9.5 USING THE REGRESSION EQUATION

If the results of the evaluation of the sample regression equation indicate that there is a relationship between the two variables of interest, we can put the regression equation to practical use. There are two ways in which the equation can be used. It can be used to *predict* what value Y is likely to assume given a particular value of X. When the normality assumption of Section 9.2 is met, a *prediction interval* for this predicted value of Y may be constructed.

We may also use the regression equation to *estimate* the mean of the subpopulation of Y values assumed to exist at any particular value of X. Again, if the assumption of normally distributed populations holds, a confidence interval for this parameter may be constructed. The predicted value of Y and the point estimate of the mean of the subpopulation of Y will be numerically equivalent for any particular value of X but, as we will see, the prediction interval will be wider than the confidence interval.

Predicting Y for a Given X If it is known, or if we are willing to assume that the assumptions of Section 9.2 are met, and when $\sigma^2_{y|x}$ is unknown, then the $100(1-\alpha)$ percent prediction interval for Y is given by

$$\hat{y} \pm t_{(1-\alpha/2)} s_{y|x} \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{9.5.1}$$

where x_p is the particular value of x at which we wish to obtain a prediction interval for Y and the degrees of freedom used in selecting t are $n-2$.

Estimating the Mean of Y for a Given X The $100(1-\alpha)$ percent confidence interval for $\mu_{y|x}$, when $\sigma^2_{y|x}$ is unknown, is given by

$$\hat{y} \pm t_{(1-\alpha/2)} s_{y|x} \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum (x_i - \bar{x})^2}} \tag{9.5.2}$$

We use MINITAB to illustrate, for a specified value of X, the calculation of a 95 percent confidence interval for the mean of Y and a 95 percent prediction interval for an individual Y measurement.

Suppose, for our present example, we wish to make predictions and estimates about AT for a waist circumference of 100 cm. In the regression dialog box click on "Options." Enter 100 in the "Prediction interval for new observations" box. Click on "Confidence limits," and click on "Prediction limits."

We obtain the following output:

Fit	Stdev.Fit	95.0% C.I.	95.0% P.I.
129.90	3.69	(122.58, 137.23)	(63.93, 195.87)

We interpret the 95 percent confidence interval (C.I.) as follows.

If we repeatedly drew samples from our population of men, performed a regression analysis, and estimated $\mu_{y|x=100}$ with a similarly constructed confidence interval, about 95 percent of such intervals would include the mean amount of deep abdominal AT for the population. For this reason we are 95 percent confident that the single interval constructed contains the population mean and that it is somewhere between 122.58 and 137.23.

Our interpretation of a prediction interval (P.I.) is similar to the interpretation of a confidence interval. If we repeatedly draw samples, do a regression analysis, and construct prediction intervals for men who have a waist circumference of 100 cm, about 95 percent of them will include the man's deep abdominal AT value. This is the probabilistic interpretation. The practical interpretation is that we are 95 percent confident that a man who has a waist circumference of 100 cm will have a deep abdominal AT area of somewhere between 63.93 and 195.87 square centimeters.

Simultaneous confidence intervals and prediction intervals can be calculated for all possible points along a fitted regression line. Plotting lines through these points will then provide a graphical representation of these intervals. Since the mean data point $(\overline{X}, \overline{Y})$ is always included in the regression equation, as illustrated by equations 9.3.2 and 9.3.3, plots of the simultaneous intervals will always provide the best estimates at the middle of the line and the error will increase toward the ends of the line. This illustrates the fact that estimation within the bounds of the data set, called *interpolation*, is acceptable, but that estimation outside of the bounds of the data set, called *extrapolation*, is not advisable since the pridiction error can be quite large. See Figure 9.5.1.

Figure 9.5.2 contains a partial printout of the SAS® simple linear regression analysis of the data of Example 9.3.1.

Resistant Line Frequently, data sets available for analysis by linear regression techniques contain one or more "unusual" observations; that is, values of x or y, or both, may be either considerably larger or considerably smaller than most of the other measurements. In the output of Figure 9.3.2, we see that the computer detected seven

