       3.12    0.006
HSmath        0.5066      0.1236       4.10    0.001

s = 0.4595      R-sq = 48.3%     R-sq(adj) = 45.4%

Analysis of Variance

SOURCE       DF          SS          MS          F        p
Regression    1      3.5454      3.5454      16.79    0.001
Error        18      3.8004      0.2111
Total        19      7.3458
```

EXHIBIT 4.10.6
MINITAB Output for Problem 4.10.1
Regression of Y = GPA on X_4 = HSenglish

```
The regression equation is
GPA = 0.249 + 0.779 HSengl

Predictor       Coef      Stdev    t-ratio       p
Constant      0.2487     0.7332       0.34   0.738
HSengl        0.7790     0.2407       3.24   0.005

s = 0.5079      R-sq = 36.8%     R-sq(adj) = 33.3%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      2.7019      2.7019      10.47   0.005
Error        18      4.6439      0.2580
Total        19      7.3458
```

4.11
Lack-of-Fit Analysis

In Sections 4.9 and 4.10 we discussed the important problem of determining which of *two* regression functions is better for predicting Y. The two regression functions being compared used different sets of predictor variables, say A and B. Both model A and model B were multiple linear regression models. For instance, in Example 4.9.4 we compared the regression function

$$\mu_Y^{(A)}(x_1, x_2, x_3, x_4) = \beta_0^A + \beta_1^A x_1 + \beta_2^A x_2 + \beta_3^A x_3 + \beta_4^A x_4$$

with the regression function

$$\mu_Y^{(B)}(x_3, x_4) = \beta_0^B + \beta_3^B x_3 + \beta_4^B x_4$$

Here A is the set $\{X_1, X_2, X_3, X_4\}$, and B is the set $\{X_3, X_4\}$ (B is nested in A). In Example 4.10.2 we compared the regression function

$$\mu_Y^{(A)}(x_3, x_4) = \beta_0^A + \beta_3^A x_3 + \beta_4^A x_4$$

with the regression function

$$\mu_Y^{(B)}(x_1, x_2) = \beta_0^B + \beta_1^B x_1 + \beta_2^B x_2$$

Here A is the set $\{X_3, X_4\}$ and B is the set $\{X_1, X_2\}$ (neither set is nested in the other). In both examples the regression functions being compared have known forms; in fact, they are both multiple *linear* regression functions.

In many real problems investigators do not know the population *regression* function $\mu_Y(x)$ or even its *form*. In such instances the investigator may want to find a *prediction* function whose mathematical form is reasonably simple and which

approximates the true, but unknown, regression function adequately for the problem at hand. It is very unlikely that population regression functions are simple functions of the form $\beta_0 + \beta_1 x$ or $\beta_0 + \beta_1 x^2$, etc., but it is often the case that they can be well approximated by simple functions such as these. Figures 4.11.1–4.11.3 illustrate the point.

FIGURE 4.11.1

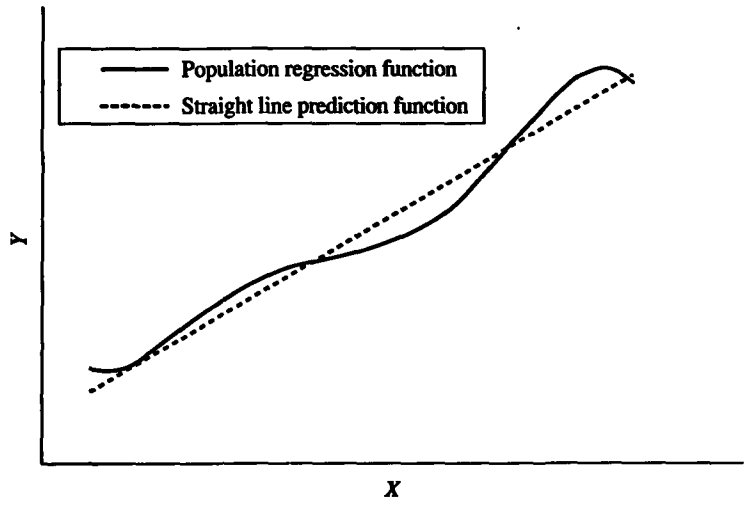


FIGURE 4.11.2

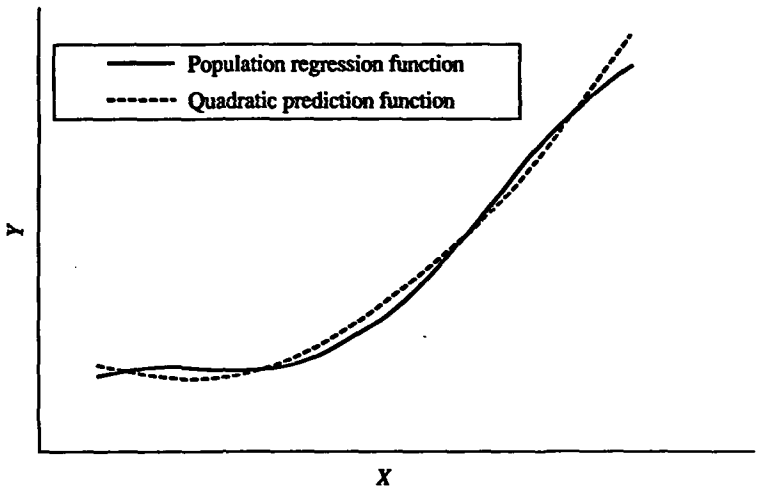
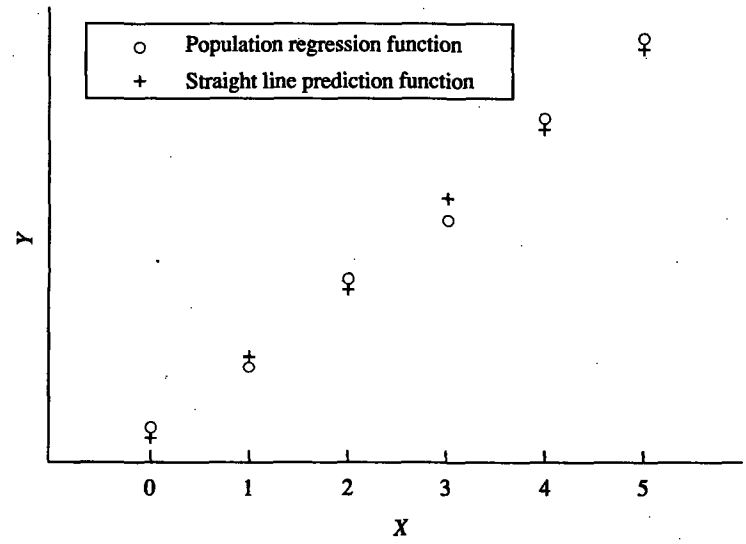


FIGURE 4.11.3



In Figure 4.11.1 a *linear* prediction function may be an adequate approximation to the regression function, and in Figure 4.11.2 a *quadratic* prediction model may be a satisfactory approximation to the regression function. In Figure 4.11.3 the population regression function is defined only at isolated values of the predictor variable X (for example when X is the number of children in a household), so the graph of the population regression function is not a continuous curve. In any event, since the population regression function $\mu_Y(x)$ can perhaps never be known exactly, we replace it with a function $P_Y(x)$ whose *form* is known and which is an adequate approximation to the population regression function. In these cases we are interested in the difference

$$\mu_Y(x) - P_Y(x)$$

to determine if indeed the function $P_Y(x)$ is an adequate approximation to the true unknown population regression function $\mu_Y(x)$, at least for the X values of interest in the investigation. When sample data are available, plots of the sample data are often useful in developing suitable classes of functions to consider in an attempt to find such an approximation to the regression function. We illustrate with two examples.

EXAMPLE 4.11.1

In using a regression function for predicting blood pressure (Y) as a function of age (X) for men between the ages of 20 and 50 of a certain ethnic background, an investigator is not sure what the regression function is. The conjecture is that the function $P_Y(x) = \beta_0 + \beta_1 x$ will provide an adequate approximation, and data are collected to check whether this is indeed the case. ■

EXAMPLE 4.11.2

An investigator wants to find a prediction function for predicting Y, the first-year maintenance cost of a new car, using the predictor variable X, the miles the car is driven the first year. He has reason to believe that the regression function $\mu_Y(x)$ of Y on X can be adequately approximated by a quadratic function of X of the form

$$P_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

Data are collected to see whether this is in fact an adequate model to use for predicting Y, the first-year maintenance cost. ■

When there is only one predictor variable, straight line functions of the form

$$P_Y(x) = \beta_0 + \beta_1 x$$

are useful in many situations. If the plot of the sample data suggests that a straight line function may serve as a good approximation to the regression function, then an obvious question of interest to the investigator is: What is the difference between the true unknown *regression* function $\mu_Y(x)$ and the proposed straight line function $P_Y(x) = \beta_0 + \beta_1 x$?

In the rest of this section we explain a procedure for answering this question.

Lack-of-Fit Analysis for Straight Line Prediction Functions

Suppose that an investigator who is interested in studying the relationship between a response variable Y and a predictor variable X for $a \leq X \leq b$ obtains a sample of size n by simple random sampling or by sampling with preselected X. Let $x_1, \ldots, x_m$ denote the distinct x values in the chosen sample. Suppose (conceptually) that the entire subpopulation corresponding to each x_i in the sample is available. We can then calculate the corresponding subpopulation means $\mu_Y(x_1), \ldots, \mu_Y(x_m)$, which we denote by $\mu_1, \ldots, \mu_m$, respectively, for ease of notation. Figure 4.11.1 shows one example of what the graph for $\mu_Y(x)$ may look like; it is obtained by plotting the mean of Y for each subpopulation corresponding to each distinct value of X in the sample. These subpopulation means μ_i need not lie exactly on a straight line, but they may lie approximately on a straight line (e.g., Figures 4.11.1 and 4.11.3). If this is so, it seems reasonable to use the least squares straight line fitted to the points $(x_1, \mu_1), \ldots, (x_m, \mu_m)$ as an approximation to the true regression function $\mu_Y(x)$. We denote this least squares straight line function by

$$P_Y(x) = \beta_0 + \beta_1 x \tag{4.11.1}$$

where

