3

Straight Line Regression

3.1 Overview

In Chapter 2 we defined the regression function $\mu_Y(x_1, \ldots, x_k)$ of a response variable Y on k predictor variables $X_1, \ldots, X_k$ and introduced many of the basic concepts underlying regression. In particular we learned that the best function for predicting the Y value of an item using the values of $X_1, \ldots, X_k$ is the regression function $\mu_Y(x_1, \ldots, x_k)$. In this chapter we focus on the simple but important special case of **straight line regression**. Accordingly, throughout this chapter, we assume that there is only one predictor variable X and that the graph of the regression function of Y on X is a *straight line*, i.e.,

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{3.1.1}$$

The quantity β_1 is the slope and β_0 is the intercept of the regression line. Thus the mean of the Y values in the subpopulation determined by $X = x$ is given by $\mu_Y(x) = \beta_0 + \beta_1 x$. Recall that $\sigma_Y(x)$ denotes the standard deviation of this subpopulation. If the entire population data are available, then we can calculate exactly the values of β_0, β_1, and $\sigma_Y(x)$ for every allowable x, but since the entire population is almost never available in a real problem, we cannot know the values of β_0, β_1, and $\sigma_Y(x)$ exactly, so we must rely on sample data to *estimate* these and other unknown quantities (parameters). In this chapter we consider point and confidence interval estimation of various quantities of interest, and we also discuss statistical tests. Section 3.3 introduces two sets of assumptions under which the theory of linear regression has been well developed. Section 3.4 discusses point estimation of parameters of interest. Methods for examining the validity of regression assumptions are discussed in Section 3.5. Confidence interval procedures and statistical tests are described in Sections 3.6 and 3.7, respectively. Section 3.8 introduces the analysis of variance. The coefficient of correlation and the coefficient of determination are described in Section 3.9. The effect of measurement errors on inferences about various model parameters is explained in Section 3.10. Section 3.11 considers the special case of the straight line regression model, where the regression line is known

to pass through the origin. Chapter exercises appear in Section 3.12. Laboratory assignments describing the use of a statistical computing package (MINITAB or SAS) for straight line regression are in Chapter 3 of the laboratory manual.

Before proceeding further, we present a detailed illustrative example where the entire population of numbers is assumed to be available, *even though it never is in a real problem*, so that you can get a better grasp of the concepts. This example will also point out how various questions of interest, arising in real problems, can be answered *exactly* when the entire population of numbers is available. Statistical inference procedures, discussed in this chapter and throughout this book, attempt to provide answers to such questions when only sample data, and not the entire population, are available.

3.2 An Example of Straight Line Regression

Table D-3 in Appendix D contains a set of data consisting of 2,600 pairs of numbers (Y, X), where Y is the score (in percent) obtained by a student on a standardized calculus test administered at a certain university, and X is the number of hours (recorded to the nearest hour) that the student spent studying for this test. These data are also stored in the file **grades.dat** on the data disk. For purposes of illustration, we suppose that these data form a *bivariate population* $\{(Y, X)\}$. The size of the population is thus 2,600. An examination of these data shows that there are 13 distinct values of X in the population, and they are $0, 1, 2, \ldots, 12$. The number of observations, the means, and the standard deviations for each of the corresponding 13 subpopulations of Y values are exhibited in Table 3.2.1. A plot of the means of the

TABLE 3.2.1

Subpopulation Counts, Means, and Standard Deviations for Population Data in Table D-3

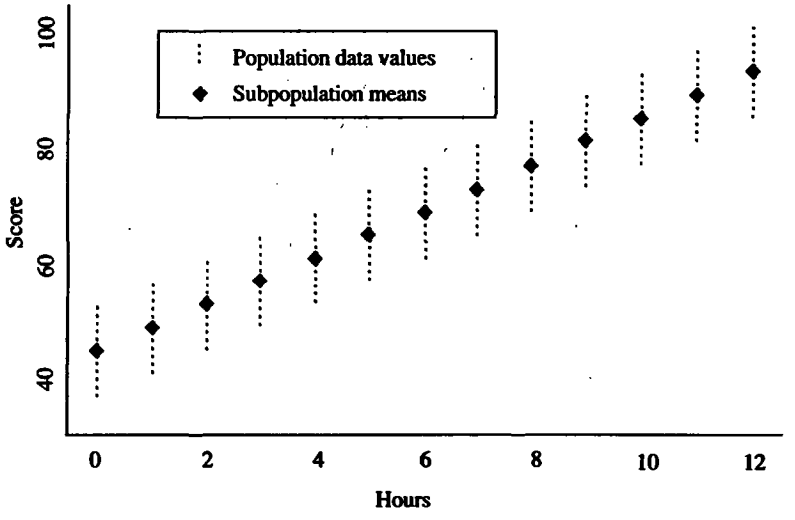
Hours X	Number of Items	Subpopulation Mean	Subpopulation Standard Deviation
0	200	45.0	2.881
1	200	49.0	2.881
2	200	53.0	2.881
3	200	57.0	2.881
4	200	61.0	2.881
5	200	65.0	2.881
6	200	69.0	2.881
7	200	73.0	2.881
8	200	77.0	2.881
9	200	81.0	2.881
10	200	85.0	2.881
11	200	89.0	2.881
12	200	93.0	2.881

Y values of these 13 subpopulations against the corresponding X values (i.e., a plot of $\mu_Y(x)$ against x for all allowable x values) is shown in Figure 3.2.1. This plot clearly shows that the regression function of Y on X is of the form $\mu_Y(x) = \beta_0 + \beta_1 x$; i.e., the subpopulation means for Y lie on a straight line when plotted against the corresponding values of X. Furthermore, we can calculate the values of β_0 and β_1 explicitly. In fact, the value of β_0 is 45.0 because the mean value of Y corresponding to $X = 0$ is 45.0 (see Table 3.2.1). Also the value of β_1 is 4.0 because the increase in the mean value of Y for a unit increase in X is easily seen to be 4.0%. Hence the population regression function is

$$\mu_Y(x) = 45.0 + 4.0x$$

Observe also that the subpopulation standard deviations are all equal to 2.881.

FIGURE 3.2.1



Note These data are specifically concocted for the purpose of illustration so that the population regression function of Y on X will be *exactly* a straight line and, in addition, the subpopulation standard deviations will all be the same. In most real problems, we cannot expect the population regression function to conform exactly to a straight line model, and the subpopulation standard deviations cannot be expected to all be exactly the same. But in many situations, these idealized conditions may be met *approximately*. You should also be aware that in actual investigations the number of subpopulations of Y values, determined by X, can be quite large, and the sizes of the subpopulations need not all be the same. In this particular example, however, we have deliberately kept the number of subpopulations rather small (13 to be precise) and the sizes of the subpopulations all equal (200 observations in each subpopulation) for ease of discussion.

Thus, because we know the entire population $\{(Y, X)\}$ in this example, we are able to determine *exactly* the values of β_0, β_1 and the subpopulation standard deviations $\sigma_Y(x)$. Any other population summary quantity (parameter) can be calculated exactly as well.

Some Questions of Interest

A student who is considering taking this calculus test may be interested in knowing the answers to the following questions:

1 What is the *average* increase in score per additional hour of studying time?

2 What is the *average* score of students who did not study at all for the test?

3 What is the *best predicted value* of the score of a student who spent 10 hours studying for this test?

4 Of all the students in the population who spent 10 hours studying for the test, what *proportion* obtained a score of 90% or above?

(3.2.1)

We give answers to these four questions by three methods.

a Answers based on the **entire population data**

b Answers based on **only population parameters**

c Answers based on **only a random sample** from the population

Of course in any real problem we can use only method (c) to obtain answers, but we give the answers to questions (1)–(4) of (3.2.1) by all three methods to help you understand that samples really can help answer questions about the population.

a Answers Based on the Entire Population Data

Answers to the preceding questions based on the entire population data are as follows:

1 The increase in the average score for each additional hour of studying time is equal to β_1, the slope of the regression line of Y on X, which has the value 4.0.

2 The average score of students who did not study at all for the test (i.e., $X = 0$) is $\mu_Y(0)$, which is equal to the intercept β_0 of the regression line, which has the value 45.0.

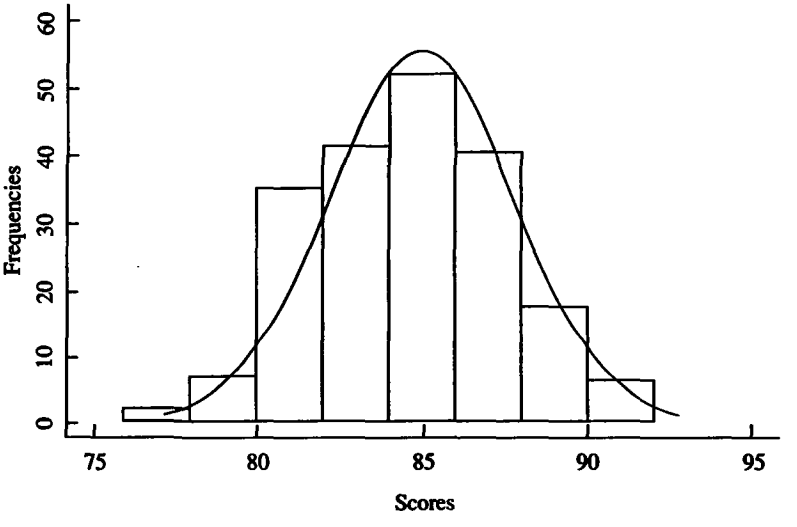
3 The best predicted value of the score of a student in this population who spent 10 hours studying for this test is $\mu_Y(10) = 45.0 + 4.0(10) = 85.0$.

4 In Table D-3 in Appendix D, an examination of the subpopulation of Y values corresponding to $X = 10$ shows that 11 out of the 200 students in this subpopulation obtained a score of 90% or above. Thus the required proportion is 0.055.

b Answers Based on Only Population Parameters

Clearly, we are able to obtain exact answers to the questions in (3.2.1) when the entire population $\{(Y, X)\}$ is available to us. In many situations, we can answer various questions concerning the population even if we do not know the entire population but know *only certain important summary quantities (parameters)* of the population. To demonstrate this in the present example, we begin by examining the histogram of the subpopulation of Y values determined by $X = 10$. The histogram is in Figure 3.2.2, which suggests that this subpopulation is approximately Gaussian. In fact, we should examine the subpopulation of Y values for each distinct X value to determine if each is approximately Gaussian.

FIGURE 3.2.2



Now suppose that we do not have the entire population $\{(Y, X)\}$ available to us, but suppose we do know that the regression function of Y on X is given by $\mu_Y(x) = 45.0 + 4.0x$ and that each subpopulation of Y values has a standard deviation equal to 2.881. Thus we know the values of β_0 and β_1, which are 45.0 and 4.0, respectively, and we also know that $\sigma_Y(x) = 2.881$ for each allowable x. Furthermore, by plotting the histogram of Y for each distinct value of X, we can demonstrate that each subpopulation of Y values is (approximately) Gaussian. With this information we can answer questions (1)–(4) in (3.2.1). Questions (1)–(3) can be answered knowing only that the regression function of Y on X is $\mu_Y(x) = 45.0 + 4.0x$. To answer question (4) we first observe that the mean Y value for the subpopulation corresponding to an X value of 10 is equal to $45.0 + 4.0(10) = 85$ and that its standard deviation is 2.881. We now use the fact that this subpopulation of Y values is approximately Gaussian. The proportion of values in a Gaussian population, with

mean equal to 85.0 and standard deviation equal to 2.881, that equals or exceeds 90 (actually 89.5, to account for the fact that the scores were rounded to the nearest integer) is equal to 0.0594 using Table T-1 in Appendix T. This is close to 0.055, the exact answer to question (4). (The reason for the slight discrepancy between the value 0.055 calculated directly from the population data and the value 0.0594 obtained using a table of Gaussian percentiles is that the theoretical Gaussian distribution is only an approximation to the actual subpopulation distribution.) Thus we can obtain answers to questions (1)–(4) in (3.2.1) if we have the appropriate population parameters β_0, β_1, and $\sigma_Y(x)$ for allowable values of x, even if we do not have the entire population.

To summarize, we can find *exact* answers to questions of interest concerning the population in Table D-3 because we have the entire population data available to us. We also see that we can obtain (nearly) exact answers based only on certain population parameters (summary quantities) because we know that the subpopulations of Y values are (nearly) Gaussian and the subpopulation standard deviations are all equal to 2.881. *It is for this reason that regression analysis focuses its attention on the estimation of various population parameters such as* β_0, β_1, $\mu_Y(x)$, $\sigma_Y(x)$, *etc.*

c Answers Based on a Random Sample

We now illustrate how to obtain answers to questions (1)–(4) in (3.2.1) by using a sample rather than the entire population. We do this by calculating, *approximately*, the values of the parameters β_0, β_1, and $\sigma_Y(x)$ and hence obtaining an approximation to the population regression line, using *sample data* from the preceding population. For this purpose we selected a sample of size 26 from the population in Table D-3 by randomly selecting two items from each subpopulation with preselected X values of $0, 1, 2, \ldots, 12$. Thus, the data are obtained by sampling with preselected X values. The sample data are displayed in Table 3.2.2, and they are also stored in the file **grades26.dat** on the data disk.

The 13 subpopulations (one subpopulation of Y values for each value of $X = 0, 1, \ldots, 12$) are displayed in Figure 3.2.3. The sample values are indicated by •. In Figure 3.2.4 the sample values are displayed together with a line that was *visually fitted* to the data. Figure 3.2.5 shows the sample data, the visually fitted line, and the population regression line $\mu_Y(x) = 45.0 + 4.0x$; of course in a real problem only the sample data are available and $\mu_Y(x)$ is not known, but we display $\mu_Y(x)$ to show how the sample data are grouped around it.

If we use the visually fitted line as the *estimate* of the population regression line, then the estimated values of β_0 and β_1 are the values of the intercept and the slope of this line which, from Figure 3.2.4, we judge to be 43.0 and 4.25, respectively (the change in Y as X changes from 0 to 12 is visually approximated as equal to 51 units, and so the slope of the line is estimated to be 51/12 = 4.25). Based on these sample estimates, we obtain the following approximate answers to questions (1)–(3) of (3.2.1).

1 On the average, the increase in score for each additional hour of studying time is estimated to be 4.25%.

2 The average score of students who did not study at all is estimated to be 43%.

TABLE 3.2.2

Grades26 Data. Sample of Size 26 from the Population Data in Table D-3 (Sampling with Preselected X Values)

Sample Item Number	Score (in percent) Y	Hours X
1	42	0
2	44	0
3	51	1
4	48	1
5	51	2
6	54	2
7	57	3
8	54	3
9	57	4
10	63	4
11	61	5
12	69	5
13	70	6
14	70	6
15	70	7
16	72	7
17	74	8
18	83	8
19	84	9
20	81	9
21	84	10
22	85	10
23	91	11
24	86	11
25	91	12
26	95	12

3 The predicted value of the score of any student who studies for 10 hours is $43 + 10 \times 4.25 = 85.5\%$.

The answer to question (4) of (3.2.1), based on sample data, depends on procedures to be discussed later (see Problem 3.4.6). Thus we see that even when we have only a *sample* of values from the population, we can obtain useful (though approximate) answers to questions of interest.

FIGURE 3.2.3

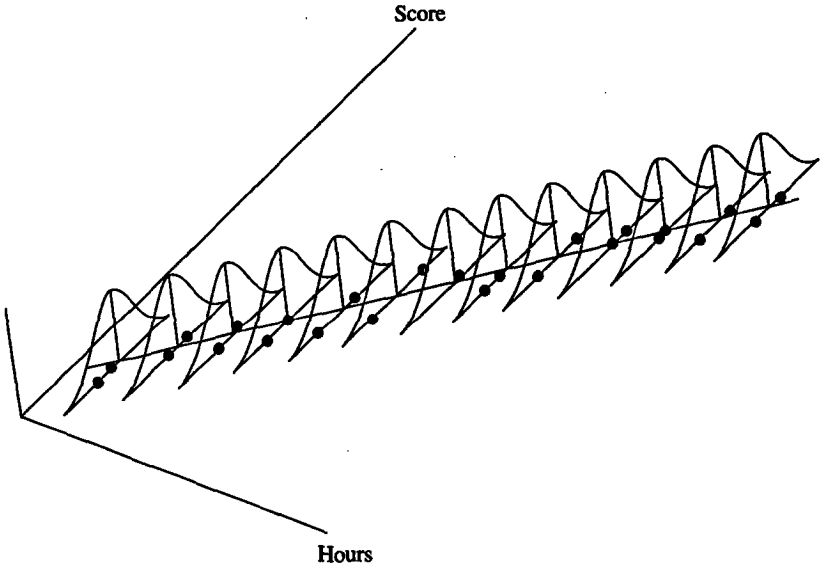


FIGURE 3.2.4

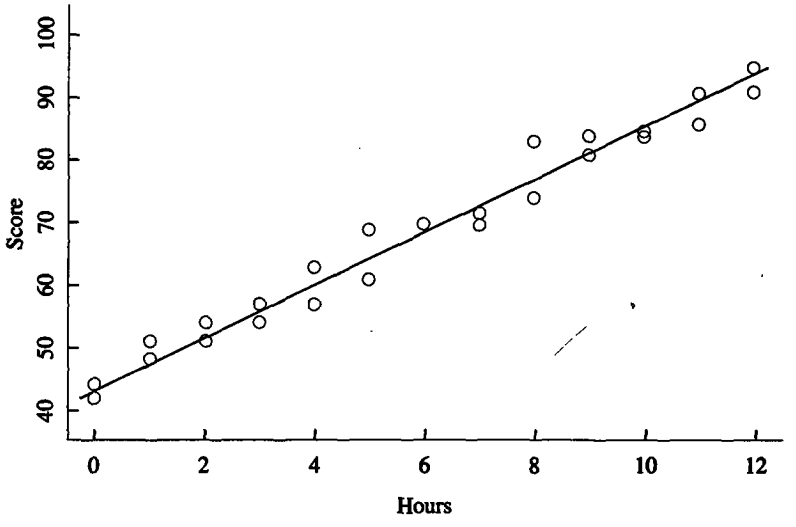
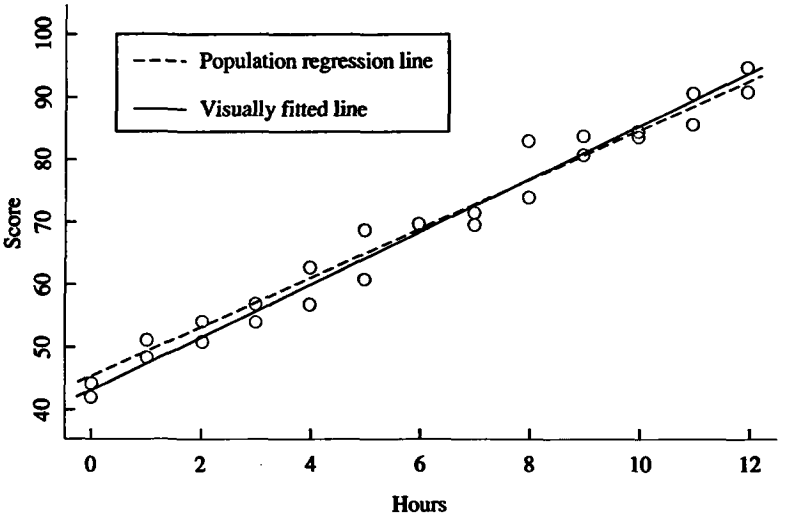


FIGURE 3.2.5



Systematic Methods of Estimation

Although we found an estimate of the population regression line using a straight line that was *visually* judged to provide a good fit to the sample data, we did this for illustration only. It is desirable to obtain an estimate of the population regression line based on a more objective and scientifically sound procedure. Several such methods are available in the literature, and one method that is widely used and has a long history is based on the so-called **method of least squares**. Another method that is becoming increasingly popular is the **method of least absolute deviations**. We use primarily the method of least squares for estimating unknown parameters because it is mathematically simpler than most alternative approaches and because estimates obtained using the method of least squares are best estimates when certain assumptions about the population and the sample are satisfied. If you are interested in circumstances under which other approaches may be desirable, you should consult more advanced books on regression.

In straight line regression, we are usually interested in estimating $\mu_Y(x)$, β_0, β_1, $\sigma_Y(x)$, and $Y(x)$, where $Y(x)$ is the Y value of an item chosen at random from the subpopulation whose X value is x. We may also be interested in estimating $\rho_{Y,X}$, μ_Y, μ_X, σ_Y, and σ_X. The estimation of $\sigma_Y(x)$ for all x is, for all practical purposes, impossible unless we make some simplifying assumptions regarding the population $\{(Y, X)\}$. In the next section we discuss two commonly used sets of assumptions regarding the population $\{(Y, X)\}$ and sampling procedures under which the theory of straight line regression has been extensively studied.

Problems 3.2

3.2.1 Using simple random sampling, a sample of size 26 was selected from the population data in Table D-3 in Appendix D, consisting of scores and hours studied for 2,600 students. The sample data are given in Table 3.2.3 and are also stored in the file **table323.dat** on the data disk. Here X is the number of hours studied, and Y is the percent score obtained on the test for each student in the sample. Examine these data. Plot Y against X, and examine this plot. Does it appear from this plot that the population regression function of Y on X is a straight line?

3.2.2 In Problem 3.2.1, visually fit a straight line to the plotted data. Use this fitted line to obtain approximate values for β_0 and β_1.

3.2.3 Use the data from Problem 3.2.1 to answer questions (1)–(3) of (3.2.1). How do the answers compare with the answers obtained using the entire population?

TABLE 3.2.3

Sample Item Number	Score (in percent) Y	Hours X
1	44	0
2	86	10
3	87	10
4	58	3
5	85	10
6	55	1
7	63	4
8	48	0
9	57	3
10	54	2
11	82	10
12	90	12
13	56	3
14	67	5
15	81	8
16	57	4
17	47	1
18	47	1
19	44	0
20	48	0
21	54	3
22	45	0
23	51	1
24	91	12
25	58	3
26	100	12

3.2.4 Compare a plot of the data in Table 3.2.3 with a plot of the sample data in Table 3.2.2 (these data are plotted in Figure 3.2.4), and with a plot of the population regression line $\mu_Y(x) = 45.0 + 4.0x$. Which set of sample data do you think best estimates the population regression line? Does this give you any reason to prefer one sampling method over the other?

3.2.5 A simple random sample of size 10 was selected from the population in Table D-3. The data are given in Table 3.2.4 and are also stored in the file **table324.dat** on the data disk. Repeat problems 3.2.1–3.2.3 for these data.

3.2.6 Which sample would you prefer, the one in Problem 3.2.5 or the one in Problem 3.2.1, to estimate the population regression function? Why?

TABLE 3.2.4

Sample Item Number	Score (in percent) Y	Hours X
1	41	1
2	59	4
3	90	11
4	88	11
5	52	2
6	53	2
7	53	1
8	63	5
9	87	10
10	74	8

3.3 Straight Line Regression Model—Assumptions (A) and (B)

To obtain useful point and confidence interval estimates and tests for parameters associated with a population $\{(Y, X)\}$, we must make some assumptions about the population and about the method used to collect the sample data. One such set of assumptions, which we refer to as assumptions (A), under which the theory for straight line regression has been well developed, is given in Box 3.3.1. Three of these assumptions concern the population and two concern the sample.

BOX 3.3.1 Assumptions (A) for Straight Line Regression

A two-variable population $\{(Y, X)\}$ is the study population.

(Population) Assumption 1 The mean of all the Y values in the subpopulation whose X value is x is denoted by $\mu_Y(x)$ and it is given by

$$\mu_Y(x) = \beta_0 + \beta_1 x \qquad \text{for } a \le x \le b$$

where β_0 and β_1 are unknown parameters and the allowable values of x lie between a and b.

(Population) Assumption 2 The standard deviations (and hence the variances) of the subpopulations do not depend on the value of X (i.e., they are the same for each subpopulation). This assumption is referred to as the assumption of **homogeneity of standard deviations** or, equivalently, **homogeneity of variances**. This common standard deviation of all the subpopulations is denoted by $\sigma_{Y|X}$. When there is no possibility of confusion, we simply write σ instead of $\sigma_{Y|X}$ to denote this common subpopulation standard deviation.

(Population) Assumption 3 Each subpopulation of Y values, determined by the distinct values of X, is a Gaussian population.

(Sample) Assumption 4 The data are obtained either by simple random sampling or by sampling with preselected X values as discussed in Section 2.3.

(Sample) Assumption 5 The X and Y values of the items in the sample are measured without error (however, see Section 3.10).

Associated with the two-variable population $\{(Y, X)\}$ are several quantities of interest. The most commonly needed quantities are β_0, β_1, $\mu_Y(x)$, $Y(x)$ (the Y value of a randomly chosen item from the subpopulation with $X = x$), and σ. Assumptions (A) in Box 3.3.1 are sufficient for making inferences about these parameters.

In some situations, we may also be interested in μ_Y, σ_Y, μ_X, σ_X, and $\rho_{Y,X}$ and, if data are obtained by sampling with preselected X values, we cannot make valid inferences about these parameters unless every subpopulation is sampled and the relative subpopulation sizes are known (which is almost never the case in real problems). **(3.3.1)**

A more restrictive set of assumptions for straight line regression, referred to as assumptions (B), given in Box 3.3.2, is sufficient for making inferences about *all* of the quantities β_0, β_1, $\mu_Y(x)$, $Y(x)$, σ, μ_Y, σ_Y, μ_X, σ_X, and $\rho_{Y,X}$.

BOX 3.3.2 Assumptions (B) for Straight Line Regression

(Population) Assumption 1 The two-variable population $\{(Y, X)\}$ that is to be studied is a **bivariate Gaussian population.**

(Sample) Assumption 2 The data are obtained by simple random sampling as discussed in Section 2.3.

(Sample) Assumption 3 The X and Y values of the items in the sample are measured without error (however, see Section 3.10).

Comments

1 For assumptions (B) in Box 3.3.2 to be met, we must obtain sample data by simple random sampling. If, instead, we obtain data by sampling with preselected X values, then no random sample from the $\{X\}$ population or the $\{Y\}$ population is available.

2 If $\{(Y, X)\}$ is a bivariate Gaussian population as in population assumption 1 of Box 3.3.2, then population assumptions 1, 2, 3 in Box 3.3.1 are automatically satisfied. Conversely, if population assumptions 1, 2, and 3 in Box 3.3.1 are satisfied and, additionally, if the one-variable population $\{X\}$ is also a Gaussian population, then population assumption 1 of Box 3.3.2 holds; i.e., the two-variable population $\{(Y, X)\}$ is bivariate Gaussian. Thus, *(population) assumptions* (B) *for straight line regression imply (population) assumptions* (A), *but the converse is not generally true.*

3 When the two-variable population $\{(Y, X)\}$ is bivariate Gaussian, then the one-variable populations $\{Y\}$ and $\{X\}$ are both Gaussian populations. However, $\{Y\}$ and $\{X\}$ may both be Gaussian populations and yet $\{(Y, X)\}$ may not be a bivariate Gaussian population. An example of this situation was given in Section 1.9.

4 Let Y_I and X_I be the Y and the X values corresponding to population item I. The predicted Y value for this item is $\mu_Y(X_I) = \beta_0 + \beta_1 X_I$, which is the mean Y value for the subpopulation with $X = X_I$. We write E_I for the difference between the actual value Y_I and the corresponding subpopulation mean $\beta_0 + \beta_1 X_I$. Thus $E_I = Y_I - (\beta_0 + \beta_1 X_I)$. Equivalently,

$$Y_I = \beta_0 + \beta_1 X_I + E_I \tag{3.3.2}$$

which is referred to as the **population regression model.** Under either assumptions (A) or assumptions (B) the population $\{E\}$ is Gaussian with mean zero and standard deviation σ.

It can never be determined if any of these assumptions are exactly satisfied in a real problem. Investigators may not know for certain that the graph of the population regression function $\mu_Y(x)$ is a straight line, but they may know that this is approximately so. The same can be said for all of the assumptions. The sample assumptions mainly concern collecting the data, and often investigators can make certain they are satisfied. On the other hand, investigators are often restricted by money, time, or other constraints, so the data collection methods may not exactly meet the requirements of randomness, etc. Sometimes the investigators who must

analyze the data and draw conclusions from them are not the ones who collected the data. They may know or suspect that 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 investigators are in the best position to determine whether the assumptions are close enough to being satisfied to allow valid conclusions to be drawn about the population under study. Investigators should always be aware of abnormalities in the data and deal with them.

The next section treats point estimation for the unknown parameters in the straight line regression model. Following this, in Section 3.5, we give methods for examining some of the assumptions in Box 3.3.1 and Box 3.3.2. If they appear not to hold, alternative procedures are sometimes available, and we discuss some of these in Chapter 8.

3.4 Point Estimation

The primary objective in a regression study is to use the sample data to obtain point and confidence interval estimates for the unknown quantities β_0, β_1, σ, $\mu_Y(x)$, and $Y(x)$ and also for various meaningful functions of these quantities. These estimates in turn aid investigators in gaining insight into quite complicated questions about the population under study. In this section we focus our attention on **point estimation.**

Recall that a point estimate of an unknown parameter is a number, computed from observed sample data, that may be used in place of the unknown value of the parameter of interest for making practical decisions. When assumptions (A) or (B) hold, it can be shown mathematically that the best estimates of β_0 and β_1 in (3.1.1) are obtained by the *method of least squares*. Using these estimates we can obtain the best estimates of other quantities of interest. We first describe the method of least squares.

Method of Least Squares

Suppose $(y_1, x_1), \ldots, (y_n, x_n)$ is a sample of size n from the bivariate population $\{(Y, X)\}$, selected using either simple random sampling or sampling with preselected X values, and the population regression function is

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{3.4.1}$$

The quantity e_i defined by

$$e_i = y_i - \mu_Y(x_i) = y_i - (\beta_0 + \beta_1 x_i) \tag{3.4.2}$$

is the prediction error when we use $\mu_Y(x_i) = \beta_0 + \beta_1 x_i$ to predict y_i for $i = 1, \ldots, n$. The relationship among the observed value y_i, the value of the regression function at x_i, viz., $\beta_0 + \beta_1 x_i$, and the prediction error e_i, is given by

$$y_i = \beta_0 + \beta_1 x_i + e_i \tag{3.4.3}$$

This is referred to as the **sample regression model.**

Since β_0 and β_1 are unknown parameters, we want to use sample data to obtain estimates of them. Under assumptions (A) or (B) it can be shown that the best estimates of β_0 and β_1 are obtained by the method of least squares. The resulting estimates are denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. The corresponding estimate of the regression function is denoted by

$$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3.4.4}$$

The prediction error when using $\hat{\mu}_Y(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$ to predict y_i is denoted by $\hat{e}_i$ and is given by

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \tag{3.4.5}$$

The quantities $\hat{e}_i$ for $i = 1, \ldots, n$ are called **residuals**. They are useful in examining the validity of the assumptions given in Box 3.3.1 as well as those given in Box 3.3.2. This is discussed in Section 3.5.

The *least squares estimates* $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen in such a way that the quantity $SSE(X)$, called the sum of squares of prediction errors when X is used to predict Y, and defined by

$$SSE(X) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2 \tag{3.4.6}$$

has the smallest possible value among all the possible choices we could make for $\hat{\beta}_0$ and $\hat{\beta}_1$. When there is no possibility of confusion, we will simply write SSE instead of $SSE(X)$ and refer to it as the **sum of squared errors** or **error sum of squares**.

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

$$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

is a good predictor of the *known sample values* y_i for each $X = x_i$ for $i = 1, 2, \ldots, n$, then we have reason to expect that it will be a good prediction function for *all* values of Y in the population corresponding to all allowable values of X. Thus we use the y_i and the x_i values of the items in the sample to assess the performance of the estimate of the population regression function at the sample points. We now enunciate the principle of least squares.

The principle of least squares states that the best estimate of the population regression function $\mu_Y(x) = \beta_0 + \beta_1 x$ is obtained by choosing $\hat{\beta}_0$ and $\hat{\beta}_1$ in (3.4.6) in such a way that the sum of squares of the prediction errors,

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3.4.7}$$

attains the *least* possible value.

Point Estimates of β_0, β_1, $\mu_Y(x)$ and $Y(x)$

It can be mathematically proven that the values of $\hat{\beta}_1$ and $\hat{\beta}_0$ that minimize $\sum \hat{e}_i^2$ in (3.4.6) are given by

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{3.4.8}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{3.4.9}$$

where

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \text{and} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

The quantities $\hat{\beta}_1$ and $\hat{\beta}_0$, defined in (3.4.8) and (3.4.9), respectively, are known as the *least squares estimates* of the population parameters β_1 and β_0. As mentioned earlier, when assumptions (A) or (B) for straight line regression are satisfied, these are in fact the best estimates of β_1 and β_0, respectively. The best estimate of $\mu_Y(x)$ is

$$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3.4.10}$$

As stated in Section 2.3, *the regression function of Y on X is also the best prediction function for predicting Y using X.* As a result, when only sample data are available, the best predicted value $\hat{Y}(x)$ of a randomly chosen observation $Y(x)$ from the subpopulation with $X = x$ is in fact equal to the estimated subpopulation mean $\hat{\mu}_Y(x)$; i.e.,

$$\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x = \hat{\mu}_Y(x) \tag{3.4.11}$$

Point Estimates for Linear Functions of β_0 and β_1

While β_0 and β_1 are important parameters in straight line regression, investigators are quite frequently interested in making inferences about certain linear combinations of β_0 and β_1. Suppose θ denotes the linear combination $a_0\beta_0 + a_1\beta_1$ of β_0 and β_1, where a_0 and a_1 are known numbers. The best point estimate of θ is equal to $\hat{\theta}$ where

$$\hat{\theta} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1 \tag{3.4.12}$$

Observe that $\mu_Y(x)$ is a quantity of the form $a_0\beta_0 + a_1\beta_1$ with $a_0 = 1$ and $a_1 = x$; β_1 is also a special case with $a_0 = 0$ and $a_1 = 1$; β_0 is a special case with $a_0 = 1$ and $a_1 = 0$; $\mu_Y(x_1) - \mu_Y(x_2) = (\beta_0 + \beta_1 x_1) - (\beta_0 + \beta_1 x_2) = (x_1 - x_2)\beta_1$ is a special case with $a_0 = 0$ and $a_1 = x_1 - x_2$.

