6

Applications of Regression I

6.1 Overview

In this chapter we discuss several applications of linear regression. Section 6.2 deals with prediction intervals. Tolerance intervals are discussed in Section 6.3, and the problem of estimating X from a knowledge of Y, commonly known as *the inverse prediction problem* or *calibration problem*, is considered in Section 6.4. Section 6.5 discusses the comparison of two or more regression lines. The intersection of two regression lines is discussed in Section 6.6, maximum and minimum of a quadratic regression function in Section 6.7, and spline regression in Section 6.8. In the laboratory manuals we present and explain programs we have written that can be used to perform the calculations required in this chapter.

Each of Sections 6.2 through 6.8 is self-contained, so sections of interest can be studied without a knowledge of the material in other sections. Only the material in Chapters 3 and 4 is prerequisite.

6.2 Prediction Intervals

Consider Example 2.2.2. In that example a car agency predicts Y, the first-year maintenance cost of new cars, based on X, the number of miles the cars will be driven. Suppose the target population is the set of all cars to be made by manufacturer A and driven between 5,000 and 20,000 miles, *next* year. A reasonable study population might be the set of similar cars made by manufacturer A, and driven between 5,000 and 20,000 miles, *last* year.

The questions asked about *next* year's cars will be answered by the agency using a sample from the study population of *last* year's cars. As usual, the *statistical* inference (point estimates, confidence intervals, etc.) from the sample to the study population is valid if the assumptions used are valid. The inference from the *study*

population to the *target* population is not statistical inference but is *judgment inference.*

Suppose that the (study) population regression function is given by $\mu_Y(x) = \beta_0 + \beta_1 x$. Then $\mu_Y(x)$ is the average first-year maintenance cost of all cars in the study population that were driven x miles the first year. Suppose, however, that you are interested not only in $\mu_Y(x)$, the *average* first-year maintenance cost of *all* cars that were driven x miles, but you are also interested in the first-year maintenance cost of a car you will purchase, which is considered to be *randomly chosen* from all cars that were driven x miles. The first-year maintenance cost of this randomly chosen car is denoted by $Y(x)$ and is the quantity that you want to determine. We call $Y(x)$ a random observation to be chosen from the subpopulation with $X = x$. We do not know $Y(x)$, but by using sample data we can obtain a point and interval estimate of it. The point estimate of this randomly chosen car is denoted by $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, which is the same as $\hat{\mu}_Y(x)$, the point estimate of the average maintenance cost of all cars driven x miles. You should not be surprised that the same quantity is used for the point estimate of $Y(x)$ and of $\mu_Y(x)$, because, as we pointed out several times, the mean of the subpopulation is the best predictor for any individual element from that subpopulation.

Obviously we don't expect the point estimate $\hat{Y}(x)$ to be exactly equal to $Y(x)$, and therefore we want to know how good the estimate is. To this end it would be very useful to have an interval

$$[L, U]$$

computed from sample data such that, with a specified degree of confidence (say, $1 - \alpha$), a future randomly chosen Y value from the subpopulation corresponding to $X = x$, i.e., $Y(x)$, will lie in this interval. We write this as

$$C[L \leq Y(x) \leq U] = 1 - \alpha$$

Such an interval is called a two-sided *prediction interval* for $Y(x)$.

There is a close resemblance between *prediction intervals* and confidence intervals. The interval is called a *prediction interval* if the quantity of interest is an observation that will be chosen at random; in this particular case it is $Y(x)$. We refer to this random observation as a *future observation* because it is randomly chosen and not observed until *after the interval is computed.* The interval is called a *confidence interval* if the quantity of interest is a fixed, unknown parameter such as the mean $\mu_Y(x)$.

The procedure for computing prediction intervals for straight line regression was discussed in Section 3.6 and for multiple regression in Section 4.6. In this section we generalize the formulas for prediction intervals to situations involving the *average* and the *sum* of h future observations. These prediction intervals follow the general form given in (4.6.1), which is

$$\hat{\theta} - (\text{table-value}) \times SE(\hat{\theta}) \leq \theta \leq \hat{\theta} + (\text{table-value}) \times SE(\hat{\theta}) \qquad \textbf{(6.2.1)}$$

where θ and $\hat{\theta}$ are $Y(x_1, \ldots, x_k)$ and $\hat{Y}(x_1, \ldots, x_k)$, respectively. In particular, suppose we have a $(k + 1)$-variable study population $\{(Y, X_1, \ldots, X_k)\}$ and we want

to predict a single future value of Y to be selected at random from the subpopulation determined by $X_1 = x_1, \ldots, X_k = x_k$; i.e., we want to estimate the value $Y(x_1, \ldots, x_k)$. The estimate $\hat{Y}(x_1, \ldots, x_k)$ and the appropriate standard error are in (4.4.10) and (4.6.6), respectively. They are repeated here.

$$\hat{Y}(x_1, \ldots, x_k) = \hat{\mu}_Y(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k \tag{6.2.2}$$

$$SE(\hat{Y}(x_1, \ldots, x_k)) = \hat{\sigma}\sqrt{1 + x^T C x} \tag{6.2.3}$$

where $x^T = [1, x_1, \ldots, x_k]$ and $C = (X^T X)^{-1}$. Note that the values $x_1, x_2, \ldots, x_k$ specify the subpopulation from which the Y value is to be obtained. The table-value for a two-sided $1 - \alpha$ prediction interval is $t_{1-\alpha/2;n-k-1}$. These are substituted into (6.2.1) to obtain a $1 - \alpha$ prediction interval for $Y(x_1, \ldots, x_k)$.

Prediction Interval for the Average of h Future Values

Suppose we want to estimate the *average* of h future Y values where the ith future Y value is to be chosen at random from the subpopulation determined by $X_1 = x_{i,1}, \ldots, X_k = x_{i,k}$ (the values $x_{i,1}, \ldots, x_{i,k}$ for $i = 1, \ldots, h$ are not necessarily the sample values) and is denoted by $Y_i(x_{i,1}, \ldots, x_{i,k})$. That is, we want a point and interval estimate of the average of $Y_i(x_{i,1}, \ldots, x_{i,k})$ for $i = 1, \ldots, h$. We denote this average by Y_A so

$$Y_A = \frac{1}{h}\sum_{i=1}^{h} Y_i(x_{i,1}, \ldots, x_{i,k})$$

The estimate and its standard error to use in (6.2.1) are in (6.2.4) and (6.2.5), respectively.

$$\hat{Y}_A = \frac{1}{h}\sum_{i=1}^{h} \hat{Y}_i(x_{i,1}, \ldots, x_{i,k}) = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_k \bar{x}_k \tag{6.2.4}$$

$$SE(\hat{Y}_A) = \hat{\sigma}\sqrt{\frac{1}{h} + \bar{x}^T C \bar{x}} \tag{6.2.5}$$

where

$$x_i^T = [1, x_{i,1}, x_{i,2}, \ldots, x_{i,k}] \qquad \bar{x}^T = [1, \bar{x}_1, \ldots, \bar{x}_k] \tag{6.2.6}$$

and

$$\bar{x}_j = \frac{1}{h}\sum_{i=1}^{h} x_{ij} \qquad \text{for } j = 1, \ldots, k \tag{6.2.7}$$

Also

$$\bar{x}^T = \frac{1}{h}[x_1^T + \ldots + x_h^T]$$

Note that x_i for $i = 1, 2, \ldots, h$ determines the h subpopulations for which we want to predict Y values. The values of $x_{i,1}, \ldots, x_{i,k}$ can be the same for all $i = 1, \ldots, h$,

or some can be the same and some different. The formulas (6.2.4) and (6.2.5) are valid under either assumptions (A) or (B).

Prediction Interval for the Sum of h Future Values

Sometimes we want to estimate the *sum* of h future Y values rather than the *average*. That is, we want a point and interval estimate of the sum of $Y_i(x_{i,1}, \ldots, x_{i,k})$ for $i = 1, \ldots, h$. We denote this sum by Y_S so

$$Y_S = \sum_{i=1}^{h} Y_i(x_{i,1}, \ldots, x_{i,k}) \tag{6.2.8}$$

The estimate of Y_S and its standard error are obtained from those for Y_A in (6.2.4) and (6.2.5) by multiplying those results by h. Specifically,

$$\hat{Y}_S = \sum_{i=1}^{h} \hat{Y}_i(x_{i,1}, \ldots, x_{i,k}) \tag{6.2.9}$$

i.e.,

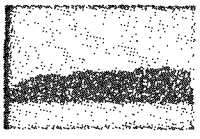
$$\hat{Y}_S = h[\hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_k \bar{x}_k] \tag{6.2.10}$$

and

$$SE(\hat{Y}_S) = \hat{\sigma}\sqrt{h + h^2\, \bar{x}^T C \bar{x}} \tag{6.2.11}$$

where $\bar{x}^T$ is defined in (6.2.6). Note that $SE(\hat{Y}_S) = h\,SE(\hat{Y}_A)$.

We illustrate the use of the preceding formulas in Task 6.2.1.



Task 6.2.1

Suppose that in Example 2.2.2 the agency that evaluates the performance of new cars also evaluates the performance of used cars (cars more than 1 year old). The objective is to predict Y, the maintenance costs of used cars the first year after they are purchased by a new owner. The predictor factors are

- X_1 = miles (in thousands) the car will be driven the first year after it is purchased
- X_2 = age (in months) of the car when it is purchased by the new owner
- X_3 = odometer reading of the car in thousands of miles at the time it was purchased

We suppose that assumptions (A) are valid where the population regression function of Y on X_1, X_2, X_3 is given by

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

The records of used car sales over the past 2 years is the study population, and a simple random sample of size 42 was obtained from this population. The data appear in Table 6.2.1 and are also in the file **usedcars.dat** on the data disk. It is assumed that any cars purchased will be chosen at random from a population of cars that is similar to the population of cars from which the sample of size 42 is obtained; i.e., the target population is judged to be similar to the study population.

TABLE 6.2.1
Used Cars Data

Observation Number	Maintenance Cost (dollars) Y	Miles Driven (thousands) X_1	Age (months) X_2	Odometer Reading (thousands of miles) X_3
1	190	6	70	70.7
2	379	11	72	70.9
3	201	6	98	108.8
4	194	8	60	49.0
5	189	4	84	95.7
6	379	8	84	96.3
7	183	6	64	64.6
8	186	7	64	64.8
9	456	19	60	49.2
10	149	17	36	17.8
11	175	5	66	68.1
12	276	8	72	71.1
13	277	8	72	70.9
14	243	7	72	70.8
15	195	7	66	68.3
16	179	10	58	52.9
17	186	8	61	58.8
18	161	9	54	43.3
19	267	11	60	49.7
20	216	9	60	48.8
21	130	15	36	18.5
22	167	7	58	52.1
23	211	18	48	31.7
24	186	7	24	59.1
25	165	9	54	43.6
26	168	9	54	43.6
27	148	6	60	49.3
28	179	7	61	58.4
29	186	7	64	64.6
30	116	13	36	18.5

(Continued)

TABLE 6.2.1
(Continued)

Observation Number	Maintenance Cost (dollars) Y	Miles Driven (thousands) X_1	Age (months) X_2	Odometer Reading (thousands of miles) X_3
31	312	9	72	71.1
32	168	6	61	58.1
33	190	7	28	49.7
34	235	5	84	95.6
35	195	8	29	48.3
36	140	9	48	32.2
37	209	9	64	68.4
38	607	14	72	95.8
39	181	5	66	68.6
40	176	7	61	58.1
41	201	10	61	58.8
42	279	8	72	70.3

A SAS output from a regression analysis of Y on X_1, X_2, and X_3 is shown in Exhibit 6.2.1 below.

EXHIBIT 6.2.1
SAS Output for Task 6.2.1

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

Model: MODEL1

            X'X Inverse, Parameter Estimates, and SSE

             INTERCEP         MILES           AGE        ODOMETER         MTCOST

INTERCEP   0.8985876124  -0.037785163  -0.005724835  -0.003282576  -198.2198953
MILES      -0.037785163  0.0028862847  -0.000127123  0.0003375752  22.355858102
AGE        -0.005724835  -0.000127123  0.0003194889  -0.000210015  -1.117485612
ODOMETER   -0.003282576  0.0003375752  -0.000210015  0.0002187719  4.8509091502
MTCOST     -198.2198953  22.355858102  -1.117485612  4.8509091502  94817.286025

Dependent Variable: MTCOST
```

EXHIBIT 6.2.1
(Continued)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	253228.33302	84409.44434	33.829	0.0001
Error	38	94817.28603	2495.19174		
C Total	41	348045.61905			

Root MSE	49.95189	R-square	0.7276
Dep Mean	219.76190	Adj R-sq	0.7061
C.V.	22.73001		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-198.219895	47.35132929	-4.186	0.0002
MILES	1	22.355858	2.68362326	8.330	0.0001
AGE	1	-1.117486	0.89285271	-1.252	0.2184
ODOMETER	1	4.850909	0.73883552	6.566	0.0001

Below are two typical problems that the agency may want to solve.

1 Someone who plans to buy a used car that will be driven 9,000 miles next year asks the agency to predict the first-year maintenance cost of this car. The car is 36 months old and has 30,200 miles on the odometer.

We compute point and interval estimates of $Y(x_1, x_2, x_3)$, *the first-year maintenance cost of a future randomly chosen car* ($h = 1$) *from the subpopulation with*

$$X_1 = x_{1,1} = 9.0, \quad X_2 = x_{1,2} = 36, \quad X_3 = x_{1,3} = 30.2$$

For the point estimate we get $\hat{Y}(9.0, 36, 30.2) = 109.25$ *by substituting* $x_{1,1} = 9.0$, $x_{1,2} = 36$, *and* $x_{1,3} = 30.2$ *into the estimated regression equation. Thus the best estimate of the first-year maintenance cost of this car is* \$109.25.

To compute a 95% prediction interval for $Y(9.0, 36, 30.2)$ *we need*

$$SE(\hat{Y}(9.0, 36, 30.2))$$

which is $\hat{\sigma}\sqrt{1 + x^T C x}$ *in (6.2.3) with* $h = 1$. *The matrix* C *is given in Exhibit 6.2.1 (see the explanation preceding Problem 4.6.1.). Since* $x^T = [1, 9.0, 36, 30.2]$, *we get*

$$x^T C x = 0.0998475$$

From Exhibit 6.2.1 we get $\hat{\sigma} = 49.95189$, *and so*

$$SE\,(\hat{Y}(9.0, 36, 30.2)) = \hat{\sigma}\sqrt{1 + x^T C x} = 52.3864$$

The table-value, obtained by linear interpolation in Table T-2 in Appendix T, is $t_{1-\alpha/2:n-k-1} = t_{0.975:38} = 2.025$. *Substituting into (6.2.1) we get*

$$C[3.17 \leq Y(9.0, 36, 30.2) \leq 215.33] = 0.95$$

If only an upper confidence bound is desired, then the purchaser has 97.5% confidence that the first-year maintenance cost of a car with $X_1 = 9, X_2 = 36, X_3 = 30.2$, *randomly chosen from the study population, will not exceed* \$215.33. *To extrapolate this to the first-year maintenance cost of a car to be purchased next year is judgment-based inference using the preceding confidence interval.*

2 A company plans to purchase two used cars. Car 1 will be driven 6,000 miles the first year and car 2 will be driven 15,000 miles. The two cars being considered for purchase have the following ages and odometer readings. Car 1 is 24 months old and has 48,900 miles on its odometer, whereas car 2 is 21 months old and has 32,100 miles on its odometer. The company wants to predict the *total* first-year maintenance cost of these two cars.

We denote the total first-year maintenance cost of these two cars by Y_S, *and the estimate is*

$$\hat{Y}_S = \hat{Y}_1(6.0, 24, 48.9) + \hat{Y}_2(15.0, 21, 32.1)$$

using (6.2.9) with $h = 2$. *By substituting the values of the predictor variables into the estimated regression equation, we get*

$$\hat{Y}_1(6.0, 24, 48.9) = 146.31 \qquad \textit{and} \qquad \hat{Y}_2(15.0, 21, 32.1) = 269.36$$

so

$$\hat{Y}_S = 146.31 + 269.36 = 415.67$$

To obtain a 95% prediction interval for Y_S, *we also need* $SE(\hat{Y}_S)$, *and to obtain this we use (6.2.11). Hence we need*

$$\bar{x}^T C \bar{x}$$

for which we first need to calculate $\bar{x}$ *in (6.2.6), where*

$$\bar{x} = (1/2)[x_1 + x_2]$$

We note that

$$x_1^T = [1, 6.0, 24, 48.9] \qquad \textit{and} \qquad x_2^T = [1, 15.0, 21, 32.1]$$

and hence

$$\bar{x}^T = (1/2)\{[1, 6.0, 24, 48.9] + [1, 15.0, 21, 32.1]\} = [1, 10.5, 22.5, 40.5]$$

The quantity $\bar{x}^T C\bar{x} = 0.2646767$, *and (note that* $h = 2$*)*

$$\sqrt{h + h^2\,\bar{x}^T C\bar{x}} = \sqrt{2 + 4(0.2646767)} = 1.748916$$

From the computer output we have $\hat{\sigma} = 49.95189$. *So*

$$SE(\hat{Y}_S) = \hat{\sigma}\sqrt{h + h^2\,\bar{x}^T C\bar{x}} = (49.95189)(1.748916) = 87.3617$$

Thus a 95% prediction interval for Y_S *is*

$$\hat{Y}_S - t_{0.975:38} SE(\hat{Y}_S) \le Y_S \le \hat{Y}_S + t_{0.975:38} SE(\hat{Y}_S)$$

This gives us

$$C[415.67 - (2.025)(87.3617) \le Y_S \le 415.67 + (2.025)(87.3617)] = 0.95$$

or

$$C[238.76 \le Y_S \le 592.58] = 0.95$$

If only an upper confidence bound is needed, the company can be 97.5% confident that the total first-year maintenance cost of two cars to be chosen at random from this study population, with specified miles driven, age, and odometer readings, will not exceed \$592.58. As usual, the extrapolation of these results to next year's cars is a subject matter inference.

Problems 6.2

Problems 6.2.1–6.2.3 refer to Task 6.2.1, for which the data are in Table 6.2.1 and are also stored in the file **usedcars.dat** on the data disk. You may use the computer output in Exhibit 6.2.1.

6.2.1 **a** What is the estimated first-year maintenance cost of a used car that will be driven 12,000 miles during the first year after purchase if it is 22 months old and has 9,300 miles on its odometer at the time of purchase by a new owner?

b What is the estimate of the *average* first-year maintenance cost of all used cars driven 12,000 miles the first year if they are 22 months old and have 9,300 miles on their odometers at the time of purchase by their respective new owners?

c Obtain a 95% prediction interval for $Y(12.0, 22, 9.3)$ in part (a). In (6.2.3) we calculated the value of the quantity $x^T Cx$, which is needed to compute this prediction interval, and it is equal to 0.1903.

d Obtain a 95% confidence interval for $\mu_Y(12.0, 22, 9.3)$ in part (b).

e Write a short report and explain how the results in parts (a)–(d) can be used to make decisions about the target population.

6.2.2 A company plans to buy two used cars for its sales staff who will drive the cars the following distances the first year after purchase: driver 1 will drive 9,000 miles, and

driver 2 will drive 18,000 miles. The two cars being considered for purchase have the following ages and odometer readings:

Car 1: age = 12 months, odometer reading = 13,700 miles

Car 2: age = 20 months, odometer reading = 24,300 miles

The company wants to hold the estimated *total* first-year maintenance cost to a minimum. Which driver should be given which car? Explain your reasoning in detail.

6.2.3 In Problem 6.2.2 estimate the *total* first-year maintenance cost of the two cars if driver 1 is given car 1 and driver 2 is given car 2.

6.2.4 This problem refers to Task 3.4.1 where crystalline forms of certain chemical compounds are used in various electronic devices and it is more desirable to have large crystals 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 Y, the weight of a crystal in grams, and X, the time in hours for 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, along with the MINITAB output from a regression analysis of Y on X, are given in Exhibit 6.2.2. These are the same data that appear in Table 3.4.2 and are also stored in the file **crystal.dat** on the data disk. Assumptions (A) are presumed to be valid, and the data were obtained by sampling with preselected X values.

EXHIBIT 6.2.2

MINITAB Output for Problem 6.2.4

obsno	weight	time
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

EXHIBIT 6.2.2
(Continued)

```
The regression equation is
weight = 0.001 + 0.503 time

Predictor       Coef      Stdev    t-ratio       p
Constant      0.0014     0.5994       0.00   0.998
time         0.50343    0.03520      14.30   0.000


s = 1.062       R-sq = 94.5%     R-sq(adj) = 94.0%


Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      230.63      230.63     204.58   0.000
Error        12       13.53        1.13
Total        13      244.16


