CHAPTER 14

Correlation Theory

CORRELATION AND REGRESSION

In Chapter 13 we considered the problem of *regression*, or *estimation*, of one variable (the dependent variable) from one or more related variables (the independent variables). In this chapter we consider the closely related problem of *correlation*, or the degree of relationship between variables, which seeks to determine *how well* a linear or other equation describes or explains the relationship between variables.

If all values of the variables satisfy an equation exactly, we say that the variables are *perfectly correlated* or that there is *perfect correlation* between them. Thus the circumferences C and radii r of all circles are perfectly correlated since $C = 2\pi r$. If two dice are tossed simultaneously 100 times, there is no relationship between corresponding points on each die (unless the dice are loaded); that is, they are *uncorrelated*. Such variables as the height and weight of individuals would show *some* correlation.

When only two variables are involved, we speak of *simple correlation* and *simple regression*. When more than two variables are involved, we speak of *multiple correlation* and *multiple regression*. This chapter considers only simple correlation. Multiple correlation and regression are considered in Chapter 15.

E-next
THE NEXT LEVEL OF EDUCATION

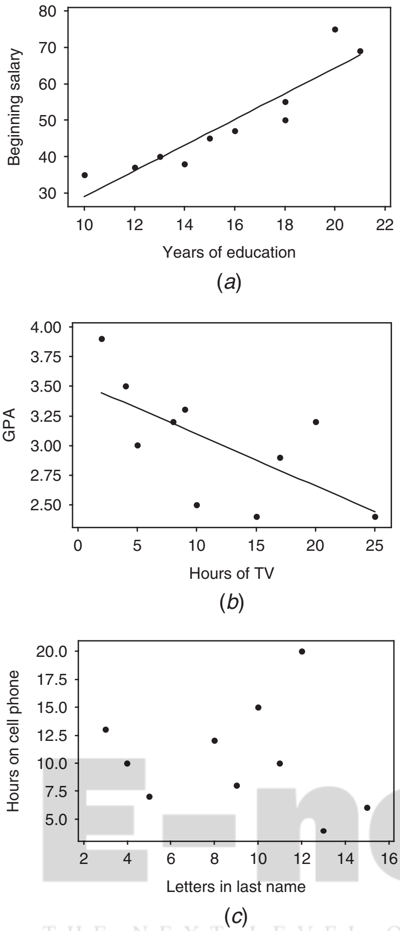
LINEAR CORRELATION

If X and Y denote the two variables under consideration, a *scatter diagram* shows the location of points (X, Y) on a rectangular coordinate system. If all points in this scatter diagram seem to lie near a line, as in Figs. 14-1(*a*) and 14-1(*b*), the correlation is called *linear*. In such cases, as we have seen in Chapter 13, a linear equation is appropriate for purposes of regression (or estimation).

If Y tends to increase as X increases, as in Fig. 14-1(*a*), the correlation is called *positive*, or *direct*, *correlation*. If Y tends to decrease as X increases, as in Fig. 14-1(*b*), the correlation is called *negative*, or *inverse*, *correlation*.

If all points seem to lie near some curve, the correlation is called *nonlinear*, and a nonlinear equation is appropriate for regression, as we have seen in Chapter 13. It is clear that nonlinear correlation can be sometimes positive and sometimes negative.

If there is no relationship indicated between the variables, as in Fig. 14-1(*c*), we say that there is *no correlation* between them (i.e., they are *uncorrelated*).



E-next
THE NEXT LEVEL OF EDUCATION

Fig. 14-1 Examples of positive correlation, negative correlation and no correlation. (*a*) Beginning salary and years of formal education are positively correlated; (*b*) Grade point average (GPA) and hours spent watching TV are negatively correlated; (*c*) There is no correlation between hours on a cell phone and letters in last name.

MEASURES OF CORRELATION

We can determine in a *qualitative* manner how well a given line or curve describes the relationship between variables by direct observation of the scatter diagram itself. For example, it is seen that a straight line is far more helpful in describing the relation between X and Y for the data of Fig. 14-1(*a*) than for the data of Fig. 14-1(*b*) because of the fact that there is less scattering about the line of Fig. 14-1(*a*).

If we are to deal with the problem of scattering of sample data about lines or curves in a *quantitative* manner, it will be necessary for us to devise *measures of correlation*

THE LEAST-SQUARES REGRESSION LINES

We first consider the problem of how well a straight line explains the relationship between two variables. To do this, we shall need the equations for the least-squares regression lines obtained in Chapter 13. As we have seen, the least-squares regression line of Y on X is

$$Y = a_0 + a_1 X \tag{1}$$

https://E-next.in

where a_0 and a_1 are obtained from the normal equations

$$\begin{aligned} \sum Y &= a_0 N \quad + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned} \tag{2}$$

which yield

$$\begin{aligned} a_0 &= \frac{(\sum Y)(\sum X^2) - (\sum X)(\sum XY)}{N \sum X^2 - (\sum X)^2} \\ a_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum X^2 - (\sum X)^2} \end{aligned} \tag{3}$$

Similarly, the regression line of X on Y is given by

$$X = b_0 + b_1 Y \tag{4}$$

where b_0 and b_1 are obtained from the normal equations

$$\begin{aligned} \sum X &= b_0 N \quad + b_1 \sum Y \\ \sum XY &= b_0 \sum X + b_1 \sum Y^2 \end{aligned} \tag{5}$$

which yield

E-next
THE NEXT LEVEL OF EDUCATION

$$\begin{aligned} b_0 &= \frac{(\sum X)(\sum Y^2) - (\sum Y)(\sum XY)}{N \sum Y^2 - (\sum Y)^2} \\ b_1 &= \frac{N \sum XY - (\sum X)(\sum Y)}{N \sum Y^2 - (\sum Y)^2} \end{aligned} \tag{6}$$

Equations (*1*) and (*4*) can also be written, respectively, as

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{and} \qquad x = \left(\frac{\sum xy}{\sum y^2}\right)y \tag{7}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

The regression equations are identical if and only if all points of the scatter diagram lie on a line. In such case there is *perfect linear correlation* between X and Y.

STANDARD ERROR OF ESTIMATE

If we let Y_{est} represent the value of Y for given values of X as estimated from equation (*1*), a measure of the scatter about the regression line of Y on X is supplied by the quantity

$$s_{Y.X} = \sqrt{\frac{\sum (Y - Y_{\text{est}})^2}{N}} \tag{8}$$

which is called the *standard error of estimate of Y on X.*

If the regression line (*4*) is used, an analogous standard error of estimate of X on Y is defined by

$$s_{X.Y} = \sqrt{\frac{\sum (X - X_{\text{est}})^2}{N}} \tag{9}$$

In general, $s_{Y.X} \neq s_{X.Y}$.

Equation (*8*) can be written

$$s^2_{Y.X} = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N} \tag{10}$$

which may be more suitable for computation (see Problem 14.3). A similar expression exists for equation (*9*).

The standard error of estimate has properties analogous to those of the standard deviation. For example, if we construct lines parallel to the regression line of Y on X at respective vertical distances $s_{Y.X}$, $2s_{Y.X}$, and $3s_{Y.X}$ from it, we should find, if N is large enough, that there would be included between these lines about 68%, 95%, and 99.7% of the sample points.

Just as a modified standard deviation given by

$$\hat{s} = \sqrt{\frac{N}{N-1}}\, s$$

was found useful for small samples, so a modified standard error of estimate given by

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-2}}\, s_{Y.X}$$

is useful. For this reason, some statisticians prefer to define equation (*8*) or (*9*) with $N - 2$ replacing N in the denominator.

E-next
LEVEL OF EDUCATION

EXPLAINED AND UNEXPLAINED VARIATION

The *total variation* of Y is defined as $\sum (Y - \bar{Y})^2$: that is, the sum of the squares of the deviations of the values of Y from the mean $\bar{Y}$. As shown in Problem 14.7, this can be written

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2 \tag{11}$$

The first term on the right of equation (*11*) is called the *unexplained variation*, while the second term is called the *explained variation*—so called because the deviations $Y_{\text{est}} - \bar{Y}$ have a definite pattern, while the deviations $Y - Y_{\text{est}}$ behave in a random or unpredictable manner. Similar results hold for the variable X.

COEFFICIENT OF CORRELATION

The ratio of the explained variation to the total variation is called the *coefficient of determination*. If there is zero explained variation (i.e., the total variation is all unexplained), this ratio is 0. If there is zero unexplained variation (i.e., the total variation is all explained), the ratio is 1. In other cases the ratio lies between 0 and 1. Since the ratio is always nonnegative, we denote it by r^2. The quantity r, called the *coefficient of correlation* (or briefly *correlation coefficient*), is given by

$$r = \pm\sqrt{\frac{\text{explained variation}}{\text{total variation}}} = \pm\sqrt{\frac{\sum (Y_{\text{est}} - \bar{Y})^2}{\sum (Y - \bar{Y})^2}} \tag{12}$$

and varies between -1 and $+1$. The $+$ and $-$ signs are used for positive linear correlation and negative linear correlation, respectively. Note that r is a dimensionless quantity; that is, it does not depend on the units employed.

By using equations (*8*) and (*11*) and the fact that the standard deviation of Y is

$$s_Y = \sqrt{\frac{\sum (Y - \bar{Y})^2}{N}} \tag{13}$$

we find that equation (*12*) can be written, disregarding the sign, as

$$r = \sqrt{1 - \frac{s_{Y.X}^2}{s_Y^2}} \qquad \text{or} \qquad s_{Y.X} = s_Y \sqrt{1 - r^2} \tag{14}$$

Similar equations exist when X and Y are interchanged.

For the case of linear correlation, the quantity r is the same regardless of whether X or Y is considered the independent variable. Thus r is a very good measure of the linear correlation between two variables.

REMARKS CONCERNING THE CORRELATION COEFFICIENT

The definitions of the correlation coefficient in equations (*12*) and (*14*) are quite general and can be used for nonlinear relationships as well as for linear ones, the only differences being that Y_{est} is computed from a nonlinear regression equation in place of a linear equation and that the $+$ and $-$ signs are omitted. In such case equation (*8*), defining the standard error of estimate, is perfectly general. Equation (*10*), however, which applies to linear regression only, must be modified. If, for example, the estimating equation is

E-next
THE NEXT LEVEL OF EDUCATION

$$Y = a_0 + a_1 X + a_2 X^2 + \cdots + a_{n-1} X^{n-1} \tag{15}$$

then equation (*10*) is replaced by

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY - \cdots - a_{n-1} \sum X^{n-1} Y}{N} \tag{16}$$

In such case the *modified standard error of estimate* (discussed earlier in this chapter) is

$$\hat{s}_{Y.X} = \sqrt{\frac{N}{N-n}}\, s_{Y.X}$$

where the quantity $N - n$ is called the number of *degrees of freedom*.

It must be emphasized that in every case the computed value of r measures the degree of the relationship relative to the type of equation that is actually assumed. Thus if a linear equation is assumed and equation (*12*) or (*14*) yields a value of r near zero, it means that there is almost no *linear correlation* between the variables. However, it does not mean that there is no correlation at all, since there may actually be a high *nonlinear correlation* between the variables. In other words, the correlation coefficient measures the goodness of fit between (1) the equation actually assumed and (2) the data. Unless otherwise specified, the term *correlation coefficient* is used to mean *linear correlation coefficient*.

It should also be pointed out that a high correlation coefficient (i.e., near 1 or -1) does not necessarily indicate a direct dependence of the variables. Thus there may be a high correlation between the number of books published each year and the number of thunderstorms each year. Such examples are sometimes referred to as *nonsense*, or *spurious*, *correlations*.

PRODUCT-MOMENT FORMULA FOR THE LINEAR CORRELATION COEFFICIENT

If a linear relationship between two variables is assumed, equation (*12*) becomes

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} \tag{17}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$ (see Problem 14.10). This formula, which automatically gives the proper sign of r, is called the *product-moment formula* and clearly shows the symmetry between X and Y.