$$\beta_1 = \frac{\sum_{i=1}^m (x_i - \bar{x})(\mu_i - \bar{\mu})}{\sum_{i=1}^m (x_i - \bar{x})^2} \quad \text{and} \quad \beta_0 = \bar{\mu} - \beta_1 \bar{x} \tag{4.11.2}$$

The function in (4.11.1), in general, is not the population regression function of Y on X (unless it so happens that the population regression function is truly a straight line), but it is the least squares straight line approximation to the regression function

at the points $x_1, \ldots, x_m$. We refer to the function $P_Y(x)$ as the *proposed prediction function*. Because the μ_i are not known, β_0 and β_1 are also unknown, but we can estimate them by collecting appropriate sample data and using the fact that $\bar{y}_i$ is an estimate of μ_i for $i = 1, \ldots, m$.

Let θ_i, $i = 1, \ldots, m$, denote the differences between the unknown population regression function $\mu_Y(x)$ and the proposed prediction function $P_Y(x)$ at the sample points $x_1, \ldots, x_m$, respectively. Thus

$$\theta_i = \mu_Y(x_i) - P_Y(x_i) = \mu_Y(x_i) - [\beta_0 + \beta_1 x_i] \qquad \textbf{(4.11.3)}$$

The quantities $\theta_1, \ldots, \theta_m$ are called lack-of-fit constants. If the θ_i are small enough so that the investigator can regard them as being negligible for the problem under consideration, then the proposed straight line function will be an adequate approximation to the unknown population regression function (which is the *best* prediction function), at least for the values of X in the sample. For this reason we want to investigate the lack-of-fit constants $\theta_1, \ldots, \theta_m$.

Remarks Traditionally, statisticians and practitioners have examined the differences $\mu_Y(x) - P_Y(x)$ by performing the following hypothesis test:

$$\text{NH: } \mu_Y(x) = \beta_0 + \beta_1 x \qquad \text{against} \qquad \text{AH: } \mu_Y(x) \neq \beta_0 + \beta_1 x \qquad \textbf{(4.11.4)}$$

If NH is rejected at level α, then the investigator might conclude that the *proposed* straight line prediction function $P_Y(x) = \beta_0 + \beta_1 x$ is not the *true* regression function. The test in (4.11.4) is often called a **lack-of-fit test.** The difficulty with this test is that the result of the test does not shed any light on the actual magnitude of the difference between the proposed prediction function and the unknown population regression function. The differences (the values of the θ_i), even if detected by the test (i.e., if NH is rejected), may be negligible for the practical problem being investigated. On the other hand, if the NH is not rejected, it does not imply that the proposed prediction function is indeed the true regression function, or even that it is a good prediction function for the problem. Thus a hypothesis test for *lack-of-fit* is of very little value when practical decisions are to be made. We therefore proceed to describe methods for obtaining point and confidence interval estimates of the lack-of-fit constants $\theta_1, \ldots, \theta_m$. This information can be used by the investigator to decide whether the differences between the proposed prediction function $P_Y(x) = \beta_0 + \beta_1 x$ and the true, but unknown, regression function $\mu_Y(x)$ are negligible for all practical purposes for the problem under study.

Estimation and Confidence Intervals for Lack-of-Fit Constants for Straight Line Prediction Functions

To obtain point and confidence interval estimates for the lack-of-fit constants $\theta_1, \ldots, \theta_m$, we must obtain sample data from the two-variable population $\{(Y, X)\}$ and make some assumptions. We make the assumptions given in Box 4.11.1.

BOX 4.11.1 (Straight Line) Lack-of-Fit Assumptions

1. The form of the regression function $\mu_Y(x)$ of Y on X is unknown.
2. The investigator is interested in determining whether the straight line function $P_Y(x) = \beta_0 + \beta_1 x$, which is the least squares approximation to the true regression function at m ($m > 2$) distinct points $x_1, \ldots, x_m$ between a and b (see Figures 4.11.1 and 4.11.3), provides a good approximation to $\mu_Y(x)$. The points $(x_1, \ldots, x_m)$ are preselected by the investigator.
3. The subpopulation of Y values determined by $X = x_i$ is a Gaussian population with mean $\mu_Y(x_i)$ (written μ_i for short) and standard deviation $\sigma_Y(x_i)$ (written simply as σ_i), both of which are unknown.
4. $\sigma_1 = \sigma_2 = \cdots = \sigma_m$ and their common value is denoted by σ.
5. From each subpopulation in (3), determined by $X = x_i$, a simple random sample of n_i items is selected. The Y values of these sample items are denoted by $y_{i,1}, y_{i,2}, \ldots, y_{i,n_i}$. Furthermore, $n_i \geq 1$ for all i, and $n_i > 1$ for at least one value of i. The sample sizes $n_1, \ldots, n_m$ are selected by the investigator. The total number of observations in the sample is $n = n_1 + \cdots + n_m$.

Remark Although part (5) requires only that $n_i > 1$ for *at least* one value of i, it is desirable to choose the values of n_i so that the quantity $\sum_{i=1}^{m}(n_i - 1) = (n - m)$ is not too small because, as we see later, the estimate of σ is based on $(n - m)$ degrees of freedom.

TABLE 4.11.1

X Values	Y Values	Mean of Y Values	Estimate of μ_i	Estimate of σ_i
x_1	$y_{1,1}, y_{1,2}, \ldots, y_{1,n_1}$	$\bar{y}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} y_{1,j}$	$\hat{\mu}_1 = \bar{y}_1$	$\hat{\sigma}_1 = \sqrt{\frac{\sum_{j=1}^{n_1}(y_{1,j} - \bar{y}_1)^2}{(n_1 - 1)}}$
x_2	$y_{2,1}, y_{2,2}, \ldots, y_{2,n_2}$	$\bar{y}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} y_{2,j}$	$\hat{\mu}_2 = \bar{y}_2$	$\hat{\sigma}_2 = \sqrt{\frac{\sum_{j=1}^{n_2}(y_{2,j} - \bar{y}_2)^2}{(n_2 - 1)}}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
x_m	$y_{m,1}, y_{m,2}, \ldots, y_{m,n_m}$	$\bar{y}_m = \frac{1}{n_m}\sum_{j=1}^{n_m} y_{m,j}$	$\hat{\mu}_m = \bar{y}_m$	$\hat{\sigma}_m = \sqrt{\frac{\sum_{j=1}^{n_m}(y_{m,j} - \bar{y}_m)^2}{(n_m - 1)}}$

A schematic representation of the sample data is given in Table 4.11.1. This table also displays *estimates* of the subpopulation means μ_i and standard deviations σ_i.

The point estimates of β_0 and β_1 we use are (substituting y_i for μ_i in (4.11.2))

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(\bar{y}_i - \bar{y})}{\sum_{i=1}^{m}(x_i - \bar{x})^2} \quad \text{and} \quad \hat{\beta}_0 = \bar{y} - \beta_1\bar{x} \tag{4.11.5}$$

where

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{i,j} \qquad \bar{y} = \frac{\bar{y}_1 + \cdots + \bar{y}_m}{m} \qquad \bar{x} = \frac{x_1 + \cdots + x_m}{m} \tag{4.11.6}$$

The point estimate of

$$P_Y(x) = \beta_0 + \beta_1 x$$

is

$$\hat{P}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

and the point estimate of θ_i is

$$\hat{\theta}_i = \bar{y}_i - \hat{P}_Y(x_i) = \bar{y}_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \quad \text{for} \quad i = 1, 2, \ldots, m \tag{4.11.7}$$