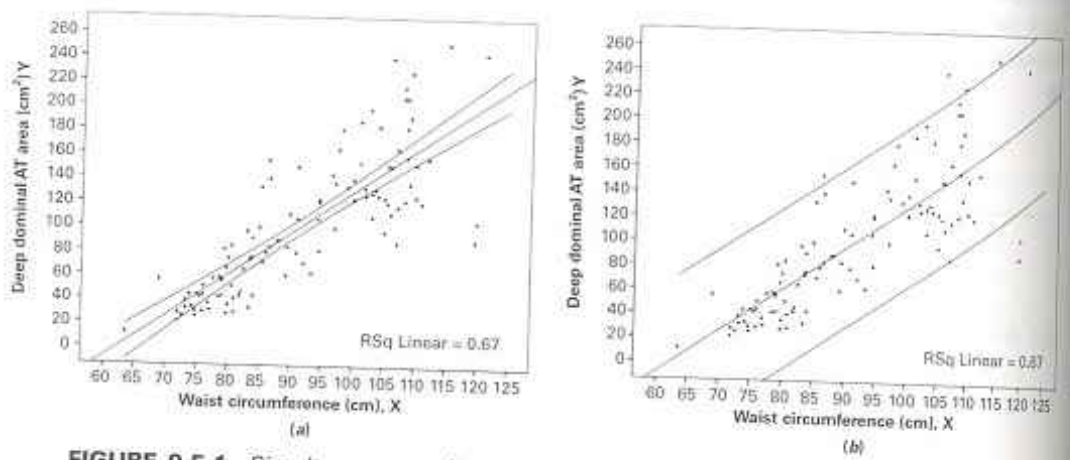

FIGURE 9.5.1 Simultaneous confidence intervals (*a*) and prediction intervals (*b*) for the data in Example 9.3.1.

```
The SAS System

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

                          Sum of           Mean
Source          DF       Squares         Square    F Value    Prob>F

Model            1   237548.51620   237548.51620    217.279    0.0001
Error          107   116981.98602     1093.28959
C Total        108   354530.50222

    Root MSE          33.06493      R-square      0.6700
    Dep Mean         101.89404      Adj R-sq      0.6670
    C.V.              32.45031

Parameter Estimates

                   Parameter       Standard      T for H0:
Variable  DF        Estimate          Error     Parameter=0   Prob > |T|

INTERCEP   1    -215.981488   21.79627076          -9.909       0.0001
X          1       3.458859    0.23465205          14.740       0.0001
```

FIGURE 9.5.2 Partial printout of the computer analysis of the data given in Example 9.3.1, using the SAS® software package.

unusual observations in the waist circumference and deep abdominal AT data shown in Table 9.3.1.

The least-squares method of fitting a straight line to data is sensitive to unusual observations, and the location of the fitted line can be affected substantially by them. Because of this characteristic of the least-squares method, the resulting least-squares line is said to lack *resistance* to the influence of unusual observations. Several methods have been devised for dealing with this problem, including one developed by John W. Tukey. The resulting line is variously referred to as *Tukey's line* and the *resistant line*.

Based on medians, which, as we have seen, are descriptive measures that are themselves resistant to extreme values, the resistant line methodology is an exploratory data analysis tool that enables the researcher to quickly fit a straight line to a set of data consisting of paired x, y measurements. The technique involves partitioning, on the basis of the independent variable, the sample measurements into three groups of as near equal size as possible: the smallest measurements, the largest measurements, and those in between. The resistant line is the line fitted in such a way that there are

Dialog box:

Stat ➤ EDA ➤ Resistant Line

Type *C2* in **Response** and *C1* in **Predictors.** Check **Residuals** and **Fits.** Click **OK.**

Session command:

```
MTB > Name C3 = 'RESI1' C4 = 'FITS1'
MTB > RLine C2 C1 'RESI1' 'FITS1';
SUBC>   MaxIterations 10.
```

Output:

Resistant Line Fit: C2 versus C1

```
Slope =      3.2869  Level =   -203.7868  Half-slope ratio =   0.690
```

FIGURE 9.5.3 MINITAB resistant line procedure and output for the data of Table 9.3.1.

an equal number of values above and below it in both the smaller group and the larger group. The resulting slope and y-intercept estimates are resistant to the effects of either extreme y values, extreme x values, or both. To illustrate the fitting of a resistant line, we use the data of Table 9.3.1 and MINITAB. The procedure and output are shown in Figure 9.5.3.

We see from the output in Figure 9.5.3 that the resistant line has a slope of 3.2869 and a y-intercept of -203.7868. The *half-slope ratio*, shown in the output as equal to .690, is an indicator of the degree of linearity between x and y. A slope, called a half-slope, is computed for each half of the sample data. The ratio of the right half-slope, b_R, and the left half-slope, b_L, is equal to b_R/b_L. If the relationship between x and y is straight, the half-slopes will be equal, and their ratio will be 1. A half-slope ratio that is not close to 1 indicates a lack of linearity between x and y.

The resistant line methodology is discussed in more detail by Hartwig and Dearing (1), Johnstone and Velleman (2), McNeil (3), and Velleman and Hoaglin (4).

EXERCISES

In each exercise refer to the appropriate previous exercise and, for the value of X indicated, (a) construct the 95 percent confidence interval for $\mu_{Y|X}$ and (b) construct the 95 percent prediction interval for Y.

9.5.1 Refer to Exercise 9.3.3 and let $X = 400$.

9.5.2 Refer to Exercise 9.3.4 and let $X = 1.6$.

9.5.3 Refer to Exercise 9.3.5 and let $X = 4.16$.

9.5.4 Refer to Exercise 9.3.6 and let $X = 29.4$.

9.5.5 Refer to Exercise 9.3.7 and let $X = 35$.

9.6 THE CORRELATION MODEL

In the classic regression model, which has been the underlying model in our discussion up to this point, only Y, which has been called the dependent variable, is required to be random. The variable X is defined as a fixed (nonrandom or mathematical) variable and is referred to as the independent variable. Recall, also, that under this model observations are frequently obtained by preselecting values of X and determining corresponding values of Y.

