Chapter 2

Inferences in Regression and Correlation Analysis

In this chapter, we first take up inferences concerning the regression parameters β_0 and β_1, considering both interval estimation of these parameters and tests about them. We then discuss interval estimation of the mean $E\{Y\}$ of the probability distribution of Y, for given X, prediction intervals for a new observation Y, confidence bands for the regression line, the analysis of variance approach to regression analysis, the general linear test approach, and descriptive measures of association. Finally, we take up the correlation coefficient, a measure of association between X and Y when both X and Y are random variables.

Throughout this chapter (excluding Section 2.11), and in the remainder of Part I unless otherwise stated, we assume that the normal error regression model (1.24) is applicable. This model is:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \textbf{(2.1)}$$

where:

β_0 and β_1 are parameters
X_i are known constants
ε_i are independent $N(0, \sigma^2)$

2.1 Inferences Concerning β_1

Frequently, we are interested in drawing inferences about β_1, the slope of the regression line in model (2.1). For instance, a market research analyst studying the relation between sales (Y) and advertising expenditures (X) may wish to obtain an interval estimate of β_1 because it will provide information as to how many additional sales dollars, on the average, are generated by an additional dollar of advertising expenditure.

At times, tests concerning β_1 are of interest, particularly one of the form:

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

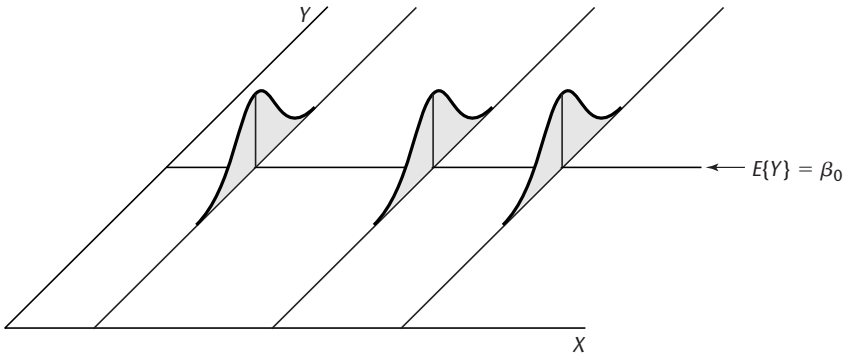


FIGURE 2.1
Regression Model (2.1) when $\beta_1 = 0$.

The reason for interest in testing whether or not $\beta_1 = 0$ is that, when $\beta_1 = 0$, there is no linear association between Y and X. Figure 2.1 illustrates the case when $\beta_1 = 0$. Note that the regression line is horizontal and that the means of the probability distributions of Y are therefore all equal, namely:

$$E\{Y\} = \beta_0 + (0)X = \beta_0$$

For normal error regression model (2.1), the condition $\beta_1 = 0$ implies even more than no linear association between Y and X. Since for this model all probability distributions of Y are normal with constant variance, and since the means are equal when $\beta_1 = 0$, it follows that the probability distributions of Y are identical when $\beta_1 = 0$. This is shown in Figure 2.1. Thus, $\beta_1 = 0$ for the normal error regression model (2.1) implies not only that there is no linear association between Y and X but also that there is no relation of any type between Y and X, since the probability distributions of Y are then identical at all levels of X.

Before discussing inferences concerning β_1 further, we need to consider the sampling distribution of b_1, the point estimator of β_1.

Sampling Distribution of b_1

The point estimator b_1 was given in (1.10a) as follows:

$$b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2} \tag{2.2}$$

The sampling distribution of b_1 refers to the different values of b_1 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from sample to sample.

For normal error regression model (2.1), the sampling distribution of b_1 is normal, with mean and variance: **(2.3)**

$$E\{b_1\} = \beta_1 \tag{2.3a}$$

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} \tag{2.3b}$$

To show this, we need to recognize that b_1 is a linear combination of the observations Y_i.

b_1 as Linear Combination of the Y_i. It can be shown that b_1, as defined in (2.2), can be expressed as follows:

$$b_1 = \sum k_i Y_i \tag{2.4}$$

where:

$$k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \tag{2.4a}$$

Observe that the k_i are a function of the X_i and therefore are fixed quantities since the X_i are fixed. Hence, b_1 is a linear combination of the Y_i where the coefficients are solely a function of the fixed X_i.

The coefficients k_i have a number of interesting properties that will be used later:

$$\sum k_i = 0 \tag{2.5}$$

$$\sum k_i X_i = 1 \tag{2.6}$$

$$\sum k_i^2 = \frac{1}{\sum (X_i - \bar{X})^2} \tag{2.7}$$

Comments

1. To show that b_1 is a linear combination of the Y_i with coefficients k_i, we first prove:

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X}) Y_i \tag{2.8}$$

This follows since:

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X}) Y_i - \sum (X_i - \bar{X}) \bar{Y}$$

But $\sum (X_i - \bar{X}) \bar{Y} = \bar{Y} \sum (X_i - \bar{X}) = 0$ since $\sum (X_i - \bar{X}) = 0$, Hence, (2.8) holds.

We now express b_1 using (2.8) and (2.4a):

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} = \sum k_i Y_i$$

2. The proofs of the properties of the k_i are direct. For example, property (2.5) follows because:

$$\sum k_i = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right] = \frac{1}{\sum (X_i - \bar{X})^2} \sum (X_i - \bar{X}) = \frac{0}{\sum (X_i - \bar{X})^2} = 0$$

Similarly, property (2.7) follows because:

$$\sum k_i^2 = \sum \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right]^2 = \frac{1}{\left[\sum (X_i - \bar{X})^2 \right]^2} \sum (X_i - \bar{X})^2 = \frac{1}{\sum (X_i - \bar{X})^2}$$

■

Normality. We return now to the sampling distribution of b_1 for the normal error regression model (2.1). The normality of the sampling distribution of b_1 follows at once from the fact that b_1 is a linear combination of the Y_i. The Y_i are independently, normally distributed

according to model (2.1), and (A.40) in Appendix A states that a linear combination of independent normal random variables is normally distributed.