If we write

$$s_{XY} = \frac{\sum xy}{N} \qquad s_X = \sqrt{\frac{\sum x^2}{N}} \qquad s_Y = \sqrt{\frac{\sum y^2}{N}} \tag{18}$$

then s_X and s_Y will be recognized as the standard deviations of the variables X and Y, respectively, while s_X^2 and s_Y^2 are their variances. The new quantity s_{XY} is called the *covariance* of X and Y. In terms of the symbols of formulas (*18*), formula (*17*) can be written

$$r = \frac{s_{XY}}{s_X s_Y} \tag{19}$$

Note that r is not only independent of the choice of units of X and Y, but is also independent of the choice of origin.

SHORT COMPUTATIONAL FORMULAS

Formula (*17*) can be written in the equivalent form

$$r = \frac{N \sum XY - (\sum X)(\sum Y)}{\sqrt{[n \sum X^2 - (\sum X)^2][N \sum Y^2 - (\sum Y)^2]}} \tag{20}$$

which is often used in computing r.

E-next THE NEXT LEVEL OF EDUCATION

For data grouped as in a *bivariate frequency table*, or *bivariate frequency distribution* (see Problem 14.17), it is convenient to use a *coding method* as in previous chapters. In such case, formula (*20*) can be written

$$r = \frac{N \sum f u_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N \sum f_X u_X^2 - (\sum f_X u_X)^2][N \sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

(see Problem 14.18). For convenience in calculations using this formula, a *correlation table* is used (see Problem 14.19).

For grouped data, formulas (*18*) can be written

$$s_{XY} = c_X c_Y \left[\frac{\sum f u_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N}\right)\left(\frac{\sum f_Y u_Y}{N}\right)\right] \tag{22}$$

$$s_X = c_X \sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N}\right)^2} \tag{23}$$

$$s_Y = c_Y \sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N}\right)^2} \tag{24}$$

where c_X and c_Y are the class-interval widths (assumed constant) corresponding to the variables X and Y, respectively. Note that (*23*) and (*24*) are equivalent to formula (*11*) of Chapter 4.

Formula (*19*) is seen to be equivalent to (*21*) if results (*22*) to (*24*) are used.

REGRESSION LINES AND THE LINEAR CORRELATION COEFFICIENT

The equation of the least-squares line $Y = a_0 + a_1 X$, the regression line of Y on X, can be written

$$Y - \bar{Y} = \frac{rs_Y}{s_X}(X - \bar{X}) \qquad \text{or} \qquad y = \frac{rs_Y}{s_X}x \tag{25}$$

Similarly, the regression line of X on Y, $X = b_0 + b_1 Y$, can be written

$$X - \bar{X} = \frac{rs_X}{s_Y}(Y - \bar{Y}) \qquad \text{or} \qquad x = \frac{rs_X}{s_Y}y \tag{26}$$

The slopes of the lines in equations (*25*) and (*26*) are equal if and only if $r = \pm 1$. In such case the two lines are identical and there is perfect linear correlation between the variables X and Y. If $r = 0$, the lines are at right angles and there is no linear correlation between X and Y. Thus the linear correlation coefficient measures the departure of the two regression lines.

Note that if equations (*25*) and (*26*) are written $Y = a_0 + a_1 X$ and $X = b_0 + b_1 Y$, respectively, then $a_1 b_1 = r^2$ (see Problem 14.22).

CORRELATION OF TIME SERIES

If each of the variables X and Y depends on time, it is possible that a relationship may exist between X and Y even though such relationship is not necessarily one of direct dependence and may produce "nonsense correlation." The correlation coefficient is obtained simply by considering the pairs of values (X, Y) corresponding to the various times and proceeding as usual, making use of the above formulas (see Problem 14.28).

It is possible to attempt to correlate values of a variable X at certain times with corresponding values of X at earlier times. Such correlation is often called *autocorrelation*.

E-next THE NEXT LEVEL OF EDUCATION

CORRELATION OF ATTRIBUTES

The methods described in this chapter do not enable us to consider the correlation of variables that are nonnumerical by nature, such as the *attributes* of individuals (e.g., hair color, eye color, etc.). For a discussion of the correlation of attributes, see Chapter 12.

SAMPLING THEORY OF CORRELATION

The N pairs of values (X, Y) of two variables can be thought of as samples from a population of all such pairs that are possible. Since two variables are involved, this is called a *bivariate population*, which we assume to be a *bivariate normal distribution*.

We can think of a theoretical population coefficient of correlation, denoted by ρ, which is estimated by the sample correlation coefficient r. Tests of significance or hypotheses concerning various values of ρ require knowledge of the sampling distribution of r. For $\rho = 0$ this distribution is symmetrical, and a statistic involving Student's distribution can be used. For $\rho \neq 0$, the distribution is skewed; in such case a transformation developed by Fisher produces a statistic that is approximately normally distributed. The following tests summarize the procedures involved:

1. **Test of Hypothesis** $\rho = 0$. Here we use the fact that the statistic

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} \tag{27}$$

has Student's distribution with $\nu = N - 2$ degrees of freedom (see Problems 14.31 and 14.32).

2. **Test of Hypothesis** $\rho = \rho_0 \neq 0$. Here we use the fact that the statistic

$$Z = \tfrac{1}{2}\log_e\left(\frac{1+r}{1-r}\right) = 1.1513\log_{10}\left(\frac{1+r}{1-r}\right) \tag{28}$$

where $e = 2.71828\ldots$, is approximately normally distributed with mean and standard deviation given by

$$\mu_Z = \tfrac{1}{2}\log_e\left(\frac{1+\rho_0}{1-\rho_0}\right) = 1.1513\log_{10}\left(\frac{1+\rho_0}{1-\rho_0}\right) \qquad \sigma_Z = \frac{1}{\sqrt{N-3}} \tag{29}$$

Equations (*28*) and (*29*) can also be used to find confidence limits for correlation coefficients (see Problems 14.33 and 14.34). Equation (*28*) is called *Fisher's Z transformation*.

3. **Significance of a Difference between Correlation Coefficients.** To determine whether two correlation coefficients, r_1 and r_2, drawn from samples of sizes N_1 and N_2, respectively, differ significantly from each other, we compute Z_1 and Z_2 corresponding to r_1 and r_2 by using equation (*28*). We then use the fact that the test statistic

$$z = \frac{Z_1 - Z_2 - \mu_{Z_1-Z_2}}{\sigma_{Z_1-Z_2}} \tag{30}$$

where

$$\mu_{Z_1-Z_2} = \mu_{Z_1} - \mu_{Z_2}$$

and

$$\sigma_{Z_1-Z_2} = \sqrt{\sigma^2_{Z_1} + \sigma^2_{Z_2}} = \sqrt{\frac{1}{N_1-3} + \frac{1}{N_2-3}}$$

is normally distributed (see Problem 14.35).

E-next
THE NEXT LEVEL OF EDUCATION

SAMPLING THEORY OF REGRESSION

The regression equation $Y = a_0 + a_1X$ is obtained on the basis of sample data. We are often interested in the corresponding regression equation for the population from which the sample was drawn. The following are three tests concerning such a population:

1. **Test of Hypothesis** $a_1 = A_1$. To test the hypothesis that the regression coefficient a_1 is equal to some specified value A_1, we use the fact that the statistic

$$t = \frac{a_1 - A_1}{s_{Y.X}/s_X}\sqrt{N-2} \tag{31}$$

has Student's distribution with $N-2$ degrees of freedom. This can also be used to find confidence intervals for population regression coefficients from sample values (see Problems 14.36 and 14.37).

2. **Test of Hypothesis for Predicted Values.** Let Y_0 denote the predicted value of Y corresponding to $X = X_0$ as estimated from the sample regression equation (i.e., $Y_0 = a_0 + a_1X_0$). Let Y_p denote the predicted value of Y corresponding to $X = X_0$ for the population. Then the statistic

$$t = \frac{Y_0 - Y_p}{s_{Y.X}\sqrt{N + 1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N-2} = \frac{Y_0 - Y_p}{\hat{s}_{X.Y}\sqrt{1 + 1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \tag{32}$$

has Student's distribution with $N-2$ degrees of freedom. From this, confidence limits for predicted population values can be found (see Problem 14.38).

3. **Test of Hypothesis for Predicted Mean Values.** Let Y_0 denote the predicted value of Y corresponding to $X = X_0$ as estimated from the sample regression equation (i.e., $Y_0 = a_0 + a_1 X_0$). Let $\bar{Y}_p$ denote the predicted *mean value* of Y corresponding to $X = X_0$ for the population. Then the statistic

$$t = \frac{Y_0 - \bar{Y}_p}{s_{Y.X}\sqrt{1 + (X_0 - \bar{X})^2/s_X^2}}\sqrt{N-2} = \frac{Y_0 - \bar{Y}_p}{\hat{s}_{Y.X}\sqrt{1/N + (X_0 - \bar{X})^2/(Ns_X^2)}} \tag{33}$$

has Student's distribution with $N - 2$ degrees of freedom. From this, confidence limits for predicted mean population values can be found (see Problem 14.39).

Solved Problems

SCATTER DIAGRAMS AND REGRESSION LINES

14.1 Table 14.1 shows [in inches (in)] the respective heights X and Y of a sample of 12 fathers and their oldest son.

(*a*) Construct a scatter diagram of the data.

(*b*) Find the least-squares regression line of the height of the father on the height of the son by solving the normal equations and by using SPSS.

(*c*) Find the least-squares regression line of the height of the son on the height of the father by solving the normal equations and by using STATISTIX.

Table 14.1

Height X of father (in)	65	63	67	64	68	62	70	66	68	67	69	71
Height Y of son (in)	68	66	68	65	69	66	68	65	71	67	68	70

E-next
THE NEXT LEVEL OF EDUCATION

SOLUTION

(*a*) The scatter diagram is obtained by plotting the points (X, Y) on a rectangular coordinate system, as shown in Fig. 14-2.

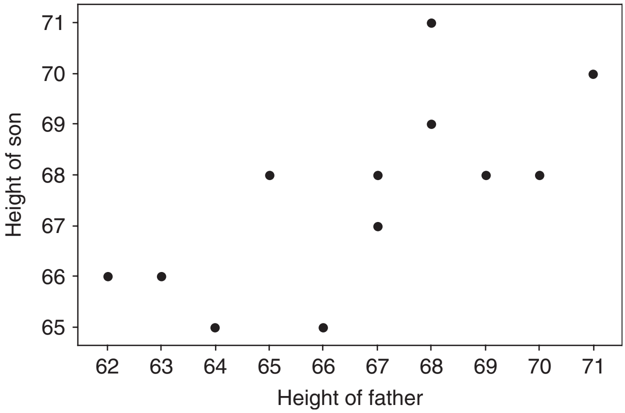


Fig. 14-2 Scatter diagram of the data in Table 14.1.

(*b*) The regression line of Y on X is given by $Y = a_0 + a_1 X$, where a_0 and a_1 are obtained by solving the normal equations.

$$\sum Y \quad = a_0 N + a_1 \sum X$$
$$\sum XY = a_0 \sum X + a_1 \sum X^2$$

https://E-next.in

The sums are shown in Table 14.2, from which the normal equations become

$$\begin{aligned} 12a_0 + 800a_1 &= 811 \\ 800a_0 + 53418a_1 &= 54107 \end{aligned}$$

from which we find that $a_0 = 35.82$ and $a_1 = 0.476$, and thus $Y = 35.82 + 0.476X$.

The SPSS pull-down **Analyze** → **Regression** → **Linear** gives the following partial output:

Coefficients[a]

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	35.825	10.178		3.520	.006
Htfather	.476	.153	.703	3.123	.011

[a] Dependent Variable: Htson.

Opposite the word (`Constant`) is the value of a_0 and opposite `Htfather` is the value of a_1.

