CHAPTER 12

Linear Regression

We know a thing when we understand it.

George Berkeley (British philosopher for whom the
California city was named)

You can observe a lot just by watching.

Yogi Berra

CONTENTS

Introductory Statistics, DOI: 10.1016/B978-0-12-374388-6.00012-0

We study the simple linear regression model which, except for random error, assumes a straight-line relationship between a response and an input variable. We use the method of least squares to estimate the parameters of this model. Assuming that the random error is normal with mean 0 and variance σ^2, we show how to test hypotheses concerning the parameters of the model. The concept of regression to the mean is introduced; we explain when it arises and how one must be careful to avoid the regression fallacy in its presence. We explain the coefficient of determination. Finally, we introduce the multiple linear regression model, which relates a response variable to a set of input variables.

It was a spring day in 1888, and Francis Galton was out for a stroll in the countryside. While walking, he considered a question that had concerned him for some time. What was the relationship between a child's physical and mental characteristics and those of the child's parents? For instance, simplifying his ideas somewhat, Galton believed that the height of a child at adulthood should have an expected value equal to the height of his or her (same-sex) parent. But if this were so, then it would follow that about one-half of the offspring of very tall (short) people would be even taller (shorter) than their parents. Thus each new generation should produce taller (as well as shorter) people than the previous generation. However, on the contrary, data indicated a stability in the heights of the population from generation to generation. How could this apparent contradiction be explained?

> It came to Galton in a flash. In his own words, "A temporary shower drove me to seek refuge by a recess in the rock by the side of a pathway. There the idea flashed across me and I forgot everything else for a moment in my great delight."

Galton's flash of insight was that the mean value of a child's characteristic (such as height) was not equal to his or her parent's height but rather was between this value and the average value of the entire population. Thus, for instance, the heights of the offspring of very tall people (called, by Galton, people "taller than mediocrity") would tend to be shorter than their parents. Similarly, the offspring of those shorter than mediocrity would tend to be taller than their parents. Galton called this insight "regression to mediocrity"; we call it *regression to the mean*.

12.1 INTRODUCTION

We are often interested in trying to determine the relationship between a pair of variables. For instance, how does the amount of money spent in advertising a new product relate to the first month's sales figures for that product? Or how does the amount of catalyst employed in a scientific experiment relate to the yield of that experiment? Or how does the height of a father relate to that of his son?

In many situations the values of the variables are not determined simultaneously in time; rather, one of the variables will be set at some value, and this will, in turn, affect the value of the second variable. For instance, the advertising budget would be set before the sales figures are determined, and the amount of catalyst to be used would be set before the resulting yield could be determined. The variable whose value is determined first is called the *input* or *independent* variable and the other is called the *response* or *dependent* variable.

Suppose that the value of the independent variable is set to equal x. Let Y denote the resulting value of the dependent variable. The simplest type of relationship between this pair of variables is a straight-line, or *linear*, relation of the form

$$Y = \alpha + \beta x \tag{12.1}$$

This model, however, supposes that (once the values of the parameters α and β are determined) it would be possible to predict exactly the response for any value of the input variable. In practice, however, such precision is almost never attainable, and the most that one can expect is that the preceding equation is valid *subject to random error*.

In Sec. 12.2 we explain precisely the meaning of the *linear regression* model, which assumes that Eq. (12.1) is valid, subject to random error. In Sec. 12.3 we show how data can be used to estimate the regression parameters α and β. The estimators presented are based on the least-squares approach to finding the best straight-line fit for a set of data pairs. Section 12.4 deals with the *error* random variable, which will be taken to be a normal random variable having mean 0 and variance σ^2. The problem of estimating σ^2 will be considered.

In Sec. 12.5 we consider tests of the statistical hypothesis that there is no linear relationship between the response variable Y and the input value x. Section 12.6 deals with the concept of *regression to the mean*. It is shown that this phenomenon arises when the value of the regression parameter β is between 0 and 1. We explain how this phenomenon will often occur in testing–retesting situations and how a careless analysis of such data can often lead one into the *regression fallacy*. In addition, we indicate in this section how regression to the mean, in conjunction with the central limit theorem and the passing of many generations, can be used to explain why biological data sets are so often approximately normally distributed.

Section 12.7 is concerned with determining an interval that, with a fixed degree of confidence, will contain a future response corresponding to a specified input. These intervals, which make use of previously obtained data, are known as *prediction intervals*. Sections 12.8 and 12.9 present, respectively, the coefficient of determination and the correlation coefficient. Both quantities can be used to indicate the degree of fit of the linear regression model to the data. An approach to assessing the validity of the linear regression model, by analyzing the residuals, is dealt with in Sec. 12.10.

In Sec. 12.11 we consider the multiple linear regression model, where one tries to predict a response not on the basis of the value of a single input variable but on the basis of the values of two or more such variables.

12.2 SIMPLE LINEAR REGRESSION MODEL

Consider a pair of variables, one of which is called the *input variable* and the other the *response variable*. Suppose that for a specified value x of the input variable the value of the response variable Y can be expressed as

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

The quantities α and β are parameters. The variable e, called the *random error*, is assumed to be a random variable having mean 0.

Definition *The relationship between the response variable* Y *and the input variable* x *specified in the preceding equation is called a* simple linear regression.

The simple linear regression relationship can also be expressed by stating that for any value x of the input variable, the response variable Y is a random variable with mean given by

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

Thus a simple linear regression model supposes a straight-line relationship between the mean value of the response and the value of the input variable. Parameters α and β will almost always be unknown and will have to be estimated from data.

To see if a simple linear regression might be a reasonable model for the relationship between a pair of variables, one should first collect and then plot data on the paired values of the variables. For instance, suppose there is available a set of data pairs (x_i, y_i), $i = 1, \ldots, n$, meaning that when the input variable was set to equal x_i, the observed value of the response variable was y_i. These points should then be plotted to see if, subject to random error, a straight-line relationship between x and y appears to be a reasonable assumption. The resulting plot is called a *scatter diagram*.

■ Example 12.1

A new type of washing machine was recently introduced in 11 department stores. These stores are of roughly equal size and are located in similar types of communities. The manufacturer varied the price charged in each store, and the following data, giving the number of units sold in 1 month for each of the different prices, resulted.

Price ($)	Number sold
280	44
290	41
300	34
310	38
320	33
330	30
340	32
350	26
360	28
370	23
380	20

A plot of the number of units sold y versus the price x for these 11 data pairs is given in Fig. 12.1. The resulting scatter diagram indicates that, subject to random error, the assumption of a straight-line relationship between the number of units sold and the price appears to be reasonable. That is, a simple linear regression model appears to be appropriate.

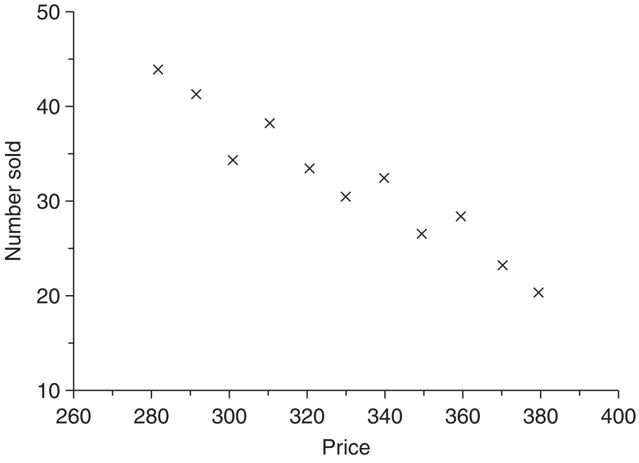

FIGURE 12.1

A scatter diagram for the data of Example 12.1.

■

Remark *The input variable x in the linear regression model is not usually thought of as being a random variable. Rather it is regarded as a constant that can be set at various values. The resulting response Y, on the other hand, is regarded as a random variable whose mean value depends in a linear way on the input x. It is for this reason that we use the capital, or uppercase, letter Y to represent the response. We use a small, or lowercase, y to denote an observed value of Y, and so y will represent the observed value of the response at the input value x.*

PROBLEMS

1. The following 12 data pairs relate y, the percentage yield of a laboratory experiment, to x, the temperature at which the experiment was conducted.

i	x_i	y_i	i	x_i	y_i
1	100	45	7	150	69
2	110	51	8	160	74
3	120	54	9	170	78
4	125	53	10	180	86
5	130	59	11	190	89
6	140	63	12	200	94

(a) Represent these data in a scatter diagram.

(b) Do you think a simple linear regression model would be appropriate for describing the relationship between percentage yield and temperature?

2. An area manager in a department store wants to study the relationship between the number of workers on duty and the value of merchandise lost to shoplifters. To do so, she assigned a different number of clerks for each of 10 weeks. The results were as follows:

Week	Number of workers	Loss
1	9	420
2	11	350
3	12	360
4	13	300
5	15	225
6	18	200
7	16	230
8	14	280
9	12	315
10	10	410

(a) Which variable should be the input variable and which should be the response?
(b) Plot the data in a scatter diagram.
(c) Does a simple linear regression model appear reasonable?

3. The following data relate the traffic density, described in the number of automobiles per mile, to the average speed of traffic on a moderately large city thoroughfare. The data were collected at the same location at 10 different times within a span of 3 months.

Density	Speed
69	25.4
56	32.5
62	28.6
119	11.3
84	21.3
74	22.1
73	22.3
90	18.5
38	37.2
22	44.6

(a) Which variable is the input and which is the response?
(b) Draw a scatter diagram.
(c) Does a simple linear regression model appear to be reasonable?

4. Repeat Prob. 3, but now let the square root of the speed, rather than the speed itself, be the response variable.

5. The use that can be obtained from a tire is affected by the air pressure in the tire. A new type of tire was tested for wear at different pressures, with the following results:

Pressure (pounds per square inch)	Mileage (thousands of miles)
30	29.4
31	32.2
32	35.9
33	38.4
34	36.6
35	34.8
36	35.0
37	32.2
38	30.5
39	28.6
40	27.4

(a) Plot the data in a scatter diagram.

(b) Does a simple linear regression model appear appropriate for describing the relation between tire pressure and miles of use?

12.3 ESTIMATING THE REGRESSION PARAMETERS

Suppose that the responses Y_i corresponding to the input values x_i, $i = 1, \ldots, n$, are to be observed and used to estimate the parameters α and β in a simple linear regression model

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

To determine estimators of α and β, we reason as follows: If A and B were the respective estimators of α and β, then the estimator of the response corresponding to the input value x_i would be $A + Bx_i$. Since the actual response is Y_i, it follows that the difference between the actual response and its estimated value is given by

$$\epsilon_i \equiv Y_i - (A + Bx_i)$$

That is, ϵ_i represents the error that would result from using estimators A and B to predict the response at input value x_i (Fig. 12.2).

Now, it is reasonable to choose our estimates of α and β to be the values of A and B that make these errors as small as possible. To accomplish this, we choose A and B to minimize the value of $\sum_{i=1}^{n} \epsilon_i^2$, the sum of the squares of the errors. The resulting estimators of α and β are called *least-square estimators*.

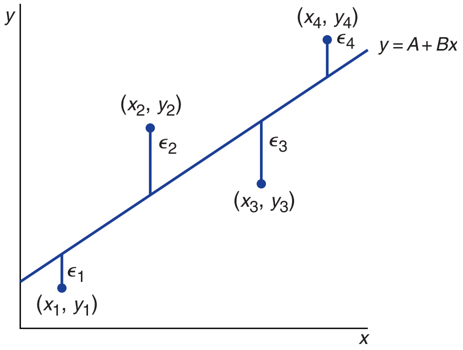

FIGURE 12.2

The errors.

Definition *For given data pairs* $(\mathrm{x_i}, \mathrm{Y_i}), \mathrm{i} = 1, \ldots, \mathrm{n}$, *the least-square estimators of* α *and* β *are the values of* A *and* B *that make*

$$\sum_{i=1}^{n} \epsilon_1^2 = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

as small as possible.

