Regression and Correlation

1 Introduction

A major objective of many statistical investigations is to establish relationships that make it possible to predict one or more variables in terms of others. Thus, studies are made to predict the potential sales of a new product in terms of its price, a patient's weight in terms of the number of weeks he or she has been on a diet, family expenditures on entertainment in terms of family income, the per capita consumption of certain foods in terms of their nutritional values and the amount of money spent advertising them on television, and so forth.

Although it is, of course, desirable to be able to predict one quantity exactly in terms of others, this is seldom possible, and in most instances we have to be satisfied with predicting averages or expected values. Thus, we may not be able to predict exactly how much money Mr. Brown will make 10 years after graduating from college, but, given suitable data, we can predict the average income of a college graduate in terms of the number of years he has been out of college. Similarly, we can at best predict the average yield of a given variety of wheat in terms of data on the rainfall in July, and we can at best predict the average performance of students starting college in terms of their I.Q.'s.

Formally, if we are given the joint distribution of two random variables X and Y, and X is known to take on the value x, the basic problem of **bivariate regression** is that of determining the conditional mean $\mu_{Y|x}$, that is, the "average" value of Y for the given value of X. The term "regression," as it is used here, dates back to Francis Galton, who employed it to indicate certain relationships in the theory of heredity. In problems involving more than two random variables, that is, in **multiple regression**, we are concerned with quantities such as $\mu_{Z|x,y}$, the mean of Z for given values of X and Y, $\mu_{X_4|x_1,\,x_2,\,x_3}$, the mean of X_4 for given values of X_1, X_2, and X_3, and so on.

Definition 1. Bivariate regression; regression equation. *If* $\mathrm{f(x, y)}$ *is the value of the joint density of two random variables* X *and* Y, ***bivariate regression*** *consists of determining the conditional density of* Y, *given* $\mathrm{X = x}$ *and then evaluating the integral*

$$\mu_{Y|x} = E(Y|x) = \int_{-\infty}^{\infty} y \cdot w(y|x)\,dy$$

The resulting equation is called the ***regression equation of*** Y ***on*** X. *Alternately, the* ***regression equation of*** X ***on*** Y *is given by*

$$\mu_{X|y} = E(X|y) = \int_{-\infty}^{\infty} x \cdot f(x|y)dy$$

In the discrete case, when we are dealing with probability distributions instead of probability densities, the integrals in the two regression equations given in Definition 1 are simply replaced by sums. When we do not know the joint probability density or distribution of the two random variables, or at least not all its parameters, the determination of $\mu_{Y|x}$ or $\mu_{X|y}$ becomes a problem of estimation based on sample data; this is an entirely different problem, which we shall discuss in Sections 3 and 4.

EXAMPLE 1

Given the two random variables X and Y that have the joint density

$$f(x,y) = \begin{cases} x \cdot e^{-x(1+y)} & \text{for } x > 0 \text{ and } y > 0 \\ 0 & \text{elsewhere} \end{cases}$$

find the regression equation of Y on X and sketch the regression curve.

Solution

Integrating out y, we find that the marginal density of X is given by

$$g(x) = \begin{cases} e^{-x} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

and hence the conditional density of Y given $X = x$ is given by

$$w(y|x) = \frac{f(x,y)}{g(x)} = \frac{x \cdot e^{-x(1+y)}}{e^{-x}} = x \cdot e^{-xy}$$

for $y > 0$ and $w(y|x) = 0$ elsewhere, which we recognize as an exponential density with $\theta = \dfrac{1}{x}$. Hence, by evaluating

$$\mu_{Y|x} = \int_0^{\infty} y \cdot x \cdot e^{-xy}\, dy$$

or by referring to the corollary of a theorem given here "The mean and the variance of the exponential distribution are given by $\mu = \theta$ and $\sigma^2 = \theta^2$," we find that the regression equation of Y on X is given by

$$\mu_{Y|x} = \frac{1}{x}$$

The corresponding regression curve is shown in Figure 1.

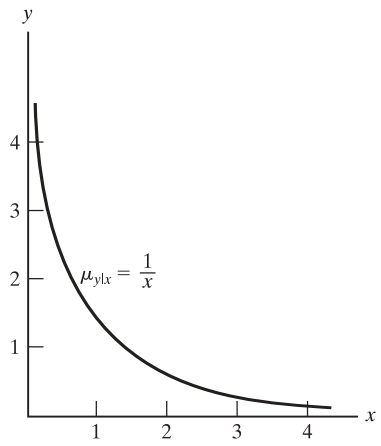

Figure 1. Regression curve of Example 1.

EXAMPLE 2

If X and Y have the multinomial distribution

$$f(x,y) = \binom{n}{x,y,n-x-y} \cdot \theta_1^x \theta_2^y (1-\theta_1-\theta_2)^{n-x-y}$$

for $x = 0,1,2,\ldots,n$, and $y = 0,1,2,\ldots,n$, with $x+y \leqq n$, find the regression equation of Y on X.

Solution

The marginal distribution of X is given by

$$g(x) = \sum_{y=0}^{n-x} \binom{n}{x,y,n-x-y} \cdot \theta_1^x \theta_2^y (1-\theta_1-\theta_2)^{n-x-y}$$

$$= \binom{n}{x} \theta_1^x (1-\theta_1)^{n-x}$$

for $x = 0,1,2,\ldots,n$, which we recognize as a binomial distribution with the parameters n and θ_1. Hence,

$$w(y|x) = \frac{f(x,y)}{g(x)} = \frac{\binom{n-x}{y} \theta_2^y (1-\theta_1-\theta_2)^{n-x-y}}{(1-\theta_1)^{n-x}}$$

for $y = 0,1,2,\ldots,n-x$, and, rewriting this formula as

$$w(y|x) = \binom{n-x}{y} \left(\frac{\theta_2}{1-\theta_1}\right)^y \left(\frac{1-\theta_1-\theta_2}{1-\theta_1}\right)^{n-x-y}$$

we find by inspection that the conditional distribution of Y given $X = x$ is a binomial distribution with the parameters $n - x$ and $\frac{\theta_2}{1-\theta_1}$, so that the regression equation of Y on X is

$$\mu_{Y|x} = \frac{(n-x)\theta_2}{1-\theta_1}$$

With reference to the preceding example, if we let X be the number of times that an even number comes up in 30 rolls of a balanced die and Y be the number of times that the result is a 5, then the regression equation becomes

$$\mu_{Y|x} = \frac{(30-x)\frac{1}{6}}{1-\frac{1}{2}} = \frac{1}{3}(30-x)$$

This stands to reason, because there are three equally likely possibilities, 1, 3, or 5, for each of the $30 - x$ outcomes that are not even.