When both Y and X are random variables, we have what is called the *correlation model*. Typically, under the correlation model, sample observations are obtained by selecting a random sample of the *units of association* (which may be persons, places, animals, points in time, or any other element on which the two measurements are taken) and taking on each a measurement of X and a measurement of Y. In this procedure, values of X are not preselected but occur at random, depending on the unit of association selected in the sample.

Although correlation analysis cannot be carried out meaningfully under the classic regression model, regression analysis can be carried out under the correlation model. Correlation involving two variables implies a co-relationship between variables that puts them on an equal footing and does not distinguish between them by referring to one as the dependent and the other as the independent variable. In fact, in the basic computational procedures, which are the same as for the regression model, we may fit a straight line to the data either by minimizing $\sum(y_i - \hat{y}_i)^2$ or by minimizing $\sum(x_i - \hat{x}_i)^2$. In other words, we may do a regression of X on Y as well as a regression of Y on X. The fitted line in the two cases in general will be different, and a logical question arises as to which line to fit.

If the objective is solely to obtain a measure of the strength of the relationship between the two variables, it does not matter which line is fitted, since the measure usually computed will be the same in either case. If, however, it is desired to use the equation describing the relationship between the two variables for the purposes discussed in the preceding sections, it does matter which line is fitted. The variable for which we wish to estimate means or to make predictions should be treated as the dependent variable; that is, this variable should be regressed on the other variable.

The Bivariate Normal Distribution Under the correlation model, X and Y are assumed to vary together in what is called a *joint distribution*. If this joint distribution is a normal distribution, it is referred to as a *bivariate normal distribution*. Inferences regarding this population may be made based on the results of samples properly drawn from it. If, on the other hand, the form of the joint distribution is known to be nonnormal, or if the form is unknown and there is no justification for assuming normality, inferential procedures are invalid, although descriptive measures may be computed.

Correlation Assumptions The following assumptions must hold for inferences about the population to be valid when sampling is from a bivariate distribution.

1. For each value of X there is a normally distributed subpopulation of Y values.
2. For each value of Y there is a normally distributed subpopulation of X values.
3. The joint distribution of X and Y is a normal distribution called the *bivariate normal distribution*.

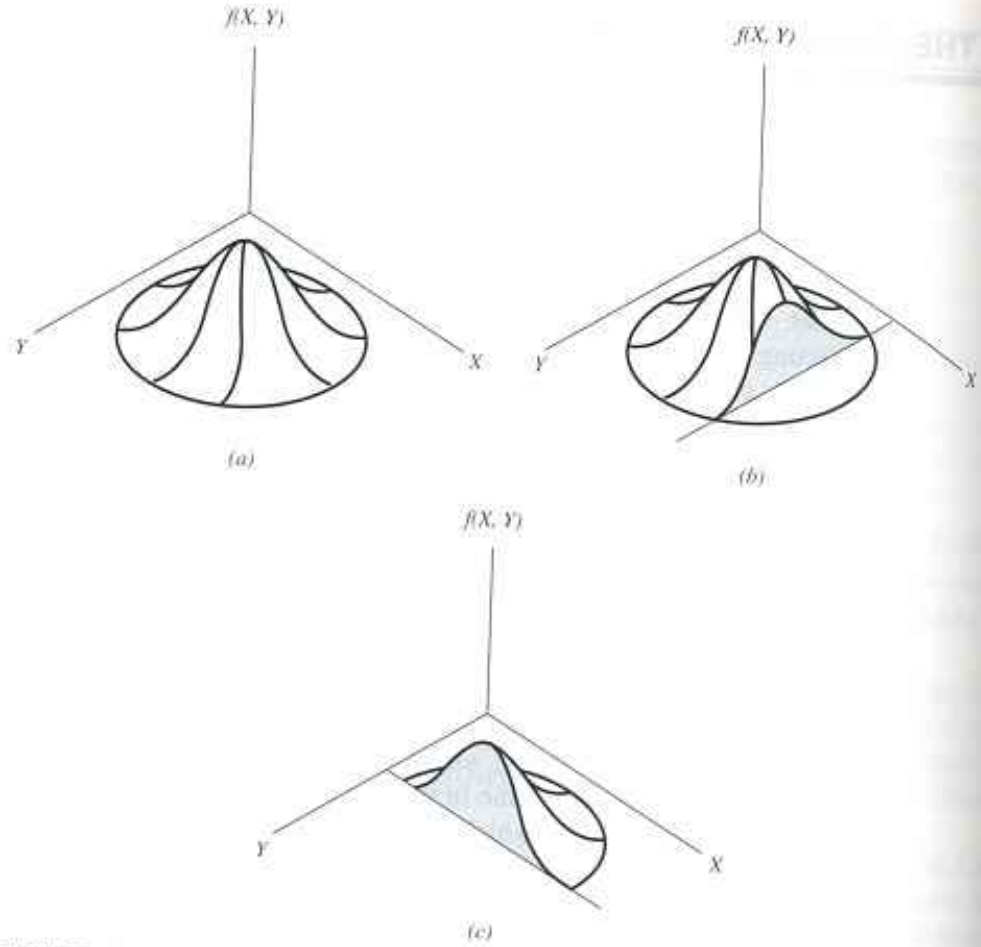

FIGURE 9.6.1 A bivariate normal distribution. (*a*) A bivariate normal distribution. (*b*) A cutaway showing normally distributed subpopulation of Y for given X. (*c*) A cutaway showing normally distributed subpopulation of X for given Y.