Remark *The reason we want* $\sum_{i=1}^{n} \epsilon_i^2$, *rather than* $\sum_{i=1}^{n} \epsilon_i$, *to be small is that the sum of the errors can be small even when individual error terms are large (since large positive and large negative errors cancel). On the other hand, this could not happen with the sum of the* ***squares*** *of the errors since none of the terms could be negative.*

It can be shown that the least-squares estimators of α and β, which we call $\hat{\alpha}$ and $\hat{\beta}$, are given by

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$$

where

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{and} \quad \overline{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$$

The line

$$y = \hat{\alpha} + \hat{\beta}x$$

is called the *estimated regression line:* $\hat{\beta}$ is the slope, and $\hat{\alpha}$ is the intercept of this line.

Notation: If we let

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

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

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

then the least-squares estimators can be expressed as

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$$

The values of $\hat{\alpha}$ and $\hat{\beta}$ can be obtained either by a pencil-and-paper computation or by using a hand calculator. In addition, Program 12-1 will compute the least-squares estimators and the estimated regression line. This program also gives the user the option of computing some other statistics whose values will be needed in the following sections.

■ Example 12.2

A large midwestern bank is planning on introducing a new word processing system to its secretarial staff. To learn about the amount of training that is needed to effectively implement the new system, the bank chose eight employees of roughly equal skill. These workers were trained for different amounts of time and were then individually put to work on a given project. The following data indicate the training times and the resulting times (both in hours) that it took each worker to complete the project.

Worker	Training time (= x)	Time to complete project (= Y)
1	22	18.4
2	18	19.2
3	30	14.5
4	16	19.0
5	25	16.6
6	20	17.7
7	10	24.4
8	14	21.0

(a) What is the estimated regression line?
(b) Predict the amount of time it would take a worker who receives 28 hours of training to complete the project.
(c) Predict the amount of time it would take a worker who receives 50 hours of training to complete the project.

Solution

(a) Rather than calculating by hand (which you will be asked to do in Prob. 2), we run Program 12-1, which computes the least-squares estimators and related statistics in simple linear regression models. We obtain the following:

First, enter the number of data pairs n, which is 8.

Next, enter the 8 successive pairs, which are:

22,	18.4
18,	19.2
30,	14.5
16,	19
25,	16.6
20,	17.7
10,	24.4
14,	21

The program computes the least-squares estimators as follows:

$$A = 27.46606$$

$$B = -0.4447002$$

The estimated regression line is as follows:

$$Y = 27.46606 - 0.4447002x$$

A plot of the scatter diagram and the resulting estimated regression line is given in Fig. 12.3.

(b) The best prediction of the completion time corresponding to the training time of 28 hours is its mean value, namely,

$$\alpha + 28\beta$$

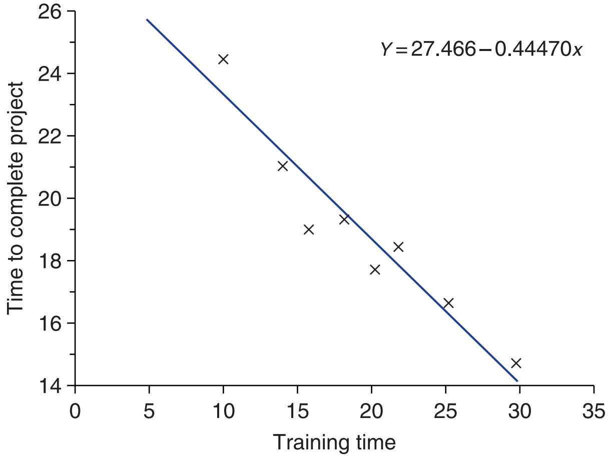

FIGURE 12.3

A scatter diagram and the estimated regression line.

By using the estimates of α and β previously derived, the predicted completion time is

$$27.466 - 28(0.445) = 15.006$$

(c) This part asks for the prediction at the input value 50, which is far greater than all the input values in our data set. As a result, even though the scatter diagram indicates that a straight-line fit should be a reasonable approximation for the range of input values considered, one should be extremely cautious about assuming that the relationship will continue to be a straight line for input values as large as 50. Thus, it is prudent not to attempt to answer part (c) on the basis of the available data. ■

Warning: *Do not use the estimated regression line to predict responses at input values that are far outside the range of the ones used to obtain this line.*

The following formulas can be useful when you are computing by hand.

$$S_{xY} = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}$$

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

PROBLEMS

1. Find, by a hand computation, the estimated regression line for the following data:

x	*y*
1	4
2	7
3	8
5	12

(a) Plot the scatter diagram, and draw the estimated regression line.
(b) Double all the data values and repeat part (a).

2. Verify the value given in Example 12.2 for the estimated regression line either by a pencil-and-paper computation or by using a hand calculator.
3. The following pairs of data represent the amounts of damages (in thousands of dollars) in fires at middle-class residences in a certain city and the distances (in miles) from these residences to the nearest fire station.

Distance	Damage
5.2	36.2
4.4	28.8
3.0	22.6
1.2	8.8
7.5	41.5
9.4	25.4

(a) Draw a scatter diagram.

(b) Try to approximate the relationship between the distance and damage by drawing a straight line through the data.

(c) Find the estimated regression line, and compare it to the line drawn in part (b).

4. Consider Prob. 1 of Sec. 12.2.

(a) Draw a straight line through the data points.

(b) Determine the estimated regression line, and compare it to the line drawn in part (a).

5. The amounts (in millions of pounds) of poultry products consumed in the United States for the years 1995 through 2002 are as follows:

25.9 26.8 27.3 27.8 29.6 30.5 30.8 32.6

(a) Letting the year be the independent variable and consumption the dependent variable, plot a scatter diagram.

(b) Find the estimated regression line. (To simplify the algebra, you can take the x variable to be the year minus 1995. That is, year 1985 has value 0, 1996 has value 1, and so on.)

(c) Plot the estimated regression line on the scatter diagram.

(d) Predict the 1994 consumption figure.

(e) Predict the 2004 consumption figure.

6. The following are the average 2003 math SAT scores in a sample of states, along with the percentage of graduating seniors who took the test.

State	Average score	Percentage
Arizona	525	38
California	519	54
Indiana	504	63
Missouri	583	8
Louisiana	559	8
Oregon	527	57
Virginia	510	71
Wisconsin	594	7
Texas	491	57
Vermont	512	70

Find the estimated regression line.

7. The following table relates the world production of wood pulp to the world production of newsprint for each of seven different years. The data come from the Statistical Division of the United Nations, New York, *Monthly Bulletin of Statistics* and are in units of 1 million metric tons.

Wood pulp	Newsprint
124.4	25.4
131.3	27.8
133.1	28.3
136.6	29.3
142.0	30.6
150.1	32.3
150.3	33.1

(a) Taking the amount of wood pulp as the independent (or input) variable, find the estimated regression line.

(b) Predict the amount of newsprint produced in a year in which 146.0 million metric tons of wood pulp is produced.

(c) Taking the amount of newsprint as the input variable, find the estimated regression line.

(d) Predict the amount of wood pulp produced in a year in which 32.0 million metric tons of newsprint is produced.

8. It is believed that the more alcohol there is in an individual's bloodstream, the slower is that person's reaction time. To test this, 10 volunteers were given different amounts of alcohol. Their blood alcohol levels were determined as percentages of their body weights. The volunteers were then tested to determine their reaction times to a given stimulus. The following data resulted.

x = amount of alcohol in blood (percent)	y = reaction time (seconds)
0.08	0.32
0.10	0.38
0.12	0.44
0.14	0.42
0.15	0.47
0.16	0.51
0.18	0.63

(a) Plot a scatter diagram.

(b) Approximate the estimated regression line by drawing a straight line through the data.

(c) What is the estimated regression line?

(d) Compare the lines in parts (b) and (c). Are their slopes nearly equal? How about the intercepts?

Predict the reaction time for an individual (not one of the original volunteers) whose blood alcohol content is

(e) 0.15

(f) 0.17

9. In Example 12.2, suppose the eight training times had been chosen in advance. How do you think the decision as to the assignment of the eight workers to these training times should have been made?

10. In an experiment designed to study the relationship between the number of alcoholic drinks consumed and blood alcohol concentration, seven individuals having the same body size were randomly assigned a certain number of alcoholic drinks. After a wait of 1 hour, their blood alcohol levels were checked. The results were as follows.

Number of drinks	Blood alcohol level
0.5	0.01
1	0.02
2	0.05
3	0.09
4	0.10
5	0.14
6	0.20

(a) Draw a scatter diagram.

(b) Find the estimated regression line, and draw it on the scatter diagram.

(c) Predict the blood alcohol level of a person, of the same general size as the people in the experiment, who had 3 drinks 1 hour ago.

(d) What if the person in part (b) had 7 drinks 1 hour ago?

11. The following data relate the per capita consumption of cigarettes in 1930 and men's death rates from lung cancer in 1950, for a variety of countries.

Country	1930 Per capita cigarette consumption	1950 Deaths per million men
Australia	480	180
Canada	500	150
Denmark	380	170
Finland	1100	350

(*Continued*)

(*Continued*)

Country	1930 Per capita cigarette consumption	1950 Deaths per million men
Great Britain	1100	460
Iceland	230	60
The Netherlands	490	240
Norway	250	90
Sweden	300	110
Switzerland	510	250
United States	1300	200

(a) Determine the estimated regression line.

Predict the number of 1950 lung cancer deaths per million men in a country whose 1930 per capita cigarette consumption was

(b) 600

(c) 850

(d) 1000

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

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

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

13. Use the data of Prob. 3 in Sec. 3.7 to predict the IQ of the daughter of a woman having an IQ of 130.

14. Use the data of Prob. 6 in Sec. 3.7 to predict the number of adults on parole in a state having 14,500 adults in prison.

15. The following data relate the proportions of coal miners who exhibit symptoms of pneumoconiosis to the number of years of working in coal mines. Use it to estimate the probability that a coal miner who has worked for 42 years will have pneumoconiosis.

Years working	Proportion having pneumoconiosis
5	0
10	0.0090
15	0.0185
20	0.0672
25	0.1542
30	0.1720
35	0.1840
40	0.2105
45	0.3570
50	0.4545

12.4 ERROR RANDOM VARIABLE

We have defined the linear regression model by the relationship

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

where α and β are unknown parameters that will have to be estimated and e is an error random variable having mean 0. To be able to make statistical inferences about the regression parameters α and β, it is necessary to make some additional assumptions concerning the error random variable e. The usual assumption, which we will be making, is that e is a normal random variable with mean 0 and variance σ^2. Thus we are assuming that the variance of the error random variable remains the same no matter what input value x is used.

Put another way, this assumption is equivalent to assuming that for any input value x, the response variable Y is a random variable that is normally distributed with mean

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

and variance

$$\text{Var}(Y) = \sigma^2$$

An additional assumption we will make is that all response variables are independent. That is, for instance, the response from input value x_1 will be assumed to be independent of the response from input value x_2.

The quantity σ^2 is an unknown that will have to be estimated from the data. To see how this can be accomplished, suppose that we will be observing the response values Y_i corresponding to the input values x_i, $i = 1, \ldots, n$. Now, for each value of i, the standardized variable

$$\frac{Y_i - E[Y_i]}{\sqrt{\operatorname{Var}(Y_i)}} = \frac{Y_i - (\alpha + \beta x_i)}{\sigma}$$

will have a standard normal distribution. Thus, since a chi-squared random variable with n degrees of freedom is defined to be the sum of the squares of n independent standard normals, we see that

$$\frac{\sum_{i=1}^{n} (Y_i - \alpha - \beta x_i)^2}{\sigma^2}$$

is chi squared with n degrees of freedom.

If we now substitute the estimators $\hat{\alpha}$ and $\hat{\beta}$ for α and β in the preceding expression, then the resulting variable will remain chi squared but will now have $n - 2$ degrees of freedom (since 1 degree of freedom will be lost for each parameter that is estimated). That is,

$$\frac{\sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2}{\sigma^2}$$

is chi squared with $n - 2$ degrees of freedom.