EXAMPLE 3

If the joint density of X_1, X_2, and X_3 is given by

$$f(x_1, x_2, x_3) = \begin{cases} (x_1 + x_2)e^{-x_3} & \text{for } 0 < x_1 < 1, 0 < x_2 < 1, x_3 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

find the regression equation of X_2 on X_1 and X_3.

Solution

The joint marginal density of X_1 and X_3 is given by

$$m(x_1, x_3) = \begin{cases} \left(x_1 + \frac{1}{2}\right)e^{-x_3} & \text{for } 0 < x_1 < 1, x_3 > 0 \\ 0 & \text{elsewhere} \end{cases}$$

Therefore,

$$\mu_{X_2|x_1,x_3} = \int_{-\infty}^{\infty} x_2 \cdot \frac{f(x_1, x_2, x_3)}{m(x_1, x_3)}\, dx_2 = \int_0^1 \frac{x_2(x_1 + x_2)}{\left(x_1 + \frac{1}{2}\right)}\, dx_2$$

$$= \frac{x_1 + \frac{2}{3}}{2x_1 + 1}$$

Note that the conditional expectation obtained in the preceding example depends on x_1 but not on x_3. This could have been expected, since there is a pairwise independence between X_2 and X_3.

2 Linear Regression

An important feature of Example 2 is that the regression equation is linear; that is, it is of the form

$$\mu_{Y|x} = \alpha + \beta x$$

where α and β are constants, called the **regression coefficients**. There are several reasons why linear regression equations are of special interest: First, they lend themselves readily to further mathematical treatment; then, they often provide good approximations to otherwise complicated regression equations; and, finally, in the case of the bivariate normal distribution, the regression equations are, in fact, linear.

To simplify the study of linear regression equations, let us express the regression coefficients α and β in terms of some of the lower moments of the joint distribution of X and Y, that is, in terms of $E(X) = \mu_1$, $E(Y) = \mu_2$, $\text{var}(X) = \sigma_1^2$, $\text{var}(Y) = \sigma_2^2$, and $\text{cov}(X, Y) = \sigma_{12}$. Then, also using the correlation coefficient

$$\rho = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

we can prove the following results.

THEOREM 1. If the regression of Y on X is linear, then

$$\mu_{Y|x} = \mu_2 + \rho \frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

and if the regression of X on Y is linear, then

$$\mu_{X|y} = \mu_1 + \rho \frac{\sigma_1}{\sigma_2}(y - \mu_2)$$

Proof Since $\mu_{Y|x} = \alpha + \beta x$, it follows that

$$\int y \cdot w(y|x)\, dy = \alpha + \beta x$$

and if we multiply the expression on both sides of this equation by $g(x)$, the corresponding value of the marginal density of X, and integrate on x, we obtain

$$\iint y \cdot w(y|x) g(x)\, dy\, dx = \alpha \int g(x)\, dx + \beta \int x \cdot g(x)\, dx$$

or

$$\mu_2 = \alpha + \beta \mu_1$$

since $w(y|x)g(x) = f(x, y)$. If we had multiplied the equation for $\mu_{Y|x}$ on both sides by $x \cdot g(x)$ before integrating on x, we would have obtained

$$\iint xy \cdot f(x, y)\, dy\, dx = \alpha \int x \cdot g(x)\, dx + \beta \int x^2 \cdot g(x)\, dx$$

or

$$E(XY) = \alpha\mu_1 + \beta E(X^2)$$

Solving $\mu_2 = \alpha + \beta\mu_1$ and $E(XY) = \alpha\mu_1 + \beta E(X^2)$ for α and β and making use of the fact that $E(XY) = \sigma_{12} + \mu_1\mu_2$ and $E(X^2) = \sigma_1^2 + \mu_1^2$, we find that

$$\alpha = \mu_2 - \frac{\sigma_{12}}{\sigma_1^2} \cdot \mu_1 = \mu_2 - \rho\frac{\sigma_2}{\sigma_1} \cdot \mu_1$$

and

$$\beta = \frac{\sigma_{12}}{\sigma_1^2} = \rho\frac{\sigma_2}{\sigma_1}$$

This enables us to write the linear regression equation of Y on X as

$$\mu_{Y|x} = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$$

When the regression of X on Y is linear, similar steps lead to the equation

$$\mu_{X|y} = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2)$$

It follows from Theorem 1 that if the regression equation is linear and $\rho = 0$, then $\mu_{Y|x}$ does not depend on x (or $\mu_{X|y}$ does not depend on y). When $\rho = 0$ and hence $\sigma_{12} = 0$, the two random variables X and Y are **uncorrelated**, and we can say that if two random variables are independent, they are also uncorrelated, but if two random variables are uncorrelated, they are not necessarily independent; the latter is again illustrated in Exercise 9.

The correlation coefficient and its estimates are of importance in many statistical investigations, and they will be discussed in some detail in Section 5. At this time, let us again point out that $-1 \leqq \rho \leqq +1$, as the reader will be asked to prove in Exercise 11, and the sign of ρ tells us directly whether the slope of a regression line is upward or downward.

3 The Method of Least Squares

In the preceding sections we have discussed the problem of regression only in connection with random variables having known joint distributions. In actual practice, there are many problems where a set of **paired data** gives the indication that the regression is linear, where we do not know the joint distribution of the random variables under consideration but, nevertheless, want to estimate the regression

coefficients α and β. Problems of this kind are usually handled by the **method of least squares**, a method of curve fitting suggested early in the nineteenth century by the French mathematician Adrien Legendre.