where $\hat{\beta}_0$ and $\hat{\beta}_1$ are computed by (4.11.5); i.e., by regressing the means $\bar{y}_1, \ldots, \bar{y}_m$ on $x_1, \ldots, x_m$, the distinct x values. Thus

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\bar{\boldsymbol{y}} \tag{4.11.8}$$

where

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} \qquad \boldsymbol{A} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_m \end{bmatrix} \qquad \bar{\boldsymbol{y}} = \begin{bmatrix} \bar{y}_1 \\ \bar{y}_2 \\ \vdots \\ \bar{y}_m \end{bmatrix} \tag{4.11.9}$$

Note If we regress the means $\bar{y}_i$ on the distinct x_i values for $i = 1, \ldots, m$, then the $\hat{\theta}_i$ in (4.11.7) are obtained by the same formulas by which the residuals $\hat{e}_i$ were computed in (4.4.12).

We can obtain confidence intervals for $\theta_1, \ldots, \theta_m$ such that *all m intervals are simultaneously correct with at least* $(1 - \alpha)$ *confidence*. The confidence interval for θ_i is of the same form as in (4.6.1), viz.,

$$\hat{\theta}_i - \text{(table-value)}\, SE(\hat{\theta}_i) \le \theta_i \le \hat{\theta}_i + \text{(table-value)}\, SE(\hat{\theta}_i) \tag{4.11.10}$$

where

$$SE(\hat{\theta}_i) = \hat{\sigma}\sqrt{v_{ii}} \tag{4.11.11}$$

The quantity v_{ii} in (4.11.11) is the ith diagonal element of the matrix $\boldsymbol{V}$, which is given by

$$\boldsymbol{V} = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}^T \tag{4.11.12}$$

with

$$\boldsymbol{Q} = \boldsymbol{I} - \boldsymbol{A}(\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T \tag{4.11.13}$$

where $\boldsymbol{I}$ is an $m \times m$ identity matrix and

$$\boldsymbol{D} = \begin{bmatrix} 1/n_1 & 0 & 0\ldots & 0 \\ 0 & 1/n_2 & 0\ldots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 1/n_m \end{bmatrix} \tag{4.11.14}$$

Thus $\boldsymbol{D}$ is a diagonal matrix with the ith diagonal element equal to $1/n_i$. Also,

$$\hat{\sigma}^2 = \frac{\sum(n_i - 1)\hat{\sigma}_i^2}{\sum(n_i - 1)} \tag{4.11.15}$$

where

$$\hat{\sigma}_i^2 = \frac{1}{(n_i - 1)} \sum_{i=1}^{n_i} (y_{i,j} - \bar{y}_i)^2 \tag{4.11.16}$$

for those x_i with $n_i > 1$ (we take $\hat{\sigma}_i^2$ to be zero if $n_i = 1$). The table-value is the smaller of the two values

$$t_{1-\alpha/2m:dfd} \quad \text{and} \quad \sqrt{(dfn)F_{1-\alpha:dfn,dfd}}$$

with $dfn = m - 2 =$ degrees of freedom for the numerator, and $dfd = n - m =$ degrees of freedom for the denominator. The table-value $t_{1-\alpha/2m:dfd}$ can be obtained from Table T-4 in Appendix T, and $F_{1-\alpha:dfn,dfd}$ can be obtained from Table T-5, also in Appendix T.

The quantity $\sum(n_i - 1)\hat{\sigma}_i^2$ in the numerator of (4.11.15) is usually referred to as the **sum of squares for pure error** (denoted by *SS(Pure error)*), and the quantity $\sum(n_i - 1)$ in the denominator of (4.11.15) is called the **degrees of freedom for pure error** denoted by *df(Pure error)*. The estimate of σ^2 is called the **mean square for pure error** and is denoted by *MS(Pure error)*. Thus we have

$$SS(Pure\ error) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i}(y_{i,j} - \bar{y}_i)^2}{\sum_{i=1}^{m}(n_i - 1)} = (n - m)\hat{\sigma}^2 \tag{4.11.17}$$

$$df(Pure\ error) = n - m \tag{4.11.18}$$

and

$$MS(Pure\ error) = \hat{\sigma}^2 \tag{4.11.19}$$

We illustrate these computations in Example 4.11.3.

EXAMPLE 4.11.3

Consider Example 4.11.1 where an investigator is studying the relationship between age and blood pressure of men of certain ethnic background. Suppose the investigator preselects five distinct values of age, chooses several men using simple random sampling from each of the five selected age groups, and records their ages (x_i values) and blood pressures (y_i values). Using these sample data, we want to check whether $P_Y(x) = \beta_0 + \beta_1 x$ is close enough to the unknown regression function $\mu_Y(x)$ so that $P_Y(x)$ can be used to predict blood pressure using age. The data appear in Table 4.11.2 and in the file **bp.dat** on the data disk. There are five subpopulations represented in the sample, so $m = 5$. Also

$$n = 25 \quad n_1 = 4 \quad n_2 = 5 \quad n_3 = 6 \quad n_4 = 6 \quad n_5 = 4$$

The distinct values of x are 25, 30, 35, 40, and 45. The estimated values of the subpopulation means corresponding to the distinct values of x are

$$\bar{y}_1 = 108.250 \quad \bar{y}_2 = 114.400 \quad \bar{y}_3 = 121.833 \quad \bar{y}_4 = 134.167 \quad \bar{y}_5 = 142.000$$

respectively.

The estimated values for the corresponding subpopulation standard deviations are $\hat{\sigma}_1 = 3.775$, $\hat{\sigma}_2 = 2.191$, $\hat{\sigma}_3 = 5.456$, $\hat{\sigma}_4 = 2.401$, and $\hat{\sigma}_5 = 3.830$. So the estimate of σ is

$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{5}(n_i - 1)\hat{\sigma}_i^2}{\sum_{i=1}^{5}(n_i - 1)}}$$

$$= \sqrt{\frac{3(3.775)^2 + 4(2.191)^2 + 5(5.456)^2 + 5(2.401)^2 + 3(3.830)^2}{3+4+5+5+3}} = 3.766$$

TABLE 4.11.2

Blood Pressure Data

Item	Blood Pressure	Age
Y	*X*	
1	104	25
2	107	25
3	113	25
4	109	25
5	114	30
6	114	30
7	114	30
8	118	30
9	112	30
10	127	35
11	125	35
12	127	35
13	117	35

Item	Blood Pressure	Age
Y	*X*	
14	114	35
15	121	35
16	132	40
17	132	40
18	133	40
19	134	40
20	136	40
21	138	40
22	141	45
23	145	45
24	145	45
25	137	45

The A matrix and the $\bar{y}$ vector in (4.11.9) are

$$A = \begin{bmatrix} 1 & 25 \\ 1 & 30 \\ 1 & 35 \\ 1 & 40 \\ 1 & 45 \end{bmatrix} \qquad \bar{y} = \begin{bmatrix} 108.250 \\ 114.400 \\ 121.833 \\ 134.167 \\ 142.000 \end{bmatrix} \tag{4.11.20}$$

A MINITAB output for calculating the regression of $\bar{y}$ on the distinct values of X appears in Exhibit 4.11.1.

From the residuals in (4.11.21) we note that $\hat{\theta}_i$ for $i = 1, \ldots, 5$ are

$$\hat{\theta}_1 = 1.57340 \quad \hat{\theta}_2 = -1.00330 \quad \hat{\theta}_3 = -2.29700 \quad \hat{\theta}_4 = 1.31030$$
$$\hat{\theta}_5 = 0.41660$$

We obtain v_{ii} for $i = 1, \ldots, 5$ from the diagonal elements of the matrix V in Exhibit 4.11.1, which was computed using the formulas in (4.11.12)–(4.11.14). We get

$$v_{11} = 0.088667 \quad v_{22} = 0.146333 \quad v_{33} = 0.141333 \quad v_{44} = 0.130333$$
$$v_{55} = 0.083333$$

To calculate the standard errors of the $\hat{\theta}_i$ we use the formula in (4.11.11) and get

$$SE(\hat{\theta}_1) = 1.12140$$
$$SE(\hat{\theta}_2) = 1.44063$$
$$SE(\hat{\theta}_3) = 1.41580$$
$$SE(\hat{\theta}_4) = 1.35959$$
$$SE(\hat{\theta}_5) = 1.08715$$

The table-value needed for computing 95% confidence intervals for θ_i is the smaller of the two values

$$\sqrt{(m-2)F_{1-\alpha:m-2,n-m}} = \sqrt{3F_{0.95:3,20}} = \sqrt{3(3.10)} = 3.05$$

and

$$t_{1-\alpha/2m:20} = t_{0.995:20} = 2.845$$

obtained from Tables T-5 and T-4 in Appendix T, respectively. Hence the required table-value is 2.845. The confidence intervals for θ_i are

$$-1.617 \le \theta_1 \le 4.764$$
$$-5.102 \le \theta_2 \le 3.095$$
$$-6.325 \le \theta_3 \le 1.731$$
$$-2.558 \le \theta_4 \le 5.178$$
$$-2.676 \le \theta_5 \le 3.509$$

EXHIBIT 4.11.1

MINITAB Output for Example 4.11.3