The quantities

$$Y_i - \hat{\alpha} - \hat{\beta} x_i \quad i = 1, \ldots, n$$

are called *residuals*. They represent the differences between the actual and the predicted responses. We will let $\mathrm{SS_R}$ denote the sum of the squares of these residuals. That is,

$$\mathrm{SS_R} = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

From the preceding result, we thus have

$$\frac{\mathrm{SS_R}}{\sigma^2}$$

is chi squared with $n - 2$ degrees of freedom.

Since the expected value of a chi-squared random variable is equal to its number of degrees of freedom, we obtain

$$\frac{E[\mathrm{SS}_R]}{\sigma^2} = n - 2$$

or

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

In other words, $\mathrm{SS}_R/(n-2)$ can be used to estimate σ^2.

$$\frac{\mathrm{SS}_R}{n-2}$$

is the estimator of σ^2.

Program 12-1 can be utilized to compute the value of SS_R.

■ Example 12.3

Consider Example 12.2 and suppose that we are interested in estimating the value of σ^2. To do so, we could again run Program 12-1, this time asking for the additional statistics. This would result in the following additional output:

$$\begin{aligned} &\mathrm{S(x,Y)} = -125.3499 \\ &\mathrm{S(x,x)} = 281.875 \\ &\mathrm{S(Y,Y)} = 61.08057 \\ &\mathrm{SS}_R \quad = 5.337465 \end{aligned}$$

THE SQUARE ROOT OF $(n-2)\mathrm{S(x, x)}/\mathrm{SS}_R$ is 17.80067

The estimate of σ^2 is $5.3375/6 = 0.8896$. ■

The following formula for SS_R is useful when you are using a calculator or computing by hand.

Computational formula for SS_R:

$$\mathrm{SS}_R = \frac{S_{xx}S_{YY} - S_{xY}^2}{S_{xx}}$$

The easiest way to compute SS_R by hand is first to determine S_{xx}, S_{xY}, and S_{YY} and then to apply the preceding formula.

PROBLEMS

1. Estimate σ^2 in Prob. 1 of Sec. 12.2.
2. Estimate σ^2 in Prob. 2 of Sec. 12.2.
3. The following data relate the speed of a particular typist and the temperature setting of his office. The units are words per minute and degrees Fahrenheit.

Temperature	Typing speed
50	63
60	74
70	79

   (a) Compute, by hand, the value of SS_R.
   (b) Estimate σ^2.
   (c) If the temperature is set at 65, what typing speed would you predict?
4. Estimate σ^2 in Prob. 3 of Sec. 12.2.
5. Estimate σ^2 in Prob. 10 of Sec. 12.3.
6. The following data give, for certain years between 1982 and 2002, the percentages of British women who were cigarette smokers.

Year	1982	1984	1988	1990	1994	1996	1998	2000	2002
Percentage	33.1	31.8	30.4	24.3	26.3	27.7	26.3	25.3	24.8

   Treat these data as coming from a linear regression model, with the input being the year and the response being the percentage. Take 1982 as the base year, so 1982 has input value $x = 0$, 1986 has input value $x = 4$, and so on.
   (a) Estimate the value of σ^2.
   (b) Predict the percentage of British women who smoked in 1997.
7. Estimate σ^2 in Prob. 11 of Sec. 12.3.
8. In data relating the ages at which 25 fathers (x) and their respective sons (Y) first began to shave, the following summary statistics resulted:

$$\overline{x} = 13.9 \quad \overline{Y} = 14.6$$

$$S_{xx} = 46.8 \quad S_{YY} = 53.3 \quad S_{xY} = 12.2$$

   (a) Determine the estimated regression line.
   (b) Predict the age at which a boy will begin to shave if his father began to shave at age 15.1 years.
   (c) Estimate σ^2.

12.5 TESTING THE HYPOTHESIS THAT $\beta = 0$

An important hypothesis to consider with respect to the simple linear regression model

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

is the hypothesis that $\beta = 0$. Its importance lies in the fact that it is equivalent to stating that a response does not depend on the value of the input; or, in other words, there is no regression on the input value.

To derive a test of

$$\text{H}_0\colon \beta = 0 \quad \text{against} \quad \text{H}_1\colon \beta \neq 0$$

first it is necessary to study the distribution of $\hat{\beta}$, the estimator of β. That is, we will clearly want to reject H_0 when $\hat{\beta}$ is far from 0 and not to reject it otherwise. To determine how far away $\hat{\beta}$ needs be from 0 to justify rejection of the null hypothesis, it is necessary to know something about its distribution.

It can be shown that $\hat{\beta}$ is normally distributed with mean and variance, respectively, given by

$$E[\hat{\beta}] = \beta$$

and

$$\text{Var}(\hat{\beta}) = \frac{\sigma^2}{S_{xx}}$$

Hence, the standardized variable

$$\frac{\hat{\beta} - \beta}{\sqrt{\sigma^2/S_{xx}}} = \sqrt{\frac{S_{xx}}{\sigma^2}}(\hat{\beta} - \beta)$$

will have a standard normal distribution.

We cannot directly base a test on the preceding fact, however, since the standardized variable involves the unknown parameter σ^2. However, if we replace σ^2 by its estimator $\text{SS}_R/(n-2)$, which is chi squared with $n-2$ degrees of freedom, then it can be shown that the resulting quantity will now have a t distribution with $n-2$ degrees of freedom. That is,

$$\sqrt{\frac{(n-2)S_{xx}}{\text{SS}_R}}(\hat{\beta} - \beta)$$

has a t distribution with $n-2$ degrees of freedom.

It follows from the preceding that if H_0 is true and so $\beta = 0$, then

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}\,\hat{\beta}$$

has a t distribution with $n - 2$ degrees of freedom. This gives rise to the following test of H_0.

A significance-level-γ test of H_0 is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } \quad |\,\mathrm{TS}\,| \geq t_{n-2,\gamma/2} \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

where

$$\mathrm{TS} = \sqrt{\frac{(n-2)S_{xx}}{SS_R}}\,\hat{\beta}$$

An equivalent way of performing this test is first to compute the value of the test statistic TS; say its value is v. The null hypothesis should then be rejected if the desired significance level γ is at least as large as the p value given by

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

where T_{n-2} is a t random variable with $n - 2$ degrees of freedom. Program 8-2 can be used to compute this latter probability.

■ Example 12.4

An individual claims that the fuel consumption of his automobile does not depend on how fast the car is driven. To test the plausibility of this hypothesis, the car was tested at various speeds between 45 and 75 miles per hour. The miles per gallon attained at each of these speeds were determined, with the following data resulting.

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

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

Solution

Suppose that a simple linear regression model

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

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

$$\text{H}_0\colon \beta = 0 \quad \text{against} \quad \text{H}_1\colon \beta \neq 0$$

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

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

The computational formula for SS_R presented at the end of Sec. 12.4 gives

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

Since

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}} = \frac{-119}{700} = -0.17$$

the value of the test statistic is

$$\text{TS} = \sqrt{\frac{5(700)}{1.527}}(-0.17) = -8.139$$

Since from App. Table D.2 $t_{5,0.005} = 4.032$, it follows that the hypothesis that $\beta = 0$ is rejected at the 1 percent level of significance. Thus, the claim that the mileage does not depend on the speed at which the car is driven is rejected. Indeed, there is clearly strong evidence that increased speeds lead to decreased efficiencies. ■

PROBLEMS

1. Test the hypothesis that $\beta = 0$ for the following data.

x	Y
3	7
8	8
10	6
13	7

Use the 5 percent level of significance.

2. The following data set presents the heights of 12 male law school classmates whose law school examination scores were roughly equal. It also gives their annual salaries 5 years after graduation. Each went into corporate law. The height is in inches, and the salary is in units of $1000.

Height	Salary
64	111
65	114
66	108
67	123
69	97
70	116
72	125
72	108
74	142
74	122
75	110
76	134

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

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

3. The following table relates the number of sunspots that appeared each year from 1970 to 1980 to the number of automobile accident deaths during that year. The data for automobile accident deaths are in units of 1000 deaths.

Year	Sunspots	Automobile deaths
70	165	54.6
71	89	53.3
72	55	56.3
73	34	49.6
74	9	47.1
75	30	45.9
76	59	48.5
77	83	50.1
78	109	52.4
79	127	52.5
80	153	53.2

Test the hypothesis that the number of automobile accident deaths is not related to the number of sunspots. Use the 5 percent level of significance.

4. An electric utility wants to estimate the relationship between the daily summer temperature and the amount of electricity used by its customers. The following data were collected.

Temperature (degrees Fahrenheit)	Electricity (millions of kilowatts)
85	22.5
90	23.7
76	20.3
91	23.4
84	24.2
94	23.5
88	22.9
85	22.4
97	26.1
86	23.1
82	22.5
78	20.9
77	21.0
83	22.6

(a) Find the estimated regression line.

(b) Predict the electricity that will be consumed tomorrow if the predicted temperature for tomorrow is 93.

(c) Test the hypothesis, at the 5 percent level of significance, that the daily temperature has no effect on the amount of electricity consumed.

Problems 5 through 8 refer to the following data relating cigarette smoking and death rates for four types of cancers in 14 states. The data are based in part on records concerning 1960 cigarette tax receipts.

Cigarette Smoking and Cancer Death Rates

		Deaths per year per 100,000 people			
State	Cigarettes per person	Bladder cancer	Lung cancer	Kidney cancer	Leukemia
California	2860	4.46	22.07	2.66	7.06
Idaho	2010	3.08	13.58	2.46	6.62
Illinois	2791	4.75	22.80	2.95	7.27
Indiana	2618	4.09	20.30	2.81	7.00
Iowa	2212	4.23	16.59	2.90	7.69
Kansas	2184	2.91	16.84	2.88	7.42
Kentucky	2344	2.86	17.71	2.13	6.41
Massachusetts	2692	4.69	22.04	3.03	6.89
Minnesota	2206	3.72	14.20	3.54	8.28
New York	2914	5.30	25.02	3.10	7.23
Alaska	3034	3.46	25.88	4.32	4.90
Nevada	4240	6.54	23.03	2.85	6.67
Utah	1400	3.31	12.01	2.20	6.71
Texas	2257	3.21	20.74	2.69	7.02

5. (a) Draw a scatter diagram of cigarettes smoked versus death rate from bladder cancer.

(b) Find the estimated regression line.

(c) Test the hypothesis, at the 5 percent level of significance, that cigarette consumption does not affect the death rate from bladder cancer.

(d) Repeat part (c) at the 1 percent level of significance.

6. (a) Draw a scatter diagram of cigarettes smoked versus death rate from lung cancer.

(b) Find the estimated regression line.

(c) Test the hypothesis, at the 5 percent level of significance, that cigarette consumption does not affect the death rate from lung cancer.

(d) Repeat part (c) at the 1 percent level of significance.

7. (a) Draw a scatter diagram of cigarettes smoked versus death rate from kidney cancer.
   (b) Find the estimated regression line.
   (c) Test the hypothesis, at the 5 percent level of significance, that cigarette consumption does not affect the death rate from kidney cancer.
   (d) Repeat part (c) at the 1 percent level of significance.
8. (a) Draw a scatter diagram of cigarettes smoked versus death rate from leukemia.
   (b) Find the estimated regression line.
   (c) Test the hypothesis, at the 5 percent level of significance, that cigarette consumption does not affect the death rate from leukemia.
   (d) Repeat part (c) at the 1 percent level of significance.
9. In Prob. 3 of Sec. 12.3, test the null hypothesis that the amount of fire damage sustained by a property does not depend on its distance to the nearest fire station. Use the 5 percent level of significance.
10. The following table gives the percentages of 15-year-old British boys and girls who are smokers, in a sample of years from 1982 to 2003. Use it to
    (a) Test, at the 5 percent level of significance, the hypothesis that the percentage of the boys who smoke is unchanging over time.
    (b) Test, at the 5 percent level of significance, the hypothesis that the percentage of the girls who smoke is unchanging over time.
    (c) Test, at the 5 percent level of significance, the hypothesis that the percentage of 15-year-olds who smoke is unchanging over time.