Table 14.2

X	Y	X^2	XY	Y^2
65	68	4225	4420	4624
63	66	3969	4158	4356
67	68	4489	4556	4624
64	65	4096	4160	4225
68	69	4624	4692	4761
62	66	3844	4092	4356
70	68	4900	4760	4624
66	65	4356	4290	4225
68	71	4624	4828	5041
67	67	4489	4489	4489
69	68	4761	4692	4624
71	70	5041	4970	4900
$\sum X = 800$	$\sum Y = 811$	$\sum X^2 = 53{,}418$	$\sum XY = 54{,}107$	$\sum Y^2 = 54{,}849$

E-next
THE NEXT LEVEL OF EDUCATION

(*c*) The regression line of X on Y is given by $X = b_0 + b_1 Y$, where b_0 and b_1 are obtained by solving the normal equations

$$\begin{aligned} \sum X &= b_0 N \quad + b_1 \sum Y \\ \sum XY &= b_0 \sum Y + b_1 \sum Y^2 \end{aligned}$$

Using the sums in Table 14.2, these become

$$\begin{aligned} 12b_0 + 811b_1 &= 800 \\ 811b_0 + 54849b_1 &= 54107 \end{aligned}$$

from which we find that $b_0 = -3.38$ and $b_1 = 1.036$, and thus $X = -3.38 + 1.036Y$

The STATISTIX pull down **Statistics** → **Linear models** → **Linear regression** gives the following partial output:

```
Statistix 8.0
Unweighted Least Squares Linear Regression of Htfather

Predictor
Variable        Coefficient     Std Error         T            P
Constant         −3.37687       22.4377         −0.15        0.8834
Htson             1.03640        0.33188         3.12        0.0108
```

Opposite the word `constant` is the value of $b_0 = -3.37687$ and opposite `Htson` is the value of $b_1 = 1.0364$.

14.2 Work Problem 14.1 using MINITAB. Construct tables giving the fitted values, Y_{est}, and the residuals. Find the sum of squares for the residuals for both regression lines.

SOLUTION

The least-squares regression line of Y on X will be found first. A part of the MINITAB output is shown below. Table 14.3 gives the fitted values, the residuals, and the squares of the residuals for the regression line of Y on X.

Table 14.3

X	Y	Fitted value Y_{est}	Residual $Y - Y_{\text{est}}$	Residual squared
65	68	66.79	1.21	1.47
63	66	65.84	0.16	0.03
67	68	67.74	0.26	0.07
64	65	66.31	−1.31	1.72
68	69	68.22	0.78	0.61
62	66	65.36	0.64	0.41
70	68	69.17	−1.17	1.37
66	65	67.27	−2.27	5.13
68	71	68.22	2.78	7.74
67	67	67.74	−0.74	0.55
69	68	68.69	−0.69	0.48
71	70	69.65	0.35	0.12
			Sum = 0	Sum = 19.70

E-next
THE NEXT LEVEL OF EDUCATION

```
MTB > Regress 'Y' on 1 predictor 'X'
```

Regression Analysis

```
The regression equation is Y = 35.8 + 0.476 X
```

The Minitab output for finding the least-squares regression line of X on Y is as follows:

```
MTB > Regress 'X' on 1 predictor 'Y'
```

Regression Analysis

```
The regression equation is X = −3.4 + 1.04 Y
```

Table 14.4 gives the fitted values, the residuals, and the squares of the residuals for the regression line of X on Y.

Table 14.4

X	Y	Fitted Value X_{est}	Residual $X - X_{\text{est}}$	Residual Squared
65	68	67.10	−2.10	4.40
63	66	65.03	−2.03	4.10
67	68	67.10	−0.10	0.01
64	65	63.99	0.01	0.00
68	69	68.13	−0.13	0.02
62	66	65.03	−3.03	9.15
70	68	67.10	2.90	8.42
66	65	63.99	2.01	4.04
68	71	70.21	−2.21	4.87
67	67	66.06	0.94	0.88
69	68	67.10	1.90	3.62
71	70	69.17	1.83	3.34
			Sum = 0	Sum = 42.85

The comparison of the sums of squares of residuals indicates that the fit for the least-squares regression line of Y on X is much better than the fit for the least-squares regression line of X on Y. Recall that the smaller the sums of squares of residuals, the better the regression model fits the data. The height of the father is a better predictor of the height of the son than the height of the son is of the height of the father.

E-next
THE NEXT LEVEL OF EDUCATION

STANDARD ERROR OF ESTIMATE

14.3 If the regression line of Y on X is given by $Y = a_0 + a_1X$, prove that the standard error of estimate $s_{Y.X}$ is given by

$$s_{Y.X}^2 = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}$$

SOLUTION

The values of Y as estimated from the regression line are given by $Y_{\text{est}} = a_0 + a_1X$. Thus

$$s_{Y.X}^2 = \frac{\sum (Y - Y_{\text{est}})^2}{N} = \frac{\sum (Y - a_0 - a_1X)^2}{N}$$

$$= \frac{\sum Y(Y - a_0 - a_1X) - a_0 \sum (Y - a_0 - a_1X) - a_1 \sum X(Y - a_0 - a_1X)}{N}$$

But $$\sum (Y - a_0 - a_1X) = \sum Y - a_0N - a_1 \sum X = 0$$

and $$\sum X(Y - a_0 - a_1X) = \sum XY - a_0 \sum X - a_1 \sum X^2 = 0$$

since from the normal equations

$$\begin{aligned} \sum Y &= a_0N + a_1 \sum X \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 \end{aligned}$$

Thus $$s_{Y.X}^2 = \frac{\sum Y(Y - a_0 - a_1X)}{N} = \frac{\sum Y^2 - a_0 \sum Y - a_1 \sum XY}{N}$$

This result can be extended to nonlinear regression equations.

https://E-next.in

14.4 If $x = X - \bar{X}$ and $y = Y - \bar{Y}$, show that the result of Problem 14.3 can be written

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1 \sum xy}{N}$$

SOLUTION

From Problem 14.3, with $X = x + \bar{X}$ and $Y = y + \bar{Y}$, we have

$$\begin{aligned}
Ns_{Y.X}^2 &= \sum Y^2 - a_0 \sum Y - a_1 \sum XY = \sum (y + \bar{Y})^2 - a_0 \sum (y + \bar{Y}) - a_1 \sum (x + \bar{X})(y + \bar{Y}) \\
&= \sum (y^2 + 2y\bar{Y} + \bar{Y}^2) - a_0(\sum y + N\bar{Y}) - a_1 \sum (xy + \bar{X}y + x\bar{Y} + \bar{X}\bar{Y}) \\
&= \sum y^2 + 2\bar{Y} \sum y + N\bar{Y}^2 - a_0 N\bar{Y} - a_1 \sum xy - a_1\bar{X} \sum y - a_1 \bar{Y} \sum x - a_1 N\bar{X}\bar{Y} \\
&= \sum y^2 + N\bar{Y}^2 - a_0 N\bar{Y} - a_1 \sum xy - a_1 N\bar{X}\bar{Y} \\
&= \sum y^2 - a_1 \sum xy + N\bar{Y}(\bar{Y} - a_0 - a_1\bar{X}) \\
&= \sum y^2 - a_1 \sum xy
\end{aligned}$$

where we have used the results $\sum x = 0$, $\sum y = 0$, and $\bar{Y} = a_0 + a_1\bar{X}$ (which follows on dividing both sides of the normal equation $\sum Y = a_0 N + a_1 \sum X$ by N).

14.5 Compute the standard error of estimate, $s_{Y.X}$, for the data of Problem 14.1 by using (*a*) the definition and (*b*) the result of Problem 14.4.

E-next
THE NEXT LEVEL OF EDUCATION

SOLUTION

(*a*) From Problem 14.1(*b*) the regression line of Y on X is $Y = 35.82 + 0.476X$. Table 14.5 lists the actual values of Y (from Table 14.1) and the estimated values of Y, denoted by Y_{est}, as obtained from the regression line; for example, corresponding to $X = 65$ we have $Y_{\text{est}} = 35.82 + 0.476(65) = 66.76$. Also listed are the values $Y - Y_{\text{est}}$, which are needed in computing $s_{Y.X}$:

$$s_{Y.X}^2 = \frac{\sum (Y - Y_{\text{est}})}{N} = \frac{(1.24)^2 + (0.19)^2 + \cdots + (0.38)^2}{12} = 1.642$$

and $s_{Y.X} = \sqrt{1.1642} = 1.28$ in.

(*b*) From Problems 14.1, 14.2, and 14.4

$$s_{Y.X}^2 = \frac{\sum y^2 - a_1 \sum xy}{N} = \frac{38.92 - 0.476(40.34)}{12} = 1.643$$

and $s_{Y.X} = \sqrt{1.643} = 1.28$ in.

Table 14.5

X	65	63	67	64	68	62	70	66	68	67	69	71
Y	68	66	68	65	69	66	68	65	71	67	68	70
Y_{est}	66.76	65.81	67.71	66.28	68.19	65.33	69.14	67.24	68.19	67.71	68.66	69.62
$Y - Y_{\text{est}}$	1.24	0.19	0.29	−1.28	0.81	0.67	−1.14	−2.24	2.81	−0.71	−0.66	0.38

14.6 (*a*) Construct two lines which are parallel to the regression line of Problem 14.1 and which are at a vertical distance $S_{Y.X}$ from it.

(*b*) Determine the percentage of data points falling between these two lines.

SOLUTION

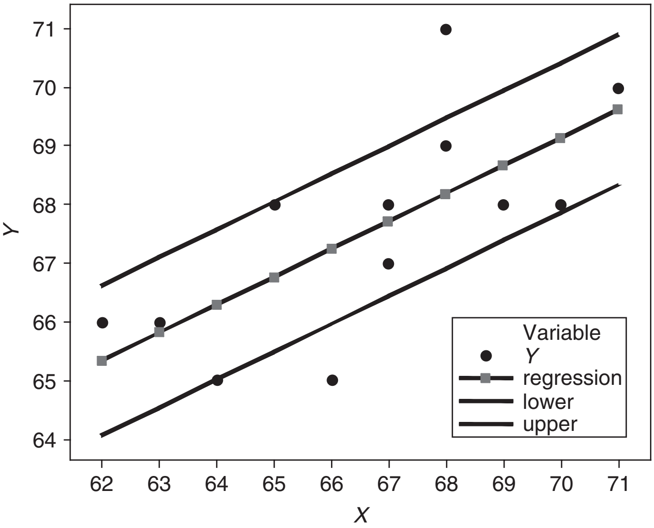


Fig. 14-3 Sixty-six percent of the data points are within $\boldsymbol{S_{Y.X}}$ of the regression line.

(*a*) The regression line $Y = 35.82 + 0.476X$, as obtained in Problem 14.1, is shown as the line with the squares on it. It is the middle of the three lines in Fig. 14-3. There are two other lines shown in Fig. 14-3. They are at a vertical distance $S_{Y.X} = 1.28$ from the regression line. They are called the lower and upper lines.

(*b*) The twelve data points are shown as black circles in Fig. 14-3. Eight of twelve or 66.7% of the data points are between the upper and lower lines. Two data points are outside the lines and two are on the lines.

E-next THE NEXT LEVEL OF EDUCATION

EXPLAINED AND UNEXPLAINED VARIATION

14.7 Prove that $\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2$.

SOLUTION

Squaring both sides of $Y - \bar{Y} = (Y - Y_{\text{est}}) + (Y_{\text{est}} - \bar{Y})$ and then summing, we have

$$\sum (Y - \bar{Y})^2 = \sum (Y - Y_{\text{est}})^2 + \sum (Y_{\text{est}} - \bar{Y})^2 + 2 \sum (Y - Y_{\text{est}})(Y_{\text{est}} - \bar{Y})$$

The required result follows at once if we can show that the last sum is zero; in the case of linear regression, this is so since

$$\begin{aligned}\sum (Y - Y_{\text{est}})(Y_{\text{est}} - \bar{Y}) &= \sum (Y - a_0 - a_1X)(a_0 + a_1X - \bar{Y}) \\ &= a_0 \sum (Y - a_0 - a_1X) + a_1 \sum X(Y - a_0 - a_1X) - \bar{Y} \sum (Y - a_0 - a_1X) = 0\end{aligned}$$

because of the normal equations $\sum (Y - a_0 - a_1X) = 0$ and $\sum X(Y - a_0 - a_1X) = 0$.