Notation

It is customary to use the notation

$$SXY = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{3.4.13}$$

$$SSX = \sum_{i=1}^{n} (x_i - \bar{x})^2 \tag{3.4.14}$$

and

$$SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.4.15}$$

so that the formula for $\hat{\beta}_1$ in (3.4.8) may be conveniently written as

$$\hat{\beta}_1 = \frac{SXY}{SSX} \tag{3.4.16}$$

Remark

The following alternate (but equivalent) expressions for SXY, SSX, and SSY are sometimes useful.

$$SXY = \sum_{i=1}^{n} x_i y_i - \frac{(\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n} = \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} \tag{3.4.17}$$

$$SSX = \sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n} = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \tag{3.4.18}$$

and

$$SSY = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} y_i)^2}{n} = \sum_{i=1}^{n} y_i^2 - n\bar{y}^2 \tag{3.4.19}$$

Point Estimate of σ

Recall that σ is the common standard deviation of the subpopulations of Y values determined by the distinct values of X. The estimate $\hat{\sigma}$ of σ can be calculated using the formula

$$\hat{\sigma} = \sqrt{\frac{SSE}{(n-2)}} \tag{3.4.20}$$

where SSE is given by any one of the following equivalent expressions:

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} [y_i - \hat{\mu}_Y(x_i)]^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The quantity $\frac{SSE}{(n-2)}$, which is under the square root symbol in (3.4.20), is called **mean square error** for predicting Y using X and is denoted by $MSE(X)$, or MSE for short.

Thus

$$MSE = \frac{SSE}{(n-2)} \tag{3.4.21}$$

With this notation the estimate $\hat{\sigma}$ of σ is given by

$$\hat{\sigma} = \sqrt{MSE} \tag{3.4.22}$$

A convenient formula for calculating *SSE* using a hand-held calculator is

$$SSE = SSY - \frac{(SXY)^2}{SSX} \tag{3.4.23}$$

Table 3.4.1 exhibits in detail how the residuals $\hat{e}_i$ enter into the calculation of $\hat{\sigma}$, which is obtained by dividing the sum of the numbers in the last column by $n-2$ and then taking the square root (see 3.4.22).

It can be easily verified that the residuals $\hat{e}_i$ in (3.4.5) must sum to zero. This fact is sometimes used to check the arithmetic involved in the calculation of $\hat{\sigma}$.

All computational formulas for evaluating $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\mu}_Y(x)$, $\hat{e}_i$, and $\hat{\sigma}$ are influenced by rounding errors, so it is advisable to carry as many significant digits as possible when performing the required arithmetical operations. The final result may be rounded to the desired number of significant digits.

The calculations required to estimate β_0, β_1, and σ may be conveniently carried out using any standard statistical computing package. We explain the use of a statistical computing package (MINITAB or SAS) for these calculations in Section 3.4 of the laboratory manual.

Terminology

For convenience, Box 3.4.1 summarizes terminology associated with straight line regression, and Box 3.4.2 summarizes the formulas for various parameter estimates of interest.

TABLE 3.4.1

Sample Item	Observed Y	Observed X	Prediction $\hat{\mu}_Y(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$	Residuals $\hat{e}_i = y_i - \hat{\mu}_Y(x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$	Residuals Squared $\hat{e}_i^2$
1	y_1	x_1	$\hat{\mu}_Y(x_1) = \hat{\beta}_0 + \hat{\beta}_1 x_1$	$\hat{e}_1 = y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1$	$(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2$
2	y_2	x_2	$\hat{\mu}_Y(x_2) = \hat{\beta}_0 + \hat{\beta}_1 x_2$	$\hat{e}_2 = y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2$	$(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2$
⋮	⋮	⋮	⋮	⋮	⋮
n	y_n	x_n	$\hat{\mu}_Y(x_n) = \hat{\beta}_0 + \hat{\beta}_1 x_n$	$\hat{e}_n = y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n$	$(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$

BOX 3.4.1

Population regression function, or simply, the regression function:

$$\mu_Y(x) = \beta_0 + \beta_1 x \qquad \text{for } a \le x \le b$$

Sample regression function:

$$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Population regression model, or simply, the regression model:

$$Y_I = \beta_0 + \beta_1 X_I + E_I \qquad \text{for } I = 1, \ldots, N$$

Sample regression model:

$$y_i = \beta_0 + \beta_1 x_i + e_i \qquad \text{for } i = 1, \ldots, n$$

A randomly chosen Y value from the subpopulation determined by $X = x$:

$$Y(x)$$

Sample prediction function, or simply, prediction function:

$$\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

Note: $\hat{\mu}_Y(x) = \hat{Y}(x)$.

BOX 3.4.2

Point Estimates of Various Population Quantities

Suppose a sample $(y_1, x_1), \ldots, (y_n, x_n)$ of size n is obtained from a study population $\{(Y, X)\}$ by simple random sampling or by sampling with preselected X values. Then

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^n y_i}{n}$$

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

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

$$SXY = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{(\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n}$$

$$= \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}$$

$$\hat{\beta}_1 = \frac{SXY}{SSX}$$

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

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = \text{ residual for sample item } i$$

$$SSE = \sum_{i=1}^{n} [y_i - \hat{\mu}_Y(x_i)]^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 = \sum_{i=1}^{n} \hat{e}_i^2$$
$$= SSY - \frac{(SXY)^2}{SSX}$$
$$MSE = \frac{SSE}{n-2}$$
$$\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{(n-2)}} = \sqrt{\frac{\sum_{i=1}^{n} \hat{e}_i^2}{(n-2)}} = \sqrt{MSE}$$

For $\theta = a_0\beta_0 + a_1\beta_1$ the point estimate is

$$\hat{\theta} = a_0\hat{\beta}_0 + a_1\hat{\beta}_1$$

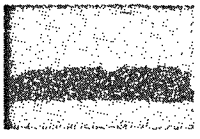
Comment

In the sample regression function in (3.4.4), we can substitute $\bar{y} - \hat{\beta}_1\bar{x}$ for $\hat{\beta}_0$ and write

$$\hat{\mu}_Y(x) = \bar{y} + \hat{\beta}_1(x - \bar{x}) \qquad \textbf{(3.4.24)}$$

Thus $\hat{\mu}_Y(\bar{x}) = \bar{y}$, which demonstrates that the graph of the sample regression function passes through the point $(\bar{x}, \bar{y})$, which is the "center" of the data.

Tasks 3.4.1 and 3.4.2 are intended to provide you with an opportunity to better grasp the concepts discussed so far in straight line regression and also to illustrate the use of the formulas for point estimation of various parameters of interest. The questions in these tasks are posed as word problems and are indicative of the types of questions arising in real applications.



Task 3.4.1

Crystalline forms of certain chemical compounds are used in various electronic devices, and it is often more desirable to have large crystals rather than small ones. Crystals of one particular compound are to be produced by a commercial process, and an investigator wants to examine the relationship between the size of a crystal, as determined by its weight Y in grams, and the number of hours X it takes the crystal to grow to its final size. The following data are from a laboratory study in which 14 crystals of various sizes were obtained by allowing the crystals to grow for different preselected amounts of time. The data are listed in Table 3.4.2 and are also stored in the file **crystal.dat** on the data disk. From the data in Table 3.4.2 we compute

$$\sum_{i=1}^{14} x_i = 210, \qquad \bar{x} = 15 \qquad \sum_{i=1}^{14} y_i = 105.74, \qquad \bar{y} = 7.5529$$

TABLE 3.4.2
Crystal Data

Crystal Number	Weight Y (in grams)	Time X (in hours)
1	0.08	2
2	1.12	4
3	4.43	6
4	4.98	8
5	4.92	10
6	7.18	12
7	5.57	14
8	8.40	16
9	8.81	18
10	10.81	20
11	11.16	22
12	10.12	24
13	13.12	26
14	15.04	28

$$\sum_{i=1}^{14}(x_i-\bar{x})^2 = SSX = 910 \qquad \sum_{i=1}^{14}(y_i-\bar{y})^2 = SSY = 244.159$$

$$\sum_{i=1}^{14}(y_i-\bar{y})(x_i-\bar{x}) = SXY = 458.12$$

A plot of the data is displayed in Figure 3.4.1.

Suppose the investigator is reasonably certain that assumptions (A) for straight line regression are (at least approximately) satisfied for x values between 2 hours and 28 hours, i.e.,

$$\mu_Y(x) = \beta_0 + \beta_1 x \qquad \text{for } 2 \le x \le 28$$

She is interested in finding answers to the following questions (our answers appear in italics).

1 Estimate how much the crystals grow per hour on the average.

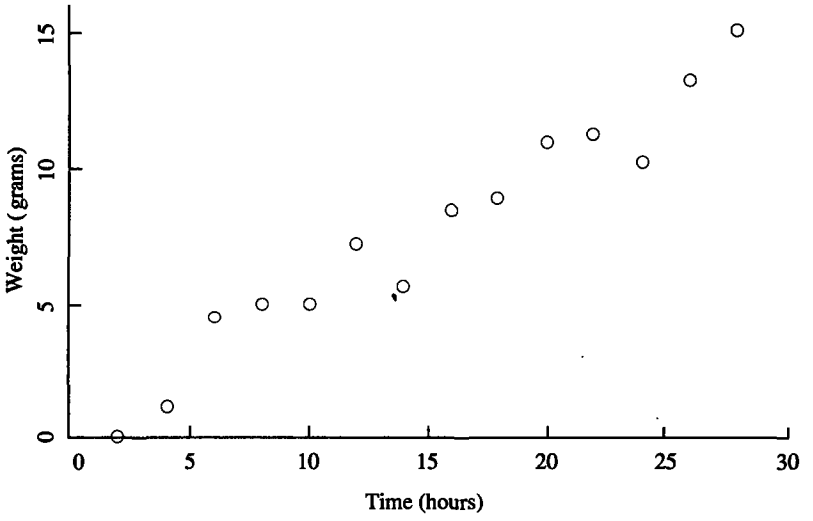
Because $\mu_Y(x) = \beta_0 + \beta_1 x$ is the regression function of Y on X, the average growth per hour is equal to β_1. The estimate of β_1, calculated using the formula in Box 3.4.2, is

$$\hat{\beta}_1 = \frac{SXY}{SSX} = \frac{458.12}{910} = 0.5034 \qquad \textit{(to 4 decimals)}$$

Also from Box 3.4.2 we get

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = 0.0014 \qquad \textit{(to 4 decimals)}$$

FIGURE 3.4.1



2 If a crystal is allowed to grow for 15 hours, what is its predicted weight?

Here we wish to predict the weight of a single crystal that is allowed to grow for 15 hours. This is considered to be a random observation, $Y(15)$, from the population of all crystals grown for 15 hours. The best predicted value of $Y(15)$ is

$$\hat{Y}(15) = \hat{\beta}_0 + \hat{\beta}_1(15)$$

Using the values $\hat{\beta}_0 = 0.0014$ and $\hat{\beta}_1 = 0.5034$ calculated in (1) above, the best predicted value for the weight of a single crystal at the end of 15 hours is $\hat{Y}(15) = 0.0014 + 0.5034(15) = 7.55$ grams.

3 The crystals are priced depending on the time taken to grow them as well as their actual weight. Crystals that are grown for 8 hours or less are priced at \$2 per gram, those that are grown between 8 hours and 16 hours are priced at \$10 per gram, and those that are grown for more than 16 hours are priced at \$16 per gram. These prices reflect the additional amount of operator intervention necessary to grow crystals for longer periods. Estimate the additional dollars that a crystal will sell for if it is allowed to grow for 24 hours instead of 12 hours.

The weight of a crystal that is grown for 24 hours is estimated to be $\hat{Y}(24) = \hat{\beta}_0 + \hat{\beta}_1(24) = 12.08$ grams, whereas the weight of a crystal grown for 12 hours is estimated to be $\hat{Y}(12) = \hat{\beta}_0 + \hat{\beta}_1(12) = 6.04$ grams. Hence the additional dollars that a crystal grown for 24 hours will fetch compared to a crystal grown for 12 hours, is estimated to be $12.08 \times 16 - 6.04 \times 10 = \$132.88$.

4 An electronic components manufacturer places an order for 100 crystals weighing 12 grams each with a tolerance of ±0.5 gram, i.e., weighing between 11.5 and 12.5 grams. How long should the crystals be allowed to grow? If 100 crystals are grown for this amount of time, how many crystals may be expected to meet the specifications?

Suppose crystals that are allowed to grow for x_0 hours have an average size equal to 12 grams. Then

$$\mu_Y(x_0) = \beta_0 + \beta_1 x_0 = 12$$

which implies that

$$x_0 = \frac{12 - \beta_0}{\beta_1}$$

The estimated value of x_0 is

$$\hat{x}_0 = \frac{12 - \hat{\beta}_0}{\hat{\beta}_1} = 23.84 \text{ hours}$$

Using the formula for $\hat{\sigma}$ in Box 3.4.2, the subpopulation standard deviation is estimated to be

$$\hat{\sigma} = \sqrt{\frac{1}{(n-2)}\left[SSY - \frac{(SXY)^2}{SSX}\right]}$$

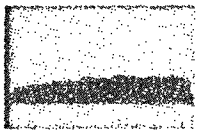
$$= \sqrt{\frac{1}{12}\left[244.159 - \frac{(458.12)^2}{910}\right]} = 1.062 \text{ grams}$$

Hence the proportion of crystals, which are grown for 23.84 hours, whose weights lie in the range from 11.5 grams to 12.5 grams, is approximately equal to the proportion of values in a Gaussian population, with mean equal to 12 and standard deviation equal to 1.062, that lie between 11.5 and 12.5. This is equivalent to the proportion of values in a standard Gaussian population that are between

$$\frac{11.5 - 12.0}{1.062} \quad \text{and} \quad \frac{12.5 - 12.0}{1.062}$$

i.e., between −0.471 and +0.471. Using Table T-1 in Appendix T, we calculate this proportion to be 0.362. If 100 crystals are grown for 23.84 hours, then we expect $100 \times 0.362 = 36$ (rounded to the nearest integer) crystals to meet the required specifications. This result is approximate because the calculations are based on estimated values of population parameters since their true values are unknown.

Task 3.4.2 further illustrates how regression techniques can be useful in real problems.



Task 3.4.2

An investigator wants to evaluate the performance of a new laboratory method for analyzing the concentration of arsenic (As) in water samples that is much cheaper than the existing method. If the new method is proven to be scientifically acceptable, then it will be adopted by environmental research groups for monitoring the quantity of As in industrial waste water. To investigate the relationship between measured concentrations of As (Y) and actual concentrations (X), the investigator makes several water samples containing *known* (preselected) amounts of As. These water samples are analyzed by a laboratory technician (who is unaware of the actual amounts of As in these solutions) using the new method of analysis. The concentrations are reported in micrograms/milliliter (μg/ml). The data are exhibited in Table 3.4.3 and are also stored in the file **arsenic.dat** on the data disk.

Suppose that, based on experience, the investigator feels assumptions (A) for straight line regression hold at least approximately. In particular the population regression function is

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

and the subpopulation standard deviations for Y are all the same, each equal to σ.

To start an analysis, *the first thing you should always do is to obtain a plot of Y against X*. This plot is shown in Figure 3.4.2. It appears from this plot that the assumption of a straight line model is quite reasonable. Next we compute some basic sums, sums of squares, and sums of crossproducts, the ingredients in the formulas for obtaining estimates of population parameters. We get

$$\sum_{i=1}^{32} x_i = 112 \qquad \bar{x} = 3.5 \qquad \sum_{i=1}^{32} y_i = 113.97 \qquad \bar{y} = 3.5616$$

$$SSX = 168 \qquad SSY = 164.95 \qquad SXY = 165.935$$

From these we compute

$$\hat{\beta}_1 = \frac{SXY}{SSX} = \frac{165.935}{168} = 0.9877$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.5616 - 0.9877(3.5) = 0.1046$$

$$\hat{\sigma} = \sqrt{\frac{1}{(n-2)}\left[SSY - \frac{(SXY)^2}{SSX}\right]} = \sqrt{\frac{1}{30}\left[164.95 - \frac{(165.935)^2}{168}\right]}$$

$$= \sqrt{\frac{1}{30}[164.95 - 163.89538]} = \sqrt{\frac{1.0546}{30}} = 0.1875$$

You should verify these calculations.

Suppose the investigator is interested in obtaining answers to the following questions.

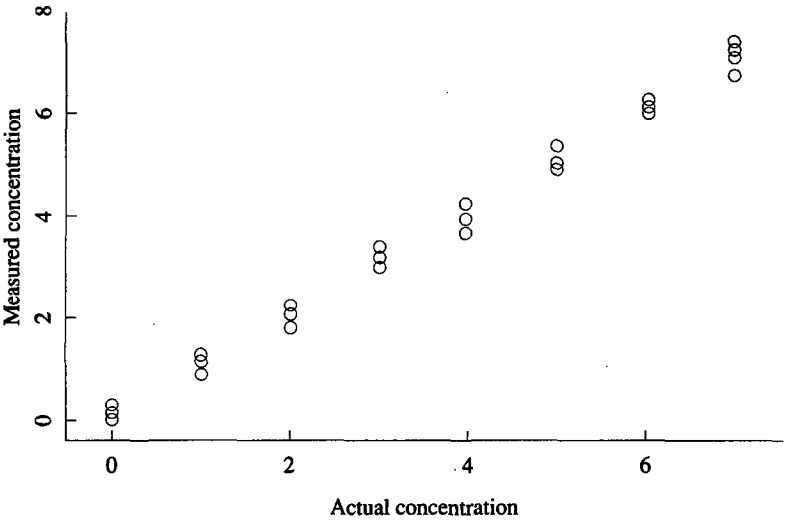
TABLE 3.4.3
Arsenic Data

Sample Item Number	Measured Concentration Y (in μg/ml)	True Concentration X (in μg/ml)
1	0.17	0
2	0.25	0
3	0.01	0
4	0.12	0
5	1.25	1
6	0.86	1
7	1.25	1
8	1.10	1
9	2.01	2
10	2.03	2
11	2.14	2
12	1.74	2
13	3.18	3
14	2.99	3
15	3.23	3
16	3.37	3
17	3.91	4
18	3.90	4
19	3.61	4
20	4.27	4
21	4.88	5
22	5.33	5
23	4.96	5
24	4.98	5
25	6.09	6
26	6.17	6
27	6.07	6
28	5.97	6
29	6.67	7
30	7.02	7
31	7.14	7
32	7.30	7

1 On the average, does the chemical analysis correctly report the absence of As when this is indeed the case? That is, is $\mu_Y(x) = 0$ when $x = 0$? Is $\beta_0 = 0$?

In individual instances, due to the presence of various kinds of disturbances or errors, the analysis may report a nonzero value of As even when the actual concentration of As is zero. The measured concentration of As for water samples containing no As (i.e., $x = 0$) is equal to β_0 on the average. *Thus the investigator wants to know whether or not β_0 is indeed zero.*

FIGURE 3.4.2



From the preceding computations, the least squares estimate $\hat{\beta}_0$ of β_0 is 0.1046. Before the investigator can decide whether or not β_0 may be regarded as equal to zero for practical purposes, he needs to know how good this estimate is. This is discussed in the section on confidence intervals.

2 On the average, does the chemical analysis result in an underestimate or an overestimate of the true As concentration, or does it provide an *unbiased* estimate of the true As concentration?

In individual instances, the presence of various kinds of errors leads to a measured value that is higher or lower than the true As concentration, and only very rarely will the measured value be exactly equal to the true value. However, if the average *of the subpopulation of measured Y values equals the true concentration X for each possible value of X, then the analysis is said to be* unbiased. *Thus, if the analysis is unbiased, then the regression function of Y on X must be $\mu_Y(x) = x$; i.e., $\beta_0 = 0$ and $\beta_1 = 1$. Here we examine whether or not the value of β_1 is equal to 1.*

From the preceding computations, the least squares estimate $\hat{\beta}_1$ of β_1 is equal to 0.9877, which is close to 1.0. Before the investigator is able to decide whether or not β_1 can be regarded as equal to one for practical purposes, he needs to know how good this estimate is. This is discussed in the section on confidence intervals.

3 Suppose, based on the calculations in (1) and (2), the investigator is fairly confident that the new method of chemical analysis for As is *unbiased*; i.e., the regression function of Y on X is $\mu_Y(x) = x$. However, in order for the method to be adopted by water quality monitoring agencies, 99% or more of the reported

concentrations must be accurate to within 1.0 μg/ml. Based on the data at hand, can he conclude that this is indeed the case?

Assuming that $\mu_Y(x) = x$ and that assumptions (A) are valid, 99% of the measured concentrations corresponding to a true concentration x will be between $\mu_Y(x) - 2.576\sigma$ and $\mu_Y(x) + 2.576\sigma$. Hence we want to know whether 2.576σ is less than 1.0. The estimate of σ calculated earlier is 0.1875, and so the estimated value of 2.576σ is equal to 0.483. This is well within the acceptable upper limit of 1.0. However, the investigator realizes that the estimate of σ is itself subject to sampling errors, and so he wants to know if the estimate is sufficiently accurate for decision-making purposes. A confidence interval for σ would be useful for this purpose. This is discussed in the section on confidence intervals.

You should bear in mind that in most instances X and Y are physical quantities that have certain units of measurement associated with them. In this connection the investigator may want to find out what happens if he decides to change the system of units. For instance, the original measurements may have been made in terms of pounds or miles, and it may be necessary to transform the measurements so they are expressed in terms of grams or meters. We discuss this next.

Effect of Change of Units on the Parameters and Their Estimates

The values of the population parameters such as β_0, β_1, and σ depend on the system of units in which Y and X are measured. If X is a measure of distance, then the units of X may be miles, kilometers, inches, feet, millimeters, etc. If X is a measure of weight, then the units of X may be tons, pounds, kilograms, etc. Similarly if Y is a measure of temperature, then the units of Y may be degrees Fahrenheit, degrees Celsius, or degrees Kelvin. Therefore we want to know what effect the choice of units has on the values of population parameters and their point estimates.

Let X_I, Y_I be the values of X and Y for the Ith population item measured using one system of units, and let X_I^*, Y_I^* denote the X and Y values of the same item measured using a second system of units. In most practical applications, different systems of units are *linearly* related. For instance,

$$\text{degrees Fahrenheit} = 32 + \tfrac{9}{5}\,(\text{degrees Celsius})$$

and

$$\text{kilometers} = 1000\ \text{meters}$$

With this in mind, suppose X_I^* and Y_I^* are defined in terms of X_I and Y_I as follows:

$$X_I^* = a + bX_I$$
$$Y_I^* = c + dY_I$$

Suppose that the population regression function is

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

when X and Y are measured using the first system of units. Also suppose the regression function is

$$\mu_{Y^*}(x^*) = \beta_0^* + \beta_1^* x^*$$

when X^* and Y^* are measured using the second system of units. Then it can be proved mathematically that

$$\beta_1^* = \frac{d}{b}\beta_1$$

and

$$\beta_0^* = c + \frac{d}{b}(b\beta_0 - a\beta_1)$$

Furthermore, if σ is the subpopulation standard deviation when the first system of units is used and σ^* denotes the subpopulation standard deviation when the second system of units is used, then

$$\sigma^* = |d|\sigma$$

The estimated values of the parameters are related in the same manner, viz.,

$$\hat{\beta}_1^* = \frac{d}{b}\hat{\beta}_1$$

$$\hat{\beta}_0^* = c + \frac{d}{b}(b\hat{\beta}_0 - a\hat{\beta}_1)$$

and

$$\hat{\sigma}^* = |d|\hat{\sigma}$$

We use the following example to demonstrate these relationships.

EXAMPLE 3.4.1

The output Y, in kilograms per hour, of a chemical process is related to the temperature X, in degrees Celsius, at which the process is run, and the population regression function has the form $\mu_Y(x) = \beta_0 + \beta_1 x$. The data in Table 3.4.4 were obtained from a pilot plant experiment using preselected X values and are also stored in the file **process1.dat** on the data disk. The estimates of β_0, β_1, and σ are calculated to be

$$\hat{\beta}_0 = -4.959 \qquad \hat{\beta}_1 = 0.0499 \qquad \hat{\sigma} = 0.4095$$

Now suppose the process output is measured in grams per second instead of kilograms per hour, and the process temperature is measured in degrees Fahrenheit instead of degrees Celsius. This gives us

$$Y_I^* = \frac{1000}{3600}Y_I \qquad \text{and} \qquad X_I^* = 32 + \frac{9}{5}X_I \qquad \text{for } I = 1, \ldots, N \qquad \textbf{(3.4.25)}$$

Let the sample data expressed in the new system of units be denoted by y_i^* and x_i^*, respectively, for $i = 1, \ldots, n$. Then using (3.4.25) we have the sample data in Table 3.4.5 in terms of the new system of units, and these data are stored in the file **process2.dat** on the data disk. Using the data in Table 3.4.5 we obtain the estimates

of the slope and the intercept relative to the new system of units as

$$\hat{\beta}_0^* = -1.624 \qquad \hat{\beta}_1^* = 0.0077 \qquad \hat{\sigma}^* = 0.1137$$

Note that we have $X_I^* = a + bX_I$ and $Y_I^* = c + dY_I$, where $a = 32$, $b = 9/5 = 1.8$, $c = 0.0$, and $d = 1000/3600 = 0.2778$. We may now verify that

$$\frac{d}{b}\hat{\beta}_1 = \frac{0.2778}{1.8}(0.0499) = 0.0077 = \hat{\beta}_1^*$$

$$c + \frac{d}{b}(b\hat{\beta}_0 - a\hat{\beta}_1) = 0 + \frac{0.2778}{1.8}[(1.8)(-4.959) - (32)(0.0499)] = -1.624 = \hat{\beta}_0^*$$

$$|d|\hat{\sigma} = (0.2778)(0.4095) = 0.1137 = \hat{\sigma}^*$$

TABLE 3.4.4
Chemical Process Data for Example 3.4.1.

Run Number	Process Output (Y) (in kg/hour)	Process Temperature (X) (° C)
1	10.0	300
2	10.5	310
3	11.5	320
4	11.5	330
5	11.9	340
6	12.3	350
7	12.8	360
8	13.1	370
9	14.4	380
10	13.9	390
11	15.7	400

TABLE 3.4.5
Transformed Data for Table 3.4.4.

Run Number	Process Output (Y^*) (g/s)	Process Temperature (X^*) (° F)
1	2.77778	572
2	2.91667	590
3	3.19444	608
4	3.19444	626
5	3.30556	644
6	3.41667	662
7	3.55556	680
8	3.63889	698
9	4.00000	716
10	3.86111	734
11	4.36111	752

Thus it is a simple matter to switch from one system of units to another as long as the two measurement systems are *linearly related.* ■

Point estimates for β_0, β_1, linear functions of β_0 and β_1, $\mu_Y(x)$, $Y(x)$, and σ, discussed in this section, are best estimates if assumptions (A) or (B) are satisfied. Valid estimates of μ_Y and σ_Y (respectively, μ_X and σ_X) are obtained using (1.6.1) and (1.6.2) provided the data are obtained by simple random sampling. These estimates are best estimates if assumptions (B) are satisfied.

Before applying any of the inference procedures discussed in this book, we recommend that the investigator carefully examine the data and, combined with his or her own prior experience and knowledge, make a judgment as to whether or not the assumptions underlying the inference procedures are at least approximately met. Statistical procedures are not meant to take the place of good subject matter judgment but to assist in this judgment.

In the next section we present some simple graphical procedures for examining the validity of some of the assumptions underlying regression. But first we explain in the following conversation the difference between fitting straight lines to data using the method of least squares and fitting straight lines by eye.

Conversation 3.4

Investigator: Good morning! Do you have time to talk to me now?

Statistician: Certainly. What's on your mind?

Investigator: A scientist I work with has a large data set that includes two variables Y and X. I plotted the data and drew a line through it by eye. It seems to him and to me that the line is a good summarization of the data. The line goes through $\bar{y} = 11.7$ and $\bar{x} = 11.7$, the means of the Y values and X values in the data set. The slope of the line is approximately (as close as we can estimate it by eye) 1.0; i.e., $\hat{\beta}_1^{eye} = 1.0$. The plot of the data and the line are shown in Figure 3.4.3. Figure 3.4.4 shows the plotted data with the line we fitted by eye (solid line) and the least squares line (dashed line). The least squares line has slope $\hat{\beta}_1 = 0.734$, which is considerably different from the slope $\hat{\beta}_1^{eye} = 1$ of the visually fitted line. Isn't this unusual?

Statistician: No, it's not unusual. If the data were obtained from a two-variable Gaussian population by simple random sampling, you would expect the data to be roughly elliptical in shape. The ellipse, the visually fitted line, and the least squares line are shown in Figure 3.4.5. The solid line that you drew by eye is a line such that the data are symmetrical around it: i.e., a line of symmetry. However, the least squares line (the dashed line) is the line that minimizes the sum of squares of *vertical* distances from the line to each point, and *the least squares line will always have a slope that is closer to zero than the line of symmetry (unless, of course, both slopes are zero).*

Both lines go through the point $(\bar{x}, \bar{y})$, where $\bar{x}$ and $\bar{y}$ are the means of the X and Y values in the data set.

Investigator: I see. But it seems to me that the line of symmetry summarizes the data better than the least squares line.

Statistician: I agree that the line of symmetry is more appealing to the eye. However, if assumptions (A) or (B) are satisfied, the least squares line is the *best line for predicting Y using X*. Thus one must be careful in using a visually fitted line.

Investigator: I think I understand. Perhaps I will come to see you again next week.

FIGURE 3.4.3

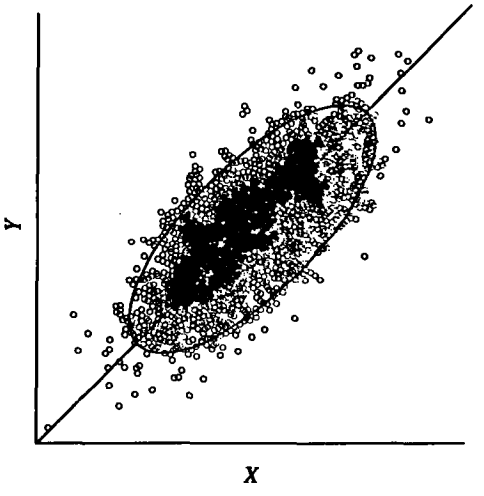


FIGURE 3.4.4

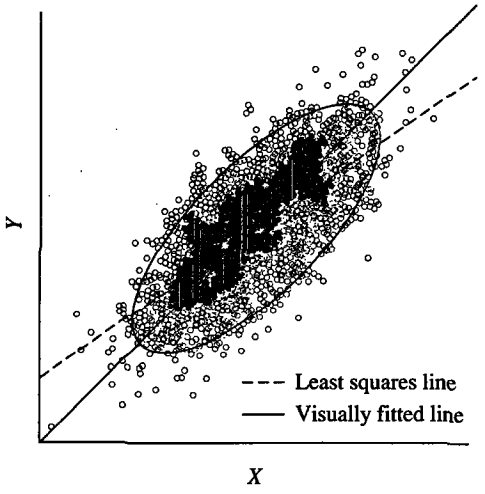
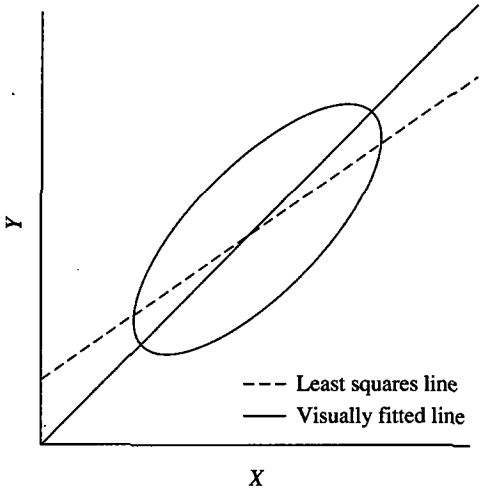


FIGURE 3.4.5



Problems 3.4

3.4.1 Consider the crystal data in Task 3.4.1. A different laboratory collects data on crystal growth at 14 preselected times. The data are given in Table 3.4.6 and are also in the file **Table346.dat** on the data disk. Assumptions (A) are presumed to be valid, and the data are obtained by sampling with preselected X values. Calculate the least squares estimates of β_0, β_1, and σ.

3.4.2 Using the results of Problem 3.4.1, answer questions (1)–(3) of Task 3.4.1.

3.4.3 In Problem 3.4.1, what is the estimate of the average weight of crystals that have grown for 19 hours? For 25 hours? For 40 hours?

3.4.4 Twenty-five cars are selected, using simple random sampling, from the car data in the file **car.dat** on the data disk (see also Table D-1 in Appendix D). The first-year maintenance cost Y for these 25 cars and the number of miles X they were driven during the first year after purchase were recorded. The following quantities were computed using these sample data: $\bar{x} = 11{,}364$; $\bar{y} = 532.44$; $SSX = 224{,}617{,}600$; $SSY = 279{,}764$; $SXY = 7{,}391{,}396$.

a Find $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\mu}_Y(x)$.

b A prospective new car buyer wants to purchase a car similar to those in the population of Table D-1 and plans to drive it 13,000 miles during the first year. What population quantity is the buyer interested in to predict the first-year maintenance cost for this car, $Y(13{,}000)$ or $\mu_Y(13{,}000)$? Explain. Obtain a point estimate for this quantity.

TABLE 3.4.6

Crystal Number	Weight Y (in grams)	Time X (in hours)
1	0.10	2
2	1.01	4
3	3.89	6
4	5.14	8
5	5.19	10
6	6.89	12
7	5.29	14
8	8.70	16
9	9.42	18
10	11.38	20
11	11.38	22
12	11.73	24
13	12.95	26
14	15.10	28

c What population quantity should buyers be interested in if they want to know the average first-year maintenance cost of all cars to be driven 16,000 miles next year? Explain.

3.4.5 Consider the simple random sample of size 26 for the grades example given in Table 3.2.2. These data are in the file **grades26.dat** on the data disk. For these data compute the following quantities: $\bar{x}$; $\bar{y}$; *SXY*; *SSX*; *SSY*; $\hat{\beta}_0$; $\hat{\beta}_1$; *SSE*; and $\hat{\sigma}$.

3.4.6 In Problem 3.4.5, compute $\hat{\mu}_Y(10)$. Using $\hat{\mu}_Y(10)$ in place of $\mu_Y(10)$ and $\hat{\sigma}$ in place of $\sigma_Y(10)$, answer question (4) of (3.2.1).

3.4.7 Sometimes it is desirable to work with the transformed data

$$x_i^* = x_i - c \qquad y_i^* = y_i - d$$

for suitably chosen constants c and d rather than the original data when the computations are done using a hand-held calculator. Many software packages also transform the data in this manner to combat rounding errors. The quantities *SSX*, *SSY*, and *SXY* can be shown to be the same whether the calculations are done using the untransformed data or the transformed data. Thus the estimates of the slope parameter β_1 and the subpopulation standard deviation σ, calculated using the transformed data, can be shown to be the same as those calculated with the untransformed data. Once $\hat{\beta}_1$ has been calculated, $\hat{\beta}_0$ can be obtained using (3.4.9). To become familiar with this transformation do the following.

a Compute y_i^* and x_i^* for the data of Problem 3.4.1 using $c = 14$ and $d = 7$.

b Using the transformed data in (a), compute *SSX*, *SSY*, and *SXY*. Using these compute $\hat{\beta}_1^*$ and $\hat{\sigma}^*$. How do these compare with $\hat{\beta}_1$ and $\hat{\sigma}$ obtained in Problem 3.4.1?