Percentage of 15-Year-Old Pupils Who Are Regular Smokers (at least 1 cigarette/week on average), England

	1982	1984	1986	1988	1990	1992	1994	1996	1998	2000	2003
Boys	24	28	18	17	25	21	26	28	19	21	18
Girls	25	28	27	22	25	25	30	33	29	26	26
All	25	28	22	20	25	23	28	30	24	23	22

11. The following table gives the U.S. per capita consumption of bananas, apples, and oranges (in pounds) in seven different years.

Bananas	Apples	Oranges	Bananas	Apples	Oranges
17.4	16.2	15.7	21.2	17.6	15.6
17.6	18.2	15.4	23.4	16.6	12.0
20.8	18.3	15.4	24.9	20.3	13.9
22.5	17.1	12.3			

Source: U.S. Department of Agriculture, *Food Consumption, Prices and Expenditures.*

Test the hypotheses that the yearly amount of bananas consumed is unrelated to the yearly amount of

(a) Apples consumed

(b) Oranges consumed

(c) Test the hypothesis that the yearly per capita amount of oranges consumed is unrelated to the yearly amount of apples consumed.

12.6 REGRESSION TO THE MEAN

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

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

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

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

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

and

$$y = x$$

A plot indicates that when $0 < \beta < 1$, the line $y = \alpha + \beta x$ is above the line $y = x$ for small values of x and is below it for large values of x. Such a plot is given in Fig. 12.4.

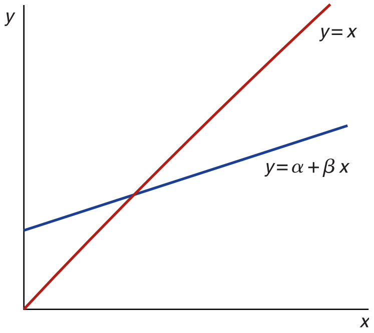

FIGURE 12.4
Regression to the mean occurs when $0 < \beta < 1$. For x small, $\alpha + \beta x > x$; for x large, $\alpha + \beta x < x$.

Example 12.5

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

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

A scatter diagram representing these data is presented in Fig. 12.5.

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

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

$$H_0\colon \beta \geq 1 \quad \text{against} \quad H_1\colon \beta < 1$$

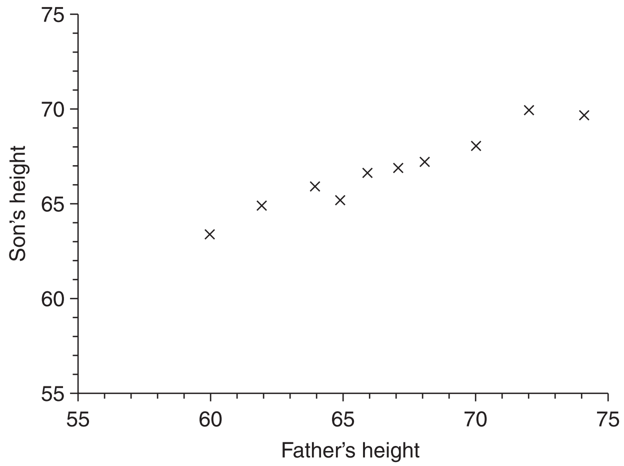

FIGURE 12.5

Scatter diagram of son's height versus father's height.

Now, this test is equivalent to a test of

$$H_0\colon \beta = 1 \quad \text{against} \quad H_1\colon \beta < 1$$

and will be based on the fact that

$$\sqrt{\frac{(n-2)S_{xx}}{SS_R}}(\hat{\beta} - \beta)$$

has a t distribution with $n - 2$ degrees of freedom.

Hence, when $\beta = 1$, the test statistic

$$TS = \sqrt{\frac{8S_{xx}}{SS_R}}(\hat{\beta} - 1)$$

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

$$\begin{array}{ll} \text{Reject } H_0 & \text{if } TS \leq -t_{8,a} \\ \text{Not reject } H_0 & \text{Otherwise} \end{array}$$

To determine the value of the test statistic TS, we run Program 12-1 and obtain the following:

The least-squares estimators are as follows

A = 35.97757

B = 0.4645573

The estimated regression line is

Y = 35.97757 + 0.4645573x

S(x,Y) = 79.71875

S(x,x) = 171.6016

S(Y,Y) = 38.53125

SS_R = 1.497325

The square root of $(n-2)S(x, x)/SS_R$ is 30.27942

From the preceding we see that

$$TS = 30.2794(0.4646 - 1) = -16.21$$

Since $t_{8,0.01} = 2.896$, we see that

$$\text{TS} < -t_{8,\,0.01}$$

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

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

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

A modern biological explanation for the phenomenon of regression to the mean would roughly go along the lines of noting that since an offspring obtains a random selection of one-half of each parent's genes, it follows that the offspring of a very tall parent would, by chance, tend to have fewer "tall" genes than its parent.

While the most important applications of the phenomenon of regression to the mean concern the relationship between the biological characteristics of an offspring and those of its parents, this phenomenon also arises in situations where we have two sets of data referring to the same variables. We illustrate it in Example 12.6.

■ Example 12.6

The following data relate the number of motor vehicle deaths occurring in 12 counties in the northwestern United States in the years 1988 and 1989.

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

The scatter diagram for this data set appears in Fig. 12.6. A glance at Fig. 12.6 indicates that in 1989 there was, for the most part, a reduction in the number

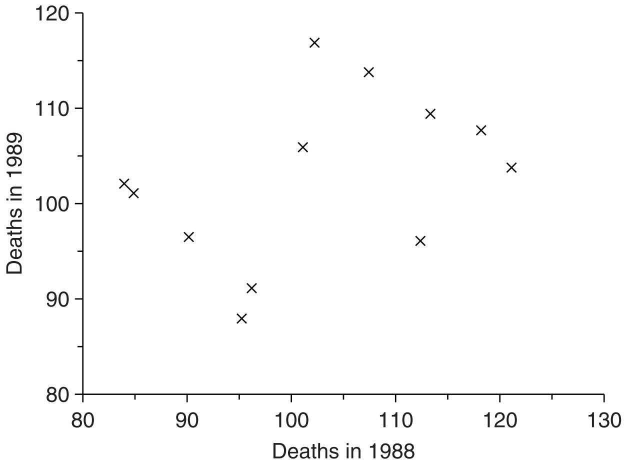

FIGURE 12.6

Scatter diagram of 1989 deaths versus 1988 deaths.

of deaths in those counties that had a large number of motor vehicle deaths in 1988. Similarly, there appears to have been an increase in those counties that had a low value in 1988. Thus, we would expect that a regression to the mean is in effect. In fact, running Program 12-1 yields the estimated regression equation

$$y = 74.589 + 0.276x$$

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

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

While the foregoing suppositions might be correct, it is important to realize that a regression to the mean would probably have occurred even if none of the counties had done anything out of the ordinary. Indeed, it could very well be the case that those counties having large numbers of casualties in 1988 were just very unlucky in that year, and thus a decrease in the next year was just a return to a more normal result for them. (For an analogy, if 9 heads result when

10 fair coins are flipped, then it is quite likely that another flip of these 10 coins will result in fewer than 9 heads.) Similarly, those counties having few deaths in 1988 might have been "lucky" that year, and a more normal result in 1989 would thus lead to an increase.

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

Regression to the mean plays a key role in the explanation of why so many communal biological data sets from a homogeneous population tend to have "normal curve" histograms. For instance, if one plotted the heights of all senior girls in a specified high school, then it is a good bet that the resulting histogram would strongly resemble the bell-shaped normal curve. One possible explanation for this combines the central limit theorem, regression to the mean, and the passing of many generations. We now sketch it.

*12.6.1 Why Biological Data Sets Are Often Normally Distributed

We will present this argument in the context of considering the heights of females in a population. We will follow this population of females over many generations. Suppose that there are initially k women, whom we will refer to as the *initial generation*, and that their heights are $x_1, \ldots, x_k$. These k values are considered to be totally arbitrary. Let d denote the largest minus the smallest of these values. For instance, if

$$k = 3 \quad x_1 = 60 \quad x_2 = 58 \quad x_3 = 66$$

then $d = 66 - 58 = 8$.

If Y denotes the height of a female child of a woman whose height is x, then we will assume the linear regression model

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

In this model we will make the usual assumption that e is an error random variable that is normally distributed with mean 0 and variance σ^2. However, whereas this assumption is often made without any real attempt at justification, it seems quite reasonable in this application because of the central limit theorem. That is, the height of a daughter of a woman of height x can be thought of as being composed of the sum of a large number of approximately independent random variables that relate, among other things, to the random set of genes that she receives as well as to environmental factors. Hence, by the central limit theorem, her height should be approximately normally distributed. We will also assume that regression to the mean is in effect, that is, that $0 < \beta < 1$.

Thus, the heights of the daughters of the k women of the initial population are all normally distributed. However, it is important to note that their mean heights are all different. For instance, a daughter of the woman of height x_1 will have a normally distributed height with mean value $\alpha + \beta x_1$, whereas the daughter of the woman of height x_2 will have a different mean height, namely, $\alpha + \beta x_2$. Thus, the heights of all the daughters do *not* come from the *same* normal distribution, and for that reason a plot of all their heights would not follow the normal curve.

However, if we now consider the difference between the largest and the smallest *mean* height of all the daughters of the initial set of women, then it is not difficult to show that

$$\text{Difference} \le \beta d$$

(If each woman of the initial set of women had at least one daughter, then this inequality would be an equality.) If we now consider the daughters of these daughters, then it can be shown that their heights will be normally distributed with differing means and a common variance. The difference between the largest and the smallest mean height of these second-generation daughters can be shown to satisfy

$$\text{Difference} \le \beta^2 d$$

Indeed, if we suppose that more and more generations have passed and we consider the women of the nth generation after the initial population, then it can be shown that the heights of the women in this generation are normally distributed with the same variance and with mean values that, while differing, are such that the difference between the largest and smallest of them satisfies

$$\text{Difference} \le \beta^n d$$

Now, since $0 < \beta < 1$, it follows that as n grows larger, $\beta^n d$ gets closer and closer to 0. Thus, after a large enough number of generations have passed, all the women in the population will have normally distributed heights with approximately the same mean and with a common variance. That is, after many generations have passed, the heights of the women will come from approximately the same normal population, and thus at this point a plot of these heights will approximately follow the bell-shaped normal curve.

PROBLEMS

1. The following data come from an experiment performed by Francis Galton. The data relate the diameter of an offspring seed to that of its parent seed in the case of a self-fertilized seed.

Diameter of parent seed	Diameter of offspring seed
15	15.3
16	16.4
17	15.5
18	16.2
19	16.0
20	17.4
21	17.5

(a) Estimate the regression parameters.
(b) Does there appear to be a regression to the mean?

2. In Example 12.6 it was shown that the estimated value of β is less than 1. Using the data of this example, test the hypothesis

$$\mathrm{H}_0\colon \beta \geq 1 \quad \text{against} \quad \mathrm{H}_1\colon \beta < 1$$

Would H_0 be rejected at the 5 percent level of significance?

3. Would you be surprised if the following data sets exhibited a regression to the mean? Would you expect them to exhibit this phenomenon? Explain your answers.
   (a) You go to 10 different restaurants that you know nothing about in advance. You eat a meal in each one and give a numerical ranking—anywhere from 0 to 100—to the quality of the meal. You then return to each of these restaurants and again give a ranking to the meal. The data consist of the two scores of each of the 10 restaurants.
   (b) At the beginning of an hour, 12 individuals check their pulse rates to determine the number of heartbeats per minute. Call these the x values. After 1 hour, they repeat this to obtain the y values.
   (c) The set considers paired data concerning different mutual funds. For each mutual fund, the x variable is the 1995 ranking of the fund, and the corresponding y variable is the 1996 ranking.
   (d) The data consist of the paired scores of 20 first-year preschoolers, with the first value in the pair being the student's test score on an IQ examination given to all entering students and the second value being the same student's score on an IQ test given at the end of the first month in school.