Mean. The unbiasedness of the point estimator b_1, stated earlier in the Gauss-Markov theorem (1.11), is easy to show:

$$E\{b_1\} = E\left\{\sum k_i Y_i\right\} = \sum k_i E\{Y_i\} = \sum k_i(\beta_0 + \beta_1 X_i)$$
$$= \beta_0 \sum k_i + \beta_1 \sum k_i X_i$$

By (2.5) and (2.6), we then obtain $E\{b_1\} = \beta_1$.

Variance. The variance of b_1 can be derived readily. We need only remember that the Y_i are independent random variables, each with variance σ^2, and that the k_i are constants. Hence, we obtain by (A.31):

$$\sigma^2\{b_1\} = \sigma^2\left\{\sum k_i Y_i\right\} = \sum k_i^2 \sigma^2\{Y_i\}$$
$$= \sum k_i^2 \sigma^2 = \sigma^2 \sum k_i^2$$
$$= \sigma^2 \frac{1}{\sum (X_i - \bar{X})^2}$$

The last step follows from (2.7).

Estimated Variance. We can estimate the variance of the sampling distribution of b_1:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

by replacing the parameter σ^2 with *MSE*, the unbiased estimator of σ^2:

$$s^2\{b_1\} = \frac{MSE}{\sum (X_i - \bar{X})^2} \tag{2.9}$$

The point estimator $s^2\{b_1\}$ is an unbiased estimator of $\sigma^2\{b_1\}$. Taking the positive square root, we obtain $s\{b_1\}$, the point estimator of $\sigma\{b_1\}$.

Comment

We stated in theorem (1.11) that b_1 has minimum variance among all unbiased linear estimators of the form:

$$\hat{\beta}_1 = \sum c_i Y_i$$

where the c_i are arbitrary constants. We now prove this. Since $\hat{\beta}_1$ is required to be unbiased, the following must hold:

$$E\{\hat{\beta}_1\} = E\left\{\sum c_i Y_i\right\} = \sum c_i E\{Y_i\} = \beta_1$$

Now $E\{Y_i\} = \beta_0 + \beta_1 X_i$ by (1.2), so the above condition becomes:

$$E\{\hat{\beta}_1\} = \sum c_i(\beta_0 + \beta_1 X_i) = \beta_0 \sum c_i + \beta_1 \sum c_i X_i = \beta_1$$

For the unbiasedness condition to hold, the c_i must follow the restrictions:

$$\sum c_i = 0 \qquad \sum c_i X_i = 1$$

Now the variance of $\hat{\beta}_1$ is, by (A.31):

$$\sigma^2\{\hat{\beta}_1\} = \sum c_i^2 \sigma^2\{Y_i\} = \sigma^2 \sum c_i^2$$

Let us define $c_i = k_i + d_i$, where the k_i are the least squares constants in (2.4a) and the d_i are arbitrary constants. We can then write:

$$\sigma^2\{\hat{\beta}_1\} = \sigma^2 \sum c_i^2 = \sigma^2 \sum (k_i + d_i)^2 = \sigma^2 \left(\sum k_i^2 + \sum d_i^2 + 2 \sum k_i d_i \right)$$

We know that $\sigma^2 \sum k_i^2 = \sigma^2\{b_1\}$ from our proof above. Further, $\sum k_i d_i = 0$ because of the restrictions on the k_i and c_i above:

$$\begin{aligned} \sum k_i d_i &= \sum k_i (c_i - k_i) \\ &= \sum c_i k_i - \sum k_i^2 \\ &= \sum c_i \left[\frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2} \right] - \frac{1}{\sum (X_i - \bar{X})^2} \\ &= \frac{\sum c_i X_i - \bar{X} \sum c_i}{\sum (X_i - \bar{X})^2} - \frac{1}{\sum (X_i - \bar{X})^2} = 0 \end{aligned}$$

Hence, we have:

$$\sigma^2\{\hat{\beta}_1\} = \sigma^2\{b_1\} + \sigma^2 \sum d_i^2$$

Note that the smallest value of $\sum d_i^2$ is zero. Hence, the variance of $\hat{\beta}_1$ is at a minimum when $\sum d_i^2 = 0$. But this can only occur if all $d_i = 0$, which implies $c_i \equiv k_i$. Thus, the least squares estimator b_1 has minimum variance among all unbiased linear estimators. ■

Sampling Distribution of $(b_1 - \beta_1)/s\{b_1\}$

Since b_1 is normally distributed, we know that the standardized statistic $(b_1 - \beta_1)/\sigma\{b_1\}$ is a standard normal variable. Ordinarily, of course, we need to estimate $\sigma\{b_1\}$ by $s\{b_1\}$, and hence are interested in the distribution of the statistic $(b_1 - \beta_1)/s\{b_1\}$. When a statistic is standardized but the denominator is an estimated standard deviation rather than the true standard deviation, it is called a *studentized statistic*. An important theorem in statistics states the following about the studentized statistic $(b_1 - \beta_1)/s\{b_1\}$:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \text{ is distributed as } t(n-2) \text{ for regression model (2.1)} \qquad \textbf{(2.10)}$$

Intuitively, this result should not be unexpected. We know that if the observations Y_i come from the same normal population, $(\bar{Y} - \mu)/s\{\bar{Y}\}$ follows the t distribution with $n - 1$ degrees of freedom. The estimator b_1, like $\bar{Y}$, is a linear combination of the observations Y_i. The reason for the difference in the degrees of freedom is that two parameters (β_0 and β_1) need to be estimated for the regression model; hence, two degrees of freedom are lost here.