3.5 Checking Assumptions

Inference procedures for regression analyses are strictly valid only when the model assumptions on which the procedures are based are satisfied. However, models are approximations to reality, and so model assumptions will never hold exactly. Nevertheless, if a model is a reasonable approximation of reality, then inferences based on the model may be adequate for real applications.

In an applied problem, an investigator usually knows that some of the assumptions are satisfied, but she cannot be sure if others are. For example, an investigator will generally know whether (sample) assumptions (A) or (B) are satisfied, but she may not know if the (population) assumptions are satisfied. In this section we discuss some procedures for examining the validity of the population assumptions.

For convenience we reproduce here the population regression model given in (3.3.2):

$$Y_I = \beta_0 + \beta_1 X_I + E_I$$

If assumptions (A) or (B) are satisfied, then the population of all E_I is a Gaussian population with mean equal to *zero* and standard deviation equal to σ (i.e., $\sigma_{Y|X}$).

Hence we would like to examine the E_I to determine if they indeed form a Gaussian population with zero mean. However the E_I are population values and are unavailable, so we consider examining the n values of E_I, which we denote by $e_1, \ldots, e_n$, corresponding to the n sample items. However, even these e_i's are not observable, but we can estimate them. An estimate of the e_i corresponding to the ith sample item is the residual $\hat{e}_i$, defined by (see (3.4.5))

$$\hat{e}_i = y_i - \hat{\mu}_Y(x_i) = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \qquad \text{for } i = 1, \ldots, n \tag{3.5.1}$$

These residuals are used to help decide whether the E_I form a Gaussian population with mean zero. If we decide that the E_I do not constitute such a population, then we can conclude that one or more of assumptions (A) or (B) are violated.

We begin by discussing some graphical procedures which—by examining the sample data, the residuals $\hat{e}_i$, and other related quantities—will help us in detecting major violations in model assumptions.

Scatter Plot of Y Against X

You should routinely plot the sample Y values against the corresponding X values and study this plot carefully before using any of the inference procedures for straight line regression. Failure of the population assumption that $\mu_Y(x)$ is of the form $\beta_0 + \beta_1 x$ is often revealed by such a plot because this assumption states that the graph of the regression function of Y on X is a straight line. If this assumption is correct, then the plotted points should all lie roughly along a straight line.

Figure 3.5.1 is a plot of the sample Y values against the corresponding X values in the crystal data of Task 3.4.1. This plot seems to support the assumption that the regression function of Y on X is of the form $\mu_Y(x) = \beta_0 + \beta_1 x$. On the other hand, if the plots were as shown in Figures 3.5.2–3.5.4, this would tend to suggest that the regression function of Y on X is not of the form $\mu_Y(x) = \beta_0 + \beta_1 x$. The plot in Figure 3.5.4 is particularly interesting because, if we ignore the two points indicated by the symbols +, all remaining points would tend to support the assumption of a straight line regression function. This is not an uncommon situation in practice. Such points are often referred to as **outliers**. We discuss outliers, and how they should be dealt with, in greater detail in Chapter 5.

Standardized Residuals

We have stated that the residuals $\hat{e}_i$ are useful for checking the validity of the model assumptions. Numerous graphical and numerical techniques for checking assumptions using residuals can be found in the regression literature [1], [2]. Most of these

FIGURE 3.5.1

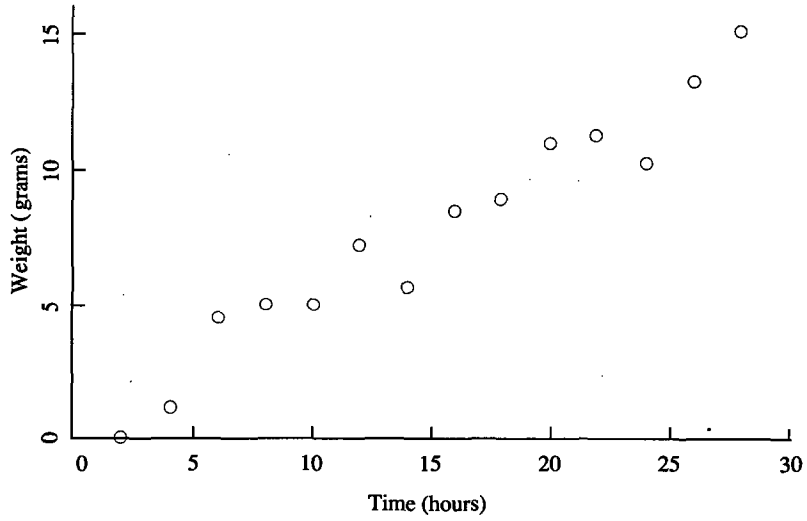


FIGURE 3.5.2

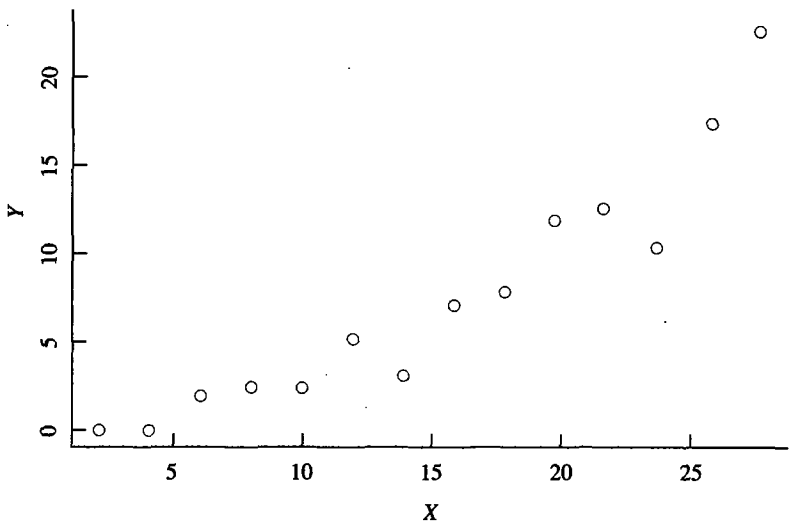


FIGURE 3.5.3

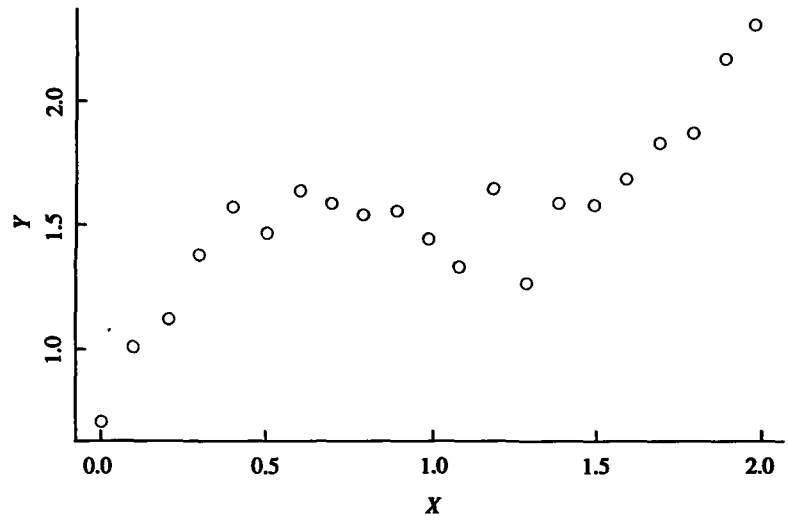
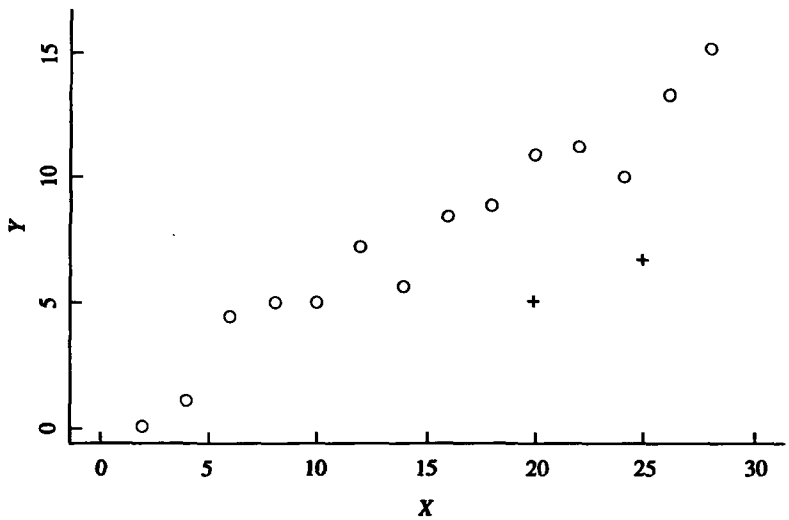


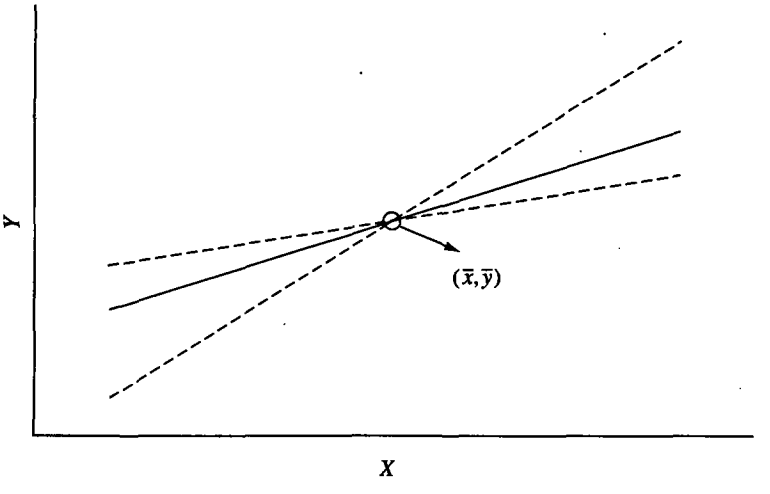
FIGURE 3.5.4



methods *standardize* the residuals $\hat{e}_i$ before using them because the residuals $\hat{e}_i$, corresponding to observations with x values that are far from the center of the range of x values, tend to vary more from sample to sample than those corresponding to observations with x values that are closer to the center. Consequently, they are not directly comparable with one another. The following explanation may help you understand the reason for this.

Recall that the estimated regression line must pass through the 'center' $(\bar{x}, \bar{y})$ of the sample data. A slight change in the estimated value of $\hat{\beta}_1$ will result in a greater change in the value of $\hat{e}_i$ for a point whose x value is far away from the middle (i.e., far from $\bar{x}$) than for a point that is closer to $\bar{x}$ (see Figure 3.5.5). You can think of the line as a see-saw with the point $(\bar{x}, \bar{y})$ (which itself varies from one sample to another) serving as the pivot. Points on the see-saw closer to the pivot move through a smaller vertical distance than points far from the pivot.

FIGURE 3.5.5



The residuals $\hat{e}_i$ may be made comparable with each other by standardizing them appropriately. It can be shown mathematically that the correct standardization procedure is to divide $\hat{e}_i$ by $\hat{\sigma}\sqrt{1 - h_{i,i}}$, for $i = 1, \ldots, n$, where

$$h_{i,i} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX} \tag{3.5.2}$$

and SSX is in (3.4.14). The quantities $h_{i,i}$ are usually referred to as **hat values**. More is said about them in Chapter 5.

We are thus led to define the **standardized residual** r_i for observation i as

$$r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{i,i}}} \tag{3.5.3}$$

If assumptions (A) or (B) are satisfied, then the standardized residuals $r_1, \ldots, r_n$ are (approximately) equivalent to a simple random sample of n observations from a Gaussian population with zero mean and unit standard deviation. (3.5.4)

This fact may be used to check the validity of some of the model assumptions.

Plotting Standardized Residuals Against Sample X Values

In some instances a plot of the standardized residuals r_i against x_i is more revealing than a plot of y_i against x_i. When assumptions (A) or (B) are satisfied, the points on the plot of r_i against x_i should be scattered in a random fashion about the horizontal line through the origin (see the dashed line in Figure 3.5.6 for an example), showing no obvious trends or other patterns. If this is found not to be the case, then one or more of these assumptions is likely to be false.

Residual plots corresponding to the scatter plots in Figures 3.5.1–3.5.4 are given in Figures 3.5.6–3.5.9, respectively. Note how the departures from linearity in the plots of Figures 3.5.2 and 3.5.3 appear more prominently in the corresponding residual plots in Figures 3.5.7 and 3.5.8. Also note that the two outliers in the plot shown in Figure 3.5.4 are easily spotted in the residual plot of Figure 3.5.9.

FIGURE 3.5.6

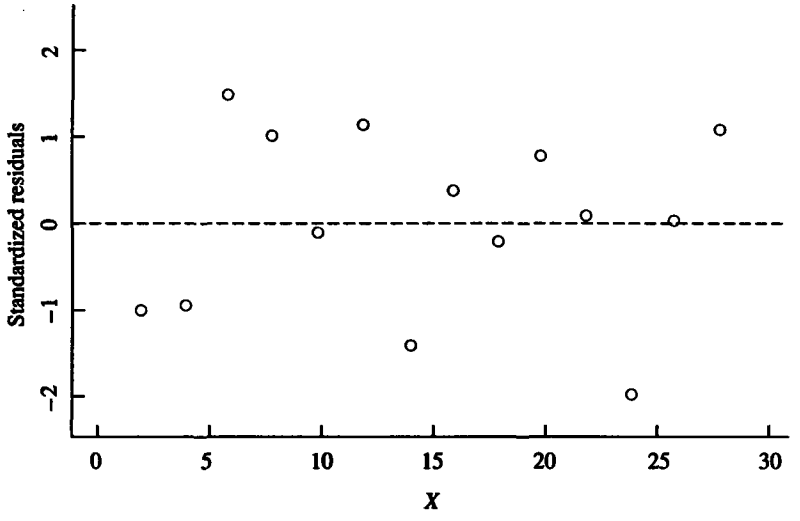


FIGURE 3.5.7

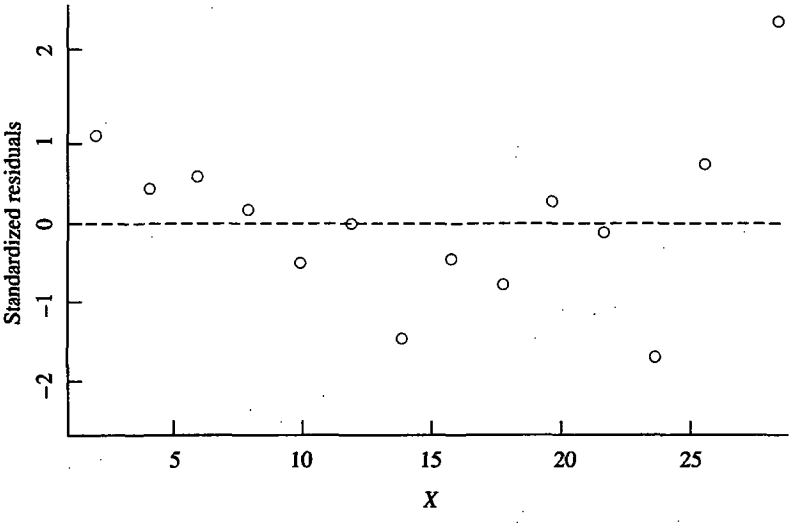


FIGURE 3.5.8

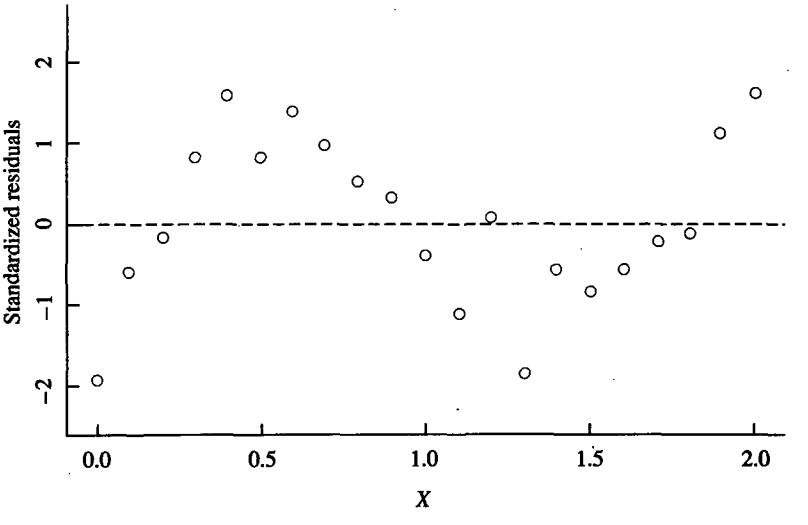
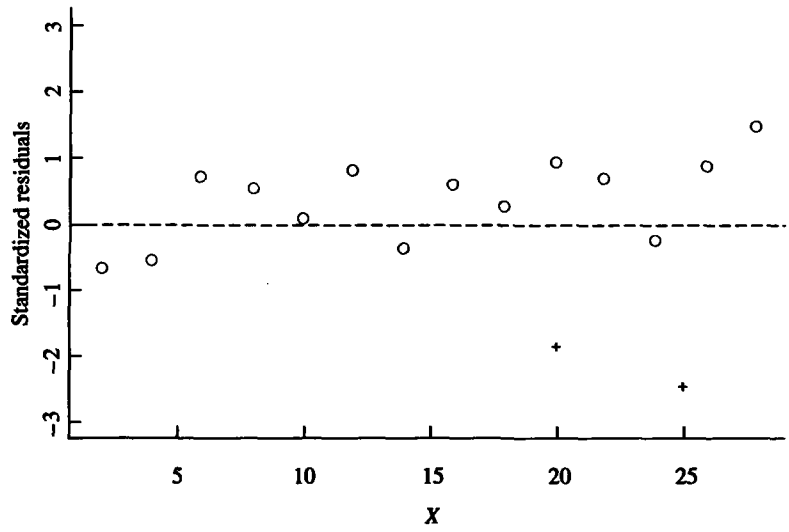


FIGURE 3.5.9



Plotting Standardized Residuals Against Fitted Values $\hat{\mu}_Y(x_i)$

The estimated subpopulation means $\hat{\mu}_Y(x_i)$ are often referred to as **fitted values** or **predicted values** corresponding to the X values, $x_1, \ldots, x_n$, in the sample.

The fitted values can be shown to be *unrelated* to (independent of) the standardized residuals r_i when assumptions (A) or (B) are satisfied. Hence the points in the plot of r_i versus $\hat{\mu}_Y(x_i)$ should be randomly scattered about the horizontal line through the origin (see the dashed line in Figure 3.5.10 for an example) with no particular pattern. Any systematic pattern observed would indicate that one or more of assumptions (A) or (B) may fail to hold.

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. In Figures 3.5.10–3.5.15 we show hypothetical plots of standardized residuals versus fitted values for illustration. In each plot the horizontal axis represents the fitted values, and the vertical axis represents standardized residuals.

Figure 3.5.10 is typical of a plot we expect when there are no violations of assumptions. Figures 3.5.11, 3.5.12, and 3.5.13 indicate a possible violation of the assumption of *homogeneity of standard deviations*; i.e., they suggest that the standard deviations of the subpopulations of Y values, determined by the predictor variable X, may not all be the same. Figure 3.5.11 suggests that the standard deviations of the Y values in the various subpopulations increase with increasing values of the subpopulation means, and Figure 3.5.12 indicates that the standard deviations of the Y values in the various subpopulations decrease with increasing values of the

FIGURE 3.5.10

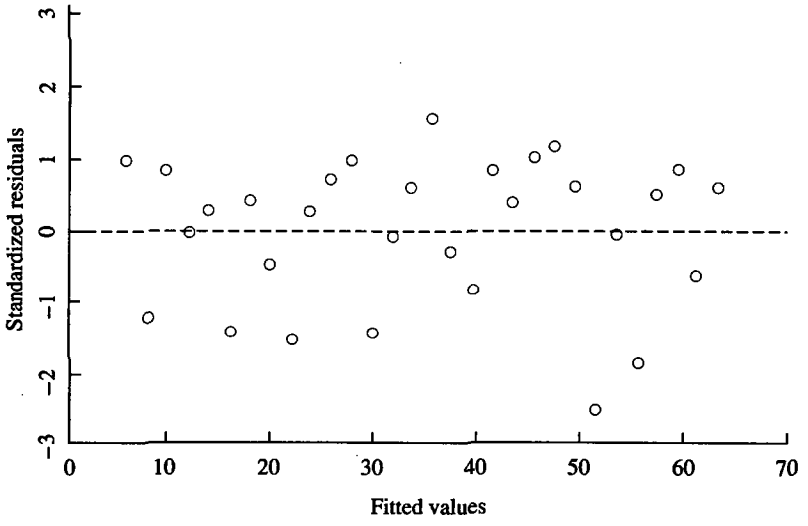


FIGURE 3.5.11

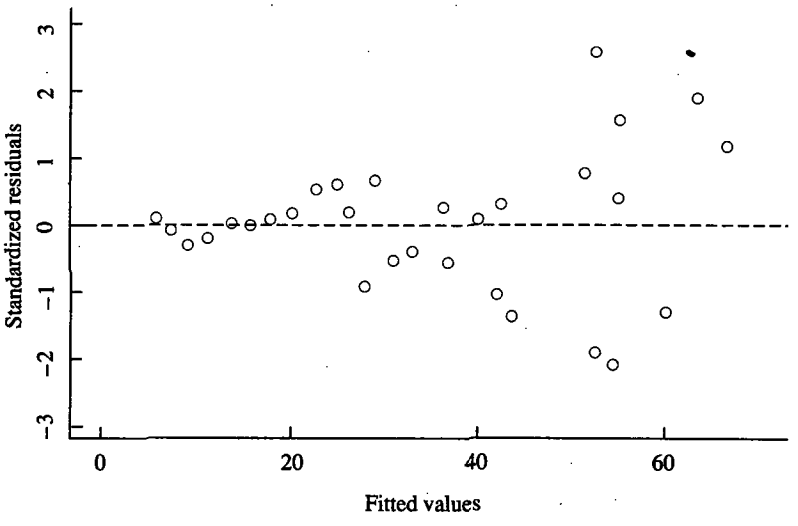


FIGURE 3.5.12

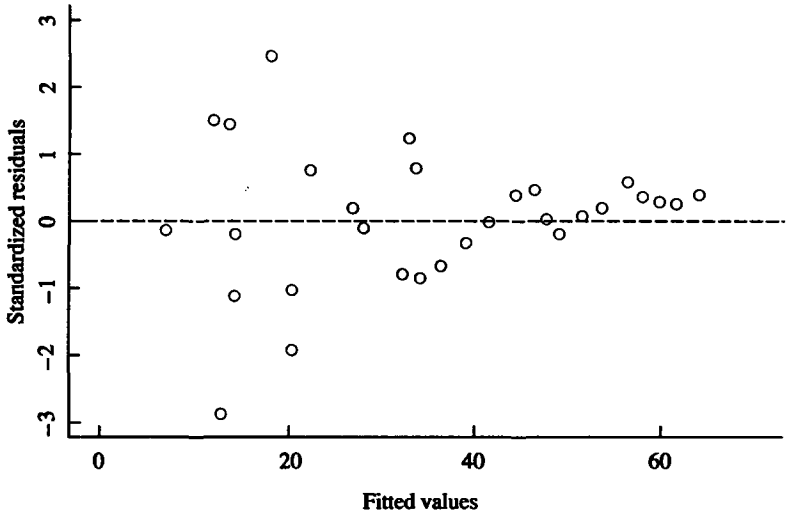


FIGURE 3.5.13

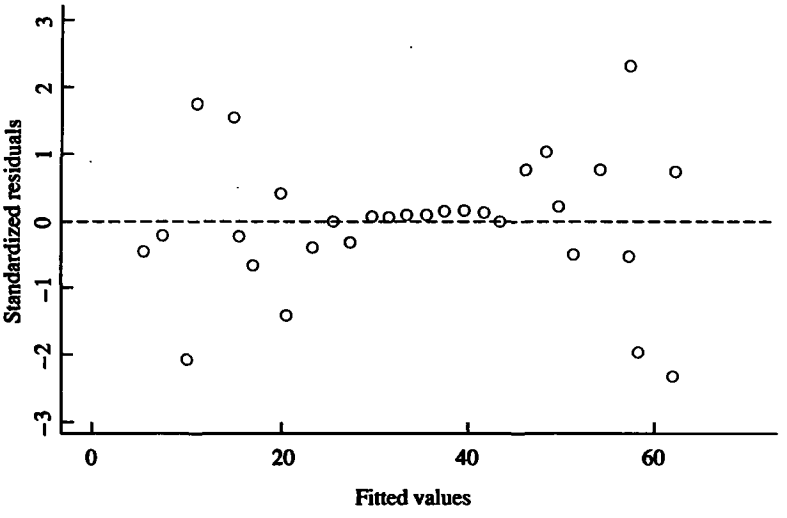


FIGURE 3.5.14

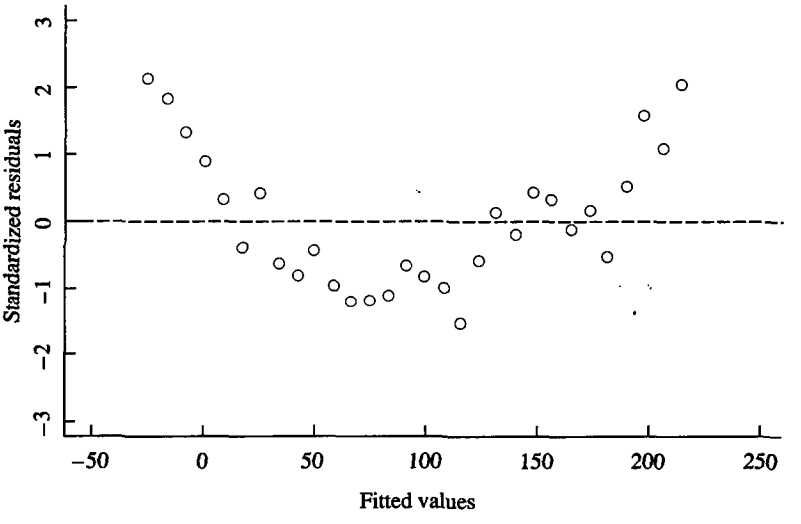
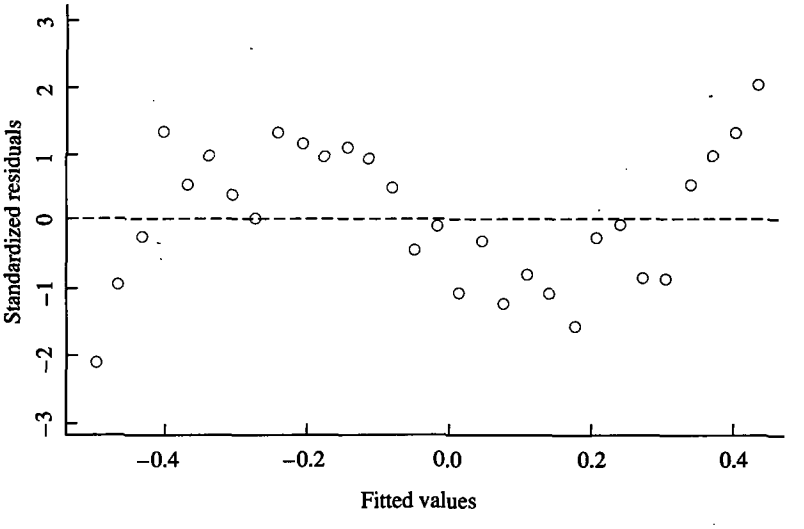


FIGURE 3.5.15



subpopulation means. Figure 3.5.13, on the other hand, suggests that the standard deviations are larger for subpopulations with small means as well as those with large

means, whereas the standard deviations are smaller for subpopulations with means near the middle of the possible range of values.

Figures 3.5.14 and 3.5.15 show plots of standardized residuals r_i versus fitted values $\hat{\mu}_{Y}(x_i)$, which indicate **lack-of-fit** (i.e., they indicate that the graph of the regression function of Y on X is not a straight line). A plot of y_i against x_i, or a plot of r_i against x_i, should also reveal this. When lack of fit is indicated by one of these plots, the investigator may want to formulate an alternate candidate for the regression function, i.e., a quadratic function, a cubic function, a logarithmic function, an exponential function, etc.

Gaussian Rankit-Plot

The **Gaussian rankit-plot** is a graphical tool for checking whether a given set of numbers appears to be a simple random sample from a Gaussian population. Although we primarily use this procedure to assess whether or not the standardized residuals from a regression analysis appear to be a simple random sample from a Gaussian population, the procedure is applicable to any set of numbers. Hereafter, unless otherwise stated, when we say **rankit-plot** we mean a **Gaussian rankit-plot.** The procedure is as follows.

To obtain a rankit-plot of a given set of numbers, say $y_1, \ldots, y_n$, we must first arrange these numbers in increasing order. The ordered data values are denoted by $y_{(1)}, \ldots, y_{(n)}$, with $y_{(1)}$ denoting the smallest of the y_i and $y_{(n)}$ the largest. A rankit-plot of the data values $y_1, \ldots, y_n$ is a plot of the ordered data values $y_{(1)}, \ldots, y_{(n)}$ against the quantities $z_1^{(n)}, \ldots, z_n^{(n)}$, called **Gaussian scores** (or **normal scores**, or **rankits**), which may be thought of as a typical sample of size n (arranged in increasing order) from a *standard* Gaussian population (a Gaussian population with mean zero and standard deviation one). If $y_1, \ldots, y_n$ is a simple random sample from a Gaussian population, then the plot of $y_{(i)}$ against $z_i^{(n)}$ should produce points that more or less all lie on a straight line. If, in addition, the y_i come from a Gaussian population with zero mean and unit standard deviation, then this line should more or less be a line through the origin with unit slope.

Figure 3.5.16 shows the rankit-plot of the data from a simple random sample of size 25 from a Gaussian population with mean 2 and standard deviation 3, while Figure 3.5.17 shows the rankit-plot for a simple random sample of size 30 from a *standard* Gaussian population. In contrast, Figure 3.5.18 is the rankit-plot for a simple random sample of size 28 from a population that is not Gaussian.

Notice how the points in Figure 3.5.16 all lie approximately on a straight line (the dashed line), and the points in Figure 3.5.17 all lie approximately on a straight line through the origin with unit slope (dotted line), but the points in Figure 3.5.18 clearly show a systematic departure from a straight line pattern.

FIGURE 3.5.16

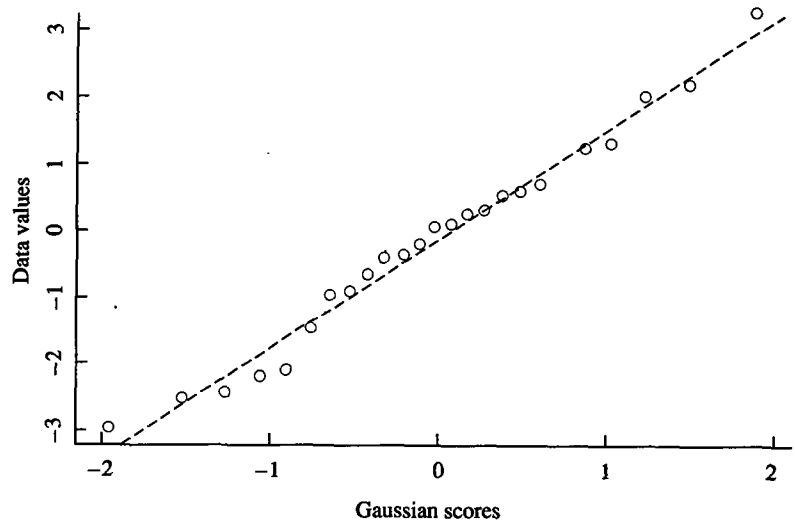


FIGURE 3.5.17

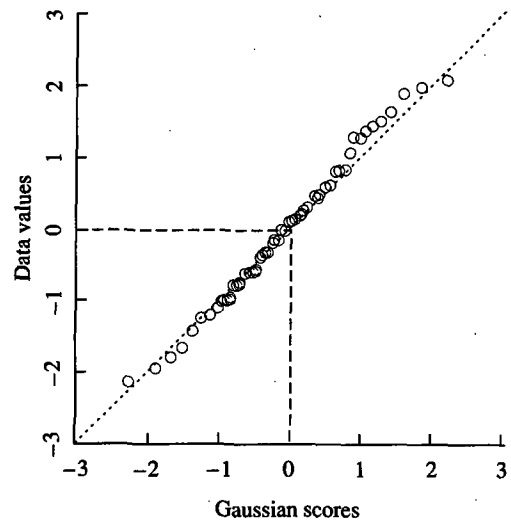
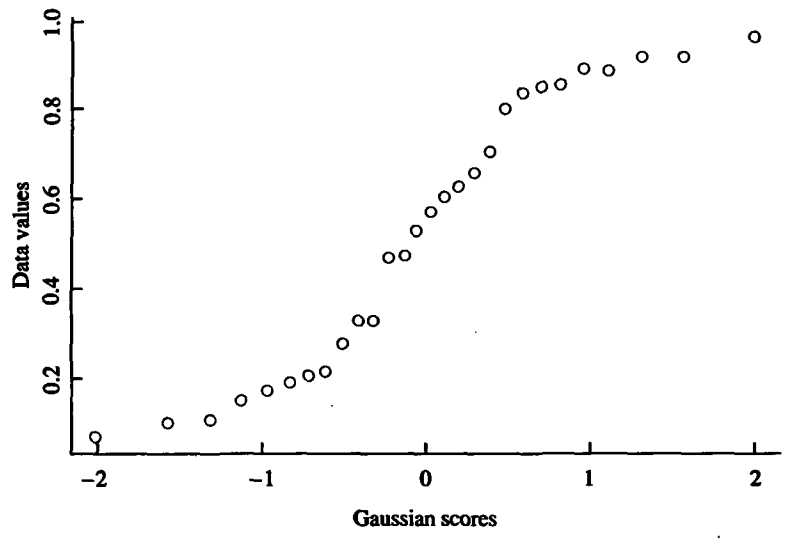


FIGURE 3.5.18



As stated in (3.5.4), when assumptions (A) or (B) are satisfied, the standardized residuals $r_1, \ldots, r_n$ are (approximately) equivalent to a simple random sample of n observations from a standard Gaussian population. So a Gaussian rankit-plot can be used to help determine if assumptions (A) or (B) appear to be violated. We discuss this in some detail, but first we discuss Gaussian scores.