4. Test

$$\mathrm{H}_0\colon \beta = 1 \quad \text{against} \quad \mathrm{H}_1\colon \beta < 1$$

for the following set of data. Use the 5 percent level of significance.

x	y
24	27
21	24
26	20
17	22
15	21
24	20
23	17

5. The following are the average 2000 and 2002 math SAT scores in a sample of states.

State	2000	2002
Arizona	523	523
California	518	517
Indiana	501	503
Missouri	577	580
Florida	486	473
Oregon	527	528
Virginia	500	506
Wisconsin	597	599
Texas	500	500
Vermont	508	510

(a) Find the estimated regression line.
(b) Does it indicate a regression to the mean?

6. Figure 12.7 presents a histogram of the heights of 8585 men. How well does it appear to fit a normal curve?

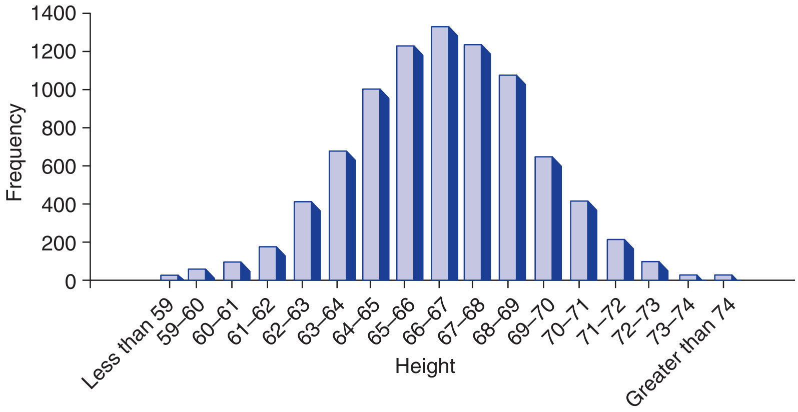

FIGURE 12.7

A histogram of heights.

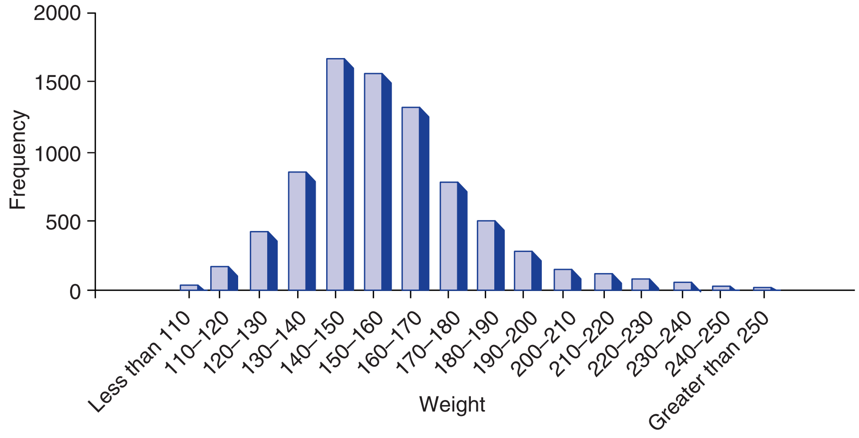

FIGURE 12.8

A histogram of weights.

7. Figure 12.8 presents a histogram of the weights of 7738 men. How well do these data fit a normal curve?

12.7 PREDICTION INTERVALS FOR FUTURE RESPONSES

Suppose, in the linear regression model, that input values x_i have led to the response values y_i, $i = 1, \ldots, n$. The best prediction of the value of a new response at input x_0 is, of course, $\hat{\alpha} + \hat{\beta}x_0$. However, rather than give a single number as the predicted value, it is often more useful to be able to present an interval that you predict, with a certain degree of confidence, will contain the response value. Such a *prediction interval* is given by the following.

Prediction interval for a response at input value x_0, based on the response values y_i at the input values x_i, $i = 1, \ldots, n$:

With $100(1 - \gamma)$ degree confidence, the response Y at the input value x_0 will lie in the interval

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\gamma/2}W$$

where $t_{n-2,\gamma/2}$ is the $100(1 - \gamma/2)$th percentile of the t distribution with $n - 2$ degrees of freedom, and

$$W = \sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\frac{SS_R}{n-2}}$$

The quantities $\hat{\alpha}, \hat{\beta}, \overline{x}, S_{xx}$, and SS_R are all computed from the data $x_i, y_i, i = 1, \ldots, n$.

■ Example 12.7

Using the data of Example 12.6, specify an interval that, with 95 percent confidence, will contain the adult height of a newborn son whose father is 70 inches tall.

Solution

From the output of Program 12-1, we obtain

$$\hat{\alpha} + 70\hat{\beta} = 68.497$$
$$W = 0.4659$$

Since from Table D.2, $t_{8,0.025} = 2.306$, we see that the 95 percent prediction interval of the height of the son of a 70-inch-tall man is

$$68.497 \pm 2.306(0.4659) = 68.497 \pm 1.074$$

That is, we can be 95 percent confident that the son's height will be between 67.423 and 69.571 inches. ■

■ Example 12.8

A company that runs a hamburger concession at a college football stadium must decide on Monday how much to order for the game that is to be played on the following Saturday. The company bases its order on the number of tickets for the game that have already been sold by Monday. The following data give the advance ticket sales and the number of hamburgers purchased for each game played this year. All data are in units of 1000.

Advance ticket sales	Hamburgers sold
29.4	19.5
21.4	16.2
18.0	15.3
25.2	18.0
32.5	20.4
23.9	16.8

If 26,000 tickets have been sold by Monday for next Saturday's game, determine a 95 percent prediction interval for the amount of hamburgers that will be sold.

Solution

Running Program 12-1 gives the following output, if we request predicted future responses and the value of the input is 26.

The predicted response is 18.04578.

W = 0.3381453

Since $t_{4,0.025} = 2.776$, we see from the output that the 95 percent prediction interval is

$$18.046 \pm 2.776(0.338) = 18.046 \pm 0.938$$

That is, with 95 percent confidence, between 17,108 and 18,984 hamburgers will be sold. ■

PROBLEMS

1. Use the following data to
   (a) Predict the response at the input value $x = 4$.
   (b) Determine an interval that contains, with 95 percent confidence, the response in part (a).

x	y
1	5
2	8
5	15

2. An official of a large automobile manufacturing firm wanted to study the relationship between a worker's age and his or her level of absenteeism. The following data concerning 10 randomly chosen employees were collected.

Age	40	28	34	27	21	38	19	55	31	35
Days missed	1	6	6	9	12	4	13	2	5	3

   (a) Predict the number of days missed by a worker aged 42.
   (b) Determine a 95 percent prediction interval for the quantity in part (a).

3. The following data were recently reported by an economist who wanted to learn about the relationship between a family's income and the proportion of that income spent on food. Each of the families consisted of a married couple with two teenage children.

Income (in $1000)	Percentage spent on food
28	35
36	33
44	32
56	29
70	23
78	19
84	17

(a) Find the estimated regression line.

(b) Predict the amount of money spent on food by a family of size 4 that earns 51,000 dollars annually.

(c) Determine a 95 percent confidence prediction interval for the amount in (b).

(d) Repeat part (c), but this time obtain a prediction interval having 99 percent confidence.

4. The following data relate the scores of 10 students on a college entrance examination to their grade-point average at the end of their first year.

Entrance examination score	Grade-point average
88	3.2
74	2.7
70	2.3
77	2.9
83	2.8
94	3.6
92	3.0
81	2.8
85	3.3
92	3.1

(a) Predict the grade-point average of a student, not listed in the given data, who scored 88 on the entrance examination.

(b) Obtain a 90 percent prediction interval for the score of the student described in part (a).

(c) Test the hypothesis, at the 5 percent level of significance, that a student's grade-point average is independent of her or his score on the entrance examination.

5. Glass plays an important role in criminal investigations, because criminal activity often results in the breakage of windows and other glass objects. Since glass fragments often lodge in the clothing of criminals, it

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

The following data relate the refractive index to the density for 12 selected pieces of glass.

Refractive index	Density	Refractive index	Density
1.514	2.480	1.516	2.484
1.515	2.482	1.517	2.486
1.516	2.480	1.518	2.495
1.517	2.490	1.519	2.498
1.517	2.482	1.522	2.511
1.520	2.505	1.525	2.520

(a) Predict the density of a fragment of glass whose refractive index is 1.520.

(b) Determine an interval that, with 95 percent confidence, will contain the density of a fragment of glass whose refractive index is 1.520.

6. The following summary data relate to the ages of puberty of 20 mother–daughter pairs. The x data refer to the mother's age and the Y data to her daughter's age at puberty.

$$\bar{x} = 12.8 \quad \overline{Y} = 12.9$$

$$S_{xx} = 36.5 \quad S_{YY} = 42.4 \quad S_{xY} = 24.4$$

(a) Find the estimated regression line.

(b) Use the computational formula given at the end of Sec. 12.4 to compute SS_R.

(c) Test, at the 5 percent level of significance, the hypothesis that $\beta = 0$.

(d) If a mother reached puberty at age 13.8, determine an interval that, with 95 percent confidence, will contain the age at which her daughter reaches puberty.

7. The following data relate the grade-point average (GPA) in accounting courses to the starting annual salary of eight 2004 accounting graduates.

Accounting GPA	Starting salary (in $1000)
3.4	42
2.5	29
3.0	33
2.8	32
3.7	40
3.5	44
2.7	30
3.1	35

(a) Predict the annual salary of a recent graduate whose grade-point average in accounting courses was 2.9.
(b) Determine an interval that, with 95 percent confidence, will contain the annual salary in part (a).
(c) Repeat parts (a) and (b) for a graduate having a 3.6 GPA.

12.8 COEFFICIENT OF DETERMINATION

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

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

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

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

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

$$SS_R = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta} x_i)^2$$

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

$$S_{YY} - SS_R$$

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

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

represents the proportion of variation in the response variables that is explained by the different input values.

Definition *The quantity* R^2 *is called the* coefficient of determination.

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

■ Example 12.9

In Example 12.5, which relates the height of a son to that of his father, the output from Program 12-1 yielded

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

Thus,

$$R^2 = 1 - \frac{1.497}{38.531} = 0.961$$

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

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

■ Example 12.10

In Example 12.8, which relates the number of hamburgers sold at a football game to the advance ticket sales for that game, Program 12-1 yielded

$$S_{YY} = 19.440 \qquad SS_R = 0.390$$

Thus,

$$R^2 = 1 - \frac{0.390}{19.440} = 0.98$$

and so 98 percent of the variation in the different amounts of hamburgers sold in the six games is explained by the advance ticket sales for these games. (Loosely put, 98 percent of the amount sold is explained by the advance ticket sales.) ■

PROBLEMS

1. A real estate brokerage gathered the following information relating the selling prices of three-bedroom homes in a particular neighborhood to the sizes of these homes. (The square footage data are in units of 1000 square feet, whereas the selling price data are in units of $1000.)

Square footage	Selling price
2.3	240
1.8	212
2.6	253
3.0	280
2.4	248
2.3	232
2.7	260

(a) Plot the data in a scatter diagram.
(b) Determine the estimated regression line.
(c) What percentage of the selling price is explained by the square footage?
(d) A house of size 2600 square feet has just come on the market. Determine an interval in which, with 95 percent confidence, the selling price of this house will lie.

2. Determine R^2 for the following data set:

x	y
2	10
3	16
5	22

3. It is difficult and time-consuming to directly measure the amount of protein in a liver sample. As a result, medical laboratories often make use of the fact that the amount of protein is related to the amount of light that would be absorbed by the sample. As a result, a spectrometer that

emits light is shined upon a solution that contains the liver sample, and the amount of light absorbed is then used to estimate the amount of protein.

This procedure was tried on five samples having known amounts of protein, with the following data resulting:

Light absorbed	Amount of protein (mg)
0.44	2
0.82	16
1.20	30
1.61	46
1.83	55

(a) Determine the coefficient of determination.
(b) Does this appear to be a reasonable way of estimating the amount of protein in a liver sample?
(c) What is the estimate of the amount of protein when the light absorbed is 1.5?
(d) Determine a prediction interval in which we can have 90 percent confidence for the quantity in part (c).