```
The regression equation is
ymeans = 63.0 + 1.75 x

Predictor       Coef      Stdev    t-ratio       p
Constant      63.043      4.255      14.82   0.001
x             1.7453     0.1192      14.65   0.001

s = 1.884       R-sq = 98.6%     R-sq(adj) = 98.2%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      761.55      761.55     214.54   0.001
Error         3       10.65        3.55
Total         4      772.20

  means      distinct      residuals
    of           x         =estimate
    y         values       of theta(i)
-------------------------------------
108.250         25           1.57340                    (4.11.21)
114.400         30          -1.00330
121.833         35          -2.29700
134.167         40           1.31030
142.000         45           0.41660
```

MATRIX $(A^TA)^{-1}$:

```
  5.100   -0.140
 -0.140    0.004
```

```
MATRIX D
 0.25000000  0.00000000  0.00000000  0.00000000  0.00000000
 0.00000000  0.20000000  0.00000000  0.00000000  0.00000000
 0.00000000  0.00000000  0.16666667  0.00000000  0.00000000
 0.00000000  0.00000000  0.00000000  0.16666667  0.00000000
 0.00000000  0.00000000  0.00000000  0.00000000  0.25000000

MATRIX V
 0.088667 -0.089333 -0.040667 -0.005333  0.046667
-0.089333  0.146333 -0.031333 -0.019000 -0.006667
-0.040667 -0.031333  0.141333 -0.026000 -0.043333
-0.005333 -0.019000 -0.026000  0.130333 -0.080000
 0.046667 -0.006667 -0.043333 -0.080000  0.083333
```

We can be at least 95% confident that all five of the confidence intervals are simultaneously correct. As a consequence, we can be at least 95% confident that the proposed straight line function $P_Y(x) = \beta_0 + \beta_1 x$ does not deviate from the true unknown regression function $\mu_Y(x)$ by more than 6.325 blood pressure units (absolute value of the lower bound for θ_3) *for any of the subpopulations represented in the sample.* ■

Cautionary Remark Remember that the lack-of-fit of the proposed straight line regression model has been investigated only for the subpopulations represented in the sample; i.e., for $X = x_1, X = x_2, \ldots, X = x_m$. Even if the proposed straight line function $P_Y(x)$ agrees *exactly* with the true unknown regression function $\mu_Y(x)$ for these selected subpopulations, we still cannot conclude from these data that the two functions are identical for all values of X. Keep this point in mind when making any practical conclusions based on the results of a lack-of-fit analysis.

In Section 4.11 in the laboratory manuals we show you how to use a program that we have supplied on the data disk to do the computations necessary to obtain point estimates and confidence intervals for the lack-of-fit constants θ_i.

The Traditional Lack-of-Fit Test for Straight Line Prediction Functions

The statistical test of (4.11.4) is usually referred to as a *lack-of-fit test.* Although we recommend against using only a statistical test to examine the lack-of-fit of a proposed function, it seems to be common practice among statisticians and investigators to use such a test. We therefore describe this test procedure for the sake of completeness.

The true regression function $\mu_Y(x)$ is unknown, but the investigator postulates that it is

$$P_Y(x) = \beta_0 + \beta_1 x$$

for some unknown constants β_0 and β_1. Assumptions for lack-of-fit analysis, given in Box 4.11.1, are presumed to be valid. Thus m values of X are preselected (which are denoted by $x_1, x_2, \ldots, x_m$) and for each i, n_i values of Y are obtained by simple random sampling from the subpopulation corresponding to $X = x_i$. A schematic representation of the sample is in Table 4.11.1. The test of

$$\text{NH:} \quad \mu_Y(x) = \beta_0 + \beta_1 x \text{ for some } \beta_0, \beta_1 \tag{4.11.22}$$

against

$$\text{AH:} \quad \mu_Y(x) \neq \beta_0 + \beta_1 x \text{ for any } \beta_0, \beta_1$$

is conducted as follows.

1 Compute the estimate of σ^2 as in (4.11.15). Recall that this is called the *mean square for pure error* and is denoted by *MS(Pure error)*. Also calculate *SS(Pure error)*, the *sum of squares for pure error* given in (4.11.17).

2 Obtain *SSE* from the regression of Y on X using *all* of the sample observations.

3 Compute the *sum of squares for lack-of-fit*, *SS*(*Lack-of-fit*), by the formula

$$SS(\textit{Lack-of-fit}) = SSE - SS(\textit{Pure error})$$

4 Compute the *mean square for lack-of-fit*, *MS*(*Lack-of-fit*), by the formula

$$MS(\textit{Lack-of-fit}) = SS(\text{Lack-of-fit})/(m-2)$$

5 Compute the test statistic F_C by

$$F_C = \frac{MS(\textit{Lack-of-fit})}{MS(\textit{Pure error})}$$

6 The P-value for the test is the value of α for which $F_C = F_{1-\alpha:m-2,n-m}$.

We illustrate the procedure for conducting the traditional lack-of-fit test in Example 4.11.4.

EXAMPLE 4.11.4

For the age and blood pressure problem discussed in Example 4.11.3, we carry out the traditional lack-of-fit test. To compute *SSE*, we used SAS and regressed Y on X. The relevant part of the computer output appears in Exhibit 4.11.2.

From (4.11.23) we get *SSE* = 340.5 (rounded to one decimal). The estimate of σ is $\hat{\sigma} = 3.766$ (from Example 4.11.3), so the sum of squares for pure error, using (4.11.17), is calculated to be $SS(\textit{Pure error}) = 20(3.766)^2 = 283.7$ with $df(\textit{Pure error}) = n - m = 20$. Also $MS(\textit{Pure error}) = 14.18$. Hence the sum of squares for lack-of-fit is $340.5 - 283.7 = 56.8$ with degrees of freedom $m - 2 = 3$. The mean square for lack-of-fit is $56.8/3 = 18.9$. The test statistic is $F_C = 18.9/14.18 = 1.33$ with 3 and 20 degrees of freedom. So from Table T-5 in Appendix T the P-value is between 0.1 and 0.5. Hence we do not reject NH at any of the commonly used α levels.

Although the null hypothesis is not rejected, we *cannot* conclude that the proposed straight line function $P_Y(x) = \beta_0 + \beta_1 x$ is correct. However, an examination of the confidence intervals for θ_i for $i = 1, \ldots, 5$ will help the investigator determine whether or not the deviations between the true unknown regression function and the proposed straight line function are small enough to be ignored for the problem under study. ■

EXHIBIT 4.11.2
SAS Output for Example 4.11.4

```
                          The SAS System  0.00 Saturday Jan 1, 1994

Dependent Variable: BP

                              Analysis of Variance

                              Sum of          Mean
Source                DF     Squares        Square      F Value      Prob>F

Model                  1   3337.24976   3337.24976      225.417      0.0001
Error                 23    340.51024     14.80479                              (4.11.23)
C Total               24   3677.76000

      Root MSE          3.84770     R-square       0.9074
      Dep Mean        124.36000     Adj R-sq       0.9034
      C.V.              3.09400

                          Parameter Estimates

                   Parameter       Standard    T for H0:
Variable  DF        Estimate          Error   Parameter=0    Prob > |T|