Gaussian Scores (or Normal Scores, or Rankits)

We now explain the Gaussian scores $z_i^{(n)}$. Suppose we repeatedly obtain simple random samples of size n from a Gaussian population with mean zero and unit standard deviation (from the population $\{Z\}$, say) and order the sample values from smallest to largest. Suppose the first sample yields the ordered observations

$$z_{1,1} \leq \cdots \leq z_{1,n}$$

and the second sample yields

$$z_{2,1} \leq \cdots \leq z_{2,n}$$

and so forth. If we repeat the process a large number of times, say m times, (i.e., m samples of size n) we will obtain m sets of ordered samples of size n each, which may be organized as in (3.5.5).

$$
\begin{array}{lccccccccc}
 & z_{1,1} & < & \cdots & < & z_{1,k} & < & \cdots & < & z_{1,n} \\
 & z_{2,1} & < & \cdots & < & z_{2,k} & < & \cdots & < & z_{2,n} \\
 & \vdots & < & \vdots & < & \vdots & < & \vdots & < & \vdots \\
 & z_{j,1} & < & \cdots & < & z_{j,k} & < & \cdots & < & z_{j,n} \\
 & \vdots & < & \vdots & < & \vdots & < & \vdots & < & \vdots \\
 & z_{m,1} & < & \cdots & < & z_{m,k} & < & \cdots & < & z_{m,n} \\
\hline
\text{Means} & z_1^{(n)} & < & \cdots & < & z_k^{(n)} & < & \cdots & < & z_n^{(n)}
\end{array}
\tag{3.5.5}
$$

Here $z_1^{(n)}$ is the mean of the *smallest* values obtained in each of these m simple random samples of size n, and $z_2^{(n)}$ is the mean of the 2nd smallest values obtained in each of these m random samples of size n, etc. In general, $z_k^{(n)}$ refers to the average of the kth smallest values in simple random samples of size n from a standard Gaussian population; i.e.,

$$z_k^{(n)} = \frac{1}{m} \sum_{j=1}^{m} z_{j,k}$$

where $z_{j,k}$ are as in (3.5.5). The values $\{z_k^{(n)}\}$ are called Gaussian scores or normal scores or rankits, and they can be obtained from various computer programs. In Section 3.5 of the MINITAB and SAS laboratory manuals we show how to use computer commands to obtain the numbers

$$z_1^{(n)} < z_2^{(n)} < \cdots < z_n^{(n)} \tag{3.5.6}$$

Rankit-Plot of Standardized Residuals

Recall from (3.5.4) that

if assumptions (A) or (B) are satisfied, then the standardized residuals r_i are (approximately) equivalent to a simple random sample from a standard Gaussian population. This can be checked graphically by examining a rankit-plot of the ordered standardized residuals, which should be (approximately) a straight line through the origin with slope equal to 1 when assumptions (A) or (B) hold.

To obtain a rankit-plot of the standardized residuals, we first order the standardized residuals r_i and denote the ordered values by $r_{(1)}, r_{(2)}, \cdots, r_{(n)}$ so that we have

$$r_{(1)} < r_{(2)} < \cdots < r_{(n)}$$

Then we compare these with the Gaussian scores in (3.5.6) by plotting the pairs $(z_i^{(n)}, r_{(i)})$ for $i = 1, \ldots, n$, with $r_{(i)}$ as ordinates and $z_i^{(n)}$ as abscissas. If the plotted points appear to deviate systematically from the line through the origin with slope equal to 1 (which should be plotted on the same graph for ease of reference), this indicates that some or all of the assumptions may be violated. Tables that contain the values $z_i^{(n)}$ for various values of n are available, but the rankit-plot of standardized residuals is more conveniently obtained by computer.

Rankit-Plots of Linear Combinations of y_i and x_i

For inferences on certain parameters (for instance the correlation coefficient $\rho_{Y,X}$ to be discussed in Section 3.9), we need assumptions (B) in Box 3.3.2. In particular, the population $\{(Y, X)\}$ is assumed to be bivariate Gaussian. To check this assumption we must investigate whether or not the sample data, $(y_1, x_1), \ldots, (y_n, x_n)$, appear to be a simple random sample from a bivariate Gaussian population. This can be done by examining the rankit-plots of linear combinations of y_i and x_i. Based on the discussions in Section 1.9, we know that a population $\{(Y, X)\}$ is bivariate Gaussian if and only if the univariate population $\{aY + bX\}$ is Gaussian for every possible choice of values for a and b. Because we have only sample data available, we can examine linear combinations $ay_i + bx_i$ of y_i and x_i in the sample. In practice it is impossible to examine *every* linear combination of y_i and x_i, but we can consider *several* linear combinations and examine the corresponding rankit-plots. For instance, if $v_i = ay_i + bx_i$ is one such linear combination, then we can examine a rankit-plot of the v_i and form an opinion about whether the v_i values appear to be a simple random sample from a Gaussian population. If we consider several such linear combinations (several different values of a and b), and if every linear combination we consider appears to be a simple random sample from a Gaussian population, then we can feel somewhat assured that the bivariate population $\{(Y, X)\}$ under consideration is at least approximately bivariate Gaussian. We give a numerical illustration of this procedure in Example 3.5.2, which is one of two examples we use to illustrate the procedures discussed in this section.

You should study a large number of rankit-plots corresponding to simple random samples from Gaussian populations, as well as non-Gaussian populations, to gain experience in judging the plots and deciding whether or not the plot indicates violations of the assumption of a Gaussian random sample.

For convenience, we list the formulas for computing $\hat{\mu}_Y(x_i)$, $\hat{e}_i$, $h_{i,i}$, and r_i, in Box 3.5.1.

BOX 3.5.1 Formulas for Computing $\hat{\mu}_Y(x_i)$, $\hat{e}_i$, $h_{i,i}$, and r_i

Fitted values $\hat{\mu}_Y(x_i)$:

$$\hat{\mu}_Y(x_i) = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Residuals $\hat{e}_i$:

$$\hat{e}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i = y_i - \hat{\mu}_Y(x_i)$$

Hat values $h_{i,i}$:

$$h_{i,i} = \left[\frac{1}{n} + \frac{(x_i - \bar{x})^2}{SSX}\right]$$

Standardized residuals r_i:

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

EXAMPLE 3.5.1

We illustrate the preceding discussions using the crystal growth data in Task 3.4.1; the data are in the file **crystal.dat** on the data disk. Recall that for these data $n = 14$, $\hat{\sigma} = 1.062$, $\bar{x} = 15$, $SSX = 910$, and the sample regression function is

$$\hat{\mu}_Y(x) = 0.0014 + 0.5034x$$

Since the X values in this data set are preselected, we know that assumptions (B) are not satisfied, so we examine the data to determine if assumptions (A) appear to be satisfied.

In Figure 3.5.19 we plot y_i, the crystal weights, against x_i, the time it takes to grow to their final sizes. This plot seems to support the assumption that the graph of the regression function of Y on X is a straight line. We calculate the fitted values $\hat{\mu}_Y(x_i)$, the residuals $\hat{e}_i$, the hat values $h_{i,i}$, and the standardized residuals r_i using the formulas in Box 3.5.1, and we list them in Table 3.5.1 along with x_i and y_i.

FIGURE 3.5.19

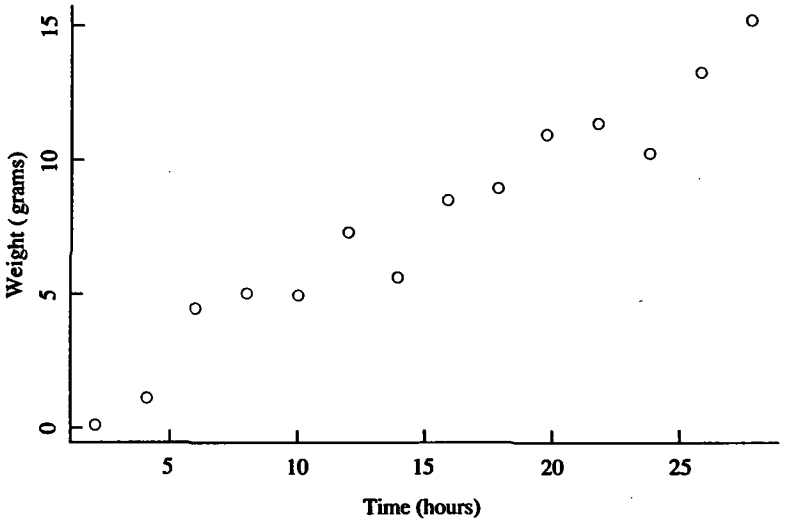


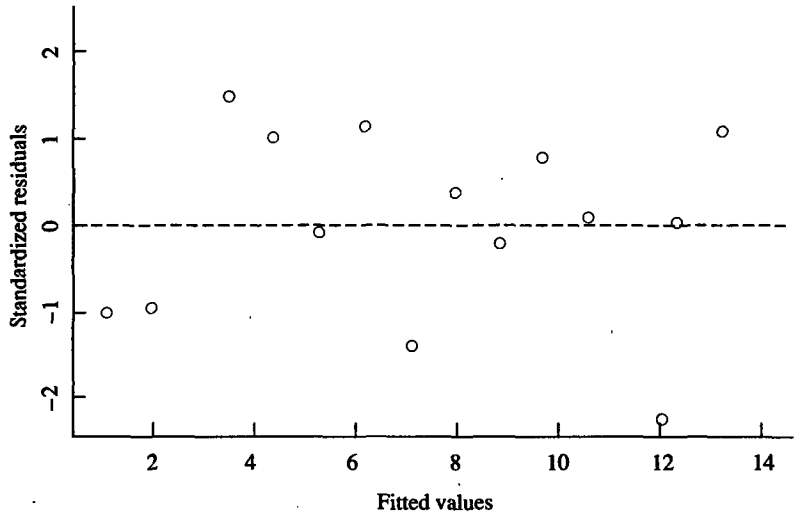
TABLE 3.5.1

Fitted Values, Residuals, Hat Values, and Standardized Residuals for Crystal Data

Weights y_i	Times x_i	Fitted Values $\hat{\mu}_Y(x_i)$	Residuals $\hat{e}_i$	Hat Values $h_{i,i}$	Standardized Residuals r_i
0.08	2	1.0083	−0.92829	0.257143	−1.01438
1.12	4	2.0151	−0.89514	0.204396	−0.94518
4.43	6	3.0220	1.40800	0.160440	1.44726
4.98	8	4.0289	0.95114	0.125275	0.95781
4.92	10	5.0357	−0.11571	0.098901	−0.11481
7.18	12	6.0426	1.13743	0.081319	1.11767
5.57	14	7.0494	−1.47943	0.072527	−1.44682
8.40	16	8.0563	0.34371	0.072527	0.33614
8.81	18	9.0631	−0.25314	0.081319	−0.24874
10.81	20	10.0700	0.74000	0.098901	0.73420
11.16	22	11.0769	0.08314	0.125275	0.08373
10.12	24	12.0837	−1.96371	0.160440	−2.01847
13.12	26	13.0906	0.02943	0.204396	0.03107
15.04	28	14.0974	0.94257	0.257143	1.02999

In Figure 3.5.20 we plot the standardized residuals r_i against the the fitted values $\hat{\mu}_Y(x_i)$.

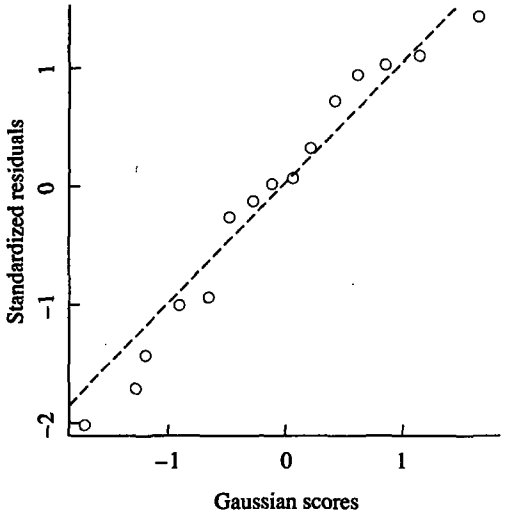
FIGURE 3.5.20



Notice that the points on this plot seem to be randomly scattered about the horizontal dashed line and show no apparent pattern. Based on this plot we do not see any evidence of violations of any of assumptions (A).

Finally we obtain the rankit-plot of r_i shown in Figure 3.5.21.

FIGURE 3.5.21



The points on this plot seem to all lie (approximately) on a straight line of unit slope through the origin (the dashed line in Figure 3.5.21), and so this plot is consistent with assumptions (A).

To summarize, for the crystal growth data in Table 3.4.2, the graphical procedures discussed in this section do not point to any violations of assumptions (A). (As stated earlier, we already know that assumptions (B) do *not* hold because the X values in this data set are preselected.) We can feel somewhat confident that the inference procedures discussed in this chapter, which are based on assumptions (A), are valid. ■

EXAMPLE 3.5.2

For another illustration, consider Example 2.2.2 where we are interested in the relationship of Y, the first-year maintenance cost of new cars, and X, the number of miles driven during the first year after purchase. Twenty cars were selected by simple random sampling from the population in Table D-1, in Appendix D, and we want to examine the plausibility of assumptions (A) or (B) for this population, *using only the sample data*. The data are given in Table 3.5.2 and are also stored in the file **car20.dat** on the data disk.

We begin by plotting y_i against x_i (see Figure 3.5.22) to see if it is reasonable to assume that the graph of the regression function of Y on X is a straight line. The plot seems to suggest that the regression function of Y on X may not be a straight line because there is some evidence of curvature. We should be able to better assess this possibility by examining a plot of the standardized residuals r_i against x_i and also of r_i against the fitted values $\hat{\mu}_Y(x_i)$. To do so, we first calculate the estimated regression function $\hat{\mu}_Y(x)$ using (3.4.10) and we get

$$\hat{\mu}_Y(x) = 177.01 + 0.031307x$$

Next we calculate the fitted values $\hat{\mu}_Y(x_i)$, residuals $\hat{e}_i$, hat values $h_{i,i}$, and standardized residuals r_i using formulas in Box 3.5.1. These are listed in Table 3.5.3.

Next we examine a plot of the standardized residuals r_i against x_i. This plot is given in Figure 3.5.23. It strongly suggests that the regression function of Y on X is not a straight line because the points are clearly not randomly scattered about the horizontal line corresponding to $r_i = 0$ (shown as the dotted line in Figure 3.5.23), and there seems to be clear evidence of curvature.

TABLE 3.5.2
Car20 Data (Sample Data from the Population in Table D-1)

Car	Maintenance Cost (in dollars) *Y*	Miles Driven *X*
1	456	11200
2	828	17300
3	500	11100
4	489	11000
5	387	6700
6	553	13700
7	531	12400
8	650	15300
9	475	11300
10	474	8200
11	533	12300
12	396	7700
13	618	14300
14	474	8800
15	639	13600
16	457	7100
17	460	8700
18	433	6500
19	621	13100
20	460	9900

FIGURE 3.5.22

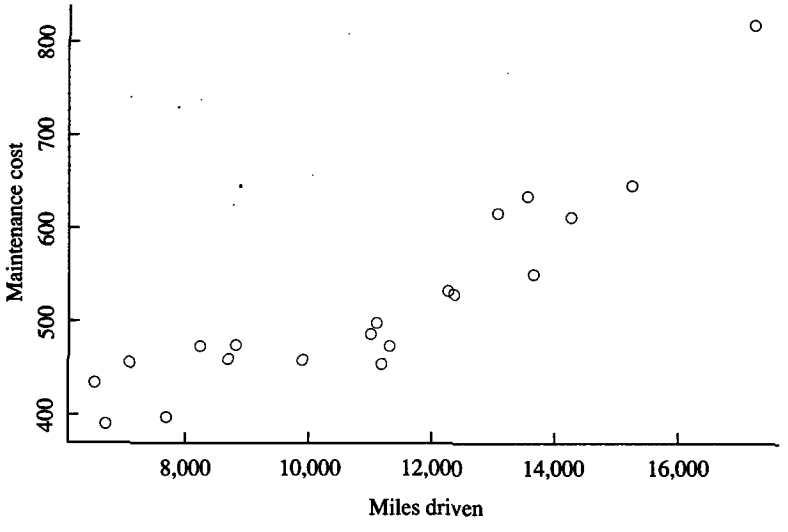


TABLE 3.5.3
Fitted Values, Residuals, Hat Values, and Standardized Residuals for car20 data

Maintenance Costs y_i	Miles Driven x_i	Fitted Values $\hat{\mu}_Y(x_i)$	Residuals e_i	Hat Values $h_{i,i}$	Standardized Residuals r_i
456	11200	527.648	−71.648	0.050206	−1.58529
828	17300	718.623	109.377	0.275645	2.77121
500	11100	524.518	−24.518	0.050046	−0.54243
489	11000	521.387	−32.387	0.050001	−0.71652
387	6700	386.766	0.234	.0.155945	0.00550
553	13700	605.917	−52.917	0.091269	−1.19700
531	12400	565.217	−34.217	0.061019	−0.76144
650	15300	656.008	−6.008	0.154964	−0.14094
475	11300	530.779	−55.779	0.050480	−1.23435
474	8200	433.726	40.274	0.095034	0.91290
533	12300	562.086	−29.086	0.059491	−0.64674
396	7700	418.073	−22.073	0.112486	−0.50523
618	14300	624.701	−6.701	0.111733	−0.15332
474	8800	452.511	21.489	0.077855	0.48254
639	13600	602.786	36.214	0.088258	0.81782
457	7100	399.288	57.712	0.137192	1.33975
460	8700	449.380	10.620	0.080433	0.23881
433	6500	380.504	52.496	0.166005	1.23954
621	13100	587.132	33.868	0.074912	0.75930
460	9900	486.949	−26.949	0.057027	−0.59842

FIGURE 3.5.23

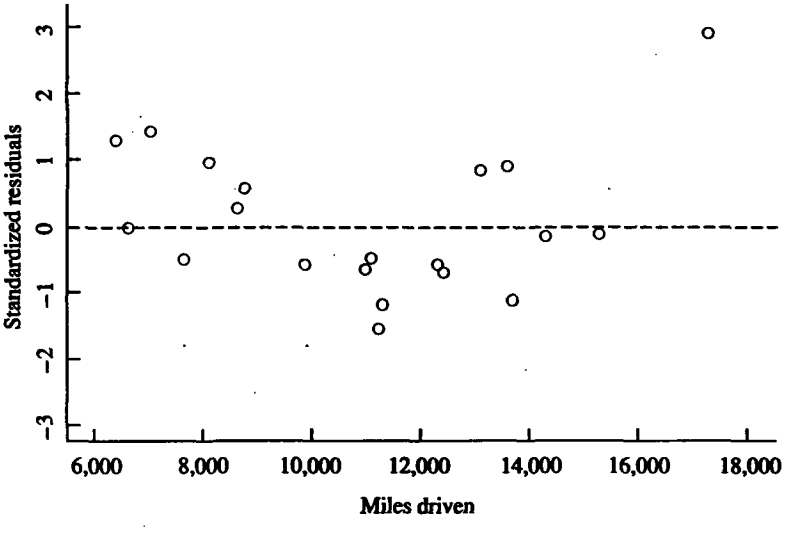
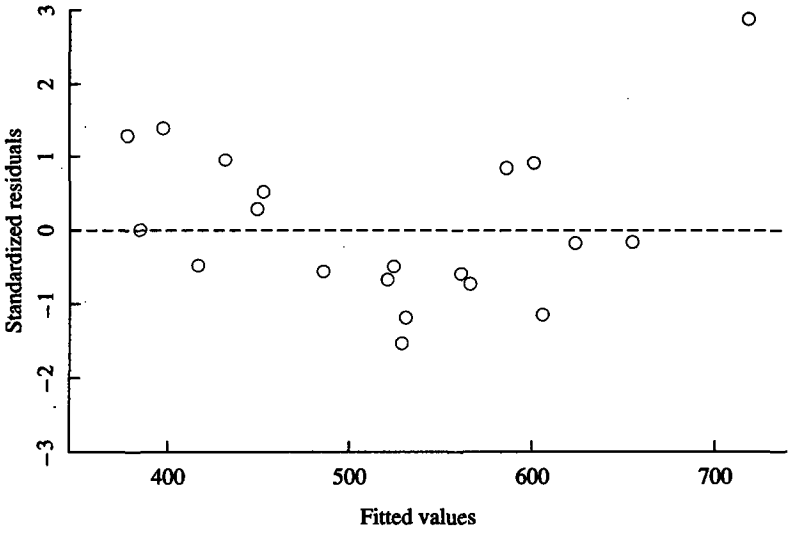


Figure 3.5.24 shows the plot of the standardized residuals r_i against fitted values $\hat{\mu}_Y(x_i)$. The points in this plot are also not randomly scattered about the horizontal dashed line; instead, they suggest the existence of curvature in the regression curve of Y on X.

FIGURE 3.5.24

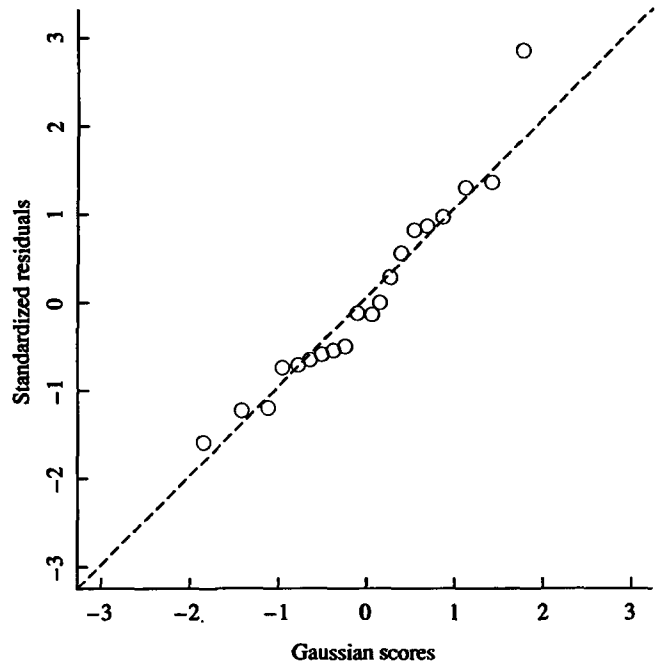


Thus we tentatively conclude that a straight line regression model is not appropriate for the data of this example. You should refer to Figure 2.3.4, which shows a scatter-plot of maintenance cost against miles driven for the population in Task 2.3.1 from which these data are sampled. Notice that the population scatter-plot definitely exhibits curvature. This should give you some confidence that a sample, properly selected, does indeed reflect aspects of the population.

For illustrative purposes, we now examine a rankit-plot of the standardized residuals in Figure 3.5.25 (the dashed line is the line through the origin with unit slope). There does not seem to be anything unusual about this plot that would make us suspect any violations of the assumptions. So based on this plot alone we are unable to detect any departure from assumptions (A) or (B).

Because the sample data in this example are obtained using simple random sampling, it is reasonable to ask if assumptions (B) appear to be satisfied; i.e., it is reasonable to examine whether the data appear to be a simple random sample from a bivariate Gaussian population. This can be done by examining whether or not various linear combinations of y_i and x_i (including y_i and x_i themselves) appear to be a simple random sample from a univariate Gaussian population (see Section 1.9) by using the rankit-plot. We illustrate this by examining rankit-plots of y_i, x_i, and two

FIGURE 3.5.25



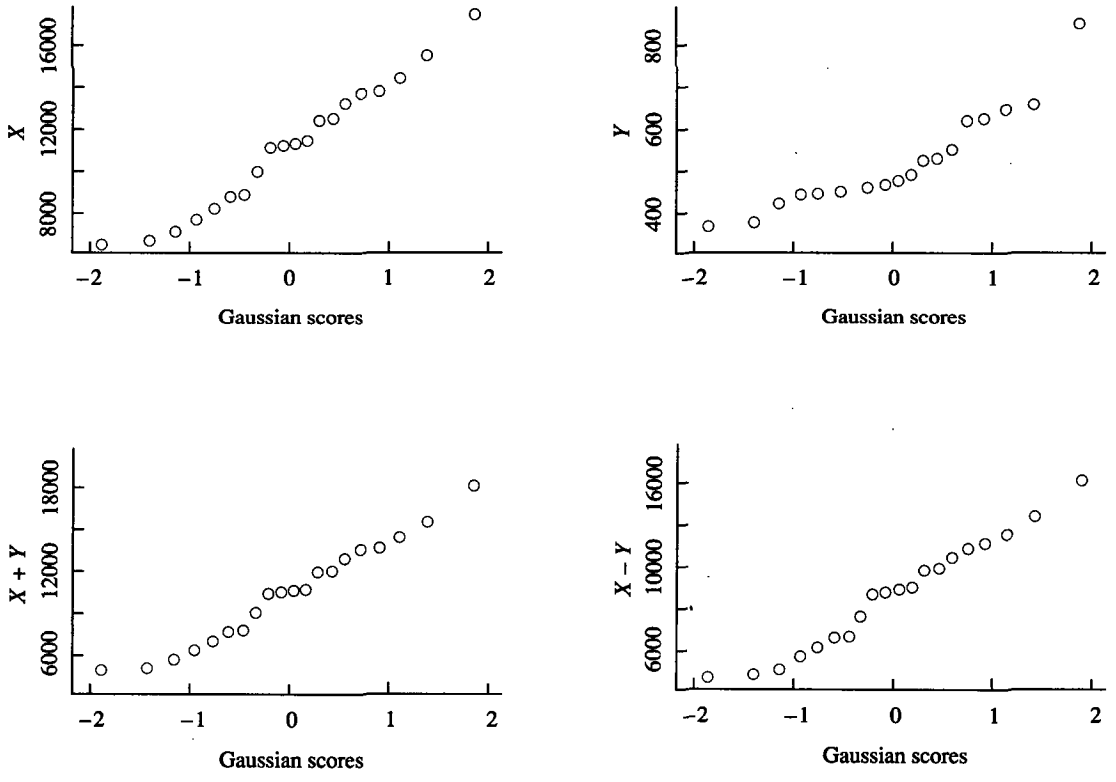
other linear combinations of y_i and x_i, namely $x_i + y_i$ and $x_i - y_i$. These four plots are given in Figure 3.5.26.

Based on these plots we conclude that it is not unreasonable to assume the x_i values are a simple random sample from a Gaussian population. The rankit-plot for Y shows a *hint* of curvature, thus suggesting that the population $\{Y\}$ may not be Gaussian.

The rankit-plots for $y_i + x_i$ and $x_i - y_i$ appear to be consistent with the Gaussian assumption. You are invited to obtain the rankit-plots for several other linear combinations of y_i and x_i for practice in judging them. However, because the sample y_i values do not appear to be a simple random sample from a Gaussian population, the data $(y_1, x_1), \ldots, (y_{20}, x_{20})$ appear not to be a simple random sample from a bivariate Gaussian population. ■

We should point out that if the population $\{(Y, X)\}$ is a bivariate Gaussian population, then the regression function of Y on X must necessarily be of the form $\mu_Y(x) = \beta_0 + \beta_1 x$. Because some of the residual plots have already supplied evidence against this possibility, we should have concluded that it is not very likely for the population $\{(Y, X)\}$ to be bivariate Gaussian. However, we proceeded with the examination of various linear combinations of y_i and x_i to illustrate the procedure.

FIGURE 3.5.26



Bear in mind that while some of the procedures for checking assumptions may not detect any violations, other procedures may. It is only with the combined evidence based on many different checks of assumptions that we can arrive at a conclusion regarding the validity (or at least, approximate validity) of the assumptions in any given situation.

Many diagnostic procedures exist for examining the validity of assumptions (A) or (B). 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 a necessarily subjective exercise, and at some point the investigator has to either conclude that assumptions (A) or (B) are satisfied (for all practical purposes) and proceed with analyses and inferences, or he/she must conclude that one or more of the assumptions are seriously violated and look for alternative procedures. For certain types of violations of assumptions, we discuss alternative procedures in later chapters.

Authors' Recommendation

When you perform a straight line regression analysis of a set of data $(y_1, x_1), \ldots, (y_n, x_n)$, we suggest you take the following steps:

1 Plot y_i against x_i. Assess the plausibility of a straight line regression model. If the plot indicates that a straight line regression model is not appropriate, then you may want to investigate alternative candidates for a regression function. Chapters 4–9 will be of assistance in this regard.

2 If a straight line regression model appears to be a plausible candidate, then calculate the sample regression function as discussed in Section 3.4. Obtain the standardized residuals r_i and the fitted values $\hat{\mu}_Y(x_i)$.

3 Plot r_i against x_i and r_i against the fitted values $\hat{\mu}_Y(x_i)$. Examine these plots for evidence of unequal subpopulation variances or of lack of fit of the model.

4 Obtain a Gaussian rankit-plot of r_i to evaluate the validity of the assumption that each subpopulation of Y values, determined by X, is a Gaussian population.

5 If you want to determine whether the bivariate population $\{(Y, X)\}$ is Gaussian, and if you obtained data by simple random sampling, then examine the Gaussian rankit-plots of y_i, x_i, and several linear combinations of y_i and x_i to assess whether or not the data appear to be a simple random sample from a bivariate Gaussian population.

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

Note: Checking assumptions as recommended here involves a great deal of computing, plotting, etc., which are easily done using one of several statistical computing packages. In Section 3.5 of the laboratory manuals we explain in detail how to do these computations using MINITAB or SAS.

Problems 3.5

Use the following information for Problems 3.5.1–3.5.6.

A coal burning power plant is located a distance of 25 miles from a national park. Emissions from the power plant contain the gas sulfur dioxide (SO_2) (which is linked to acid rain), and consequently a certain fraction of the emitted SO_2 is transported through the atmosphere to the national park. A certain amount of background SO_2 that is not emitted by the power plant is always present at the national park. To assess the power plant's SO_2 contribution to the national park, recordings were made of X, the SO_2 output by the plant in tons/hour, as well as Y, the SO_2 concentrations at the national park in micrograms/cubic meter ($\mu g/m^3$), at various randomly selected

times during a particular year. The data are given in Table 3.5.4 and are also stored in the file **so2.dat** on the data disk.

From these data we compute the following quantities.

$$\sum_{i=1}^{14} x_i = 55.760 \qquad \sum_{i=1}^{14} y_i = 123.35$$

$$\sum_{i=1}^{14} x_i^2 = 253.907 \qquad \sum_{i=1}^{14} y_i^2 = 1220.27$$

$$\sum_{i=1}^{14} y_i x_i = 549.355$$

The variations in the Y values for a given X value occur because of changes in the wind direction and wind speed, variations in the rate of dispersion of the gas, etc. We want to find out if higher levels of SO_2 at the national park are associated with higher levels of SO_2 emitted by the power plant. In similar investigations, assumptions (B) have been used, but for these data the experimenter wants to assess the validity of these assumptions.

3.5.1 Plot the values of SO_2 at the national park (y_i) against the values of SO_2 emissions at the power plant (x_i). After examining this plot, do you think a straight line regression function $\mu_Y(x) = \beta_0 + \beta_1 x$ seems reasonable?

3.5.2 Obtain the least squares estimates of β_0 and β_1. Also obtain an estimate of σ.

3.5.3 Table 3.5.5 contains y_i, x_i, the fitted values $\hat{\mu}_Y(x_i)$, the residuals $\hat{e}_i$, the hat values $h_{i,i}$, and the standardized residuals r_i. These quantities are also stored in the file **table355.dat** on the data disk. Plot the standardized residuals r_i against the fitted

TABLE 3.5.4
Power Plant SO_2 Data

Time	Y ($\mu g/m^3$)	X (tons/hour)
1	5.21	1.92
2	7.36	3.92
3	16.26	6.80
4	10.10	6.32
5	5.80	2.00
6	8.06	4.32
7	4.76	2.40
8	6.93	2.96
9	9.36	3.52
10	10.90	4.24
11	12.48	5.12
12	11.70	5.84
13	7.44	3.60
14	6.99	2.80

TABLE 3.5.5
Fitted Values, Residuals, Hat Values, and Standardized Residuals for Power Plant SO_2 Data

y_i	x_i	Fitted Values	Residuals	Hat Values	Standardized Residuals
5.21	1.92	5.0465	0.16352	0.205148	0.12115
7.36	3.92	8.6960	−1.33601	0.071553	−0.91583
16.26	6.80	13.9513	2.30865	0.320817	1.85031
10.10	6.32	13.0755	−2.97546	0.243072	−2.25895
5.80	2.00	5.1925	0.60754	0.194978	0.44725
8.06	4.32	9.4259	−1.36592	0.075000	−0.93807
4.76	2.40	5.9224	−1.16237	0.150159	−0.83283
6.93	2.96	6.9442	−0.01424	0.104305	−0.00994
9.36	3.52	7.9661	1.39389	0.078161	0.95892
10.90	4.24	9.2799	1.62006	0.073506	1.11171
12.48	5.12	10.8857	1.59426	0.112062	1.11750
11.70	5.84	12.1996	−0.49957	0.179808	−0.36435
7.44	3.60	8.1121	−0.67209	0.076035	−0.46183
6.99	2.80	6.6523	0.33773	0.115395	0.23718

values $\hat{\mu}_Y(x_i)$. Does this plot suggest any violations of assumptions (B)? If so, what assumption seems to be violated? Round the numbers to two decimals for plotting.

3.5.4 Plot the standardized residuals r_i against x_i. What does this plot suggest regarding assumptions (B)?

3.5.5 The ordered standardized residuals $r_{(i)}$ and the corresponding Gaussian scores $z_i^{(n)}$, where $n = 14$, appear in Table 3.5.6. Obtain a rankit-plot of these standardized residuals. What do you conclude from this plot? For plotting purposes you may suitably round the values of Gaussian scores.

3.5.6 Examine the plausibility of the assumption that the data (y_i, x_i) are a simple random sample from a bivariate Gaussian population by obtaining and examining the rankit-plots of y_i, x_i, $x_i + y_i$, and $x_i - y_i$. The Gaussian scores for y_i, x_i, $x_i + y_i$, $x_i - y_i$ appear in Table 3.5.7. Based on these plots, what is your conclusion as to whether (y_i, x_i) is a simple random sample from a bivariate Gaussian population? For plotting purposes you may suitably round the values of Gaussian scores.