Comment

We can show that the studentized statistic $(b_1 - \beta_1)/s\{b_1\}$ is distributed as t with $n-2$ degrees of freedom by relying on the following theorem:

> For regression model (2.1), SSE/σ^2 is distributed as χ^2 with $n-2$ degrees of freedom and is independent of b_0 and b_1. **(2.11)**

First, let us rewrite $(b_1 - \beta_1)/s\{b_1\}$ as follows:

$$\frac{b_1 - \beta_1}{\sigma\{b_1\}} \div \frac{s\{b_1\}}{\sigma\{b_1\}}$$

The numerator is a standard normal variable z. The nature of the denominator can be seen by first considering:

$$\frac{s^2\{b_1\}}{\sigma^2\{b_1\}} = \frac{\dfrac{MSE}{\sum(X_i - \bar{X})^2}}{\dfrac{\sigma^2}{\sum(X_i - \bar{X})^2}} = \frac{MSE}{\sigma^2} = \frac{\dfrac{SSE}{n-2}}{\sigma^2}$$

$$= \frac{SSE}{\sigma^2(n-2)} \sim \frac{\chi^2(n-2)}{n-2}$$

where the symbol $\sim$ stands for "is distributed as." The last step follows from (2.11). Hence, we have:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim \frac{z}{\sqrt{\dfrac{\chi^2(n-2)}{n-2}}}$$

But by theorem (2.11), z and χ^2 are independent since z is a function of b_1 and b_1 is independent of $SSE/\sigma^2 \sim \chi^2$. Hence, by (A.44), it follows that:

$$\frac{b_1 - \beta_1}{s\{b_1\}} \sim t(n-2)$$

This result places us in a position to readily make inferences concerning β_1. ■

Confidence Interval for β_1

Since $(b_1 - \beta_1)/s\{b_1\}$ follows a t distribution, we can make the following probability statement:

$$P\{t(\alpha/2; n-2) \le (b_1 - \beta_1)/s\{b_1\} \le t(1-\alpha/2; n-2)\} = 1 - \alpha \qquad \textbf{(2.12)}$$

Here, $t(\alpha/2; n-2)$ denotes the $(\alpha/2)100$ percentile of the t distribution with $n-2$ degrees of freedom. Because of the symmetry of the t distribution around its mean 0, it follows that:

$$t(\alpha/2; n-2) = -t(1-\alpha/2; n-2) \qquad \textbf{(2.13)}$$

Rearranging the inequalities in (2.12) and using (2.13), we obtain:

$$P\{b_1 - t(1-\alpha/2; n-2)s\{b_1\} \le \beta_1 \le b_1 + t(1-\alpha/2; n-2)s\{b_1\}\} = 1 - \alpha \qquad \textbf{(2.14)}$$

Since (2.14) holds for all possible values of β_1, the $1-\alpha$ confidence limits for β_1 are:

$$b_1 \pm t(1-\alpha/2; n-2)s\{b_1\} \qquad \textbf{(2.15)}$$

Example

Consider the Toluca Company example of Chapter 1. Management wishes an estimate of β_1 with 95 percent confidence coefficient. We summarize in Table 2.1 the needed results obtained earlier. First, we need to obtain $s\{b_1\}$:

$$s^2\{b_1\} = \frac{MSE}{\sum(X_i - \bar{X})^2} = \frac{2{,}384}{19{,}800} = .12040$$
$$s\{b_1\} = .3470$$

This estimated standard deviation is shown in the MINITAB output in Figure 2.2 in the column labeled Stdev corresponding to the row labeled X. Figure 2.2 repeats the MINITAB output presented earlier in Chapter 1 and contains some additional results that we will utilize shortly.

For a 95 percent confidence coefficient, we require $t(.975; 23)$. From Table B.2 in Appendix B, we find $t(.975; 23) = 2.069$. The 95 percent confidence interval, by (2.15), then is:

$$3.5702 - 2.069(.3470) \le \beta_1 \le 3.5702 + 2.069(.3470)$$
$$2.85 \le \beta_1 \le 4.29$$

Thus, with confidence coefficient .95, we estimate that the mean number of work hours increases by somewhere between 2.85 and 4.29 hours for each additional unit in the lot.

Comment

In Chapter 1, we noted that the scope of a regression model is restricted ordinarily to some range of values of the predictor variable. This is particularly important to keep in mind in using estimates of the slope β_1. In our Toluca Company example, a linear regression model appeared appropriate for lot sizes between 20 and 120, the range of the predictor variable in the recent past. It may not be

TABLE 2.1 Results for Toluca Company Example Obtained in Chapter 1.

$n = 25$	$\bar{X} = 70.00$
$b_0 = 62.37$	$b_1 = 3.5702$
$\hat{Y} = 62.37 + 3.5702X$	$SSE = 54{,}825$
$\sum(X_i - \bar{X})^2 = 19{,}800$	$MSE = 2{,}384$
$\sum(Y_i - \bar{Y})^2 = 307{,}203$	

FIGURE 2.2 Portion of MINITAB Regression Output—Toluca Company Example.

```
The regression equation is
Y = 62.4 + 3.57 X

Predictor       Coef       Stdev     t-ratio         p
Constant       62.37       26.18        2.38     0.026
X             3.5702      0.3470       10.29     0.000

s = 48.82       R-sq = 82.2%      R-sq(adj) = 81.4%

Analysis of Variance

SOURCE       DF          SS          MS         F         p
Regression    1      252378      252378    105.88     0.000
Error        23       54825        2384
Total        24      307203
```

reasonable to use the estimate of the slope to infer the effect of lot size on number of work hours far outside this range since the regression relation may not be linear there. ■

Tests Concerning β_1

Since $(b_1 - \beta_1)/s\{b_1\}$ is distributed as t with $n - 2$ degrees of freedom, tests concerning β_1 can be set up in ordinary fashion using the t distribution.