INTERCEP   1      62.310987     4.20380966        14.823        0.0001
AGE        1       1.762756     0.11740836        15.014        0.0001
```

Problems 4.11

4.11.1 The bowl-life Y (in seconds) of a breakfast cereal, which is usually eaten with milk, is defined to be the amount of time that the cereal will retain its crunchiness. This bowl-life depends on the temperature X (in degrees Celsius) of the milk. The regression function $\mu_Y(x)$ of Y on X is unknown, but the investigator postulates that the regression function of Y on X is adequately approximated by a straight line prediction function of the form

$$P_Y(x) = \beta_0 + \beta_1 x \qquad (4.11.24)$$

which is the least squares approximation to $\mu_Y(x)$ for $x = 40, 45, 50$, and 55. To analyze the lack-of-fit of the proposed model, data were collected at these preselected values of X by adding the milk to 5 ounces of the cereal under each test condition. The data are displayed in Table 4.11.3 and are also stored in the file **cereal.dat** on the data disk.

Because there are multiple observations in at least one of the four subpopulations represented in the sample, it is possible to perform a lack-of-fit analysis using the methods of Section 4.11.

Several quantities that are needed for a lack-of-fit analysis in this problem follow.

$$\bar{y} = 9.06667$$
$$\bar{x} = 47.5000$$
$$SSY = 195.653$$
$$SSX = 750.000$$
$$SXY = -257.500$$

TABLE 4.11.3
Cereal Bowl-Life Data

Sample Item Number	Bowl-Life Y (minutes)	Temperature X (°C)
1	13.8	40.0
2	14.8	40.0
3	11.1	40.0
4	11.3	40.0
5	9.7	40.0
6	8.7	40.0
7	12.5	45.0
8	12.7	45.0
9	10.9	45.0
10	10.5	45.0
11	6.5	45.0
12	7.1	45.0
13	10.5	50.0
14	10.2	50.0
15	8.7	50.0
16	8.2	50.0
17	6.5	50.0
18	5.2	50.0
19	8.7	55.0
20	8.7	55.0
21	6.8	55.0
22	6.6	55.0
23	3.6	55.0
24	4.3	55.0

$$SSE = 107.245$$
$$SS(\textit{Pure error}) = 107.130$$

The V matrix is

$$\begin{bmatrix} 0.050000 & -0.066667 & -0.016667 & 0.033333 \\ -0.066667 & 0.116667 & -0.033333 & -0.016667 \\ -0.016667 & -0.033333 & 0.116667 & -0.066667 \\ 0.033333 & -0.016667 & -0.066667 & 0.050000 \end{bmatrix}$$

a What is the value of m, the number of distinct subpopulations represented in the sample?

b What is n? What are the values of $n_1, \ldots, n_m$?

c Calculate the estimates of the subpopulation means $\mu_i = \mu_Y(x_i)$ and the estimates of the subpopulation standard deviations $\sigma_i = \sigma_Y(x_i)$ for each subpopulation represented in the sample.

d Calculate the pure error estimate of σ, and show that it is equal to $\hat{\sigma} = 2.31441$.

e Estimate the lack-of-fit constants $\theta_1, \ldots, \theta_m$.

f Show that $SE(\hat{\theta}_i)$ are

$$0.517518, \quad 0.790522, \quad 0.790522, \quad 0.517518$$

respectively.

g Show that the two-sided confidence intervals for $\theta_1, \ldots, \theta_m$, such that we have at least 90% confidence that all of them are simultaneously correct, are

Lower	Upper
−1.25265	1.10272
−1.69058	1.90730
−1.79058	1.80730
−1.21933	1.13604

Use $t_{0.9875:20} = 2.423$ and $F_{0.9:2,20} = 2.59$.

h Find a number d such that we can say with at least 90% confidence that the population regression function and the proposed linear prediction function will not differ by more than d units in any of the m subpopulations included in the sample.

i Suppose an investigator decides that the proposed prediction function in (4.11.24) is an adequate approximation of the true regression function if the differences $\theta_1, \ldots, \theta_m$ between the two functions can be shown to be less than or equal to 2 minutes. Based on the lack-of-fit analysis in (g), can the prediction function in (4.11.24) be regarded as close enough to the true regression function for the problem under study?

j Perform a traditional lack-of-fit test of the model in (4.11.24). What is your conclusion using $\alpha = .10$?

Conversation 4.11

Investigator: Good morning. Do you have time to talk to me?

Statistician: Certainly. How can I help you?

Investigator: I want to discuss lack-of-fit. Can you explain the difference between the following?

1 Checking the lack-of-fit of a straight line model and

2 Checking whether model A is better than model B for predicting Y where the two models are

$$\text{Model A: } \mu_Y^{(A)}(x) = \beta_0^A + \beta_1^A x$$
$$\text{Model B: } \mu_Y^{(B)} = \beta_0^B$$

Statistician: In the simplest terms, in (1) you are checking to determine whether a proposed model $P_Y(x) = \beta_0 + \beta_1 x$ is close enough to the true *unknown* regression function $\mu_Y(x)$ so that $P_Y(x)$ can be used in place of $\mu_Y(x)$. In (2), you are assuming model A is correct and checking to see whether it is better than model B for predicting Y.

For case (2), both models are specified (they of course both contain unknown parameters), but in case (1) you do not specify the *true* regression function $\mu_Y(x)$ because you do not know what it is. You just want to know if the model you propose, namely $P_Y(x) = \beta_0 + \beta_1 x$, can be used in place of the true unknown regression model $\mu_Y(x)$.

Investigator: One of the people I work for says that he would rather use the *test* for lack-of-fit than the simultaneous confidence intervals you propose.

Statistician: Why is that?

Investigator: He says the test is easier. All he has to do to test for lack-of-fit is compute a P-value, and on the basis of that he can reject or not reject the hypothesis that the proposed model $P_Y(x) = \beta_0 + \beta_1 x$ is equal to the unknown true model $\mu_Y(x)$. On the other hand, after the confidence intervals are obtained for the lack-of-fit constants, he has to spend a considerable amount of time determining whether any of the differences are important in his problem.

Statistician: That's exactly right. It seems to me that an investigator would be much more comfortable with the decision if he examined the confidence intervals than if he just used the P-value.

To perform a test, he would examine two hypotheses:

$$\text{NH: } \beta_0 + \beta_1 x \quad \textit{is the true model}$$

against

$$\text{AH: } \beta_0 + \beta_1 x \quad \textit{is not the true model}$$

However, if he looked at the confidence intervals, they would help him decide whether the proposed model $P_Y(x) = \beta_0 + \beta_1 x$ is adequate for his problem. It is unlikely that $P_Y(x) = \beta_0 + \beta_1 x$ is exactly the true model, but it may be close enough to be useful in a specified problem.

Investigator: I see your point, but the investigator says if NH is rejected, he will assume that the model $\beta_0 + \beta_1 x$ is not adequate for his problem, but if NH is not rejected, he will assume the model is adequate.

Statistician: You recall our previous conversation in Chapter 1 where we decided that results can be quite different if confidence intervals are used instead of tests. So if the investigator insists on using a *P*-value to make the decision, why don't you give him the *P*-value and the confidence intervals and ask him to examine them also.

Investigator: I will do that, but he may say that a test for lack-of-fit is easier to compute than simultaneous confidence intervals.

Statistician: Tell him that isn't the case if he has the MINITAB or SAS macro provided on the data disk and discussed in Section 4.11 in the laboratory manuals. This macro computes the simultaneous confidence intervals he needs.

Investigator: I'll explain that to him.

4.12 Exercises

4.12.1 While performing experiments to study the absorption of a certain drug in mice, an investigator administered a specified dose of the drug to a laboratory mouse and determined the drug concentration in blood samples drawn from the mouse at times ranging from 20 to 420 minutes. The drug concentrations C in the blood and the times T when blood was drawn are given in Exhibit 4.12.1 along with a MINITAB printout from a regression analysis of these data. The data also appear in the file **mouse.dat** on the data disk.

If we let

$$Y = log_{10}(C) \qquad X_1 = log_{10}(T) \qquad X_2 = X_1^2 = [log_{10}(T)]^2$$

the regression function is given by

$$\mu_Y(t) = \beta_0 + \beta_1 log_{10}(t) + \beta_2 [log_{10}(t)]^2$$

i.e., by

$$\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \qquad (4.12.1)$$

Population assumptions (A) are presumed to hold for $\{(Y, X_1, X_2)\}$.

EXHIBIT 4.12.1

MINITAB Output for Mouse Data

RAW AND TRANSFORMED DATA

C conc	T time	Y	X1	X2
1.02	20	0.008600	1.30103	1.69268
1.08	40	0.033424	1.60206	2.56660
1.10	60	0.041393	1.77815	3.16182
1.06	90	0.025306	1.95424	3.81906
0.95	150	-0.022276	2.17609	4.73537
0.77	210	-0.113509	2.32222	5.39270
0.60	300	-0.221849	2.47712	6.13613
0.42	420	-0.376751	2.62325	6.88144

The regression equation is

Y = - 1.31 + 1.62 X1 - 0.479 X2

Predictor	Coef	Stdev	t-ratio	p
Constant	-1.3101	0.2005	-6.53	0.001
X1	1.6222	0.2091	7.76	0.001
X2	-0.47926	0.05266	-9.10	0.000

s = 0.02441 R-sq = 98.1% R-sq(adj) = 97.4%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	0.156172	0.078086	131.03	0.000
Error	5	0.002980	0.000596		
Total	7	0.159152			

a Plot C against T.

b Plot Y against X_1.

c The estimated regression equation can be written as

$$\hat{\mu}_{log_{10}(C)}(t) = \hat{\mu}_Y(t) = -1.31 + 1.62\, log_{10}(t) - 0.479[log_{10}(t)]^2$$

Compute $\hat{\mu}_Y(30)$. On the graph in (b), superimpose the graph of $\hat{\mu}_Y(t)$.

d If β_2 is zero in the regression function in (4.12.1), it would mean that this regression function is a straight line in X_1. On the other hand, if β_2 is nonzero, then the regression function in (4.12.1) is a quadratic function of X_1. Suppose the investigator will consider the quadratic term to be negligible for this problem if β_2 is less than 0.0002 in magnitude. In that case a linear regression function

would be used for this problem. Compute an appropriate 95% confidence interval for β_2 and state what the investigator's conclusion will be.

e Suppose the investigator wants to use a statistical test to help determine whether the data provide evidence (at $\alpha = 0.05$) suggesting that the regression function in (4.12.1) is a quadratic function and not a straight line function of X_1 (i.e., $\beta_2 = 0$). Formulate an appropriate statistical test to help the investigator determine this and carry out the test. What is the P-value for this test? What is your conclusion based on this test? Compare this with your answer for part (d).

4.12.2 An organization that evaluates the performance of automobiles wants to predict the first-year maintenance cost Y of a new car as a function of the number of miles X the car will be driven. With this in mind a sample of 17 cars was selected and the owners were asked to report maintenance costs after the cars were driven a specified number of miles; i.e., the data were obtained by sampling with preselected X values to cover a wide range of miles driven. The values of X (in miles) and Y (in dollars) are recorded. The data and computer output appear in Exhibit 4.12.2 and are also stored in **car17.dat** on the data disk. Assumptions (A) are presumed valid for the population $\{(Y, X)\}$, with the regression function of Y on X given by

$$\mu_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{4.12.2}$$

which can be written as $\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ where $x_1 = x$ and $x_2 = x^2$.

If $\beta_2 = 0$ then the regression function reduces to

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{4.12.3}$$

a Plot Y against X.

b Assume that the regression function of Y on X is given in (4.12.3).

i What are the estimates of β_0, β_1, $\sigma_{Y|X}$, $\mu_Y(x)$, and $Y(x)$?

You will notice that there are nine different subpopulations represented in the sample data, with two or more observations from six of the nine subpopulations, so it is possible to carry out a lack-of-fit analysis. The investigator wants to know if the model

$$P_Y(x) = \beta_0 + \beta_1 x$$

is an adequate approximation for the regression function. Questions (ii)–(vi) pertain to this.

ii Find *SS(Pure error)*, *degrees of freedom(Pure error)*, and *MS(Pure error)*.

EXHIBIT 4.12.2
MINITAB Output for Problem 4.12.2

carno	Y mtcost	X1 miles	X2 [miles]^2
1	272	3000	9000000
2	300	5000	25000000
3	287	7000	49000000
4	327	9000	81000000
5	330	10000	100000000
6	386	14000	196000000
7	442	18000	324000000
8	522	22000	484000000
9	604	25000	625000000
10	266	3000	9000000
11	313	7000	49000000
12	336	10000	100000000
13	328	10000	100000000
14	367	14000	196000000
15	397	14000	196000000
16	483	18000	324000000
17	537	22000	484000000

Regression of Y on $X_1 = X$ and $X_2 = X^2$

The regression equation is
Y = 259 + 0.00331 X1 +0.00000042 X2.

Predictor	Coef	Stdev	t-ratio	p
Constant	258.72	12.36	20.93	0.000
X1	0.003309	0.002084	1.59	0.135
X2	0.00000042	0.00000007	5.58	0.000

s = 12.85 R-sq = 98.6% R-sq(adj) = 98.4%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	161689	80845	489.28	0.000
Error	14	2313	165		
Total	16	164002			

EXHIBIT 4.12.2
(Continued)

Regression of Y on $X_1 = X$