TABLE 3.5.6

Ordered Standardized Residuals	Gaussian Scores
−2.25895	−1.70991
−0.93807	−1.20448
−0.91583	−0.89743
−0.83283	−0.65862
−0.46183	−0.45321
−0.36435	−0.26585
−0.00994	−0.08767
0.12115	0.08767
0.23718	0.26585
0.44725	0.45321
0.95892	0.65862
1.11171	0.89743
1.11750	1.20448
1.85031	1.70991

TABLE 3.5.7

Item	y_i	Gaussian Scores for y_i	x_i	Gaussian Scores for x_i	$x_i + y_i$	Gaussian Scores for $(x_i + y_i)$	$x_i - y_i$	Gaussian Scores for $(x_i - y_i)$
1	5.21	−1.20448	1.92	−1.70991	7.13	−1.70991	−3.29	1.20448
2	7.36	−0.26585	3.92	0.08767	11.28	−0.08767	−3.44	0.89743
3	16.26	1.70991	6.80	1.70991	23.06	1.70991	−9.46	−1.70991
4	10.10	0.45321	6.32	1.20448	16.42	0.65862	−3.78	0.45321
5	5.80	−0.89743	2.00	−1.20448	7.80	−0.89743	−3.80	0.26585
6	8.06	0.08767	4.32	0.45321	12.38	0.08767	−3.74	0.65862
7	4.76	−1.70991	2.40	−0.89743	7.16	−1.20448	−2.36	1.70991
8	6.93	−0.65862	2.96	−0.45321	9.89	−0.45321	−3.97	−0.08767
9	9.36	0.26585	3.52	−0.26585	12.88	0.26585	−5.84	−0.45321
10	10.90	0.65862	4.24	0.26585	15.14	0.45321	−6.66	−0.89743
11	12.48	1.20448	5.12	0.65862	17.60	1.20448	−7.36	−1.20448
12	11.70	0.89743	5.84	0.89743	17.54	0.89743	−5.86	−0.65862
13	7.44	−0.08767	3.60	−0.08767	11.04	−0.26585	−3.84	0.08767
14	6.99	−0.45321	2.80	−0.65862	9.79	−0.65862	−4.19	−0.26585

3.6 Confidence Intervals

An interval estimate of a parameter, along with a point estimate, is generally very useful for making decisions because an interval estimate implicitly tells the investigator how well the parameter in question is being estimated, i.e., how close the point estimate is likely to be to the true value. In this connection you should recall the discussions in Section 1.6. In this section we discuss procedures for computing confidence intervals for β_0, β_1, $\mu_Y(x)$, $Y(x)$, σ, and linear functions of the form $a_0\beta_0 + a_1\beta_1$. The results are valid under both assumptions (A) and (B). Note that confidence intervals for μ_Y, σ_Y, μ_X, and σ_X are obtained using the formulas given in Table 1.6.2 and are valid under assumptions (B).

General Form of Confidence Intervals for β_0, β_1, $\mu_Y(x)$, $Y(x)$, and Linear Functions $a_0\beta_0 + a_1\beta_1$

The general form for a $1-\alpha$ two-sided confidence interval for β_0, β_1, $Y(x)$, $\mu_Y(x)$, and the linear function $a_0\beta_0 + a_1\beta_1$ (let θ denote any one of these quantities and $\hat{\theta}$ denote the point estimate of θ) is

$$\hat{\theta} - (\text{table-value}) \times SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + (\text{table-value}) \times SE(\hat{\theta}) \qquad (3.6.1)$$

where table-value is the quantity $t_{1-\alpha/2:df}$ obtained from a student's t-table (Table T-2 in Appendix T) with df = degrees of freedom = $n-2$ in the case of straight line regression. The quantity $SE(\hat{\theta})$, called the standard error of $\hat{\theta}$, is an estimate of the *precision* of $\hat{\theta}$ and is calculated from sample data.

The point estimates for β_1, β_0, $\mu_Y(x)$, $Y(x)$, and $a_0\beta_0 + a_1\beta_1$ are given in Box 3.4.2. Their respective standard errors are given in (3.6.2)–(3.6.6).

$$SE(\hat{\beta}_1) = \hat{\sigma}/\sqrt{SSX} \qquad (3.6.2)$$

$$SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SSX}} \qquad (3.6.3)$$

$$SE(\hat{\mu}_Y(x)) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}} \qquad (3.6.4)$$

$$SE(\hat{Y}(x)) = \hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{SSX}} \qquad (3.6.5)$$

$$SE(a_0\hat{\beta}_0 + a_1\hat{\beta}_1) = \hat{\sigma}\sqrt{\frac{a_0^2}{n} + \frac{(a_1 - a_0\bar{x})^2}{SSX}} \qquad (3.6.6)$$

The following relationship between $SE(\hat{\mu}_Y(x))$ and $SE(\hat{Y}(x))$ is sometimes useful.

$$SE(\hat{Y}(x)) = \sqrt{\hat{\sigma}^2 + [SE(\hat{\mu}_Y(x))]^2} \qquad (3.6.7)$$

Note that $SE(\hat{Y}(x))$, is really an abbreviation for $SE(\hat{Y}(x) - Y(x))$.

Remarks

1 $SE(\hat{\beta}_0)$, $SE(\hat{\beta}_1)$, and $SE(\hat{\mu}_Y(x))$ can be obtained from $SE(a_0\hat{\beta}_0 + a_1\hat{\beta}_1)$ by substituting the appropriate values for a_0 and a_1.

2 By studying (3.6.4) we see that the further that x is from $\bar{x}$, the larger $SE(\hat{\mu}_Y(x))$ becomes. This can be intuitively explained as follows. Recall that the estimated regression line always passes through the point $(\bar{x}, \bar{y})$. Think of the estimated line as a see-saw with its pivot at $(\bar{x}, \bar{y})$. A slight rotation of the line about the pivot causes large deviations far away from the pivot but only small deviations close to the pivot (see Figure 3.5.5 and the discussion pertaining to it). As a result, points on the line (which are the estimated means of subpopulations) close to $(\bar{x}, \bar{y})$ are estimated with greater precision compared to points far away from $(\bar{x}, \bar{y})$. A similar explanation applies to $SE(\hat{Y}(x))$.

3 In any particular application, it is important for you to determine whether a confidence interval for $\mu_Y(x)$ or a confidence interval for $Y(x)$ is required. Some authors call the confidence interval for $Y(x)$ a *prediction interval* because the term confidence interval is traditionally reserved for parameters, but $Y(x)$ is a random variable (it is the Y value of a randomly chosen item from the subpopulation corresponding to $X = x$).

4 Note that even though $\hat{\mu}_Y(x) = \hat{Y}(x)$, i.e.,

$$\hat{\mu}_Y(x) = \hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

the standard error of $\hat{\mu}_Y(x)$ is not the same as the standard error of $\hat{Y}(x)$. This is so because there is greater uncertainty in predicting a single randomly chosen value from the subpopulation corresponding to $X = x$ than in estimating the mean of the entire subpopulation.

5 For each quantity being estimated, the investigator should be given a point estimate, its standard error, and the degrees of freedom associated with this standard error for obtaining the table t-value. The investigator can substitute these quantities into (3.6.1) and compute the desired confidence intervals by selecting the confidence coefficient $1 - \alpha$ that he wants to use.

Confidence Intervals for σ

A two-sided $1 - \alpha$ confidence interval for σ is given by

$$C\left[\sqrt{\frac{(df)\,\hat{\sigma}^2}{\chi^2_{1-\alpha/2:df}}} \le \sigma \le \sqrt{\frac{(df)\,\hat{\sigma}^2}{\chi^2_{\alpha/2:df}}}\right] = 1 - \alpha \qquad \textbf{(3.6.8)}$$

where $df = n - 2$ and $\chi^2_{\alpha/2:df}$ and $\chi^2_{1-\alpha/2:df}$ are obtained from Table T-3 in Appendix T. The formula for a confidence interval for σ is not of the general form

given in (3.6.1). In fact the general form for a two-sided $(1-\alpha)$ confidence interval for a parameter that is a standard deviation (such as σ_Y, σ_X, σ, etc.) is

$$\sqrt{\frac{(df)\ (\text{estimated standard deviation})^2}{\chi^2_{1-\alpha/2:df}}} \leq \sigma \leq \sqrt{\frac{(df)\ (\text{estimated standard deviation})^2}{\chi^2_{\alpha/2:df}}} \tag{3.6.9}$$

where df represents the number of degrees of freedom associated with the estimated standard deviation. Note that the number of degrees of freedom associated with $\hat{\sigma}_X$ or $\hat{\sigma}_Y$ is $n-1$, whereas that associated with $\hat{\sigma}$ is $n-2$. It is also worth observing that while the confidence intervals for the quantities, β_0, β_1, $\mu_Y(x)$, $Y(x)$, and $a_0\beta_0 + a_1\beta_1$ are symmetric about the corresponding point estimates, this is not so in the case of σ. However, the confidence interval for σ in (3.6.9) is equal-tailed, contains the point estimate $\hat{\sigma}$, and gives us some indication of how close $\hat{\sigma}$ may be to the true value σ.

One-Sided Confidence Bounds

In this section, the discussion so far has been about two-sided confidence intervals for parameters of interest. However, in some applications one-sided confidence bounds are more useful because an investigator may be interested in only the lower bound or the upper bound for a parameter in a decision-making situation. As discussed in Section 1.6, if a confidence interval is equal-tailed, we can obtain one-sided confidence bounds with confidence coefficient $1-\alpha$ by first constructing a two-sided confidence interval with confidence coefficient $1-2\alpha$ and reading off either the lower or the upper endpoint as appropriate. This is valid for all of the quantities, β_0, β_1, $\mu_Y(x)$, $Y(x)$, and $a_0\beta_0 + a_1\beta_1$, as well as σ, σ_Y, and σ_X, because these confidence intervals are *equal-tailed.*

The following two tasks illustrate the computations required for obtaining confidence intervals in the context of real applications.

Task 3.6.1

Consider the problem discussed in Task 3.4.2 where an investigator wants to evaluate the performance of a new laboratory method, which is less expensive than the current method, for analyzing the concentration of arsenic (As) in water samples. The data appear in Table 3.4.3. For convenience they are reproduced in Table 3.6.1 and are also stored in the file **arsenic.dat** on the data disk.

TABLE 3.6.1
Arsenic Data

Sample Item Number	Measured Concentration Y (μg/ml)	True Concentration X (μg/ml)
1	0.17	0
2	0.25	0
3	0.01	0
4	0.12	0
5	1.25	1
6	0.86	1
7	1.25	1
8	1.10	1
9	2.01	2
10	2.03	2
11	2.14	2
12	1.74	2
13	3.18	3
14	2.99	3
15	3.23	3
16	3.37	3
17	3.91	4
18	3.90	4
19	3.61	4
20	4.27	4
21	4.88	5
22	5.33	5
23	4.96	5
24	4.98	5
25	6.09	6
26	6.17	6
27	6.07	6
28	5.97	6
29	6.67	7
30	7.02	7
31	7.14	7
32	7.30	7

Suppose, based on experience, the investigator feels that population assumptions (A) for straight line regression should hold, at least approximately, where the data are obtained by sampling with preselected X values. In particular, the population regression function is

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

and the subpopulation standard deviations for Y are all the same, each equal to σ.

Some basic sums, sums of squares, and sums of crossproducts—the ingredients in the formulas for obtaining point and confidence interval estimates of population parameters—are as follows:

$$\sum_{i=1}^{n} x_i = 112 \qquad \bar{x} = 3.5 \qquad \sum_{i=1}^{n} y_i = 113.97 \qquad \bar{y} = 3.56156$$

$$SSX = 168 \qquad SSY = 164.95 \qquad SXY = 165.935$$
$$SSE = 1.0544 \qquad MSE = 0.03515$$

From these we compute the point estimates

$$\hat{\beta}_1 = \frac{SXY}{SSX} = \frac{165.935}{168} = 0.98771$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 3.56156 - 0.98771(3.5) = 0.10458$$
$$\hat{\mu}_Y(x) = \hat{\beta}_0 + \hat{\beta}_1 x = 0.10458 + 0.98771x = \hat{Y}(x)$$
$$\hat{\sigma} = \sqrt{MSE} = \sqrt{0.03515} = 0.1875$$

We now calculate appropriate confidence intervals for β_0, β_1, and σ that will help the investigator obtain practical answers to the following three questions (using $1 - \alpha = 0.95$). In Section 3.6 of the laboratory manual we show how a computer can be used to calculate all quantities needed in this section.

1 On the average, does the chemical analysis correctly report the absence of As when this is indeed the case? (That is, is $\mu_Y(x) = 0$ when $x = 0$? In other words is $\beta_0 = 0$?)

To answer this question we first compute a 95% two-sided confidence interval for β_0. Using the formula in (3.6.3) we compute

$$SE(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SSX}} = (0.1875)\sqrt{\frac{1}{32} + \frac{(3.5)^2}{168}} = 0.0605$$

From Table T-2 in Appendix T we obtain $t_{1-\alpha/2:30} = t_{0.975:30} = 2.042$, corresponding to $1 - \alpha = 0.95$ and degrees of freedom $= n - 2 = 30$. Hence a 95% two-sided confidence interval for β_0 is given by

$$C[-0.019 \le \beta_0 \le 0.228] = 0.95$$

Based on this confidence interval, the investigator can decide whether β_0 can be considered close enough to zero to be regarded as negligible for all practical purposes for this problem.

2 Now suppose the investigator wants to examine how close β_1 is to 1 in the model $\mu_Y(x) = \beta_0 + \beta_1 x$. Compute a 95% two-sided confidence interval for β_1 to help determine this.

Using the formula in (3.6.2) we calculate

$$SE(\hat{\beta}_1) = \hat{\sigma}/\sqrt{SSX} = 0.1875/\sqrt{168} = 0.01447$$

From Table T-2 in Appendix T we obtain $t_{1-\alpha/2:30} = t_{0.975:30} = 2.042$. *Hence a 95% confidence interval is given by the confidence statement*

$$C[0.958 \leq \beta_1 \leq 1.017] = 0.95$$

On the basis of this information, the investigator can decide whether β_1 *can be regarded as close enough to 1.0 for all practical purposes in this problem.*

3 Suppose that, based on data and previous experience, the investigator is fairly confident that the new method of chemical analysis for As is unbiased; i.e., the regression function of Y on X is $\mu_Y(x) = x$. However, for the method to be adopted by water quality monitoring agencies, they want to know whether a proportion 0.99 or more of the measured concentrations will be accurate to within 1.0 μg/ml. Use the data at hand to help the investigator make a decision.

If $\mu_Y(x) = x$, *then 99% of the measured concentrations corresponding to a true concentration x will be between* $x - z_{0.995}\sigma$ *and* $x + z_{0.995}\sigma$, *i.e., between* $x - 2.575\sigma$ *and* $x + 2.575\sigma$. *Hence we want to know whether* 2.575σ *is less than 1.0. The estimate of* σ *from the preceding calculations is equal to 0.1875, and so the estimated value of* 2.575σ *is equal to 0.483. This is well within the acceptable upper limit of 1.0. However, the investigator realizes that the estimate 0.1875 of* σ *is subject to sampling errors, so he wants to know if the estimate is sufficiently accurate for decision-making purposes. To investigate this the investigator wants a 95% one-sided upper confidence bound for* σ. *We use the formula for a 90% two-sided confidence bound for* σ *given in (3.6.8), and from this we can obtain a one-sided 95% upper confidence bound. The necessary* χ^2 *table-values, obtained from Table T-3 in Appendix T, are* $\chi^2_{0.05:30} = 18.493$ *and* $\chi^2_{0.95:30} = 43.773$. *Thus we get the confidence statement*

$$C\left[\sqrt{\frac{30\,(0.1875)^2}{43.773}} \leq \sigma \leq \sqrt{\frac{30\,(0.1875)^2}{18.493}}\right] = 0.90$$

i.e.,

$$C[0.1552 \leq \sigma \leq 0.2388] = 0.90$$

and hence

$$C[0.3996 \leq 2.575\sigma < 0.6149] = 0.90$$

Thus the investigator has 95% confidence that $2.575\sigma \leq 0.6149$. *Based on this confidence statement the investigator will perhaps conclude that the variability in the new method of analysis for measuring As is within acceptable limits and may be adopted for routine use in environmental monitoring.*

Task 3.6.2

A sample of size 24 was randomly selected using simple random sampling from a population of individuals of a particular ethnic group with ages ranging from 21 to 70 years, and their ages and blood pressures were recorded. The response variable is the average systolic blood pressure Y at 8 A.M. over a two-week period, and the predictor variable is age X of the individual in years. The sample data are listed in Table 3.6.2. They are also stored in the file **agebp.dat** on the data disk.

It is supposed that assumptions (B) in Box 3.3.2 are valid. (You are encouraged to examine the validity of these assumptions using the methods of Section 3.5.) In particular, the regression function of Y on X is of the form $\mu_Y(x) = \beta_0 + \beta_1 x$, where β_0 and β_1 are unknown parameters. Also the subpopulation standard deviations are all equal, and their common value is denoted by σ, which is also an unknown parameter.

TABLE 3.6.2
Age and Blood Pressure Data

Sample Item Number	Blood Pressure (systolic) Y	Age (in years) X
1	116	34
2	112	26
3	151	51
4	161	58
5	122	34
6	129	40
7	119	31
8	158	57
9	144	46
10	150	53
11	111	29
12	148	50
13	135	40
14	126	34
15	172	67
16	100	23
17	139	47
18	135	42
19	163	61
20	128	38
21	159	57
22	177	66
23	135	42
24	149	53

Some of the basic quantities needed to answer the following questions using sample data are

$$\sum_{i=1}^{24} x_i = 1{,}079 \qquad \sum_{i=1}^{24} y_i = 3{,}339$$

$$\sum_{i=1}^{24} (x_i - \bar{x})^2 = SSX = 3{,}608.96 \qquad \sum_{i=1}^{24} (y_i - \bar{y})^2 = SSY = 9{,}514.63$$

$$\sum_{i=1}^{24} (x_i - \bar{x})(y_i - \bar{y}) = SXY = 5{,}805.13 \qquad SSE = 176.896 \qquad MSE = 8.04075$$

$$\hat{\beta}_1 = 1.6085 \qquad \hat{\beta}_0 = 66.8081 \qquad \hat{\sigma} = 2.8356 \qquad \hat{\sigma}_Y = 20.3391$$

1 What is the average age of the individuals in the population?

The average age of the individuals in the population is μ_X; it cannot be exactly determined from the sample data but must be estimated. The estimated average age is $\hat{\mu}_X = \bar{x} = 45$ years (rounded to the nearest integer). Note that if data had been obtained by sampling with preselected X values, then, in general, no valid estimate of μ_X would be available (see (3.3.1)).

2 What is the average blood pressure of the individuals in the population?

The average blood pressure of the individuals in the population is μ_Y; it cannot be exactly determined from the sample data. The estimated value of μ_Y is $\hat{\mu}_Y = \bar{y} = 139$ units (to the nearest integer). Note that if data had been obtained by sampling with preselected X values, then, in general, no valid estimate of μ_Y would be available (see (3.3.1.))

3 Without using age as a predictor factor, predict the blood pressure of a randomly chosen individual from this population. Compute a number d such that you can be 95% confident that the predicted value will be within d units of the actual value.

Without using age X as a predictor factor, the best predicted value of the blood pressure Y of a randomly chosen individual from the population is $\hat{Y} = \bar{y} = 139.125$ and the standard error of $\hat{Y}$ is $\hat{\sigma}_Y\sqrt{\frac{1}{n} + 1} = 20.3391\sqrt{1.04167} = 20.7585$. (As we explained earlier, this is actually the standard error of $\hat{Y} - Y$.) From Table T-2 in Appendix T we get $t_{1-\alpha/2:n-1} = t_{0.975:23} = 2.069$. So a 95% confidence statement for Y is

$$C[96.176 \leq Y \leq 182.074] = 0.95$$

i.e.,

$$C[|Y - 139.125| \leq 42.95] = 0.95$$

(You should verify this.) Hence we can be 95% confident that the predicted value of 139.125 will be within $d = 42.95$ units of the actual Y value.

4 What is the average blood pressure of all individuals in the population who are 65 years old? How well can we estimate it based on the sample data (use $1 - \alpha = 0.95$)?

The average blood pressure of all individuals in the population who are 65 years old cannot be determined exactly from sample values. However, based on sample data, we estimate the average blood pressure of all individuals in the population who are 65 years old to be $\hat{\mu}_Y(65) = \hat{\beta}_0 + \hat{\beta}_1(65) = 66.8081 + 1.6085(65) = 171.361$. *To find out how good this sample estimate is we calculate its standard error using the formula for* $SE(\hat{\mu}_Y(x))$ *in (3.6.4). We first calculate the value of* $\hat{\sigma}$ *as 2.8356. So* $SE(\hat{\mu}_Y(65)) = 1.109$. *To obtain a 95% confidence interval we use Table T-2 in Appendix T to find that* $t_{1-\alpha/2:n-2} = t_{0.975:22} = 2.074$. *So using (3.6.1) we obtain a 95% two-sided confidence interval for* $\mu_Y(65)$ *as*

$$[171.361 - 2.074 \times 1.109, 171.361 + 2.074 \times 1.109]$$

which is [169.06, 173.66]. Thus we are 95% confident that the average blood pressure of all the individuals in the population who are 65 years old is in the interval [169.06, 173.66].

5 On the average, does blood pressure increase with age? If so, by how much?

The slope of the line relating the average blood pressure $\mu_Y(x)$ *of the subpopulation with* $X = x$ *to age X is the parameter* β_1. *The estimated value of* β_1 *is 1.6085, which means that the average blood pressure is estimated to increase with age at the rate of 1.6085 units per year. A confidence interval for* β_1 *will tell us how good our estimate of* β_1 *is. We first calculate* $SE(\hat{\beta}_1)$ *using (3.6.2) and obtain the value 0.0472. Hence a 95% two-sided confidence interval for* β_1 *is* $[1.6085 - 2.074 \times 0.0472,\ 1.6085 + 2.074 \times 0.0472]$, *i.e., [1.5106, 1.7064]. So we can be 95% confident that the average rate of increase of blood pressure with age is between 1.51 units and 1.71 units per year (rounded to two decimal places).*

6 Based on the sample data, can we estimate the average blood pressure of the subpopulation of all newborn babies?

Newborn babies are zero years old, and so the average blood pressure of all newborn babies would be $\mu_Y(0)$, *which is equal to* β_0. *The calculations for the estimate of* β_0 *using the sample data would yield the value 66.808. However, there is no guarantee that the population regression function* $\mu_Y(x) = \beta_0 + \beta_1 x$ *is valid for all values of X. Noting that the smallest age in the sample is 23, we immediately know not to extrapolate the estimated regression function to obtain an estimate of the average blood pressure of newborn babies. In short, while we can carry out the calculations for* $\hat{\mu}_Y(0)$, *the result will not necessarily be meaningful because this involves an excessive amount of extrapolation.*

7 Calculate a 95% two-sided prediction interval for the blood pressure of an individual who is 60 years old.

If an individual is randomly chosen from the subpopulation of all individuals who are 60 years old, this individual's blood pressure value would be predicted to be in the interval

$$[\hat{Y}(60) - 2.074\,SE(\hat{Y}(60)),\ \hat{Y}(60) + 2.074\,SE(\hat{Y}(60))]$$

with 95% confidence. The value of $\hat{Y}(60)$ *is* $\hat{\mu}_Y(60) = \hat{\beta}_0 + \hat{\beta}_1(60) = 163.32$, *and the value of* $SE(\hat{Y}(60))$, *using (3.6.5), is 2.98. So this prediction interval is calculated to be [157.139, 169.501], and the confidence statement is*

$$C[157.14 \leq Y(60) \leq 169.50] = 0.95$$

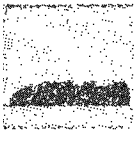
Simultaneous Confidence Intervals for $\mu_Y(x)$ for Several Values of x

There are situations when we want confidence intervals for $\mu_Y(x)$ for several (or all) values of x such that we have confidence $1 - \alpha$ that all the intervals are simultaneously correct (refer to Section 1.6). For example, in evaluating the performance of a rocket during a certain 30-second interval (say, $15 \leq X \leq 45$), the linear model $\mu_Y(x) = \beta_0 + \beta_1 x$ is assumed to be valid, where Y is the speed of the rocket in feet per second and X is time after the launch in seconds. Data are collected from test firings of small models and estimates are obtained for β_1, β_0, and σ by the formulas in (3.4.8), (3.4.9), and (3.4.22), respectively. A decision must be made about a piece of equipment that is to be installed on the rocket, and the correct decision depends on knowing the average speed at 15 seconds, 20 seconds, 25 seconds, 30 seconds, 35 seconds, 40 seconds, and 45 seconds after the launch. Thus confidence intervals are needed for $\mu_Y(x)$ for $x = 15, 20, 25, 30, 35, 40$, and 45 seconds such that we have confidence at least $1 - \alpha$ that *all* the confidence intervals are simultaneously correct. Thus we can have confidence $1 - \alpha$ that the decision is correct.

To obtain confidence intervals for $\mu_Y(x)$ for m different values of X so that we can have at least $1 - \alpha$ confidence that they are *all* simultaneously correct, we use the formula in (3.6.1) with the *table-value* given by

$$\text{table-value} = \text{smaller of the two values } t_{1-\alpha/2m:n-2} \text{ and } \sqrt{2F_{1-\alpha:2,n-2}}$$

Table T-4 in Appendix T contains table t-values for m **simultaneous** two-sided confidence intervals with *simultaneous confidence coefficient* greater than or equal to $1 - \alpha$ for $m = 2, 3, \ldots, 6$. Table T-5 contains F values.



Problems 3.6

3.6.1 For the population data in Table D-1 in Appendix D, which is also in the file **car.dat** on the data disk, what parameters must be estimated to obtain the difference between the average first-year maintenance costs of cars driven 15,000 miles and cars driven 10,000 miles?

3.6.2 Parts (a)–(e) refer to the data in Table 3.5.2, which were obtained by simple random sampling from the population data in Table D-1 in Appendix D. These data are also in the file **car20.dat** on the data disk. Assumptions (B) are presumed to hold so the regression function of maintenance cost Y on miles driven X is of the form

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

For these data we have computed the following quantities: $\bar{y} = 521.7$; $\bar{x} = 11{,}010$; $SSY = 210{,}568$; $SSX = 175{,}337{,}952$; $SXY = 5{,}489{,}360$.

a Obtain a point estimate for the quantity of interest in Problem 3.6.1.

b Obtain a 90% confidence interval for the quantity in Problem 3.6.1.

c A company has purchased three cars of the same make as those in Example 3.5.2. One of the cars is to be driven 5,000 miles during its first year, and the other two cars will be driven 12,500 miles each. Predict the total first-year maintenance cost for all three cars together.

d Compute a 90% two-sided confidence interval for the average maintenance cost of all cars driven 12,500 miles during their first year.

e The predictor factor X (miles driven) will be considered to be a useful predictor of Y if σ is less than \$50. Compute an 80% confidence statement to help the investigator arrive at a decision.

3.6.3 Consider the arsenic data of Task 3.4.2. Recall that the investigator is interested in determining whether or not the new method of analysis for determining the concentration of As is unbiased. Compute confidence intervals for β_0 and β_1 in Task 3.4.2 such that both confidence intervals are simultaneously correct with 90% confidence. Based on these, help the investigator arrive at an appropriate decision. Give reasons.

3.6.4 Questions (a)–(c) refer to the crystal data of Task 3.4.1.

a Calculate an appropriate 90% confidence statement to help the investigator decide whether the average rate of growth of the crystals is at least 0.4 gram per hour.

b Compute a 95% confidence interval for the weight of a crystal that is allowed to grow for 15 hours.

c Obtain simultaneous two-sided confidence intervals for the weights $Y(10)$ and $Y(25)$ of two individual crystals, one of which is to be grown for 10 hours and the other for 25 hours, such that the investigator can be at least 90% confident that *both* confidence intervals are simultaneously correct. (*Hint*: Use the Bonferroni method described in Box 1.6.3.)

3.7 Tests

Investigators often use statistical tests in an attempt to arrive at answers to practical questions. To do so the investigator is required to formulate a pair of hypotheses, called the null hypothesis (NH) and the alternative hypothesis (AH), and to perform appropriate statistical calculations to obtain a measure, called the **significance probability** or the ***P*-value** of the evidence contained in the sample data *against* NH. (The results of these calculations do not tell the investigator how much evidence is provided by the sample data in favor of the null hypothesis.) If the evidence against NH is strong, which would be the case if the *P*-value is small, then it is customary to *reject* the null hypothesis. In the contrary case—i.e., when the *P*-value is not small—the null hypothesis is not rejected. Neither is it accepted without further analysis.

At this point we refer you back to Section 1.6 (in particular, to the conversation). The discussion in Section 1.6 should convince you that statistical tests are often inappropriate for making practical decisions. Confidence intervals can be much more useful for this purpose.

Formulas and procedures for some statistical tests of size α that pertain to the straight line regression model are summarized in Boxes 3.7.1–3.7.5. The procedures are valid under assumptions (A) or (B).

BOX 3.7.1 Tests for β_0

Let q be a specified number. Compute the statistic

$$t_C = \frac{\hat{\beta}_0 - q}{SE(\hat{\beta}_0)}$$

a For testing NH: $\beta_0 = q$ versus AH: $\beta_0 \neq q$, the P-value is the value of α such that $|t_C| = t_{1-\alpha/2:n-2}$.

b For testing NH: $\beta_0 \leq q$ versus AH: $\beta_0 > q$, the P-value is the value of α such that $t_C = t_{1-\alpha:n-2}$.

c For testing NH: $\beta_0 \geq q$ versus AH: $\beta_0 < q$, the P-value is the value of α such that $-t_C = t_{1-\alpha:n-2}$.

BOX 3.7.2 Tests for β_1

Let q be a specified number. Compute the statistic

$$t_C = \frac{\hat{\beta}_1 - q}{SE(\hat{\beta}_1)}$$

a For testing NH: $\beta_1 = q$ versus AH: $\beta_1 \neq q$, the P-value is the value of α such that $|t_C| = t_{1-\alpha/2:n-2}$.

b For testing NH: $\beta_1 \leq q$ versus AH: $\beta_1 > q$, the P-value is the value of α such that $t_C = t_{1-\alpha:n-2}$.

c For testing NH: $\beta_1 \geq q$ versus AH: $\beta_1 < q$, the P-value is the value of α such that $-t_C = t_{1-\alpha:n-2}$.

BOX 3.7.3 Tests for $\mu_Y(x)$

Let q be a specified number. Compute the statistic

$$t_C = \frac{\hat{\mu}_Y(x) - q}{SE(\hat{\mu}_Y(x))}$$

a For testing NH: $\mu_Y(x) = q$ versus AH: $\mu_Y(x) \neq q$, the P-value is the value of α such that $|t_C| = t_{1-\alpha/2:n-2}$.

b For testing NH: $\mu_Y(x) \leq q$ versus AH: $\mu_Y(x) > q$, the P-value is the value of α such that $t_C = t_{1-\alpha:n-2}$.

c For testing NH: $\mu_Y(x) \geq q$ versus AH: $\mu_Y(x) < q$, the P-value is the value of α such that $-t_C = t_{1-\alpha:n-2}$.

BOX 3.7.4 Tests for $a_0\beta_0 + a_1\beta_1$

Let q be a specified number. Compute the statistic

$$t_C = \frac{(a_0\hat{\beta}_0 + a_1\hat{\beta}_1) - q}{SE(a_0\hat{\beta}_0 + a_1\hat{\beta}_1)}$$

a For testing NH: $a_0\beta_0 + a_1\beta_1 = q$ versus AH: $a_0\beta_0 + a_1\beta_1 \neq q$, the P-value is the value of α such that $|t_C| = t_{1-\alpha/2:n-2}$.

b For testing NH: $a_0\beta_0 + a_1\beta_1 \leq q$ versus AH: $a_0\beta_0 + a_1\beta_1 > q$, the P-value is the value of α such that $t_C = t_{1-\alpha:n-2}$.

c For testing NH: $a_0\beta_0 + a_1\beta_1 \geq q$ versus AH: $a_0\beta_0 + a_1\beta_1 < q$, the P-value is the value of α such that $-t_C = t_{1-\alpha:n-2}$.

The P-values for the tests in Boxes 3.7.1–3.7.4 can be found (at least approximately) by consulting Table T-2 in Appendix T.

BOX 3.7.5 Tests for σ

Let q be a specified positive number. Compute the statistic

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

a 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

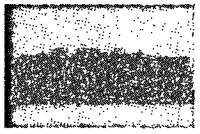
$$\chi^2_C = \chi^2_{\alpha/2:n-2} \quad \text{or} \quad \chi^2_C = \chi^2_{1-\alpha/2:n-2}$$

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

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

c For testing NH: $\sigma \geq q$ versus AH: $\sigma < q$, the P-value is the value of α such that $\chi^2_C = \chi^2_{\alpha:n-2}$.

P-values for the tests in Box 3.7.5 can be found (at least approximately) by consulting Table T-3 in Appendix T. We illustrate these procedures in the following task.



Task 3.7.1

Consider the problem described in Task 3.4.2 in which an investigator wants to evaluate the performance of a new laboratory method for analyzing the concentration of arsenic (As) in water samples. The data appear in Table 3.4.3 and also in the file **arsenic.dat** on the data disk. Recall that the regression function is $\mu_Y(x) = \beta_0 + \beta_1 x$, and the investigator is interested in knowing

- Whether or not β_0 is zero
- Whether or not β_1 is equal to one

Some investigators attempt to answer these questions using statistical tests as follows:

1 Can we conclude that β_0 is zero for all practical purposes?

This question is sometimes translated as follows: How much evidence do the data provide against the hypothesis that $\beta_0 = 0$?

To answer this question, some statisticians and investigators carry out a test of

$$NH: \beta_0 = 0 \qquad against \qquad AH: \beta_0 \neq 0$$

In Task 3.6.1 we obtained $\hat{\beta}_0 = 0.10458$ and $SE(\hat{\beta}_0) = 0.0605$. From Box 3.7.1 we get

$$t_C = \frac{0.10458 - 0}{0.0605} = 1.729$$

Because $n = 32$ for this problem, the degrees of freedom are $n - 2 = 30$ and, using Table T-2 in Appendix T, the P-value for this test is between 0.05 and 0.10. If the investigator uses an α value equal to 0.05, then NH will not be rejected. If an α value equal to 0.10 or greater is used, then NH will be rejected and the investigator will conclude that β_0 is not equal to zero. This hypothesis test does not help the investigator make a practical decision. If NH is not rejected, then the investigator is still not sure whether β_0 is close enough to zero to be considered equal to zero from a practical point of view or whether there are not enough data to arrive at a practical decision. On the other hand, the confidence interval for β_0 given in part (1) of Task 3.6.1 should help the investigator in this regard.