To illustrate this technique, let us consider the following data on the number of hours that 10 persons studied for a French test and their scores on the test:

Hours studied x	*Test score* y
4	31
9	58
10	65
14	73
4	37
7	44
12	60
22	91
1	21
17	84

Plotting these data as in Figure 2, we get the impression that a straight line provides a reasonably good fit. Although the points do not all fall exactly on a straight line, the overall pattern suggests that the average test score for a given number of hours studied may well be related to the number of hours studied by means of an equation of the form $\mu_{Y|x} = \alpha + \beta x$.

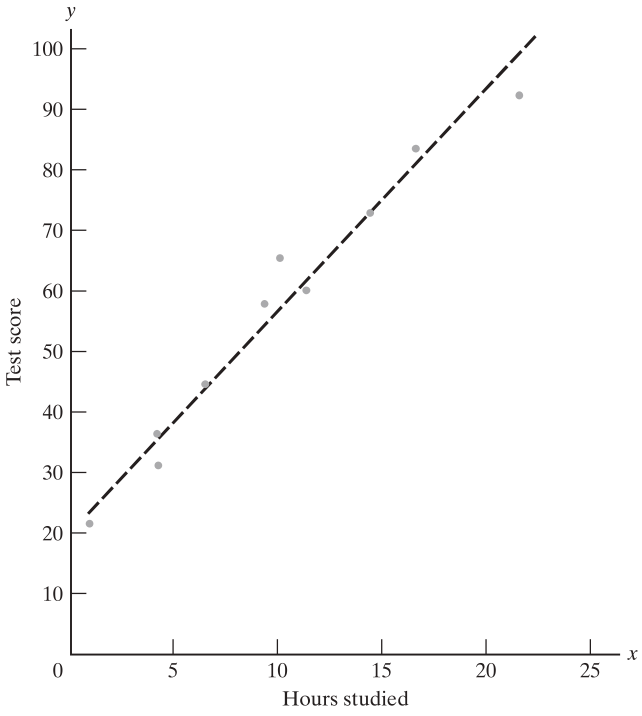

Figure 2. Data on hours studied and test scores.

Once we have decided in a given problem that the regression is approximately linear and the joint density of x and y is unknown, we face the problem of estimating the coefficients α and β from the sample data. In other words, we face the problem of obtaining estimates $\hat{\alpha}$ and $\hat{\beta}$ such that the estimated regression line $\hat{y} = \hat{\alpha} + \hat{\beta}x$ in some sense provides the best possible fit to the given data.

Denoting the vertical deviation from a point to the estimated regression line by e_i, as indicated in Figure 3, the least squares criterion on which we shall base this "goodness of fit" is defined as follows:

DEFINITION 2. LEAST SQUARES ESTIMATE. *If we are given a set of paired data*

$$\{(x_i, y_i); i = 1, 2, \ldots, n\}$$

The ***least squares estimates*** *of the regression coefficients in bivariate linear regression are those that make the quantity*

$$q = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2$$

a minimum with respect to $\hat{\alpha}$ *and* $\hat{\beta}$.

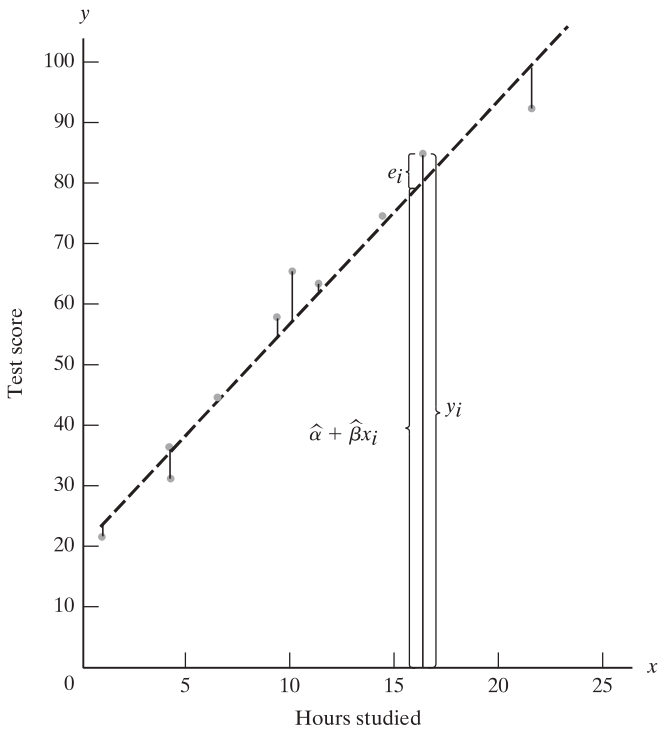

Figure 3. Least squares criterion.

Finding the minimum by differentiating partially with respect to $\hat{\alpha}$ and $\hat{\beta}$ and equating these partial derivatives to zero, we obtain

$$\frac{\partial q}{\partial \hat{\alpha}} = \sum_{i=1}^{n}(-2)[y_i - (\hat{\alpha} + \hat{\beta}x_i)] = 0$$

and

$$\frac{\partial q}{\partial \hat{\beta}} = \sum_{i=1}^{n}(-2)x_i[y_i - (\hat{\alpha} + \hat{\beta}x_i)] = 0$$

which yield the system of **normal equations**

$$\sum_{i=1}^{n} y_i = \hat{\alpha}n + \hat{\beta} \cdot \sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i y_i = \hat{\alpha} \cdot \sum_{i=1}^{n} x_i + \hat{\beta} \cdot \sum_{i=1}^{n} x_i^2$$

Solving this system of equations by using determinants or the method of elimination, we find that the least squares estimate of β is

$$\hat{\beta} = \frac{n\left(\sum_{i=1}^{n} x_i y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Then we can write the least squares estimate of α as

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i - \hat{\beta} \cdot \sum_{i=1}^{n} x_i}{n}$$

by solving the first of the two normal equations for $\hat{\alpha}$. This formula for $\hat{\alpha}$ can be simplified as

$$\hat{\alpha} = \bar{y} - \hat{\beta} \cdot \bar{x}$$

To simplify the formula for $\hat{\beta}$ as well as some of the formulas we shall meet in Sections 4 and 5, let us introduce the following notation:

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

$$S_{yy} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

and

$$S_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)$$

We can thus write the following theorem.