The C matrix is

 .318681318      -.016483516
-.016483516       .001098901
```

a If a crystal is allowed to grow for 15 hours, estimate $Y(15)$, its weight.

b Obtain a 90% two-sided prediction interval for the weight of the crystal in part (a).

Problems (c), (d), (e), and (f) depend on the fact that the crystals are priced based on the time taken to grow them as well on 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.

c Estimate the *total* weight of three crystals where crystal 1 is to be grown for 3 hours, crystal 2 is to be grown for 5 hours, and crystal 3 is to be grown for 13 hours.

d Obtain a lower bound L in part (c) such that you have 95% confidence that the total weight of all three crystals is greater than L.

e A customer orders a crystal that is to grow for 20 hours. Find the point estimate of the dollar value of this crystal.

f Obtain a lower bound L in part (e) such that you have 90% confidence that the dollar value of this crystal is greater than L.

g Obtain a point estimate of the *average* weight of all crystals grown for 20 hours.

h Obtain a 90% two-sided confidence interval for the *average* weight of the crystals in part (g).

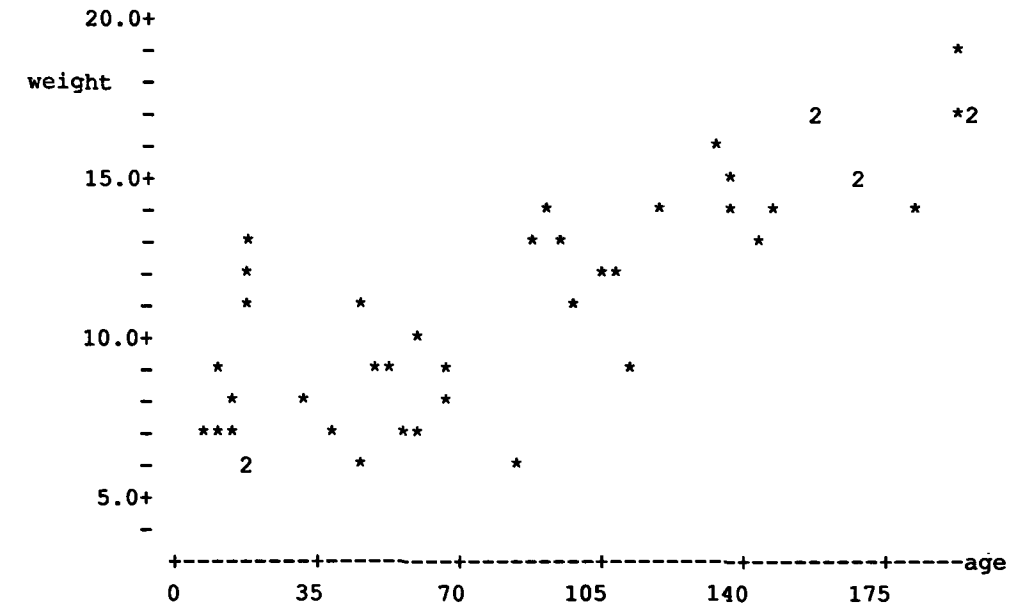
6.2.5 A scientist is interested in Y, the weight in pounds of babies belonging to a certain ethnic group as a function of X, the age in days. A simple random sample was selected from the records of babies born in two hospitals in a certain large city for the past three years, and their weights and ages were recorded. The data are given in Table 6.2.2 and are also stored in the file **ethnic.dat** on the data disk.

The target population that the scientist wants to investigate is the entire set of babies in the United States that belong to this ethnic group. The study population consists of the records in the hospitals previously referred to. A MINITAB output from a regression of Y on X, including a plot of Y versus X is given in Exhibit 6.2.3. Assumptions (A) are presumed valid.

TABLE 6.2.2
Ethnic Babies Data

Observation Number	Weight Y	Age X	Observation Number	Weight Y	Age X
1	6.60	7	23	12.77	88
2	7.44	10	24	14.04	92
3	9.10	12	25	13.13	96
4	8.16	13	26	11.14	99
5	7.16	14	27	12.02	105
6	6.41	16	28	12.11	108
7	11.79	17	29	8.84	113
8	12.94	17	30	13.64	120
9	10.92	18	31	15.67	132
10	5.97	18	32	15.36	135
11	8.19	30	33	14.41	138
12	6.99	37	34	13.39	142
13	5.70	45	35	14.02	147
14	11.43	46	36	17.15	156
15	9.48	49	37	17.21	159
16	9.49	52	38	14.61	167
17	6.59	56	39	15.01	168
18	6.93	58	40	14.32	183
19	9.78	59	41	17.05	191
20	7.54	67	42	18.88	194
21	8.50	67	43	16.77	195
22	6.44	84	44	17.26	196

EXHIBIT 6.2.3
MINITAB Output for Problem 6.2.5.



The regression equation is
weight = 6.81 + 0.0517 age

Predictor	Coef	Stdev	t-ratio	p
Constant	6.8124	0.5624	12.11	0.000
age	0.051738	0.005219	9.91	0.000

s = 2.104 R-sq = 70.1% R-sq(adj) = 69.3%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	434.90	434.90	98.29	0.000
Error	42	185.84	4.42		
Total	43	620.74			

The C matrix is

0.07148029	-0.0005477867
-0.00054778	0.0000061549

a Exhibit the population parameter that represents the average weight of all babies in the study population who are 30 days old. Repeat this for babies who are 60 days old.

b What is the estimated average weight of babies who are 60 days old?

c Compute a 90% two-sided confidence interval for the quantity in part (b).

d Compute a 90% two-sided interval for the weight of a randomly chosen baby who is 60 days old.

6.3 Tolerance Intervals

Until now, in straight line regression, we have been concentrating on $\mu_Y(x)$, the mean of the Y values in the subpopulation where $X = x$, and on a randomly chosen value $Y(x)$ from this subpopulation. We now turn our attention to the actual distribution of the Y values in the subpopulation where $X = x$. In particular we are interested in finding a number $\lambda_p(x)$ such that a proportion p of the Y values in the subpopulation corresponding to $X = x$ will be less than this number (see Figure 6.3.1). Then a proportion $1 - p$ of the Y values in this subpopulation will be greater than $\lambda_p(x)$ and, for any p such that $0 \leq p \leq 1$, a proportion p will lie in the interval

$$[\lambda_{(1-p)/2}(x), \lambda_{(1+p)/2}(x)]$$

(see Figure 6.3.2). Note that in Figure 6.3.2 the area to the left of $\lambda_{(1-p)/2}(x)$ is $(1-p)/2$ and the area to the left of $\lambda_{(1+p)/2}(x)$ is $(1+p)/2$, so the area between $\lambda_{(1-p)/2}(x)$ and $\lambda_{(1+p)/2}(x)$ is equal to $(1+p)/2 - (1-p)/2 = p$. We illustrate with two examples.

FIGURE 6.3.1

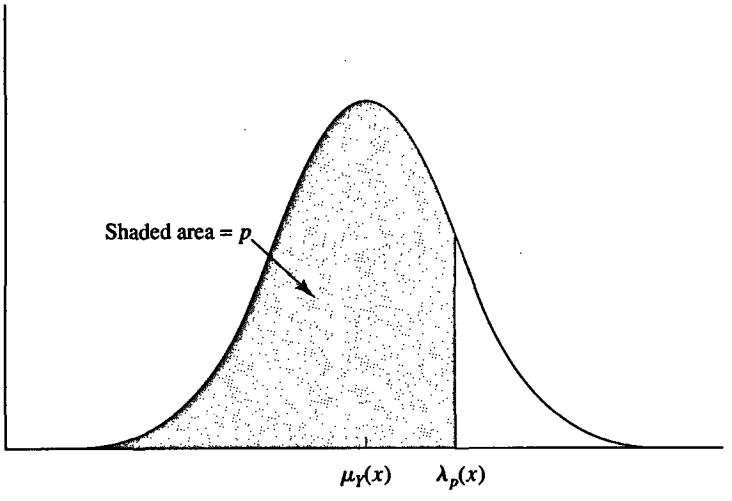
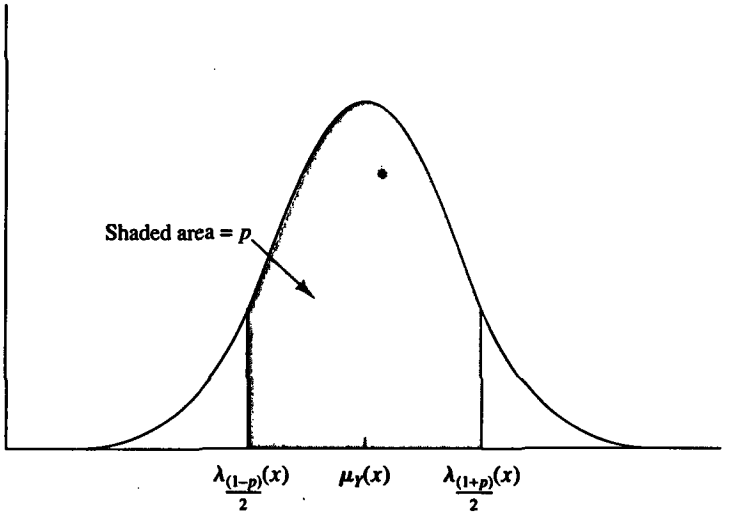


FIGURE 6.3.2



EXAMPLE 6.3.1

Blood cholesterol level Y in adults who have taken a certain drug for a year tends to change with age X. Suppose that in a study population the regression function is given by $\mu_Y(x) = \beta_0 + \beta_1 x$ for $35 \leq x \leq 60$. Based on a simple random sample of size n, we can estimate $\mu_Y(x)$ for any specified age group in $35 \leq x \leq 60$. This gives us information about the average cholesterol levels for various age groups. For a given age group with $X = x$ and a given proportion p, we may want to determine a number $\lambda_p(x)$ such that *a proportion p of this subpopulation will have cholesterol values below this number.* For example, we may want to make statements of the form "99% (i.e., $p = 0.99$) of the individuals in this population who are 40 years old (i.e., $x = 40$) have cholesterol levels below $\lambda_{0.99}(40)$." The number $\lambda_{0.99}(40)$ gives us an idea of the upper extremes of cholesterol levels for individuals in the age group $x = 40$, and this information can be extremely useful in identifying individuals with abnormally high cholesterol levels. ■

EXAMPLE 6.3.2

A company that manufactures steel rods is studying how X_1, the hardness of the steel rods, and X_2, the amount of carbon used in the alloy, affect Y, the strength of the rods. For a given value of $X_1 = x_1$ and $X_2 = x_2$, and for $p = 0.001$, the company wants to determine the number $\lambda_p(x_1, x_2) = \lambda_{0.001}(x_1, x_2)$ such that a proportion of only 0.001 (i.e., 0.1%) of the steel rods produced will have strength less than that number; i.e., 99.9% of the rods will have strength greater than $\lambda_{0.001}(x_1, x_2)$. ■

Tolerance Points

Consider a $(k+1)$-variable study population $\{(Y, X_1, \ldots, X_k)\}$. We are interested in the following quantities for the subpopulation with $X_1 = x_1, \ldots, X_k = x_k$.

1 The number

$$\lambda_p(x_1, \ldots, x_k) \tag{6.3.1}$$

such that a proportion p of the subpopulation Y values are below $\lambda_p(x_1, \ldots, x_k)$, and hence a proportion $1-p$ of the subpopulation Y values are above $\lambda_p(x_1, \ldots, x_k)$. This is called the *p*th *tolerance point* or *p*th *percentile* for the subpopulation (i.e., the area between $-\infty$ and $\lambda_p(x_1, \ldots, x_k)$ is p) (see Figure 6.3.1.).

2 The two numbers $\lambda_{(1-p)/2}(x_1, \ldots, x_k)$ and $\lambda_{(1+p)/2}(x_1, \ldots, x_k)$ such that a proportion p of the subpopulation Y values are in the interval

$$[\lambda_{(1-p)/2}(x_1, \ldots, x_k), \lambda_{(1+p)/2}(x_1, \ldots, x_k)] \tag{6.3.2}$$

This is called a *two-sided* (symmetric) *population tolerance interval* (See Figure 6.3.2).

Under population assumptions (A) or (B), the pth subpopulation tolerance point corresponding to $X_1 = x_1, \ldots, X_k = x_k$ is

$$\lambda_p(x_1, \ldots, x_k) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + z_p \sigma \tag{6.3.3}$$

where z_p is the pth percentile of the standard Gaussian population. Because $\lambda_p(x_1, \ldots, x_k)$ in (6.3.3) contains unknown parameters β_i and σ, we must use sample values to compute point and interval estimates for it. So data are collected and $\hat{\beta}_i$ and $\hat{\sigma}$ are computed by formulas in Chapter 4. A point estimate of $\lambda_p(x_1, \ldots, x_k)$ is

$$\hat{\lambda}_p(x_1, \ldots, x_k) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k + z_p \hat{\sigma} \tag{6.3.4}$$

We also compute a confidence interval for $\lambda_p(x_1, \ldots, x_k)$ of the form

$$C[L_p \leq \lambda_p(x_1, \ldots, x_k) \leq U_p] = 1 - \alpha \tag{6.3.5}$$

In (6.3.5) and elsewhere we use L_p for $L_p(x_1, \ldots, x_k)$ for ease of notation. Similarly we write U_p for $U_p(x_1, \ldots, x_k)$. We can attach one of the following three meanings to the confidence interval in (6.3.5).

1 We have confidence $1-\alpha$ that the interval L_p to U_p includes $\lambda_p(x_1, \ldots, x_k)$.

2 If only an upper confidence bound is needed, then we have $1-\alpha/2$ confidence that *at least* $100p\%$ of the Y values in the subpopulation corresponding to $X_1 = x_1, \ldots, X_k = x_k$ are below U_p because $100p\%$ of Y values in this subpopulation are below $\lambda_p(x_1, \ldots, x_k)$, which in turn is smaller than U_p with confidence $1-\alpha/2$.

3 If only a lower confidence bound is needed, then we have $1-\alpha/2$ confidence that *at least* $100(1-p)\%$ of the Y values in the subpopulation corresponding to $X_1 = x_1, \ldots, X_k = x_k$ are above L_p because $100(1-p)\%$ of the Y values are above $\lambda_p(x_1, \ldots, x_k)$, which in turn is greater than L_p with confidence $1-\alpha/2$.

Also, using the Bonferroni method we can obtain numbers $L_{(1-p)/2}$ and $U_{(1+p)/2}$ such that

$$C[L_{(1-p)/2} \le \lambda_{(1-p)/2}(x_1, \ldots, x_k) \quad \text{and} \quad \lambda_{(1+p)/2}(x_1, \ldots, x_k) \le U_{(1+p)/2}] \ge 1 - \alpha$$

But, since a proportion p of the subpopulation values are between $\lambda_{(1-p)/2}(x_1, \ldots, x_k)$ and $\lambda_{(1+p)/2}(x_1, \ldots, x_k)$ we have confidence greater than or equal to $1 - \alpha$ that at least $100p\%$ of the Y values in the subpopulation with $X_1 = x_1, \ldots, X_k = x_k$ are between the values $L_{(1-p)/2}$ and $U_{(1+p)/2}$. For example, if $p = 0.80$, then $(1-p)/2 = 0.10$ and $(1+p)/2 = 0.90$, and we have confidence greater than $1 - \alpha$ that at least $100p\% = 80\%$ of the Y values in the subpopulation with $X_1 = x_1, \ldots, X_k = x_k$ will be between $L_{0.10}$ and $U_{0.90}$.

The instructions for computing L_p and U_p are in Box 6.3.1.

BOX 6.3.1 Instructions for Computing L_p and U_p

Assumptions (A) or (B) are presumed to be valid, and a sample of size n is selected from a $(k+1)$-variable study population $\{(Y, X_1, \ldots, X_k)\}$. The sample values and the corresponding X matrix are

Sample			
Y	X_1	...	X_k
y_1	$x_{1,1}$	$\cdots$	$x_{1,k}$
y_2	$x_{2,1}$	$\cdots$	$x_{2,k}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
y_n	$x_{n,1}$	$\cdots$	$x_{n,k}$

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,k} \\ 1 & x_{2,1} & \cdots & x_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,k} \end{bmatrix}$$

1 The confidence coefficient $1 - \alpha$ is specified by the investigator.

2 The number p is specified by the investigator.

3 The subpopulation of interest has $X_1 = x_1, \ldots, X_k = x_k$ and, using these values, the vector x is defined by

$$x^T = [1, x_1, \ldots, x_k] \tag{6.3.6}$$

4 The statistics $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$ and $\hat{\sigma}$ are computed by the formulas in Chapter 4 using the sample data.

5 The matrix $C = (X^T X)^{-1}$ is computed where X is given above. The quantity A is computed where

$$A = \sqrt{x^T C x}$$

and where x is given in (6.3.6).

6 z_p is obtained from Table T-1 in Appendix T where z_p is the pth percentile point of a standard normal population.

7 δ_p is computed where $\delta_p = -z_p/A$.

8 Look up $t_{1-\alpha/2:n-k-1;(\delta_p)}$ and $t_{\alpha/2:n-k-1;(\delta_p)}$. The quantity $t_{\gamma:n-k-1;(\delta_p)}$ (for $\gamma = 1-\alpha/2$ and for $\gamma = \alpha/2$) is obtained from Table T-8 in Appendix T and is the γ percentile point of the noncentral t distribution with $n-k-1$ degrees of freedom and noncentrality δ_p. The quantity δ_p is in (7). The quantity $t_{\gamma:n-k-1;(\delta_p)}$ can also be obtained from several statistical computing packages. It is useful to note that $t_{1-\alpha/2:n-k-1;(\delta)} = -t_{\alpha/2:n-k-1;(-\delta)}$.

9 Compute $g_{p,1-\alpha/2}$ and $g_{p,\alpha/2}$, where $g_{p,1-\alpha/2} = -At_{1-\alpha/2:n-k-1;(\delta_p)}$ and $g_{p,\alpha/2} = -At_{\alpha/2:n-k-1;(\delta_p)}$.

10 The confidence bounds L_p and U_p in (6.3.5) are given in (6.3.7) and (6.3.8), respectively.

$$L_p = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k + \hat{\sigma} g_{p,1-\alpha/2} \qquad \textbf{(6.3.7)}$$

$$U_p = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_k x_k + \hat{\sigma} g_{p,\alpha/2} \qquad \textbf{(6.3.8)}$$

To help understand these computations, we illustrate the procedures using an artificial example.

EXAMPLE 6.3.3

Assume that the data in Table 6.3.1 were obtained from a two-variable target population $\{(Y, X)\}$ by sampling with preselected X values. Suppose assumptions (A) are satisfied. The data are also stored in the file **table631.dat** on the data disk.

We want the following quantities.

a A point estimate of $\lambda_{0.80}(3.0)$, the number such that 80% of the Y values in the subpopulation with $X = 3.0$ is less than that number.

b A 95% two-sided confidence interval for $\lambda_{0.80}(3.0)$.

c A point estimate of $\lambda_{0.20}(3.0)$, the number such that 20% of the subpopulation with $X = 3.0$ is less than that number.

d A 95% two-sided confidence interval for $\lambda_{0.20}(3.0)$.

For this problem $k = 1$ and, as usual, the matrix X can be obtained from the column of x_i values given in Table 6.3.1 by attaching a column of 1's as the first column. A SAS output from a regression analysis of Y on X follows in Exhibit 6.3.1.

TABLE 6.3.1

Observation Number	Y	X
1	4.81	1.0
2	3.60	1.1
3	4.90	1.3
4	3.05	1.6
5	3.44	1.8
6	3.17	1.8
7	3.34	1.8
8	1.61	2.1
9	1.22	2.4
10	0.20	2.6
11	1.56	2.6
12	0.55	2.7
13	−2.56	2.9
14	−0.34	3.0
15	−2.56	3.5
16	−2.96	3.6
17	−1.04	4.1
18	−4.64	5.2

EXHIBIT 6.3.1

SAS Output for Example 6.3.3

```
                              The SAS System
                                              00:00 Saturday,  Jan 1, 1994
Model: MODEL1

              X'X Inverse, Parameter Estimates, and SSE

                    INTERCEP                X                Y

INTERCEP       0.3598685805      -0.121455309      6.9909129346
X              -0.121455309      0.0484744028      -2.405464142
Y              6.9909129346      -2.405464142      17.808145181