2 Can we conclude that $\beta_1 = 1$ for all practical purposes?

This question is often translated as follows: How much evidence do the data provide against the hypothesis that $\beta_1 = 1$?

To answer this question, some statisticians and investigators carry out a test of

$$NH: \quad \beta_1 = 1 \qquad against \qquad AH: \quad \beta_1 \neq 1$$

In Task 3.6.1 we obtained $\hat{\beta}_1 = 0.98771$ *and* $SE(\hat{\beta}_1) = 0.01447$. *From Box 3.7.2 we get*

$$t_C = \frac{0.98771 - 1.0}{0.01447} = -0.85$$

From Table T-2 in Appendix T the P-value for this test is approximately 0.40, so the data do not provide sufficient evidence (at any reasonable value of α *such as 0.2, 0.1, 0.05, etc.) to conclude that* $\beta_1 \neq 1.0$.

Again, this hypothesis test does not help the investigator make a practical decision. If NH is not rejected, then the investigator is still not sure whether β_1 *is close enough to 1 to be considered equal to 1 from a practical point of view or whether there are not enough data to arrive at a practical decision. On the other hand, the confidence interval for* β_1 *given in part (2) of Task 3.6.1 should help the investigator in this regard.*

Task 3.7.2

Now consider the problem discussed in Task 3.4.1 where an investigator is interested in predicting Y, the weight of crystals used in electronic devices as a function of X, the number of hours the crystals grow. We perform an appropriate statistical test in an attempt to find an answer to the following question.

1 Suppose that crystals are allowed to grow for 24 hours instead of 12 hours. Do the data provide enough evidence to conclude, on the average, that the amount of money the larger crystals can be sold for is more than \$50 over what the smaller crystals would fetch?

The average weight of crystals that are allowed to grow for a total of 24 hours is $\mu_Y(24)$ *grams, whereas the crystals grown for 12 hours have an average weight of* $\mu_Y(12)$ *grams. Thus the additional money that the larger crystals would bring (on the average) is equal to* $16\mu_Y(24) - 10\mu_Y(12) = 6\beta_0 + 264\beta_1$ *dollars. This leads to the pair of hypotheses*

$$NH: \quad 6\beta_0 + 264\beta_1 \leq 50 \qquad against \qquad AH: \quad 6\beta_0 + 264\beta_1 > 50$$

So $q = 50$ *and let* $\theta = 6\beta_0 + 264\beta_1$.

Now recall that in Task 3.4.1 we computed the following quantities: $\bar{x} = 15$; $SSX = 910$; $\hat{\beta}_0 = 0.0014$; $\hat{\beta}_1 = 0.5034$; *and* $\hat{\sigma} = 1.062$. *Thus we have* $\hat{\theta} = 6\hat{\beta}_0 + 264\hat{\beta}_1 = 132.906$. *Also using (3.6.6) we calculate* $SE(\hat{\theta}) = 6.36$. *Using the test statistic in Box 3.7.4 for testing* θ *we get*

$$t_C = \frac{\hat{\theta} - 50}{SE(\hat{\theta})} = 13.04$$

Because $n = 14$ for this problem, the degrees of freedom are 12 and the P-value for this test, obtained from Table T-2 in Appendix T, is less than 0.0005, indicating that the data contain strong evidence against the null hypothesis in favor of the alternative hypothesis.

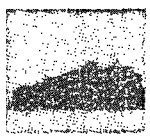
It is instructive to compute a one-sided 95% lower confidence bound for θ. As usual, this may be obtained as the lower endpoint of a 90% two-sided confidence interval, which is given by

$$C[121.58 \leq \theta \leq 144.240] = 0.90$$

Thus we get

$$C[121.58 \leq \theta] = 0.95$$

Therefore we are 95% confident that the larger crystals can be sold for at least \$121 more than the smaller crystals, so we would be led to conclude that they can be sold for at least \$50 more than the smaller crystals. Note that the confidence interval supports the result of the test, but it gives considerably more information than the test.



Problems 3.7

3.7.1 A particular brand of cough syrup comes in $\frac{1}{4}$-litre bottles and the manufacturer recommends that after a bottle is unsealed it be kept under cool conditions. The shelf-life of the cough syrup in question is dependent on the temperature at which it is stored. The quality control laboratory of the manufacturing company has obtained the data in Table 3.7.1 on the shelf-life of the cough syrup in question. The data are also in the file **shelflif.dat** on the data disk.

Some basic quantities that are needed for obtaining point estimates and confidence intervals are as follows:

$$\sum_{i=1}^{18} y_i = 11341 \qquad \sum_{i=1}^{18} x_i = 387$$

$$\sum_{i=1}^{18} (y_i - \bar{y})^2 = SSY = 96294.9 \qquad \sum_{i=1}^{18} (x_i - \bar{x})^2 = SSX = 484.5$$

$$\sum_{i=1}^{18} (y_i - \bar{y})(x_i - \bar{x}) = SXY = -6663.49$$

The regression function of shelf-life Y on storage temperature X is assumed to be a straight line $\mu_Y(x) = \beta_0 + \beta_1 x$ for values of x in the range 10°C to 35°C, and assumptions (A) are presumed to be satisfied.

a Define an appropriate target population for this investigation.

b Define an appropriate study population for this investigation.

TABLE 3.7.1
Shelf-Life Data

Bottle Number	Shelf-Life (Y, in days)	Storage Temperature (X, in °C)
1	727	13
2	760	14
3	730	15
4	716	16
5	683	17
6	665	18
7	641	19
8	663	20
9	653	21
10	615	22
11	585	23
12	614	24
13	592	25
14	564	26
15	537	27
16	537	28
17	552	29
18	507	30

c Are the data in this investigation obtained by simple random sampling or by sampling with preselected X values?

d Plot y_i versus x_i. Examine this plot and decide whether a straight line regression model seems reasonable.

e The director of the laboratory wants to determine whether the data provide evidence (at the 0.05 level) indicating that shelf-life does indeed depend on storage temperature, so he decides to use a statistical test. State an appropriate pair of hypotheses, suitably designating one as the null hypothesis and the other as the alternative hypothesis, and calculate the P-value for this test. What is your conclusion?

f Estimate, if possible, the average shelf-life for this cough syrup if it is to be stored at 0°C.

g Estimate the average shelf-life for this cough syrup if it is to be stored at 15°C. Also compute a 95% confidence interval for this quantity.

h Answer part (e) using an appropriate confidence interval instead of a hypothesis test.

i Do the data provide evidence (at the 0.05 level) indicating that the average shelf-life for bottles of cough syrup stored at 13°C is at least 650 days? Carry out an appropriate statistical test and state your conclusions.

j Construct an appropriate confidence interval to answer part (i).

3.8 Analysis of Variance

For the straight line regression model (and also for more general linear regression models), it is customary to summarize, in a table, certain key numerical quantities that are useful for making inferences. The process of calculating and examining these key numerical quantities is called an **analysis of variance.** The resulting table containing these quantities is called an analysis of variance table. The first key quantity is

$$SSY = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{3.8.1}$$

which was defined in (3.4.15). Recall that SSY can also be written as

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

The quantity $\sum_{i=1}^{n} y_i^2$ is called the *uncorrected total sum of squares* for Y, the quantity $n\bar{y}^2$ is called *correction for the mean*, and SSY is called the *corrected total sum of squares for* Y. The adjective *corrected* is often omitted, and SSY is usually simply referred to as the *total sum of squares for* Y.

We know that the best predictor of the Y value of a randomly chosen item from the population, in the absence of any knowledge about its X value, is μ_Y, the mean Y value of all items in the population, and that σ_Y is a measure of how well μ_Y represents the entire population. If data are obtained by simple random sampling, then we can estimate μ_Y by the sample mean $\bar{y}$, and σ_Y by

$$\hat{\sigma}_Y = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{(n-1)}} = \sqrt{\frac{SSY}{(n-1)}}$$

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.

The second key quantity is $SSE(X)$, or SSE for short, the sum of squares of the prediction errors when the sample regression function of Y on X is used to predict the Y values of the sample items. Recall that we defined SSE in (3.4.6), but for convenience we reproduce the definition here.

$$SSE = \sum_{i=1}^{n} \hat{e}_i^2 = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \tag{3.8.2}$$

and it has $(n-2)$ degrees of freedom associated with it. The quantity $SSE/(n-2)$ is often denoted by $MSE(X)$, or simply by MSE, and it is referred to as the *mean squared error* or *residual mean square*. Thus

$$MSE(X) = MSE = \frac{SSE(X)}{n-2} \tag{3.8.3}$$

More generally, when a quantity that is a sum of squares of estimated errors of prediction is divided by its associated degrees of freedom, the resulting quantity is referred to as a mean square error. Recall that $\sqrt{MSE} = \hat{\sigma}$ is the estimate of σ.

The third key quantity is the difference $SSY - SSE$, which is the amount by which the total sum of squares SSY is reduced, by using the regression of Y on X, to obtain SSE. This difference is called the *sum of squares due to regression* and is denoted by $SSR(X)$, or simply as SSR. Thus

$$SSR = SSY - SSE \tag{3.8.4}$$

It can also be easily verified that SSR is in fact equal to the quantity $\hat{\beta}_1^2(SSX)$.

In general, the quantity

$$\frac{SSR}{\text{degrees of freedom associated with } SSR}$$

is referred to as the *mean square due to regression.* For straight line regression, the degrees of freedom associated with SSR is 1, and hence the mean square due to regression, which is denoted by $MSR(X)$, or MSR for short, is the same as the sum of squares due to regression.

The quantities SSY, SSR, and SSE are generally displayed in a table, called an ANalysis Of VAriance (ANOVA) table, as in Table 3.8.1.

TABLE 3.8.1
ANOVA for Straight Line Regression

Source	Degrees of Freedom df	Sum of Squares SS	Mean square MS	Computed F-Value
Regression	1	SSR	MSR	$F_C = \frac{MSR}{MSE}$
Error	$n-2$	SSE	MSE	
Total	$n-1$	SSY	MSY	

The statistic F_C in the last column of Table 3.8.1 is sometimes used to test NH: $\beta_1 = 0$ against AH: $\beta_1 \neq 0$. The P-value for the test is equal to the value of α for which $F_C = F_{1-\alpha:1,n-2}$. This test is equivalent to the t-test of NH: $\beta_1 = 0$ against AH: $\beta_1 \neq 0$, which is a special case of the test of NH: $\beta_1 = q$ against AH: $\beta_1 \neq q$ described in Box 3.7.2. The reason is the following. It can be shown that the square of the student's t table-value $t_{1-\alpha/2:m}$ is equal to the F table-value $F_{1-\alpha:1,m}$; i.e.,

$$t_{1-\alpha/2:m}^2 = F_{1-\alpha:1,m}$$

Thus to test NH: $\beta_1 = 0$ with AH: $\beta_1 \neq 0$, instead of using the test statistic

$$t_C = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = \frac{\hat{\beta}_1\sqrt{SSX}}{\hat{\sigma}}$$

and comparing it with a student's t table-value, we could use the statistic

$$t_C^2 = F_C = \frac{\hat{\beta}_1^2 SSX}{\hat{\sigma}^2} = \frac{MSR}{MSE}$$

and compare it with an F table-value. The P-value based on the t-test will be identical to the P-value based on the corresponding F-test.

Caution:

Do not forget the fact that the F-test in an ANOVA table for straight line regression can be used only to test NH: $\beta_1 = 0$ against AH: $\beta_1 \neq 0$ and cannot be used to test NH: $\beta_1 = q$ against AH: $\beta_1 \neq q$ for *nonzero* values of q, nor can it be used for *one-sided* tests of hypotheses regarding β_1.

EXAMPLE 3.8.1

For the age and blood pressure data in Table 3.6.2 we compute an analysis of variance table and illustrate the F-test for NH: $\beta_1 = 0$ against AH: $\beta_1 \neq 0$. The analysis of variance is displayed in Table 3.8.2. The entries are obtained from the computations in Task 3.6.2. The P-value corresponding to $F_C = 1161.31$ with 1 and 22 degrees of freedom is less than 0.01 using Table T-5 in Appendix T. The hypothesis that $\beta_1 = 0$ can be tested using the t-statistic in Box 3.7.2. From Task 3.6.2 we get

$$t_C = \frac{1.6085 - 0}{0.0472} = 34.078$$

again yielding a P-value less than 0.01 (actually less than 0.0005). Note also that $t_C^2 = 1161.38 = F_C$ (to within rounding error). If assumptions (B) are satisfied, then $\hat{\sigma}_Y = \sqrt{MSY} = \sqrt{9514.6/23} = 20.34$ is a valid estimate of σ_Y. ■

TABLE 3.8.2

ANOVA for Age and Blood Pressure Data in Table 3.6.2

Source	Degrees of Freedom *df*	Sum of Squares *SS*	Mean square *MS*	Computed *F*-Value
Regression	1	9337.7	9337.7	$F_C = \frac{MSR}{MSE} = 1161.31$
Error	22	176.9	8.0	
Total	23	9514.6	413.68	

Problems 3.8

3.8.1 The following questions refer to the shelf-life data in Table 3.7.1, which are also stored in the file **shelflif.dat** on the data disk.

a Present an analysis of variance table.

b Use F_C from the ANOVA table in part (a) to test NH: $\beta_1 = 0$ against AH: $\beta_1 \neq 0$. What is the P-value for this test? Interpret the result.

c Calculate t_C for testing NH against AH in part (b). What is the P-value for this test? Interpret the result.

d Verify that the square of t_C in part (c) is equal to F_C in (b). Further verify that the P-value calculated from the t statistic in part (c) is the same as that calculated from the F statistic in part (b).

e What conclusion do you draw regarding β_1 based on the test in part (b)?

f Compute a 99% confidence interval for β_1. How will you use this confidence interval to *decide* whether or not β_1 is close enough to zero to be considered negligible for this problem?

g Write a short paragraph outlining your conclusions in parts (b)–(f) and give reasons for your statements.

3.8.2 The following refer to the age and blood pressure data discussed in Task 3.6.2. The data are in Table 3.6.2 and also in the file **agebp.dat** on the data disk.

a Present an analysis of variance table.

b Use F_C from the ANOVA in part (a) to test NH: $\beta_1 = 0$ against AH: $\beta_1 \neq 0$. Compute the P-value for this test. Interpret the result.

c Compute t_C for testing NH against AH in part (b). Compute the P-value for this test. Interpret the result.

d Verify that the square of t_C in part (c) is equal to F_C in part (b). Further verify that the P-value calculated from the t statistic in part (c) is the same as that calculated from the F statistic in part (b).

e What conclusion do you draw regarding β_1 based on the test in part (b)?

f Compute a 99% confidence interval for β_1. How will you use this confidence interval to *decide* whether or not β_1 is close enough to zero to be considered negligible for this problem?

g Write a paragraph outlining your conclusions in parts (b)–(f) and justify how you reached them.

3.8.3 Consider the grades26 data given in Table 3.2.2, which are also stored in the file **grades26.dat** on the data disk. Repeat parts (a)–(g) of Problem 3.8.2 for these data.

3.9
Coefficient of Determination and Coefficient of Correlation

We have seen that for a two-variable population $\{(Y, X)\}$, the best prediction of the Y value of a randomly chosen item, given that its X value is x is $\mu_Y(x)$, the value of the regression function of Y on X evaluated at $X = x$. If the X value of the item in question is not used, then the best prediction of the Y value of the item is μ_Y. Clearly, investigators have a choice. They can use $\mu_Y(x)$ to predict the Y value of the selected item, or they can use μ_Y to predict its Y value. Of course to use $\mu_Y(x)$ to predict the Y value of the item, we must know its X value, and there may be some

costs involved in determining the X value. Also, there is no guarantee that using $\mu_Y(x)$ rather than μ_Y to predict Y will improve the prediction sufficiently to justify the cost associated with measuring or observing X. The following example makes the point clearer.

EXAMPLE 3.9.1

Suppose a physician advising a patient with a brain tumor wants to predict the length of time the patient will live if no surgery is performed to remove the tumor. Also suppose that this patient may be regarded as a randomly chosen subject from a population of subjects afflicted with the same type of brain tumor who elected to forego surgery. We assume that the durations between diagnosis and death (referred to as *survival times*) are available for this population of subjects. It is thought that survival time Y is related to tumor severity score X on a scale of 1 to 10, which can be determined from various brain scans. The physician has two options. He can use μ_Y, the average survival time of all the patients in the population, to predict the survival time (i.e., the Y value of his patient, in which case there is no need to know the value of the severity score for this patient's tumor), or he can measure the severity score of the tumor, say its value is $X = 7$, and use $\mu_Y(7)$, the average survival time for all such tumor patients whose tumor severity at the time of diagnosis was equal to 7, to predict the patient's survival time. If $\mu_Y(x)$ is not much better than μ_Y for predicting the Y values (i.e., the survival times), then he has to decide whether the cost of obtaining the X value can be justified. ■

Thus the decision to choose between μ_Y and $\mu_Y(x)$ for predicting Y usually depends, at least in part, on (a) the cost of observing the X value, and (b) the improvement in prediction that is made possible by using the X value. In this connection the investigator may be interested in knowing the answers to the following questions:

1 How good is μ_Y as a predictor of the Y value of an item that is to be randomly chosen from the population?

2 How good is $\mu_Y(x)$ as a predictor of the Y value of an item that is to be randomly chosen from the population (note that in order to use $\mu_Y(x)$ to predict the Y value of an item, we must know its X value)? **(3.9.1)**

3 How much better is $\mu_Y(x)$ than μ_Y for predicting the Y value of a randomly chosen item?

4 Is μ_Y an adequate predictor of Y?

5 Is $\mu_Y(x)$ an adequate predictor of Y?

We answer these questions using population standard deviations of prediction errors as *summary measures* of how good predictors are.

1 The quantity σ_Y is the measure of how good μ_Y is as a predictor of the value of Y.

2 The quantity $\sigma = \sigma_{Y|X}$ is the measure of how good $\mu_Y(x)$ is as a predictor of the value of Y because, if we know that the X value of the chosen item is x, then we restrict our attention to the subpopulation of all items with $X = x$; $\mu_Y(x)$ is the mean and σ is the standard deviation for this subpopulation. Recall that to use $\mu_Y(x)$, we must know the X value for the item whose Y value is being predicted.

3 σ_Y/σ, or σ_Y^2/σ^2, or $\sigma_Y - \sigma$, or $\sigma_Y^2 - \sigma^2$ (or some other meaningful function of σ_Y and σ) is a measure of how much better $\mu_Y(x)$ is than μ_Y for predicting the value of Y. In this book, we will use σ_Y/σ to describe how much better $\mu_Y(x)$ is than μ_Y for predicting Y.

4 Whether or not μ_Y is adequate for predicting the value of Y depends on the particular problem. An investigator may consider μ_Y to be an adequate predictor of the value of Y if most of the Y values, say at least a proportion p of the population, lie close to μ_Y, say within a distance of d units from μ_Y (p and d are specified by the investigator). It can be shown that when population assumptions (B) for straight line regression are satisfied, a proportion p of the population values lie in the interval $\mu_Y - z_{(1+p)/2}\sigma_Y$ to $\mu_Y + z_{(1+p)/2}\sigma_Y$. So at least a proportion p of the population values will lie in the interval $\mu_Y - d$ to $\mu_Y + d$ provided

$$z_{(1+p)/2}\sigma_Y < d \tag{3.9.2}$$

i.e., if

$$\sigma_Y < d/z_{(1+p)/2} \tag{3.9.3}$$

5 As in item (4), an investigator may consider $\mu_Y(x)$ to be an adequate predictor of the Y value of a randomly chosen item whose X value is known to be equal to x if at least a proportion p of the Y values in the subpopulation with $X = x$ is within d units from the predicted value $\mu_Y(x)$. When population assumptions (A) or (B) for straight line regression are satisfied, this will be true provided that

$$z_{(1+p)/2}\sigma < d \tag{3.9.4}$$

i.e., if

$$\sigma < d/z_{(1+p)/2} \tag{3.9.5}$$

Note Keep in mind that $\mu_Y(x)$ and μ_Y may *both* be adequate for predicting Y or that *neither one* may be adequate. Also note that we are discussing population prediction functions. In practice we use estimates of these prediction functions based on sample data. The sample prediction functions do not perform as well as the population prediction functions, but if the estimates are based on sufficiently large samples, then we can expect the sample prediction functions to perform almost as well as the population prediction functions.

A further complication arises in judging the adequacy of various prediction functions, viz., in practice we do not know the values of σ_Y or σ. However, sample data may be used to obtain confidence bounds for these quantities that will allow us to

determine whether or not the prediction function is adequate. The following task illustrates this point.

Task 3.9.1

To illustrate the preceding ideas, consider the crystal data of Task 3.4.1 where we want to predict the weight Y of a crystal using the amount of time X for which the crystal is grown.

1 Suppose that $\mu_Y(x)$ will be considered an adequate predictor of the Y values if, for each allowable value x of X, a proportion $p = 0.90$ or more of the Y values lie within 0.5 gram of the predicted values. Compute a 95% confidence statement to help decide whether or not $\mu_Y(x)$ is an adequate predictor of Y.

Here $p = 0.90$ and $d = 0.5$. Since assumptions (A) are presumed to be valid, we can conclude using (3.9.5) that $\mu_Y(x)$ is an adequate predictor of Y provided

$$\sigma < d/z_{(1+p)/2}$$

Because $z_{(1+p)/2} = z_{0.95} = 1.645$ from Table T-1 in Appendix T, we can conclude $\mu_Y(x)$ is an adequate predictor of Y if

$$\sigma < 0.5/z_{0.95} = 0.5/1.645 = 0.304$$

Unfortunately we do not know the precise value of σ, and so we do not know whether or not $\sigma \leq 0.304$. In Task 3.4.1 we calculated the point estimate of σ and obtained $\hat{\sigma} = 1.062$. Using (3.6.8) we get the following 95% two-sided confidence interval for σ:

$$C[0.76 \leq \sigma \leq 1.75] = 0.95$$

Thus we can be 95% confident that σ is between 0.76 and 1.75; Using this we would perhaps conclude that $\mu_Y(x)$ is not an adequate predictor of Y.

There are instances where it is difficult, or even impossible, to specify a criterion of adequacy for a prediction function. In such instances, the investigator may be interested in *comparing* the standard deviations of the prediction errors corresponding to each regression function under consideration. In the present situation there are two possible quantities, μ_Y and $\mu_Y(x)$, for predicting Y, and hence the investigator may be interested in comparing σ_Y with σ. She may want either to examine σ_Y and σ individually or to examine some function of σ_Y and σ that may be particularly meaningful in a given problem. One function of σ_Y and σ that has found widespread use in the literature is the coefficient of determination, which we discuss next.

Coefficient of Determination

A commonly used measure that summarizes the performance of $\mu_Y(x)$ as a predictor of Y, *relative* to μ_Y, is the *coefficient of determination* of Y with X, denoted by $\eta^2_{Y,X}$.

This is defined by (η is the Greek letter eta)

$$\eta^2_{Y,X} = \frac{\sigma_Y^2 - \sigma^2}{\sigma_Y^2} \tag{3.9.6}$$

Recall that σ_Y is the standard deviation of the prediction errors when μ_Y is used to predict Y, whereas σ is the standard deviation of the prediction errors when $\mu_Y(x)$ is used to predict Y. It can be shown that, under assumptions (A) or (B), σ^2 cannot be greater than σ_Y^2, and therefore the quantity on the right-hand side of (3.9.6) cannot be negative. Thus $\eta^2_{Y,X}$ is the *proportional reduction in the variance of prediction errors when using $\mu_Y(x)$ rather than μ_Y to predict Y.*

An alternative measure of relative performance of $\mu_Y(x)$ relative to μ_Y is $\delta_{Y,X}$, the *proportional reduction in the standard deviation of prediction errors*, and it is defined by

$$\delta_{Y,X} = \frac{\sigma_Y - \sigma}{\sigma_Y}$$

These two measures are related by the equation

$$\delta_{Y,X} = 1 - \sqrt{1 - \eta^2_{Y,X}}$$

so that either quantity can be obtained from a knowledge of the other. Because $\eta^2_{Y,X}$ is the measure that is traditionally used by statisticians and practitioners, we consider only this measure although we believe that $\delta_{Y,X}$ is also a meaningful measure.

Relation of $\eta^2_{Y,X}$ to (Pearson's) Coefficient of Correlation $\rho_{Y,X}$

Recall that Pearson's coefficient of correlation (also called the *simple correlation coefficient* or *product moment correlation coefficient*), defined in (1.5.1) and denoted by $\rho_{Y,X}$, is a measure of the *linear association* between Y and X. It can be shown that *when the regression function of Y on X is of the form*

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

(in particular, when population assumptions (A) or (B) are satisfied), the coefficient of determination, $\eta^2_{Y,X}$, of Y with X, is in fact the square of Pearson's correlation coefficient $\rho_{Y,X}$, i.e., $\eta^2_{Y,X} = \rho^2_{Y,X}$.

This is the reason that the symbol $\rho^2_{Y,X}$ is often used to denote the coefficient of determination of Y with X, but you should be aware that if the regression function of Y on X is not of the form $\mu_Y(x) = \beta_0 + \beta_1 x$ (i.e., it is not linear in x), then $\eta^2_{Y,X}$ is not equal to $\rho^2_{Y,X}$. To avoid any possibility of confusion, we should use the symbol $\eta^2_{Y,X}$ to denote the coefficient of determination of Y with X. However, because this chapter deals with only straight line regression, we are allowed to use $\eta^2_{Y,X}$ and $\rho^2_{Y,X}$ interchangeably in the discussions that follow. Accordingly, *we use the symbol $\rho^2_{Y,X}$ to denote the coefficient of determination in the rest of this chapter.*

Suppose that (population) assumptions (A) or (B) are satisfied. Then $\mu_Y(x) = \beta_0 + \beta_1 x$, and the statements in Box 3.9.1 are true.

BOX 3.9.1 **Properties of $\rho^2_{Y,X}$**

1. $\sigma \leq \sigma_Y$ and consequently $\rho^2_{Y,X} \geq 0$.
2. $\rho^2_{Y,X} = 0$ if and only if $\beta_1 = 0$; i.e., if and only if $\sigma = \sigma_Y$.
3. When $\beta_1 \neq 0$, the sign of $\rho_{Y,X}$ is the same as the sign of β_1.
4. If $\rho^2_{Y,X} = 0$, then $\mu_Y(x) = \beta_0 + \beta_1 x$ is no better for predicting Y than μ_Y is.
5. $\rho^2_{Y,X} = 1$ if and only if $\mu_Y(x)$ is a *perfect predictor* of Y. In this case $\sigma = 0$.
6. The larger the value of $\rho^2_{Y,X}$ is, the better the prediction of Y will be using X; i.e., the predicted values will tend to be closer to the true values.

Remark As stated in item (4) of Box 3.9.1, the statement $\rho^2_{Y,X} = 0$ means that the regression function of Y on X, namely $\mu_Y(x) = \beta_0 + \beta_1 x$, is no better for predicting Y than the population mean μ_Y is. However, *if $\mu_Y(x)$ is not of the form $\beta_0 + \beta_1 x$, then we cannot draw such a conclusion.*

When the straight line regression model does not hold, it is quite possible that $\rho^2_{Y,X} = 0$ and $\eta^2_{Y,X} \neq 0$, so the function $\mu_Y(x)$ may be a *much better predictor* (in fact, $\mu_Y(x)$ may even be a perfect predictor) of Y than μ_Y is. An illustration of this is given in the conversation later in this section.

Point Estimation for $\rho_{Y,X}$

The quantities $\rho_{Y,X}$ and $\rho^2_{Y,X}$ are population parameters, and the discussion about them has centered around their use and meaning in the population. Valid point estimates of $\rho^2_{Y,X}$ and $\rho_{Y,X}$ can be calculated from sample data according to the formulas given in (3.9.7) and (3.9.8), respectively, provided that assumptions (B) are satisfied. In particular, the data must be obtained by simple random sampling.

$$\hat{\rho}^2_{Y,X} = \frac{SSY - SSE(X)}{SSY} = \frac{SSR(X)}{SSY} = \frac{(SXY)^2}{(SSX)(SSY)} \tag{3.9.7}$$

and

$$\hat{\rho}_{Y,X} = \frac{SXY}{\sqrt{(SSX)(SSY)}} \tag{3.9.8}$$

If data are obtained by sampling with preselected X values, then no valid estimate of $\rho_{Y,X}$ or $\rho^2_{Y,X}$ is available from the sample data.

As we have stated several times, in the case of straight line regression, investigators can use $\rho^2_{Y,X}$ to decide whether $\mu_Y(x)$ is better than μ_Y for predicting Y. Technically, if population assumptions (A) or (B) are satisfied and $\rho^2_{Y,X} \neq 0$, then we can conclude that $\mu_Y(x)$ is better than μ_Y to predict Y because if $\rho^2_{Y,X} \neq 0$, then it follows that $\sigma < \sigma_Y$. However, in practice, we generally want to know *how much smaller* σ is than σ_Y. To find this out, we can examine the estimated values of σ and σ_Y. An alternate approach is to compute a confidence interval for the *ratio* σ_Y/σ and use this information in making the required decision about how much smaller σ is than σ_Y.

It turns out that to compute a confidence interval for the ratio σ_Y/σ, it is prudent to compute a confidence interval for $\rho_{Y,X}$ as an intermediate step because the table-values that are necessary are readily available in the case of $\rho_{Y,X}$, but that is not the case for σ_Y/σ. Hence we first describe the procedure for obtaining a confidence interval for $\rho_{Y,X}$.

Confidence Interval for $\rho_{Y,X}$

The procedure for computing a two-sided $1-\alpha$ confidence interval for $\rho_{Y,X}$ is given in Box 3.9.2, and it is valid when assumptions (B) are satisfied.

BOX 3.9.2 **Two-sided $1-\alpha$ Confidence Interval for $\rho_{Y,X}$**

1. Denote the estimated correlation coefficient $\hat{\rho}_{Y,X}$ by r.
2. Select the chart in Table T-7 in Appendix T corresponding to the desired $1-\alpha$.
3. Find the number corresponding to the computed correlation coefficient r.
4. Go vertically up the graph along the line at r until you encounter the first curve corresponding to sample size n.
5. Go horizontally from this point on the curve toward the left margin (the ρ margin) until you encounter the vertical axis, say at the point corresponding to $\rho = L$. Then the number L is a $1-\alpha/2$ lower confidence bound for $\rho_{Y,X}$; i.e., $C[L \le \rho_{Y,X}] = 1-\alpha/2$.
6. Go vertically up the graph along the line at r until you encounter the second curve corresponding to sample size n.
7. Go horizontally from this point on the curve toward the left margin (the ρ margin) until you encounter the vertical axis, say at the point corresponding to $\rho = U$. Then the number U is a $1-\alpha/2$ upper confidence bound for $\rho_{Y,X}$, i.e., $C[\rho_{Y,X} \le U] = 1-\alpha/2$.
8. A two-sided $1-\alpha$ confidence interval for $\rho_{Y,X}$ is $L \le \rho_{Y,X} \le U$, and the confidence statement is $C[L \le \rho_{Y,X} \le U] = 1-\alpha$.

Note When $\hat{\rho}_{Y,X}$ is close to zero or one, the charts in Table T-7 are difficult to read. However, if you are careful in reading the charts, the procedure is adequate for most problems.

Confidence Interval for σ_Y/σ

The procedure for computing a two-sided confidence interval for σ_Y/σ is given in Box 3.9.3, and it is valid under assumptions (B).

BOX 3.9.3 Two-Sided $1-\alpha$ Confidence Interval for σ_Y/σ

1 Obtain L as in Box 3.9.2 corresponding to the desired value of $1-\alpha$.

2 Obtain U as in Box 3.9.2 corresponding to the desired value of $1-\alpha$.

3 Let Q_1 denote the larger, and Q_2 the smaller, of the two numbers L^2 and U^2, respectively.

4 Let $U_0 = \frac{1}{\sqrt{1-Q_1}}$.

5 If L and U are of the same sign, then let $L_0 = \frac{1}{\sqrt{1-Q_2}}$; if they are of opposite signs, then let $L_0 = 1$.

6 A $1-\alpha$ two-sided confidence statement for σ_Y/σ is given by

$$C[L_0 \leq \sigma_Y/\sigma \leq U_0] = 1-\alpha$$

L_0 is a $1-\alpha/2$ lower confidence bound, and U_0 is a $1-\alpha/2$ upper confidence bound, respectively, for σ_Y/σ.

The following example illustrates the computations discussed in this section.

EXAMPLE 3.9.2

Consider the power plant SO_2 data given in Table 3.5.4 where we want to study the association between SO_2 concentrations Y at a national park and SO_2 emissions X from a nearby power plant. The data are also in the file **so2.dat** on the data disk. Suppose the investigator wants to know how much improvement in prediction is possible if $\mu_Y(x)$, the regression function of Y on X, is used to predict Y instead of μ_Y; i.e., the investigator wants to know how much smaller σ is than σ_Y. Thus an appropriate population parameter of interest is $\eta^2_{Y,X}$; which is equal to $\rho^2_{Y,X}$ because the regression function of Y on X is assumed to be a straight line. For these data we have

$$SSX = 31.823 \qquad SSY = 133.468 \qquad SXY = 58.0696$$

Using (3.9.8) we obtain

$$\hat{\rho}_{Y,X} = \frac{SXY}{\sqrt{(SSX)(SSY)}} = \frac{58.0696}{\sqrt{(31.823)(133.468)}} = 0.8910$$

Therefore $\hat{\rho}^2_{Y,X} = \hat{\eta}^2_{Y,X} = 0.7939$. Thus we estimate that the prediction error *variance* can be reduced 79.39% by using $\mu_Y(x)$ rather than μ_Y to predict Y.

A 95% two-sided confidence interval for $\rho_{Y,X}$ is obtained, using the procedure described in Box 3.9.2, as follows. From the chart in Table T-7 we obtain $L = 0.77$ and $U = 0.95$ (approximately), corresponding to $n = 14$ and $\hat{\rho}_{Y,X} = 0.89$. Hence we get the following confidence statement:

$$C[0.77 \leq \rho_{Y,X} \leq 0.95] = 0.95$$

Now let us use the procedure in Box 3.9.3 to compute a two-sided 95% confidence interval for σ_Y/σ. We have, as previously calculated, $L = 0.77$ and $U = 0.95$ so

that $Q_1 = 0.9025$ and $Q_2 = 0.5929$. Because L and U are of the same sign, we get $L_0 = 1/\sqrt{1 - 0.5929} = 1.57$, whereas $U_0 = 1/\sqrt{1 - 0.9025} = 3.20$. Thus we have the confidence statement

$$C[1.57 \leq \sigma_Y/\sigma \leq 3.20] = 0.95$$

In particular if we want only a lower confidence bound for σ_Y/σ, we have

$$C[1.57 \leq \sigma_Y/\sigma] = 0.975$$

so we have 97.5% confidence that the standard deviation of the prediction errors when using μ_Y to predict Y is at least 1.57 times as big as the standard deviation of the prediction errors when using $\mu_Y(x)$ to predict Y. The investigator can use this information to decide whether to use $\mu_Y(x)$ or μ_Y to predict Y. ▪

Remark While the coefficient of determination may be useful for comparing the performance of $\mu_Y(x)$ *relative to* μ_Y for predicting the values of Y, it cannot be used to determine whether or not $\mu_Y(x)$ is an *adequate* prediction function. This is illustrated in the following example.