```
The regression equation is
Y = 201 + 0.0146 X1

Predictor       Coef       Stdev      t-ratio        p
Constant      200.68       11.57        17.35    0.000
X1          0.0146230   0.0008238        17.75    0.000

s = 22.29       R-sq = 95.5%     R-sq(adj) = 95.2%

Analysis of Variance

SOURCE        DF          SS          MS          F        p
Regression     1      156551      156551     315.12    0.000
Error         15        7452         497
Total         16      164002
```

iii The matrix V defined in (4.11.12), the estimates of the lack-of-fit constants $\hat{\theta}_i$ given in (4.11.7), and the standard errors of the $\hat{\theta}_i$ given in (4.11.11) follow:

$$V = \begin{bmatrix} 0.403773 & -0.216777 & -0.073971 & -0.154221 & -0.030178 & -0.021355 & -0.012766 & 0.009490 & 0.096004 \\ -0.216777 & 0.692293 & -0.167147 & -0.226212 & -0.104619 & -0.065611 & -0.030664 & 0.018966 & 0.099769 \\ -0.073971 & -0.167147 & 0.439303 & -0.130442 & -0.027233 & -0.021770 & -0.024194 & -0.010920 & 0.016374 \\ -0.154220 & -0.226211 & -0.130443 & 0.812562 & -0.090897 & -0.072117 & -0.065053 & -0.041275 & -0.032345 \\ -0.030178 & -0.104619 & -0.027233 & -0.090897 & 0.329122 & -0.004859 & -0.019136 & -0.016192 & -0.036008 \\ -0.021355 & -0.065611 & -0.021770 & -0.072117 & -0.004859 & 0.330090 & -0.022910 & -0.023325 & -0.098143 \\ -0.012766 & -0.030664 & -0.024194 & -0.065053 & -0.019136 & -0.022910 & 0.444380 & -0.067048 & -0.202610 \\ 0.009490 & 0.018966 & -0.010920 & -0.041275 & -0.016192 & -0.023324 & -0.067048 & 0.412542 & -0.282240 \\ 0.096004 & 0.099769 & 0.016374 & -0.032345 & -0.036008 & -0.098143 & -0.202610 & -0.282240 & 0.439199 \end{bmatrix}$$

The estimates of the lack-of-fit constants θ_i are

$$\hat{\theta}_1 = 21.9485 \quad \hat{\theta}_2 = 23.1066 \quad \hat{\theta}_3 = -6.7354$$
$$\hat{\theta}_4 = -9.5773 \quad \hat{\theta}_5 = -20.1649 \quad \hat{\theta}_6 = -27.8488$$
$$\hat{\theta}_7 = -8.3661 \quad \hat{\theta}_8 = -1.0499 \quad \hat{\theta}_9 = 28.6871$$

Show that the standard errors of the estimates $\hat{\theta}_i$ are

$$SE(\hat{\theta}_1) = 9.5429 \quad SE(\hat{\theta}_2) = 12.4956 \quad SE(\hat{\theta}_3) = 9.9540$$
$$SE(\hat{\theta}_4) = 13.5376 \quad SE(\hat{\theta}_5) = 8.6157 \quad SE(\hat{\theta}_6) = 8.6284$$
$$SE(\hat{\theta}_7) = 10.0113 \quad SE(\hat{\theta}_8) = 9.6460 \quad SE(\hat{\theta}_9) = 9.9528$$

iv Show that simultaneous confidence intervals for $\theta_1, \ldots, \theta_9$, with confidence coefficients greater than or equal to 95%, are given by (use $t_{0.9972:8} = 3.7856$ and $F_{0.95:7,8} = 3.501$)

θ	Lower	Upper
1	−13.9194	57.8164
2	−23.8593	70.0725
3	−44.1485	30.6777
4	−60.4597	41.3051
5	−52.5478	12.2180
6	−60.2795	4.5819
7	−45.9945	29.2623
8	−37.3053	35.2055
9	−8.7214	66.0956

v An investigator will use a straight line model for this problem if none of the lack-of-fit constants θ_i, $i = 1, \ldots, 9$ exceed \$75 in magnitude. Using the results from (iv), would the investigator use a straight line model?

vi Perform the traditional lack-of-fit test for a straight line model using $\alpha = 0.05$. What is your conclusion? Compare this with your conclusion in part (v).

c Assume that the regression function of Y on X is given by (4.12.2). What are the estimates of $\beta_0, \beta_1, \beta_2, \sigma_{Y|X_1,X_2}, \mu_Y(x_1, x_2)$, and $Y(x_1, x_2)$?

d Plot the estimates of the regression functions in (4.12.2) and (4.12.3) on the same graph. Also show the observed data points on this graph.

4.12.3 The height (Y) in inches of a plant during the first few days after germination is related to the temperature (X_1) in degrees Fahrenheit at which it is grown and the time (X_2) in days after germination. Assume that the regression function of Y on X_1 and X_2 is of the form

$$\mu_Y(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \qquad \textbf{(4.12.4)}$$

which can be written as

$$\mu_Y(x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where $x_3 = x_1 x_2$.

Twenty plants were included in an experiment. Each plant was grown at a prechosen temperature (60°F, 70°F, 80°F, 90°F, or 100°F) for a preselected number of days (6 days or 12 days), at the end of which its height was recorded. The data and computer output are given in Exhibit 4.12.3 and are also stored in the file **plant.dat** on the data disk. Assumptions (A) are presumed to hold, and data were obtained by preselecting the values of X_1 and X_2.

The computer output in Exhibit 4.12.4 (obtained using MINITAB) lists the values of $X_1, X_2, X_1X_2, Y, \hat{\mu}_Y(x_1, x_2)$ = fits, $\hat{e}_i$ = residuals, r_i = standardized residuals, and Gaussian scores (nscores) = $z_i^{(n)}$ for the model in (4.12.4).

EXHIBIT 4.12.3

MINITAB Output for Plant Growth Data

Y	X1	X2	X1X2
3.11	60	6	360
2.04	60	6	360
4.36	60	12	720
4.60	60	12	720
2.98	70	6	420
3.65	70	6	420
6.31	70	12	840
7.05	70	12	840
4.21	80	6	480
4.31	80	6	480
7.86	80	12	960
8.45	80	12	960
4.86	90	6	540
4.25	90	6	540
9.63	90	12	1080
9.59	90	12	1080
5.66	100	6	600
5.28	100	6	600
10.89	100	12	1200
11.23	100	12	1200

The regression equation is
Y = 1.70 - 0.0203 X1 - 0.548 X2 + 0.0151 X1X2

Predictor	Coef	Stdev	t-ratio	p
Constant	1.697	1.513	1.12	0.279
X1	-0.02030	0.01862	-1.09	0.292
X2	-0.5477	0.1595	-3.43	0.003
X1X2	0.015100	0.001963	7.69	0.000

s = 0.3724 R-sq = 98.4% R-sq(adj) = 98.1%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	3	140.149	46.716	336.84	0.000
Error	16	2.219	0.139		
Total	19	142.368			

EXHIBIT 4.12.4
Diagnostics for the Model in (4.12.4) for the Plant Growth Data

X1 temp	X2 time	X1X2 (temp)(time)	Y height	fits	residuals	stdresid	nscores
60	6	360	3.11	2.629	0.481000	1.54373	1.40377
60	6	360	2.04	2.629	-0.589000	-1.89035	-1.87129
60	12	720	4.36	4.779	-0.419000	-1.34475	-1.12690
60	12	720	4.60	4.779	-0.179000	-0.57449	-0.58740
70	6	420	2.98	3.332	-0.352000	-1.02520	-0.91718
70	6	420	3.65	3.332	0.318000	0.92618	0.91718
70	12	840	6.31	6.388	-0.078000	-0.22718	-0.18593
70	12	840	7.05	6.388	0.662000	1.92808	1.87129
80	6	480	4.21	4.035	0.175000	0.49533	0.44602
80	6	480	4.31	4.035	0.275000	0.77837	0.74198
80	12	960	7.86	7.997	-0.137000	-0.38777	-0.31325
80	12	960	8.45	7.997	0.453000	1.28219	1.12690
90	6	540	4.86	4.738	0.122000	0.35533	0.31325
90	6	540	4.25	4.738	-0.488000	-1.42130	-1.40377
90	12	1080	9.63	9.606	0.024000	0.06990	0.18593
90	12	1080	9.59	9.606	-0.016000	-0.04660	-0.06165
100	6	600	5.66	5.441	0.219000	0.70286	0.58740
100	6	600	5.28	5.441	-0.161000	-0.51672	-0.44602
100	12	1200	10.89	11.215	-0.325000	-1.04306	-0.74198
100	12	1200	11.23	11.215	0.014999	0.04814	0.06165

a Plot height versus temperature using different symbols for points corresponding to different times while using the same symbol for points corresponding to the same times.

b Plot height versus time using different symbols for points corresponding to different temperatures while using the same symbol for points corresponding to the same temperatures.

c How were the data obtained for this study—by simple random sampling or by sampling with preselected X_1, X_2 values ?

d Carry out a residual analysis for the model in (4.12.4) to examine the validity of assumptions (A) for this problem. What are your conclusions?

e Estimate $\beta_0, \beta_1, \beta_2, \beta_3, \sigma_{Y|X_1,X_2,X_3}, \mu_Y(x_1, x_2, x_3)$, and $Y(x_1, x_2, x_3)$.

f For the model in (4.12.4) test NH: $\beta_2 = \beta_3 = 0$ against AH: at least one of β_2, β_3 is nonzero. Use $\alpha = 0.05$. You may need the results from Exhibit 4.12.5. State your conclusion.

g Using the model in (4.12.4) estimate the average height in inches, at the end of 10 days, of plants that are grown at 65° F. Construct a 95% two-sided confidence interval for this mean height. Describe in words the meaning of this confidence interval.

EXHIBIT 4.12.5
MINITAB Output for Regression of Y on X_1 for the Plant Growth Data