Dependent Variable: Y
```

EXHIBIT 6.3.1
(Continued)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	119.36728	119.36728	107.247	0.0001
Error	16	17.80815	1.11301		
C Total	17	137.17543			

Root MSE	1.05499	R-square	0.8702
Dep Mean	0.96389	Adj R-sq	0.8621
C.V.	109.45167		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	6.990913	0.63287992	11.046	0.0001
X	1	-2.405464	0.23227667	-10.356	0.0001

The quantities needed for parts (a) and (b) as described in Box 6.3.1, are exhibited below.

1 $1 - \alpha = 0.95$, so $\alpha = 0.05$, $\alpha/2 = .025$, and $1 - \alpha/2 = 0.975$.

2 $p = 0.80$, so $1 - p = 0.20$.

3 The subpopulation for which the tolerance point is to be evaluated has $X = 3.0$, so $\boldsymbol{x}^T = [1,\ 3.0]$.

4 From the computer output in Exhibit 6.3.1 we get

$$\hat{\beta}_0 = 6.9909 \qquad \hat{\beta}_1 = -2.4055 \qquad \hat{\sigma} = 1.055$$

5 From the matrix $\boldsymbol{C}$ and the vector $\boldsymbol{x}^T = [1,\ 3.0]$ above, we compute

$$A = \sqrt{\boldsymbol{x}^T \boldsymbol{C} \boldsymbol{x}} = \sqrt{0.0674} = 0.2596$$

6 $z_p = z_{0.80} = 0.8416$ (from SAS).

7 $\delta_p = \delta_{0.80} = -0.8416/0.2596 = -3.242$.

8 $t_{0.975:16;(-3.242)} = -1.254$ and $t_{0.025:16;(-3.242)} = -6.174$ from Table T-8 in Appendix T.

9 $g_{0.80,0.975} = -(0.2596)(-1.254) = 0.3255$ and
$g_{0.80,0.025} = -(0.2596)(-6.174) = 1.6028$.

Also $\hat{\beta}_0 + \hat{\beta}_1 x = 6.9909 + (-2.4055)(3.0) = -0.2256$. We are now in a position to calculate the needed quantities.

a The point estimate of $\lambda_{0.80}(3.0)$ is

$$\hat{\lambda}_{0.80}(3.0) = -0.2256 + (0.8416)(1.055) = 0.6623$$

b From (6.3.7) and (6.3.8) we get

$$L_{0.80} = -0.2256 + (0.3255)(1.055) = 0.1178$$

and

$$U_{0.80} = -0.2256 + (1.6028)(1.055) = 1.465$$

so a 95% confidence interval for $\lambda_{0.80}(3.0)$ is given by

$$C[0.1178 \leq \lambda_{0.80}(3.0) \leq 1.465] = 0.95$$

c The point estimate of $\lambda_{0.20}(3.0)$, using (6.3.4) with $z_{0.20} = -0.8416$, is

$$\hat{\lambda}_{0.20}(3.0) = -0.2256 + (-0.8416)(1.055) = -1.113$$

d To obtain $L_{0.20}$ and $U_{0.20}$ we get $\delta_{0.20} = -(-0.8416)/0.2596 = 3.242$. From Table T-8 in Appendix T we obtain $t_{0.975:16;(3.242)} = -t_{0.025:16;(-3.242)} = 6.174$, and $t_{0.025:16;(3.242)} = 1.254$. So $g_{0.20,0.975} = -(0.2596)(6.174) = -1.6028$, and $g_{0.20,0.025} = -(0.2596)(1.254) = -0.3255$. From (6.3.7) and (6.3.8) we get

$$L_{0.20} = -0.2256 - 1.6028(1.055) = -1.917$$

and

$$U_{0.20} = -0.2256 - 0.3255(1.055) = -0.5690$$

so a 95% confidence interval for $\lambda_{0.20}(3.0)$ is given by

$$C[-1.917 \leq \lambda_{0.20}(3.0) \leq -0.5690] = 0.95 \quad \blacksquare$$

Problems 6.3

6.3.1 **a** In Example 6.3.3 compute a point estimate of $\lambda_{0.85}(3.0)$.

b In Example 6.3.3 compute a 90% two-sided confidence interval for $\lambda_{0.85}(3.0)$. Write a paragraph explaining what this means.

c In Example 6.3.3 compute a 90% two-sided confidence interval for $\lambda_{0.15}(3.0)$ and state in words what this means.

6.3.2 In Example 6.3.3 compute numbers L and U such that you have confidence greater than or equal to 90% that at least a proportion $p = 0.7$ of the subpopulation of Y values with $X = 3.0$ is between L and U.

6.3.3 This problem refers to Problem 6.2.5 where a scientist is interested in Y, the weight in pounds of babies belonging to a certain ethnic group, as a function of X, the age in days. A simple random sample was selected from the records of weights and ages of

babies in two hospitals in a certain large city. The data are given in Table 6.2.2 and are also stored in the file **ethnic.dat** on the data disk. Assumptions (A) are presumed to hold, and the regression function is

$$\mu_Y(x) = \beta_0 + \beta_1 x \qquad \text{for} \qquad 5 \le x \le 200$$

A SAS output from a regression analysis of Y on X is given in Exhibit 6.3.2.

EXHIBIT 6.3.2
SAS Output for Problem 6.3.3

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

Model: MODEL1

X'X Inverse, Parameter Estimates, and SSE

	INTERCEP	AGE	WEIGHT
INTERCEP	0.0714802886	-0.000547787	6.8123997925
AGE	-0.000547787	6.1549067E-6	0.0517375917
WEIGHT	6.8123997925	0.0517375917	185.83817639

Dependent Variable: WEIGHT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	434.90154	434.90154	98.289	0.0001
Error	42	185.83818	4.42472		
C Total	43	620.73972			

Root MSE	2.10350	R-square	0.7006
Dep Mean	11.41705	Adj R-sq	0.6935
C.V.	18.42422		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	Parameter=0	Prob > |T|
INTERCEP	1	6.812400	0.56238790	12.113	0.0001
AGE	1	0.051738	0.00521859	9.914	0.0001

a Parents of a 30-day-old baby who is in the study population bring their baby to the hospital and want to know if their baby is underweight. Do you consider this baby's weight to be a random observation from the study population? Explain.

b In part (a) would you compare this baby's weight with the average of all babies in the study population who are 30 days old to help answer their question? Explain.

c To help answer the parents' question in part (a), suppose the physician informs them that 90% of all babies in this study population who are 30 days old weigh more than c pounds. Which of the following subpopulation parameters is the quantity c?

i $Y(30)$

ii $\mu_Y(30)$

iii $\lambda_{0.90}(30)$

iv $\lambda_{0.10}(30)$

d Compute an appropriate interval for the correct quantity in part (c) such that we have 95% confidence that the interval contains this quantity. Write a short report for the physician to give to the parents.

e Suppose two other parents whose baby is in the study population bring their 60-day-old baby to the hospital and want to know if their baby is underweight. The physician weighs the baby and notices that it weighs 7.3 pounds. To determine how usual or unusual this weight is, which of the following parameters is of interest to the physician?

i $Y(x)$

ii $\mu_Y(x)$

iii $\lambda_p(x)$

What is the correct value of x for the physician to use?

f Suppose that a baby's weight is considered to be unusually low if its weight is in the lower 5% of the subpopulation for its age. What value of p should be used in part (e) for the physician to consider the baby's weight unusual?

g Compute an appropriate 95% one-sided confidence bound for the proper quantity in part (e) and, in a written report, interpret this bound so the family can understand it.

6.4
Calibration and Regulation for Straight Line Regression

Most of the discussion in this book has been about the population regression function $\mu_Y(x)$ or about predicting Y as a function of X, and the statistical inferences are valid if assumptions (A) or (B) are satisfied. However, there are several very important applications where the primary interest is not predicting Y as a function of X, but rather to predict X as a function of Y. It may seem at first that we could use all the formulas developed in Chapter 3 and just interchange Y and X. But as we shall see, it is sometimes not that simple. Two such applications are discussed in this section

where we want to predict X as a function of Y, but it is not appropriate to simply interchange the roles of Y and X. The applications are referred to as **calibration** and **regulation**, respectively.

We illustrate with two examples.

EXAMPLE 6.4.1

Determining small quantities of toxic substances in water samples is sometimes a very difficult problem for analytical chemists. Typically analytical methods do not recover all of a given toxic substance present in the water sample. Suppose for a given chemical procedure a chemist knows that the relationship between X, the actual amount of mercury present in water samples and Y, the amount of mercury recovered by the analytical method, is given by the regression function $\mu_Y(x) = \beta_0 + \beta_1 x$. A set of standard solutions containing *known* amounts $x_1, x_2, \ldots, x_n$ of mercury, is prepared and is subjected to chemical analysis. Then $y_1, y_2, \ldots, y_n$, the amounts of mercury recovered from these solutions, are recorded. Clearly data are obtained by sampling with preselected X values because the values $x_1, x_2, \ldots, x_n$ were preselected (solutions with *known* amounts of mercury were prepared and used). So it may be reasonable to suppose that assumptions (A) hold. The sample can be presumed to have been obtained from a target population of interest. Estimates of the parameters β_0, β_1, and σ can be calculated using the sample data $(y_1, x_1), \ldots, (y_n, x_n)$. Then a water sample containing an *unknown* amount x_0 of mercury is subjected to the same chemical analysis, and the amount of mercury recovered is measured to be y_0. The chemist wants to determine x_0, the actual amount of mercury present. Thus it is necessary to estimate x_0, the subpopulation from which a single sample value y_0 was obtained. ■

EXAMPLE 6.4.2

Suppose a company is investigating a new food supplement for increasing the weight of chickens. The research scientist decides that the regression function $\mu_Y(x) = \beta_0 + \beta_1 x$ relates the average weight Y gained in pounds and X, the time in weeks the chickens have been fed the new food supplement. An experiment is conducted with five groups of chickens, where each group of chickens is fed the new ration for a different amount of time, say $X = 2, 4, 6, 8$, and 10 weeks, respectively, and the weights Y recorded. The data are used to obtain estimates of β_0, β_1, and σ. Since the X values were preselected (the number of weeks each chicken was fed the new ration was preselected), it may be reasonable to suppose that assumptions (A) apply where data are obtained by preselected X values. The company wants to determine x_0 so that in their advertisements for the new product they can claim, "If you want your chickens to gain an *average* of 10 pounds, feed them the new supplement for x_0 weeks." In other words, the investigator wants to determine x_0 for which $\mu_Y(x_0) = 10$. ■

The two examples are slightly different. In Example 6.4.1 a chemist wants to determine x_0, the X value for the subpopulation from which a *single* sample value y_0 is observed. In Example 6.4.2 a scientist wants to determine x_0, the value of X for the subpopulation whose *mean* $\mu_Y(x_0)$ is equal to a specified value, say m_0. The first problem, determining x_0 for a given y_0, is referred to as a *calibration* problem, whereas the second problem, determining x_0 for a given value of $\mu_Y(x_0)$, is referred to as a *regulation* problem. Both are *inverse estimation* problems, i.e., estimating the value of X given some information about Y.

As stated earlier, the first thing that may cross your mind is to interchange Y and X, use the regression function of X on Y given by $\mu_X(y) = \beta_0^* + \beta_1^* y$, and use the formulas in Chapter 3 for point estimates and confidence intervals for $\mu_X(y)$ and $X(y)$. This is the correct way to proceed if the problem is such that assumptions (B) are met, but if the data are obtained by preselecting the X values, as they are in many cases (for instance in Examples 6.4.1 and 6.4.2), then assumptions (A) are not satisfied when Y and X are interchanged. It is easy to see why this procedure is not justified, because if X and Y were interchanged, assumptions (A) require the observed X values to be a simple random sample from a Gaussian population for each specified value of Y, but X values are preselected and thus cannot be a simple random sample from a Gaussian (or any other) population.

This section applies to problems when assumptions (A) are valid, data are obtained by preselected X values, and the unknown value x_0 corresponding to an observed value y_0 (or corresponding to a specified value m_0 of $\mu_Y(x)$) is to be determined.

We discuss calibration and regulation separately and exhibit formulas for point and confidence interval estimates for x_0 in each case.

Calibration

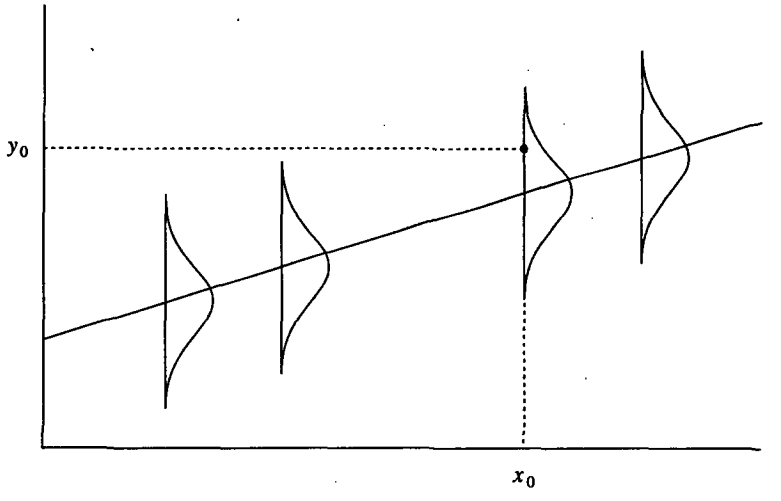
Suppose we are required to calibrate a new instrument, for example a thermometer. An experiment is conducted, and readings $y_1, y_2, \ldots, y_n$ on the new instrument are taken at *preselected, known* temperatures $x_1, x_2, \ldots, x_n$; thus the data are obtained by sampling with preselected X values. We suppose that assumptions (A) are valid where the population regression function is

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

The data are $(y_1, x_1), \ldots, (y_n, x_n)$. From the data we compute $\hat{\beta}_0, \hat{\beta}_1, \hat{\mu}_Y(x)$, and $\hat{\sigma}$ by the formulas in Chapter 3. To use the new instrument, we need to observe a reading y_0 on it and determine the true temperature x_0. Actually we do this every-time we read any gauge, not just a thermometer. Another way to view this is as follows: let $y(x_0)$ denote the value of a random observation (i.e., a reading on a new thermometer to be calibrated) from a subpopulation determined by the unknown value x_0 (unknown *true* temperature). We assume that the subpopulation from which $y(x_0)$ was obtained is Gaussian with unknown mean $\mu_Y(x_0)$ and unknown standard deviation σ. The problem is to determine x_0, the X value for the subpopulation from

which $y(x_0)$ was selected; i.e., to predict the true temperature x_0 when the reading on the new instrument is $y(x_0)$ (see Figure 6.4.1.).

FIGURE 6.4.1



The estimated value for $Y(x_0)$ at the point x_0 (unknown) based on sample data is given by

$$\hat{Y}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 \tag{6.4.1}$$

Assuming that $\hat{\beta}_1 \neq 0$ and solving (6.4.1) for x_0, we get

$$x_0 = \frac{\hat{Y}(x_0) - \hat{\beta}_0}{\hat{\beta}_1} \tag{6.4.2}$$

If the observed value y_0 is substituted for $\hat{Y}(x_0)$ in (6.4.2), we get the point estimate of x_0 as

$$\hat{x}_0 = \frac{y_0 - \hat{\beta}_0}{\hat{\beta}_1} \tag{6.4.3}$$

To obtain a $1-\alpha$ confidence region for x_0, we compute the quantities in Box 6.4.1.

BOX 6.4.1 Confidence Region Computations for Calibration Problems

Carry out steps (1)–(5) below.

1 Compute

$$A = \hat{\beta}_1^2 - \frac{\hat{\sigma}^2 t_{1-\alpha/2:n-2}^2}{SSX}$$

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

2 Compute

$$B = A(1 + \frac{1}{n}) + \frac{[y_0 - \bar{y}]^2}{SSX}$$

3 Compute

$$C = \hat{\beta}_1[y_0 - \bar{y}]$$

4 Compute

$$D = t_{1-\alpha/2:n-2}\hat{\sigma}$$

5 If $A \neq 0$ and $B > 0$, compute L and U where

$$L = \bar{x} + \frac{C - D\sqrt{B}}{A}$$

$$U = \bar{x} + \frac{C + D\sqrt{B}}{A}$$

A $1 - \alpha$ confidence region for x_0 is given by

a	$-\infty < x_0 < \infty$		if $A < 0$ and $B \leq 0$	
b	$-\infty < x_0 \leq U$	and $L \leq x_0 < \infty$	if $A < 0$ and $B > 0$	(6.4.4)
c	$L \leq x_0 \leq U$		if $A > 0$	

If we obtain either (a) or (b) in (6.4.4), then the result will certainly be unsatisfactory because the confidence region is not a finite interval. However, even if we obtain the confidence interval (c), the result may still be unsatisfactory if the width $U - L$ is so large that we cannot make a useful decision about x_0. Note that the confidence interval in (c) results if and only if $A > 0$. This condition will hold when

$$\left[\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\right]^2 > t_{1-\alpha/2:n-2}^2$$

which in turn implies that the $1 - \alpha$ confidence interval for β_1 does not include zero, and we have $1 - \alpha$ confidence that $\beta_1 \neq 0$. If $A \leq 0$, then this implies that a $1 - \alpha$ confidence interval for β_1 will include zero. This does not necessarily imply that $\beta_1 = 0$, but based on sample data we are unable to rule out this possibility at the $1 - \alpha$ confidence level. Clearly if β_1 is indeed 0, then the regression function of Y on X is $\mu_Y(x) = \beta_0$; i.e., x is not in the regression function, so y_0 does not contain any information about x_0. If the result is (a) or (b) in (6.4.4), then either the sample

size is too small and a larger sample must be obtained to get a confidence interval as in (c), or β_1 is indeed zero in $\mu_Y(x) = \beta_0 + \beta_1 x$ and consequently y_0 does not contain any information about x_0. This latter situation does not occur in practice if the investigator knows, a priori, that $\beta_1 \neq 0$ in the regression function of Y on X.

The following example illustrates the computations described in Box 6.4.1.

EXAMPLE 6.4.3

A mercury-in-glass thermometer that is designed to measure temperatures between 95°F and 110°F is being calibrated to determine whether it is reliable enough for a physician to use. It is placed in a water bath at a known constant temperature X, and the corresponding thermometer reading Y is recorded. This is repeated at several known temperatures. The data from the calibration experiment are given in Table 6.4.1 and are also stored in the file **thermom.dat** on the data disk.

Clearly the data are obtained by sampling with preselected X values, and we suppose that assumptions (A) hold.

The thermometer is later used to measure the temperature of a patient in a hospital's emergency room. We want to estimate the true temperature x_0 of the patient if the thermometer reads $y_0 = 104$. We will also obtain a 95% two-sided confidence interval for the true temperature, x_0. MINITAB output from a regression analysis of Y on X is given in Exhibit 6.4.1.

From Table T-2 in Appendix T we get $t_{0.975:6} = 2.447$, and from the data we compute $\bar{x} = 103$, $\bar{y} = 102.975$, and $SSX = 168$. From the computer output we obtain

$$\hat{\beta}_0 = -3.140 \qquad \hat{\beta}_1 = 1.03024 \qquad \hat{\sigma} = 0.2344$$

Thus

$$\hat{Y}(x) = -3.140 + 1.03024x$$

TABLE 6.4.1

Thermometer Calibration Data

Observation Number	Thermometer Reading Y	Known Temperature X
1	95.71	96
2	98.16	98
3	99.52	100
4	102.09	102
5	103.79	104
6	106.18	106
7	108.14	108
8	110.21	110

EXHIBIT 6.4.1
MINITAB Output for Example 6.4.3

```
The regression equation is
reading = - 3.14 + 1.03 knowntmp

Predictor       Coef      Stdev    t-ratio        p
Constant      -3.140      1.865      -1.68    0.143
knowntmp     1.03024    0.01809      56.96    0.000

s = 0.2344      R-sq = 99.8%     R-sq(adj) = 99.8%
```

and

$$\hat{x}_0 = \frac{104 - (-3.140)}{1.03024} = 103.995$$

To compute a 95% confidence region for x_0, we follow the instructions in Box 6.4.1. From steps (1)–(4) of Box 6.4.1 we get

$$A = 1.0594 \qquad B = 1.1981 \qquad C = 1.0560 \qquad D = 0.57358$$

Since $A \neq 0$ and $B > 0$, we compute L and U as described in step (5) of Box 6.4.1 and obtain

$$L = 103.4 \quad \text{and} \quad U = 104.6$$

Since $A > 0$, case (c) in Box 6.4.1 applies, and we get a 95% confidence interval for x_0 as

$$C[103.4 \leq x_0 \leq 104.6] = .95$$

Thus a physician can be 95% confident that the patient's temperature is between 103.4 and 104.6 degrees. ■

Regulation

This application is very similar to the calibration problem except that we want to determine x_0 corresponding to a specified *average* value of Y, i.e., corresponding to a specified value of $\mu_Y(x_0)$ denoted by m_0. We suppose that assumptions (A) hold with $\mu_Y(x) = \beta_0 + \beta_1 x$, and data are obtained by sampling with preselected X values.

Example 6.4.2 illustrates one such situation. Consider the following question: For how many weeks (i.e., what is the value of x_0?) should one feed the supplement to the chickens so that the *average* weight gain for the entire subpopulation of chickens is, say, 10 pounds? Here $m_0 = 10$ and x_0 is the unknown quantity to be estimated. Note that we are not trying to predict the value of x_0 corresponding to the observed weight y_0 of a single chicken; if we were, then we would apply the results from *calibration* just discussed. What we want here is to *regulate* the average

weight gain of chickens by feeding them the supplement for an appropriate number of weeks, viz., x_0 weeks.

We estimate the value of x_0 corresponding to the specified average value m_0 of Y by setting $\mu_Y(x_0)$ equal to m_0 in the regression function $\mu_Y(x_0) = \beta_0 + \beta_1 x_0$ and solving for x_0. This gives

$$x_0 = \frac{m_0 - \beta_0}{\beta_1}$$

Then we substitute the estimates of β_0 and β_1, obtained by formulas in Chapter 3, and get $\hat{x}_0$, which is

$$\hat{x}_0 = \frac{m_0 - \hat{\beta}_0}{\hat{\beta}_1} \qquad \textbf{(6.4.5)}$$

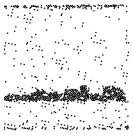
Thus the estimate of x_0 is the same as in (6.4.3) except m_0 takes the place of y_0. A confidence region for x_0 is computed by following the instructions for calibration problems given in Box 6.4.1, except B and C are slightly modified. The B and C values to use in Box 6.4.1, for *regulation* problems are

$$B = \frac{A}{n} + \frac{(m_0 - \bar{y})^2}{SSX} \quad \text{and} \quad C = \hat{\beta}_1(m_0 - \bar{y}) \qquad \textbf{(6.4.6)}$$

In summary, the results in this section are used when:

1. The unknown value x_0 of X is to be predicted corresponding to an observed value y_0 of Y (calibration) or for a specified value m_0 of the average Y value for the subpopulation with $X = x_0$ (regulation).
2. The regression function is $\mu_Y(x) = \beta_0 + \beta_1 x$ and $\beta_1 \neq 0$.
3. Assumptions (A) are presumed to be valid and the data are obtained by sampling with preselected X values.
4. Remember, if assumptions (B) are satisfied, then the data are collected by simple random sampling. Hence, to estimate and to obtain confidence intervals for x_0 corresponding to an observed value y_0 of Y, you simply interchange Y and X and use the standard procedures in Chapter 3 for regressing X on Y. In this case x_0 is in fact $X(y_0)$, a randomly chosen x value from the subpopulation determined by $Y = y_0$.

Note One-sided $1 - \alpha/2$ confidence bounds for x_0 cannot be obtained by using only the upper bound or only the lower bound from the $1 - \alpha$ confidence region for x_0 discussed in this section.



Problems 6.4

6.4.1 The temperature of a reaction chamber is regulated by adjusting a circular dial that has markings running from 0 to 100 engraved on it. The relationship between X, the

dial readings, and Y, the temperature of the reaction chamber, is given by

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

To examine this relationship an experiment is conducted in which the dial is set at various preselected levels and the actual temperature of the reaction chamber is measured. Assumptions (A) are presumed to be valid where the data are collected by sampling with preselected X values. The data are shown in Table 6.4.2 and are also stored in the file **chamber.dat** on the data disk.

The computer output from a regression of Y on X for these data is given in Exhibit 6.4.2. If it is desired to have the *average* temperature of the reaction chamber to be 400°F, compute the estimate of x_0, the dial setting that would achieve this objective.

TABLE 6.4.2
Reaction Chamber Data.

Observation Number	Chamber Temperature Y(°F)	Dial Setting X
1	206.36	0
2	225.52	10
3	252.18	20
4	289.33	30
5	318.11	40
6	349.49	50
7	383.03	60
8	410.70	70
9	444.40	80
10	469.14	90
11	501.16	100

EXHIBIT 6.4.2
MINITAB Output for Problem 6.4.1

```
The regression equation is
chambtmp = 198 + 3.03 dialset

Predictor       Coef      Stdev    t-ratio       p
Constant     198.456      2.282      86.98   0.000
dialset      3.02982    0.03857      78.56   0.000

s = 4.045       R-sq = 99.9%     R-sq(adj) = 99.8%
```

6.4.2 In Problem 6.4.1 compute a 99% confidence region for x_0.

6.4.3 In Example 6.4.3 suppose that a patient's temperature is measured to be 100°. What is the estimate of her actual temperature?

6.4.4 In Problem 6.4.3 obtain a 90% confidence region for her actual temperature.

6.4.5 Consider Task 3.4.1 where crystalline forms of certain chemical compounds are used in various electronic devices and it is often more desirable to have large crystals 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 Y, the weight in grams of a crystal, and X, the time in hours taken for 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. Assumptions (A) are presumed to be valid where the X values are preselected, and the regression function of Y on X is

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

The data are given in Table 3.4.2. For convenience, they are listed below and are also stored in the file **crystal.dat** on the data disk. A computer output for the regression of Y on X is given in Exhibit 6.4.3.

Crystal Data of Table 3.4.2

Crystal Number	Weight Y (grams)	Time X (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

A company orders one crystal that must weigh approximately 5 grams and wants to know what the cost will be. In order for the manufacturer to determine what to charge for the crystal, it wants to know the value of x_0 such that the average weight of crystals that are grown x_0 hours is 5 grams. Estimate the value of x_0.

6.4.6 In Problem 6.4.5 obtain a 90% confidence region for x_0.

6.4.7 In Problem 3.5.1 we considered a coal burning power plant located at a distance of 25 miles from a national park. The emissions from the power plant contain the gas

EXHIBIT 6.4.3
MINITAB Output for Problem 6.4.5

```
The regression equation is
weight = 0.001 + 0.503 time

Predictor       Coef      Stdev    t-ratio       p
Constant      0.0014     0.5994       0.00   0.998
time         0.50343    0.03520      14.30   0.000