EXAMPLE 3.9.3

Consider the problem discussed in Task 3.6.2 concerning the relationship between age and blood pressure. The data are in Table 3.6.2 and also in the file **agebp.dat** on the data disk. The investigator wants to know how much improvement in prediction is possible if the regression function $\mu_Y(x)$ of blood pressure on age is used for predicting blood pressure instead of μ_Y, the mean blood pressure of all individuals in the population (i.e., the investigator wants to know how much smaller σ is than σ_Y). Thus an appropriate population parameter of interest is $\eta^2_{Y,X}$, which is equal to $\rho^2_{Y,X}$ since the regression function of Y on X is a straight line. Using (3.9.8) we obtain

$$\hat{\rho}_{Y,X} = \frac{SXY}{\sqrt{(SSX)(SSY)}} = \frac{5805.13}{\sqrt{(3608.96)(9514.63)}} = 0.9907$$

The fact that the estimated value of $\rho^2_{Y,X}$ is 0.981 means that the improvement in prediction, by using $\mu_Y(x)$ rather than μ_Y to predict Y, appears to be substantial. *However, this does not necessarily imply that age is an adequate predictor of blood pressure.* That depends on the particular application at hand. For instance, suppose the investigator wants to predict blood pressures of individuals accurately to within $d = 5.0$ units for at least a proportion $p = 0.99$ of the individuals in the population. If he uses $\mu_Y(x)$ to predict the Y values, then from (3.9.5) this would be true provided that

$$\sigma < 5/z_{(1+p)/2}$$

i.e., if

$$\sigma < 5/z_{0.995} = 5/2.575 = 1.942$$

because $z_{0.995} = 2.575$ from Table T-1 in Appendix T . Unfortunately, we do not know the exact value of σ based on the sample data, but for these data we have

$\hat{\sigma} = 2.836$, and a 95% confidence interval for σ is given by the confidence statement $C[2.193 \leq \sigma \leq 4.014] = 0.95$. Using this confidence statement the investigator would perhaps conclude that $\mu_Y(x)$ is not an adequate prediction function for this problem.

Observe that $\hat{\sigma}_Y = 20.34$, which is much greater than $\hat{\sigma} = 2.836$. Consequently using age to predict blood pressure does appear to reduce the prediction error considerably, but not quite enough to make it an adequate predictor of blood pressure for this problem. ■

Authors' Recommendation

The use of $\rho_{Y,X}$ or $\rho^2_{Y,X}$ to determine whether X is an *adequate* predictor of Y is incorrect. Instead the investigator should use σ to judge the adequacy of $\mu_Y(x)$ as a predictor of Y in the context of the study. When an investigator needs to decide whether $\mu_Y(x)$ or μ_Y should be used to predict Y, we recommend that the investigator examine both $\hat{\sigma}$ and $\hat{\sigma}_Y$ and appropriate functions of them. *Do not make this decision based only on the estimated value of* $\rho_{Y,X}$ *or* $\rho^2_{Y,X}$.

Conversation 3.9

Investigator: I'd like to get your help in interpreting correlation coefficients. One of our scientists says some of his data suggest that the population correlation coefficient between Y and X is about 0.98, and that this means factor X is an adequate predictor of Y. Is that true?

Statistician: No, it is not necessarily true. For example, suppose that $\sigma_Y = 100$ and $\sigma = 20$. If the regression function is $\mu_Y(x) = \beta_0 + \beta_1 x$, then

$$\rho^2_{Y,X} = \frac{(10{,}000 - 400)}{10{,}000} = 0.96$$

and so the magnitude of $\rho_{Y,X}$ is $\sqrt{0.96} = 0.98$. But if the scientist decides that $\mu_Y(x)$ is an adequate predictor of Y only if σ is less than 10 units, then X is not an adequate predictor of Y for this problem (because $\sigma = 20$ in this example) even though $\rho_{Y,X} = 0.98$. The thing to remember is that $\rho_{Y,X}$ (or $\rho^2_{Y,X}$) *by itself* is of almost no value in determining whether factor X is an *adequate* predictor of Y. What we do learn from a knowledge of $\rho^2_{Y,X}$, in the case of straight line regression, is *relatively how much better* $\mu_Y(x) = \beta_0 + \beta_1 x$ *is than* μ_Y *as a predictor of* Y.

Investigator: I see what you're saying, but if $\rho_{Y,X} = 0$, doesn't that mean that factor X is of no value in predicting Y?

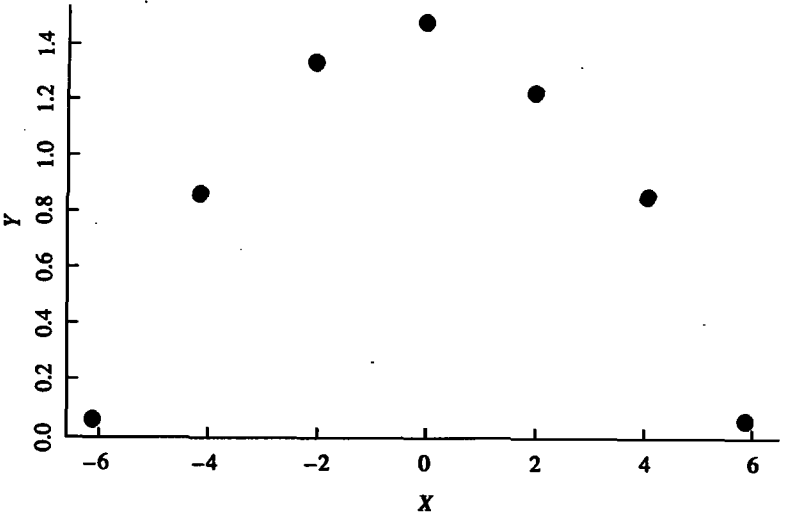
Statistician: It does if we know *a priori* that the regression function is $\mu_Y(x) = \beta_0 + \beta_1 x$ (i.e., if the regression function is linear in x). Otherwise $\rho_{Y,X} = 0$ cannot be interpreted

to mean that X is of no use in predicting Y. I have concocted some data that I will show you and let you make up your own mind as to whether factor X is useful for predicting Y. The data are

Y	X
1.335	−1.9
0.050	5.9
1.230	1.9
0.052	−6.1
0.858	4.0
1.489	0.3
0.861	−4.2

and the estimated correlation coefficient, $\hat{\rho}_{Y,X}$ of Y and X, is 0.008; i.e., $\hat{\rho}^2_{Y,X} = 0.000064$. So the data suggest that the square of the correlation coefficient between X and Y is essentially equal to zero. We'll plot the data (see Figure 3.9.1) to see what it looks like.

FIGURE 3.9.1



It appears that there is a strong relationship between y_i and x_i. In fact, it appears from the graph that Y can be predicted very well by some function of X, in this case by the square of X. So factor X is clearly very useful in predicting Y even though $\rho^2_{Y,X}$ is essentially zero.

Investigator: I'm getting more discouraged by the minute. First you tell me that a correlation coefficient of 0.98 may *not* indicate that X is an adequate predictor of Y, and then you tell me that even if $\rho^2_{Y,X} = 0$, factor X may be quite useful for predicting Y.

Statistician: That is correct. The correlation coefficient of X with Y is mainly useful when the regression function is $\mu_Y(x) = \beta_0 + \beta_1 x$ (i.e., linear in x), and in other situations it can in fact be misleading. Of course when the population $\{(Y, X)\}$ is bivariate Gaussian, then the regression function of Y on X is of the form $\mu_Y(x) = \beta_0 + \beta_1 x$, and in that case correlation can be a useful summary measure for that population. However, even in that case a correlation coefficient, when used alone, does not tell you how good factor X is for predicting Y. It is the value of σ that tells you how good factor X is for predicting Y. So my advice for the scientists you work with is this: *Do not use the correlation coefficient to decide whether factor X is useful for predicting Y*. On the other hand, if you want to determine how much better $\mu_Y(x) = \beta_0 + \beta_1 x$ is than μ_Y for predicting Y, then compare σ_Y with σ. One way to compare them is by examining their ratio, and we see that

$$\rho^2_{Y,X} = \frac{\sigma_Y^2 - \sigma^2}{\sigma_Y^2} = 1 - \left(\frac{\sigma}{\sigma_Y}\right)^2 \tag{3.9.9}$$

From this we get

$$\sqrt{1 - \rho^2_{Y,X}} = \frac{\sigma}{\sigma_Y} \tag{3.9.10}$$

Thus we see that correlation coefficients can be useful in determining how much better $\mu_Y(x) = \beta_0 + \beta_1 x$ is than μ_Y for predicting Y. Of course we've been discussing populations and population parameters, but in a real problem sample values have to be used and the required population quantities have to be estimated.

Investigator: Suppose I have two predictor factors X_1 and X_2, and the regression functions of Y on X_1 and Y on X_2 are both linear. Suppose that ρ^2_{Y,X_1} is two times larger than ρ^2_{Y,X_2}. Does this mean that X_1 is two times better than X_2 for predicting Y?

Statistician: No, it doesn't. Because we are using $\sigma_{Y|X_1}$ and $\sigma_{Y|X_2}$ as the appropriate measures of how good X_1 and X_2, respectively, are for predicting Y, it is better to compare them with one another. From equation (3.9.10) we get

$$\frac{\sqrt{1 - \rho^2_{Y,X_1}}}{\sqrt{1 - \rho^2_{Y,X_2}}} = \frac{\sigma_{Y|X_1}/\sigma_Y}{\sigma_{Y|X_2}/\sigma_Y} = \frac{\sigma_{Y|X_1}}{\sigma_{Y|X_2}}$$

So if $\rho^2_{Y,X_1} = 0.90$ and $\rho^2_{Y,X_2} = 0.45$, then $\rho^2_{Y,X_1} = 2\rho^2_{Y,X_1}$ and

$$\frac{\sigma_{Y|X_1}}{\sigma_{Y|X_2}} = \sqrt{0.10/0.55} = \sqrt{0.18182} = 0.4264$$

On the other hand, if $\rho^2_{Y,X_1} = 0.40$ and $\rho^2_{Y,X_2} = 0.20$, then again $\rho^2_{Y,X_1} = 2\rho^2_{Y,X_2}$ but

$$\frac{\sigma_{Y|X_1}}{\sigma_{Y|X_2}} = \sqrt{0.60/0.80} = \sqrt{0.75} = 0.866$$

So $\sigma_{Y|X_1} = 0.4264\sigma_{Y|X_2}$ when $\rho^2_{Y,X_1} = 2\rho^2_{Y,X_2}$ with $\rho^2_{Y,X_1} = 0.90$ and $\rho^2_{Y,X_2} = 0.45$, but $\sigma_{Y|X_1} = 0.866\sigma_{Y|X_2}$ when $\rho^2_{Y,X_1} = 2\rho^2_{Y,X_2}$ with $\rho^2_{Y,X_1} = 0.40$ and $\rho^2_{Y,X_2} = 0.20$. Thus if one value of ρ^2 is twice as large as another, that does not tell us how much better X_1 is than X_2 for predicting Y.

Investigator: I understand what you're saying, but I'll have to think about all this. Thank you for your time. Perhaps I'll have more questions next week.

Problems 3.9

3.9.1 Consider the sample data in Table 3.2.3, which is also in the file **table323.dat** on the data disk. Assumptions (B) are presumed valid.

- **a** Plot y_i against x_i to determine whether it appears that the regression function of Y on X is linear in X. Based on this plot, do you think it is?
- **b** Estimate $\rho_{Y,X}$ the coefficient of correlation between Y and X. Based on this estimate can you tell
  - **i** how good X is for predicting Y?
  - **ii** how much better it is to use $\mu_Y(x)$ rather than μ_y for predicting Y?
  - **iii** *relatively* how much better it is to use $\mu_Y(x)$ for predicting Y than to use μ_Y?
- **c** Estimate σ and σ_Y. Based on these estimates can you decide whether $\mu_Y(x)$ is an adequate predictor of Y?

3.9.2 Consider the blood pressure data given in Table 3.6.2.

a–c Repeat (a)–(c) of Problem 3.9.1 for these data.

- **d** The investigator wants to know whether age is an adequate predictor of blood pressure Y. In other words he wants to know if $\mu_Y(x) = \beta_0 + \beta_1 x$ is an adequate predictor of Y. He will decide that $\mu_Y(x)$ is an adequate predictor of Y if a proportion $p = 0.95$ of the blood pressures of all individuals who are x years old is within 10 units of the predicted value $\mu_Y(x)$. What information about σ is needed for the investigator to determine whether or not $\mu_Y(x) = \beta_0 + \beta_1 x$ is an adequate predictor of Y?
- **e** In (d) compute an appropriate 90% confidence statement that will help the investigator decide whether $\mu_Y(x) = \beta_0 + \beta_1 x$ is an adequate predictor of Y.

3.9.3 Consider the crystal data in Table 3.4.2.

a–c Repeat parts (a)–(c) of Problem 3.9.1 for these data.

d The investigator wants to decide whether X, the number of hours that crystals grow, is a good predictor of Y, the weight of crystals. In particular she wants to decide whether $\mu_Y(x) = \beta_0 + \beta_1 x$ is an adequate predictor of crystal weight Y. She will decide that $\mu_Y(x)$ is indeed an adequate predictor of Y if she determines that a proportion $p = 0.99$ of all crystals that grow x hours will weigh within 1 gram of the predicted value $\mu_Y(x)$. What information about σ is needed for the investigator for her to determine whether or not $\mu_Y(x)$ is an adequate predictor of Y?

e In (d) compute an appropriate 95% confidence statement that will help the investigator decide whether $\mu_Y(x)$ is an adequate predictor of Y.

3.10
Regression Analysis When There Are Measurement Errors

Thus far in this chapter we have used assumptions (A) or assumptions (B), which among other things require that the response variable Y as well as the predictor variable X can be observed without measurement error. For instance, in Task 3.4.1 it is assumed that the measured (observed) values of crystal weight Y and time X are the true values. The results based on these assumptions are generally satisfactory when measurement errors are present but negligible. However, in many instances the measurement errors are not negligible and it becomes necessary to explicitly account for the presence of these errors in the theoretical development and in the application of linear regression analysis. In this section we consider regression when there are errors in measuring the predictor variable X and/or the response variable Y.

Note In this section we consider several two-variable populations, so to avoid possible confusion we use the more complete notation $\sigma_{Y|X}$, instead of the simpler notation σ, to denote the standard deviation of the Y values in any subpopulation determined by X.

Measurement Error

Let $(y_1, x_1), \cdots, (y_n, x_n)$ represent sample data from the population $\{(Y, X)\}$ obtained either by simple random sampling or by sampling with preselected X values.

Suppose the true Y value of sample item i is y_i and the *observed* (or recorded or measured) Y value is denoted by v_i. The measurement error in observing the ith sample value of the response variable Y is defined by

D E F I N I T I O N
Y measurement error (for sample item i) = observed Y value (v_i) − true Y value (y_i). ■

So we write

$$d_i = v_i - y_i \quad \text{for} \quad i = 1, \ldots, n \tag{3.10.1}$$

Thus d_i is the error in measuring the Y value of the ith sample item.

Suppose the true X value of sample item i is x_i and the *observed* (or recorded or measured) X value is denoted by u_i. The measurement error in observing the ith sample value of the predictor variable X is defined by

D E F I N I T I O N

X measurement error (for sample item i) = observed X value (u_i) − true X value (x_i). ■

So we write

$$e_i = u_i - x_i \quad \text{for} \quad i = 1, \ldots, n \tag{3.10.2}$$

Thus e_i is the error in measuring the X value of the ith sample item.

In this situation it is useful to introduce some additional notation so that we can easily distinguish between true values and measured values. Suppose, as usual, Y_I, X_I denote the true Y and X values for population item I. Now imagine that we measure the Y and X values of each population item. The (conceptual) measured Y value for population item I is denoted by V_I, and the (conceptual) measured X value for this item is denoted by U_I. Thus we have a conceptual two-variable population $\{(V, U)\}$ consisting of the measured values of Y and X for each population item. The true Y and X values for sample item i are denoted by y_i and x_i, respectively. Likewise, the measured Y and X values for sample item i are denoted by v_i and u_i, respectively. Keep in mind that the quantities y_i, x_i will remain unknown due to imprecise measuring techniques, but the quantities v_i, u_i will be known for each sample item.

The assumptions in Box 3.10.1 are made about the measurement errors.

B O X 3.10.1 **Assumptions about Measurement Errors**

1 The errors d_i in measuring the response variable Y are assumed to be a random sample from a Gaussian population with mean zero and standard deviation σ_d.

2 The errors e_i in measuring the predictor variable X are assumed to be a random sample from a Gaussian population with mean zero and standard deviation σ_e.

3 For the *Berkson model* to be discussed, the errors are all assumed to be statistically independent of each other and of Y and U (i.e., the value of any one error is not related to the values of any other errors and is not related to the values of U nor to the values of Y).

4 For the *classical errors in variables model* to be discussed, the errors are all assumed to be statistically independent of each other and of Y and X (i.e., the value of any one error is not related to the values of any other errors and is not related to the values of X nor to the values of Y).

One of the principal reasons for studying the population $\{(Y, X)\}$ is to obtain valid point estimates and confidence intervals *(by **valid** we mean point estimates that are unbiased, or nearly so, and confidence intervals with specified confidence coefficients, or nearly so)* for various population quantities of interest. When assumptions (A) or assumptions (B) are satisfied, the regression function of Y on X is given by

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{3.10.3}$$

with subpopulation standard deviations all equal to $\sigma_{Y|X}$. In this case we are typically interested in the following population quantities:

$$\beta_0, \quad \beta_1, \quad \mu_Y(x) = \beta_0 + \beta_1 x, \quad Y(x), \quad a_0\beta_0 + a_1\beta_1 \tag{3.10.4}$$

where a_0 and a_1 are specified constants, and x is a specified value of X. Often we are also interested in

$$\sigma_{Y|X} \qquad \rho_{Y,X} \tag{3.10.5}$$

If there are measurement errors in X and/or Y, it may *not* be possible to obtain valid estimates for all the quantities in (3.10.4) and (3.10.5).

When the two-variable population $\{(Y, X)\}$ satisfies population assumptions (A) (respectively, population assumptions (B)), and when the measurement errors satisfy the conditions in Box 3.10.1, it can be shown that the two-variable population $\{(V, U)\}$ must also satisfy population assumptions (A) (respectively, population assumptions (B)). In particular, the regression function of V on U is also a straight line function. We denote this function by

$$\mu_V(u) = \beta_0^* + \beta_1^* u \tag{3.10.6}$$

where β_0^*, β_1^* may or may not be different from β_0, β_1 in (3.10.3). The standard deviation of the subpopulations of V values determined by U is denoted by $\sigma_{V|U}$. More is said about this later. Because we want to predict Y using the measured value U of X, it is useful to consider another two-variable population, viz., the population $\{(Y, U)\}$. Again we can mathematically prove that this population satisfies population assumptions (A) (respectively, population assumptions (B)) and that the regression function of Y on U is given by

$$\mu_Y(u) = \beta_0^* + \beta_1^* u \tag{3.10.7}$$

where the regression coefficients β_0^* and β_1^* in (3.10.7) are *identical* to the regression coefficients in (3.10.6). It is this fact that enables us to use U to predict Y even though we cannot observe the true Y values of the sample items. Since we can observe the values of V and U for each sample item, it is possible to obtain valid estimates for $\mu_V(u)$. But $\mu_V(u)$ is identical to $\mu_Y(u)$. Thus $\hat{\mu}_V(u)$ can be used to predict the true Y value of a randomly chosen item from the study population. Specifically, we can use sample data v_i, u_i to estimate β_0^*, β_1^* and $\mu_V(u)$, and we can use this to predict the true Y values using the measured values of U. The standard deviation of the subpopulation of Y values determined by U is denoted by $\sigma_{Y|U}$.

We now examine the consequences of measurement errors in some detail. First we consider the case when there are errors in measuring the response variable Y only, but the predictor variable X is measured without error.

Measurement Errors in the Response Variable Y But Not in the Predictor Variable X

Suppose $\{(Y, X)\}$ is the population under study, and assumptions (A) (respectively, assumptions (B)) are satisfied except that the values of the response variable Y are measured with error. Here we assume that the values of the predictor variable X are measured without error. Thus the true X values $x_1, \ldots, x_n$ of the sample items are available; however, the true Y values $y_1, \ldots, y_n$ of these items are not available, but their measured values $v_1, \ldots, v_n$ are. Thus we consider the population $\{(V, X)\}$, which also satisfies population assumptions (A) (respectively, population assumptions (B)) with the regression function given by

$$\mu_V(x) = \beta_0 + \beta_1 x$$

where β_0, β_1 are the same as in (3.10.3).

In this situation, point estimates for the quantities in (3.10.4) are computed as usual using the formulas in Box 3.4.2, and confidence intervals are computed using the formulas (3.6.1)–(3.6.7) using the data $(v_1, x_1), \ldots, (v_n, x_n)$ in place of $(y_1, x_1), \ldots, (y_n, x_n)$ because $y_1, \ldots, y_n$ are unavailable; the quantity $\hat{\sigma}_{V|X}$ is used in place of $\hat{\sigma}$ (i.e, $\hat{\sigma}_{Y|X}$) in (3.6.2)–(3.6.7). The results are valid if the measurement errors satisfy the assumptions in Box 3.10.1. However, $\sigma_{V|X} \geq \sigma_{Y|X}$, and so these point estimates tend to have larger standard errors, and the confidence intervals tend to be wider than when there are no errors in observing Y.

On the other hand, there is no valid point estimate or confidence interval for $\sigma_{Y|X}$ or for $\rho_{Y,X}$ in (3.10.5) unless additional information is available about the value of σ_d, the standard deviation of the errors in measuring Y.

When measurement errors are present in the predictor variable X, the procedures are somewhat more complicated than when there are measurement errors only in the response variable Y. When there are measurement errors in the predictor variable X, we consider two distinct models: (1) the Berkson model, named after Joseph Berkson, who first discussed it in detail, and (2) the classical errors in variables model. These are discussed next.

Berkson Model and Classical Errors in Variables Model

Suppose $\{(Y, X)\}$ is the population under study and assumptions (A) (respectively, assumptions (B)) are satisfied except that measurement errors are present in the predictor variable X. Measurement errors *may* also be present in the response variable Y. Then the *Berkson model* applies when data are obtained by preselecting the *measured* values of X, (i.e., by sampling with preselected U values), for example when the recorded values $u_1, \ldots, u_n$ of the predictor variable X are values on a dial or a gauge, etc., that are set at preselected levels. *Although the dial or gauge settings are the values that are recorded for X, the true X values may be different because the gauge or dial may be in error.* On the other hand, the *classical errors in variables model* applies if data are obtained by simple random sampling. In this case it is impossible to preselect the observed (recorded) value for the predictor variable X because the items are randomly selected. We illustrate these two situations with examples.

EXAMPLE 3.10.1 Berkson Model

Consider the problem where an engineer is interested in studying how the temperature X used in manufacturing aluminum cans is related to the strength Y of the cans. Data are collected by making aluminum cans at various temperatures and measuring their strength. It is known that there are nonnegligible errors in measuring X. The process is run by setting the temperature gauge at a preselected value, say 300°C, and measuring Y, the strength of the can; the temperature gauge is then set at another value, say 325°C, and the Y value is measured; etc. Even though the temperature gauge is set at (say) 300°C, the actual temperature at which the process runs may not be exactly 300°C; i.e., even though 300°C is recorded by observing the gauge, this may not be the true temperature at which the process was run because the gauge may be in error. The true temperature may be 298°C or 304°C, etc. In this problem there may also be errors in measuring the response variable Y. ■

EXAMPLE 3.10.2 Classical Model

Suppose a biologist is interested in the relationship between the number of damaged blood cells Y and blood sugar levels X in female rats. A random sample of rats is obtained for the study, and a small vial of blood is drawn from each rat. The vials of blood are sent to a laboratory for analysis. Both the counts of damaged cells and the blood sugar levels reported by the laboratory are likely to be different from the unknown true values because of measurement errors. For instance, the laboratory may have recorded the number of damaged cells by examining only a small drop of blood (instead of the entire vial of blood) under a microscope. Similarly, the blood sugar value determined by the laboratory may also be in error. Thus the measured values of X and Y, which we denote by $u_1, \ldots, u_n$ and $v_1, \ldots, v_n$, respectively, are not the true values, and the measurement errors may be nonnegligible. ■

Consequences of Measurement Errors: Berkson Model

Suppose $\{(Y, X)\}$ is the population under study and assumptions (A) are satisfied, except the values of the predictor variable X, and perhaps the values of the response variable Y, are observed or measured with errors. Further suppose that the data are obtained by preselecting the *measured* values of X, i.e., preselecting the values of $u_1, \ldots, u_n$, using dials or gauges that are subject to measurement errors. If the measurement errors satisfy the assumptions in Box 3.10.1, then it can be shown that the quantities β_0^* and β_1^* in (3.10.6) and in (3.10.7) are the same as β_0 and β_1, respectively, in (3.10.3). *So valid estimates for* β_0, β_1, $\mu_Y(x)$, $Y(x)$, *and* $a_0\beta_0 + a_1\beta_1$ *in (3.10.4) can be obtained by the formulas in Box 3.4.2 using the measured data values* $(v_1, u_1), \ldots, (v_n, u_n)$ *in place of the true values* $(y_1, x_1), \ldots, (y_n, x_n)$. Also, confidence intervals for all the quantities in (3.10.4), except $Y(x)$, can be obtained using the formulas (3.6.1), (3.6.2), (3.6.3), (3.6.4), and (3.6.6) with $\hat{\sigma}_{V|U}$ replacing $\hat{\sigma} = \hat{\sigma}_{Y|X}$.

On the other hand there are no valid prediction intervals for $Y(x)$ *and there are no valid point estimates or valid confidence intervals for* $\sigma_{Y|X}$ *or* $\rho_{Y,X}$ *unless additional information is available about the standard deviations* σ_d *and* σ_e *of the measurement errors.*

Consequences of Measurement Errors: Classical Model

Suppose the study population is $\{(Y, X)\}$ and assumptions (B) hold, except that the X values, and perhaps the Y values, are observed with errors due to imprecise measurement procedures. We suppose that the measurement errors satisfy the assumptions given in Box 3.10.1. Thus data are obtained by simple random sampling; i.e., a simple random sample of n items is chosen from the population and the Y values and the X values are measured for each item. Because the items in the sample are randomly sampled from the population, the X values *cannot be preselected.* The measured Y and X values for the sample items are denoted by $(v_1, u_1), \ldots, (v_n, u_n)$ as usual.

Since population assumptions (B) hold for the study population $\{(Y, X)\}$, the regression function of Y on X is a straight line function. We have denoted this regression function by

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

(see (3.10.3)). However, under the classical errors in variables model, unlike the Berkson model, the quantities β_0^* and β_1^* in (3.10.6) and (3.10.7) are in general different from the quantities β_0 and β_1 in (3.10.3). Therefore, no valid point estimates or confidence interval estimates are available for β_0 or β_1 unless we have information about the size of the measurement errors for X (such as a knowledge of σ_e). Moreover, because the X value of a randomly chosen item cannot be observed (due to errors in measurement), it does not make sense to attempt to use the true X value of the chosen item to predict its Y value. Fortunately, it is possible to obtain valid predictions of the Y value of a randomly chosen item using U, the measured value of X for that item. This is discussed next.

The usual formulas for straight line regression given in Box 3.4.2 are used to compute valid point estimates for β_0^*, β_1^*, $\mu_V(u) = \beta_0^* + \beta_1^* u$, and $a_0\beta_0^* + a_1\beta_1^*$, whereas formulas (3.6.1)–(3.6.4) and (3.6.6) are used for confidence intervals with v_i and u_i replacing y_i and x_i, respectively (i.e., regress V on U), and with $\hat{\sigma}_{V|U}$ taking the place of $\hat{\sigma} = \hat{\sigma}_{Y|X}$. The quantity $\hat{\mu}_V(u)$ is also the point estimate for $Y(u)$. Because $\mu_V(u) = \mu_Y(u)$, the computed confidence interval for $\mu_V(u)$ is a valid confidence interval for $\mu_Y(u)$. However, no valid estimate of $\sigma_{Y|X}$ or $\rho_{Y,X}$ is available, and no valid confidence interval is available for $Y(u)$ without additional information regarding the measurement error standard deviations σ_d and σ_e.

Summary

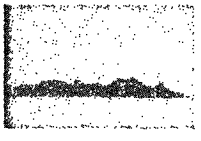
For the classical model where assumptions (B) apply, except that there are measurement errors in the predictor variable X (and possibly in the response variable Y), and where the measurement errors satisfy the assumptions in Box 3.10.1, there are *no* valid estimates or confidence intervals for the regression coefficients β_0 and β_1 in the population regression function of Y on X, given in (3.10.3), without additional information regarding the measurement error standard deviations σ_d and σ_e. However, if the observed sample data v_i, u_i are used in place of the true but unobservable y_i, x_i, respectively, in the formulas in

Box 3.4.2, valid point estimates are obtained for β_0^* and β_1^* (see (3.10.6) and (3.10.7)) and $a_0\beta_0^* + a_1\beta_1^*$, respectively. Valid confidence intervals for β_1^*, β_0^*, and $a_0\beta_0^* + a_1\beta_1^*$ are obtained using (3.6.1) with (3.6.2), (3.6.3), and (3.6.6), respectively, and replacing $\hat{\sigma}_{Y|X}$ in these formulas by $\hat{\sigma}_{V|U}$. Using v_i and u_i, we can also obtain a point estimate and a confidence interval for $\mu_V(u)$, and *these give a valid point estimate and a valid confidence interval for* $\mu_Y(u)$. Also we can compute a point estimate and a confidence interval (prediction interval) for $V(u)$. The point estimate of $V(u)$ is also a valid point estimate of $Y(u)$, but the prediction interval for $V(u)$ is not a valid prediction interval for $Y(u)$. Furthermore, no valid point estimates or confidence intervals exist for

$$\mu_Y(x) = \beta_0 + \beta_1 x, \qquad \rho_{Y,X}, \qquad \sigma_{Y|X}$$

unless information is available about the values of σ_d and σ_e, the standard deviations of the measurement errors associated with Y and X, respectively.

In the following two tasks, we discuss several typical problems encountered when measurement error is present.



Task 3.10.1

In this task we discuss the Berkson Model using the setup of Example 3.10.1. We suppose that an engineer preselected 5 different sets of *measured* values for temperature and ran the process by setting the temperature gauge at each of these preselected values. The *measured values* of the strength of the aluminum cans, along with the *measured values* of the predictor variable, are displayed in Table 3.10.1 and are also stored in the file **cans.dat** on the data disk.

Assumptions (A) are presumed to be valid except that Y and X cannot be observed due to measurement errors. We suppose that these measurement errors satisfy the assumptions in Box 3.10.1. Thus the regression function of the true values of the response variable Y on the true values of the predictor variable X is of the form

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{3.10.8}$$

Since the U values were obtained by setting the temperature gauge at several preselected values, Berkson's model is applicable, and valid point estimates and confidence intervals for β_0, β_1, and $\mu_Y(x)$ can be calculated by using the measured data values $(v_1, u_1), \ldots, (v_{15}, u_{15})$. We simply use the formulas in Box 3.4.2 for point estimates and the formulas (3.6.1), (3.6.2), (3.6.3), and (3.6.4) for confidence intervals with v_i and u_i replacing y_i and x_i, respectively, in the calculations, and $\hat{\sigma}_{V|U}$ taking the place of $\hat{\sigma}_{Y|X}$.

TABLE 3.10.1
Aluminum Cans Data

Run Number	Observed Strength (newtons) V	Observed Temperature (°C) U
1	18.6	300.0
2	26.3	300.0
3	31.5	300.0
4	20.0	400.0
5	29.2	400.0
6	32.9	400.0
7	29.2	500.0
8	32.5	500.0
9	41.9	500.0
10	31.5	600.0
11	37.6	600.0
12	41.1	600.0
13	34.7	700.0
14	43.2	700.0
15	44.5	700.0

The following quantities can be calculated easily.

$$\bar{v} = 32.9800 \qquad \bar{u} = 500 \qquad SSV = 860.044$$

$$SSU = 300{,}000 \qquad SUV = 12{,}010.0$$

$$SSE(U) = 379.244 \qquad MSE(U) = 29.1726 \qquad \hat{\sigma}_{V|U} = 5.40117$$

$$\hat{\beta}_0 = 12.963 \qquad \hat{\beta}_1 = 0.040033 \qquad SE(\hat{\beta}_0) = 5.124 \qquad SE(\hat{\beta}_1) = 0.009861$$

Now consider the following questions.

1 What is the estimate of the true average strength of aluminum cans that were manufactured with a true temperature of 350° C?

The true average strength of aluminum cans that were manufactured with a true temperature of 350°C is $\mu_Y(350) = \beta_0 + 350\beta_1$. *This is estimated by* $\hat{\beta}_0 + 350\hat{\beta}_1 = 26.975$ *newtons.*

2 What is the estimate of the *true average* strength of aluminum cans that were manufactured with the temperature dial set at 350° C?

The answer again is $\hat{\mu}_Y(350) = 26.975$ *newtons, because* $\hat{\mu}_V(350) = \hat{\beta}_0 + \hat{\beta}_1(350) = 12.963 + 0.040033(350) = 26.975$ *newtons and* $\hat{\mu}_V(u) = \hat{\mu}_Y(u)$ *for any allowable value of u, the temperature dial setting.*

3 What is a 95% two-sided confidence interval for $\mu_Y(u)$, where $u = 350°$ C is the temperature dial setting during the manufacturing process?

We first compute a 95% two-sided confidence interval for $\mu_V(350)$. *We get*

$$C[22.58 \leq \mu_V(350) \leq 31.37] = 0.95$$

Because $\mu_V(u) = \mu_Y(u)$, *a 95% two-sided confidence interval for* $\mu_Y(350)$ *is given by*

$$C[22.58 \leq \mu_Y(350) \leq 31.37] = 0.95$$

4 What is a 95% two-sided confidence interval for $\mu_Y(x)$, where $x = 350°$ C is the true temperature during the manufacturing process?