Example 1

Two-Sided Test A cost analyst in the Toluca Company is interested in testing, using regression model (2.1), whether or not there is a linear association between work hours and lot size, i.e., whether or not $\beta_1 = 0$. The two alternatives then are:

$$\begin{aligned} H_0&: \beta_1 = 0 \\ H_a&: \beta_1 \neq 0 \end{aligned} \tag{2.16}$$

The analyst wishes to control the risk of a Type I error at $\alpha = .05$. The conclusion H_a could be reached at once by referring to the 95 percent confidence interval for β_1 constructed earlier, since this interval does not include 0.

An explicit test of the alternatives (2.16) is based on the test statistic:

$$t^* = \frac{b_1}{s\{b_1\}} \tag{2.17}$$

The decision rule with this test statistic for controlling the level of significance at α is:

$$\begin{aligned} &\text{If } |t^*| \leq t(1 - \alpha/2; n - 2), \text{ conclude } H_0 \\ &\text{If } |t^*| > t(1 - \alpha/2; n - 2), \text{ conclude } H_a \end{aligned} \tag{2.18}$$

For the Toluca Company example, where $\alpha = .05$, $b_1 = 3.5702$, and $s\{b_1\} = .3470$, we require $t(.975; 23) = 2.069$. Thus, the decision rule for testing alternatives (2.16) is:

$$\begin{aligned} &\text{If } |t^*| \leq 2.069, \text{ conclude } H_0 \\ &\text{If } |t^*| > 2.069, \text{ conclude } H_a \end{aligned}$$

Since $|t^*| = |3.5702/.3470| = 10.29 > 2.069$, we conclude H_a, that $\beta_1 \neq 0$ or that there is a linear association between work hours and lot size. The value of the test statistic, $t^* = 10.29$, is shown in the MINITAB output in Figure 2.2 in the column labeled t-ratio and the row labeled X.

The two-sided P-value for the sample outcome is obtained by first finding the one-sided P-value, $P\{t(23) > t^* = 10.29\}$. We see from Table B.2 that this probability is less than .0005. Many statistical calculators and computer packages will provide the actual probability; it is almost 0, denoted by 0+. Thus, the two-sided P-value is $2(0+) = 0+$. Since the two-sided P-value is less than the specified level of significance $\alpha = .05$, we could conclude H_a directly. The MINITAB output in Figure 2.2 shows the P-value in the column labeled p, corresponding to the row labeled X. It is shown as 0.000.

Comment

When the test of whether or not $\beta_1 = 0$ leads to the conclusion that $\beta_1 \neq 0$, the association between Y and X is sometimes described to be a linear statistical association. ■

Example 2

One-Sided Test Suppose the analyst had wished to test whether or not β_1 is positive, controlling the level of significance at $\alpha = .05$. The alternatives then would be:

$$\begin{aligned} H_0&: \beta_1 \leq 0 \\ H_a&: \beta_1 > 0 \end{aligned}$$

and the decision rule based on test statistic (2.17) would be:

$$\text{If } t^* \le t(1-\alpha; n-2), \text{ conclude } H_0$$
$$\text{If } t^* > t(1-\alpha; n-2), \text{ conclude } H_a$$

For $\alpha = .05$, we require $t(.95; 23) = 1.714$. Since $t^* = 10.29 > 1.714$, we would conclude H_a, that β_1 is positive.

This same conclusion could be reached directly from the one-sided P-value, which was noted in Example 1 to be 0+. Since this P-value is less than .05, we would conclude H_a.

Comments

1. The P-value is sometimes called the observed level of significance.

2. Many scientific publications commonly report the P-value together with the value of the test statistic. In this way, one can conduct a test at any desired level of significance α by comparing the P-value with the specified level α.

3. Users of statistical calculators and computer packages need to be careful to ascertain whether one-sided or two-sided P-values are reported. Many commonly used labels, such as PROB or P, do not reveal whether the P-value is one- or two-sided.

4. Occasionally, it is desired to test whether or not β_1 equals some specified nonzero value β_{10}, which may be a historical norm, the value for a comparable process, or an engineering specification. The alternatives now are:

$$\begin{aligned} H_0&: \beta_1 = \beta_{10} \\ H_a&: \beta_1 \ne \beta_{10} \end{aligned} \tag{2.19}$$

and the appropriate test statistic is:

$$t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}} \tag{2.20}$$

The decision rule to be employed here still is (2.18), but it is now based on t^* defined in (2.20).

Note that test statistic (2.20) simplifies to test statistic (2.17) when the test involves $H_0: \beta_1 = \beta_{10} = 0$. ■

2.2 Inferences Concerning β_0

As noted in Chapter 1, there are only infrequent occasions when we wish to make inferences concerning β_0, the intercept of the regression line. These occur when the scope of the model includes $X = 0$.

Sampling Distribution of b_0

The point estimator b_0 was given in (1.10b) as follows:

$$b_0 = \bar{Y} - b_1 \bar{X} \tag{2.21}$$

The sampling distribution of b_0 refers to the different values of b_0 that would be obtained with repeated sampling when the levels of the predictor variable X are held constant from

sample to sample.

For regression model (2.1), the sampling distribution of b_0 is normal, with mean and variance: **(2.22)**

$$E\{b_0\} = \beta_0 \quad \textbf{(2.22a)}$$

$$\sigma^2\{b_0\} = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] \quad \textbf{(2.22b)}$$

The normality of the sampling distribution of b_0 follows because b_0, like b_1, is a linear combination of the observations Y_i. The results for the mean and variance of the sampling distribution of b_0 can be obtained in similar fashion as those for b_1.