s = 1.062       R-sq = 94.5%     R-sq(adj) = 94.0%

Analysis of Variance

SOURCE       DF      SS        MS        F       p
Regression    1   230.63    230.63   204.58   0.000
Error        12    13.53      1.13
Total        13   244.16
```

sulfur dioxide (SO_2), which is linked to acid rain. A certain fraction of the emitted SO_2 will be transported through the atmosphere to the national park. There is always a certain amount of background SO_2 that is present at the national park that is not emitted by the power plant. In order to assess the SO_2 contribution by the power plant to the national park, the SO_2 output X by the plant, in tons/hour, as well as the SO_2 concentrations Y at the national park, in micrograms/cubic meter, were recorded at various randomly selected times during a particular year. The data are given in Table 3.5.4 and are reproduced here for convenience. They are also stored in the file **SO2.dat** on the data disk.

Power Plant SO_2 Data

Observation Number	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

From these data we compute the following:

$$\sum_{i=1}^{14} x_i = 55.76 \qquad \sum_{i=1}^{14} y_i = 123.35$$

$$\sum_{i=1}^{14} x_i^2 = 253.9072 \qquad \sum_{i=1}^{14} y_i^2 = 1220.2711$$

$$\sum_{i=1}^{14} x_i y_i = 549.3552$$

In similar investigations assumptions (B) have been used so we presume they are valid for this problem. On a certain day, the SO_2 concentration measured at the park was $y_0 = 10.5$ micrograms/cubic meter. Predict the SO_2 emission rate, x_0, by the power plant on this day. Should you consider the model

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

or the model

$$\mu_X(y) = \beta_0^* + \beta_1^* y$$

Note the assumptions!

6.4.8 *In Problem 6.4.7 compute a 90% confidence region for x_0. Note the assumptions and* use the proper model to compute the confidence interval.

6.5 Comparison of Several Straight Line Regressions—Identical, Parallel, and Intersecting Lines

In many applied problems we want to compare two or more population regression functions. Consider, for instance, the regression function relating annual salary (Y) and the number of years of experience (X) for computer programmers in California. We may want to examine the regression functions for female and male programmers separately. We may also want to compare the two regression functions and study the differences, if any, between them. In this section we discuss the comparison of several straight line regression functions. Questions of practical interest can often be answered by determining how much the lines differ in slope, intercept, or both. Then, based on these differences, practical decisions can be made. The procedures discussed in this section are sometimes referred to as *regression using dummy variables*. We illustrate with two examples.

EXAMPLE 6.5.1

A study is conducted to determine the relationship between age X and blood pressure Y of individuals between the ages of 30 and 50. It is assumed that the population regression function is a straight line, but both females and males were included in the

study so it was decided that a separate regression function for each was appropriate. The models are given below.

Note To simplify notation in this section, we use α and β instead of β_0 and β_1 for regression coefficients.

$$\text{Females: } \mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x \qquad 30 \le x \le 50$$
$$\text{Males: } \mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x \qquad 30 \le x \le 50$$

We want to study the two models separately and together. In examining the two models together, we want to determine whether the value of α_1 is close enough to the value of α_2, and the value of β_1 is close enough to the value of β_2 so that, for this problem, the relationship between age and blood pressure is considered to be the same for males as it is for females, in the range $30 \le x \le 50$. ■

EXAMPLE 6.5.2

Previous experience indicates that if a commercial fertilizer is applied to wheat fields, the yield Y in bushels per acre is linearly related to X, the amount of fertilizer applied per acre. An experiment is conducted in which three varieties of wheat are used. It is suspected that the fertilizer may affect the yield differently for each variety. So there are three regression functions as given below.

$$\text{Variety 1: } \mu_Y^{(1)} = \alpha_1 + \beta_1 x \qquad 0 \le x \le 4$$
$$\text{Variety 2: } \mu_Y^{(2)} = \alpha_2 + \beta_2 x \qquad 0 \le x \le 4$$
$$\text{Variety 3: } \mu_Y^{(3)} = \alpha_3 + \beta_3 x \qquad 0 \le x \le 4$$

An investigator may be interested in obtaining answers to the following questions.

1 Are the average yields of the three varieties of wheat the same when the same amount of fertilizer is applied to each variety? To help determine this it is sometimes recommended that one test whether the three regression functions are the same. The null hypothesis considered is

$$\text{NH: } \alpha_1 = \alpha_2 = \alpha_3 \quad \text{and} \quad \beta_1 = \beta_2 = \beta_3 \tag{6.5.1}$$

2 If the amount of fertilizer applied to each variety is increased (or decreased) by one unit, is the change in *average* yield the same for each variety? In terms of population parameters, this is true if and only if $\beta_1 = \beta_2 = \beta_3$, so to answer this question it is sometimes recommended that one should test whether or not the three regressions lines have the same slope (i.e., that the three lines are parallel). The null hypothesis considered is

$$\text{NH: } \beta_1 = \beta_2 = \beta_3 \tag{6.5.2}$$

3 Are the average yields for the three varieties the same when a specified amount x_0 of fertilizer is applied? To help answer this question it is sometimes recommended that one should test whether or not the three lines intersect in a common

specified point with $X = x_0$. The null hypothesis for this test is

$$\text{NH: } \mu_Y^{(1)}(x_0) = \mu_Y^{(2)}(x_0) = \mu_Y^{(3)}(x_0) \tag{6.5.3}$$

which is equivalent to

$$\text{NH: } \alpha_1 + \beta_1 x_0 = \alpha_2 + \beta_2 x_0 = \alpha_3 + \beta_3 x_0 \tag{6.5.4}$$

If the specified point x_0 is 0, then one would be testing whether or not the regression lines have the same intercept (i.e., whether or not the three varieties have the same average yield if no fertilizer is applied).

4 Most quantities of interest in applied problems such as these can be represented as linear combinations of the α_i and β_j given by

$$\boldsymbol{d}^T\boldsymbol{\beta} = \sum_{i=1}^{3}(a_i\alpha_i + b_i\beta_i) = a_1\alpha_1 + b_1\beta_1 + a_2\alpha_2 + b_2\beta_2 + a_3\alpha_3 + b_3\beta_3 \tag{6.5.5}$$

where $\boldsymbol{d}^T = [a_1, b_1, a_2, b_2, a_3, b_3]$, $\boldsymbol{\beta}^T = [\alpha_1, \beta_1, \alpha_2, \beta_2, \alpha_3, \beta_3]$, and the a_i and b_i are specified constants. For example, suppose the investigator is interested in the difference between the average yields of variety 1 and variety 2 when 1.5 units of fertilizer per acre are applied to fields growing each variety. This means that the investigator wants to examine

$$\begin{aligned}\mu_Y^{(1)}(1.5) - \mu_Y^{(2)}(1.5) &= (\alpha_1 + 1.5\beta_1) - (\alpha_2 + 1.5\beta_2) \\ &= \alpha_1 + 1.5\beta_1 - \alpha_2 - 1.5\beta_2 = \boldsymbol{d}^T\boldsymbol{\beta}\end{aligned}$$

In this case $\boldsymbol{d}^T = [1, 1.5, -1, -1.5, 0, 0]$. Suppose, instead, the investigator is interested in the difference between the average yields of varieties 2 and 3 when 1.5 units of fertilizer per acre are applied to fields where variety 2 is grown, and 3.0 units of fertilizer per acre are applied to fields where variety 3 is grown. In this case the investigator is interested in

$$\mu_Y^{(2)}(1.5) - \mu_Y^{(3)}(3.0) = \alpha_2 + 1.5\beta_2 - \alpha_3 - 3.0\beta_3 = \boldsymbol{d}^T\boldsymbol{\beta}$$

which is the linear combination in (6.5.5) with $\boldsymbol{d}^T = [0, 0, 1, 1.5, -1, -3.0]$. ■

As stated above, statistical tests (of hypotheses) are often used to determine whether regression lines are identical, parallel, or intersect in a common (specified) point. We, however, recommend that tests not be used for these purposes because, as explained earlier in several places, tests alone do not give much information. Not only do tests give very little information, but it is inconceivable that the three varieties of wheat will have *exactly* the same average yield or have exactly the same values of α_i, β_i, etc. So rather than ask if the three lines are exactly identical, or have exactly the same slopes or intercepts, it seems more useful to consider the question: "What are the differences in the average yields, or what are the differences in the slopes or intercepts, of *each pair* of varieties?" With this information, an investigator can decide whether these differences are close enough to zero to be considered neglible for the problem under study. In fact, one would generally want to make a full examination of the regression lines by examining all *pairs* to see how different the lines are. This full examination requires several decisions, and one may

want to have a specified confidence, say $1 - \alpha$, that *collectively* all of the decisions are correct. This can be done by using simultaneous confidence intervals, and will be explained next.

To illustrate procedures for answering the questions just discussed, we can consider an arbitrary number (say H) of regression lines, but to simplify notation we let $H = 3$ and consider only three lines. You should have no difficulty in extending the results to any value of H.

Suppose samples of sizes n_1, n_2, and n_3 are obtained from study populations 1, 2, and 3, respectively, and suppose that either assumptions (A) or (B) apply for each population. The sample values may be organized as shown in Table 6.5.1.

TABLE 6.5.1

Sample from Population 1		Sample from Population 2		Sample from Population 3	
$y_{1,1}$	$x_{1,1}$	$y_{1,2}$	$x_{1,2}$	$y_{1,3}$	$x_{1,3}$
$y_{2,1}$	$x_{2,1}$	$y_{2,2}$	$x_{2,2}$	$y_{2,3}$	$x_{2,3}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$y_{n_1,1}$	$x_{n_1,1}$	$y_{n_2,2}$	$x_{n_2,2}$	$y_{n_3,3}$	$x_{n_3,3}$

We use the notation σ_h to represent the subpopulation standard deviation in population h for $h = 1, 2, 3$, and we assume that $\sigma_1 = \sigma_2 = \sigma_3$; this common value is denoted by σ.

The three straight line regression functions are

$$\begin{aligned} &\text{Model 1:} \quad \mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x \qquad a \le x \le b \\ &\text{Model 2:} \quad \mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x \qquad a \le x \le b \\ &\text{Model 3:} \quad \mu_Y^{(3)}(x) = \alpha_3 + \beta_3 x \qquad a \le x \le b \end{aligned} \tag{6.5.6}$$

To examine the three regression functions individually, we can write

$$\boldsymbol{y}_h = \boldsymbol{X}_h \boldsymbol{\beta}_h + \boldsymbol{e}_h \qquad h = 1, 2, 3 \tag{6.5.7}$$

where

$$\boldsymbol{y}_h = \begin{bmatrix} y_{1,h} \\ y_{2,h} \\ \vdots \\ y_{n_h,h} \end{bmatrix} \qquad \boldsymbol{X}_h = \begin{bmatrix} 1 & x_{1,h} \\ 1 & x_{2,h} \\ \vdots & \vdots \\ 1 & x_{n_h,h} \end{bmatrix} \qquad \boldsymbol{\beta}_h = \begin{bmatrix} \alpha_h \\ \beta_h \end{bmatrix}$$

and

$$\boldsymbol{C}_h = (\boldsymbol{X}_h^T \boldsymbol{X}_h)^{-1} \tag{6.5.8}$$

All the procedures in Chapter 3 can be applied to each model separately to obtain point estimates and confidence intervals for α_h, β_h, and σ_h for $h = 1, 2, 3$ by regressing Y on X for each sample.

To examine the three regression functions *collectively* to determine how much they differ in slope, intercept, or both, we can compute confidence intervals for $\alpha_i - \alpha_j$, for $\beta_i - \beta_j$, and for $\mu_Y^{(i)}(x_0) - \mu_Y^{(j)}(x_0^*)$, for all i and j $(i \neq j)$, where x_0 and x_0^* are specified constants which may or may not be the same. These differences can all be written as special cases of the linear function $\boldsymbol{d}^T\boldsymbol{\beta}$ in (6.5.5).

A confidence interval for $\boldsymbol{d}^T\boldsymbol{\beta}$ is

$$\boldsymbol{d}^T\hat{\boldsymbol{\beta}} - (\text{table-value})SE(\boldsymbol{d}^T\hat{\boldsymbol{\beta}}) \leq \boldsymbol{d}^T\boldsymbol{\beta} \leq \boldsymbol{d}^T\hat{\boldsymbol{\beta}} + (\text{table-value})SE(\boldsymbol{d}^T\hat{\boldsymbol{\beta}}) \tag{6.5.9}$$

where

$$SE(\boldsymbol{d}^T\hat{\boldsymbol{\beta}}) = \hat{\sigma}\sqrt{\boldsymbol{d}^T\boldsymbol{C}\boldsymbol{d}} \tag{6.5.10}$$

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_1 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{C}_2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{C}_3 \end{bmatrix}$$

and $\boldsymbol{C}_h$ is given in (6.5.8). Also the estimate of σ is given by

$$\begin{aligned} \hat{\sigma} &= \sqrt{\frac{(n_1-2)\hat{\sigma}_1^2 + (n_2-2)\hat{\sigma}_2^2 + (n_3-2)\hat{\sigma}_3^2}{(n_1-2)+(n_2-2)+(n_3-2)}} \\ &= \sqrt{\frac{SSE(1)+SSE(2)+SSE(3)}{(n_1-2)+(n_2-2)+(n_3-2)}} \end{aligned} \tag{6.5.11}$$

where $SSE(h)$ denotes the **sum of squared errors** for group h, for $h = 1, 2, 3$. The numerator of the expression under the square root sign in (6.5.11) is called the *pooled sum of squared errors* and is denoted by $SSE(pooled)$. More generally, when H straight line regressions are being compared, we have

$$\begin{aligned} SSE(pooled) &= \sum_{h=1}^{H}(n_h-2)\hat{\sigma}_h^2 = (n_1-2)\hat{\sigma}_1^2 + \cdots + (n_H-2)\hat{\sigma}_H^2 \\ &= SSE(1) + \cdots + SSE(H) \end{aligned} \tag{6.5.12}$$

The quantity

$$(n_1-2)+(n_2-2)+(n_3-2) = \sum_{h=1}^{3}(n_h-2) = n-2(3) = n-6$$

(where $n = n_1 + n_2 + n_3$) in the denominator of the expression under the square root in (6.5.11) is called the *pooled degrees of freedom* (written df $(pooled)$) for estimating σ. More generally, when there are H straight regressions being compared, the pooled degrees of freedom for estimating σ is given by

$$df(pooled) = (n_1-2) + \cdots + (n_H-2) = n-2H \tag{6.5.13}$$

where $n = n_1 + \cdots + n_H$ is the total number of observations. The quantities $\hat{\alpha}_h$, $\hat{\beta}_h$, $\hat{\sigma}_h$, and $\boldsymbol{C}_h$ can be obtained by regressing Y on X for each sample.

For a confidence interval on a *single* linear function $d^T\beta$ in (6.5.5), the table-value is $t_{1-\alpha/2:df(pooled)}$ when H straight lines are being compared ($H = 3$ for the case we are considering here).

If confidence intervals are desired for m distinct linear combinations $d_i^T\beta$, for $i = 1, \ldots, m$ such that one has at least $1 - \alpha$ confidence that *all m intervals are simultaneously correct*, then one procedure uses (6.5.9) with

$$\text{table-value} = t_{1-\alpha/2m:df(pooled)} \tag{6.5.14}$$

which is the $1 - \alpha/2m$ percentile of a student's t population with $df(pooled) = n - 2H$ degrees of freedom, which we denote by v. For other procedures see [23].

Note For obtaining the value of m in (6.5.14), two linear combinations of parameters, say θ_1 and θ_2, *are distinct* if there do not exist constants a and b such that $\theta_1 = a\theta_2 + b$. For instance, the linear combinations $\theta_1 = \beta_1 - \beta_2$ and $\theta_2 = 2\beta_1 - 2\beta_2$ are *not* distinct. Likewise, $\theta_1 = \beta_1 - \beta_2$ and $\theta_2 = 3(\beta_2 - \beta_1) + 6$ are not distinct. On the other hand, the linear combinations $\theta_1 = \alpha_1 - \beta_1$ and $\theta_2 = \alpha_1 - \beta_3$ are distinct. We leave these for you to verify.

The table-values $t_{1-\alpha/2m:v}$ are in Table T-4 in Appendix T for $m = 2, \ldots, 6$. The confidence statements for $d_i^T\beta$ for $i = 1, 2, \ldots, m$ are

$$C\left[\begin{array}{c} d_i^T\hat{\beta} - t_{1-\alpha/2m:v}SE(d_i^T\hat{\beta}) \le d_i^T\beta \le d_i^T\hat{\beta} + t_{1-\alpha/2m:v}SE(d_i^T\hat{\beta}) \\ \text{simultaneously for all } i = 1, \ldots, m \end{array}\right] \ge 1 - \alpha \tag{6.5.15}$$

If $m = 1$, then in (6.5.15) the $\ge$ sign is replaced with the = sign. We can use (6.5.15) with $m = 6$ to obtain confidence intervals for

$$\alpha_1 - \alpha_2 \qquad \alpha_1 - \alpha_3 \qquad \alpha_2 - \alpha_3 \qquad \beta_1 - \beta_2 \qquad \beta_1 - \beta_3 \qquad \beta_2 - \beta_3 \tag{6.5.16}$$

for the three regression functions in (6.5.6), and we have confidence greater than or equal to $1 - \alpha$ that all *six* of the intervals are correct. By examining the confidence intervals for these six quantities, an investigator can determine how much any two of the regression lines differ in slope, intercept, or both.

To determine the difference between two straight line regression functions, say $\mu_Y^{(1)}(x)$ and $\mu_Y^{(2)}(x)$ for x values in the range $a \le x \le b$, it may be useful to compute simultaneous confidence intervals for the differences $\mu_Y^{(1)}(a) - \mu_Y^{(2)}(a)$ and $\mu_Y^{(1)}(b) - \mu_Y^{(2)}(b)$. If both these differences are acceptably small (at a given confidence level), then the investigator may conclude that the two lines can be considered to be equivalent for the problem under consideration. For instance, if for some constant d we have

$$|\mu_Y^{(1)}(a) - \mu_Y^{(2)}(a)| \le d$$

and

$$|\mu_Y^{(1)}(b) - \mu_Y^{(2)}(b)| \le d$$

where $a < b$, then the two regression functions differ by less than d units for all x in the interval $a \le x \le b$ (see Figures 6.5.1 and 6.5.2).

FIGURE 6.5.1

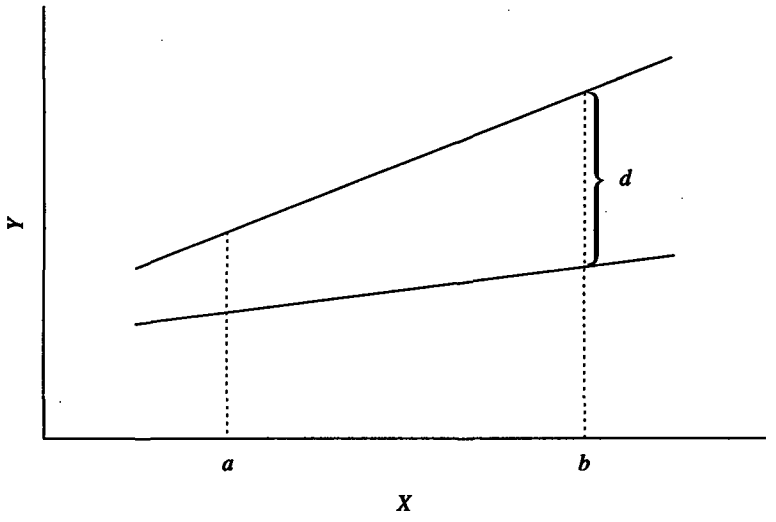
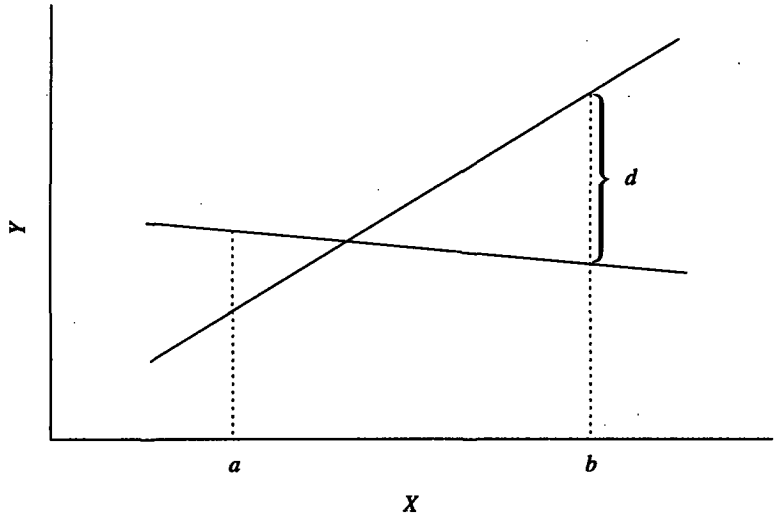


FIGURE 6.5.2



We illustrate the procedures discussed in this section with Example 6.5.3.

EXAMPLE 6.5.3

Various amounts X of a new food supplement were fed to three different breeds of chickens for 6 weeks to determine its effect on the hardness Y of egg shells. A straight line regression function is assumed to hold for each breed. The three regression functions are

$$\text{Breed 1:}\quad \mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x \qquad 0 \le x \le 20$$
$$\text{Breed 2:}\quad \mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x \qquad 0 \le x \le 20$$
$$\text{Breed 3:}\quad \mu_Y^{(3)}(x) = \alpha_3 + \beta_3 x \qquad 0 \le x \le 20$$

The problem is to study these regression functions together and separately to determine how they differ in slope (change in average hardness of egg shells per unit of the new food supplement used), in intercept (average hardness of egg shells for each breed if the new food supplement is not used), or both. Assumptions (A) are presumed to hold, and the data are obtained by sampling with preselected X values. The investigator wants to be at least 95% confident that the decisions are simultaneously correct. The data are given in Table 6.5.2 and are also stored in the file **eggshell.dat** on the data disk.

A regression analysis is carried out for each breed and the results are summarized in the MINITAB output in Exhibit 6.5.1. From these data we get the following quantities:

$$\hat{\alpha}_1 = 5.9662 \quad \hat{\beta}_1 = 3.0494 \quad \hat{\sigma}_1 = 1.330 \quad SSE(1) = 17.7$$
$$\hat{\alpha}_2 = 6.4673 \quad \hat{\beta}_2 = 1.0948 \quad \hat{\sigma}_2 = 0.9015 \quad SSE(2) = 4.876$$
$$\hat{\alpha}_3 = 5.0225 \quad \hat{\beta}_3 = 0.26344 \quad \hat{\sigma}_3 = 1.284 \quad SSE(3) = 11.538$$

TABLE 6.5.2

Eggshell Data

Observation Number	Breed 1		Breed 2		Breed 3	
	Y_1	X_1	Y_2	X_2	Y_3	X_3
1	8.42	1	9.86	3	6.52	2
2	14.68	3	9.54	3	5.11	5
3	21.42	5	11.96	4	7.75	7
4	25.45	6	12.46	5	6.84	8
5	27.14	7	11.38	6	7.65	10
6	30.53	8	14.69	8	9.49	15
7	34.51	9	16.48	9	7.03	16
8	34.52	9	20.11	12	9.41	18
9	33.24	10			12.01	20
10	39.63	11				
11	43.98	12				
12	47.77	14				

EXHIBIT 6.5.1

MINITAB Output for Example 6.5.3

REGRESSION ANALYSIS FOR BREED 1

The regression equation is
Y1 = 5.97 + 3.05 X1

Predictor	Coef	Stdev	t-ratio	p
Constant	5.9662	0.9289	6.42	0.000
X1	3.0494	0.1068	28.54	0.000

s = 1.330 R-sq = 98.8% R-sq(adj) = 98.7%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	1440.6	1440.6	814.61	0.000
Error	10	17.7	1.8		
Total	11	1458.3			

MATRIX C1

0.487896719	-0.051102743
-0.051102743	0.006455083

REGRESSION ANALYSIS FOR BREED 2

The regression equation is
Y2 = 6.47 + 1.09 X2

Predictor	Coef	Stdev	t-ratio	p
Constant	6.4673	0.7386	8.76	0.000
X2	1.0948	0.1066	10.27	0.000

s = 0.9015 R-sq = 94.6% R-sq(adj) = 93.7%

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	1	85.703	85.703	105.47	0.000
Error	6	4.876	0.813		
Total	7	90.579			

MATRIX C2

0.67132867	-0.08741259
-0.08741259	0.01398601

EXHIBIT 6.5.1
(Continued)

```
                REGRESSION ANALYSIS FOR BREED 3

The regression equation is
Y3 = 5.02 + 0.263 X3

Predictor       Coef       Stdev    t-ratio        p
Constant      5.0225      0.9193       5.46    0.000
X3           0.26344     0.07250       3.63    0.008

s = 1.284       R-sq = 65.4%     R-sq(adj) = 60.4%

Analysis of Variance

SOURCE       DF          SS          MS          F        p
Regression    1      21.761      21.761      13.20    0.008
Error         7      11.538       1.648
Total         8      33.298