THEOREM 2. Given the sample data $\{(x_i, y_i); i = 1, 2, \ldots, n\}$, the coefficients of the least squares line $\hat{y} = \hat{\alpha} + \hat{\beta}x$ are

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

and

$$\hat{\alpha} = \overline{y} - \hat{\beta} \cdot \overline{x}$$

EXAMPLE 4

With reference to the data in the table in Section 3,

(a) find the equation of the least squares line that approximates the regression of the test scores on the number of hours studied;

(b) predict the average test score of a person who studied 14 hours for the test.

Solution

(a) Omitting the limits of summation for simplicity, we get $n = 10$, $\Sigma x = 100$, $\Sigma x^2 = 1{,}376$, $\Sigma y = 564$, and $\Sigma xy = 6{,}945$ from the data. Thus

$$S_{xx} = 1{,}376 - \frac{1}{10}(100)^2 = 376$$

and

$$S_{xy} = 6{,}945 - \frac{1}{10}(100)(564) = 1{,}305$$

Thus, $\hat{\beta} = \dfrac{1{,}305}{376} = 3.471$ and $\hat{\alpha} = \dfrac{564}{10} - 3.471 \cdot \dfrac{100}{10} = 21.69$, and the equation of the least squares line is

$$\hat{y} = 21.69 + 3.471x$$

(b) Substituting $x = 14$ into the equation obtained in part (a), we get

$$\hat{y} = 21.69 + 3.471(14) = 70.284$$

or $\hat{y} = 70$, rounded to the nearest unit.

Since we did not make any assumptions about the joint distribution of the random variables with which we were concerned in the preceding example, we cannot judge the "goodness" of the prediction obtained in part (b); also, we cannot judge the "goodness" of the estimates $\hat{\alpha} = 21.69$ and $\hat{\beta} = 3.471$ obtained in part (a). Problems like this will be discussed in Section 4.

The least squares criterion, or, in other words, the method of least squares, is used in many problems of curve fitting that are more general than the one treated in this section. Above all, it will be used in Sections 6 and 7 to estimate the coefficients of **multiple regression equations** of the form

$$\mu_{Y|x_1,\ldots,x_k} = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

Exercises

1. With reference to Example 1, show that the regression equation of X on Y is

$$\mu_{X|y} = \frac{2}{1+y}$$

Also sketch the regression curve.

2. Given the joint density

$$f(x,y) = \begin{cases} \frac{2}{5}(2x+3y) & \text{for } 0<x<1 \text{ and } 0<y<1 \\ 0 & \text{elsewhere} \end{cases}$$

find $\mu_{Y|x}$ and $\mu_{X|y}$.

3. Given the joint density

$$f(x,y) = \begin{cases} 6x & \text{for } 0<x<y<1 \\ 0 & \text{elsewhere} \end{cases}$$

find $\mu_{Y|x}$ and $\mu_{X|y}$.

4. Given the joint density

$$f(x,y) = \begin{cases} \dfrac{2x}{(1+x+xy)^3} & \text{for } x>0 \text{ and } y>0 \\ 0 & \text{elsewhere} \end{cases}$$

show that $\mu_{Y|x} = 1 + \dfrac{1}{x}$ and that $\text{var}(Y|x)$ does not exist.

5. This question has been intentionally omitted for this edition.

6. This question has been intentionally omitted for this edition.

7. Given the joint density

$$f(x,y) = \begin{cases} 2 & \text{for } 0<y<x<1 \\ 0 & \text{elsewhere} \end{cases}$$

show that

(a) $\mu_{Y|x} = \dfrac{x}{2}$ and $\mu_{X|y} = \dfrac{1+y}{2}$;

(b) $E(X^m Y^n) = \dfrac{2}{(n+1)(m+n+2)}$.

Also,

(c) verify the results of part (a) by substituting the values of μ_1, μ_2, σ_1, σ_2, and ρ, obtained with the formula of part (b), into the formulas of Theorem 1.

8. Given the joint density

$$f(x,y) = \begin{cases} 24xy & \text{for } x>0, y>0, \text{ and } x+y<1 \\ 0 & \text{elsewhere} \end{cases}$$

show that $\mu_{Y|x} = \frac{2}{3}(1-x)$ and verify this result by determining the values of μ_1, μ_2, σ_1, σ_2, and ρ and by substituting them into the first formula of Theorem 1.

9. Given the joint density

$$f(x,y) = \begin{cases} 1 & \text{for } -y<x<y \text{ and } 0<y<1 \\ 0 & \text{elsewhere} \end{cases}$$

show that the random variables X and Y are uncorrelated but not independent.

10. Show that if $\mu_{Y|x}$ is linear in x and $\text{var}(Y|x)$ is constant, then $\text{var}(Y|x) = \sigma_2^2(1-\rho^2)$.

11. This question has been intentionally omitted for this edition.

12. Given the random variables X_1, X_2, and X_3 having the joint density $f(x_1,x_2,x_3)$, show that if the regression of X_3 on X_1 and X_2 is linear and written as

$$\mu_{X_3|x_1,x_2} = \alpha + \beta_1(x_1-\mu_1) + \beta_2(x_2-\mu_2)$$

then

$$\alpha = \mu_3$$

$$\beta_1 = \frac{\sigma_{13}\sigma_2^2 - \sigma_{12}\sigma_{23}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$$

$$\beta_2 = \frac{\sigma_{23}\sigma_1^2 - \sigma_{12}\sigma_{13}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$$

where $\mu_i = E(X_i)$, $\sigma_i^2 = \text{var}(X_i)$, and $\sigma_{ij} = \text{cov}(X_i, X_j)$. [*Hint*: Proceed as in Section 2, multiplying by $(x_1-\mu_1)$ and $(x_2-\mu_2)$, respectively, to obtain the second and third equations.]