4. Determine the coefficient of determination for the data of Prob. 1 of Sec. 12.2.
5. Determine the coefficient of determination for the data of Example 12.6.
6. A new-car dealer is interested in the relationship between the number of salespeople working on a weekend and the number of cars sold. Data were gathered for six consecutive Sundays:

Number of salespeople	Number of cars sold
5	22
7	20
4	15
2	9
4	17
8	25

(a) Determine the estimated regression line.
(b) What is the coefficient of determination?
(c) How much of the variation in the number of automobiles sold is explained by the number of salespeople?
(d) Test the null hypothesis that the mean number of sales does not depend on the number of salespeople working.

7. Find the coefficient of determination in Prob. 8 of Sec. 12.4.

12.9 SAMPLE CORRELATION COEFFICIENT

Consider a set of data pairs (x_i, Y_i), $i = 1, \ldots, n$. In Sec. 3.7 we defined the *sample correlation coefficient* of this data set by

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

It was noted that r provided a measure of the degree to which high values of x are paired with high values of Y and low values of x with low values of Y. A value of r near $+1$ indicated that large x values were strongly associated with large Y values and small x values were strongly associated with small Y values, whereas a value near -1 indicated that large x values were strongly associated with small Y values and small x values with large Y values.

In the notation of this chapter, r would be expressed as

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

Upon using the identity

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

that was presented at the end of Sec. 12.4, we can show that the absolute value of the sample correlation coefficient r can be expressed as

$$|r| = \sqrt{1 - \frac{SS_R}{S_{YY}}}$$

That is,

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

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

All this gives additional meaning to the sample correlation coefficient. For instance, if a data set has its sample correlation coefficient r equal to 0.9, then this implies that a simple linear regression model for these data explains 81 percent (since $R^2 = 0.9^2 = 0.81$) of the variation in the response values. That is, 81 percent of the variation in the response values is explained by the different input values.

PROBLEMS

1. Determine the coefficient of determination and the sample correlation coefficient for the following data sets of paired values.

(a)

x	y
2	4
3	5
5	9

(b)

x	y
4	2
5	3
9	5

What does this lead you to conclude?

*2. Show that the sample correlation coefficient of a given set of data pairs (u_i, v_i) is the same regardless of whether the u_i are considered to be the input values or the response values.

3. Find the sample correlation coefficient when the coefficient of determination and the estimated regression line are
   (a) $R^2 = 0.64, \quad y = 2x + 4$
   (b) $R^2 = 0.64, \quad y = 2x - 4$
   (c) $R^2 = 0.64, \quad y = -2x + 0.4$
   (d) $R^2 = 0.64, \quad y = -2x - 0.4$

4. If the sample correlation coefficient is 0.95, how much of the variation in the responses is explained by the different input values?

5. The following data relate the ages of wives and husbands when they were married. Before you look at the data, would you expect a positive, negative, or near-zero value of the sample correlation coefficient?

Wife's age	18	24	40	33	30	25
Husband's age	21	29	51	30	36	25

   (a) Letting the wife's age be the input, find the estimated regression line for determining the husband's age.
   (b) Letting the husband's age be the input, find the estimated regression line for determining the wife's age.
   (c) Determine the coefficient of determination and the sample correlation coefficient for the situation described in part (a).
   (d) Determine the coefficient of determination and the sample correlation coefficient for the situation described in part (b).

6. Find the sample correlation coefficient in Prob. 6 of Sec. 12.7.

12.10 ANALYSIS OF RESIDUALS: ASSESSING THE MODEL

The initial step for ascertaining whether the simple linear regression model

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

where e is a normal random variable with mean 0 and variance 1, is appropriate in a given situation is to investigate the scatter diagram. Indeed, this is often sufficient to convince one that the regression model is or is not correct. When the scatter diagram does not by itself rule out the preceding model, then the least-squares estimators A and B should be computed and the residuals $Y_i - (A + Bx_i)$, $i = 1, \ldots, n$, analyzed. The analysis begins by normalizing, or standardizing, the residuals by dividing them by $\sqrt{SS_R/(n-2)}$, the estimate of the standard deviation of the Y_i. The resulting quantities

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

are called the *standardized residuals*.

When the simple linear regression model is correct, the standardized residuals are approximately independent standard normal random variables and thus should be randomly distributed about 0 with about 95 percent of their values being between -2 and $+2$ (since $P\{-1.96 < Z < 1.96\} = 0.95$). In addition, a plot of the standardized residuals should not indicate any distinct pattern. Indeed, any indication of a distinct pattern should make one suspicious about the validity of the assumed simple linear regression model.

Figure 12.9 presents three different scatter diagrams and their associated standardized residuals. The first of these, as indicated both by its scatter diagram and the random nature of its standardized residuals, appears to fit the straight-line model quite well. The second residual plot shows a discernible pattern, in that the residuals appear to be first decreasing and then increasing as the input level increases. This often means that higher-order (than just linear) terms are needed to describe the relationship between the input and response. Indeed, this is also indicated by the scatter diagram in this case. The third standardized residual plot also shows a pattern, in that the absolute value of the residuals, and thus their squares, appear to be increasing, as the input level increases. This often indicates that the variance of the response is not constant but increases with the input level.

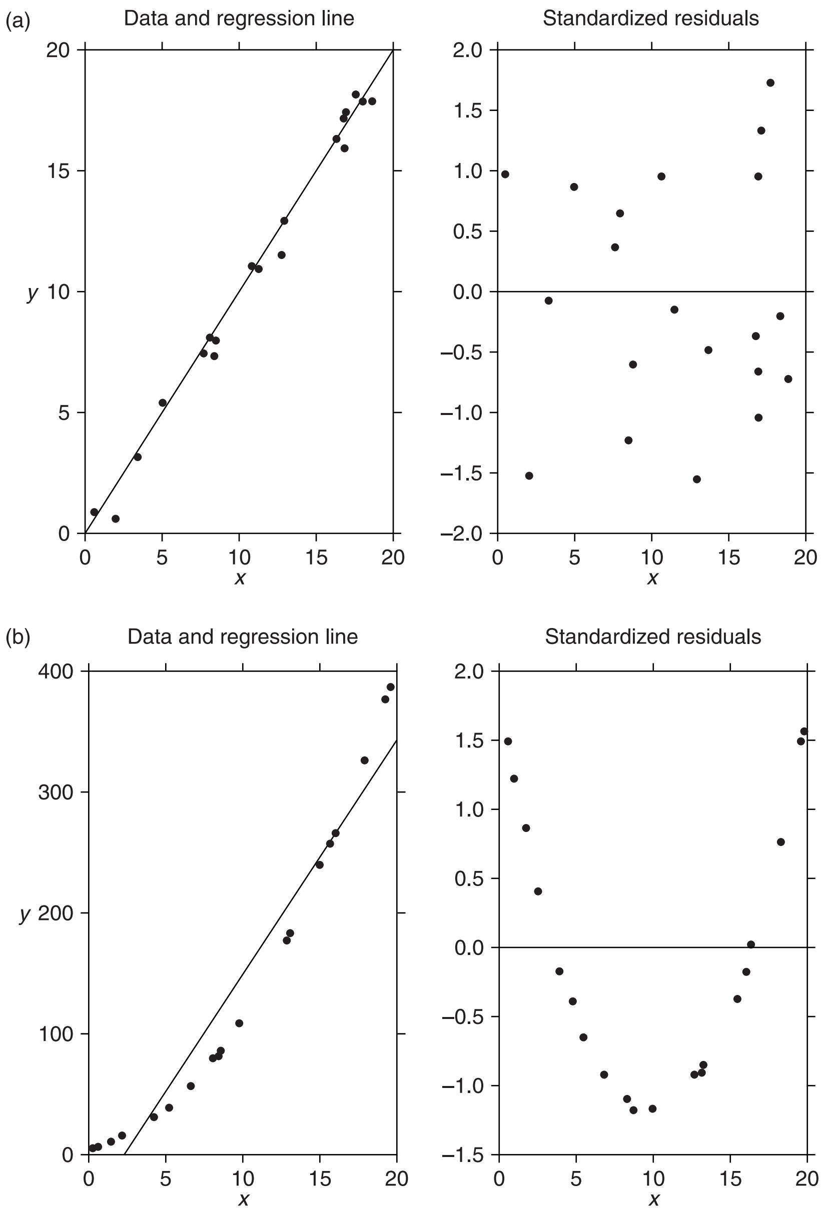

FIGURE 12.9

Three scatter diagrams and their associated standardized residuals. (*Continued*)

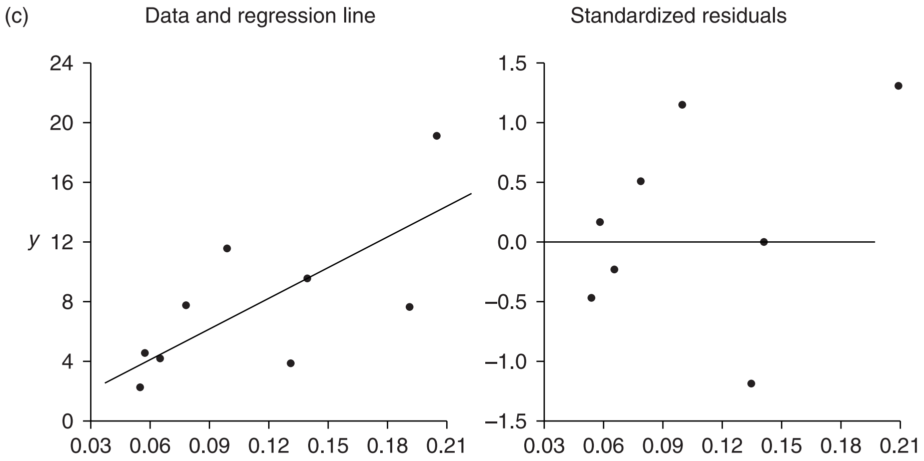

FIGURE 12.9

PROBLEMS

1. Plot the standardized residuals using the data from Example 1 of Sec. 12.3. What conclusions can be drawn about the assumption of a simple linear regression model?
2. Plot the standardized residuals using the data from Example 6 of Sec. 12.3. What conclusions can be drawn about the assumption of a simple linear regression model?

12.11 MULTIPLE LINEAR REGRESSION MODEL

Up to now we have been concerned with predicting the value of a response on the basis of the value of a single input variable. However, in many situations the response is dependent on a multitude of input variables.

■ Example 12.11

In laboratory experiments two factors that often affect the percentage yield of the experiment are the temperature and the pressure at which the experiment is conducted. The following data detail the results of four independent experiments. For each experiment, we have the temperature (in degrees Fahrenheit) at which the experiment is run, the pressure (in pounds per square inch), and the percentage yield.

Experiment	Temperature	Pressure	Percentage yield
1	140	210	68
2	150	220	82
3	160	210	74
4	130	230	80

Suppose that we are interested in predicting the response value Y on the basis of the values of the k input variables $x_1, x_2, \ldots, x_k$.

Definition *The* multiple linear regression *model supposes that the response* Y *is related to the input values* $\mathrm{x_i}$, $\mathrm{i} = 1, \ldots, \mathrm{k}$, *through the relationship*

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

In this expression, $\beta_0, \beta_1, \ldots, \beta_k$ are *regression parameters* and e is an *error* random variable that has mean 0. The regression parameters will not be initially known and must be estimated from a set of data.

Suppose that we have at our disposal a set of n responses corresponding to n different sets of the k input values. Let y_i denote the ith response, and let the k input values corresponding to this response be $x_{i1}, x_{i2}, \ldots, x_{ik}$, $i = 1, \ldots, n$. Thus, for instance, y_1 was the response when the k input values were $x_{11}, x_{12}, \ldots, x_{1k}$. The data set is presented in Fig. 12.10.