 MATRIX C3

 0.512756910   -0.035790220
-0.035790220    0.003189227
```

To get the point estimate of σ we can use the formula in (6.5.11) to *pool* the estimates $\hat{\sigma}_h$. We get

$$\hat{\sigma} = \sqrt{\frac{17.7 + 4.876 + 11.538}{10 + 6 + 7}} = 1.218 \tag{6.5.17}$$

Using (6.5.10), we get

$$SE(\hat{\alpha}_1 - \hat{\alpha}_2) = 1.218\sqrt{0.487897 + 0.671329} = 1.311$$
$$(\text{note } \boldsymbol{d}^T = [1, 0, -1, 0, 0, 0])$$
$$SE(\hat{\alpha}_1 - \hat{\alpha}_3) = 1.218\sqrt{0.487897 + 0.512757} = 1.218$$
$$(\text{note } \boldsymbol{d}^T = [1, 0, 0, 0, -1, 0])$$
$$SE(\hat{\alpha}_2 - \hat{\alpha}_3) = 1.218\sqrt{0.671329 + 0.512757} = 1.325$$
$$(\text{note } \boldsymbol{d}^T = [0, 0, 1, 0, -1, 0])$$
$$SE(\hat{\beta}_1 - \hat{\beta}_2) = 1.218\sqrt{0.006455 + 0.013986} = 0.1741$$
$$(\text{note } \boldsymbol{d}^T = [0, 1, 0, -1, 0, 0])$$
$$SE(\hat{\beta}_1 - \hat{\beta}_3) = 1.218\sqrt{0.006455 + 0.003189} = 0.1196$$
$$(\text{note } \boldsymbol{d}^T = [0, 1, 0, 0, 0, -1])$$
$$SE(\hat{\beta}_2 - \hat{\beta}_3) = 1.218\sqrt{0.013986 + 0.003189} = 0.1596$$
$$(\text{note } \boldsymbol{d}^T = [0, 0, 0, 1, 0, -1]).$$

To compute the appropriate table-value we observe that $1-\alpha = 0.95$, $\alpha = 0.05$, $m = 6$, and $df(pooled) = \nu = n_1 + n_2 + n_3 - 6 = 23$. From Table T-4 in Appendix T the required table-value is $t_{1-\alpha/2m:\nu} = t_{0.995833:23} = 2.886$. So the confidence statement is

$$C\begin{bmatrix} -4.285 \le \alpha_1 - \alpha_2 \le 3.283 \\ -2.572 \le \alpha_1 - \alpha_3 \le 4.459 \\ -2.379 \le \alpha_2 - \alpha_3 \le 5.269 \\ 1.452 \le \beta_1 - \beta_2 \le 2.457 \\ 2.441 \le \beta_1 - \beta_3 \le 3.131 \\ 0.371 \le \beta_2 - \beta_3 \le 1.292 \end{bmatrix} \ge 0.95$$

We have at least 95% confidence that all six of these intervals are correct. On the basis of these, an investigator can decide, for the problem at hand, how the three lines, or any two of them, differ in slope, intercept, or both.

Suppose the investigator is willing to consider the regression lines to be equivalent for this problem provided that they do not differ by more than ten units in the interval $0 \le x \le 20$. We can compute simultaneous confidence intervals (say, with confidence coefficient equal to 0.95) for the differences

$$\begin{aligned}
\mu_Y^{(1)}(0) - \mu_Y^{(2)}(0) &\quad \text{which is } \alpha_1 - \alpha_2 \\
\mu_Y^{(1)}(0) - \mu_Y^{(3)}(0) &\quad \text{which is } \alpha_1 - \alpha_3 \\
\mu_Y^{(2)}(0) - \mu_Y^{(3)}(0) &\quad \text{which is } \alpha_2 - \alpha_3 \\
\mu_Y^{(1)}(20) - \mu_Y^{(2)}(20) &\quad \text{which is } \alpha_1 + 20\beta_1 - \alpha_2 - 20\beta_2 \\
\mu_Y^{(1)}(20) - \mu_Y^{(3)}(20) &\quad \text{which is } \alpha_1 + 20\beta_1 - \alpha_3 - 20\beta_3 \\
\mu_Y^{(2)}(20) - \mu_Y^{(3)}(20) &\quad \text{which is } \alpha_2 + 20\beta_2 - \alpha_3 - 20\beta_3
\end{aligned}$$

You should verify the following confidence statement.

$$C\begin{bmatrix} -4.285 \le \mu_Y^{(1)}(0) - \mu_Y^{(2)}(0) \le 3.283 \\ -2.572 \le \mu_Y^{(1)}(0) - \mu_Y^{(3)}(0) \le 4.459 \\ -2.379 \le \mu_Y^{(2)}(0) - \mu_Y^{(3)}(0) \le 5.269 \\ 31.743 \le \mu_Y^{(1)}(20) - \mu_Y^{(2)}(20) \le 45.439 \\ 52.5297 \le \mu_Y^{(1)}(20) - \mu_Y^{(3)}(20) \le 60.7961 \\ 11.8571 \le \mu_Y^{(2)}(20) - \mu_Y^{(3)}(20) \le 24.2869 \end{bmatrix} \ge 0.95$$

Using this information the investigator would perhaps conclude:

a When no supplement is added, the average hardnesses differ by no more than 5.269 units, which is of no practical importance (less than 10 units) for this problem.

b When 20 units of the supplement are used, the differences between the average hardnesses are greater than 10 units for each pair of breeds, which *is* of practical importance for this problem.

c The three regression lines are not equivalent for values of x in the range from 0 to 20 units. ■

In Section 6.5 of the laboratory manuals we discuss computer programs we have written (supplied on the data disk) to carry out the computations discussed in this section.

Another problem of interest in this and similar examples is finding the point where two regression lines intersect. This is discussed in the next section.

Problems 6.5

6.5.1 This is a continuation of Task 3.4.1. Crystals of various sizes are used in electronic devices, and there is a linear relationship between Y, the weight in grams of a crystal, and X, the time in hours it takes the crystal to grow. Three procedures, denoted here as 1, 2, and 3, are used to grow the crystals, and we want to determine how the procedures differ. A straight line regression model is assumed to hold for each procedure with subpopulation standard deviation σ_h for $h = 1, 2, 3$.

$$\text{Procedure 1: } \mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x \qquad 1 \le x \le 30$$
$$\text{Procedure 2: } \mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x \qquad 1 \le x \le 30$$
$$\text{Procedure 3: } \mu_Y^{(3)}(x) = \alpha_3 + \beta_3 x \qquad 1 \le x \le 30$$

The data and a computer output containing relevant regression analyses are given in Exhibit 6.5.2. The data are also stored in the file **crystal3.dat** on the data disk. We suppose that assumptions (A) are valid and that the data are obtained by preselecting the X values.

a Exhibit the estimates of β_h and α_h for $h = 1, 2, 3$.

b Exhibit the estimates of σ_i for $i = 1, 2, 3$.

c Compute the estimate of σ. Assume $\sigma_1 = \sigma_2 = \sigma_3 = \sigma$.

d Exhibit the vector d^T for which $SE(\hat{\beta}_1 - \hat{\beta}_2) = \hat{\sigma}\sqrt{d^T C d}$ in (6.5.10).

e Exhibit the vector d^T for which $SE(\hat{\alpha}_2 - \hat{\alpha}_3) = \hat{\sigma}\sqrt{d^T C d}$ in (6.5.10).

f What value of m should be used to compute simultaneous confidence intervals for

$$\alpha_1 - \alpha_2 \text{ and } \beta_1 - \beta_2$$

such that one has confidence ≥ 0.90 that they are both correct?

g What value of m should be used to compute simultaneous confidence intervals for

$$\alpha_1 - \alpha_3,\ 2\alpha_1 - 2\alpha_3,\ \beta_1 - \beta_3, \text{ and } \beta_1 + \beta_2 - 2\beta_3$$

such that one has confidence ≥ 0.90 that all four intervals are correct?

EXHIBIT 6.5.2

MINITAB Output for Problem 6.5.1

	Procedure 1		Procedure 2		Procedure 3	
obsno	Y1	X1	Y2	X2	Y3	X3
1	0.10	2	0.31	2	2.57	4
2	0.95	3	2.74	4	4.96	6
3	4.79	5	5.93	7	7.23	8
4	5.02	6	7.98	9	8.41	9
5	5.56	7	10.00	11	11.02	11
6	5.79	7	12.36	13	11.31	12
7	6.31	8	14.94	15	12.56	13
8	7.58	9	16.02	16	17.86	17
9	8.19	9	16.87	17	20.14	19
10	9.37	10	19.13	19	32.67	29

REGRESSION ANALYSIS FOR PROCEDURE 1

The regression equation is
Y1 = - 1.88 + 1.10 X1

Predictor	Coef	Stdev	t-ratio	p
Constant	-1.8817	0.4945	-3.80	0.005
X1	1.09814	0.07008	15.67	0.000

s = 0.5536 R-sq = 96.8% R-sq(adj) = 96.5%

The matrix C1 is

```
 0.79807692   -0.10576923
-0.10576923    0.01602564
```

REGRESSION ANALYSIS FOR PROCEDURE 2

The regression equation is
Y2 = - 1.87 + 1.11 X2

Predictor	Coef	Stdev	t-ratio	p
Constant	-1.8710	0.1290	-14.50	0.000
X2	1.10611	0.01030	107.44	0.000

s = 0.1766 R-sq = 99.9% R-sq(adj) = 99.9%

The matrix C2 is

```
 0.534172050   -0.038422305
-0.038422305    0.003400204
```

EXHIBIT 6.5.2
(Continued)

```
              REGRESSION ANALYSIS FOR PROCEDURE 3

The regression equation is
Y3 = - 2.47 + 1.20 X3

Predictor       Coef      Stdev    t-ratio       p
Constant     -2.4721     0.2426     -10.19   0.000
X3           1.19883    0.01666      71.97   0.000

s = 0.3663      R-sq = 99.8%     R-sq(adj) = 99.8%

The matrix C3 is

 0.438792390  -0.026468156
-0.026468156   0.002067825
```

h What value of m should be used to compute simultaneous confidence intervals for

$$\mu_Y^{(1)}(2) - \mu_Y^{(2)}(2) \text{ and } \mu_Y^{(1)}(5) - \mu_Y^{(2)}(5)$$

such that one has confidence $\geq 1 - \alpha$ that they are both correct?

i Compute confidence intervals for the quantities in part (f) such that one has at least 90% confidence that the two intervals are simultaneously correct.

6.5.2 In Problem 6.5.1 suppose that an investigator is interested in comparing the three procedures. She decides that for the problem at hand, the slopes can be considered to be equivalent for her purposes if they differ by less than 0.30 unit in magnitude. She also decides that if the intercepts differ by less than 0.2 unit in magnitude they can be considered to be equivalent for this problem. Make a complete analysis and decide which, if any, of the regression lines are equivalent in intercept, slope, or both. Use an appropriate 90% simultaneous confidence statement to help make this decision. Write a short report explaining your results.

6.5.3 In Example 6.5.2 an investigator wants to determine the difference between the average yields of varieties 1 and 2 if the fields where variety 1 is grown receive no fertilizer and the fields where variety two is grown receive 1.5 units of fertilizer per acre. Exhibit the population parameters for which an investigator would want point estimates and confidence intervals to make this determination.

6.5.4 In Example 6.5.2 suppose there are six varieties to be compared, instead of three, and the investigator wants to examine the differences between each distinct pair of the α_i and each distinct pair of the β_j by obtaining confidence intervals for the differences of all distinct pairs with a simultaneous confidence coefficient of at least 0.95. What

value of m should be used in (6.5.5)? Write out these m distinct quantities $\alpha_i - \alpha_j$ and $\beta_i - \beta_j$.

6.5.5 In Problem 6.5.1 suppose the investigator is willing to consider the population regression lines to be equivalent for practical purposes, provided that they do not differ by more than 2 units for $1 \leq x \leq 30$ hours. To help make a decision, exhibit the population quantities in which the investigator would be interested.

6.6 Intersection of Two Straight Line Regression Functions

As we indicated at the end of the previous section, there are situations when an investigator may want to find the point, say x_0, where two population regression lines intersect. We illustrate one such situation in Example 6.6.1.

EXAMPLE 6.6.1

Consider the problem in Example 2.2.2 where a company is studying the relationship between Y, the first-year maintenance cost of new cars, and X, the number of miles the car will be driven the first year. Suppose the company is interested in cars made by two manufacturers, say manufacturers 1 and 2. It is assumed that the population regression function of Y on X is a straight line for each make of car. The two regression functions are given by

$$\text{Manufacturer 1: } \mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x \qquad a \leq x \leq b$$
$$\text{Manufacturer 2: } \mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x \qquad a \leq x \leq b$$

The investigator wants to determine the value of X, say $X = x_0$, where the two lines intersect because to the right of x_0 one make of car will have lower average first-year maintenance costs, and to the left of x_0 the other make of car will have lower average first-year maintenance costs. If x_0 is outside the interval $[a, b]$, then cars made by one of the manufacturers have lower average first-year maintenance costs than the other everywhere in the interval. For example, in Figure 6.6.1 cars made by manufacturer 1 have lower average first-year maintenance costs if the miles driven is between x_0 and b, and cars made by manufacturer 2 have lower average first-year maintenance costs if the number of miles driven is between a and x_0. In Figure 6.6.2, cars made by manufacturer 1 have lower average first-year maintenance costs in the entire interval from a to b. ■

To find the point x_0 where the two regression lines intersect, we set $\mu_Y^{(1)}(x) = \mu_Y^{(2)}(x)$, solve for x, and denote the solution by x_0. We get

$$x_0 = \frac{\alpha_1 - \alpha_2}{\beta_2 - \beta_1}$$

Thus the two regression lines intersect at $X = x_0$ (we assume $\beta_1 \neq \beta_2$).

To compute point and confidence interval estimates of x_0 we obtain samples from populations 1 and 2. We suppose that assumptions (A) or (B) are valid for each

FIGURE 6.6.1

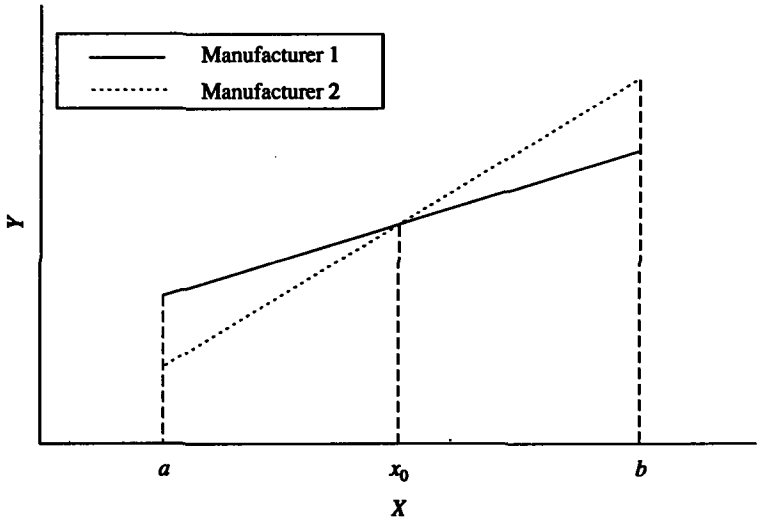
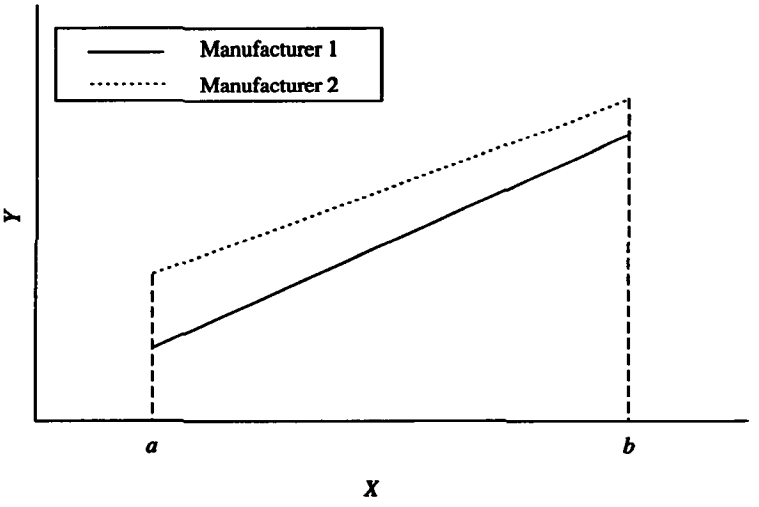


FIGURE 6.6.2



population, and we assume that the two subpopulation standard deviations are the same; i.e., $\sigma_1 = \sigma_2$, with their common value denoted by σ.

Let $(y_{i,1}, x_{i,1})$ for $i = 1, \ldots, n_1$ denote the n_1 sample values from population (1), and let $(y_{i,2}, x_{i,2})$ for $i = 1, \ldots, n_2$ denote the n_2 sample values from population (2). To get a point estimate of x_0, compute estimates of the unknown parameters for each

population by formulas in Chapter 3. The point estimate of x_0 is

$$\hat{x}_0 = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\hat{\beta}_2 - \hat{\beta}_1} \tag{6.6.1}$$

In Box 6.6.1 we give the instructions for computing a $1-\alpha$ confidence region for x_0.

BOX 6.6.1 Instructions for Computing a Confidence Region for x_0, the Point of Intersection of Two Straight Line Regressions

Carry out steps 1–8 below.

1 Compute $\bar{x}_1$, the mean of the X values from sample 1, and $\bar{x}_2$, the mean of the X values from sample 2. Also compute

$$SSX(1) = \sum_{i=1}^{n_1}(x_{i,1} - \bar{x}_1)^2 \qquad Q_1 = \sum_{i=1}^{n_1} x_{i,1}^2$$

and

$$SSX(2) = \sum_{i=1}^{n_2}(x_{i,2} - \bar{x}_2)^2 \qquad Q_2 = \sum_{i=1}^{n_2} x_{i,2}^2$$

2 Compute $SSE(1)$ and $SSE(2)$, the sum of squared errors for samples 1 and 2, respectively.

3 $\hat{\sigma}^2 = \dfrac{(n_1-2)\hat{\sigma}_1^2 + (n_2-2)\hat{\sigma}_2^2}{(n_1-2)+(n_2-2)} = \dfrac{SSE(1)+SSE(2)}{(n_1-2)+(n_2-2)}$

4 $A = [\hat{\beta}_1 - \hat{\beta}_2]^2 - \left[\dfrac{1}{SSX(1)} + \dfrac{1}{SSX(2)}\right]\hat{\sigma}^2 F_{1-\alpha:1,n_1+n_2-4}$

5 $B = [\hat{\alpha}_1 - \hat{\alpha}_2][\hat{\beta}_1 - \hat{\beta}_2] + \left[\dfrac{\bar{x}_1}{SSX(1)} + \dfrac{\bar{x}_2}{SSX(2)}\right]\hat{\sigma}^2 F_{1-\alpha:1,n_1+n_2-4}$

6 $C = (\hat{\alpha}_1 - \hat{\alpha}_2)^2 - \left[\dfrac{Q_1}{n_1 SSX(1)} + \dfrac{Q_2}{n_2 SSX(2)}\right]\hat{\sigma}^2 F_{1-\alpha:1,n_1+n_2-4}$

7 $D = B^2 - AC$

8 If $A \neq 0$ and $D > 0$, compute L and U where $L = -\dfrac{(B+\sqrt{D})}{A}$ and $U = -\dfrac{(B-\sqrt{D})}{A}$.

A $1-\alpha$ confidence region for x_0 is

a	$-\infty < x_0 < \infty$			if $D \leq 0$
b	$-\infty < x_0 \leq U$	and	$L \leq x_0 < \infty$	if $D > 0$ and $A < 0$
c	$L \leq x_0 \leq U$			if $A > 0, D > 0$

(6.6.2)

The resulting confidence region can take any one of the forms (a), (b), or (c) in (6.6.2). If the result is given by (a), then the confidence region consists of the entire range between minus infinity and infinity. If the result is given by (b), then the confidence region consists of two disconnected intervals! Only case (c) results in a finite width confidence *interval*. If (a) or (b) is obtained, then the results are unsatisfactory and a larger sample is required to obtain a confidence interval as in (c). However, even in case (c), if the interval is too wide to draw useful conclusions, then the result is considered unsatisfactory and a larger sample is required to obtain a shorter confidence interval.

We illustrate the computations discussed above in Example 6.6.2.

EXAMPLE 6.6.2

In Example 6.5.3 suppose that we want to compare the hardness Y of egg shells for breeds 2 and 3 for values of X in the range from 2 to 20 units. To help make this comparison, we want to determine x_0, the point where the regression lines for breeds 2 and 3 intersect. We will find the point estimate of x_0 and a 95% confidence region for x_0. The data are given in Table 6.6.1 and are also stored in the file **eggshell.dat** on the data disk. Exhibit 6.6.1 gives a SAS output containing the results of regressing Y on X for breeds 2 and 3, respectively. Assumptions (A) are presumed to be valid, *and the data are obtained by sampling with preselected X values.*

We also compute

$$\bar{x}_2 = 6.25 \qquad \bar{x}_3 = 11.22 \qquad SSX(2) = 71.5 \qquad SSX(3) = 313.556 \qquad (6.6.3)$$

TABLE 6.6.1

Observation Number	Breed 2		Breed 3	
	Y_2	X_2	Y_3	X_3
1	9.86	3	6.52	2
2	9.54	3	5.11	5
3	11.96	4	7.75	7
4	12.46	5	6.84	8
5	11.38	6	7.65	10
6	14.69	8	9.49	15
7	16.48	9	7.03	16
8	20.11	12	9.41	18
9			12.01	20

EXHIBIT 6.6.1
SAS Output for Example 6.6.2

Regression of Y on X for Breed 2

The SAS System

00:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: Y2

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	85.70291	85.70291	105.466	0.0001
Error	6	4.87569	0.81261		
C Total	7	90.57860			

Root MSE	*0.90145*	*R-square*	*0.9462*
Dep Mean	13.31000	Adj R-sq	0.9372
C.V.	6.77274		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	6.467343	0.73860086	8.756	0.0001
X2	1	*1.094825*	0.10660785	10.270	0.0001

Regression of Y on X for Breed 3

Model: MODEL1
Dependent Variable: Y3

Analysis of Variance

Source	DF	Sum of *Squares*	Mean *Square*	F Value	Prob>F
Model	1	21.76050	21.76050	13.202	0.0084
Error	7	11.53778	1.64825		
C Total	8	33.29829			

EXHIBIT 6.6.1
(Continued)

Root MSE	1.28384	R-square	0.6535
Dep Mean	7.97889	Adj R-sq	0.6040
C.V.	16.09051		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	5.022537	0.91932262	5.463	0.0009
X3	1	0.263437	0.07250283	3.633	0.0084

From the computer output for each breed we obtain

$$\hat{\alpha}_2 = 6.467343 \quad \hat{\beta}_2 = 1.094825 \quad \hat{\alpha}_3 = 5.022537 \quad \hat{\beta}_3 = 0.263437 \tag{6.6.4}$$

$$\hat{\sigma}_2 = 0.90145 \quad \hat{\sigma}_3 = 1.28384 \quad SSE(2) = 4.87569 \quad SSE(3) = 11.53778 \tag{6.6.5}$$

Using (6.6.1) we get

$$\hat{x}_0 = \frac{(6.467343 - 5.022537)}{(0.263437 - 1.094825)} = -1.738 \tag{6.6.6}$$

Hence the estimate of x_0 indicates that the two regression lines do not intersect in the range from 2 to 20 units of food supplement. We now compute a 95% confidence region for x_0 by following the instructions in Box 6.6.1 using subscripts 2 and 3 (breeds 2 and 3) in place of subscripts 1 and 2, respectively.

1 $\bar{x}_2 = 6.25 \quad \bar{x}_3 = 11.22 \quad SSX(2) = 71.5 \quad SSX(3) = 313.556$
$Q_2 = 384 \quad Q_3 = 1447$

2 $SSE(2) = 4.87569 \quad SSE(3) = 11.53778$

3 $\hat{\sigma}^2 = \dfrac{(n_2 - 2)\hat{\sigma}_2^2 + (n_3 - 2)\hat{\sigma}_3^2}{(n_2 - 2) + (n_3 - 2)} = \dfrac{SSE(2) + SSE(3)}{(n_2 - 2) + (n_3 - 2)} = 1.263$

The table-value is $F_{0.95:1,13} = 4.67$. Using this we get

4 $A = 0.5900$

5 $B = 1.9272$

6 $C = -4.8900$

7 $D = 6.6000$

Since $A > 0$ and $D > 0$, the confidence region is of the form $L \le x_0 \le U$ where

8 $L = -7.62 \quad U = 1.09$

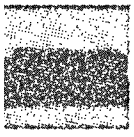
The required confidence statement is

$$C[-7.62 \leq x_0 \leq 1.09] = 0.95$$

Thus we conclude that the average hardness of egg shells will be greater for breed 2 than for breed 3 when the amount of food supplement is in the range from 2 to 20 units. ■

Caution One-sided $1 - \alpha/2$ confidence bounds for x_0 cannot be obtained by using only the upper bound or only the lower bound from the $1 - \alpha$ confidence region for x_0 discussed in this section.

You will have undoubtedly noticed that the calculations required to compute a confidence region for x_0 are quite cumbersome. For this reason we provide a computer program on the data disk for carrying out these computations. The use of this program is explained in Section 6.6 of the laboratory manuals.



Problems 6.6

6.6.1 For the breed data in Example 6.5.3, estimate the point x_0 where the regression lines for breed 1 and breed 2 intersect.

6.6.2 For the breed data in Example 6.5.3, estimate the point x_0 where the regression lines for breed 1 and breed 3 intersect.

6.6.3 In Problem 6.6.1 compute a 95% confidence region for x_0. What conclusions can you draw about the average hardness of egg shells for breeds 1 and 2? Explain. Below are the values of A, B, and C required in the calculations of Box 6.6.1.

$$A = 3.69095 \qquad B = -0.101865 \qquad C = -7.09432$$

You should verify these values.

6.6.4 In Problem 6.5.1 estimate the point x_0 where the two lines representing procedures 1 and 2 intersect.

6.6.5 In Problem 6.6.4 find a 90% confidence region for x_0. What conclusions can you draw from this confidence region? Below are values of A, B, and C in Box 6.6.1. Be sure to check these values.

$$A = -0.009932 \qquad B = 0.07428 \qquad C = -0.685414$$

6.7 Maximum or Minimum of a Quadratic Regression Model

In many practical applications we want to find the value of the predictor variable X that would maximize or minimize the average response $\mu_Y(x)$. For instance, we may be interested in maximizing the average breaking strength of an alloy by controlling

the amount of carbon in it. Or we may want to minimize the number of pests in agricultural plots by using suitable amounts of insecticides, etc. If the regression function of Y on X is a straight line over $a \leq x \leq b$, the interval of interest, then the maximum value or the minimum value of $\mu_Y(x)$ in the interval $a \leq x \leq b$ must occur either at a or at b. Thus the problem is easily solved in the case of straight line regression. However, in a number of applied problems, the regression function may not be a straight line, but it may be well approximated by a quadratic function

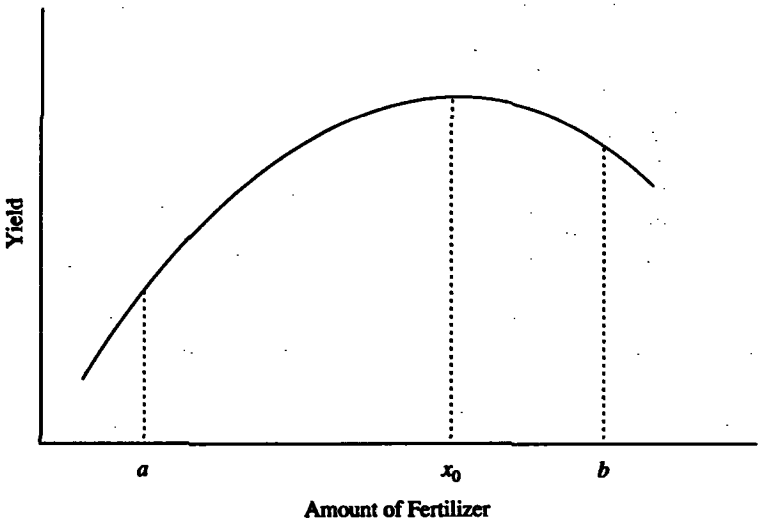
$$\mu_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2$$

In such situations, finding the value of x that would minimize or maximize the value of $\mu_Y(x)$ is not as simple as it is in the case of a straight line regression. This is the topic of our discussion in this section. We begin with an example.

EXAMPLE 6.7.1

Suppose an agricultural scientist is testing a new fertilizer to see how it affects the yield of corn. It is assumed that corn yield Y is related to the amount of fertilizer X through a quadratic regression function. This is reasonable because it is known that as increasing amounts of the fertilizer are applied the yield increases, but if too much fertilizer is used the yield will then decrease (see Figure 6.7.1). The scientist wants to determine x_0, the amount of fertilizer to apply, so that the average yield is a maximum.

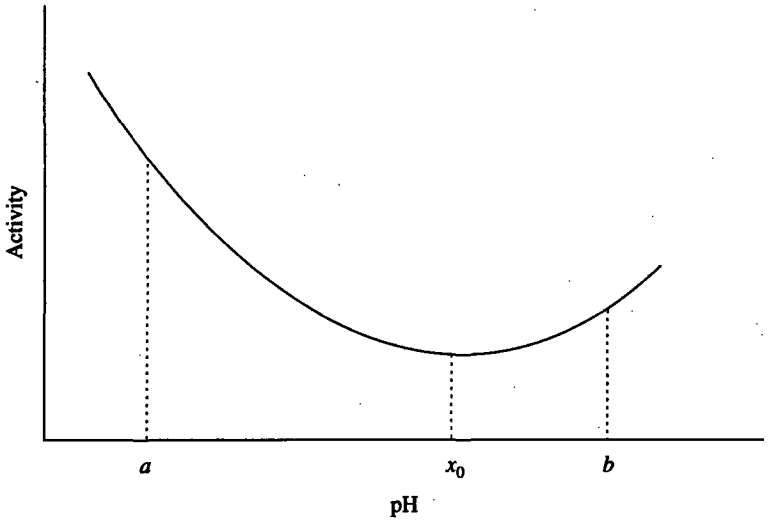
FIGURE 6.7.1



In some problems the quadratic function may be of the form shown in Figure 6.7.2, in which case the maximum will occur at one of the endpoints, a or b, of the interval $a \leq x \leq b$ of interest. In the figure the maximum occurs at a. ■

In certain problems it is the minimum point of a quadratic function that is of interest. This is illustrated in the following example.

FIGURE 6.7.2



EXAMPLE 6.7.2

The activity level Y in a growing medium of a certain disease-causing bacteria is related to the pH level X of the growing medium according to a quadratic regression function similar to the one in Figure 6.7.2. A scientist wants to determine x_0 such that when the pH level of the growing medium equals x_0, the average activity level of the bacteria is a *minimum*. The results from such a study will be used to prepare drugs to slow down the progress of the disease in affected humans. ■

For this problem, assumptions (A) are presumed to hold with

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

which we write as a multiple linear regression function

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

where we denote x by x_1 and x^2 by x_2. With a little algebra we can verify that

$$\mu_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 = \beta_2\left(x + \frac{\beta_1}{2\beta_2}\right)^2 + \left(\beta_0 - \frac{\beta_1^2}{4\beta_2}\right)$$

From this it is seen that $\mu_Y(x)$ attains a maximum at

$$x = \frac{-\beta_1}{2\beta_2}$$

when $\beta_2 < 0$, whereas $\mu_Y(x)$ attains a minimum at

$$x = \frac{-\beta_1}{2\beta_2}$$

when $\beta_2 > 0$ (see Figures 6.7.1 and 6.7.2). We denote the value where the quadratic function $\mu_Y(x)$ attains its maximum or minimum by x_0. Thus

$$x_0 = \frac{-\beta_1}{2\beta_2}$$

We want to obtain a point and confidence interval estimate of x_0. A sample of size n, denoted by $(y_1, x_1), \ldots, (y_n, x_n)$, is selected either by simple random sampling or by sampling with preselected X values, and the data are organized as in Table 6.7.1.

TABLE 6.7.1

Y	$X_1 = X$	$X_2 = X^2$
y_1	$x_{1,1} = x_1$	$x_{1,2} = x_1^2$
y_2	$x_{2,1} = x_2$	$x_{2,2} = x_2^2$
$\vdots$	$\vdots$	$\vdots$
y_n	$x_{n,1} = x_n$	$x_{n,2} = x_n^2$

From this we get the X matrix

$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} \\ 1 & x_{2,1} & x_{2,2} \\ \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{bmatrix}$$

As usual we let C denote $(X^TX)^{-1}$ where C is a 3 by 3 matrix given by

$$C = \begin{bmatrix} c_{1,1} & c_{1,2} & c_{1,3} \\ c_{2,1} & c_{2,2} & c_{2,3} \\ c_{3,1} & c_{2,3} & c_{3,3} \end{bmatrix} \tag{6.7.1}$$

Point estimates of β_0, β_1, β_2, $\mu_Y(x)$, and $\sigma = \sigma_{Y|X_1,X_2}$ are obtained by using the formulas in Section 4.4. The point estimate of x_0 is

$$\hat{x}_0 = -\frac{\hat{\beta}_1}{2\hat{\beta}_2} \tag{6.7.2}$$

A $1-\alpha$ confidence region for x_0 can be computed by following the instructions given in Box 6.7.1.

BOX 6.7.1 Instructions for Computing a Confidence Region for the Maximum or the Minimum Point of a Quadratic Regression Function

Carry out steps 1-6:

1 Compute

$$T = \hat{\sigma}^2 F_{1-\alpha:1,n-3}$$

2 Compute

$$A = 4(\hat{\beta}_2^2 - c_{3,3}T)$$

3 Compute

$$B = 2(\hat{\beta}_1\hat{\beta}_2 - c_{2,3}T)$$

4 Compute

$$D = \hat{\beta}_1^2 - c_{2,2}T$$

In the preceding formulas, $c_{i,j}$ is the (i,j)th element of the matrix $\boldsymbol{C}$ in (6.7.1).

5 Compute

$$G = B^2 - AD$$

6 If $G > 0$ and $A \neq 0$, compute L and U, where

$$L = \frac{-B - \sqrt{G}}{A} \quad \text{and} \quad U = \frac{-B + \sqrt{G}}{A}$$

A $1-\alpha$ confidence region for x_0 is given by

a	$-\infty < x_0 < \infty$	if $G \leq 0$	
b	$-\infty < x_0 \leq U$ and $L \leq x_0 < \infty$	if $G > 0$ and $A < 0$	(6.7.3)
c	$L \leq x_0 \leq U$	if $G > 0$ and $A > 0$	

If (a) or (b) is obtained in (6.7.3), the results will certainly be unsatisfactory and more observations will be required to obtain finite bounds such as in (c). However, (c) may also be unsatisfactory if the confidence interval is too wide for decision-making purposes.

The computations are illustrated in Example 6.7.3. We provide a computer program on the data disc for carrying out the preceding calculations, and the use of this program is explained in Section 6.7 of the laboratory manuals.

EXAMPLE 6.7.3

The output Y of an industrial process that manufactures sulfuric acid depends on X, the temperature at which the process is run. Past experience indicates that the regression function of output on temperature can be closely represented using a quadratic (i.e., a second-degree polynomial) function. To determine the temperature at which the maximum rate of production is achieved, a scientist carries out an experiment using several different process temperatures. The quantity of sulfuric acid produced in a day for each temperature is recorded. The data are given in Table 6.7.2 and are also stored in the file **sulfuric.dat** on the data disk. The (conceptual) target population can be defined as the collection of all possible pairs of numbers (Y, X), with Y equal to total daily output of sulfuric acid and X is any temperature that could be used. The sample can be considered to have been obtained from this target population using sampling with preselected X values.

TABLE 6.7.2

Sulfuric Acid Data

Observation Number	Daily Acid Production Y (tons)	Temperature X (°C)
1	1.93	100
2	2.22	125
3	2.85	150
4	2.69	175
5	3.01	200
6	3.82	225
7	3.91	250
8	3.65	275
9	3.71	300
10	3.40	325
11	3.71	350
12	2.57	375
13	2.71	400

Assumptions (A) are presumed valid with

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

We want a point estimate and a 95% confidence region for x_0, the temperature at which to run the process so the output will be maximized. A SAS output containing the results of the regression analysis is given in Exhibit 6.7.1. The predictor variables are $X_1 = X$ and $X_2 = X^2$. Exhibit 6.7.1 also contains a plot of the data. From the plot it seems reasonable to assume that a quadratic model is appropriate for this problem. From the computer output we get

$$\hat{x}_0 = \frac{(-0.0355190809)}{2(-0.000065223)} = 272.29$$

EXHIBIT 6.7.1

SAS Output for Example 6.7.3