13. Find the least squares estimate of the parameter β in the regression equation $\mu_{Y|x} = \beta x$.

14. Solve the normal equations in Section 3 simultaneously to show that

$$\hat{\alpha} = \frac{\left(\sum_{i=1}^{n} x_i^2\right)\left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} x_i y_i\right)}{n\left(\sum_{i=1}^{n} x_i^2\right) - \left(\sum_{i=1}^{n} x_i\right)^2}$$

15. When the x's are equally spaced, the calculation of $\hat{\alpha}$ and $\hat{\beta}$ can be simplified by coding the x's by assigning them the values $\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots$ when n is odd, or the values $\ldots, -5, -3, -1, 1, 3, 5, \ldots$ when n is even. Show that with this coding the formulas for $\hat{\alpha}$ and $\hat{\beta}$ become

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} y_i}{n} \quad \text{and} \quad \hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$$

16. The method of least squares can be used to fit curves to data. Using the method of least squares, find the normal equations that provide least squares estimates of α, β, and γ when fitting a quadratic curve of the form $y = \mathrm{a} + \mathrm{b}x + \gamma x^2$ to paired data.

4 Normal Regression Analysis

When we analyze a set of paired data $\{(x_i, y_i); 1, 2, \ldots, n\}$ by **regression analysis**, we look upon the x_i as constants and the y_i as values of corresponding independent random variables Y_i. This clearly differs from **correlation analysis**, which we shall take up in Section 5, where we look upon the x_i and the y_i as values of corresponding random variables X_i and Y_i. For example, if we want to analyze data on the ages and prices of used cars, treating the ages as known constants and the prices as values of random variables, this is a problem of regression analysis. On the other hand, if we want to analyze data on the height and weight of certain animals, and height and weight are both looked upon as random variables, this is a problem of correlation analysis.

This section will be devoted to some of the basic problems of **normal regression analysis**, where it is assumed that for each fixed x_i the conditional density of the corresponding random variable Y_i is the normal density

$$w(y_i|x_i) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left[\frac{y_i - (\alpha + \beta x_i)}{\sigma}\right]^2} \qquad -\infty < y_i < \infty$$

where α, β, and σ are the same for each i. Given a random sample of such paired data, normal regression analysis concerns itself mainly with the estimation of σ and the regression coefficients α and β, with tests of hypotheses concerning these three parameters, and with predictions based on the estimated regression equation $\hat{y} = \hat{\alpha} + \hat{\beta}x$, where $\hat{\alpha}$ and $\hat{\beta}$ are estimates of α and β.

To obtain maximum likelihood estimates of the parameters α, β, and σ, we partially differentiate the likelihood function (or its logarithm, which is easier) with respect to α, β, and σ, equate the expressions to zero, and then solve the resulting system of equations. Thus, differentiating

$$\ln L = -n \cdot \ln \sigma - \frac{n}{2} \cdot \ln 2\pi - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2$$

partially with respect to α, β, and σ and equating the expressions that we obtain to zero, we get

$$\frac{\partial \ln L}{\partial \alpha} = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)] = 0$$

$$\frac{\partial \ln L}{\partial \beta} = \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} x_i [y_i - (\alpha + \beta x_i)] = 0$$

$$\frac{\partial \ln L}{\partial \sigma} = -\frac{n}{\sigma} + \frac{1}{\sigma^3} \cdot \sum_{i=1}^{n} [y_i - (\alpha + \beta x_i)]^2 = 0$$

Since the first two equations are equivalent to the two normal equations in an earlier page, the maximum likelihood estimates of α and β are identical with the least squares estimate of Theorem 2. Also, if we substitute these estimates of α and β into the equation obtained by equating $\frac{\partial \ln L}{\partial \sigma}$ to zero, it follows immediately that the maximum likelihood estimate of σ is given by

$$\hat{\sigma} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} [y_i - (\hat{\alpha} + \hat{\beta} x_i)]^2}$$

This can also be written as

$$\hat{\sigma} = \sqrt{\frac{1}{n}(S_{yy} - \hat{\beta} \cdot S_{xy})}$$

as the reader will be asked to verify in Exercise 17.

Having obtained maximum likelihood estimators of the regression coefficients, let us now investigate their use in testing hypotheses concerning α and β and in constructing confidence intervals for these two parameters. Since problems concerning β are usually of more immediate interest than problems concerning α (β is the slope of the regression line, whereas α is merely the y-intercept; also, the null hypothesis $\beta = 0$ is equivalent to the null hypothesis $\rho = 0$), we shall discuss here some of the sampling theory relating to $\hat{\mathrm{B}}$, where B is the capital Greek letter *beta*. Corresponding theory relating to $\hat{\mathrm{A}}$, where A is the capital Greek letter *alpha*, will be treated in Exercises 20 and 22.

To study the sampling distribution of $\hat{\mathrm{B}}$, let us write

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

$$= \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_{xx}} \right) Y_i$$

which is seen to be a linear combination of the n independent normal random variables Y_i. $\hat{\mathrm{B}}$ itself has a normal distribution with the mean

$$E(\hat{\mathrm{B}}) = \sum_{i=1}^{n} \left[\frac{x_i - \overline{x}}{S_{xx}} \right] \cdot E(Y_i | x_i)$$

$$= \sum_{i=1}^{n} \left[\frac{x_i - \overline{x}}{S_{xx}} \right] (\alpha + \beta x_i) = \beta$$

and the variance

$$\text{var}(\hat{B}) = \sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{S_{xx}}\right]^2 \cdot \text{var}(Y_i|x_i)$$

$$= \sum_{i=1}^{n}\left[\frac{x_i - \bar{x}}{S_{xx}}\right]^2 \cdot \sigma^2 = \frac{\sigma^2}{S_{xx}}$$

In order to apply this theory to test hypotheses about β or construct confidence intervals for β, we shall have to use the following theorem.

THEOREM 3. Under the assumptions of normal regression analysis, $\frac{n\hat{\sigma}^2}{\sigma^2}$ is a value of a random variable having the chi-square distribution with $n-2$ degrees of freedom. Furthermore, this random variable and $\hat{B}$ are independent.

A proof of this theorem is referred to at the end of this chapter.