The result can similarly be shown valid for nonlinear regression by using a least-squares curve given by $Y_{\text{est}} = a_0 + a_1X + a_2X^2 + \cdots + a_nX^n$.

14.8 Compute (*a*) the total variation, (*b*) the unexplained variation, and (*c*) the explained variation for the data in Problem 14.1.

SOLUTION

The least-squares regression line is $Y_{\text{est}} = 35.8 + 0.476X$. From Table 14.6, we see that the total variation $= \sum (Y - \bar{Y})^2 = 38.917$, the unexplained variation $= \sum (Y - Y_{\text{est}})^2 = 19.703$, and the explained variation $= \sum (Y_{\text{est}} - \bar{Y})^2 = 19.214$.

Table 14.6

Y	Y_{est}	$(Y - \bar{Y})^2$	$(Y - Y_{\text{est}})^2$	$(Y_{\text{est}} - \bar{Y})^2$
68	66.7894	0.1739	1.46562	0.62985
66	65.8366	2.5059	0.02669	3.04986
68	67.7421	0.1739	0.06650	0.02532
65	66.3130	6.6719	1.72395	1.61292
69	68.2185	2.0079	0.61074	0.40387
66	65.3602	2.5059	0.40930	4.94068
68	69.1713	0.1739	1.37185	2.52257
65	67.2657	6.6719	5.13361	0.10065
71	68.2185	11.6759	7.73672	0.40387
67	67.7421	0.3399	0.55075	0.02532
68	68.6949	0.1739	0.48286	1.23628
70	69.6476	5.8419	0.12416	4.26273
$\bar{Y} = 67.5833$		Sum = 38.917	Sum = 19.703	Sum = 19.214

The following output from MINITAB gives these same sums of squares. They are shown in bold. Note the tremendous amount of computation that the software saves the user.

```
MTB > Regress 'Y' 1 'X';
SUBC> Constant;
SUBC> Brief 1.
```

Regression Analysis

```
The regression equation is
Y = 35.8 + 0.476 X

Analysis of Variance

Source            DF       SS        MS        F        P
Regression         1   19.214    19.214     9.75    0.011
Residual Error    10   19.703     1.970
Total             11   38.917
```

E-next
THE NEXT LEVEL OF EDUCATION

COEFFICIENT OF CORRELATION

14.9 Use the results of Problem 14.8 to find (*a*) the coefficient of determination and (*b*) the coefficient of correlation.

SOLUTION

(*a*) Coefficient of determination $= r^2 = \dfrac{\text{explained variation}}{\text{total variation}} = \dfrac{19.214}{38.917} = 0.4937$

(*b*) Coefficient of correlation $= r = \pm\sqrt{0.4937} = \pm 0.7027$

Since X and Y are directly related, we choose the plus sign and have two decimal places $r = 0.70$.

14.10 Prove that for linear regression the coefficient of correlation between the variables X and Y can be written

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

where $x = X - \bar{X}$ and $y = Y - \bar{Y}$.

SOLUTION

The least-squares regression line of Y on X can be written $Y_{\text{est}} = a_0 + a_1 X$ or $y_{\text{est}} = a_1 x$, where [see Problem 13.15(*a*)]

$$a_1 = \frac{\sum xy}{\sum x^2} \quad \text{and} \quad y_{\text{est}} = Y_{\text{est}} - \bar{Y}$$

Then
$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{\sum (Y_{\text{est}} - \bar{Y})^2}{\sum (Y - \bar{Y})^2} = \frac{\sum y_{\text{est}}^2}{\sum y^2}$$

$$= \frac{\sum a_1^2 x^2}{\sum y^2} = \frac{a_1^2 \sum x^2}{\sum y^2} = \left(\frac{\sum xy}{\sum x^2}\right)^2 \frac{\sum x^2}{\sum y^2} = \frac{(\sum xy)^2}{(\sum x^2)(\sum y^2)}$$

and
$$r = \pm \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

However, since the quantity
$$\frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

is positive when y_{est} increases as x increases (i.e., positive linear correlation) and negative when y_{est} decreases as x increases (i.e., negative linear correlation), it *automatically* has the correct sign associated with it. Hence we define the coefficient of linear correlation to be

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

This is often called the *product-moment formula* for the linear correlation coefficient.

E-next
THE NEXT LEVEL OF EDUCATION

PRODUCT-MOMENT FORMULA FOR THE LINEAR CORRELATION COEFFICIENT

14.11 Find the coefficient of linear correlation between the variables X and Y presented in Table 14.7.

Table 14.7

X	1	3	4	6	8	9	11	14
Y	1	2	4	4	5	7	8	9

SOLUTION

The work involved in the computation can be organized as in Table 14.8.

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{84}{\sqrt{(132)(56)}} = 0.977$$

This shows that there is a very high linear correlation between the variables, as we have already observed in Problems 13.8 and 13.12.

Table 14.8

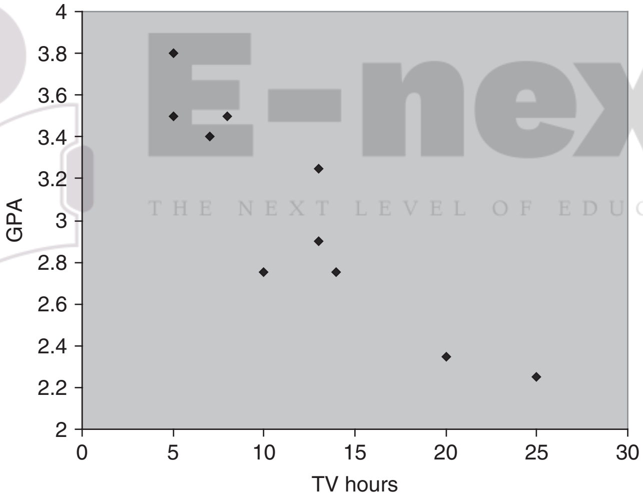
X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
1	1	−6	−4	36	24	16
3	2	−4	−3	16	12	9
4	4	−3	−1	9	3	1
6	4	−1	−1	1	1	1
8	5	1	0	1	0	0
9	7	2	2	4	4	4
11	8	4	3	16	12	9
14	9	7	4	49	28	16
$\sum X = 56$ $\bar{X} = 56/8 = 7$	$\sum Y = 40$ $\bar{Y} = 40/8 = 5$			$\sum x^2 = 132$	$\sum xy = 84$	$\sum y^2 = 56$

https://E-next.in

14.12 In order to investigate the connection between grade point average (GPA) and hours of TV watched per week, the pairs of data in Table 14.9 and Fig. 14-4 were collected and EXCEL was used to make a scatter plot of the data. The data was collected on 10 high school students and X is the number of hours the subject spends watching TV per week (TV hours) and Y is their GPA:

Table 14.9

TV hours	GPA
20	2.35
5	3.8
8	3.5
10	2.75
13	3.25
7	3.4
13	2.9
5	3.5
25	2.25
14	2.75



E-next
THE NEXT LEVEL OF EDUCATION

Fig. 14-4 EXCEL scatter plot of data in Problem 14.12.

Use EXCEL to compute the correlation coefficient of the two variables and verify it by using the product-moment formula.

SOLUTION

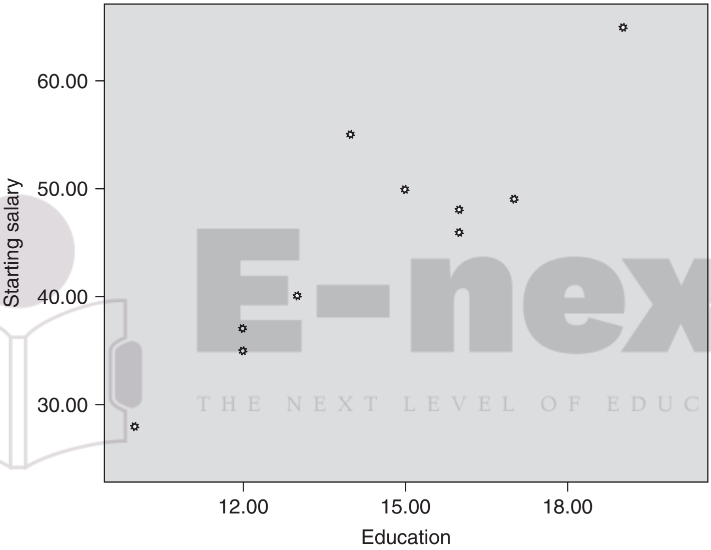
The EXCEL function `=CORREL(E2:E11,F2:F11)` is used to find the correlation coefficient, if the TV hours are in `E2:E11` and the GPA values are in `F2:F11`. The correlation coefficient is -0.9097. The negative value means that the two variables are negatively correlated. That is, as more hours of TV are watched, the lower is the GPA.

14.13 A study recorded the starting salary (in thousands), Y, and years of education, X, for 10 workers. The data and a SPSS scatter plot are given in Table 14.10 and Fig. 14-5.

https://E-next.in

Table 14.10

Starting salary	Years of Education
35	12
46	16
48	16
50	15
40	13
65	19
28	10
37	12
49	17
55	14



E-next
THE NEXT LEVEL OF EDUCATION

Fig. 14-5 SPSS scatter plot for Problem 14.13.

Use SPSS to compute the correlation coefficient of the two variables and verify it by using the product-moment formula.

SOLUTION

Correlations

		startsal	education
startsal	Pearson Correlation	1	.891**
	Sig. (2-tailed)		.001
	N	10	10
education	Pearson Correlation	.891**	1
	Sig. (2-tailed)	.001	
	N	10	10

**Correlation is significant at the 0.001 level (2-tailed).

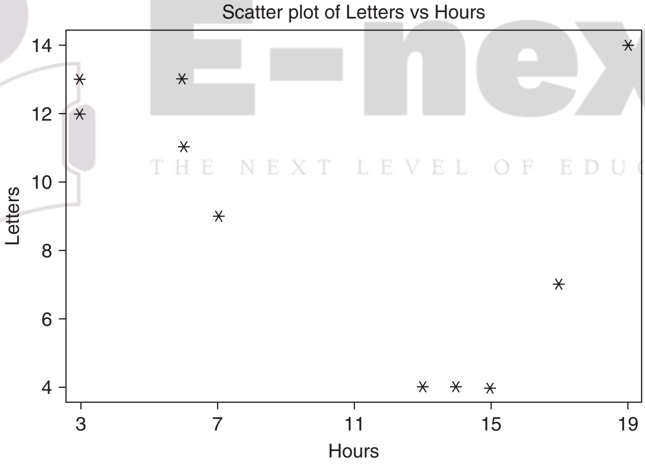
The SPSS pull-down **Analyze** → **Correlate** → **Bivariate** gives the correlation by the product-moment formula. It is also called the Pearson correlation.

The output above gives the coefficient of correlation, $r = 0.891$.

14.14 A study recorded the hours per week on a cell phone, Y, and letters in the last name, X, for 10 students. The data and a STATISTIX scatter plot are given in Table 14.11 and Fig. 14-6.

Table 14.11

Hours on Cell Phone	Letters in Last Name
6	13
6	11
3	12
17	7
19	14
14	4
15	4
3	13
13	4
7	9



E-next
THE NEXT LEVEL OF EDUCATION

Fig. 14-6 STATISTIX scatter plot of data in Table 14.11.

Use STATISTIX to compute the correlation coefficient of the two variables and verify it by using the product-moment formula.