```
                                The SAS System
                                                    00:00 Saturday, Jan 1, 1994
        Plot of TONS*TEMP.  Legend: A = 1 obs, B = 2 obs, etc.
     4 +
       |                              A    A
       |                                   A    A         A
       |
       |                                             A
       |
       |                         A
     3 +
       |              A
       |                   A                                              A
  TONS |                                                          A
       |
       |        A
     2 +  A
       |
       |
       |
       |
       |
     1 +
       ---+----+----+----+----+----+----+----+----+----+----+----+----+----+--
         100  125  150  175  200  225  250  275  300  325  350  375  400
                                      TEMP


                                The SAS System
                                                    00:00 Saturday, Jan 1, 1994
Model: MODEL1
                 X'X Inverse, Parameter Estimates, and SSE
                  INTERCEP            TEMP           TEMP2            TONS
 INTERCEP     4.3206793207   -0.036563437   0.0000687313   -1.141878122
 TEMP        -0.036563437   0.0003284715   -6.393606E-7   0.0355190809
 TEMP2        0.0000687313   -6.393606E-7   1.2787213E-9   -0.000065223
 TONS        -1.141878122   0.0355190809   -0.000065223   0.8366037962
Dependent Variable: TONS
```

EXHIBIT 6.7.1
(Continued)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	4.28849	2.14424	25.630	0.0001
Error	10	0.83660	0.08366		
C Total	12	5.12509			

Root MSE	0.28924	R-square	0.8368
Dep Mean	3.09077	Adj R-sq	0.8041
C.V.	9.35822		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-1.141878	0.60122348	-1.899	0.0867
TEMP	1	0.035519	0.00524214	6.776	0.0001
TEMP2	1	-0.000065223	0.00001034	-6.306	0.0001

A glance at the plot confirms that the maximum occurs when the temperature is close to 272.

For confidence region calculations we need certain elements of the matrix C (the first three rows and the first three columns in the matrix labeled X'X Inverse) given in the computer output of Exhibit 6.7.1. The table-value that we need is $F_{0.95:1,10} = 4.96$. The quantities mentioned in Box 6.7.1 are

$$\begin{aligned} &1 \quad T = 0.414955 \\ &2 \quad A = 0.0000000148936 \\ &3 \quad B = -0.00000410269 \\ &4 \quad D = 0.00112530 \\ &5 \quad G = 0.0000000000000722759 \end{aligned} \tag{6.7.4}$$

Since $G > 0$ we get

$$L = 257.4 \quad \text{and} \quad U = 293.5$$

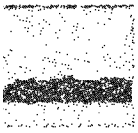
Thus the maximum yield is estimated to occur at the temperature 272°C. A 95% two-sided confidence interval for x_0, the temperature at which the maximum yield occurs, is given by

$$C[257.4 \le x_0 \le 293.5] = 0.95$$

Note These calculations are *very sensitive to rounding errors*, as is the case with most polynomial models, and it is advisable to keep as many significant figures

as possible for all intermediate calculations. Most good computer programs for regression analysis use special numerical techniques to minimize problems due to rounding. ■

Caution One-sided $1 - \alpha/2$ confidence bounds for x_0 cannot be obtained by using only the upper bound or only the lower bound from the $1 - \alpha$ confidence region for x_0 discussed in this section.



Problems 6.7

6.7.1 A study was done to determine how Y, the crushing strength of concrete, is affected by X, the amount of sand (in cubic inches) used in the mixture for a fixed amount of cement. The data in Table 6.7.3 were obtained by crushing concrete cylinders made with various amounts of sand and measuring the strength, which is the number of tons of force that concrete cylinders withstood before crumbling. The data are also stored in the file **concrete.dat** on the data disk. Assumptions (A) are presumed to be valid with $\mu_Y(x)$ given by

$$\mu_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \qquad 1 \le x \le 60$$

The data may be considered to have been obtained by sampling with preselected X values from the (conceptual) target population consisting of all possible pairs of numbers (Y, X) (with Y equal to crushing strength and X equal to amount of sand added) that could be observed.

TABLE 6.7.3
Concrete Crushing Strength Data

Observation Number	Strength Y (tons)	Amount of Sand X (cubic inches)
1	2.2	1
2	3.7	5
3	5.3	10
4	5.8	15
5	6.4	20
6	7.1	25
7	8.2	30
8	7.9	35
9	6.2	40
10	4.8	50
11	3.9	60

A MINITAB output from a regression analysis of Y on $X_1 = X$ and $X_2 = X^2$ is given in Exhibit 6.7.2.

EXHIBIT 6.7.2
MINITAB Output for Problem 6.7.1

```
The regression equation is
strength = 2.16 + 0.328  sand - 0.00515  sand^2

Predictor        Coef       Stdev    t-ratio        p
Constant       2.1628      0.4717       4.59    0.000
sand          0.32785     0.03714       8.83    0.000
sand^2     -0.0051511   0.0006031      -8.54    0.000