Making use of this theorem as well as the result proved earlier that $\hat{B}$ has a normal distribution with the mean β and the variance $\frac{\sigma^2}{S_{xx}}$, we find that the definition of the t distribution leads to the following theorem.

THEOREM 4. Under the assumptions of normal regression analysis,

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

is a value of a random variable having the t distribution with $n-2$ degrees of freedom.

Based on this statistic, let us now test a hypothesis about the regression coefficient β.

EXAMPLE 5

With reference to the data in the table in Section 3 pertaining to the amount of time that 10 persons studied for a certain test and the scores that they obtained, test the null hypothesis $\beta = 3$ against the alternative hypothesis $\beta > 3$ at the 0.01 level of significance.

Solution

1. H_0: $\beta = 3$
   H_1: $\beta > 3$
   $\alpha = 0.01$
2. Reject the null hypothesis if $t \geqq 2.896$, where t is determined in accordance with Theorem 4 and 2.896 is the value of $t_{0.01,8}$ obtained from the Table IV of "Statistical Tables."

3. Calculating $\sum y^2 = 36{,}562$ from the original data and copying the other quantities from Section 3, we get

$$S_{yy} = 36{,}562 - \frac{1}{10}(564)^2 = 4{,}752.4$$

and

$$\hat{\sigma} = \sqrt{\frac{1}{10}[4{,}752.4 - (3.471)(1{,}305)]} = 4.720$$

so that

$$t = \frac{3.471 - 3}{4.720}\sqrt{\frac{8 \cdot 376}{10}} = 1.73$$

4. Since $t = 1.73$ is less than 2.896, the null hypothesis cannot be rejected; we cannot conclude that on the average an extra hour of study will increase the score by more than 3 points.

Letting $\hat{\Sigma}$ be the random variable whose values are $\hat{\sigma}$, we have

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

according to Theorem 4. Writing this as

$$P\left[\hat{B} - t_{\alpha/2,\,n-2} \cdot \hat{\Sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} < \beta < \hat{B} + t_{\alpha/2,\,n-2} \cdot \hat{\Sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}\right] = 1 - \alpha$$

we arrive at the following confidence interval formula.

THEOREM 5. Under the assumptions of normal regression analysis,

$$\hat{\beta} - t_{\alpha/2,\,n-2} \cdot \hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}} < \beta < \hat{\beta} + t_{\alpha/2,\,n-2} \cdot \hat{\sigma}\sqrt{\frac{n}{(n-2)S_{xx}}}$$

is a $(1-\alpha)100\%$ confidence interval for the parameter β.

EXAMPLE 6

With reference to the same data as in Example 5, construct a 95% confidence interval for β.

Solution

Copying the various quantities from Example 4 and Section 4 and substituting them together with $t_{0.025,8} = 2.306$ into the confidence interval formula of Theorem 5, we get

$$3.471 - (2.306)(4.720)\sqrt{\frac{10}{8(376)}} < \beta < 3.471 + (2.306)(4.720)\sqrt{\frac{10}{8(376)}}$$

or

$$2.84 < \beta < 4.10$$

Since most realistically complex regression problems require fairly extensive calculations, they are virtually always done nowadays by using appropriate computer software. A printout obtained for our illustration using MINITAB software is shown in Figure 4; as can be seen, it provides not only the values of $\hat{\alpha}$ and $\hat{\beta}$ in the column headed COEFFICIENT, but also estimates of the standard deviations of the sampling distributions of $\hat{A}$ and $\hat{B}$ in the column headed ST. DEV. OF COEF. Had we used this printout in Example 5, we could have written the value of the t statistic directly as

$$t = \frac{3.471 - 3}{0.2723} = 1.73$$

and in Example 6 we could have written the confidence limits directly as $3.471 \pm (2.306)(0.2723)$.

```
MTB  > NAME C1 = 'X'
MTB  > NAME C2 = 'Y'
MTB  > SET C1
DATA > 4  9  1Ø  14  4  7  12  22  1  17
MTB  > SET C2
DATA > 31  58  65  73  37  44  6Ø  91  21  84
MTB  > REGR C2 1 C1

  THE REGRESSION EQUATION IS
  Y = 21.7 + 3.47 X

                                ST. DEV.     T-RATIO =
  COLUMN     COEFFICIENT       OF COEF.     COEF/S.D.
             21.693              3.194          6.79
  X           3.47Ø7             Ø.2723        12.74
```

Figure 4. Computer printout for Examples 4, 5, and 6.

Exercises

17. Making use of the fact that $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$ and $\hat{\beta} = \dfrac{S_{xy}}{S_{xx}}$, show that

$$\sum_{i=1}^{n}[y_i - (\hat{\alpha} + \hat{\beta}x_i)]^2 = S_{yy} - \hat{\beta}S_{xy}$$

18. Show that
(a) $\hat{\Sigma}^2$, the random variable corresponding to $\hat{\sigma}^2$, is not an unbiased estimator of σ^2;
(b) $S_e^2 = \dfrac{n \cdot \hat{\Sigma}^2}{n-2}$ is an unbiased estimator of σ^2.
The quantity s_e is often referred to as the **standard error of estimate**.

19. Using s_e (see Exercise 18) instead of $\hat{\sigma}$, rewrite
(a) the expression for t in Theorem 4;
(b) the confidence interval formula of Theorem 5.

20. Under the assumptions of normal regression analysis, show that
(a) the least squares estimate of α in Theorem 2 can be written in the form

$$\hat{\alpha} = \sum_{i=1}^{n}\left[\frac{S_{xx} + n\bar{x}^2 - n\bar{x}x_i}{nS_{xx}}\right]y_i$$

(b) $\hat{A}$ has a normal distribution with

$$E(\hat{A}) = \alpha \quad \text{and} \quad \text{var}(\hat{A}) = \frac{(S_{xx} + n\bar{x}^2)\sigma^2}{nS_{xx}}$$

21. This question has been intentionally omitted for this edition.

22. Use the result of part (b) of Exercise 20 to show that

$$z = \frac{(\hat{\alpha} - \alpha)\sqrt{nS_{xx}}}{\sigma\sqrt{S_{xx} + n\bar{x}^2}}$$

is a value of a random variable having the standard normal distribution. Also, use the first part of Theorem 3 and the fact that $\hat{A}$ and $\dfrac{n\hat{\Sigma}^2}{\sigma^2}$ are independent to show that

$$t = \frac{(\hat{\alpha} - \alpha)\sqrt{(n-2)S_{xx}}}{\hat{\sigma}\sqrt{S_{xx} + n\bar{x}^2}}$$

is a value of a random variable having the t distribution with $n-2$ degrees of freedom.