```
The regression equation is
Y = - 3.23 + 0.116 X1

Predictor       Coef      Stdev    t-ratio       p
Constant      -3.232      2.855      -1.13   0.272
X1           0.11560    0.03514       3.29   0.004

s = 2.223       R-sq = 37.5%     R-sq(adj) = 34.1%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      53.453      53.453      10.82   0.004
Error        18      88.915       4.940
Total        19     142.368
```

4.12.4 An investigator is interested in studying how the height Y at age 18 years of a group of people who have lived in mountain isolation for several generations is related to the following variables.

X_1 = Length at birth
X_2 = Mother's height at age 18
X_3 = Father's height at age 18
X_4 = Maternal grandmother's height at age 18
X_5 = Maternal grandfather's height at age 18
X_6 = Paternal grandmother's height at age 18
X_7 = Paternal grandfather's height at age 18

All heights and lengths are in inches. A simple random sample of 20 males of age 18 or more was drawn, and all the above information was recorded. The data and computer output are given in Exhibit 4.12.6 and are also stored in the file **age18.dat** on the data disk. Assumptions (B) are presumed to hold.

EXHIBIT 4.12.6
Data and MINITAB Output for Problem 4.12.4

Sample Item Number	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7
1	67.2	19.7	60.5	70.3	65.7	69.3	65.7	67.3
2	69.1	19.6	64.9	70.4	62.6	69.6	64.6	66.4
3	67.0	19.4	65.4	65.8	66.2	68.8	64.0	69.4
4	72.4	19.4	63.4	71.9	60.7	68.0	64.9	67.1
5	63.6	19.7	65.1	65.1	65.5	65.5	61.8	70.9
6	72.7	19.6	65.2	71.1	63.5	66.2	67.3	68.6
7	68.5	19.8	64.3	67.9	62.4	71.4	63.4	69.4
8	69.7	19.7	65.3	68.8	61.5	66.0	62.4	67.7
9	68.4	19.7	64.5	68.7	63.9	68.8	62.3	68.8
10	70.4	19.9	63.4	70.3	65.9	69.0	63.7	65.1
11	67.5	18.9	63.3	70.4	63.7	68.2	66.2	68.5
12	73.3	20.8	66.2	70.2	65.4	66.6	61.7	64.0
13	70.0	20.3	64.9	68.8	65.2	70.2	62.4	67.0
14	69.8	19.7	63.5	70.3	63.1	64.4	65.1	67.0
15	63.6	19.9	62.0	65.5	64.1	67.7	62.1	66.5
16	64.3	19.6	63.5	65.2	63.9	70.0	64.2	64.5
17	68.5	21.3	66.1	65.4	64.8	68.4	66.4	70.8
18	70.5	20.1	64.8	70.2	65.3	65.5	63.7	66.9
19	68.1	20.2	62.6	68.6	63.7	69.8	66.7	68.0
20	66.1	19.2	62.2	67.3	63.6	70.9	63.6	66.7

The regression of Y on $X_1, X_2, X_3, X_4, X_5, X_6, X_7$

```
The regression equation is
Y = - 78.3 + 1.37 X1 + 0.782 X2 + 1.05 X3 - 0.120 X4 +
    0.091 X5 + 0.088 X6 - 0.102 X7

Predictor       Coef      Stdev    t-ratio       p

Constant      -78.27      26.96      -2.90   0.013
X1            1.3718     0.5207       2.63   0.022
X2            0.7824     0.1992       3.93   0.002
X3            1.0514     0.1358       7.74   0.000
X4           -0.1199     0.1717      -0.70   0.498
X5            0.0914     0.1301       0.70   0.496
X6            0.0883     0.1613       0.55   0.594
X7           -0.1017     0.1549      -0.66   0.524

s = 1.004       R-sq = 91.7%     R-sq(adj) = 86.9%
```

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	7	133.657	19.094	18.95	0.000
Error	12	12.088	1.007		
Total	19	145.746			

THE C MATRIX IS

721.671	-1.494	-2.020	-2.062	-2.279	-2.030	-0.557	-1.496
-1.494	0.269	-0.047	0.011	-0.027	-0.001	-0.014	0.016
-2.020	-0.047	0.039	-0.003	0.004	0.007	0.010	-0.010
-2.062	0.011	-0.003	0.018	0.006	0.005	-0.009	0.010
-2.279	-0.027	0.004	0.006	0.029	0.002	0.001	0.001
-2.030	-0.001	0.007	0.005	0.002	0.017	-0.002	0.002
-0.557	-0.014	0.010	-0.009	0.001	-0.002	0.026	-0.011
-1.496	0.016	-0.010	0.010	0.001	0.002	-0.011	0.024

The regression of Y on X_2 and X_3

The regression equation is
Y = - 61.2 + 0.895 X2 + 1.06 X3

Predictor	Coef	Stdev	t-ratio	p
Constant	-61.20	14.55	-4.21	0.001
X2	0.8947	0.1768	5.06	0.000
X3	1.0556	0.1189	8.88	0.000

s = 1.131 R-sq = 85.1% R-sq(adj) = 83.3%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	124.009	62.005	48.49	0.000
Error	17	21.736	1.279		
Total	19	145.746			

EXHIBIT 4.12.6

Data and MINITAB Output for Problem 4.12.4