s = 0.6300      R-sq = 90.7%     R-sq(adj) = 88.4%

The C matrix is

 .5606449894334   -.0369011838509    .0004973796640
-.0369011838509    .0034763981595   -.0000540772111
 .0004973796640   -.0000540772111    .0000009164621
```

a Plot Y against X and assess the appropriateness of the quadratic regression model.

b Estimate $\beta_0, \beta_1, \beta_2, \sigma$.

c Find a point estimate of x_0, the amount of sand that gives the maximum crushing strength.

d Find a 90% confidence region for x_0. You are given the values of A, B, and D defined in Box 6.7.1, as follows:

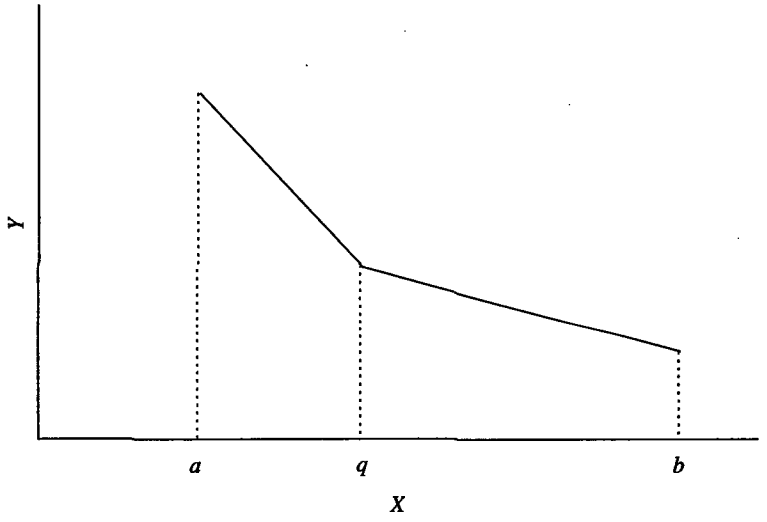
$$A = 0.0001011 \qquad B = -0.00323 \qquad D = 0.102715$$

You must verify these calculations.

6.8 Linear Splines

In some applications the regression function cannot be adequately described by a straight line or a polynomial function over the entire range of interest. If $[a, b]$ denotes the interval of interest for X, it may be that for some value q (between a and b) the regression function is a straight line between a and q, and is a *different* straight line between q and b and that the two lines intersect at q (see Figure 6.8.1).

FIGURE 6.8.1



The point q (which is assumed to be known) where the two lines intersect is called a *knot-point*, and the graph consisting of the two lines is called a linear spline with a knot-point at q. We could consider linear splines with several knot-points, and we could consider polynomial splines with several knot points. However, in this book we discuss only linear splines with one knot-point. We illustrate with two examples.

EXAMPLE 6.8.1

An investigator notices that for small companies the volume (Y) of sales in dollars tends to increase as a function of X, the dollars spent on advertising. The rate of increase in sales is rapid for the first several thousand dollars spent on advertising, but it slows down at some point. This suggests that a linear spline between a and b, similar to that shown in Figure 6.8.2, may be used as a prediction function to model the relationship between the average yearly sales and dollars spent on advertising. ■

EXAMPLE 6.8.2

A computer company leases a large number of personal computers to small businesses, and the company wants to predict next year's total maintenance cost of all computers leased. The computers are from 1 to 7 years old. It is known that the maintenance cost Y of a personal computer is a function of its age X. During the first 3 years the maintenance cost tends to increase less rapidly than during the next 4 years. This suggests that an appropriate model may be a linear spline regression function such as the one shown in Figure 6.8.3 with a knot-point at $q = 3$. ■

FIGURE 6.8.2

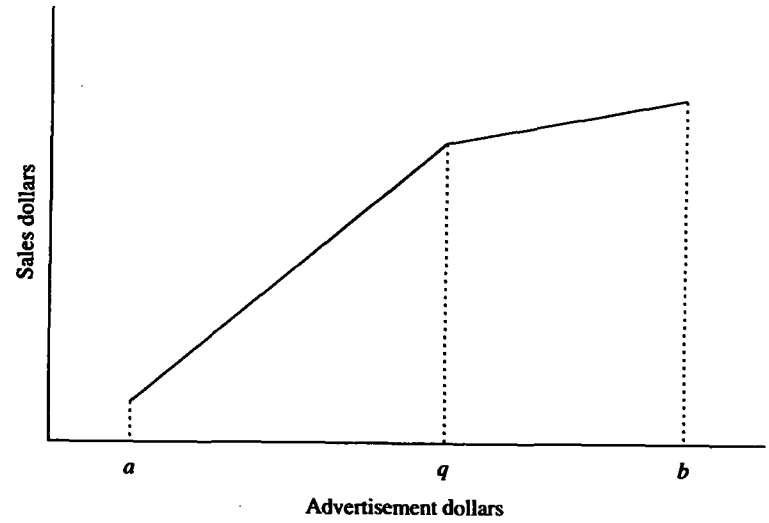


FIGURE 6.8.3

A linear spline regression function $\mu_Y(x)$ with a single knot-point at $x = q$ is defined by

$$\mu_Y(x) = \begin{cases} \mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x & \text{for } a \le x \le q \\ \mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x & \text{for } q \le x \le b \end{cases} \tag{6.8.1}$$

Because the lines are connected at $x = q$, where q is a known number, it follows that $\mu_Y^{(1)}(q) = \mu_Y^{(2)}(q)$, and we get

$$\alpha_1 + q\beta_1 = \alpha_2 + q\beta_2$$

We solve for α_2 and get

$$\alpha_2 = \alpha_1 + q(\beta_1 - \beta_2) \tag{6.8.2}$$

Thus, while it appears that there are four unknown parameters $\alpha_1, \alpha_2, \beta_1$, and β_2 in the model in (6.8.1), in fact only three parameters are needed because one of these parameters can be expressed in terms of the other three.

Suppose an investigator wants to estimate $\mu_Y(x)$ in (6.8.1) using sample data

$$(y_1, x_1), (y_2, x_2), \ldots, (y_n, x_n)$$

where $x_1, \ldots, x_m$ $(m < n)$ are in the interval $a \le x \le q$ and $x_{m+1}, \ldots, x_n$ are in the interval $q < x \le b$. The sample regression model is written out in detail below.

$$\begin{array}{lll} y_1 & = \alpha_1 + \beta_1 x_1 & + e_1 \\ \vdots & \vdots \quad \vdots & \vdots \\ y_m & = \alpha_1 + \beta_1 x_m & + e_m \\ y_{m+1} & = \alpha_2 + \beta_2 x_{m+1} & + e_{m+1} \\ \vdots & = \vdots \quad \vdots & \vdots \\ y_n & = \alpha_2 + \beta_2 x_n & + e_n \end{array} \tag{6.8.3}$$

Recall that $e_1, \ldots, e_n$ are the deviations of $y_1, \ldots, y_n$ from their respective subpopulation means. Assumptions (A) are presumed valid, and the data are obtained either by simple random sampling or by sampling with preselected X values. Substituting for α_2 using (6.8.2), the equations in (6.8.3) for $y_1, \ldots, y_n$ can be rewritten as follows:

$$\begin{array}{lll} y_1 & = \alpha_1 + \beta_1 x_1 & + e_1 \\ \vdots & \vdots & \vdots \\ y_m & = \alpha_1 + \beta_1 x_m & + e_m \\ y_{m+1} & = \alpha_1 + q\beta_1 + \beta_2(x_{m+1} - q) & + e_{m+1} \\ \vdots & \vdots & \vdots \\ y_n & = \alpha_1 + q\beta_1 + \beta_2(x_n - q) & + e_n \end{array} \tag{6.8.4}$$

To use the theory in Chapter 4 to compute point and interval estimates, the data are first organized as follows:

Y	X_1	X_2
y_1	x_1	0
y_2	x_2	0
$\vdots$	$\vdots$	$\vdots$
y_m	x_m	0
y_{m+1}	q	$x_{m+1}-q$
y_{m+2}	q	$x_{m+2}-q$
$\vdots$	$\vdots$	$\vdots$
y_n	q	x_n-q

(6.8.5)

Observe that the data for X_1 and X_2 in (6.8.5) correspond to the coefficients of β_1 and β_2 in (6.8.4). The X matrix is obtained by putting a column of 1's as the first column, values of X_1 in the second column, and the values of X_2 in the third column. Thus, in matrix notation, we can write (6.8.4) as

$$y = X\beta + e \tag{6.8.6}$$

where

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad X = \begin{bmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ \vdots & \vdots & \vdots \\ 1 & x_m & 0 \\ 1 & q & x_{m+1}-q \\ \vdots & \vdots & \vdots \\ 1 & q & x_n-q \end{bmatrix} \quad \beta = \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \beta_2 \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{6.8.7}$$

Hence we get

$$\hat{\beta} = \begin{bmatrix} \hat{\alpha}_1 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = (X^T X)^{-1} X^T y \tag{6.8.8}$$

and

$$\hat{\alpha}_2 = \hat{\alpha}_1 + q\hat{\beta}_1 - q\hat{\beta}_2 = d^T\hat{\beta} \tag{6.8.9}$$

where $d^T = [1, q, -q]$. So from (6.8.1) we get

$$\hat{\mu}_Y(x) = \begin{cases} \hat{\mu}_Y^{(1)}(x) = \hat{\alpha}_1 + \hat{\beta}_1 x & \text{for } a \le x \le q \\ \hat{\mu}_Y^{(2)}(x) = \hat{\alpha}_2 + \hat{\beta}_2 x & \text{for } q \le x \le b \end{cases} \tag{6.8.10}$$

Note that

$$\mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x = \boldsymbol{d}^T\boldsymbol{\beta} \quad \text{where} \quad \boldsymbol{d}^T = [1, x, 0] \text{ for } a \le x \le q$$

and

$$\mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x = \alpha_1 + \beta_1 q + \beta_2(x - q) = \boldsymbol{d}^T\boldsymbol{\beta}$$
$$\text{where } \boldsymbol{d}^T = [1, q, x - q] \text{ for } q \le x \le b$$

Since the spline model of this section is a special case of the multiple linear regression model, all formulas for point and confidence interval estimation discussed in Chapter 4 can be used for this model.

If we let $\boldsymbol{C}$ denote $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ and let $\hat{\sigma}$ denote the estimate of the standard deviations of the subpopulations, then

$$\hat{\sigma} = \sqrt{\frac{SSE}{(n-3)}} \tag{6.8.11}$$

where SSE is the error sum of squares. The standard errors to be used in confidence interval calculations for $\alpha_1, \alpha_2, \beta_1, \beta_2$, and $\mu_Y(x)$ are

$$SE(\hat{\alpha}_1) = \hat{\sigma}\sqrt{c_{1,1}} \qquad SE(\hat{\beta}_1) = \hat{\sigma}\sqrt{c_{2,2}} \qquad SE(\hat{\beta}_2) = \hat{\sigma}\sqrt{c_{3,3}} \tag{6.8.12}$$

$$SE(\hat{\alpha}_2) = \hat{\sigma}\sqrt{\boldsymbol{d}^T\boldsymbol{C}\boldsymbol{d}} \quad (\text{where } \boldsymbol{d}^T = [1, q, -q])$$
$$= \hat{\sigma}\sqrt{c_{1,1} + q^2c_{2,2} + q^2c_{3,3} + 2qc_{1,2} - 2qc_{1,3} - 2q^2c_{2,3}} \tag{6.8.13}$$

and

$$SE(\hat{\mu}_Y(x)) = \hat{\sigma}\sqrt{\boldsymbol{d}^T\boldsymbol{C}\boldsymbol{d}} \qquad (\text{where } \boldsymbol{d}^T = [1, x, 0])$$
$$= \hat{\sigma}[c_{1,1} + 2xc_{1,2} + x^2c_{2,2}]^{1/2} \qquad \text{for } a \le x \le q \tag{6.8.14}$$

$$SE(\hat{\mu}_Y(x)) = \hat{\sigma}\sqrt{\boldsymbol{d}^T\boldsymbol{C}\boldsymbol{d}} \qquad (\text{where } \boldsymbol{d}^T = [1, q, x - q])$$
$$= \hat{\sigma}[c_{1,1} + q^2c_{2,2} + (x-q)^2c_{3,3} + 2qc_{1,2}$$
$$+ 2q(x-q)c_{2,3} + 2(x-q)c_{1,3}]^{1/2} \qquad \text{for } q < x \le b \tag{6.8.15}$$

In the preceding equations, $c_{i,j}$ is the (i,j)th element of the matrix C. We illustrate the procedure in Example 6.8.3.

EXAMPLE 6.8.3

Consider the problem discussed in Example 6.8.1 where Y is annual sales and X is the number of dollars spent on advertising in one year. Data for this problem were obtained by simple random sampling of sales and advertising records of a study population consisting of several hundred small companies and are given in Table 6.8.1. These data are also stored in the file **sales.dat** on the data disk. Suppose that a spline model with a knot point at $q = 50$ is appropriate for this problem. We presume that assumptions (A) are valid. We exhibit the computations for a linear spline regression with a knot-point at $q = 50$ and find estimates of α_i and β_i and their standard errors. We will also exhibit the computations required to estimate $\mu_Y(65)$ and obtain the corresponding standard error.

TABLE 6.8.1
Advertising and Sales Data

Observation Number	Sales Y (thousands of dollars)	Advertisement Budget X (thousands of dollars)
1	260	12
2	328	25
3	376	30
4	356	35
5	404	41
6	399	41
7	404	41
8	414	44
9	428	45
10	436	46
11	439	47
12	452	47
13	465	55
14	461	59
15	475	64
16	462	66
17	472	73
18	456	74
19	490	83
20	496	87

The population regression function is

$$\mu_Y(x) = \begin{cases} \mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x & \text{for } 0 \le x \le 50 \\ \mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x & \text{for } 50 \le x \le 100 \end{cases} \tag{6.8.16}$$

The y vector and the X matrix, described in (6.8.7), are

$$y = \begin{bmatrix} 260 \\ 328 \\ 376 \\ 356 \\ 404 \\ 399 \\ 404 \\ 414 \\ 428 \\ 436 \\ 439 \\ 452 \\ 465 \\ 461 \\ 475 \\ 462 \\ 472 \\ 456 \\ 490 \\ 496 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 12 & 0 \\ 1 & 25 & 0 \\ 1 & 30 & 0 \\ 1 & 35 & 0 \\ 1 & 41 & 0 \\ 1 & 41 & 0 \\ 1 & 41 & 0 \\ 1 & 44 & 0 \\ 1 & 45 & 0 \\ 1 & 46 & 0 \\ 1 & 47 & 0 \\ 1 & 47 & 0 \\ 1 & 50 & 5 \\ 1 & 50 & 9 \\ 1 & 50 & 14 \\ 1 & 50 & 16 \\ 1 & 50 & 23 \\ 1 & 50 & 24 \\ 1 & 50 & 33 \\ 1 & 50 & 37 \end{bmatrix} \tag{6.8.17}$$

The regression of Y on X_1 and X_2 (where X_1 and X_2 are columns 2 and 3, respectively, of the preceding matrix X) can be obtained by the formulas in Chapter 4. Exhibit 6.8.1 gives a MINITAB output containing the results of a regression analysis of Y on X_1 and X_2. A plot of Y against X for the data in Table 6.8.1 is also given.

The estimates of the regression coefficients are

$$\hat{\alpha}_1 = 201.45 \qquad \hat{\beta}_1 = 5.0218 \qquad \hat{\beta}_2 = 0.9658$$

From these, and using (6.8.9), we get

$$\hat{\alpha}_2 = \hat{\alpha}_1 + q\hat{\beta}_1 - q\hat{\beta}_2$$

or

$$\hat{\alpha}_2 = 201.45 + 50(5.0218) - 50(0.9658) = 404.25$$

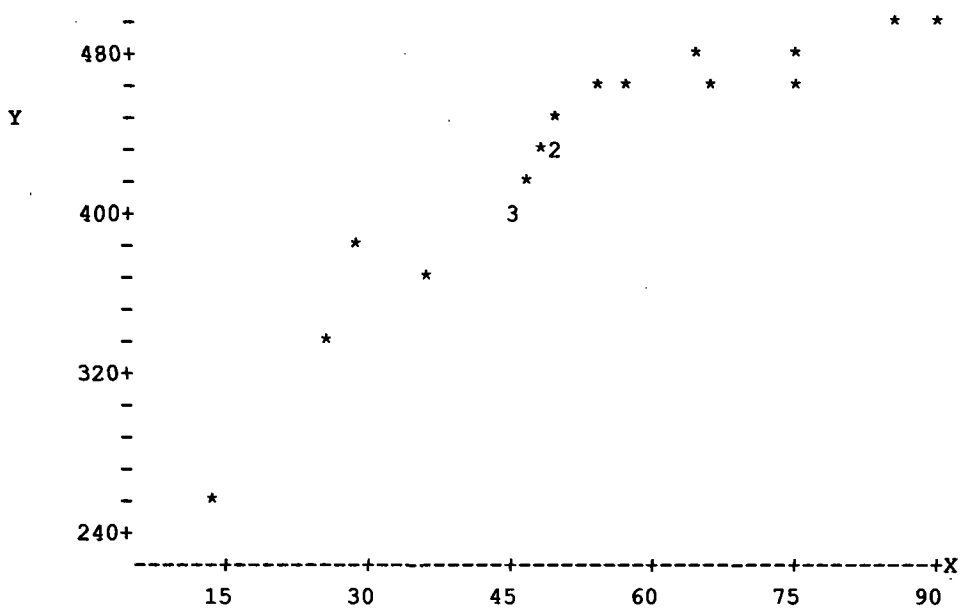
Furthermore, the estimate of σ is 11.05.

The standard errors of the estimated regression coefficients are obtained from (6.8.12) and (6.8.13) and are as follows:

$$SE(\hat{\alpha}_1) = 11.70 \qquad SE(\hat{\beta}_1) = 0.2875 \qquad SE(\hat{\beta}_2) = 0.2398 \qquad SE(\hat{\alpha}_2) = 15.28$$

EXHIBIT 6.8.1

MINITAB Output for Example 6.8.3

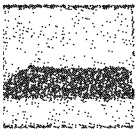



```
The regression equation is
  Y = 201 + 5.02 X1 + 0.966 X2

Predictor       Coef       Stdev     t-ratio       p
Constant      201.45       11.70       17.22   0.000
X1            5.0218      0.2875       17.47   0.000
X2            0.9658      0.2398        4.03   0.001

s = 11.05       R-sq = 96.8%     R-sq(adj) = 96.5%

The C matrix is

 1.1211817651      -.0266383640       .0082330906
 -.0266383640       .0006769450      -.0002816381
  .0082330906      -.0002816381       .0004711620
```

The standard errors for $\hat{\alpha}_1$, $\hat{\beta}_1$, and $\hat{\beta}_2$ can also be obtained directly from Exhibit 6.8.1. From these we can obtain confidence intervals for α_1, β_1, α_2, and β_2 in the usual manner.

To obtain the estimate of $\mu_Y(65)$ we first note that 65 is greater than 50 and so $\mu_Y(65) = \mu_Y^{(2)}(65) = \alpha_2 + 65\beta_2$. Hence $\hat{\mu}_Y(65) = \hat{\alpha}_2 + 65\hat{\beta}_2 = 467.03$. To obtain $SE(\hat{\mu}_Y(65))$ we apply the formula in (6.8.15). Hence $SE(\hat{\mu}_Y(65)) = \hat{\sigma}\sqrt{\boldsymbol{d}^T \boldsymbol{C}\boldsymbol{d}}$ where $\boldsymbol{d}^T = [1, q, x - q] = [1, 50, 65 - 50] = [1, 50, 15]$. You should verify that $\boldsymbol{d}^T \boldsymbol{C}\boldsymbol{d} = 0.0803$. Thus we get $SE(\hat{\mu}_Y(65)) = 11.05\sqrt{0.0803} = 3.13$.

The data disk contains a program written by us that can be used to do the calculations required in this section. The use of this program is explained in Section 6.8 of the laboratory manuals. ■



Problems 6.8

6.8.1 Consider the sulfuric acid data in Example 6.7.3, reproduced for convenience in Table 6.8.2. Assumptions (A) are presumed to hold, and the data are obtained by sampling with preselected X values. Suppose $\mu_Y(x)$ is a linear spline regression function given in (6.8.1) with a knot-point at $q = 270$. Exhibit the $\boldsymbol{X}$ matrix in (6.8.7) for these data.

6.8.2 A MINITAB output for the regression of Y on X_1 and X_2 (see (6.8.5)–(6.8.7)) in Problem 6.8.1 is given in Exhibit 6.8.2.

a Exhibit the point estimate and a 95% confidence interval for σ.

TABLE 6.8.2
Sulfuric Acid Data

Observation Number	Daily Acid Production Y (tons)	Temperature X (°C)
1	1.93	100
2	2.22	125
3	2.85	150
4	2.69	175
5	3.01	200
6	3.82	225
7	3.91	250
8	3.65	275
9	3.71	300
10	3.40	325
11	3.71	350
12	2.57	375
13	2.71	400

EXHIBIT 6.8.2
MINITAB Output for Problem 6.8.2

```
The regression equation is
Y = 0.780 + 0.0120 X1 - 0.0103 X2

Predictor       Coef      Stdev    t-ratio       p
Constant      0.7798     0.3269       2.39   0.038
X1          0.012033   0.001616       7.45   0.000
X2         -0.010350   0.002179      -4.75   0.000

s = 0.2795      R-sq = 84.8%     R-sq(adj) = 81.7%

The C matrix is

     1.36823595149      -.00647157429       .00401126291
     -.00647157429       .00003341185      -.00002697841
      .00401126291      -.00002697841       .00006075838
```

b Exhibit point estimates and 90% confidence intervals for α_1, β_1, α_2, and β_2.

c Exhibit the point estimate and 95% confidence interval for $\mu_Y(175)$.

d Exhibit the point estimate and 95% confidence interval for $\mu_Y(375)$.

6.8.3 A quadratic regression function $\mu_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2$ was fitted to the data in Problem 6.8.1. The corresponding MINITAB output with $X1 = X$ and $X2 = X^2$ is given in Exhibit 6.8.3. Which fits better, a spline function or a quadratic regression function? Explain.

EXHIBIT 6.8.3
MINITAB Output for Problem 6.8.3

```
The regression equation is
Y = - 1.14 + 0.0355 X1 -0.000065 X2

Predictor         Coef       Stdev    t-ratio       p
Constant       -1.1419      0.6012      -1.90   0.087
X1            0.035519    0.005242       6.78   0.000
X2         -0.00006522  0.00001034      -6.31   0.000

s = 0.2892      R-sq = 83.7%     R-sq(adj) = 80.4%

The C matrix is

    4.320679320679321    -.036563436563437     .000068731268731
    -.036563436563437     .000328471528472    -.000000639360639
     .000068731268731    -.000000639360639     .000000001278721
```

6.8.4 **a** Obtain the residual plots corresponding to the regression functions in Problems 6.8.2 and 6.8.3.

b Study these plots carefully and decide which of the two models results in residuals that are consistent with assumptions (A) for regression.

6.9 Exercises

6.9.1 Past research indicates that in track-and-field events such as 100-m, 200-m, and 400-m races, the performances of athletes during practice a day before a race and their performances during the race are closely related. The data given in Table 6.9.1 were obtained from a simple random sample of female athletes chosen from Division I schools. X is time, in seconds, to run 400 m in practice, and Y is time, in seconds to run 400 m during a race. The data are also stored in the file **track.dat** on the floppy disk. Assumptions (B) are presumed to hold.

A MINITAB output containing the regression of Y on X is given in Exhibit 6.9.1.

a Which of the following quantities can be validly estimated from the preceding data: $\mu_Y(x)$, β_0, β_1, σ, σ_Y, σ_X, μ_Y, and μ_X ?

b Find the point estimate of each parameter in (a) for which a valid point estimate can be computed.

TABLE 6.9.1
Track Data

Observation Number	Race Times (seconds) Y	Practice Times (seconds) X
1	50.97	51.85
2	49.47	51.05
3	52.07	52.47
4	51.17	51.82
5	51.38	52.94
6	50.26	51.33
7	51.11	51.90
8	49.15	50.17
9	49.97	51.24
10	49.51	50.43
11	50.63	52.00
12	49.63	50.88
13	50.81	51.31

EXHIBIT 6.9.1

MINITAB Output for Exercise 6.9.1

```
The regression equation is
racetime  = - 2.23 + 1.02 practime

Predictor       Coef      Stdev    t-ratio       p
Constant      -2.226      7.537      -0.30   0.773
practime      1.0234     0.1464       6.99   0.000

s = 0.3949      R-sq = 81.6%     R-sq(adj) = 80.0%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    1      7.6268      7.6268      48.90   0.000
Error        11      1.7158      0.1560
Total        12      9.3426

The C matrix is

 364.1887817    -7.0712953
  -7.0712953     0.1373293
```

c If an athlete runs the 400-m course in 52.08 seconds during practice a day before the race, predict her time for running 400 m during the race. Also compute a 95% lower prediction bound for her time.

d Four athletes are chosen to form a team to run a 4 by 400-m relay where each of the four athletes runs 400 m. Their times for 400 m during practice were 51.32, 52.07, 52.58, 51.96 seconds, respectively. Predict this team's total time for the relay during the race and compute a 99% lower bound for it.

e In (d) compute a 95% two-sided prediction interval for total team time.

6.9.2 The relationship between systolic blood pressure Y and weight X is known to be approximately linear for a population of people who have taken a certain medication for 1 year. The weights are in the range from 120 pounds to 250 pounds. A simple random sample of individuals was chosen from this population, and their weights and blood pressures were recorded. The data are shown in Table 6.9.2 and are also stored in the file **bpweight.dat** on the data disk. Assumptions (B) are presumed to hold.

A SAS output containing the results of a regression analysis of Y on X is given in Exhibit 6.9.2.

a Estimate $\lambda_{0.99}(210)$, the number such that 99% of the people in the population who weigh 210 pounds have blood pressure lower than this. Find a 95% two-sided confidence interval for $\lambda_{0.99}(210)$.

TABLE 6.9.2
Blood Pressure and Weight Data

Observation Number	Blood Pressure *Y*	Weight *X*
1	127	175
2	120	189
3	149	245
4	140	233
5	107	126
6	128	194
7	163	247
8	146	234
9	146	232
10	124	160
11	101	142
12	129	178
13	120	176
14	127	205
15	98	132
16	120	188
17	151	245
18	105	126
19	110	160
20	120	176

EXHIBIT 6.9.2
SAS Output for Exercise 6.9.2

```
                              The SAS System

                                        00:00 Saturday, Jan  1, 1994

Model: MODEL1

              X'X Inverse, Parameter Estimates, and SSE

                  INTERCEP          WEIGHT              BP

INTERCEP       1.175733109    -0.005983168    47.731630338
WEIGHT        -0.005983168       0.0000318     0.4189124085
BP            47.731630338    0.4189124085    624.47022328

Dependent Variable: BP
```

EXHIBIT 6.9.2
(Continued)

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	5518.47978	5518.47978	159.067	0.0001
Error	18	624.47022	34.69279		
C Total	19	6142.95000			

Root MSE	5.89006	R-square	0.8983
Dep Mean	126.55000	Adj R-sq	0.8927
C.V.	4.65433		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	47.731630	6.38666283	7.474	0.0001
WEIGHT	1	0.418912	0.03321491	12.612	0.0001

b A person who weighs 240 pounds has a blood pressure of 210. Compute an appropriate 80% confidence interval to help decide whether his blood pressure is in the upper 5% of the blood pressures of all people in this population who weigh 240 pounds.

c Calculate a 95% confidence interval for $\lambda_{0.99}(160)$, which is the number such that 99% of the people in the population who weigh 160 pounds will have blood pressure below this number.

6.9.3 In certain chemical assays the unknown concentration of a given compound in a chemical solution is measured indirectly by making use of the fact that when a beam of light of a given intensity passes through this chemical solution, the amount Y of light transmitted decreases with increasing concentrations X of the compound in the solution. The measurement system is calibrated by using solutions of known concentration and recording the intensity of transmitted light. The calibration data are given in Table 6.9.3 and are also stored in the file assay.dat on the data disk.

Here X is concentration in parts per million (ppm) and Y is intensity of transmitted light in appropriate units. Suppose that assumptions (A) apply and the data are obtained by sampling with preselected X values (i.e., preparing solutions with specified concentrations). The population regression function is of the form

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

TABLE 6.9.3
Chemical Assay Data

Observation Number	Intensity of Light Y	Concentration (parts per million) X
1	102.930	0
2	99.971	1
3	100.601	2
4	81.673	3
5	88.179	4
6	90.340	5
7	82.643	6
8	80.984	7
9	70.672	8
10	76.386	9
11	68.437	10
12	67.204	11
13	63.110	12
14	53.971	13
15	62.395	14
16	54.063	15
17	48.696	16
18	53.478	17
19	55.605	18
20	28.731	19
21	49.159	20

The (conceptual) study population is the collection of pairs of numbers (Y, X) (where Y is intensity of transmitted light and X is concentration) obtained from all possible solutions that could have been prepared. A MINITAB output containing the results of a regression analysis of Y on X is given in Exhibit 6.9.3.

EXHIBIT 6.9.3
MINTAB Output for Exercise 6.9.3

```
The regression equation is
intensty = 101 - 3.04 concentr

Predictor       Coef      Stdev    t-ratio       p
Constant     100.840      2.555      39.46   0.000
concentr      -3.0400     0.2186     -13.91   0.000

s = 6.066       R-sq = 91.1%     R-sq(adj) = 90.6%
```

EXHIBIT 6.9.3
(Continued)

```
Analysis of Variance

SOURCE       DF          SS          MS          F        p
Regression    1      7116.2      7116.2     193.42    0.000
Error        19       699.0        36.8
Total        20      7815.2


The C Matrix is

 .177489177      -.012987013
-.012987013       .001298701
```

a Estimate $\mu_Y(x)$, the average amount of light transmitted when the concentration is x ppm.

b A solution containing an unknown concentration x_0 of the compound gives a Y reading of 55 units. Estimate the unknown concentration x_0 and also compute a 95% confidence region for x_0.

6.9.4 The hardness Y of steel ball bearings is related to the rate X at which they were cooled after they were made. Data were collected to determine this relationship by making ball bearings using various known rates of cooling and then measuring the hardness. Data obtained from a simple random sample of a week's output of ball bearings (which is the study population) at each rate of cooling are given in Table 6.9.4 and are also stored in the file **ballbear.dat** on the data disk.

Suppose assumptions (A) hold and the data were obtained by sampling with preselected X values. The population regression function is

$$\mu_Y(x) = \beta_0 + \beta_1 x \tag{6.9.1}$$

A MINITAB output containing the results of a regression analysis of Y on X is given in Exhibit 6.9.4.

a Plot Y versus X.

b Estimate $\mu_Y(x)$.

c If the manufacturer wants to produce ball bearings with an average hardness of 35.00 units, estimate the required cooling rate x_0 and also compute a 95% confidence region for x_0.

d Calculate a 95% confidence interval for $\lambda_{0.01}(25)$, the number such that 1% of all ball bearings made using a cooling rate of 25°C will have a hardness less than that number (i.e., 99% of all ball bearings made using a cooling rate of 25°C will have a hardness greater than that number).

e Write a short report summarizing the results of parts (a), (b), (c), and (d). The language of the report should be such that people who are not well acquainted with the meaning of confidence intervals, tolerance intervals, etc., can understand the results.

TABLE 6.9.4
Ball Bearing Data

Observation Number	Hardness Index Y	Rate of cooling (deg C per minute) X
1	48.60	10
2	47.80	10
3	47.60	15
4	46.70	15
5	46.20	20
6	45.70	20
7	46.55	25
8	46.57	25
9	46.49	30
10	41.82	30
11	41.40	35
12	42.10	35
13	42.01	40
14	41.67	40
15	38.96	45
16	40.97	45
17	38.71	50
18	37.00	50
19	35.88	55
20	36.25	55
21	39.23	60
22	34.18	60
23	34.59	65
24	37.56	65
25	33.49	70
26	33.93	70
27	31.02	75
28	31.57	75
29	26.99	80
30	28.38	80

EXHIBIT 6.9.4
MINITAB Output for Exercise 6.9.4

```
The regression equation is
hardness = 51.9 - 0.271 coolrate

Predictor       Coef      Stdev    t-ratio       p
Constant     51.8648     0.6631      78.22   0.000
coolrate    -0.27113    0.01328     -20.41   0.000
```

EXHIBIT 6.9.4
(Continued)

```
s = 1.572       R-sq = 93.7%     R-sq(adj) = 93.5%

Analysis of Variance

SOURCE        DF          SS          MS          F        p
Regression     1      1029.1      1029.1     416.61    0.000
Error         28        69.2         2.5
Total         29      1098.3