An estimator of $\sigma^2\{b_0\}$ is obtained by replacing σ^2 by its point estimator MSE:

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] \quad \textbf{(2.23)}$$

The positive square root, $s\{b_0\}$, is an estimator of $\sigma\{b_0\}$.

Sampling Distribution of $(b_0 - \beta_0)/s\{b_0\}$

Analogous to theorem (2.10) for b_1, a theorem for b_0 states:

$$\frac{b_0 - \beta_0}{s\{b_0\}} \text{ is distributed as } t(n-2) \text{ for regression model (2.1)} \quad \textbf{(2.24)}$$

Hence, confidence intervals for β_0 and tests concerning β_0 can be set up in ordinary fashion, using the t distribution.

Confidence Interval for β_0

The $1 - \alpha$ confidence limits for β_0 are obtained in the same manner as those for β_1 derived earlier. They are:

$$b_0 \pm t(1 - \alpha/2; n-2)s\{b_0\} \quad \textbf{(2.25)}$$

Example

As noted earlier, the scope of the model for the Toluca Company example does not extend to lot sizes of $X = 0$. Hence, the regression parameter β_0 may not have intrinsic meaning here. If, nevertheless, a 90 percent confidence interval for β_0 were desired, we would proceed by finding $t(.95; 23)$ and $s\{b_0\}$. From Table B.2, we find $t(.95; 23) = 1.714$. Using the earlier results summarized in Table 2.1, we obtain by (2.23):

$$s^2\{b_0\} = MSE\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right] = 2{,}384\left[\frac{1}{25} + \frac{(70.00)^2}{19{,}800}\right] = 685.34$$

or:

$$s\{b_0\} = 26.18$$

The MINITAB output in Figure 2.2 shows this estimated standard deviation in the column labeled Stdev and the row labeled Constant.

The 90 percent confidence interval for β_0 is:

$$62.37 - 1.714(26.18) \le \beta_0 \le 62.37 + 1.714(26.18)$$

$$17.5 \le \beta_0 \le 107.2$$

We caution again that this confidence interval does not necessarily provide meaningful information. For instance, it does not necessarily provide information about the "setup" cost (the cost incurred in setting up the production process for the part) since we are not certain whether a linear regression model is appropriate when the scope of the model is extended to $X = 0$.

2.3 Some Considerations on Making Inferences Concerning β_0 and β_1

Effects of Departures from Normality

If the probability distributions of Y are not exactly normal but do not depart seriously, the sampling distributions of b_0 and b_1 will be approximately normal, and the use of the t distribution will provide approximately the specified confidence coefficient or level of significance. Even if the distributions of Y are far from normal, the estimators b_0 and b_1 generally have the property of *asymptotic normality*—their distributions approach normality under very general conditions as the sample size increases. Thus, with sufficiently large samples, the confidence intervals and decision rules given earlier still apply even if the probability distributions of Y depart far from normality. For large samples, the t value is, of course, replaced by the z value for the standard normal distribution.

Interpretation of Confidence Coefficient and Risks of Errors

Since regression model (2.1) assumes that the X_i are known constants, the confidence coefficient and risks of errors are interpreted with respect to taking repeated samples in which the X observations are kept at the same levels as in the observed sample. For instance, we constructed a confidence interval for β_1 with confidence coefficient .95 in the Toluca Company example. This coefficient is interpreted to mean that if many independent samples are taken where the levels of X (the lot sizes) are the same as in the data set and a 95 percent confidence interval is constructed for each sample, 95 percent of the intervals will contain the true value of β_1.

Spacing of the X Levels

Inspection of formulas (2.3b) and (2.22b) for the variances of b_1 and b_0, respectively, indicates that for given n and σ^2 these variances are affected by the spacing of the X levels in the observed data. For example, the greater is the spread in the X levels, the larger is the quantity $\sum(X_i - \bar{X})^2$ and the smaller is the variance of b_1. We discuss in Chapter 4 how the X observations should be spaced in experiments where spacing can be controlled.

Power of Tests

The power of tests on β_0 and β_1 can be obtained from Appendix Table B.5. Consider, for example, the general test concerning β_1 in (2.19):

$$H_0: \beta_1 = \beta_{10}$$
$$H_a: \beta_1 \neq \beta_{10}$$

for which test statistic (2.20) is employed:

$$t^* = \frac{b_1 - \beta_{10}}{s\{b_1\}}$$

and the decision rule for level of significance α is given in (2.18):

$$\text{If } |t^*| \le t(1-\alpha/2; n-2), \text{ conclude } H_0$$
$$\text{If } |t^*| > t(1-\alpha/2; n-2), \text{ conclude } H_a$$

The power of this test is the probability that the decision rule will lead to conclusion H_a when H_a in fact holds. Specifically, the power is given by:

$$\text{Power} = P\{|t^*| > t(1-\alpha/2; n-2) \mid \delta\} \tag{2.26}$$

where δ is the *noncentrality measure*—i.e., a measure of how far the true value of β_1 is from β_{10}:

$$\delta = \frac{|\beta_1 - \beta_{10}|}{\sigma\{b_1\}} \tag{2.27}$$

Table B.5 presents the power of the two-sided t test for $\alpha = .05$ and $\alpha = .01$, for various degrees of freedom df. To illustrate the use of this table, let us return to the Toluca Company example where we tested:

$$H_0: \beta_1 = \beta_{10} = 0$$
$$H_a: \beta_1 \ne \beta_{10} = 0$$

Suppose we wish to know the power of the test when $\beta_1 = 1.5$. To ascertain this, we need to know σ^2, the variance of the error terms. Assume, based on prior information or pilot data, that a reasonable planning value for the unknown variance is $\sigma^2 = 2{,}500$, so $\sigma^2\{b_1\}$ for our example would be:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum(X_i - \bar{X})^2} = \frac{2{,}500}{19{,}800} = .1263$$

or $\sigma\{b_1\} = .3553$. Then $\delta = |1.5 - 0| \div .3553 = 4.22$. We enter Table B.5 for $\alpha = .05$ (the level of significance used in the test) and 23 degrees of freedom and interpolate linearly between $\delta = 4.00$ and $\delta = 5.00$. We obtain:

$$.97 + \frac{4.22 - 4.00}{5.00 - 4.00}(1.00 - .97) = .9766$$

Thus, if $\beta_1 = 1.5$, the probability would be about .98 that we would be led to conclude H_a ($\beta_1 \ne 0$). In other words, if $\beta_1 = 1.5$, we would be almost certain to conclude that there is a linear relation between work hours and lot size.