Example 12.12

In Example 12.11 there are two input variables, the temperature and the pressure, and so $k = 2$. There are four experimental results, and so $n = 4$. The value

Set	Input 1	Input 2	...	Input k	Response
1	x_{11}	x_{12}	$\ldots$	x_{1k}	y_1
2	x_{21}	x_{22}	$\ldots$	x_{2k}	y_2
3	x_{31}	x_{32}	$\ldots$	x_{3k}	y_3
$\vdots$					
n	x_{n1}	x_{n2}	$\ldots$	x_{nk}	y_n

FIGURE 12.10

Data on n experiments.

x_{i1} refers to the temperature and x_{i2} to the pressure of experiment i. The value y_i is the percentage yield (response) of experiment i. Thus, for instance,

$$x_{31} = 160 \quad x_{32} = 210 \quad y_3 = 74$$ ■

To estimate the regression parameters again, as in the case of simple linear regression, we use the method of least squares. That is, we start by noting that if B_0, $B_1, \ldots, B_k$ are estimators of the regression parameters β_0, $\beta_1, \ldots, \beta_k$, then the estimate of the response when the input values are $x_{i1}, x_{i2}, \ldots, x_{ik}$ is given by

$$\text{Estimated response} = B_0 + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik}$$

Since the actual response was y_i, we see that the difference between the actual response and what would have been predicted if we had used the estimators B_0, $B_1, \ldots, B_k$ is

$$\epsilon_i = y_i - (B_0 + B_1 x_{i1} + B_2 x_{i2} + \cdots + B_k x_{ik})$$

Thus, ϵ_i can be regarded as the *error* that would have resulted if we had used the estimators B_i, $i = 0, \ldots, k$. The estimators that make the sum of the squares of the errors as small as possible are called the *least-squares estimators*.

The least-squares estimators of the regression parameters are the choices of B_i that make

$$\sum_{i=1}^{n} \epsilon_i^2$$

as small as possible.

The actual computations needed to obtain the least-squares estimators are algebraically messy and will not be presented here. Instead we refer to Program 12-2 to do the computations for us. The outputs of this program are the estimates of the regression parameters. In addition, the program provides predicted response values for specified sets of input values. That is, if the user enters the values x_1, $x_2, \ldots, x_k$, then the computer will print out the value of $B(0) + B(1)x_1 + \cdots + B(k)x_k$, where $B(0), B(1), \ldots, B(k)$ are the least-squares estimators of the regression parameters.

■ Example 12.13

The following data relate the suicide rate y to the population size x_1 and the yearly divorce rate x_2 in eight different cities.

Location	Population (thousands)	Divorce rate per 100,000	Suicide rate per 100,000
Akron, OH	679	30.4	11.6
Anaheim, CA	1420	34.1	16.1
Buffalo, NY	1349	17.2	9.3
Austin, TX	296	26.8	9.1
Chicago, IL	3975	29.1	8.4
Columbia, SC	323	18.7	7.7
Detroit, MI	2200	32.6	11.3
Gary, IN	633	32.5	8.4

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

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

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

(b) Predict the suicide rate in a county having a population of 400,000 people and a divorce rate of 28.4 divorces yearly for every 1000 people.

Historical Perspective

Method of Least Squares

The first publication detailing the method of least squares was due to the French mathematical scientist Adrien-Marie Legendre in 1805. Legendre presented the method in the appendix to his book *Nouvelles methodes pour la determination des orbites des cometes (New Methods for Determining the Orbits of Comets)*. After explaining the method, Legendre worked out an example, using data from the 1795 survey of the French meridian arc, an example in which $k = 2$ and $n = 5$. In 1809 Karl Friedrich Gauss published a justification of the method of least squares that highlighted the normal as the distribution of the error term. In his paper Gauss started a controversy by claiming that he had been using the method since 1795. Gauss claimed that he had used the method of least squares in 1801 to locate the missing asteroid Ceres. This asteroid, the largest in the solar system and the first to be discovered, was spotted by the Italian astronomer Giuseppe Piazzi of the Palerno Observatory on January 1, 1801. Piazzi observed it for 40 consecutive days at which time the asteroid, which had a very low luminosity, disappeared from view. In the hope that other scientists would be able to determine its path, Piazzi published the data concerning his observations. Months later the news and data reached the attention of Gauss. In a short time, and without any explanation of his method, Gauss published a predicted orbit for the asteroid. Shortly afterward, Ceres was found in almost the exact position predicted by Gauss.

The ensuing priority dispute between Legendre and Gauss became rather heated. In the 1820 edition of his book, Legendre added an attack on Gauss that he attributed to the anonymous writer Monsieur***. Gauss, in turn, solicited testimony from colleagues to the effect that he had told them of his method before 1805. Present-day scholars for the most part accept Gauss' claim that he knew and used the method of least squares before Legendre. (Gauss is famous for often letting many years go by before publishing his results.) However, most scholars also feel that priority should be determined by the earliest date of publication and so the credit for the discovery of the method of least squares rightfully belongs to Legendre.

Solution

Running Program 12-2 gives the following output:

The estimates of the regression coefficients are as follows

B(0) = 3.686646

B(1) = −2.411092E−04

B(2) = .2485504

If the two input values are 400 and 28.4, the predicted response is 10.64903. That is, the estimated multiple regression equation is

$$Y = 3.6866 - 0.00024x_1 + 0.24855x_2$$

The predicted suicide rate is

$$\begin{aligned} y &= 3.6866 - 0.00024 \times 400 + 0.24855 \times 28.4 \\ &= 10.649 \end{aligned}$$

That is, we predict that in such a county the yearly suicide rate is 10.649 per 100,000 residents. Since the population size is 400,000, this means a prediction of 42.596 suicides per year. ■

12.11.1 Dummy Variables for Categorical Data

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

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

Whereas each of the first four variables will take on a value in some interval, the final variable is a categorical variable that indicates whether the person under consideration has or does not have a certain characteristic (which, in our case, is whether the person is a smoker or not). To determine which category the person belongs to, we let

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

We can now try to fit the multiple regression model

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

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

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

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

and

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

In other words, in using the model with a dummy variable we are assuming that if a smoker and nonsmoker had the same values for the four quantitative variables x_1, x_2, x_3, x_4 then the difference between their mean cholesterol levels would always be a constant, no matter what the values of x_1, x_2, x_3, x_4. Put another way, using the language of the analysis of variance the model supposes that there is no interaction in determining cholesterol level between the quantitative variables and whether or not the person is a smoker. (Thus, for instance, the dummy variable model assumes that the amount that one is overweight has the same effect on raising the cholesterol on a smoker as it does on a nonsmoker.) Because this seems like a large assumption, it is typically preferable when the data set is large enough to use two regression models rather than combining into one model by the use of a dummy variable. (Although it is true that the standard multiple regression model assumes that there is no interaction between the different variable values, this assumption seems more questionable when one of the variables is a dummy variable indicating a qualitative characteristic than when all the variables are quantitative.)

In situations, however, where there are multiple qualitative characteristics that the researcher feels are relevant it might be necessary to utilize dummy variables, for otherwise the data set may become too fragmented to yield reliable estimates of the regression parameters. So, for instance, if the cholesterol researcher felt that the sex of the person was also a relevant factor, then the researcher could utilize a multiple regression model having two dummy variables, namely x_5 and

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

PROBLEMS

1. The following data relate the selling price y to the living space x_1, the lot size x_2, and the number of bathrooms x_3 for 10 recently sold homes in a common area.

Selling price (thousands of dollars)	House size (square feet)	Lot size (acres)	Number of bathrooms
170	1300	0.25	1
177	1450	0.30	1.5
191	1600	0.30	2
194	1850	0.45	2
202	2100	0.40	2
210	2000	0.40	2.5
214	2100	0.50	2
228	2400	0.50	2.5
240	2700	0.50	2.5
252	2600	0.70	3

(a) Fit a multiple linear regression model to the data.
(b) Predict the selling price of a home of 2500 square feet whose lot size is 0.4 acres and that has two bathrooms.
(c) What if the house in part (b) had three bathrooms?

2. In Example 12.11, predict the yield of an experiment run at a temperature of 150 degrees Fahrenheit and with a pressure of 215 pounds per square inch.
3. Fit a multiple linear regression model to the following data set:

x_1	x_2	x_3	x_4	y
1	3	5	9	121
2	4	4	10	165
1.5	8	2	14	150
3	9	3	8	170
1	11	4	12	140

Predict the value of a response taken at the input values

$$x_1 = 2 \quad x_2 = 7 \quad x_3 = 3 \quad x_4 = 13$$

4. The following data set refers to Stanford heart transplants. It relates the survival time of patients after receiving a heart transplant to their age and to a mismatch score that is used as an indicator of how well the transplanted heart should match the recipient.

Survival time (days)	Mismatch Score	Age
624	1.32	51.0
1350	0.87	54.1
64	1.89	54.6
46	0.61	42.5
1024	1.13	43.4
280	1.12	49.5
10	2.76	55.3
60	0.69	64.5
836	1.58	45.0
136	1.62	52.0
730	0.96	58.4
39	1.38	42.8

(a) Fit a multiple linear regression model to these data.
(b) Estimate the survival time of a 50-year-old heart transplant patient whose mismatch score is 1.46.

5. A steel company will be producing cold-reduced sheet steel consisting of 0.15 percent copper and produced at an annealing temperature of 1150 degrees Fahrenheit. The company is interested in estimating the

mean hardness of this steel. It collected the following data on 10 different specimens of sheet steel that had been produced at different copper contents and annealing temperatures.

Hardness	Copper content	Annealing temperature
79.2	0.02	1050
64.0	0.03	1200
55.7	0.03	1250
56.3	0.04	1300
58.6	0.10	1300
49.8	0.09	1450
51.1	0.12	1400
61.0	0.09	1200
70.4	0.15	1100
84.3	0.16	1000

Estimate the mean hardness of the steel to be produced.

6. In the subsection on dummy variables it was supposed that two of the variables in a cholesterol study referred to the number of pounds the subject was overweight and the number he or she was underweight. Do you think this was a good idea as opposed to, say, just having a single variable equal to the weight of the person? What if the single variable was the number of pounds (positive or negative) that an individual was overweight?
7. Consider a multiple regression model where the researcher is planning to use both smoking and sex as categorical variables. Discuss having dummy variables for both of these quantities versus having 4 dummy variables x_5, x_6, x_7, x_8 defined as

$$x_5 = \begin{cases} 1, & \text{if person is a male smoker} \\ 0, & \text{otherwise} \end{cases}$$

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

$$x_7 = \begin{cases} 1, & \text{if person is a male nonsmoker} \\ 0, & \text{otherwise} \end{cases}$$

$$x_8 = \begin{cases} 1, & \text{if person is a female nonsmoker} \\ 0, & \text{otherwise} \end{cases}$$

KEY TERMS

Simple linear regression: A model that relates a response variable Y to an input variable x by the equation

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

The quantities α and β are parameters of the regression model, and e is an error random variable.

Dependent variable: Another term for the response variable.

Independent variable: Another term for the input variable.

Method of least squares: A method for obtaining estimators of the regression parameters α and β. It chooses as estimators those values that make the sum of the squares of the differences between the observed and the predicted responses as small as possible.

Regression to the mean: This phenomenon occurs when the regression parameter β is strictly between 0 and 1. This makes the mean response corresponding to the input level x larger than x when x is small and smaller than x when x is large. This phenomenon is common in testing–retesting situations.

Regression fallacy: The belief in testing–retesting situations that the phenomenon of regression to the mean has a significant cause when it is actually just a byproduct of random fluctuations.

Coefficient of determination: A statistic whose value indicates the proportion of the variation in the response values that is caused by the different input values.

Sample correlation coefficient: Its absolute value is the square root of the coefficient of determination. Its sign is the same as that of the estimator of the regression parameter β.

Multiple linear regression: A model that relates a response variable Y to a set of k input variables $x_1, \ldots, x_k$ by the equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

SUMMARY

The simple linear regression model relates the value of a *response* random variable Y to the value of an *input* variable x by the equation

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

The parameters α and β are *regression parameters* that have to be estimated from data. The quantity e is an *error* random variable that has expected value 0.