The C matrix is

 .17797619048    -.00321428571
-.00321428571     .00007142857
```

6.9.5 We want to determine whether or not Y, the number of days of prison sentence for thefts whose total value is under \$1,000 dollars for first time offenders, is related to X, the amount of money stolen. Sample data from thefts falling in this category in three different states in the United States, denoted as states 1, 2, and 3, are obtained from last year's police records in each state. These data are given in Table 6.9.5 and are also stored in the file **prison.dat** on the data disk.

TABLE 6.9.5
Prison Data

Observation Number	Days in Prison Y	Dollars Stolen X	State
1	44	367	1
2	81	855	1
3	43	284	1
4	40	305	1
5	38	215	1
6	44	308	1
7	49	433	1
8	51	455	1
9	49	454	1
10	57	429	1
11	47	345	1
12	37	167	1
13	67	689	1
14	55	499	1
15	43	538	2

(Continued)

TABLE 6.9.5
(Continued)

Observation Number	Days in Prison Y	Dollars Stolen X	State
16	32	290	2
17	53	759	2
18	55	734	2
19	40	499	2
20	44	541	2
21	42	474	2
22	51	940	2
23	35	314	2
24	39	351	2
25	51	703	2
26	37	459	2
27	50	732	3
28	52	556	3
29	53	960	3
30	39	134	3
31	55	826	3
32	53	738	3
33	37	403	3
34	46	511	3
35	45	699	3
36	49	778	3
37	37	530	3
38	35	140	3
39	38	429	3
40	48	554	3
41	50	672	3
42	30	125	3
43	29	124	3

There are three populations, one for each of the three states. The population regression function for state (i) is

$$\mu_Y^{(i)}(x) = \alpha_i + \beta_i x \qquad \text{for } i = 1, 2, 3$$

The subpopulation standard deviation for state (i) is σ_i for $i = 1, 2, 3$. We suppose that assumptions (A) are valid for each state and that data are obtained for each state by simple random sampling. The study and target populations are the same for this problem because the investigator wants to study the police records for these three states for the past 2 years. A MINITAB output that contains the results of regression analysis for each state is given in Exhibit 6.9.5. In this Exhibit Y_1, X_1 denote the response variable and the predictor variable for state (1). Likewise Y_2, X_2 and Y_3, X_3 refer to states (2) and (3), respectively.

EXHIBIT 6.9.5

MINITAB Output for Exercise 6.9.5

REGRESSION ANALYSIS FOR STATE (1)

The regression equation is
Y1 = 23.5 + 0.0643 X1

Predictor	Coef	Stdev	t-ratio	p
Constant	23.472	1.941	12.10	0.000
X1	0.064322	0.004312	14.92	0.000

s = 2.823 R-sq = 94.9% R-sq(adj) = 94.5%

The C matrix is:

.472391779577	-.000967008598
-.000967008598	.000002332148

ymean = 50.1429
xmean = 414.643
SSY = 1869.71
SSX = 428789
SXY = 27580.7

REGRESSION ANALYSIS FOR STATE (2)

The regression equation is
Y2 = 24.5 + 0.0345 X2

Predictor	Coef	Stdev	t-ratio	p
Constant	24.496	2.805	8.73	0.000
X2	0.034543	0.004820	7.17	0.000

s = 3.171 R-sq = 83.7% R-sq(adj) = 82.1%

The C matrix is:

.782652676944	-.001271104532
-.001271104532	.000002310399

y mean = 43.5000
x mean = 550.167
SSY = 617.000
SSX = 432826
SXY = 14951.0

EXHIBIT 6.9.5
(Continued)

REGRESSION ANALYSIS FOR STATE (3)

The regression equation is
Y3 = 29.3 + 0.0278 X3

Predictor	Coef	Stdev	t-ratio	p
Constant	29.309	2.246	13.05	0.000
X3	0.027802	0.003844	7.23	0.000

s = 4.089 R-sq = 77.7% R-sq(adj) = 76.2%

The C matrix is:

.3015545702755	-.0004630712260
-.0004630712260	.0000008834262

y mean = 43.8824
x mean = 524.176
SSY = 1125.76
SSX = 1131957
SXY = 31470.4

a Estimate $\mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x$; $\mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x$; $\mu_Y^{(3)}(x) = \alpha_3 + \beta_3 x$; and $\sigma_1, \sigma_2, \sigma_3$.

b If we assume that $\sigma_1 = \sigma_2 = \sigma_3$ and denote their common value by σ, estimate this common standard deviation.

c If $|\mu_Y^{(i)}(x) - \mu_Y^{(j)}(x)| \leq 20$ days for i and j and for all x between 200 and 1,000 dollars, then the population regression function of Y on X for state (i) would be considered to be equivalent to that for state (j) for the purposes of this problem. Compute appropriate 95% simultaneous confidence intervals that an investigator can use to help decide whether the three states (or any two states) have equivalent population regression functions.

d What population parameters must be examined to help determine how much the average sentences differ for each pair of states when \$1,000 is stolen?

6.9.6 An investigator wants to study how the salaries of high school teachers who teach in the public school system of a large city are related to experience (in years employed) and determine what the differences are, if any, between the salaries of male teachers and female teachers. A simple random sample of male teachers and a simple random sample of female teachers are chosen, and their monthly salaries Y and years of experience X are recorded. The data for both males and females are given in Table 6.9.6. The data are also stored in the file **salaries.dat** on the data disk. The study populations of items in this problem are all male teachers in this school system and

all female teachers in the school system, respectively, the year that the sample was collected. These are also the target populations of items for this problem.

Suppose that assumptions (A) are valid for each population (male and female). The population regression functions for males and females are

$$\text{females:} \quad \mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x$$
$$\text{males:} \quad \mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x$$

The results of a regression analysis, done separately for males and females, is given in Exhibit 6.9.6.

TABLE 6.9.6
Salary Data

Observation Number	Salary (thousands of dollars) Y	Experience (years) X	Sex (1 = females 2 = males)
1	25.1	0	1
2	41.3	17	1
3	29.6	5	1
4	40.7	15	1
5	36.1	9	1
6	40.2	15	1
7	34.5	8	1
8	28.9	5	1
9	37.1	13	1
10	42.0	20	1
11	36.7	11	1
12	24.8	1	1
13	33.0	6	2
14	33.4	7	2
15	54.9	23	2
16	53.2	20	2
17	46.8	18	2
18	61.2	27	2
19	40.9	11	2
20	38.9	10	2
21	61.7	29	2
22	53.5	23	2
23	30.7	4	2
24	53.2	22	2
25	58.6	25	2
26	37.8	9	2
27	58.3	25	2

EXHIBIT 6.9.6

MINITAB Output for Exercise 6.9.6

REGRESSION ANALYSIS FOR FEMALES

The regression equation is
salary = 25.3 + 0.954 yearsexp

Predictor	Coef	Stdev	t-ratio	p
Constant	25.2873	0.7711	32.79	0.000
yearsexp	0.95422	0.06626	14.40	0.000

s = 1.398 R-sq = 95.4% R-sq(adj) = 94.9%

The C matrix is:

0.3043641	-0.0222888
-0.0222888	0.0022476

mean y = 34.7500
mean x = 9.91667
SSY = 424.650
SSX = 444.917
SXY = 424.550

REGRESSION ANALYSIS FOR MALES

The regression equation is
salary = 25.6 + 1.28 yearsexp

Predictor	Coef	Stdev	t-ratio	p
Constant	25.6121	0.6921	37.01	0.000
yearsexp	1.28154	0.03618	35.42	0.000

s = 1.154 R-sq = 99.0% R-sq(adj) = 98.9%

The C matrix is:

0.35984004	-0.01697915
-0.01697915	0.00098335

y mean = 47.7400
x mean = 17.2667
SSY = 1687.46
SSX = 1016.93
SXY = 1303.24

a Plot the estimated regression lines for both males and females on the same graph.

b What is the difference between the average salaries for males and females with the same number of years of experience? Express your answer in terms of population parameters.

c What is the difference between the average starting salaries of males and females? Express your answer in terms of population parameters. Find a point estimate and a 95% confidence interval for this quantity.

d What population parameters would you examine to determine whether the disparity between male and female salaries remains constant at every experience level (years) or whether it changes with years of experience for the years from 0 to 30 (i.e., $0 \leq x \leq 30$)? Find a 95% confidence interval for this quantity. Explain.

e What population parameters would you examine to determine whether the salaries ever become equal in the range $0 \leq x \leq 30$? Explain.

f The investigator will conclude that there is evidence of a *systematic* salary differential if the difference between the male and female average annual salaries exceeds \$500 anywhere in the range $0 \leq x \leq 30$.

i Write the population parameters needed to determine whether there is a systematic salary differential between males and females.

ii Do the data provide evidence of a systematic salary differential between males and females (i.e., estimate the quantities in (i))? If so, in which direction?

iii Compute appropriate 95% simultaneous confidence intervals to help arrive at a conclusion in (i).

6.9.7 The gas mileage Y in miles per gallon (mpg) for cars depends on X, the speed in miles per hour (mph) at which they are driven. A study was conducted to evaluate the gas mileage, at various speeds, of cars (with four cylinders) made by two leading manufacturers, 1 and 2. The data are given in Table 6.9.7 and are also stored in the file **mpg.dat** on the data disk.

We suppose that assumptions (A) are valid for both populations under consideration. The data were obtained by sampling with preselected X values. The regression functions are

$$\mu_Y^{(1)}(x) = \alpha_1 + \beta_1 x$$
$$\mu_Y^{(2)}(x) = \alpha_2 + \beta_2 x$$

where x is in the interval $25 \leq x \leq 65$. A MINITAB output is given in Exhibit 6.9.7 to help you answer questions of interest.

List the population parameters that must be estimated to answer each of questions (a), (b), and (c).

a What is the difference between the average gas mileages for cars made by manufacturer 1 and manufacturer 2 when driven at the speed of 25 miles per hour?

b What is the difference between the average gas mileages for cars made by manufacturer 1 and manufacturer 2 when driven at the speed of 60 miles per hour?

TABLE 6.9.7

Miles per Gallon Data

Observation Number	Mpg Y	Mph X	Manufacturer
1	30.4	30	1
2	28.9	35	1
3	28.6	40	1
4	29.2	45	1
5	27.1	50	1
6	26.8	55	1
7	26.6	60	1
8	25.1	65	1
9	28.2	30	2
10	28.0	35	2
11	27.2	40	2
12	26.9	45	2
13	27.6	50	2
14	27.0	55	2
15	26.1	60	2
16	26.5	65	2

EXHIBIT 6.9.7

MINITAB Output for Exercise 6.9.7

```
   REGRESSION ANALYSIS FOR CARS OF MANUFACTURER (1)

The regression equation is
mile/gal = 34.2 - 0.134 mile/hr

Predictor        Coef       Stdev    t-ratio        p
Constant      34.1821      0.8804      38.83    0.000
mile/hr      -0.13357     0.01802      -7.41    0.000

s = 0.5839      R-sq = 90.2%      R-sq(adj) = 88.5%

 The C matrix is

      2.27380943       -0.04523810
     -0.04523810        0.00095238

y mean = 27.8375
x mean = 47.5000
SSY    = 20.7787
SSX    = 1050.00
SXY    = -140.250
```

EXHIBIT 6.9.7
(Continued)

REGRESSION ANALYSIS FOR CARS OF MANUFACTURER (1)

```
The regression equation is
mile/gal = 29.6 - 0.507 mile/hr

Predictor        Coef          Stdev    t-ratio        p
Constant      29.5964         0.5931      49.90    0.000
mile/hr      -0.05071        0.01214      -4.18    0.006

s = 0.3933      R-sq = 74.4%     R-sq(adj) = 70.2%

 The C matrix is:

     2.27380943     -0.04523810
    -0.04523810      0.00095238

y mean = 27.1875
x mean = 47.5000
SSY    = 3.62875
SSX    = 1050.00
SXY    =-53.2500
```

c What is the difference between the slopes of the two population regression lines?

d Obtain point estimates for the quantities of interest in parts (a), (b), and (c).

e Obtain 90% two-sided confidence intervals for the quantities of interest in parts (a), (b), and (c). You must decide whether one-at-a-time intervals or simultaneous intervals are appropriate.

f Estimate where the two lines intersect; i.e., compute the point estimate of x_0 such that $\mu_Y^{(1)}(x_0) = \mu_Y^{(2)}(x_0)$. Also compute a 90% two-sided confidence region for x_0.

g Write a short report discussing which population of cars, those made by manufacturer 1 or 2, gives better gas mileage on the average, and by how much, for highway driving (i.e., for driving at speeds between 50 and 65 miles per hour).

6.9.8 A study was conducted to understand the relationship between the nickel-to-iron ratio Y in oat plants and X, their age in days after emergence. From past experience it is known that a polynomial of degree less than or equal to 3 will give an adequate fit. Assumptions (A) are presumed to be valid, and the data are obtained by sampling with preselected X values. The population regression function is

$$\mu_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$$

where some of the β_i's may be negligible. The data are given in Table 6.9.8 and are also stored in the file **nickel.dat** on the data disk.

TABLE 6.9.8
Nickel-to-Iron Ratio Data

Observation Number	Nickel-to-Iron Ratio Y	Age (days) X
1	0.08	0
2	0.71	5
3	0.69	10
4	0.96	15
5	1.02	20
6	1.13	25
7	1.16	30
8	1.16	35
9	1.13	40
10	1.19	45
11	1.25	50
12	1.17	55
13	1.24	60
14	1.08	65
15	1.07	70
16	1.02	75
17	0.73	80

The SAS output in Exhibit 6.9.8 contains the regression analyses of Y on X corresponding to polynomial models of degrees 3, 2, and 1, respectively. In the computer output, $X1$ denotes X, $X2$ denotes X^2, and $X3$ denotes X^3.

a Fit polynomial models of degrees 3, 2, 1, and 0, in this order, and evaluate how good each model is by computing the mean squared error (*MSE*) for each model. Plot each fitted regression function along with the observed data. Which model would you choose after examining the plots and each *MSE*?

b Estimate all the parameters for the model chosen in (a).

c Use the model chosen in (a) and exhibit the population parameter that is the average nickel-to-iron ratio of oat plants at the end of 12 days after emergence.

d Estimate the quantity of interest in (c). Also compute a two-sided 80% confidence interval for this quantity.

e Compute a two-sided 80% confidence interval for the nickel-to-iron ratio of a single oat plant 12 days after emergence if this plant is to be chosen at random from all plants 12 days after emergence. Use the model chosen in (a).

EXHIBIT 6.9.8

SAS Output for Exercise 6.9.8

The SAS System

REGRESSION OF Y ON X1, X2, X3

00:00 Saturday, Jan 1, 1994

Model: MODEL1
Dependent Variable: RATIO

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	1.23880	0.41293	37.829	0.0001
Error	13	0.14191	0.01092		
C Total	16	1.38071			

Root MSE	0.10448	R-square	0.8972
Dep Mean	0.98765	Adj R-sq	0.8735
C.V.	10.57853		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.262497	0.08255493	3.180	0.0072
X	1	0.052611	0.00921737	5.708	0.0001
X2	1	-0.000837	0.00027267	-3.071	0.0089
X3	1	0.000003402	0.00000224	1.520	0.1523

REGRESSION OF Y ON X1, X2

Model: MODEL2
Dependent Variable: RATIO

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	2	1.21357	0.60678	50.825	0.0001
Error	14	0.16714	0.01194		
C Total	16	1.38071			

EXHIBIT 6.9.8
(Continued)

Root MSE	0.10926	R-square	0.8789
Dep Mean	0.98765	Adj R-sq	0.8617
C.V.	11.06305		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.333942	0.07098660	4.704	0.0003
X	1	0.039938	0.00411591	9.703	0.0001
X2	1	-0.000429	0.00004964	-8.642	0.0001

REGRESSION OF Y ON X1

Model: MODEL3
Dependent Variable: RATIO

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	1	0.32189	0.32189	4.560	0.0496
Error	15	1.05881	0.07059		
C Total	16	1.38071			

Root MSE	0.26568	R-square	0.2331
Dep Mean	0.98765	Adj R-sq	0.1820
C.V.	26.90064		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	0.762941	0.12338876	6.183	0.0001
AGE	1	0.005618	0.00263066	2.135	0.0496

6.9.9 The growth rate of newly hatched turkeys is known to depend on the amount of vitamin A in the diet. An experiment is conducted to evaluate this relationship. Twelve newly hatched turkeys were grouped into six pairs, and each pair was put on a diet consisting of a specified amount of vitamin A. The average weight gain Y in pounds per week (for the first three weeks) and the vitamin A dosage Z for each turkey are given in Table 6.9.9 and are also stored in the file **turkey.dat** on the data disk.

TABLE 6.9.9
Turkey Growth Data

Observation Number	Weight Gain Y (lb/week)	Vitamin A Z (units/g of diet)
1	0.169	1.5
2	0.137	1.5
3	0.219	3.0
4	0.221	3.0
5	0.278	6.0
6	0.289	6.0
7	0.328	12.0
8	0.317	12.0
9	0.287	24.0
10	0.336	24.0
11	0.274	48.0
12	0.286	48.0

From previous investigations it is known that the regression function

$$\beta_0 + \beta_1 \log_{10} z + \beta_2 (\log_{10} z)^2$$

is appropriate for modeling the relation between weight gain and vitamin dosage. The population under study is $\{(Y, Z)\}$, but if we let $X = \log_{10}(Z)$, then the regression function becomes

$$\mu_Y(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{6.9.2}$$

Thus it is convenient to work with the population $\{(Y, X)\}$ as the study population. We suppose that assumptions (A) hold for the study population $\{(Y, X)\}$ (at least approximately). A MINITAB output obtained by regressing Y on X and X^2 is shown in Exhibit 6.9.9. In the computer output $X1$ denotes X and $X2$ denotes X^2.
The sum of squares for pure error = 0.001908.

a Were the data for this study obtained by simple random sampling or by sampling with preselected X values? Explain.

b Estimate $\mu_Y(x)$.

EXHIBIT 6.9.9

MINITAB Output for Exercise 6.9.9

```
The regression equation is
Y = 0.0850 + 0.379 X1 - 0.156 X2

Predictor       Coef      Stdev    t-ratio       p
Constant     0.08498    0.01585       5.36   0.000
X1           0.37935    0.03987       9.52   0.000
X2          -0.15577    0.02089      -7.46   0.000

s = 0.01636     R-sq = 94.4%     R-sq(adj) = 93.2%

Analysis of Variance

SOURCE       DF          SS          MS          F       p
Regression    2    0.040583    0.020291      75.85   0.000
Error         9    0.002408    0.000268
Total        11    0.042990

The C matrix is

  0.93869811    -2.10459471    0.97548175
 -2.10459471     5.94146967   -3.02917099
  0.97548175    -3.02917099    1.63092554
```

c What parameter would you test equal to zero to test whether or not a straight line regression model is as good as the quadratic model in (6.9.2) for predicting Y?

d Carry out the test in (c). Find the P-value and reject NH if it is less than 0.05. What is your conclusion based on this test?

e Estimate the number x_0 for which $\mu_Y(x_0)$, the average growth rate, is a maximum.

f Calculate a 95% confidence region for x_0 in (e).

g Calculate a 95% confidence region for z_0, the dosage of vitamin A at which the average growth rate $\mu_Y(z)$ is a maximum.

6.9.10 The weight of a newborn baby increases quite rapidly during the first 100 days after birth, and after that it slows down somewhat. Data for seventeen babies were obtained from the past 3 years' records for babies born at a certain hospital. Suppose that assumptions (A) are valid where the data were obtained by simple random sampling. The regression function $\mu_Y(x)$ is given by the following spline model with a knot-point at $x = 100$.

$$\mu_Y(x) = \begin{cases} \alpha_1 + \beta_1 x & 0 < x \leq 100 \\ \alpha_2 + \beta_2 x & 100 < x \leq 200 \end{cases} \qquad \textbf{(6.9.3)}$$

The data are given in Table 6.9.10 and are also stored in the file **babywt.dat** on the data disk. A MINITAB output is given in Exhibit 6.9.10 to help answer various questions of interest.

a Estimate $\alpha_1, \beta_1, \alpha_2, \beta_2$, and σ. Plot y_i against x_i. Exhibit the X matrix in (6.8.7).

b In (a) exhibit the population quantities needed to determine the difference between the *average growth rate* of newborn babies during the first 100 days and the *average growth rate* during the next 100 days.

c In (b) compute a 90% two-sided confidence interval for the quantity of interest.

d Instead of using the preceding spline model, assume that a quadratic model given by

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

holds for $0 \leq x \leq 200$. Estimate this regression function. Exhibit 6.9.11 contains a MINITAB output for a quadratic regression model along with the fits, residuals, and standardized residuals. Plot the residuals. Compare this residual plot with the residual plot in (a). Which model (a quadratic or the spline) is better for this problem? Give your reasons in a short report.

TABLE 6.9.10
Weights of Newborn Babies.

Observation Number	Weight Y (pounds)	Age X (days)
1	7.5	7
2	7.9	12
3	8.4	18
4	10.1	45
5	11.5	67
6	12.8	88
7	13.4	92
8	13.9	99
9	13.5	105
10	14.5	108
11	14.7	120
12	15.8	147
13	16.1	156
14	16.5	159
15	16.9	167
16	17.3	183
17	17.6	195

EXHIBIT 6.9.10

MINITAB Output for Exercise 6.9.10

```
The regression equation is
Y = 7.08 + 0.0673 X1 + 0.0423 X2

Predictor       Coef       Stdev    t-ratio       p
Constant      7.0810      0.1356      52.21   0.000
X1          0.067321    0.001825      36.89   0.000
X2          0.042319    0.001873      22.60   0.000

s = 0.2178      R-sq = 99.6%     R-sq(adj) = 99.6%

Analysis of Variance

SOURCE       DF          SS          MS         F       p
Regression    2     175.155      87.577   1845.97   0.000
Error        14       0.664       0.047
Total        16     175.819

 The C matrix is

   .38767760595    -.00453033498    .00096764899
  -.00453033498     .00007021359   -.00003688170
   .00096764899    -.00003688170    .00007392925
```

EXHIBIT 6.9.11

MINITAB Output for (d) in Exercise 6.9.10

```
The regression equation is
Y = 6.92 + 0.0801 X1 -0.000128 X2

Predictor       Coef       Stdev    t-ratio       p
Constant      6.9199      0.1528      45.30   0.000
X1          0.080080    0.003443      23.26   0.000
X2       -0.00012802  0.00001694      -7.56   0.000

s = 0.2229      R-sq = 99.6%     R-sq(adj) = 99.5%
```

EXHIBIT 6.9.11
(Continued)

Analysis of Variance

SOURCE	DF	SS	MS	F	p
Regression	2	175.124	87.562	1763.12	0.000
Error	14	0.695	0.050		
Total	16	175.819			

ROW	Y	X1	X2	fits	residual	stdresid
1	7.5	7	49	7.4742	0.025764	0.14462
2	7.9	12	144	7.8625	0.037526	0.20115
3	8.4	18	324	8.3199	0.080087	0.41234
4	10.1	45	2025	10.2643	-0.164321	-0.78694
5	11.5	67	4489	11.7107	-0.210653	-1.00600
6	12.8	88	7744	12.9756	-0.175645	-0.84295
7	13.4	92	8464	13.2038	0.196205	0.94185
8	13.9	99	9801	13.5932	0.306801	1.47206
9	13.5	105	11025	13.9170	-0.416988	-1.99805
10	14.5	108	11664	14.0754	0.424573	2.03240
11	14.7	120	14400	14.6861	0.013862	0.06599
12	15.8	147	21609	15.9254	-0.125438	-0.59224
13	16.1	156	24336	16.2971	-0.197060	-0.93540
14	16.5	159	25281	16.4163	0.083673	0.39861
15	16.9	167	27889	16.7231	0.176897	0.85592
16	17.3	183	33489	17.2875	0.012501	0.06499
17	17.6	195	38025	17.6678	-0.067781	-0.39812