SOLUTION

The pull-down "**Statistics** → **Linear models** → **correlations(Pearson)**" gives the following output:

```
Statistix 8.0
Correlations (Pearson)
                     Hours

Letters            -0.4701
P-VALUE             0.1704
```

https://E-next.in

The coefficient of correlation is $r = -0.4701$. There is no significant correlation between these two variables.

14.15 Show that the linear correlation coefficient is given by

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

SOLUTION

Writing $x = X - \bar{X}$ and $y = Y - \bar{Y}$ in the result of Problem 14.10, we have

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sqrt{[\sum(X-\bar{X})^2][\sum(Y-\bar{Y})^2]}} \tag{34}$$

But
$$\begin{aligned}\sum(X-\bar{X})(Y-\bar{Y}) &= \sum(XY - \bar{X}Y - X\bar{Y} + \bar{X}\bar{Y}) = \sum XY - \bar{X}\sum Y - \bar{Y}\sum X + N\bar{X}\bar{Y}\\ &= \sum XY - N\bar{X}\bar{Y} - N\bar{Y}\bar{X} + N\bar{X}\bar{Y} = \sum XY - N\bar{X}\bar{Y}\\ &= \sum XY - \frac{(\sum X)(\sum Y)}{N}\end{aligned}$$

since $\bar{X} = (\sum X)/N$ and $\bar{Y} = (\sum Y)/N$. Similarly,

$$\begin{aligned}\sum(X-\bar{X})^2 &= \sum(X^2 - 2X\bar{X} + \bar{X}^2) = \sum X^2 - 2\bar{X}\sum X + N\bar{X}^2\\ &= \sum X^2 - \frac{2(\sum X)^2}{N} + \frac{(\sum X)^2}{N} = \sum X^2 - \frac{(\sum X)^2}{N}\end{aligned}$$

and
$$\sum(Y-\bar{Y})^2 = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

Thus equation (34) becomes

$$r = \frac{\sum XY - (\sum X)(\sum Y)/N}{\sqrt{[\sum X^2 - (\sum X)^2/N][\sum Y^2 - (\sum Y)^2/N]}} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

14.16 The relationship between being overweight and level of high blood pressure was researched in obese adults. Table 14.12 gives the number of pounds overweight and the number of units over 80 of the diastolic blood pressure. A SAS scatter plot is given in Fig. 14-7.

Table 14.12

Pounds over Weight	Units over 80
75	15
86	13
88	10
125	27
75	20
30	5
47	8
150	31
114	28
68	22

E-next THE NEXT LEVEL OF EDUCATION

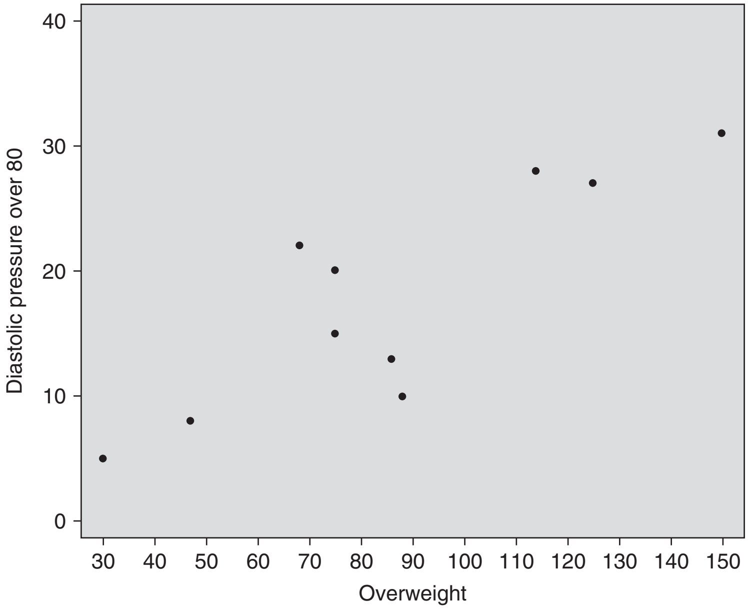


Fig. 14-7 SAS scatter plot for Problem 14.16.

Use SAS to compute the correlation coefficient of the two variables and verify it by using the product-moment formula.

SOLUTION

The SAS pull-down **Statistics** → **Descriptive** → **Correlations** gives the correlation procedure, part of which is shown.

```
                    The CORR Procedure
                         Overwt                    Over80

Overwt                   1.00000                   0.85536
Overwt                                             0.0016
Over80                   0.85536                   1.00000
Over80                   0.0016
```

The output gives the coefficient of correlation as 0.85536. There is a significant correlation between how much overweight the individual is and how far over 80 their diastolic blood pressure is.

E-next
THE NEXT LEVEL OF EDUCATION

CORRELATION COEFFICIENT FOR GROUPED DATA

14.17 Table 14.13 shows the frequency distributions of the final grades of 100 students in mathematics and physics. Referring to this table, determine:

(*a*) The number of students who received grades of 70–79 in mathematics and 80–89 in physics.

(*b*) The percentage of students with mathematics grades below 70.

(*c*) The number of students who received a grade of 70 or more in physics and of less than 80 in mathematics.

(*d*) The percentage of students who passed at least one of the subjects; assume that the minimum passing grade is 60.

https://E-next.in

SOLUTION

(*a*) In Table 14.13, proceed down the column headed 70–79 (mathematics grade) to the row marked 80–89 (physics grade), where the entry is 4, which is the required number of students.

Table 14.13

		Mathematics Grades						
		40–49	50–59	60–69	70–79	80–89	90–99	Total
	90–99				2	4	4	10
Physics Grades	80–89			1	4	6	5	16
	70–79			5	10	8	1	24
	60–69	1	4	9	5	2		21
	50–59	3	6	6	2			17
	40–49	3	5	4				12
	Total	7	15	25	23	20	10	100

(*b*) The total number of students with mathematics grades below 70 is the number with grades 40–49 + the number with grades 50–59 + the number with grades 60–69 = 7 + 15 + 25 = 47. Thus the required percentage of students is 47/100 = 47%.

(*c*) The required number of students is the total of the entries in Table 14.14 (which represents part of Table 14.13). Thus the required number of students is 1 + 5 + 2 + 4 + 10 = 22.

(*d*) Table 14.15 (taken from Table 14.13) shows that the number of students with grades below 60 in both mathematics and physics is 3 + 3 + 6 + 5 = 17. Thus the number of students with grades 60 or over in either physics or mathematics or in both is 100 − 17 = 83, and the required percentage is 83/100 = 83%.

Table 14.14

		Mathematics Grades	
		60–69	70–79
Physics Grades	90–99		2
	80–89	1	4
	70–79	5	10

Table 14.15

		Mathematics Grades	
		40–49	50–59
Physics Grades	50–59	3	6
	40–49	3	5

Table 14.13 is sometimes called a *bivariate frequency table*, or *bivariate frequency distribution*. Each square in the table is called a *cell* and corresponds to a pair of classes or class intervals. The number indicated in the cell is called the *cell frequency*. For example, in part (*a*) the number 4 is the frequency of the cell corresponding to the pair of class intervals 70–79 in mathematics and 80–89 in physics.

The totals indicated in the last row and last column are called *marginal totals*, or *marginal frequencies*. They correspond, respectively, to the class frequencies of the separate frequency distributions of the mathematics and physics grades.

E-next
THE NEXT LEVEL OF EDUCATION

14.18 Show how to modify the formula of Problem 14.15 for the case of data grouped as in the bivariate frequency table (Table 14.13).

SOLUTION

For grouped data, we can consider the various values of the variables X and Y as coinciding with the class marks, while f_X and f_Y are the corresponding class frequencies, or marginal frequencies, shown in the last row and column of the bivariate frequency table. If we let f represent the various cell frequencies corresponding to the pairs of class marks (X, Y), then we can replace the formula of Problem 14.15 with

$$r = \frac{N\sum fXY - (\sum f_X X)(\sum f_Y Y)}{\sqrt{[N\sum f_X X^2 - (\sum f_X X)^2][N\sum f_Y Y^2 - (\sum f_Y Y)^2]}} \tag{35}$$

If we let $X = A + c_X u_X$ and $Y = B + c_Y u_Y$, where c_X and c_Y are the class-interval widths (assumed constant) and A and B are arbitrary class marks corresponding to the variables, formula (*35*) becomes formula (*21*) of this chapter:

$$r = \frac{N\sum fu_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}} \tag{21}$$

This is the *coding method* used in previous chapters as a short method for computing means, standard deviations, and higher moments.

14.19 Find the coefficient of linear correlation of the mathematics and physics grades of Problem 14.17.

SOLUTION

We use formula (*21*). The work can be arranged as in Table 14.16, which is called a *correlation table*. The sums $\sum f_X$, $\sum f_X u_X$, $\sum f_X u_X^2$, $\sum f_Y$, $\sum f_Y u_Y$, and $\sum f_Y u_Y^2$ are obtained by using the coding method, as in earlier chapters.

The number in the corner of each cell in Table 14.16 represents the product $fu_X u_Y$, where f is the cell frequency. The sum of these corner numbers in each row is indicated in the corresponding row of the last column. The sum of these corner numbers in each column is indicated in the corresponding column of the last row. The final totals of the last row and last column are equal and represent $\sum fu_X u_Y$.

From Table 14.16 we have

$$r = \frac{N\sum fu_X u_Y - (\sum f_X u_X)(\sum f_Y u_Y)}{\sqrt{[N\sum f_X u_X^2 - (\sum f_X u_X)^2][N\sum f_Y u_Y^2 - (\sum f_Y u_Y)^2]}}$$

$$= \frac{(100)(125) - (64)(-55)}{\sqrt{[(100(236) - (64)^2][(100)(253) - (-55)^2]}} = \frac{16{,}020}{\sqrt{(19{,}504)(22{,}275)}} = 0.7686$$

14.20 Use Table 14.16 to compute (*a*) s_X, (*b*) s_Y, and (*c*) s_{XY} and thus to verify the formula $r = s_{XY}/(s_X s_Y)$.

SOLUTION

(*a*) $$s_X = c_X\sqrt{\frac{\sum f_X u_X^2}{N} - \left(\frac{\sum f_X u_X}{N}\right)^2} = 10\sqrt{\frac{236}{100} - \left(\frac{64}{100}\right)^2} = 13.966$$

(*b*) $$s_Y = c_Y\sqrt{\frac{\sum f_Y u_Y^2}{N} - \left(\frac{\sum f_Y u_Y}{N}\right)^2} = 10\sqrt{\frac{253}{100} - \left(\frac{-55}{100}\right)^2} = 14.925$$

(*c*) $$s_{XY} = c_X c_Y\left[\frac{\sum fu_X u_Y}{N} - \left(\frac{\sum f_X u_X}{N}\right)\left(\frac{\sum f_Y u_Y}{N}\right)\right] = (10)(10)\left[\frac{125}{100} - \left(\frac{64}{100}\right)\left(\frac{-55}{100}\right)\right] = 160.20$$

Thus the standard deviations of the mathematics and physics grades are 14.0 and 14.9, respectively, while their covariance is 160.2. The correlation coefficient r is therefore

E-next THE NEXT LEVEL OF EDUCATION

Table 14.16

			Mathematics Grades X									
		X	44.5	54.5	64.5	74.5	84.5	94.5				
	Y	u_X u_Y	-2	-1	0	1	2	3	f_Y	$f_Y u_Y$	$f_Y u_y^2$	Sum of corner numbers in each row
Physics Grades Y	94.5	2				2 $\boxed{4}$	4 $\boxed{16}$	4 $\boxed{24}$	10	20	40	44
	84.5	1			1 $\boxed{0}$	4 $\boxed{4}$	6 $\boxed{12}$	5 $\boxed{15}$	16	16	16	31
	74.5	0			5 $\boxed{0}$	10 $\boxed{0}$	8 $\boxed{0}$	1 $\boxed{0}$	24	0	0	0
	64.5	-1	1 $\boxed{2}$	4 $\boxed{4}$	9 $\boxed{0}$	5 $\boxed{-5}$	2 $\boxed{-4}$		21	-21	21	-3
	54.5	-2	3 $\boxed{12}$	6 $\boxed{12}$	6 $\boxed{0}$	2 $\boxed{-4}$			17	-34	68	20
	44.5	-3	3 $\boxed{18}$	5 $\boxed{15}$	4 $\boxed{0}$				12	-36	108	33
f_X			7	15	25	23	20	10	$\sum f_X = \sum f_Y$ $= N = 100$	$\sum f_Y u_Y$ $= -55$	$\sum f_Y u_Y^2$ $= 253$	$\sum f u_X u_Y$ $= 125$
$f_X u_X$			-14	-15	0	23	40	30	$\sum f_X u_X$ $= 64$			
$f_X u_X^2$			28	15	0	23	80	90	$\sum f_X u_X^2$ $= 236$	Check		
Sum of corner numbers in each column			32	31	0	-1	24	39	$\sum f u_X u_Y$ $= 125$			

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{160.20}{(13.966)(14.925)} = 0.7686$$

agreeing with Problem 14.19.