The power of tests concerning β_0 can be obtained from Table B.5 in completely analogous fashion. For one-sided tests, Table B.5 should be entered so that one-half the level of significance shown there is the level of significance of the one-sided test.

2.4 Interval Estimation of $E\{Y_h\}$

A common objective in regression analysis is to estimate the mean for one or more probability distributions of Y. Consider, for example, a study of the relation between level of piecework pay (X) and worker productivity (Y). The mean productivity at high and medium levels of piecework pay may be of particular interest for purposes of analyzing the benefits obtained from an increase in the pay. As another example, the Toluca Company was interested in the mean response (mean number of work hours) for a range of lot sizes for purposes of finding the optimum lot size.

Let X_h denote the level of X for which we wish to estimate the mean response. X_h may be a value which occurred in the sample, or it may be some other value of the predictor variable within the scope of the model. The mean response when $X = X_h$ is denoted by $E\{Y_h\}$. Formula (1.12) gives us the point estimator $\hat{Y}_h$ of $E\{Y_h\}$:

$$\hat{Y}_h = b_0 + b_1 X_h \tag{2.28}$$

We consider now the sampling distribution of $\hat{Y}_h$.

Sampling Distribution of $\hat{Y}_h$

The sampling distribution of $\hat{Y}_h$, like the earlier sampling distributions discussed, refers to the different values of $\hat{Y}_h$ that would be obtained if repeated samples were selected, each holding the levels of the predictor variable X constant, and calculating $\hat{Y}_h$ for each sample.

For normal error regression model (2.1), the sampling distribution of $\hat{Y}_h$ is normal, with mean and variance: **(2.29)**

$$E\{\hat{Y}_h\} = E\{Y_h\} \tag{2.29a}$$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{2.29b}$$

Normality. The normality of the sampling distribution of $\hat{Y}_h$ follows directly from the fact that $\hat{Y}_h$, like b_0 and b_1, is a linear combination of the observations Y_i.

Mean. Note from (2.29a) that $\hat{Y}_h$ is an unbiased estimator of $E\{Y_h\}$. To prove this, we proceed as follows:

$$E\{\hat{Y}_h\} = E\{b_0 + b_1 X_h\} = E\{b_0\} + X_h E\{b_1\} = \beta_0 + \beta_1 X_h$$

by (2.3a) and (2.22a).

Variance. Note from (2.29b) that the variability of the sampling distribution of $\hat{Y}_h$ is affected by how far X_h is from $\bar{X}$, through the term $(X_h - \bar{X})^2$. The further from $\bar{X}$ is X_h, the greater is the quantity $(X_h - \bar{X})^2$ and the larger is the variance of $\hat{Y}_h$. An intuitive explanation of this effect is found in Figure 2.3. Shown there are two sample regression lines, based on two samples for the same set of X values. The two regression lines are assumed to go through the same $(\bar{X}, \bar{Y})$ point to isolate the effect of interest, namely, the effect of variation in the estimated slope b_1 from sample to sample. Note that at X_1, near $\bar{X}$, the fitted values $\hat{Y}_1$ for the two sample regression lines are close to each other. At X_2, which is far from $\bar{X}$, the situation is different. Here, the fitted values $\hat{Y}_2$ differ substantially.

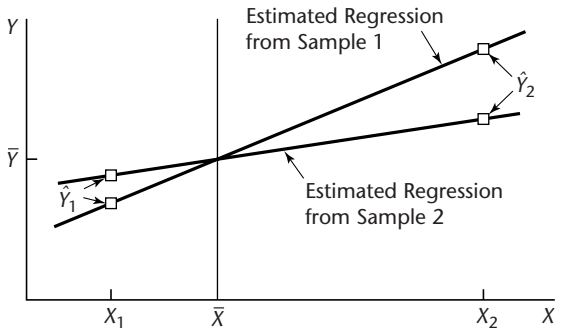


FIGURE 2.3
Effect on $\hat{Y}_h$ of Variation in b_1 from Sample to Sample in Two Samples with Same Means $\bar{Y}$ and $\bar{X}$.

Thus, variation in the slope b_1 from sample to sample has a much more pronounced effect on $\hat{Y}_h$ for X levels far from the mean $\bar{X}$ than for X levels near $\bar{X}$. Hence, the variation in the $\hat{Y}_h$ values from sample to sample will be greater when X_h is far from the mean than when X_h is near the mean.

When MSE is substituted for σ^2 in (2.29b), we obtain $s^2\{\hat{Y}_h\}$, the estimated variance of $\hat{Y}_h$:

$$s^2\{\hat{Y}_h\} = MSE\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{2.30}$$

The estimated standard deviation of $\hat{Y}_h$ is then $s\{\hat{Y}_h\}$, the positive square root of $s^2\{\hat{Y}_h\}$.

Comments

1. When $X_h = 0$, the variance of $\hat{Y}_h$ in (2.29b) reduces to the variance of b_0 in (2.22b). Similarly, $s^2\{\hat{Y}_h\}$ in (2.30) reduces to $s^2\{b_0\}$ in (2.23). The reason is that $\hat{Y}_h = b_0$ when $X_h = 0$ since $\hat{Y}_h = b_0 + b_1 X_h$.