4. The subpopulations of Y values all have the same variance.
5. The subpopulations of X values all have the same variance.

The bivariate normal distribution is represented graphically in Figure 9.6.1. In this illustration we see that if we slice the mound parallel to Y at some value of X, the cutaway reveals the corresponding normal distribution of Y. Similarly, a slice through the mound parallel to X at some value of Y reveals the corresponding normally distributed subpopulation of X.

9.7 THE CORRELATION COEFFICIENT

The bivariate normal distribution discussed in Section 9.6 has five parameters, σ_x, σ_y, μ_x, μ_y, and ρ. The first four are, respectively, the standard deviations and means associated with the individual distributions. The other parameter, ρ, is called the population

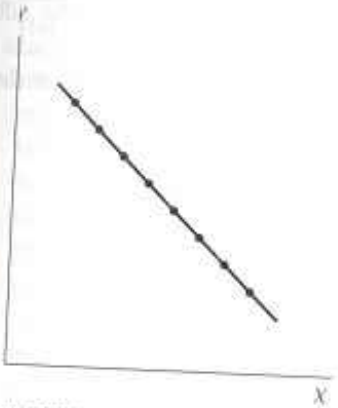

FIGURE 9.7.1 Scatter diagram for $r = -1$.

correlation coefficient and measures the strength of the linear relationship between X and Y.

The population correlation coefficient is the positive or negative square root of ρ^2, the population coefficient of determination previously discussed, and since the coefficient of determination takes on values between 0 and 1 inclusive, ρ may assume any value between -1 and $+1$. If $\rho = 1$ there is a perfect direct linear correlation between the two variables, while $\rho = -1$ indicates perfect inverse linear correlation. If $\rho = 0$ the two variables are not linearly correlated. The sign of ρ will always be the same as the sign of β_1, the slope of the population regression line for X and Y.

The sample correlation coefficient, r, describes the linear relationship between the sample observations on two variables in the same way that ρ describes the relationship in a population. The sample correlation coefficient is the square root of the sample coefficient of determination that was defined earlier.

Figures 9.4.5(d) and 9.4.5(c), respectively, show typical scatter diagrams where $r \to 0$ ($r^2 \to 0$) and $r = +1$ ($r^2 = 1$). Figure 9.7.1 shows a typical scatter diagram where $r = -1$.

We are usually interested in knowing if we may conclude that $\rho \neq 0$, that is, that X and Y are linearly correlated. Since ρ is usually unknown, we draw a random sample from the population of interest, compute r, the estimate of ρ, and test $H_0: \rho = 0$ against the alternative $\rho \neq 0$. The procedure will be illustrated in the following example.

EXAMPLE 9.7.1

The purpose of a study by Kwast-Rabben et al. (A-7) was to analyze somatosensory evoked potentials (SEPs) and their interrelations following stimulation of digits I, III, and V in the hand. The researchers wanted to establish reference criteria in a control population. Thus, healthy volunteers were recruited for the study. In the future this information could be quite valuable as SEPs may provide a method to demonstrate functional disturbances in patients with suspected cervical root lesion who have pain and sensory symptoms. In the study, stimulation below-pain-level intensity was applied to the fingers.

Recordings of spinal responses were made with electrodes fixed by adhesive electrode cream to the subject's skin. One of the relationships of interest was the correlation between a subject's height (cm) and the peak spinal latency (Cv) of the SEP. The data for 155 measurements are shown in Table 9.7.1.

TABLE 9.7.1 Height and Spine SEP Measurements (Cv) from Stimulation of Digit I for 155 Subjects Described in Example 9.7.1

Height	Cv	Height	Cv	Height	Cv
149	14.4	168	16.3	181	15.8
149	13.4	168	15.3	181	18.8
155	13.5	168	16.0	181	18.6
155	13.5	168	16.6	182	18.0
156	13.0	168	15.7	182	17.9
156	13.6	168	16.3	182	17.5
157	14.3	168	16.6	182	17.4
157	14.9	168	15.4	182	17.0
158	14.0	170	16.6	182	17.5
158	14.0	170	16.0	182	17.8
160	15.4	170	17.0	184	18.4
160	14.7	170	16.4	184	18.5
161	15.5	171	16.5	184	17.7
161	15.7	171	16.3	184	17.7
161	15.8	171	16.4	184	17.4
161	16.0	171	16.5	184	18.4
161	14.6	172	17.6	185	19.0
161	15.2	172	16.8	185	19.6
162	15.2	172	17.0	187	19.1
162	16.5	172	17.6	187	19.2
162	17.0	173	17.3	187	17.8
162	14.7	173	16.8	187	19.3
163	16.0	174	15.5	188	17.5
163	15.8	174	15.5	188	18.0
163	17.0	175	17.0	189	18.0
163	15.1	175	15.6	189	18.8
163	14.6	175	16.8	190	18.3
163	15.6	175	17.4	190	18.6
163	14.6	175	17.6	190	18.8
164	17.0	175	16.5	190	19.2
164	16.3	175	16.6	191	18.5
164	16.0	175	17.0	191	18.5
164	16.0	176	18.0	191	19.0
165	15.7	176	17.0	191	18.5
165	16.3	176	17.4	194	19.8