REGRESSION LINES AND THE CORRELATION COEFFICIENT

14.21 Prove that the regression lines of Y on X and of X on Y have equations given, respectively, by (*a*) $Y - \bar{Y} = (rs_Y/s_X)(X - \bar{Y})$ and (*b*) $X - \bar{X} = (rs_X/s_Y)(Y - \bar{Y})$.

SOLUTION

(*a*) From Problem 13.15(*a*), the regression line of Y on X has the equation

$$y = \left(\frac{\sum xy}{\sum x^2}\right)x \qquad \text{or} \qquad Y - \bar{Y} = \left(\frac{\sum xy}{\sum x^2}\right)(X - \bar{X})$$

Then, since $$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} \qquad \text{(see Problem 14.10)}$$

E-next THE NEXT LEVEL OF EDUCATION

we have $$\frac{\sum xy}{\sum x^2} = \frac{r\sqrt{(\sum x^2)(\sum y^2)}}{\sum x^2} = \frac{r\sqrt{\sum y^2}}{\sqrt{\sum x^2}} = \frac{rs_Y}{s_X}$$

and the required result follows.

(*b*) This follows by interchanging X and Y in part (*a*).

14.22 If, the regression lines of Y on X and of X on Y are given, respectively, by $Y = a_0 + a_1X$ and $X = b_0 + b_1Y$, prove that $a_1b_1 = r^2$.

SOLUTION

From Problem 14.21, parts (*a*) and (*b*),

$$a_1 = \frac{rs_Y}{s_X} \quad \text{and} \quad b_1 = \frac{rs_X}{s_Y}$$

Thus $$a_1b_1 = \left(\frac{rs_Y}{s_X}\right)\left(\frac{rs_X}{s_Y}\right) = r^2$$

This result can be taken as the starting point for a definition of the linear correlation coefficient.

14.23 Use the result of Problem 14.22 to find the linear correlation coefficient for the data of Problem 14.1.

SOLUTION

E-next
THE NEXT LEVEL OF EDUCATION

From Problem 14.1 [parts (*b*) and (*c*), respectively] $a_1 = 484/1016 = 0.476$ and $b_1 = 484/467 = 1.036$. Thus $Y^2 = a_1b_1 = (384/1016)(484/467)$ and $r = 0.7027$.

14.24 For the data of Problem 14.19, write the equations of the regression lines of (*a*) Y on X and (*b*) X on Y.

SOLUTION

From the correlation table (Table 14.16) of Problem 14.19 we have

$$\bar{X} = A + c_X \frac{\sum f_X u_X}{N} = 64.5 + \frac{(10)(64)}{100} = 70.9$$
$$\bar{Y} = B + c_Y \frac{\sum f_Y u_Y}{N} = 74.5 + \frac{(10)(-55)}{100} = 69.0$$

From the results of Problem 14.20, $s_X = 13.966$, $s_Y = 14.925$, and $r = 0.7686$. We now use Problem 14.21, parts (*a*) and (*b*), to obtain the equations of the regression lines.

(*a*) $$Y - \bar{Y} = \frac{rs_Y}{s_X}(X - \bar{X}) \qquad Y - 69.0 = \frac{(0.7686)(14.925)}{13.966}(X - 70.9) = 0.821(X - 70.9)$$

(*b*) $$X - \bar{X} = \frac{rs_X}{s_Y}(Y - \bar{Y}) \qquad X - 70.9 = \frac{(0.7686)(13.966)}{14.925}(Y - 69.0) = 0.719(Y - 69.0)$$

14.25 For the data of Problem 14.19, compute the standard errors of estimate (*a*) $s_{Y.X}$ and (*b*) $s_{X.Y}$. Use the results of Problem 14.20.

SOLUTION

(*a*) $$s_{Y.X} = s_Y\sqrt{1 - r^2} = 14.925\sqrt{1 - (0.7686)^2} = 9.548$$

(*b*) $$s_{X.Y} = s_X\sqrt{1 - r^2} = 13.966\sqrt{1 - (0.7686)^2} = 8.934$$

14.26 Table 14.17 shows the United States consumer price indexes for food and medical-care costs during the years 2000 through 2006 compared with prices in the base years, 1982 to 1984 (mean taken as 100). Compute the correlation coefficient between the two indexes and give the MINITAB computation of the coefficient.

Table 14.17

Year	2000	2001	2002	2003	2004	2005	2006
Food	167.8	173.1	176.2	180.0	186.2	190.7	195.2
Medical	260.8	272.8	285.6	297.1	310.1	323.2	336.2

Source: Bureau of Labor Statistics.

SOLUTION

Denoting the index numbers for food and medical care as X and Y, respectively, the calculation of the correlation coefficient can be organized as in Table 14.18. (Note that the year is used only to specify the corresponding values of X and Y.)

Table 14.18

X	Y	$x = X - \bar{X}$	$y = Y - \bar{Y}$	x^2	xy	y^2
167.8	260.8	−13.5	−37.2	182.25	502.20	1383.84
173.1	272.8	− 8.2	−25.2	67.24	206.64	635.04
176.2	285.6	− 5.1	−12.4	26.01	63.24	153.76
180.0	297.1	− 1.3	− 0.9	1.69	1.17	0.81
186.2	310.1	4.9	12.1	24.01	59.29	46.41
190.7	323.2	9.4	25.2	88.36	236.88	635.04
195.2	336.2	13.9	38.2	193.21	530.98	1459.24
$\bar{X} = 181.3$	$\bar{Y} = 298.0$			Sum = 582.77	Sum = 1600.4	Sum = 4414.14

E-next
THE NEXT LEVEL OF EDUCATION

Then by the product-moment formula

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}} = \frac{1600.4}{\sqrt{(582.77)(4414.14)}} = 0.998$$

After putting the X values in `C1` and the Y values in `C2`, the MINITAB command **`correlation C1 C2`** produces the correlation coefficient which is the same that we computed.

```
Correlations: X, Y

Pearson correlation of X and Y =0.998
P-Value=0.000
```

NONLINEAR CORRELATION

14.27 Fit a least-squares parabola of the form $Y = a_0 + a_1X + a_2X^2$ to the set of data in Table 14.19. Also give the MINITAB solution.

SOLUTION

The normal equations (*23*) of Chapter 13 are

$$\begin{aligned} \sum Y &= a_0 N + a_1 \sum X + a_2 \sum X^2 \\ \sum XY &= a_0 \sum X + a_1 \sum X^2 + a_2 \sum X^3 \\ \sum X^2 Y &= a_0 \sum X^2 + a_1 \sum X^3 + a_2 \sum X^4 \end{aligned} \tag{36}$$

Table 14.19

X	1.2	1.8	3.1	4.9	5.7	7.1	8.6	9.8
Y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

The work involved in computing the sums can be arranged as in Table 14.20. Then, since $N = 8$, the normal equations (*36*) become

$$\begin{aligned} 8a_0 + 42.2a_1 + 291.20a_2 &= 46.4 \\ 42.2a_0 + 291.20a_1 + 2275.35a_2 &= 230.42 \\ 291.20a_0 + 2275.35a_1 + 18971.92a_2 &= 1449.00 \end{aligned} \tag{37}$$

Solving, $a_0 = 2.588$, $a_1 = 2.065$, and $a_2 = -0.2110$; hence the required least-squares parabola has the equation

$$Y = 2.588 + 2.065X - 0.2110X^2$$

E-next THE NEXT LEVEL OF EDUCATION

Table 14.20

X	Y	X^2	X^3	X^4	XY	X^2Y
1.2	4.5	1.44	1.73	2.08	5.40	6.48
1.8	5.9	3.24	5.83	10.49	10.62	19.12
3.1	7.0	9.61	29.79	92.35	21.70	67.27
4.9	7.8	24.01	117.65	576.48	38.22	187.28
5.7	7.2	32.49	185.19	1055.58	41.04	233.93
7.1	6.8	50.41	357.91	2541.16	48.28	342.79
8.6	4.5	73.96	636.06	5470.12	38.70	332.82
9.8	2.7	96.04	941.19	9223.66	26.46	259.31
$\sum X$ $= 42.2$	$\sum Y$ $= 46.4$	$\sum X^2$ $= 291.20$	$\sum X^3$ $= 2275.35$	$\sum X^4$ $= 18{,}971.92$	$\sum XY$ $= 230.42$	$\sum X^2Y$ $= 1449.00$

The values for Y are entered into `C1`, the Values for X are entered into `C2`, and the values for X^2 are entered into `C3`. The MINITAB pull-down **Stat → Regression → Regression** is given. The least-squares parabola is given as part of the output as follows.

```
The regression equation is Y = 2.59 + 2.06 X - 0.211 Xsquared
```

This is the same solution as given by solving the normal equations.

14.28 Use the least-squares parabola of Problem 14.27 to estimate the values of Y from the given values of X.

SOLUTION

For $X = 1.2$, $Y_{\text{est}} = 2.588 + 2.065(1.2) - 0.2110(1.2)^2 = 4.762$. Other estimated values are obtained similarly. The results are shown in Table 14.21 together with the actual values of Y.

Table 14.21

Y_{est}	4.762	5.621	6.962	7.640	7.503	6.613	4.741	2.561
Y	4.5	5.9	7.0	7.8	7.2	6.8	4.5	2.7

14.29 (*a*) Find the linear correlation coefficient between the variables X and Y of Problem 14.27.

(*b*) Find the nonlinear correlation coefficient between these variables, assuming the parabolic relationship obtained in Problem 14.27.

(*c*) Explain the difference between the correlation coefficients obtained in parts (*a*) and (*b*).

(*d*) What percentage of the total variation remains unexplained by assuming a parabolic relationship between X and Y?

SOLUTION

(*a*) Using the calculations already obtained in Table 14.20 and the added fact that $\sum Y^2 = 290.52$, we find that

$$r = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$
$$= \frac{(8)(230.42) - (42.2)(46.4)}{\sqrt{[(8)(291.20) - (42.2)^2][(8)(290.52) - (46.4)^2]}} = -0.3743$$

(*b*) From Table 14.20, $\bar{Y} = (\sum Y)/N = 46.4/8 = 5.80$; thus the total variation is $\sum(Y - \bar{Y})^2 = 21.40$. From Table 14.21, the explained variation is $\sum(Y_{\text{est}} - \bar{Y})^2 = 21.02$. Thus

$$r^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{21.02}{21.40} = 0.9822 \qquad \text{and} \qquad r = 0.9911 \quad \text{or} \quad 0.99$$

(*c*) The fact that part (*a*) shows a linear correlation coefficient of only -0.3743 indicates that there is practically no *linear relationship* between X and Y. However, there is a very good *nonlinear relationship* supplied by the parabola of Problem 14.27, as indicated by the fact that the correlation coefficient in part (*b*) is 0.99.