23. Use the results of Exercises 20 and 21 and the fact that $E(\hat{\mathrm{B}}) = \beta$ and $\mathrm{var}(\hat{\mathrm{B}}) = \dfrac{\sigma^2}{S_{xx}}$ to show that $\hat{Y}_0 = \hat{\mathrm{A}} + \hat{\mathrm{B}}x_0$ is a random variable having a normal distribution with the mean

$$\alpha + \beta x_0 = \mu_{Y|x_0}$$

and the variance

$$\sigma^2\left[\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

Also, use the first part of Theorem 3 as well as the fact that $\hat{Y}_0$ and $\dfrac{n\hat{\Sigma}^2}{\sigma^2}$ are independent to show that

$$t = \frac{(\hat{y}_0 - \mu_{Y|x_0})\sqrt{n-2}}{\hat{\sigma}\sqrt{1 + \dfrac{n(x_0 - \bar{x})^2}{S_{xx}}}}$$

is a value of a random variable having the t distribution with $n-2$ degrees of freedom.

24. Derive a $(1-\alpha)100\%$ confidence interval for $\mu_{Y|x_0}$, the mean of Y at $x = x_0$, by solving the double inequality $-t_{\alpha/2,\,n-2} < t < t_{\alpha/2,\,n-2}$ with t given by the formula of Exercise 23.

25. Use the results of Exercises 20 and 21 and the fact that $E(\hat{\mathrm{B}}) = \beta$ and $\mathrm{var}(\hat{\mathrm{B}}) = \dfrac{\sigma^2}{S_{xx}}$ to show that $Y_0 - (\hat{\mathrm{A}} + \hat{\mathrm{B}}x_0)$ is a random variable having a normal distribution with zero mean and the variance

$$\sigma^2\left[1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right]$$

Here Y_0 has a normal distribution with the mean $\alpha + \beta x_0$ and the variance σ^2; that is, Y_0 is a future observation of Y corresponding to $x = x_0$. Also, use the first part of Theorem 3 as well as the fact that $Y_0 - (\hat{\mathrm{A}} + \hat{\mathrm{B}}x_0)$ and $\dfrac{n\hat{\Sigma}^2}{\sigma^2}$ are independent to show that

$$t = \frac{[y_0 - (\hat{\alpha} + \hat{\beta}x_0)]\sqrt{n-2}}{\hat{\sigma}\sqrt{1 + n + \dfrac{n(x_0 - \bar{x})^2}{S_{xx}}}}$$

is a value of a random variable having the t distribution with $n-2$ degrees of freedom.

26. Solve the double inequality $-t_{\alpha/2,n-2} < t < t_{\alpha/2,n-2}$ with t given by the formula of Exercise 25 so that the middle term is y_0 and the two limits can be calculated without knowledge of y_0. Note that although the resulting double inequality may be interpreted like a confidence interval, it is not designed to estimate a parameter; instead, it provides **limits of prediction** for a future observation of Y that corresponds to the (given or observed) value x_0.

5 Normal Correlation Analysis

In normal correlation analysis we drop the assumption that the x_i are fixed constants, analyzing the set of paired data $\{(x_i, y_i);\ i = 1, 2, \ldots, n\}$, where the x_i's and y_i's are values of a random sample from a bivariate normal population with the parameters μ_1, μ_2, σ_1, σ_2, and ρ. To estimate these parameters by the method of maximum likelihood, we shall have to maximize the likelihood

$$L = \prod_{i=1}^{n} f(x_i, y_i)$$

and to this end we shall have to differentiate L, or $\ln L$, partially with respect to μ_1, μ_2, σ_1, σ_2, and ρ, equate the resulting expressions to zero, and then solve the resulting system of equations for the five parameters. Leaving the details to the reader, let us merely state that when $\dfrac{\partial \ln L}{\partial \mu_1}$ and $\dfrac{\partial \ln L}{\partial \mu_2}$ are equated to zero, we get

$$-\frac{\sum_{i=1}^{n}(x_i-\mu_1)}{\sigma_1^2}+\frac{\rho\sum_{i=1}^{n}(y_i-\mu_2)}{\sigma_1\sigma_2}=0$$

and

$$-\frac{\rho\sum_{i=1}^{n}(x_i-\mu_1)}{\sigma_1\sigma_2}+\frac{\sum_{i=1}^{n}(y_i-\mu_2)}{\sigma_2^2}=0$$

Solving these two equations for μ_1 and μ_2, we find that the maximum likelihood estimates of these two parameters are

$$\hat{\mu}_1=\bar{x} \quad \text{and} \quad \hat{\mu}_2=\bar{y}$$

that is, the respective sample means. Subsequently, equating $\frac{\partial \ln L}{\partial \sigma_1}$, $\frac{\partial \ln L}{\partial \sigma_2}$, and $\frac{\partial \ln L}{\partial \rho}$ to zero and substituting $\bar{x}$ and $\bar{y}$ for μ_1 and μ_2, we obtain a system of equations whose solution is

$$\hat{\sigma}_1=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}, \qquad \hat{\sigma}_2=\sqrt{\frac{\sum_{i=1}^{n}(y_i-\bar{y})^2}{n}}$$

$$\hat{\rho}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

(A detailed derivation of these maximum likelihood estimates is referred to at the end of this chapter.) It is of interest to note that the maximum likelihood estimates of σ_1 and σ_2 are identical with the one obtained for the standard deviation of the univariate normal distribution; they differ from the respective sample standard deviations s_1 and s_2 only by the factor $\sqrt{\frac{n-1}{n}}$.