2. To derive $\sigma^2\{\hat{Y}_h\}$, we first show that b_1 and $\bar{Y}$ are uncorrelated and, hence, for regression model (2.1), independent:

$$\sigma\{\bar{Y}, b_1\} = 0 \tag{2.31}$$

where $\sigma\{\bar{Y}, b_1\}$ denotes the covariance between $\bar{Y}$ and b_1. We begin with the definitions:

$$\bar{Y} = \sum\left(\frac{1}{n}\right) Y_i \qquad b_1 = \sum k_i Y_i$$

where k_i is as defined in (2.4a). We now use (A.32), with $a_i = 1/n$ and $c_i = k_i$; remember that the Y_i are independent random variables:

$$\sigma\{\bar{Y}, b_1\} = \sum\left(\frac{1}{n}\right) k_i \sigma^2\{Y_i\} = \frac{\sigma^2}{n}\sum k_i$$

But we know from (2.5) that $\sum k_i = 0$. Hence, the covariance is 0.

Now we are ready to find the variance of $\hat{Y}_h$. We shall use the estimator in the alternative form (1.15):

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\{\bar{Y} + b_1(X_h - \bar{X})\}$$

Since $\bar{Y}$ and b_1 are independent and X_h and $\bar{X}$ are constants, we obtain:

$$\sigma^2\{\hat{Y}_h\} = \sigma^2\{\bar{Y}\} + (X_h - \bar{X})^2\sigma^2\{b_1\}$$

Now $\sigma^2\{b_1\}$ is given in (2.3b), and:

$$\sigma^2\{\bar{Y}\} = \frac{\sigma^2\{Y_i\}}{n} = \frac{\sigma^2}{n}$$

Hence:

$$\sigma^2\{\hat{Y}_h\} = \frac{\sigma^2}{n} + (X_h - \bar{X})^2\frac{\sigma^2}{\sum(X_i - \bar{X})^2}$$

which, upon a slight rearrangement of terms, yields (2.29b). ■

Sampling Distribution of $(\hat{Y}_h - E\{Y_h\})/s\{\hat{Y}_h\}$

Since we have encountered the t distribution in each type of inference for regression model (2.1) up to this point, it should not be surprising that:

$$\frac{\hat{Y}_h - E\{Y_h\}}{s\{\hat{Y}_h\}} \text{ is distributed as } t(n-2) \text{ for regression model (2.1)} \qquad \textbf{(2.32)}$$

Hence, all inferences concerning $E\{Y_h\}$ are carried out in the usual fashion with the t distribution. We illustrate the construction of confidence intervals, since in practice these are used more frequently than tests.

Confidence Interval for $E\{Y_h\}$

A confidence interval for $E\{Y_h\}$ is constructed in the standard fashion, making use of the t distribution as indicated by theorem (2.32). The $1 - \alpha$ confidence limits are:

$$\hat{Y}_h \pm t(1 - \alpha/2; n-2)s\{\hat{Y}_h\} \qquad \textbf{(2.33)}$$

Example 1

Returning to the Toluca Company example, let us find a 90 percent confidence interval for $E\{Y_h\}$ when the lot size is $X_h = 65$ units. Using the earlier results in Table 2.1, we find the point estimate $\hat{Y}_h$:

$$\hat{Y}_h = 62.37 + 3.5702(65) = 294.4$$

Next, we need to find the estimated standard deviation $s\{\hat{Y}_h\}$. We obtain, using (2.30):

$$s^2\{\hat{Y}_h\} = 2{,}384\left[\frac{1}{25} + \frac{(65 - 70.00)^2}{19{,}800}\right] = 98.37$$

$$s\{\hat{Y}_h\} = 9.918$$

For a 90 percent confidence coefficient, we require $t(.95; 23) = 1.714$. Hence, our confidence interval with confidence coefficient .90 is by (2.33):

$$294.4 - 1.714(9.918) \le E\{Y_h\} \le 294.4 + 1.714(9.918)$$

$$277.4 \le E\{Y_h\} \le 311.4$$

We conclude with confidence coefficient .90 that the mean number of work hours required when lots of 65 units are produced is somewhere between 277.4 and 311.4 hours. We see that our estimate of the mean number of work hours is moderately precise.

Example 2

Suppose the Toluca Company wishes to estimate $E\{Y_h\}$ for lots with $X_h = 100$ units with a 90 percent confidence interval. We require:

$$\hat{Y}_h = 62.37 + 3.5702(100) = 419.4$$
$$s^2\{\hat{Y}_h\} = 2{,}384\left[\frac{1}{25} + \frac{(100 - 70.00)^2}{19{,}800}\right] = 203.72$$
$$s\{\hat{Y}_h\} = 14.27$$
$$t(.95; 23) = 1.714$$

Hence, the 90 percent confidence interval is:

$$419.4 - 1.714(14.27) \leq E\{Y_h\} \leq 419.4 + 1.714(14.27)$$
$$394.9 \leq E\{Y_h\} \leq 443.9$$

Note that this confidence interval is somewhat wider than that for Example 1, since the X_h level here ($X_h = 100$) is substantially farther from the mean $\bar{X} = 70.0$ than the X_h level for Example 1 ($X_h = 65$).