(*d*)
$$\frac{\text{Unexplained variation}}{\text{Total variation}} = 1 - r^2 = 1 - 0.9822 = 0.0178$$

Thus 1.78% of the total variation remains unexplained. This could be due to random fluctuations or to an additional variable that has not been considered.

14.30 Find (*a*) s_Y and (*b*) $s_{Y.X}$ for the data of Problem 14.27.

SOLUTION

(*a*) From Problem 14.29(*a*), $\sum(Y - \bar{Y})^2 = 21.40$. Thus the standard deviation of Y is

$$s_Y = \sqrt{\frac{\sum(Y - \bar{Y})^2}{N}} = \sqrt{\frac{21.40}{8}} = 1.636 \quad \text{or} \quad 1.64$$

(*b*) **First method**

Using part (*a*) and Problem 14.29(*b*), the standard error of estimate of Y on X is

$$s_{Y.X} = s_Y\sqrt{1-r^2} = 1.636\sqrt{1-(0.9911)^2} = 0.218 \quad \text{or} \quad 0.22$$

Second method

Using Problem 14.29,

$$s_{Y.X} = \sqrt{\frac{\sum(Y-Y_{\text{est}})^2}{N}} = \sqrt{\frac{\text{unexplained variation}}{N}} = \sqrt{\frac{21.40-21.02}{8}} = 0.218 \quad \text{or} \quad 0.22$$

Third method

Using Problem 14.27 and the additional calculation $\sum Y^2 = 290.52$, we have

$$s_{Y.X} = \sqrt{\frac{\sum Y^2 - a_0\sum Y - a_1\sum XY - a_2\sum X^2Y}{N}} = 0.218 \qquad \text{or} \qquad 0.22$$

SAMPLING THEORY OF CORRELATION

14.31 A correlation coefficient based on a sample of size 18 was computed to be 0.32. Can we conclude at significance levels of (*a*) 0.05 and (*b*) 0.01 that the corresponding population correlation coefficient differs from zero?

SOLUTION

We wish to decide between the hypotheses $H_0 : \rho = 0$ and $H_1 : \rho > 0$.

$$t = \frac{r\sqrt{N-2}}{\sqrt{1-r^2}} = \frac{0.32\sqrt{18-2}}{\sqrt{1-(0.32)^2}} = 1.35$$

E-next THE NEXT LEVEL OF EDUCATION

(*a*) Using a one-tailed test of Student's distribution at the 0.05 level, we would reject H_0 if $t > t_{.95} = 1.75$ for $(18-2) = 16$ degrees of freedom. Thus we cannot reject H_0 at the 0.05 level.

(*b*) Since we cannot reject H_0 at the 0.05 level, we certainly cannot reject it at the 0.01 level.

14.32 What is the minimum sample size necessary in order that we may conclude that a correlation coefficient of 0.32 differs significantly from zero at the 0.05 level?

SOLUTION

Using a one-tailed test of Student's distribution at the 0.05 level, the minimum value of N must be such that

$$\frac{0.32\sqrt{N-2}}{\sqrt{1-(0.32)^2}} = t_{.95}$$

for $N-2$ degrees of freedom. For an infinite number of degrees of freedom, $t_{.95} = 1.64$ and hence $N = 25.6$.

$$\begin{array}{llll}
\text{For } N = 26: & \nu = 24 & t_{.95} = 1.71 & t = 0.32\sqrt{24}/\sqrt{1-(0.32)^2} = 1.65 \\
\text{For } N = 27: & \nu = 25 & t_{.95} = 1.71 & t = 0.32\sqrt{25}/\sqrt{1-(0.32)^2} = 1.69 \\
\text{For } N = 28: & \nu = 26 & t_{.95} = 1.71 & t = 0.32\sqrt{26}/\sqrt{1-(0.32)^2} = 1.72
\end{array}$$

Thus the minimum sample size is $N = 28$.

14.33 A correlation coefficient on a sample of size 24 was computed to be $r = 0.75$. At the 0.05 significance level, can we reject the hypothesis that the population correlation coefficient is as small as (*a*) $\rho = 0.60$ and (*b*) $\rho = 0.50$?

SOLUTION

(*a*)
$$Z = 1.1513 \log\left(\frac{1+0.75}{1-0.75}\right) = 0.9730 \qquad \mu_Z = 1.1513 \log\left(\frac{1+0.60}{1-0.60}\right) = 0.6932$$

and
$$\sigma_Z = \frac{1}{\sqrt{N-3}} = \frac{1}{\sqrt{21}} = 0.2182$$

Thus
$$z = \frac{Z - \mu_Z}{\sigma_Z} = \frac{0.9730 - 0.6932}{0.2182} = 1.28$$

Using a one-tailed test of the normal distribution at the 0.05 level, we would reject the hypothesis only if z were greater than 1.64. Thus we cannot reject the hypothesis that the population correlation coefficient is as small as 0.60.

(*b*) If $\rho = 0.50$, then $\mu_Z = 1.1513 \log 3 = 0.5493$ and $z = (0.9730 - 0.5493)/0.2182 = 1.94$. Thus we can reject the hypothesis that the population correlation coefficient is as small as $\rho = 0.50$ at the 0.05 level.

14.34 The correlation coefficient between the final grades in physics and mathematics for a group of 21 students was computed to be 0.80. Find the 95% confidence limits for this coefficient.

SOLUTION

Since $r = 0.80$ and $N = 21$, the 95% confidence limits for μ_Z are given by

$$Z \pm 1.96\sigma_Z = 1.1513 \log\left(\frac{1+r}{1-r}\right) \pm 1.96\left(\frac{1}{\sqrt{N-3}}\right) = 1.0986 \pm 0.4620$$

Thus μ_Z has the 95% confidence interval 0.5366 to 1.5606. Now if

$$\mu_Z = 1.1513 \log\left(\frac{1+\rho}{1-\rho}\right) = 0.5366 \qquad \text{then} \qquad \rho = 0.4904$$

and if
$$\mu_Z = 1.1513 \log\left(\frac{1+\rho}{1-\rho}\right) = 1.5606 \qquad \text{then} \qquad \rho = 0.9155$$

Thus the 95% confidence limits for ρ are 0.49 and 0.92.

E-next THE NEXT LEVEL OF EDUCATION

14.35 Two correlation coefficients obtained from samples of size $N_1 = 28$ and $N_3 = 35$ were computed to be $r_1 = 0.50$ and $r_2 = 0.30$, respectively. Is there a significant difference between the two coefficients at the 0.05 level?

SOLUTION

$$Z_1 = 1.1513 \log\left(\frac{1+r_1}{1-r_1}\right) = 0.5493 \qquad Z_2 = 1.1513 \log\left(\frac{1+r_2}{1-r_2}\right) = 0.3095$$

and
$$\sigma_{Z_1 - Z_2} = \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}} = 0.2669$$

We wish to decide between the hypotheses $H_0\colon \mu_{Z_1} = \mu_{Z_2}$ and $H_1\colon \mu_{Z_1} \neq \mu_{Z_2}$. Under hypothesis H_0,

$$z = \frac{Z_1 - Z_2 - (\mu_{Z_1} - \mu_{Z_2})}{\sigma_{Z_1 - Z_2}} = \frac{0.5493 - 0.3095 - 0}{0.2669} = 0.8985$$

Using a two-tailed test of the normal distribution, we would reject H_0 only if $z > 1.96$ or $z < -1.96$. Thus we cannot reject H_0, and we conclude that the results are not significantly different at the 0.05 level.

14.36 In Problem 14.1 we found the regression equation of Y on X to be $Y = 35.82 + 0.476X$. Test the null hypothesis at the 0.05 significance level that the regression coefficient of the population regression equation is 0.180 versus the alternative hypothesis that the regression coefficient exceeds 0.180. Perform the test without the aid of computer software as well as with the aid of MINITAB computer software.

SOLUTION

$$t = \frac{a_1 - A_1}{S_{Y.X}/S_X}\sqrt{N-2} = \frac{0.476 - 0.180}{1.28/2.66}\sqrt{12-2} = 1.95$$

since $S_{Y.X} = 1.28$ (computed in Problem 14.5) and $S_X = \sqrt{(\sum x^2)/N} = \sqrt{84.68/12} = 2.66$. Using a one-tailed test of Student's distribution at the 0.05 level, we would reject the hypothesis that the regression coefficient is 0.180 if $t > t_{.95} = 1.81$ for $(12-2) = 10$ degrees of freedom. Thus, we reject the null hypothesis.

The MINITAB output for this problem is as follows.

```
MTB > Regress 'Y' 1 'X';
SUBC>  Constant;
SUBC>  Predict c7.
```

Regression Analysis

```
The regression equation is
Y = 35.8 + 0.476 X

Predictor        Coef        StDev         T         P
Constant        35.82        10.18      3.52     0.006
X              0.4764       0.1525      3.12     0.011

S = 1.404     R-Sq = 49.4%     R-Sq(adj) = 44.3%

Analysis of Variance

Source            DF          SS          MS        F         P
Regression         1      19.214      19.214     9.75     0.011
Residual Error    10      19.703       1.970
Total             11      38.917

Predicted Values
Fit       StDev Fit        95.0% CI                 95.0% PI
66.789    0.478        (65.724,   67.855)    (63.485,   70.094)
69.171    0.650        (67.723,   70.620)    (65.724,   72.618)
```

The following portion of the output gives the information needed to perform the test of hypothesis.

```
Predictor        Coef        StDev         T         P
Constant        35.82        10.18      3.52     0.006
X              0.4764       0.1525      3.12     0.011
```

The computed test statistic is found as follows:

$$t = \frac{0.4764 - 0.180}{0.1525} = 1.94$$

The computed t value shown in the output, **3.12**, is used for testing the null hypothesis that the regression coefficient is 0. To test any other value for the regression coefficient requires a computation like the

E-next THE NEXT LEVEL OF EDUCATION

one shown. To test that the regression coefficient is 0.25, for example, the computed value of the test statistic would equal

$$t = \frac{0.4764 - 0.25}{0.1525} = 1.48$$

The null hypothesis that the regression coefficient equals 0.25 would not be rejected.

14.37 Find the 95% confidence limits for the regression coefficient of Problem 14.36. Set the confidence interval without the aid of any computer software as well as with the aid of MINITAB computer software.

SOLUTION

The confidence interval may be expressed as

$$a_1 \pm \frac{t}{\sqrt{N-2}}\left(\frac{S_{Y.X}}{S_X}\right)$$

Thus the 95% confidence limits for A_1 (obtained by setting $t = \pm t_{.975} = \pm 2.23$ for $12 - 2 = 10$ degrees of freedom) are given by

$$a_1 \pm \frac{2.23}{\sqrt{12-2}}\left(\frac{S_{Y.X}}{S_X}\right) = 0.476 \pm \frac{2.23}{\sqrt{10}}\left(\frac{1.28}{2.66}\right) = 0.476 \pm 0.340$$

That is, we are 95% confident that A_1 lies between 0.136 and 0.816.

The following portion of the MINITAB output from Problem 14.36 gives the information needed to set the 95% confidence interval.

```
Predictor        Coef         StDev          T          P
Constant        35.82         10.18       3.52      0.006
X              0.4764        0.1525       3.12      0.011
```

The term

$$\frac{1}{\sqrt{N-2}}\left(\frac{S_{Y.X}}{S_X}\right)$$

is sometimes called the standard error associated with the estimated regression coefficient. The value for this standard error is shown in the output as `0.1525`. To find the 95% confidence interval, we multiply this standard error by $t_{.975}$ and then add and subtract this term from $a_1 = 0.476$ to obtain the following confidence interval for A_1:

$$0.476 \pm 2.23(0.1525) = 0.476 \pm 0.340$$

14.38 In Problem 14.1, find the 95% confidence limits for the heights of sons whose fathers' heights are (*a*) 65.0 and (*b*) 70.0 in. Set the confidence interval without the aid of any computer software as well as with the aid of MINITAB computer software.