The estimate $\hat{\rho}$, called the **sample correlation coefficient**, is usually denoted by the letter r, and its calculation is facilitated by using the following alternative, but equivalent, computing formula.

Theorem 6. If $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ are the values of a random sample from a bivariate population, then

$$r=\frac{S_{xy}}{\sqrt{S_{xx}\cdot S_{yy}}}$$

Since ρ measures the strength of the linear relationship between X and Y, there are many problems in which the estimation of ρ and tests concerning ρ are of special interest. When $\rho = 0$, the two random variables are uncorrelated, and, as we have

already seen, in the case of the bivariate normal distribution this means that they are also independent. When ρ equals $+1$ or -1, it follows from the relationship

$$\sigma^2_{Y|x} = \sigma^2 = \sigma_2^2(1-\rho^2)$$

where $\sigma = 0$, and this means that there is a perfect linear relationship between X and Y. Using the invariance property of maximum likelihood estimators, we can write

$$\hat{\sigma}^2 = \hat{\sigma}_2^2(1-r^2)$$

which not only provides an alternative computing formula for finding $\hat{\sigma}^2$, but also serves to tie together the concepts of regression and correlation. From this formula for $\hat{\sigma}^2$ it is clear that when $\hat{\sigma}^2 = 0$, that is, when the set of data points $\{(x_i, y_i); i = 1, 2, \ldots, n\}$ fall on a straight line, then r will equal $+1$ or -1. We take $r = +1$ when the line has a positive slope and $r = -1$ when it has a negative slope. In order to interpret values of r between 0 and $+1$ or 0 and -1, we solve the preceding equation for r^2 and multiply by 100, getting

$$100r^2 = \frac{\hat{\sigma}_2^2 - \hat{\sigma}^2}{\hat{\sigma}_2^2} \cdot 100$$

where $\hat{\sigma}_2^2$ measures the total variation of the y's, $\hat{\sigma}^2$ measures the conditional variation of the y's for fixed values of x, and hence $\hat{\sigma}_2^2 - \hat{\sigma}^2$ measures that part of the total variation of the y's that is accounted for by the relationship with x. *Thus*, $100r^2$ *is the percentage of the total variation of the* y*'s that is accounted for by the relationship with* x. For instance, when $r = 0.5$, then 25 percent of the variation of the y's is accounted for by the relationship with x; when $r = 0.7$, then 49 percent of the variation of the y's is accounted for by the relationship with x; and we might thus say that a correlation of $r = 0.7$ is almost "twice as strong" as a correlation of $r = 0.5$. Similarly, we might say that a correlation of $r = 0.6$ is "nine times as strong" as a correlation of $r = 0.2$.

EXAMPLE 7

Suppose that we want to determine on the basis of the following data whether there is a relationship between the time, in minutes, it takes a secretary to complete a certain form in the morning and in the late afternoon:

Morning x	*Afternoon* y
8.2	8.7
9.6	9.6
7.0	6.9
9.4	8.5
10.9	11.3
7.1	7.6
9.0	9.2
6.6	6.3
8.4	8.4
10.5	12.3

Compute and interpret the sample correlation coefficient.

Solution
From the data we get $n = 10$, $\Sigma x = 86.7$, $\Sigma x^2 = 771.35$, $\Sigma y = 88.8$, $\Sigma y^2 = 819.34$, and $\Sigma xy = 792.92$, so

$$S_{xx} = 771.35 - \frac{1}{10}(86.7)^2 = 19.661$$

$$S_{yy} = 819.34 - \frac{1}{10}(88.8)^2 = 30.796$$

$$S_{xy} = 792.92 - \frac{1}{10}(86.7)(88.8) = 23.024$$

and

$$r = \frac{23.024}{\sqrt{(19.661)(30.796)}} = 0.936$$

This is indicative of a positive association between the time it takes a secretary to perform the given task in the morning and in the late afternoon, and this is also apparent from the **scattergram** of Figure 5. Since $100r^2 = 100(0.936)^2 = 87.6$, we can say that almost 88 percent of the variation of the y's is accounted for by the implicit linear relationship with x.

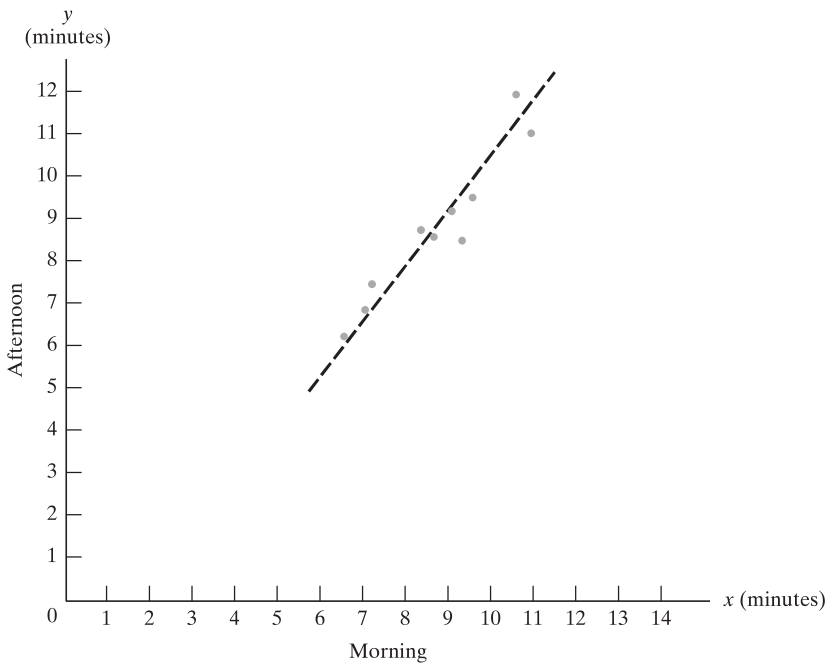

Figure 5. Scattergram of data of Example 7.

Since the sampling distribution of R for random samples from bivariate normal populations is rather complicated, it is common practice to base confidence intervals for ρ and tests concerning ρ on the statistic