(Continued)

Height	Cv	Height	Cv	Height	Cv
165	17.4	176	18.2	194	18.8
165	17.0	176	17.3	194	18.4
165	16.3	177	17.2	194	19.0
166	14.1	177	18.3	195	18.0
166	14.2	179	16.4	195	18.2
165	14.7	179	16.1	196	17.6
166	13.9	179	17.6	196	18.3
166	17.2	179	17.8	197	18.9
167	16.7	179	16.1	197	19.2
167	16.5	179	16.0	200	21.0
167	14.7	179	16.0	200	19.2
167	14.3	179	17.5	202	18.6
167	14.8	179	17.5	202	18.6
167	15.0	180	18.0	182	20.0
167	15.5	180	17.9	190	20.0
167	15.4	181	18.4	190	19.5
168	17.3	181	16.4		

Source: Olga Kwast-Rabben, Ph.D. Used with permission.

Solution: The scatter diagram and least-squares regression line are shown in Figure 9.7.2.

Let us assume that the investigator wishes to obtain a regression equation to use for estimating and predicting purposes. In that case the sample correlation coefficient will be obtained by the methods discussed under the regression model.

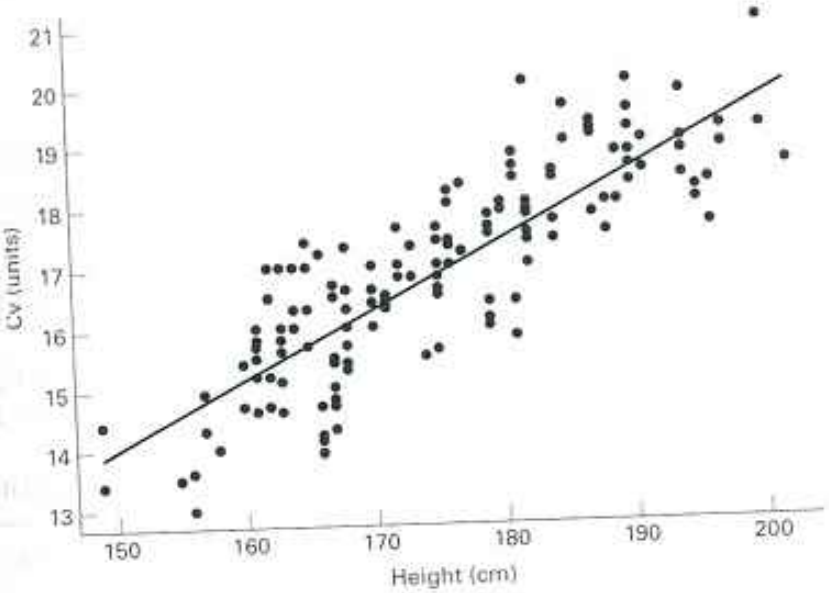

FIGURE 9.7.2 Height and cervical (spine) potentials in digit I stimulation for the data described in Example 9.7.1.

```
The regression equation is
Cv = -3.20 + 0.115 Height

Predictor        Coef     SE Coef         T        P
Constant       -3.198       1.016     -3.15    0.002
Height       0.114567    0.005792     19.78    0.000

S = 0.8573      R-Sq = 71.9%      R-Sq(adj) = 71.7%

Analysis of Variance

Source            DF          SS          MS          F      P
Regression         1      287.56      287.56     391.30  0.000
Residual Error   153      112.44        0.73
Total            154      400.00

Unusual Observations
Obs    Height         Cv        Fit    SE Fit    Residual    St Resid
 39       166    14.1000    15.8199    0.0865     -1.7199     -2.02R
 42       166    13.9000    15.8199    0.0865     -1.9199     -2.25R
105       181    15.8000    17.5384    0.0770     -1.7384     -2.04R
151       202    18.6000    19.9443    0.1706     -1.3443     -1.60 X
152       202    18.6000    19.9443    0.1706     -1.3443     -1.60 X
153       182    20.0000    17.6529    0.0798      2.3471      2.75R

R denotes an observation with a large standardized residual
X denotes an observation whose X value gives it large influence.
```

FIGURE 9.7.3 MINITAB output for Example 9.7.1 using the simple regression procedure.