SOLUTION

Since $t_{.975} = 2.23$ for $(12 - 2) = 10$ degrees of freedom, the 95% confidence limits for Y_P are given by

$$Y_0 \pm \frac{2.23}{\sqrt{N-2}}\, S_{Y.X}\sqrt{N + 1 + \frac{(X_0 - \bar{X})^2}{S_X^2}}$$

where $Y_0 = 35.82 + 0.476X_0$, $S_{Y.X} = 1.28$, $S_X = 2.66$, and $N = 12$.

(*a*) If $X_0 = 65.0$, then $Y_0 = 66.76$ in. Also, $(X_0 - \bar{X})^2 = (65.0 - 66.67)^2 = 2.78$. Thus the 95% confidence limits are

$$66.76 \pm \frac{2.23}{\sqrt{10}}\,(1.28)\sqrt{12 + 1 + \frac{2.78}{2.66^2}} = 66.76 \pm 3.30 \text{ in}$$

That is, we can be 95% confident that the sons' heights are between 63.46 and 70.06 in.

E-next
THE NEXT LEVEL OF EDUCATION

(*b*) If $X_0 = 70.0$, then $Y_0 = 69.14$ in. Also, $(X_0 - \bar{X})^2 = (70.0 - 66.67)^2 = 11.09$. Thus the 95% confidence limits are computed to be 69.14 ± 3.45 in; that is, we can be 95% confident that the sons' heights are between 65.69 and 72.59 in.

The following portion of the MINITAB output found in Problem 14.36 gives the confidence limits for the sons' heights.

```
Predicted Values
Fit       StDev Fit      95.0% CI                95.0% PI
66.789       0.478    (65.724,   67.855)   (63.485,    70.094)
69.171       0.650    (67.723,   70.620)   (65.724,    72.618)
```

The confidence interval for individuals are sometimes referred to as prediction intervals. The 95% prediction intervals are shown in bold. These intervals agree with those computed above except for rounding errors.

14.39 In Problem 14.1, find the 95% confidence limits for the mean heights of sons whose fathers' heights are (*a*) 65.0 in and (*b*) 70.0 in. Set the confidence interval without the aid of any computer software as well as with the aid of MINITAB computer software.

SOLUTION

Since $t_{.975} = 2.23$ for 10 degrees of freedom, the 95% confidence limits for $\bar{Y}_{\mathrm{P}}$ are given by

$$Y_0 \pm \frac{2.23}{\sqrt{10}} S_{Y.X} \sqrt{1 + \frac{(X_0 - \bar{X})^2}{S_X^2}}$$

where $Y_0 = 35.82 + 0.476X_0$, $S_{Y.X} = 1.28$, and $S_X = 2.66$.

(*a*) For $X_0 = 65.0$, we find the confidence limits to be 66.76 ± 1.07 or 65.7 and 67.8.

(*b*) For $X_0 = 70.0$, we find the confidence limits to be 69.14 ± 1.45 or 67.7 and 70.6.

The following portion of the MINITAB output found in Problem 14.36 gives the confidence limits for the mean heights.

```
Predicted Values
Fit       StDev Fit       95.0% CI                 95.0% PI
66.789       0.478    (65.724,    67.855)    (63.485,    70.094)
69.171       0.650    (67.723,    70.620)    (65.724,    72.618)
```

E-next
THE NEXT LEVEL OF EDUCATION

Supplementary Problems

LINEAR REGRESSION AND CORRELATION

14.40 Table 14.22 shows the first two grades (denoted by *X* and *Y*, respectively) of 10 students on two short quizzes in biology.

(*a*) Construct a scatter diagram.

(*b*) Find the least-squares regression line of *Y* on *X*.

(*c*) Find the least-squares regression line of *X* on *Y*.

(*d*) Graph the two regression lines of parts (*b*) and (*c*) on the scatter diagram of part (*a*).

14.41 Find (*a*) $s_{Y.X}$ and (*b*) $s_{X.Y}$ for the data in Table 14.22.

Table 14.22

Grade on first quiz (X)	6	5	8	8	7	6	10	4	9	7
Grade on second quiz (Y)	8	7	7	10	5	8	10	6	8	6

14.42 Compute (a) the total variation in Y, (b) the unexplained variation in Y, and (c) the explained variation in Y for the data of Problem 14.40.

14.43 Use the results of Problem 14.42 to find the correlation coefficient between the two sets of quiz grades of Problem 14.40.

14.44 Find the correlation coefficient between the two sets of quiz grades in Problem 14.40 by using the product-moment formula, and compare this finding with the correlation coefficient given by SPSS, SAS, STATISTIX, MINITAB, and EXCEL.

14.45 Find the covariance for the data of Problem 14.40(a) directly and (b) by using the formula $s_{XY} = rs_X s_Y$ and the result of Problem 14.43 or Problem 14.44.

14.46 Table 14.23 shows the ages X and the systolic blood pressures Y of 12 women.

(a) Find the correlation coefficient between X and Y using the product-moment formula, EXCEL, MINITAB, SAS, SPSS, and STATISTIX.

(b) Determine the least-squares regression equation of Y on X by solving the normal equations, and by using EXCEL, MINITAB, SAS, SPSS, and STATISTIX.

(c) Estimate the blood pressure of a woman whose age is 45 years.

E-next
THE NEXT LEVEL OF EDUCATION

Table 14.23

Age (X)	56	42	72	36	63	47	55	49	38	42	68	60
Blood pressure (Y)	147	125	160	118	149	128	150	145	115	140	152	155

14.47 Find the correlation coefficients for the data of (a) Problem 13.32 and (b) Problem 13.35.

14.48 The correlation coefficient between two variables X and Y is $r = 0.60$. If $s_X = 1.50$, $s_Y = 2.00$, $\bar{X} = 10$, and $\bar{Y} = 20$, find the equations of the regression lines of (a) Y on X and (b) X on Y.

14.49 Compute (a) $s_{Y.X}$ and (b) $s_{X.Y}$ for the data of Problem 14.48.

14.50 If $s_{Y.X} = 3$ and $s_Y = 5$, find r.

14.51 If the correlation coefficient between X and Y is 0.50, what percentage of the total variation remains unexplained by the regression equation?

14.52 (a) Prove that the equation of the regression line of Y on X can be written

$$Y - \bar{Y} = \frac{s_{XY}}{s_X^2}(X - \bar{X})$$

(b) Write the analogous equation for the regression line of X on Y.

14.53 (*a*) Compute the correlation coefficient between the corresponding values of X and Y given in Table 14.24.

Table 14.24

X	2	4	5	6	8	11
Y	18	12	10	8	7	5

(*b*) Multiply each X value in the table by 2 and add 6. Multiply each Y value in the table by 3 and subtract 15. Find the correlation coefficient between the two new sets of values, explaining why you do or do not obtain the same result as in part (*a*).

14.54 (*a*) Find the regression equations of Y on X for the data considered in Problem 14.53, parts (*a*) and (*b*).

(*b*) Discuss the relationship between these regression equations.

14.55 (*a*) Prove that the correlation coefficient between X and Y can be written

$$r = \frac{\overline{XY} - \bar{X}\bar{Y}}{\sqrt{[\overline{X^2} - \bar{X}^2][\overline{Y^2} - \bar{Y}^2]}}$$

(*b*) Using this method, work Problem 14.1.

14.56 Prove that a correlation coefficient is independent of the choice of origin of the variables or the units in which they are expressed. (*Hint:* Assume that $X' = c_1X + A$ and $Y' = c_2Y + B$, where c_1, c_2, A, and B are any constants, and prove that the correlation coefficient between X' and Y' is the same as that between X and Y.)

14.57 (*a*) Prove that, for linear regression,

$$\frac{s^2_{Y.X}}{s^2_Y} = \frac{s^2_{X.Y}}{s^2_X}$$

(*b*) Does the result hold for nonlinear regression?

E-next THE NEXT LEVEL OF EDUCATION

CORRELATION COEFFICIENT FOR GROUPED DATA

14.58 Find the correlation coefficient between the heights and weights of the 300 U.S. adult males given in Table 14.25, a frequency table.

Table 14.25

		Heights X (in)				
		59–62	63–66	67–70	71–74	75–78
Weights Y (lb)	90–109	2	1			
	110–129	7	8	4	2	
	130–149	5	15	22	7	1
	150–169	2	12	63	19	5
	170–189		7	28	32	12
	190–209		2	10	20	7
	210–229			1	4	2

14.59 (*a*) Find the least-squares regression equation of Y on X for the data of Problem 14.58.

(*b*) Estimate the weights of two men whose heights are 64 and 72 in, respectively.

14.60 Find (*a*) $s_{Y.X}$ and (*b*) $s_{X.Y}$ for the data of Problem 14.58.

14.61 Establish formula (*21*) of this chapter for the correlation coefficient of grouped data.

CORRELATION OF TIME SERIES

14.62 Table 14.26 shows the average annual expenditures per consumer unit for health care and the per capita income for the years 1999 through 2004. Find the correlation coefficient.

Table 14.26

Year	1999	2000	2001	2002	2003	2004
Health care cost	1959	2066	2182	2350	2416	2574
Per capita income	27939	29845	30574	30810	31484	33050

Source: Bureau of Labor Statistics and U.S. Bureau of Economic Analysis.

14.63 Table 14.27 shows the average temperature and precipitation in a city for the month of July during the years 2000 through 2006. Find the correlation coefficient.

Table 14.27

Year	2000	2001	2002	2003	2004	2005	2006
Temperature (°F)	78.1	71.8	75.6	72.7	75.3	73.6	75.1
Precipitation (in)	6.23	3.64	3.42	2.84	1.83	2.82	4.04

E-next
THE NEXT LEVEL OF EDUCATION

SAMPLING THEORY OF CORRELATION

14.64 A correlation coefficient based on a sample of size 27 was computed to be 0.40. Can we conclude at significance levels of (*a*) 0.05 and (*b*) 0.01, that the corresponding population correlation coefficient differs from zero?

14.65 A correlation coefficient based on a sample of size 35 was computed to be 0.50. At the 0.05 significance level, can we reject the hypothesis that the population correlation coefficient is (*a*) as small as $\rho = 0.30$ and (*b*) as large as $\rho = 0.70$?

14.66 Find the (*a*) 95% and (*b*) 99% confidence limits for a correlation coefficient that is computed to be 0.60 from a sample of size 28.

14.67 Work Problem 14.66 if the sample size is 52.

14.68 Find the 95% confidence limits for the correlation coefficients computed in (*a*) Problem 14.46 and (*b*) Problem 14.58.

14.69 Two correlation coefficients obtained from samples of size 23 and 28 were computed to be 0.80 and 0.95, respectively. Can we conclude at levels of (*a*) 0.05 and (*b*) 0.01 that there is a significant difference between the two coefficients?

SAMPLING THEORY OF REGRESSION

14.70 On the basis of a sample of size 27, a regression equation of Y on X was found to be $Y = 25.0 + 2.00X$. If $s_{Y.X} = 1.50$, $s_X = 3.00$, and $\bar{X} = 7.50$, find the (*a*) 95% and (*b*) 99% confidence limits for the regression coefficient.

14.71 In Problem 14.70, test the hypothesis that the population regression coefficient at the 0.01 significance level is (*a*) as low as 1.70 and (*b*) as high as 2.20.

14.72 In Problem 14.70, find the (*a*) 95% and (*b*) 99% confidence limits for Y when $X = 6.00$.

14.73 In Problem 14.70, find the (*a*) 95% and (*b*) 99% confidence limits for the mean of all values of Y corresponding to $X = 6.00$.

14.74 Referring to Problem 14.46, find the 95% confidence limits for (*a*) the regression coefficient of Y on X, (*b*) the blood pressures of all women who are 45 years old, and (*c*) the mean of the blood pressures of all women who are 45 years old.

E-next
THE NEXT LEVEL OF EDUCATION