```
THE C MATRIX IS

  165.675    -1.669    -0.855
   -1.669     0.024     0.002
   -0.855     0.002     0.011
```

The regression of Y on X_1, X_2, and X_3

```
The regression equation is
Y = - 78.2 + 1.35 X1 + 0.692 X2 + 1.10 X3

Predictor       Coef      Stdev    t-ratio       p
Constant     -78.23      13.24      -5.91   0.000
X1           1.3503     0.4474       3.02   0.008
X2           0.6925     0.1602       4.32   0.001
X3          1.10250    0.09908      11.13   0.000

s = 0.9305      R-sq = 90.5%     R-sq(adj) = 88.7%


Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    3     131.894      43.965      50.78   0.000
Error        16      13.852       0.866
Total        19     145.746

THE C MATRIX IS

  202.461    -2.917    -1.233    -0.957
   -2.917     0.231    -0.035     0.008
   -1.233    -0.035     0.030     0.000
   -0.957     0.008     0.000     0.011
```

For parts (a) and (b), suppose that assumptions (B) hold and that the model is

$$\text{model A: } \mu_Y^{(A)}(x_1, x_2, x_3, x_4, x_5, x_6, x_7) = \beta_0^A + \beta_1^A x_1 + \beta_2^A x_2 + \beta_3^A x_3 + \beta_4^A x_4 + \beta_5^A x_5 + \beta_6^A x_6 + \beta_7^A x_7 \quad \textbf{(4.12.5)}$$

Answer the following questions. For ease of notation let σ_A denote $\sigma_{Y|X_1,X_2,X_3,X_4,X_5,X_6,X_7}$ and $\rho^2_{Y(A)}$ denote $\rho^2_{Y(X_1,X_2,X_3,X_4,X_5,X_6,X_7)}$.

a What might be an appropriate target population of interest? Is the target population identical with the study population? Explain.

b **i** Estimate β_i^A $(i = 0, 1, 2, 3, 4, 5, 6, 7)$, and $\mu_Y^{(A)}(x_1, x_2, x_3, x_4, x_5, x_6, x_7)$.

ii Estimate σ_A.

iii Compute 95% two-sided confidence intervals for β_i^A, $i = 0, 1, 2, 3, 4, 5, 6, 7$. Explain in words the meaning of the computed confidence interval for β_1^A.

iv The investigator wants to examine the difference in the average heights at age 18 between two groups of subjects whose lengths at birth differed by 1 inch, but the subjects have the same set of values for $X_2, X_3, X_4, X_4, X_5, X_6, X_7$. Express this difference in terms of the *parameters* of the multiple regression function in (4.12.5) and obtain a 95% confidence interval for it.

v To evaluate how good the regression function in (4.12.5) is for predicting Y, the investigator wants to compare σ_A with σ_Y. Calculate a two-sided confidence interval for σ_Y/σ_A with confidence coefficient greater than or equal to 90%. Explain in words the meaning of this confidence statement.

vi To assist the investigator in evaluating how good the regression function in (4.12.5) is for predicting Y, *relative to* μ_Y (when no predictors are used), compute a point estimate for $\rho^2_{Y(A)}$.

vii Predict the height at age 18 of an *individual* who belongs to the subpopulation with

X_1 = length at birth = 20 inches
X_2 = mother's height at age 18 = 60 inches
X_3 = father's height at age 18 = 72 inches
X_4 = maternal grandmother's height at age 18 = 61 inches
X_5 = maternal grandfather's height at age 18 = 71 inches
X_6 = paternal grandmother's height at age 18 = 62 inches
X_7 = paternal grandfather's height at age 18 = 70 inches

viii In (vii) estimate the *average* height at age 18 of all individuals in the subpopulation.

ix Compute a 95% lower confidence bound for the height at age 18 of an individual randomly chosen from the subpopulation in (vii). The value of $x^T C x$, should you need it, is given to be 2.5321.

x In (vii) compute a 95% lower confidence bound for the average height at age 18 of all individuals in the subpopulation.

xi Explain why the answers to (vii) and (viii) are the same but the lower bounds in (ix) and (x) are not equal to each other.

c Since assumptions (B) are presumed to hold for the eight-variable population $\{(Y, X_1, \ldots, X_7)\}$, it follows that assumptions (B) hold for the three-variable population $\{(Y, X_2, X_3)\}$ and also for the four-variable population $\{(Y, X_1, X_2, X_3)\}$. Exhibit 4.12.6 gives the computer outputs for model C and model D defined in (4.12.6) and (4.12.7), respectively.

i Consider model C defined by

$$\mu_Y^{(C)}(x_2, x_3) = \beta_0^C + \beta_2^C x_2 + \beta_3^C x_3 \qquad (4.12.6)$$

Note that this model uses only mother's height at age 18 and father's height at age 18 to predict son's height at age 18. Estimate β_0^C, β_2^C, β_3^C, and σ_C.

ii Consider model D defined by

$$\mu_Y^{(D)}(x_1, x_2, x_3) = \beta_0^D + \beta_1^D x_1 + \beta_2^D x_2 + \beta_3^D x_3 \qquad \textbf{(4.12.7)}$$

Note that this model uses length at birth, mother's height at age 18, and father's height at age 18 to predict son's height at age 18. Estimate β_0^D, β_1^D, β_2^D, β_3^D, and σ_D.

iii To evaluate how much better model D is than model C for predicting Y, the investigator wants to compare σ_D with σ_C. Calculate an approximate 90% two-sided confidence interval for σ_C/σ_D. Explain in words the meaning of this confidence statement.

For (iv)–(xiii), use model C in (4.12.6).

iv Express the difference (in terms of population parameters) in the average heights at age 18 between two subpopulations of individuals if every mother of the first group of individuals was 1 inch taller (at age 18) than every mother of the individuals in the second group, but fathers' heights are the same for both subpopulations. Estimate this difference and compute a 95% lower confidence bound for it.

v An investigator is interested in determining whether the average height at age 18 of a subpopulation of individuals is at least 1/4 inch greater than the average height at age 18 of another subpopulation of individuals if every mother of the first group of individuals was 1 inch taller (at age 18) than every mother of the individuals in the second group, but fathers' heights are the same for both subpopulations. Formulate an appropriate pair of hypotheses and carry out the test. Calculate the P-value for this test. State your conclusions using $\alpha = 0.05$.

vi Which is more informative—the confidence bound in (iv) or the hypothesis test in (v)? Why?

vii What is the difference between the average heights at age 18 of two subpopulations if every father of the first group of individuals was 1 inch taller (at age 18) than every father of the individuals in the second group, given mothers' heights are the same for both subpopulations? Estimate this difference and obtain a 95% upper confidence bound for it.

viii Do the data provide evidence (at $\alpha = 0.05$) indicating that the average height at age 18 of a subpopulation of individuals is at most 1 inch greater than the average height at age 18 of another subpopulation of individuals if every father of the first group of individuals was 1 inch taller (at age 18) than every father of the individuals in the second group, if the mothers' heights are the same for both subpopulations? Formulate an appropriate pair of hypotheses and carry out the test. Calculate the P-value for this test. State your conclusion using $\alpha = 0.05$.

ix Which is more informative—the confidence bound in (vii) or the hypothesis test in (viii)? Why?

x Consider the subpopulation of all individuals with $X_2 = 58$. For this subpopulation, estimate the coefficient of determination of Y with X_3. The quantity $SSE(X_2)$ is equal to 122.491.

xi Compute a two-sided 90% confidence interval for $\sigma_{Y|X_2,X_3}$.

xii Predict the height at age 18 of a child who is now 2 years old if it is known that his mother was 60 inches tall at age 18 and his father was 72 inches tall at age 18. Also compute a two-sided 95% confidence interval for the height at age 18 of *this* child.

xiii Consider all individuals in the subpopulation determined by

mother's height at age 18 = $X_2 = 60$
father's height at age 18 = $X_3 = 72$

Estimate the mean height at age 18; i.e., $\mu_Y^{(C)}(60, 72)$, of the individuals in this subpopulation. Also compute a 90% two-sided confidence interval for this mean height.

d For (i)–(iv) use model D in (4.12.7).

i Estimate the height at age 18 of an individual in the population whose length at birth is 20 inches, whose mother's height at age 18 was 60 inches, and whose father's height at age 18 was 72 inches.

ii Compute a 95% two-sided confidence interval for the height at age 18 of a randomly chosen baby belonging to the subpopulation in (i).

iii Compute a 95% two-sided confidence interval for the average height at age 18 of all individuals belonging to the subpopulation in (i).

iv Estimate $\rho^2_{Y(X_1)|X_2,X_3}$. Explain in words the meaning of $\rho^2_{Y(X_1)|X_2,X_3}$.