The Regression Equation

Let us assume that we wish to predict Cv levels from knowledge of heights. In that case we treat height as the independent variable and Cv level as the dependent variable and obtain the regression equation and correlation coefficient with MINITAB as shown in Figure 9.7.3. For this example $r = \sqrt{.719} = .848$. We know that r is positive because the slope of the regression line is positive. We may also use the MINITAB correlation procedure to obtain r as shown in Figure 9.7.4.

The printout from the SAS® correlation procedure is shown in Figure 9.7.5. Note that the SAS® procedure gives descriptive measures for each variable as well as the p value for the correlation coefficient.

When a computer is not available for performing the calculations, r may be obtained by means of the following formulas:

$$r = \sqrt{\frac{\hat{\beta}_1^2[\sum x_i^2 - (\sum x_i)^2/n]}{\sum y_i^2 - (\sum y_i)^2/n}} \tag{9.7.1}$$

Data:

```
C1: Height
C2: Cv
```

Dialog Box:

Stat ➤ Basic Statistics ➤ Correlation

Type *C1 C2* in **Variables.** Click **OK.**

Session command:

```
MTB > Correlation C1 C2.
```

OUTPUT:

Correlations: Height, Cv

```
Pearson correlation of Height and Cv = 0.848
P-Value = 0.000
```

FIGURE 9.7.4 MINITAB procedure for Example 9.7.1 using the correlation command.

The CORR Procedure

2 Variables: HEIGHT CV

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
HEIGHT	155	175.04516	11.92745	27132	149.00000	202.00000
CV	155	16.85613	1.61165	2613	13.00000	21.00000

Pearson Correlation Coefficients, N = 155

Prob > |r| under H0: Rho=0

	HEIGHT	CV
HEIGHT	1.00000	0.84788 <.0001
CV	0.84788 <.0001	1.00000

FIGURE 9.7.5 SAS® printout for Example 9.7.1.

An alternative formula for computing r is given by

$$r = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{9.7.2}$$

An advantage of this formula is that r may be computed without first computing b. This is the desirable procedure when it is not anticipated that the regression equation will be used.

Remember that the sample correlation coefficient, r, will always have the same sign as the sample slope, b. ∎

EXAMPLE 9.7.2

Refer to Example 9.7.1. We wish to see if the sample value of $r = .848$ is of sufficient magnitude to indicate that, in the population, height and Cv SEP levels are correlated.

Solution: We conduct a hypothesis test as follows.

1. **Data.** See the initial discussion of Example 9.7.1.
2. **Assumptions.** We presume that the assumptions given in Section 9.6 are applicable.
3. **Hypotheses.**

$$H_0: \rho = 0$$
$$H_A: \rho \neq 0$$

4. **Test statistic.** When $\rho = 0$, it can be shown that the appropriate test statistic is

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{9.7.3}$$

5. **Distribution of test statistic.** When H_0 is true and the assumptions are met, the test statistic is distributed as Student's t distribution with $n - 2$ degrees of freedom.
6. **Decision rule.** If we let $\alpha = .05$, the critical values of t in the present example are $\pm$ 1.9754 (by interpolation). If, from our data, we compute a value of t that is either greater than or equal to +1.9754 or less than or equal to −1.9754, we will reject the null hypothesis.
7. **Calculation of test statistic.** Our calculated value of t is

$$t = .848\sqrt{\frac{153}{1-.719}} = 19.787$$

8. **Statistical decision.** Since the computed value of the test statistic does exceed the critical value of t, we reject the null hypothesis.

9. **Conclusion.** We conclude that, in the population, height and SEP levels in the spine are linearly correlated.
10. ***p* value.** Since $t = 19.787 > 2.6085$ (interpolated value of t for 153, .995), we have for this test, $p < .005$. ■

A Test for Use When the Hypothesized ρ Is a Nonzero Value

The use of the t statistic computed in the above test is appropriate only for testing $H_0: \rho = 0$. If it is desired to test $H_0: \rho = \rho_0$, where ρ_0 is some value other than zero, we must use another approach. Fisher (5) suggests that r be transformed to z_r as follows:

$$z_r = \frac{1}{2} \ln \frac{1 + r}{1 - r} \tag{9.7.4}$$

where ln is a natural logarithm. It can be shown that z_r is approximately normally distributed with a mean of $z_\rho = \frac{1}{2}\ln\{(1 + \rho)/(1 - \rho)\}$ and estimated standard deviation of

$$\sigma_{z_r} = \frac{1}{\sqrt{n - 3}} \tag{9.7.5}$$

To test the null hypothesis that ρ is equal to some value other than zero, the test statistic is

$$Z = \frac{z_r - z_\rho}{1/\sqrt{n - 3}} \tag{9.7.6}$$

which follows approximately the standard normal distribution.

To determine z_r for an observed r and z_ρ for a hypothesized ρ, we consult Table I, thereby avoiding the direct use of natural logarithms.