The quantity $\mu_Y(x)$ *for* $x = 350$ *is* $\beta_0 + 350\beta_1$ *and is identical to the quantity* $\mu_V(u)$ *with* $u = 350$. *A 95% two-sided confidence interval was computed for* $\mu_V(u)$ *with* $u = 350$ *in question (3). So the required confidence interval for* $\mu_Y(x)$ *with* $x = 350$ *is given by*

$$C[22.58 \leq \mu_Y(350) \leq 31.37] = 0.95$$

5 Compute a point estimate and 95% confidence interval for $\sigma_{Y|X}$.

The answer to this question cannot be obtained without some information about the values of σ_d *and* σ_e, *the standard deviations of the errors in measuring Y and X, respectively.*

In the following task we illustrate the procedures for the classical errors in variables model.

Task 3.10.2

A researcher in the exercise science department of a university conducted a study to evaluate the relationship between dietary fat and body fat of competitive runners who ran at least 12 hours per week. A random sample of 18 such runners was obtained, and their body fat Y (in percent) and dietary fat X (in percent), were measured. It is known that there are measurement errors in both Y and X, so the measured dietary fat (the measured value of X) is denoted by U, and the measured body fat (the measured value of Y) by V.

The classical errors in variables model is presumed to apply for this problem. Thus the regression function of Y on X is of the form

$$\mu_Y(x) = \beta_0 + \beta_1 x \qquad \textbf{(3.10.9)}$$

But since the true X values are not observable, we cannot obtain valid estimates of β_0 and β_1 without additional information concerning the measurement error standard deviations σ_d and σ_e. However, it is possible to obtain valid predictions of the true

values Y using the measured values U. In fact, the regression function of Y on U is given by

$$\mu_Y(u) = \beta_0^* + \beta_1^* u \tag{3.10.10}$$

where β_0^* and β_1^* are also the regression coefficients in the regression function of V on U; i.e.,

$$\mu_V(u) = \mu_Y(u) = \beta_0^* + \beta_1^* u \tag{3.10.11}$$

Note that β_0^* and β_1^* in (3.10.11) are not the same as the corresponding quantities β_0, β_1 in (3.10.9).

The researcher wants to estimate the regression function $\mu_Y(u)$ for predicting an individual's true body fat Y using u, the measured value of dietary fat. The data are in Table 3.10.2 and also in the file **fat.dat** on the data disk.

The following key quantities can be calculated.

$$\bar{v} = 10.3167 \qquad \bar{u} = 25.8333 \qquad SSV = 36.2650$$

$$SSU = 1010.50 \qquad SUV = 117.450$$

$$SSE(U) = 22.6138 \qquad MSE(U) = 1.4134 \qquad \hat{\sigma}_{V|U} = 1.1889$$

$$\hat{\beta}_1^* = 0.116230 \qquad SE(\hat{\beta}_1^*) = 0.03740 \qquad \hat{\beta}_0^* = 7.31407 \qquad SE(\hat{\beta}_0^*) = 1.006$$

TABLE 3.10.2
Fat Data

Runner	Measured Body Fat V (%)	Measured Dietary Fat U (%)
1	9.8	22
2	11.7	22
3	8.0	14
4	9.7	21
5	10.9	32
6	7.8	26
7	9.7	30
8	11.6	21
9	8.6	17
10	11.2	35
11	12.3	35
12	10.2	24
13	12.0	24
14	11.6	36
15	10.4	20
16	10.8	37
17	11.5	35
18	7.9	14

A careful study of the following questions and our answers to them will help you understand the concepts related to the classical errors in variables model.

1 What is the estimate of the true average body fat of runners with a true dietary fat of 25%; i.e., what is the estimate of $\mu_Y(x)$ for $x = 25$?

The required true average body fat is $\mu_Y(x)$ with $x = 25$; i.e., $\beta_0 + 25\beta_1$. But we do not have valid estimates for β_0 and β_1 without additional information about the value of σ_e, the standard deviation of the errors in measuring X. So we cannot answer this question with available information.

2 What is the estimate of $\mu_Y(x)$, the true average body fat of runners who have a measured dietary fat of $u = 25$?

The answer is $\hat{\mu}_Y(u) = \hat{\mu}_V(u) = \hat{\beta}_0^* + \hat{\beta}_1^* u$ with $u = 25$, which gives $7.31407 + 0.11623(25) = 10.22$ for the required estimate.

3 What is a 95% two-sided confidence interval for $\mu_Y(u)$, where $u = 25$ is the measured dietary fat?

We first compute a 95% two-sided confidence interval for $\mu_V(u)$ by using formulas (3.6.1) and (3.6.4) with $\hat{\sigma}_{V|U} = 1.1889$ substituted for $\hat{\sigma} = \hat{\sigma}_{Y|X}$. We get

$$C[9.622 \leq \mu_V(25) \leq 10.818] = 0.95$$

But since $\mu_Y(u) = \mu_V(u)$, the 95% two-sided confidence interval for $\mu_Y(u)$ with $u = 25$ is also [9.622, 10.818]. ***Note that for this to be a valid confidence interval, the body fat $u = 25$ must be measured by the same procedure that was used to obtain the sample values u_i.***

4 What is a 95% two-sided confidence interval for $\mu_Y(x)$, where $x = 25$ is the true dietary fat of runners?

The answer to this question cannot be obtained without some information about the value of σ_e, the standard deviation of the errors in measuring X.

5 If an individual's dietary fat is measured to be $u = 25$, estimate this individual's true body fat; i.e., obtain $\hat{Y}(u)$.

The formula in (3.4.11) can be used to obtain $\hat{Y}(u)$, the point estimate for the true body fat, using the measured dietary fat $u = 25$. We get

$$\hat{Y}(u) = \hat{\mu}_Y(u) = \hat{\beta}_0^* + 25\hat{\beta}_1^* = 7.31407 + 0.11623(25) = 10.22$$

6 In question (5), compute a 95% two-sided confidence interval for the true value of this individual's body fat if the measured value of dietary fat is $u = 25$; i.e., compute a 95% two-sided confidence interval for $Y(u)$ with $u = 25$.

A valid 95% confidence interval for $Y(u)$ for any specified value of u cannot be determined based on available information.

7 Compute a point estimate and 95% confidence interval for $\sigma_{Y|X}$ and for $\rho_{Y,X}$.

The required point estimates and confidence intervals cannot be obtained without additional information about the values of σ_d and σ_e, the standard deviations of the errors in measuring Y and X, respectively.

Note A valid point estimate and confidence interval for $\sigma_{V|U}$ can be obtained using the usual formulas in (3.4.20) and (3.6.8) with y_i, x_i replaced by v_i, u_i. Likewise, a valid point estimate and confidence interval for $\rho_{V,U}$ can be obtained by using the formula in (3.9.8) and the procedure given in Box 3.9.2, respectively, with y_i, x_i replaced by v_i, u_i.

Conversation 3.10

Investigator: Good afternoon. I have some questions about regression when there are measurement errors present. Is this a good time to discuss them with you?

Statistician: Certainly.

Investigator: In Example 3.10.1, suppose the gauge is extremely accurate so there is no error in measuring X, but there is error in measuring Y, the strength of the aluminum cans. (I am presuming that assumptions (A) are satisfied, except that Y cannot be observed due to errors in measurement.) The observed value of Y is denoted by V. Can I ignore the fact that there are measurement errors and proceed as if there are none?

Statistician: You can regress V on X and use all the formulas in Sections 3.4 and 3.6 for point estimates and confidence intervals, *except* there is no valid estimate for $\sigma_{Y|X}$ (and $\rho_{Y,X}$).

Investigator: Does that matter?

Statistician: Only if you need to know $\sigma_{Y|X}$ to make your decisions.

Investigator: Because $\hat{\sigma}_{Y|X}$ (i.e., $\hat{\sigma}$) appears in the formulas for $SE(\hat{\beta}_i)$, and hence in confidence intervals for β_i, don't I need to compute it?

Statistician: For the standard errors and confidence intervals you refer to, the quantity $\hat{\sigma}_{Y|X}$ (i.e., $\hat{\sigma}$) is replaced with $\hat{\sigma}_{V|X}$, and this can be computed. In fact, this is equal to $\sqrt{MSE(X)}$, and it is computed, as usual, by regressing V on X.

Investigator: If assumptions (B) are satisfied except there are errors in measuring X (and perhaps in measuring Y), you state that one cannot estimate the β_i in (3.10.3) and hence one cannot use the estimate of

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

to predict Y using the true value x of X. It appears to me that even if the β_i are known exactly, one cannot use

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{3.10.12}$$

in (3.10.3) to predict Y because the true value of X cannot be observed so I have no X value to use in (3.10.12). So why are we ever interested in it?

Statistician: You are correct in noting that (3.10.12) cannot be used to predict Y when the true x is unavailable. However, if β_0 and β_1 are known, then you can use the quantity $\beta_0 + \beta_1 u$ to predict Y where u is the measured value corresponding to x. So if β_0 and β_1 are known, they can certainly be useful even if the true x is not known. Furthermore, the investigator might want to examine (3.10.12) to see how the average value of Y is affected by the values of X. To illustrate, consider Example 3.10.2. An investigator might want to know how much the true average number of damaged blood cells $\mu_Y(x)$ changes when the true amount of blood sugar X changes by one unit. The answer is β_1 in (3.10.12). Additionally, the investigator may want to use the estimate of β_0 or β_1 in another formula in another context. For these reasons it is useful to know the values of β_0 and β_1, but no *valid* estimate of β_0 or β_1 is available for this problem without some additional knowledge about the measurement error standard deviation σ_e.

In practice we can *actually predict* the true average value of Y by using U, the measured value of the predictor, and this can be done as usual by regressing V on U (i.e., use the formulas in Sections 3.4 and 3.6) to obtain point and confidence interval estimates for the β_i and for $\mu_V(u)$. Then use the fact that $\mu_V(u) = \mu_Y(u)$ to obtain a point estimate and confidence interval for $\mu_Y(u)$.

Investigator: I see what you're saying. In general we *can* obtain valid estimates for many of the quantities we want even if there are measurement errors in Y and X.

Statistician: That is correct, but we have to understand the underlying assumptions, limitations, and interpretations.

Investigator: You've said that in some problems the measurement errors are small enough to be considered negligible. How do I determine whether *measurement errors are small enough to be considered negligible* in a given problem?

Statistician: It isn't possible to give a simple answer that will be appropriate for every problem. What you need to do is first specify the objectives of the study and the quantities needed for making decisions, and then examine how the estimates of these quantities will change as a result of errors in measurement. If these changes are small enough so that the decisions aren't affected, then the measurement errors can be considered negligible for that problem. These calculations are best done with the help of a professional statistician.

Investigator: You have stated that some answers aren't available unless σ_e, the standard deviation of the errors in measuring X, is available. Will I ever know it?

Statistician: Probably not, but sometimes an estimate of σ_e is available if the study is planned carefully. This subject is discussed in more advanced books. See reference [7].

Investigator: I see. Are there any other important points I should know in connection with measurement errors?

Statistician: There are many other important considerations you should be aware of in connection with measurement errors, but we will have to discuss them some other time.

However, I'd like to bring to your attention the fact that, in both the Berkson model and the classical model, $\sigma_{Y|X} \leq \sigma_{V|U}$. You can use the measured data to compute a $1-\alpha$ upper confidence bound for $\sigma_{V|U}$ and get the confidence statement

$$C[\sigma_{V|U} \leq \sqrt{SSE(U)/\chi^2_{\alpha:n-2}}] = 1-\alpha$$

Then it follows that

$$C[\sigma_{Y|X} \leq \sqrt{SSE(U)/\chi^2_{\alpha:n-2}}] \geq 1-\alpha$$

is a valid confidence statement. This confidence statement about $\sigma_{Y|X}$ can be useful in practical applications.

Investigator: I think I understand. Thank you. Perhaps I'll come again when I have more questions.

Problems 3.10

3.10.1 A researcher wants to evaluate how different soil temperatures X affect the rate of growth Y of a particular variety of cabbage plants. He conducts an experiment in a greenhouse using seven different soil temperatures (in degrees Fahrenheit). The soil temperatures are controlled by thermostats, but the actual soil temperatures are not necessarily the same as the temperatures at which the thermostats are set. The temperature setting of the thermostats is denoted by U, whereas the true soil temperature is denoted by X. The measurement of growth rate (rate of change of biomass in grams per week) is also subject to errors, and the measured growth rate is denoted by V, whereas the true growth rate is denoted by Y. Due to nonnegligible errors in measurement, only U and V are observable. The data from the experiment are given in Table 3.10.3 and are also stored in the file cabbage.dat on the data disk.

It is presumed that (population) assumptions (A) hold for the two-variable population $\{(Y,X)\}$ for $60 \leq X \leq 90$. In particular, the regression function of Y on X is of the form

$$\mu_Y(x) = \beta_0 + \beta_1 x \quad \text{for} \quad 60 \leq x \leq 90 \tag{3.10.13}$$

We further suppose that the measurement errors satisfy the assumptions given in Box 3.10.1. The results from a regression analysis of V on U are as follows:

$$\bar{v} = 46.7500 \qquad \bar{u} = 75.0000 \qquad SSV = 399.235$$

$$SSU = 1400.00 \qquad SUV = 578.996$$

$$SSE(U) = 159.78 \qquad MSE(U) = 13.315 \qquad \hat{\sigma}_{V|U} = 3.6489$$

$$\hat{\beta}_1^* = 0.4136 \qquad SE(\hat{\beta}_1^*) = 0.09752 \qquad \hat{\beta}_0^* = 15.732 \qquad SE(\hat{\beta}_0^*) = 7.379$$

TABLE 3.10.3
Cabbage Data

Plant Number	Growth Rate V (grams/week)	Temperature U (°F)
1	35.1	60
2	43.6	60
3	45.3	65
4	37.3	65
5	44.5	70
6	49.5	70
7	46.8	75
8	47.7	75
9	50.2	80
10	51.5	80
11	45.8	85
12	53.8	85
13	54.4	90
14	49.0	90

a Using the sample data in Table 3.10.3 estimate the parameters β_0 and β_1. Obtain a two-sided 85% confidence interval for the average growth rate of this cabbage plant when the thermostat is set at 72° F. Use $t_{0.925;12} = 1.538$.

b What is the change in the average growth rate of this particular variety of cabbage plants if the thermostat setting for soil temperature is increased by 5° F? Obtain a two-sided 80% confidence interval for this quantity.

c If an increase of 5° F in the thermostat setting for soil temperature fails to result in an increase in average growth rate of at least 5 grams per week, then the researcher will conclude that soil temperature is not an important factor relative to his objectives. What will his decision be, based on the confidence interval in part (b)?

d Suppose the researcher conducts a statistical test to reach a decision in part (c). State the appropriate null and alternative hypotheses for this purpose. Carry out the test and report the P-value. What should his decision be if he uses $\alpha = 0.10$?

e Predict the growth rate of a cabbage plant if the soil temperature is set at 75° F. Also, if possible, obtain a two-sided 95% confidence interval for the growth rate of this cabbage plant.

f Obtain a valid estimate, if possible, for the quantity $\sigma_{Y|X}$.

g Obtain a valid estimate, if possible, for the quantity $\sigma_{V|U}$.

3.10.2 Suppose a researcher wants to predict the first-year maintenance cost Y for minivans purchased next year, using the number of miles X the minivan will be driven during its first year after purchase. The target population is clearly a future population and is unavailable for sampling. Instead, the researcher uses a similar population of minivans from last year as the study population. She selects a simple random sample of

15 minivans from the study population and records the first-year maintenance costs and the miles driven. However, because of inaccurate odometers, the recorded values of miles driven are not the true values but are subject to nonnegligible measurement errors. The maintenance costs may also contain nonnegligible errors becuse of inaccurate record keeping of the minivan owners. Let Y and X represent the true maintenance cost and true miles driven (which are unavailable), respectively, and let V and U represent the corresponding observed, i.e., recorded values. The data are given in Table 3.10.4 and are also stored in the file **minivan.dat** on the data disk.

It is presumed that (population) assumptions (B) hold for the two-variable population $\{(Y, X)\}$. In particular the regression function of Y on X is of the form

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{3.10.14}$$

We further suppose that the measurement errors satisfy the assumptions given in Box 3.10.1 so that the regression function of V on U is of the form

$$\mu_V(u) = \beta_0^* + \beta_1^* u \tag{3.10.15}$$

and the regression function of Y on U is also of the form

$$\mu_Y(u) = \beta_0^* + \beta_1^* u \tag{3.10.16}$$

where the parameters β_0^* and β_1^* in (3.10.15) are the same as those in (3.10.16).

a What is the estimated regression function of Y on U?

b What is the average true first-year maintenance cost for all minivans that will be driven *exactly* 12,500 miles during the first year (give the appropriate population

TABLE 3.10.4
Minivan Data

Van	Maintenance Costs V (dollars)	Miles Driven U
1	652	16500
2	422	8000
3	724	14200
4	746	18400
5	571	9300
6	644	13900
7	548	11000
8	553	13400
9	792	17200
10	739	16500
11	742	18400
12	763	18700
13	698	17700
14	568	10100
15	663	16300

quantity)? Obtain a valid point estimate of this quantity if possible. If no valid point estimate exists, state why.

c What is the average true first-year maintenance cost of all minivans whose odometer readings indicate that they were driven 12,500 miles during the first year after purchase (give the population quantity)? Obtain a valid point estimate of this quantity if possible. If not, state the reason why it is not possible.

d Is a valid two-sided 90% confidence interval available for the quantity in part (b)? If so, compute it. If not, explain why not.

e Is a valid two-sided 95% confidence interval available for the quantity in part (c)? If so, compute it. If not, explain why not.

3.11 Regression Through the Origin

Recall that when assumptions (A) or (B) for straight line regression are valid, the regression function of Y on X is of the form

$$\mu_Y(x) = \beta_0 + \beta_1 x$$

In some situations the investigator may know, based on subject matter considerations, that the intercept β_0 must equal zero. In such cases the population regression function reduces to

$$\mu_Y(x) = \beta_1 x \tag{3.11.1}$$

and the graph of this function is a straight line that passes through the origin. In this situation, the formulas for computing $\hat{\beta}_1$, $\hat{\mu}_Y(x)$ $\hat{Y}(x)$, $\hat{\sigma}$, $SE(\hat{\beta}_1)$, $SE(\hat{\mu}_Y(x))$, and $SE(\hat{Y}(x))$ must be modified. The appropriate formulas are as follows:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2} \tag{3.11.2}$$

$$\hat{\mu}_Y(x) = \hat{\beta}_1 x \tag{3.11.3}$$

$$\hat{Y}(x) = \hat{\beta}_1 x \tag{3.11.4}$$

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\left[\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n x_i y_i)^2}{\sum_{i=1}^n x_i^2}\right]} \tag{3.11.5}$$

$$SE(\hat{\beta}_1) = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^n x_i^2}} \tag{3.11.6}$$

$$SE(\hat{\mu}_Y(x)) = \frac{|x|\hat{\sigma}}{\sqrt{\sum_{i=1}^n x_i^2}} \tag{3.11.7}$$

$$SE(\hat{Y}(x)) = \hat{\sigma}\sqrt{1 + \frac{x^2}{\sum_{i=1}^n x_i^2}} \tag{3.11.8}$$

$$= \sqrt{\hat{\sigma}^2 + [SE(\hat{\mu}_Y(x))]^2} \tag{3.11.9}$$

The quantity $SSE(X)$ is now given by

$$SSE(X) = \sum_{i=1}^{n}(y_i - \hat{\beta}_1 x_i)^2 \tag{3.11.10}$$

which on simplification reduces to

$$SSE(X) = \sum_{i=1}^{n} y_i^2 - \frac{(\sum_{i=1}^{n} x_i y_i)^2}{\sum_{i=1}^{n} x_i^2} \tag{3.11.11}$$

The mean square error $MSE(X)$ is now given by

$$MSE(X) = \frac{SSE(X)}{n-1} \tag{3.11.12}$$

which is the estimate of σ^2.

Note that the quantity $SSE(X)$ is divided by $n-1$ rather than by $n-2$ as in (3.4.21) to obtain $MSE(X)$ because the number of degrees of freedom associated with $SSE(X)$ is $n-1$ and not $n-2$. More generally, *the number of degrees of freedom associated with a sum of squares of errors is calculated by subtracting from the sample size n, the number of unknown parameters (the β's) in the population regression function.* Thus, when the population regression function is $\mu_Y(x) = \beta_0 + \beta_1 x$, the degrees of freedom associated with the sum of squares of errors

$$SSE(X) = \sum(y_i - \hat{\mu}_Y(x_i))^2 = \sum(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

is $n-2$, and when the population regression function is $\mu_Y(x) = \beta_0$ (in which case $\beta_0 = \mu_Y$), the degrees of freedom associated with the sum of squares of errors

$$SSY = \sum(y_i - \hat{\mu}_Y(x_i))^2 = \sum(y_i - \bar{y})^2$$

is $n-1$.

Formulas for confidence intervals for β_1, $\mu_Y(x)$, and $Y(x)$ are the same as in (3.6.1), with the understanding that the standard errors to be used are those given in (3.11.6), (3.11.7), and (3.11.8), respectively, and the degrees of freedom to be used for finding table-values is $n-1$. The corresponding procedures for statistical tests for β_1 and $\mu_Y(x)$ are in Boxes 3.7.2 and 3.7.3. The formula for confidence intervals for σ is in (3.6.8), where the degrees of freedom for finding table-values is $n-1$. The procedure for statistical tests for σ is given in Box 3.7.5, where again the degrees of freedom are $n-1$ instead of $n-2$.

The following example illustrates the computations for straight line regression through the origin.

EXAMPLE 3.11.1

It is well known (based on the laws of physics) that when an object is dropped from rest, it will fall because of gravity with continually increasing speed. The speed Y, at the end of X seconds after release, may vary from trial to trial because of air resistance and other random experimental errors. However, the average speed Y (in feet per second) of the object, after a specified amount of elapsed time X in seconds,

is given by the regression function

$$\mu_Y(x) = \beta_1 x$$

The quantity β_1 is a constant and is referred to as the acceleration due to gravity. If y_i and x_i are the observed distances and times, respectively, the sample regression model is

$$y_i = \beta_1 x_i + e_i$$

In an experiment conducted to estimate the value of β_1, an object is dropped from rest repeatedly, and each time its speed at the end of a preselected amount of time is recorded. The data appear in Table 3.11.1 and are also stored in the file **gravity.dat** on the data disk.

TABLE 3.11.1
Gravity Data

Trial Number	Y (ft/sec)	X (sec)
1	63	2
2	128	4
3	194	6
4	257	8
5	322	10
6	387	12
7	451	14

We compute the following quantities:

$$\sum y_i = 1,802 \qquad \sum x_i = 56$$

$$\sum x_i y_i = 18,036 \qquad \sum x_i^2 = 560 \qquad \sum y_i^2 = 580,892$$

$$\hat{\beta}_1 = 32.2071 \qquad \hat{\sigma} = 0.8136 \qquad SE(\hat{\beta}_1) = 0.0344$$

A 95% two-sided confidence interval for β_1 (using $t_{0.975:6} = 2.447$ from Table T-2 as the table-value) is given by the confidence statement

$$C[32.2071 - (2.447)(0.0344) \leq \beta_1 \leq 32.2071 + (2.447)(0.0344)] = 0.95$$

i.e.,

$$C[32.12 \leq \beta_1 \leq 32.29] = 0.95$$

A 90% two-sided confidence interval for σ (using $\chi^2_{0.05:6} = 1.635$ and $\chi^2_{0.95:6} = 12.592$ from Table T-3 as the table-values) is

$$\sqrt{\frac{6(0.8136)^2}{12.592}} \leq \sigma \leq \sqrt{\frac{6(0.8136)^2}{1.635}}$$

i.e.,

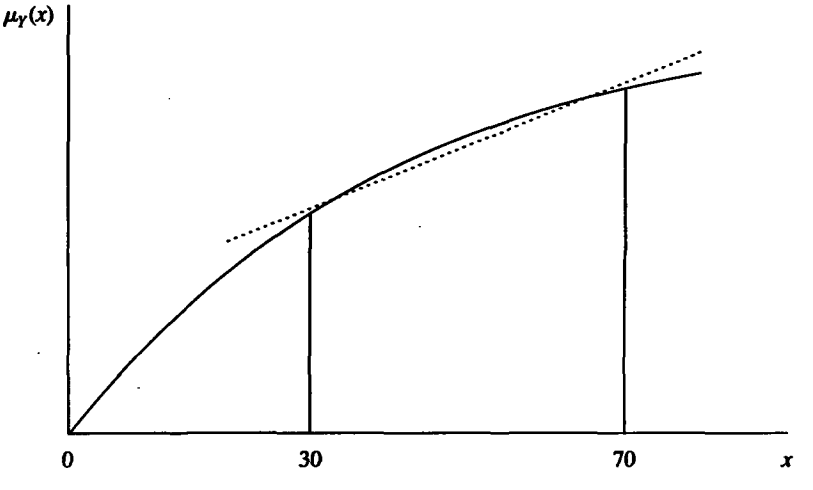
$$0.562 \leq \sigma \leq 1.559$$

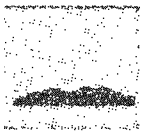
Suppose we want to test (using $\alpha = .01$) NH: $\beta_1 = 32$ against AH: $\beta_1 \neq 32$. We compute $t_C = \dfrac{\hat{\beta}_1 - 32}{SE(\hat{\beta}_1)} = 6.02$ and $t_T = t_{1-\alpha/2:6} = t_{0.995:6} = 3.707$, so $t_C > t_T$ and NH is rejected at $\alpha = 0.01$. ■

Caution

For some situations it may seem that Y must equal zero whenever X is zero and that a straight line regression function with no intercept term is the correct model (i.e., $\mu_Y(x) = \beta_1 x$). This is not always the case. For example, when studying the growth Y of a population of plants as a function of time X, it is true that at time $X = 0$ none of the plants have grown, so the Y value is zero for each plant. Consequently, $\mu_Y(0) = 0$, so the model $\mu_Y(x) = \beta_1 x$ may appear to be correct. If the study period is 30 days to 70 days from germination, the model with no intercept would be very inadequate because in these situations the growth model is often of the form given in Figure 3.11.1 (i.e., not a straight line but a nonlinear curve passing through the origin). In the range from $X = 30$ to $X = 70$, a model that is linear in X may fit quite well (dotted line in Figure 3.11.1), but if it is forced through the origin (i.e., if β_0 is required to be zero), the fit will be bad. So if we use a straight line model for this situation for $30 \leq x \leq 70$, we do not require β_0 to be zero even though we know that $\mu_Y(0) = 0$.

FIGURE 3.11.1





Problems 3.11

3.11.1 Consider the crystal data of Task 3.4.1. These data are given in Table 3.4.2 and are also stored in the file **crystal.dat** on the data disk. You should refer to Task 3.4.1 for a description of these data. Suppose that assumptions (A) hold with the regression function $\mu_Y(x)$ given by

$$\mu_Y(x) = \beta_1 x \qquad \text{for} \qquad 0 \le x \le 30 \tag{3.11.13}$$

i.e., a straight line through the origin.

- **a** Plot Y against X and visually evaluate whether or not a straight line through the origin appears to be a reasonable model for these data.
- **b** Obtain point estimates for β_1 and σ.
- **c** On the average, what is the increase in weight of crystals for each additional hour of growth? Give the population parameter that answers the question.
- **d** Compute a two-sided 80% confidence interval for the quantity of interest in part (c).
- **e** Compute a two-sided 90% confidence interval for σ.
- **f** The regression function $\mu_Y(x)$ in (3.11.13) will be considered adequate for predicting the weights of crystals provided the predicted weights will be within 3 grams of the true weights for at least a proportion $p = 0.9$ of the crystals in the population. Compute an appropriate 95% confidence interval to help the investigator determine whether or not $\mu_Y(x)$ in (3.11.13) is adequate for predicting values of Y.

3.11.2 Consider the arsenic data of Task 3.4.2. These data are given in Table 3.4.3 and are also stored in the file **arsenic.dat** on the data disk. Refer back to Task 3.4.2 for a description of these data. Suppose assumptions (A) hold with the regression function of Y on X given by

$$\mu_Y(x) = \beta_1 x \qquad \text{for} \qquad 0 \le x \le 7, \tag{3.11.14}$$

i.e., a straight line through the origin. The chemical analysis will be deemed to provide unbiased estimates of the true As concentrations if the slope β_1 is between 0.96 and 1.04. Compute an appropriate 99% confidence interval to help the investigator decide whether or not the chemical analysis provides unbiased estimates of As concentrations.

3.12 Exercises

3.12.1 Give two examples (preferably from your own field) in which you want to study the relationship between a response variable Y and one or more factor variables for the purpose of prediction or other reasons. Describe carefully the target population,

the study population, and convenient sampling methods for each example. If possible describe the form of $\mu_Y(x)$.

3.12.2 The SAT scores X and the the corresponding GPAs (Y) at the end of the first term for a simple random sample of ten students from a major university are as follows:

SAT Scores	321	358	640	270	443	669	582	451	791	594
GPA	1.97	2.19	2.98	1.89	2.66	3.14	2.20	2.02	3.07	3.11

a If assumptions (B) are valid, for which of the following parameters can a valid estimate be obtained?

$$\beta_0, \quad \beta_1, \quad \sigma, \quad \mu_Y(x), \quad \mu_Y, \quad \mu_X, \quad \sigma_X, \quad \sigma_Y, \quad \rho_{Y,X}$$

b Find point estimates of all parameters in part (a) for which valid estimates are available. Show all your calculations.

c Find 90% two-sided confidence intervals for all parameters in part (a), except $\rho_{Y,X}$, for which valid confidence intervals are available. Compute a 95% two-sided confidence interval for $\rho_{Y,X}$.

d Plot y_i versus x_i. From the plot do you have reason to believe that the regression function of Y on X is a straight line function?

e Estimate how much better $\mu_Y(x)$ is than μ_Y for predicting GPA.

f Using population parameters, write an expression for the difference between the average GPA of all students whose SAT score is 550 and the average GPA of all students whose SAT score is 500.

g Obtain a point estimate and a 95% two-sided confidence interval for the quantity in part (f).

h Repeat parts (f) and (g) for SAT scores of 700 and 600 (instead of 550 and 500), respectively.

3.12.3 A simple random sample of twenty U.S. males was selected, and the following information was recorded for each individual.

X = Number of grams of fat consumed per day. (This is usually calculated from a detailed record of food intake over a period of several days and then averaged to obtain fat intake per day).
Y = Total cholesterol in blood in milligrams per deciliter. (This is obtained from a blood test).

Suppose assumptions (B) for straight line regression are satisfied so the model is $\mu_Y(x) = \beta_0 + \beta_1 x$. The data are given in Table 3.12.1 and are also stored in the file **chol.dat** on the data disk.

TABLE 3.12.1
Cholesterol Data

Sample Item Number	Total Cholesterol Y, (in mg/dl)	Daily Fat Intake X, (in g)
1	130	21
2	163	29
3	169	43
4	136	52
5	187	56
6	193	64
7	170	77
8	115	81
9	196	84
10	237	93
11	214	98
12	239	101
13	258	107
14	283	109
15	242	113
16	289	120
17	298	127
18	271	134
19	297	148
20	316	157

Some basic numerical summaries that you will need are: $\sum x_i = 1{,}814$

$\sum y_i = 4{,}403 \quad SSX = 27{,}674.2 \quad SSY = 72{,}098.6 \quad SXY = 39{,}495.9$

a Plot y_i against x_i.

b Estimate the regression function of total cholesterol on daily fat intake.

c Plot this regression line on the same graph as in part (a).

d Compute the standard errors for the intercept and slope estimates (i.e., compute $SE(\hat{\beta}_0)$ and $SE(\hat{\beta}_1)$).

e Estimate σ.

f Construct a 95% upper confidence bound for the population intercept β_0.

g Construct a 99% two-sided confidence interval for the population slope β_1. Does it appear that $\beta_1 = 0$? Why?

h *Construct a 95% two-sided confidence interval for σ.*

i Construct a 99.5% lower confidence bound for the population slope β_1. Interpret this confidence bound (see part (g)).

j Find the estimate of the average value of total cholesterol for those people whose daily fat intake is 50 grams (i.e., find $\hat{\mu}_Y(50)$). Find the estimate of the cholesterol of an individual chosen at random from this subpopulation (i.e., $\hat{Y}(50)$).

Construct a two-sided 95% confidence interval for $Y(50)$ and $\mu_Y(50)$. Obtain a 95% two-sided confidence interval for the cholesterol level of an individual whose daily fat intake is 60 grams.

k Construct a 97.5% upper confidence bound for the mean cholesterol level of the subpopulation for which the daily fat intake is 25 grams.

l Estimate the correlation coefficient between total cholesterol and fat intake (i.e., find $\hat{\rho}_{Y,X}$). Construct a 95% two-sided confidence interval for $\rho_{Y,X}$, the population correlation coefficient between Y and X. What does this say about the linear relationship between cholesterol and fat intake?

m Perform the following test of hypothesis:

$$\text{NH: } \beta_1 \leq 2 \text{ against AH: } \beta_1 > 2$$

Compute the P-value. What is your interpretation of the result of this test if you decide to reject NH when $P < 0.005$?

n Use the confidence interval in part (g) to decide between NH and AH in part (m). What is your interpretation?

3.12.4 This is a continuation of Exercise 3.12.3.

a Estimate σ_Y and σ.

b Exhibit an analysis of variance table.

c From the analysis of variance table, obtain the computed F statistic F_C for the test of NH: $\beta_1 = 0$ versus AH: $\beta_1 \neq 0$.

d What is the P-value for this test?

e Obtain a 95% confidence interval for σ_Y/σ.

f What do you conclude about the relationship between cholesterol and fat intake?

3.12.5 In Exercise 3.12.4 suppose that a company manufacturing margarine makes the following claim: The difference between the average blood cholesterol level for the subpopulation of individuals consuming 100 grams of fat per day and the average cholesterol level of the subpopulation of individuals consuming 50 grams of fat per day does not exceed 40 milligrams per deciliter (mg/dl). If the manufacturer's claim is true, then perhaps some people would be willing to include extra fat in their diets, thinking that the resulting increase in cholesterol is small enough so that there is no need for concern.

a State the appropriate null hypothesis and alternative hypothesis to test the manufacturer's claim.

b Use the data of Exercise 3.12.3 and calculate the P-value for this test.

c What do you conclude about the manufacturer's claim?

d Compute an appropriate confidence interval and make a decision about the manufacturer's claim.