The *method of least squares* is used to estimate the regression parameters α and β. Suppose that experiments are run at the input levels $x_i, i = 1, \ldots, n$. Let $Y_i, i = 1, \ldots, n$, denote the corresponding outputs. The least-squares approach is to choose as estimators of α and β the values of A and B that make

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

as small as possible. The values of A and B that accomplish this—call these values $\hat{\alpha}$ and $\hat{\beta}$—are given by

$$\hat{\beta} = \frac{S_{xY}}{S_{xx}}$$

$$\hat{\alpha} = \overline{Y} - \hat{\beta}\overline{x}$$

where $\overline{x}$ and $\overline{Y}$ are the average values of the x_i's and the Y_i's, respectively, and

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

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

The straight-line relationship

$$y = \hat{\alpha} + \hat{\beta}x$$

is called the *estimated regression line*.

The error random variable e is assumed to be a normal random variable with expected value 0 and variance σ^2, where σ^2 is unknown and needs to be estimated from the data. The estimator of σ^2 is

$$\frac{SS_R}{n-2}$$

where the quantity SS_R, called the *sum of the squares of the residuals*, is defined by

$$SS_R = \sum_{i=1}^{n} (Y_i - \hat{\alpha} - \hat{\beta}x_i)^2$$

The quantities $Y_i - \hat{\alpha} - \hat{\beta}x_{i,}$ representing the difference between the actual response and its predicted value under the least-squares estimators, are called the *residuals*.

The following is a useful computational formula for finding SS_R when using a hand calculator.

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

where

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

If the regression parameter β is equal to 0, then the value of a response will not be affected by its input value x. To see if this hypothesis is plausible, we test

$$\mathrm{H}_0\colon \beta = 0 \quad \text{against} \quad \mathrm{H}_1\colon \beta \neq 0$$

The significance-level-γ test is based on the test statistic

$$\mathrm{TS} = \sqrt{\frac{(n-2)S_{xx}}{\mathrm{SS}_R}}\,\hat{\beta}$$

and is to

$$\begin{array}{ll} \text{Reject } \mathrm{H}_0 & \text{if } |\mathrm{TS}| \geq t_{n-2,\gamma/2} \\ \text{Not reject } \mathrm{H}_0 & \text{otherwise} \end{array}$$

Equivalently, if the value of TS is v, then the p value is given by

$$p \text{ value } = 2P\{T_{n-2} \geq |v|\}$$

where T_{n-2} is a t random variable with $n - 2$ degrees of freedom.

The phenomenon of *regression to the mean* is said to occur when the regression parameter β lies between 0 and 1. When this is the case, the expected response corresponding to the input value x will be greater than x when x is small and will be less than x when x is large.

The phenomenon of regression to the mean is often seen in testing–retesting situations involving a homogeneous population. This is because some of those being tested will, purely by chance, do significantly better or worse than is their norm. In the repeated test they will often obtain a more normal result. Thus, those scoring high on the first test often come down somewhat on the second while those scoring low on the first test often improve on the second. The belief that something significant has caused the regression to the mean (for instance, that the lower-scoring students studied much harder for the retest while the higher-scoring ones were complacent) when in fact it was just due to random fluctuations about the mean value is called the *regression fallacy.*

The input–response data pairs $(x_i, y_i), i = 1, \ldots, n$, can be used to provide a *prediction interval* that, with a prescribed degree of confidence, will contain a future response at the input value x_0. Specifically, we can assert, with $100(1 - \gamma)$ percent confidence, that the response at the input value x_0 will lie in the interval

$$\hat{\alpha} + \hat{\beta}x_0 \pm t_{n-2,\gamma/2}W$$

where

$$W = \sqrt{\left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right] \frac{SS_R}{(n-2)}}$$

The quantities $\hat{\alpha}$, $\hat{\beta}$, $\overline{x}$, S_{xx}, and W are all based on the data pairs (x_i, y_i), $i = 1, \ldots, n$, and can be obtained by running Program 12-1.

The quantity R^2 defined by

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

is called the *coefficient of determination.* Its value, which will always lie between 0 and 1, can be interpreted as the proportion of the variation in the response values that is explained by the different input values.

The quantity r, defined by

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

is called the *sample correlation coefficient*. Aside from its sign (either positive or negative) it is equal to the square root of the coefficient of determination. That is,

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

The quantities

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

are called the *standardized residuals.* A plot of these residuals can be used to assess the accuracy of the linear regression model.

The *multiple linear regression* model relates a response random variable Y to a set of input variables $x_1, \ldots, x_k$ according to the equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + e$$

In this equation, $\beta_0, \beta_1, \ldots, \beta_k$ are regression parameters and e is an error random variable having mean 0.

The regression parameters are estimated from data by using the method of least squares. That is, the estimators are chosen to minimize the sum of the squares of the differences between the actual observed response values and their predicted values. Program 12-2 can be used to obtain these estimates. This program will also return predicted values of responses corresponding to arbitrarily entered input values.

REVIEW PROBLEMS

1. The following relates the breaking strength of eight pieces of rope and the percentage of that rope that is nylon (rather than cotton).

Percentage nylon	Breaking strength (pounds)
0	160
10	240
20	325
20	340
30	395
40	450
50	510
50	520

(a) Plot the data in a scatter diagram.
(b) Give the estimated regression line.
(c) Predict the breaking strength of a new piece of rope that is 50 percent nylon.
(d) Give an interval that, with 95 percent confidence, will contain the breaking strength of a piece of rope that is 50 percent nylon.

2. It is generally accepted that by increasing the number of units it produces, a manufacturer can often decrease its cost per unit. The following relates the manufacturing cost per unit to the number of units produced.

Number of units	10	20	50	100	150	200
Cost per unit	9.4	9.2	9.0	8.5	8.1	7.4

(a) Predict the cost per unit when a production run of 125 units is called for.
(b) Estimate the variance of the cost in part (a).
(c) Give an interval that, with 99 percent confidence, will contain the cost per unit when a production run of 110 units is used.

3. Use the data relating to the first 20 women on the data set given in App. A. Let the input variable be the weight and the response variable be the systolic blood pressure.
   (a) Estimate the regression parameters.
   (b) Give a 95 percent prediction interval for the systolic blood pressure of a female student who weighs 120 pounds.
   (c) Find all female students in App. A that weigh between 119 and 121 pounds. What percentage of them have systolic blood pressures that fall within the interval given in (c)?
4. A set of 10 married couples are randomly chosen from a given community, and the 20 individuals are given an IQ test. Number the couples, and let x_i and y_i denote the score of the wife and of the husband of couple i. Do you think that a plot of the resulting scatter diagram will indicate a regression to the mean? Explain.
5. Experienced flight instructors have claimed that praise for an exceptionally fine landing is typically followed by a poorer landing on the next attempt, whereas criticism of a faulty landing is typically followed by an improved landing. Should we thus conclude that verbal praise tends to lower performance levels whereas verbal criticism tends to raise them? Or is some other explanation possible?
6. The following data relate the average number of cigarettes smoked daily to the number of free radicals found in the lungs of eight individuals.

Number of cigarettes	Free radicals
0	94
10	144
14	182
5	120
18	240
20	234
30	321
40	400

   (a) Represent this data set in a scatter diagram.
   (b) Fit a straight line to the data "by hand."
   (c) Determine the estimated regression line, and compare it to the one drawn in part (b).
   (d) Predict the number of free radicals in someone who smokes an average of 26 cigarettes daily.
   (e) Determine a prediction interval which, with 95 percent confidence, will contain the amount of free radicals in an individual who smokes an average of 26 cigarettes daily.

7. The following data give the gasoline retail prices per gallon in the United States for each of the years from 1990 to 2002.

Year	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002
Price	1.16	1.14	1.13	1.11	1.11	1.15	1.23	1.23	1.06	1.17	1.51	1.46	1.36

(a) Find the estimated regression line.
(b) Test the hypothesis, at the 5 percent level of significance, that $\beta = 0$.

8. The following are the average scores of college-bound high school students on the science and reasoning section of the American College Testing (ACT) examination in certain years through 2003, excluding 1998.

Year	1996	1997	1999	2000	2001	2002	2003
Score	20.9	21.0	21.0	21.0	21.0	20.8	20.8

Source: *High Schools Profile Report,* ACT Program, Iowa City, IA.

(a) Predict the 1998 average score.
(b) Find a 95 percent prediction interval for that score.

9. The following table gives the percentage of workers in manufacturing who were union members in both 1984 and 1989 for a random sample of nine states.

State	**1984**	**1989**
Alabama	27.3	23.8
Colorado	10.9	9.5
Illinois	40.9	29.8
Kentucky	27.0	21.5
Minnesota	25.7	16.4
New Jersey	25.4	24.4
Texas	15.9	13.8
Wisconsin	31.1	23.0
New York	50.4	47.2

Source: *Manufacturing Climates Study,* Grant/Thornton, Chicago, annual.

(a) The percentage of Ohio's manufacturing workers who were union members was 41.6 in 1984. Predict the 1989 percentage.
(b) Oklahoma's union membership percentage was 17.5 in 1984. Construct an interval that, with 95 percent confidence, contains Oklahoma's membership percentage in 1989.

10. The following give the body mass index (BMI) and systolic blood pressure of 8 randomly chosen men who do not take any blood pressure medication.

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

Give an interval that, with 95 percent confidence, will include the systolic blood pressure of a man who does not take blood pressure medication and whose BMI is 26.0.

11. The tensile strength of a certain synthetic fiber is thought to be related to the percentage of cotton in the fiber and to the drying time of the fiber. A study of eight pieces of fiber yielded the following results.

Percentage of cotton	Drying time	Tensile strength
13	2.1	212
15	2.2	221
18	2.5	230
20	2.4	219
18	3.2	245
20	3.3	238
17	4.1	243
18	4.3	242

(a) Fit a multiple regression equation, with tensile strength being the response and the percentage of cotton and the drying time being the input variables.

(b) Predict the tensile strength of a synthetic fiber having 22 percent cotton whose drying time is 3.5.

12. The following data refer to the seasonal wheat yield per acre at eight different locations, all having roughly the same quality soil. The data relate the wheat yield at each location to the seasonal amount of rainfall and the amount of fertilizer used per acre.

Rainfall (inches)	Fertilizer (pounds per acre)	Wheat yield
15.4	100	46.6
18.2	85	45.7
17.6	95	50.4
18.4	140	66.5
24.0	150	82.1
25.2	100	63.7
30.3	120	75.8
31.0	80	58.9

(a) Estimate the regression parameters.

(b) Estimate the additional yield in wheat for each additional inch of rain.

(c) Estimate the additional yield in wheat for each additional pound of fertilizer.

(d) Predict the wheat yield in a year having 26 inches of rain if the amount of fertilizer used that year was 130 pounds per acre.

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

Yearly income (thousands of dollars)	Years on the job	Job satisfaction
47	8	5.6
42	4	6.3
54	12	6.8
48	9	6.7
56	16	7.0
59	14	7.7
53	10	7.0
62	15	8.0
66	22	7.8

(a) Estimate the regression parameters.

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

(c) Predict the job satisfaction of an employee who has spent 5 years on the job and earns a yearly salary of $51,000.

14. Suppose in Prob. 13 that job satisfaction was related to years on the job, and so the following data would have resulted:

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

(a) Estimate the regression parameters α and β.

(b) What is the qualitative relationship between years of service and job satisfaction? That is, based on the given data, what appears to happen to job satisfaction as service increases?

(c) Compare your answer to part (b) to the answer you obtained in part (b) of Prob. 13.

(d) What conclusion, if any, can you draw from your answer in part (c)?

15. The correct answer to Prob. 5 of Sec. 12.5 is to reject the hypothesis that cigarette consumption and bladder cancer rates are unrelated. Does this imply that cigarette smoking directly leads to an increased risk of contracting bladder cancer, or can you think of another explanation? (*Hint*: Is there another variable you can think of that is statistically associated with both smoking and bladder cancer? What type of data collection and statistical procedure would you recommend to increase our knowledge about the factors affecting bladder cancer rates?)