Suppose in our present example we wish to test

$$H_0: \rho = .80$$

against the alternative

$$H_A: \rho \neq .80$$

at the .05 level of significance. By consulting Table I (and interpolating), we find that for

$$r = .848, \qquad z_r = 1.24726$$

and for

$$\rho = .80, \qquad z_\rho = 1.09861$$

Our test statistic, then, is

$$Z = \frac{1.24726 - 1.09861}{1/\sqrt{155 - 3}} = 1.83$$

Since 1.83 is less than the critical value of $z = 1.96$, we are unable to reject H_0. We conclude that the population correlation coefficient may be .80.

For sample sizes less than 25, Fisher's Z transformation should be used with caution, if at all. An alternative procedure from Hotelling (6) may be used for sample sizes equal to or greater than 10. In this procedure the following transformation of r is employed:

$$z^* = z_r - \frac{3z_r + r}{4n} \tag{9.7.7}$$

The standard deviation of z^* is

$$\sigma_{z^*} = \frac{1}{\sqrt{n-1}} \tag{9.7.8}$$

The test statistic is

$$Z^* = \frac{z^* - \zeta^*}{1/\sqrt{n-1}} = (z^* - \zeta^*)\sqrt{n-1} \tag{9.7.9}$$

where

$$\zeta^* \text{ (pronounced zeta)} = z_\rho - \frac{(3z_\rho + \rho)}{4n}$$

Critical values for comparison purposes are obtained from the standard normal distribution.

In our present example, to test $H_0: \rho = .80$ against $H_A: \rho \neq .80$ using the Hotelling transformation and $\alpha = .05$, we have

$$z^* = 1.24726 - \frac{3(1.24726) + .848}{4(155)} = 1.2339$$

$$\zeta^* = 1.09861 - \frac{3(1.09861) + .8}{4(155)} = 1.0920$$

$$Z^* = (1.2339 - 1.0920)\sqrt{155 - 1} = 1.7609$$

Since 1.7609 is less than 1.96, the null hypothesis is not rejected, and the same conclusion is reached as when the Fisher transformation is used.

Alternatives In some situations the data available for analysis do not meet the assumptions necessary for the valid use of the procedures discussed here for testing hypotheses about a population correlation coefficient. In such cases it may be more appropriate to use the Spearman rank correlation technique discussed in Chapter 13.

Confidence Interval for ρ Fisher's transformation may be used to construct $100(1 - \alpha)$ percent confidence intervals for ρ. The general formula for a confidence interval

$$\text{estimator} \pm (\text{reliability factor})(\text{standard error})$$

is employed. We first convert our estimator, r, to z_r, construct a confidence interval about z_ρ, and then reconvert the limits to obtain a $100(1 - \alpha)$ percent confidence interval about ρ. The general formula then becomes

$$z_r \pm z(1/\sqrt{n-3}) \tag{9.7.10}$$

For our present example the 95 percent confidence interval for z_ρ is given by

$$1.24726 \pm 1.96(1/\sqrt{155-3})$$
$$1.08828,\ 1.40624$$

Converting these limits (by interpolation in Appendix Table I), which are values of z_r, into values of r gives

z_r	r
1.08828	.7962
1.40624	.8866

We are 95 percent confident, then, that ρ is contained in the interval .7962 to .88866. Because of the limited entries in the table, these limits must be considered as only approximate.

EXERCISES

In each of the following exercises:

(a) Prepare a scatter diagram.

(b) Compute the sample correlation coefficient.

(c) Test H_0: $\rho = 0$ at the .05 level of significance and state your conclusions.

(d) Determine the p value for the test.

(e) Construct the 95 percent confidence interval for ρ.

9.7.1 The purpose of a study by Brown and Persley (A-8) was to characterize acute hepatitis A in patients more than 40 years old. They performed a retrospective chart review of 20 subjects who were diagnosed with acute hepatitis A, but were not hospitalized. Of interest was the use of age (years) to predict bilirubin levels (mg/dl). The following data were collected.

Age (Years)	Bilirubin (mg/dl)	Age (Years)	Bilirubin (mg/dl)
78	7.5	44	7.0
72	12.9	42	1.8
81	14.3	45	.8
59	8.0	78	3.8
64	14.1	47	3.5
48	10.9	50	5.1
46	12.3	57	16.5

(*Continued*)

Age (Years)	Bilirubin (mg/dl)	Age (Years)	Bilirubin (mg/dl)
42	1.0	52	3.5
58	5.2	58	5.6
52	5.1	45	1.9

Source: Geri R. Brown, M.D. Used with permission.

9.7.2 Another variable of interest in the study by Reiss et al. (A-3) (see Exercise 9.3.4) was partial thromboplastin (aPTT), the standard test used to monitor heparin anticoagulation. Use the data in the following table to examine the correlation between aPTT levels as measured by the CoaguCheck point-of-care assay and standard laboratory hospital assay in 90 subjects receiving heparin alone, heparin with warfarin, and warfarin and exoenoxaparin.

Heparin		Warfarin		Warfarin and Exoenoxaparin	
CoaguCheck aPTT	Hospital aPTT	CoaguCheck aPTT	Hospital aPTT	CoaguCheck aPTT	Hospital aPTT
49.3	71.4	18.0	77.0	56.5	46.5
57.9	86.4	31.2	62.2	50.7	34.9
59.0	75.6	58.7	53.2	37.3	28.0
77.3	54.5	75.2	53.0	64.8	52.3
42.3	57.7	18.0	45.7	41.2	37.5
44.3	59.5	82.6	81.1	90.1	47.1
90.0	77.2	29.6	40.9	23.1	27.1
55.4	63.3	82.9	75.4	53.2	40.6
20.3	27.6	58.7	55.7	27.3	37.8
28.7	52.6	64.8	54.0	67.5	50.4
64.3	101.6	37.9	79.4	33.6	34.2
90.4	89.4	81.2	62.5	45.1	34.8
64.3	66.2	18.0	36.5	56.2	44.2
89.8	69.8	38.8	32.8	26.0	28.2
74.7	91.3	95.4	68.9	67.8	46.3
150.0	118.8	53.7	71.3	40.7	41.0
32.4	30.9	128.3	111.1	36.2	35.7
20.9	65.2	60.5	80.5	60.8	47.2
89.5	77.9	150.0	150.0	30.2	39.7
44.7	91.5	38.5	46.5	18.0	31.3
61.0	90.5	58.9	89.1	55.6	53.0
36.4	33.6	112.8	66.7	18.0	27.4
52.9	88.0	26.7	29.5	18.0	35.7
57.5	69.9	49.7	47.8	78.3	62.0
39.1	41.0	85.6	63.3	75.3	36.7
74.8	81.7	68.8	43.5	73.2	85.3
32.5	33.3	18.0	54.0	42.0	38.3
125.7	142.9	92.6	100.5	49.3	39.8
77.1	98.2	46.2	52.4	22.8	42.3
143.8	108.3	60.5	93.7	35.8	36.0

Source: Curtis E. Haas, Pharm. D. Used with permission.

9.7.3 In the study by Parker et al. (A-4) (see Exercise 9.3.5), the authors also looked at the change in AUC (area under the curve of plasma concentration of digoxin) when comparing digoxin levels taken with and without grapefruit juice. The following table gives the AUC when digoxin was consumed with water (ng · hr/ml) and the change in AUC compared to the change in AUC when digoxin is taken with grapefruit juice (GFJ, %).

Water AUC Level (ng · hr/ml)	Change in AUC with GFJ (%)
6.96	17.4
5.59	24.5
5.31	8.5
8.22	20.8
11.91	−26.7
9.50	−29.3
11.28	−16.8

Source: Robert B. Parker, Pharm. D. Used with permission.

9.7.4 An article by Tuzson et al. (A-9) in *Archives of Physical Medicine and Rehabilitation* reported the following data on peak knee velocity in walking (measured in degrees per second) at flexion and extension for 18 subjects with cerebral palsy.

Flexion (°/s)	Extension (°/s)
100	100
150	150
210	180
255	165
200	210
185	155
440	440
110	180
400	400
160	140
150	250
425	275
375	340
400	400
400	450
300	300
300	300
320	275

Source: Ann E. Tuzson, Kevin P. Granata, and Mark F. Abel, "Spastic Velocity Threshold Constrains Functional Performance in Cerebral Palsy," *Archives of Physical Medicine and Rehabilitation*, *84* (2003), 1363–1368.

9.7.5 Amyotrophic lateral sclerosis (ALS) is characterized by a progressive decline of motor function. The degenerative process affects the respiratory system. Butz et al. (A-10) investigated the longitudinal impact of nocturnal noninvasive positive-pressure ventilation on patients with ALS. Prior to treatment, they measured partial pressure of arterial oxygen (Pao_2) and partial pressure of arterial carbon dioxide ($Paco_2$) in patients with the disease. The results were as follows:

$Paco_2$	Pao_2
40.0	101.0
47.0	69.0
34.0	132.0
42.0	65.0
54.0	72.0
48.0	76.0
53.6	67.2
56.9	70.9
58.0	73.0
45.0	66.0
54.5	80.0
54.0	72.0
43.0	105.0
44.3	113.0
53.9	69.2
41.8	66.7
33.0	67.0
43.1	77.5
52.4	65.1
37.9	71.0
34.5	86.5
40.1	74.7
33.0	94.0
59.9	60.4
62.6	52.5
54.1	76.9
45.7	65.3
40.6	80.3
56.6	53.2
59.0	71.9

Source: M. Butz, K. H. Wollinsky, U. Widemuth-Catrinescu, A. Sperfeld, S. Winter, H. H. Mehrkens, A. C. Ludolph, and H. Schreiber, "Longitudinal Effects of Noninvasive Positive-Pressure Ventilation in Patients with Amyotrophic Lateral Sclerosis," *American Journal of Medical Rehabilitation*, 82 (2003) 597–604.

9.7.6 A simple random sample of 15 apparently healthy children between the ages of 6 months and 15 years yielded the following data on age, X, and liver volume per unit of body weight (ml/kg), Y.

X	Y	X	Y
.5	41	10.0	26
.7	55	10.1	35
2.5	41	10.9	25
4.1	39	11.5	31

(*Continued*)