Comments

1. Since the X_i are known constants in regression model (2.1), the interpretation of confidence intervals and risks of errors in inferences on the mean response is in terms of taking repeated samples in which the X observations are at the same levels as in the actual study. We noted this same point in connection with inferences on β_0 and β_1.
2. We see from formula (2.29b) that, for given sample results, the variance of $\hat{Y}_h$ is smallest when $X_h = \bar{X}$. Thus, in an experiment to estimate the mean response at a particular level X_h of the predictor variable, the precision of the estimate will be greatest if (everything else remaining equal) the observations on X are spaced so that $\bar{X} = X_h$.
3. The usual relationship between confidence intervals and tests applies in inferences concerning the mean response. Thus, the two-sided confidence limits (2.33) can be utilized for two-sided tests concerning the mean response at X_h. Alternatively, a regular decision rule can be set up.
4. The confidence limits (2.33) for a mean response $E\{Y_h\}$ are not sensitive to moderate departures from the assumption that the error terms are normally distributed. Indeed, the limits are not sensitive to substantial departures from normality if the sample size is large. This robustness in estimating the mean response is related to the robustness of the confidence limits for β_0 and β_1, noted earlier.
5. Confidence limits (2.33) apply when a single mean response is to be estimated from the study. We discuss in Chapter 4 how to proceed when several mean responses are to be estimated from the same data. ■

2.5 Prediction of New Observation

We consider now the prediction of a new observation Y corresponding to a given level X of the predictor variable. Three illustrations where prediction of a new observation is needed follow.

1. In the Toluca Company example, the next lot to be produced consists of 100 units and management wishes to predict the number of work hours for this particular lot.

2. An economist has estimated the regression relation between company sales and number of persons 16 or more years old from data for the past 10 years. Using a reliable demographic projection of the number of persons 16 or more years old for next year, the economist wishes to predict next year's company sales.
3. An admissions officer at a university has estimated the regression relation between the high school grade point average (GPA) of admitted students and the first-year college GPA. The officer wishes to predict the first-year college GPA for an applicant whose high school GPA is 3.5 as part of the information on which an admissions decision will be based.

The new observation on Y to be predicted is viewed as the result of a new trial, independent of the trials on which the regression analysis is based. We denote the level of X for the new trial as X_h and the new observation on Y as $Y_{h(\text{new})}$. Of course, we assume that the underlying regression model applicable for the basic sample data continues to be appropriate for the new observation.

The distinction between estimation of the mean response $E\{Y_h\}$, discussed in the preceding section, and prediction of a new response $Y_{h(\text{new})}$, discussed now, is basic. In the former case, we estimate the *mean* of the distribution of Y. In the present case, we predict an *individual outcome* drawn from the distribution of Y. Of course, the great majority of individual outcomes deviate from the mean response, and this must be taken into account by the procedure for predicting $Y_{h(\text{new})}$.

Prediction Interval for $Y_{h(\text{new})}$ when Parameters Known

To illustrate the nature of a *prediction interval* for a new observation $Y_{h(\text{new})}$ in as simple a fashion as possible, we shall first assume that all regression parameters are known. Later we drop this assumption and make appropriate modifications.

Suppose that in the college admissions example the relevant parameters of the regression model are known to be:

$$\beta_0 = .10 \qquad \beta_1 = .95$$
$$E\{Y\} = .10 + .95X$$
$$\sigma = .12$$

The admissions officer is considering an applicant whose high school GPA is $X_h = 3.5$. The mean college GPA for students whose high school average is 3.5 is:

$$E\{Y_h\} = .10 + .95(3.5) = 3.425$$

Figure 2.4 shows the probability distribution of Y for $X_h = 3.5$. Its mean is $E\{Y_h\} = 3.425$, and its standard deviation is $\sigma = .12$. Further, the distribution is normal in accord with regression model (2.1).

Suppose we were to predict that the college GPA of the applicant whose high school GPA is $X_h = 3.5$ will be between:

$$E\{Y_h\} \pm 3\sigma$$
$$3.425 \pm 3(.12)$$

so that the prediction interval would be:

$$3.065 \leq Y_{h(\text{new})} \leq 3.785$$

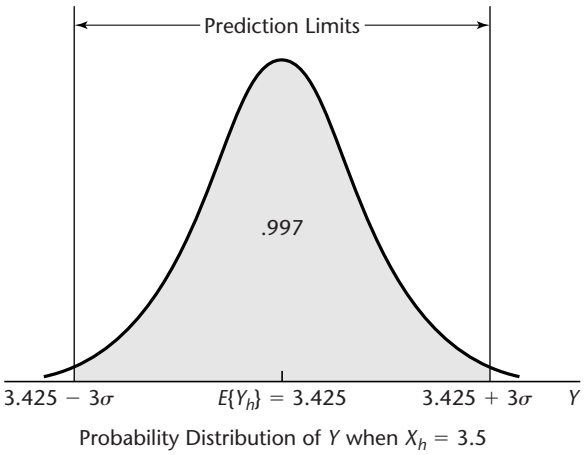


FIGURE 2.4
Prediction of $Y_{h(\text{new})}$ when Parameters Known.

Since 99.7 percent of the area in a normal probability distribution falls within three standard deviations from the mean, the probability is .997 that this prediction interval will give a correct prediction for the applicant with high school GPA of 3.5. While the prediction limits here are rather wide, so that the prediction is not too precise, the prediction interval does indicate to the admissions officer that the applicant is expected to attain at least a 3.0 GPA in the first year of college.

The basic idea of a prediction interval is thus to choose a range in the distribution of Y wherein most of the observations will fall, and then to declare that the next observation will fall in this range. The usefulness of the prediction interval depends, as always, on the width of the interval and the needs for precision by the user.

In general, when the regression parameters of normal error regression model (2.1) are known, the $1 - \alpha$ prediction limits for $Y_{h(\text{new})}$ are:

$$E\{Y_h\} \pm z(1 - \alpha/2)\sigma \tag{2.34}$$

In centering the limits around $E\{Y_h\}$, we obtain the narrowest interval consistent with the specified probability of a correct prediction.

Prediction Interval for $Y_{h(\text{new})}$ when Parameters Unknown

When the regression parameters are unknown, they must be estimated. The mean of the distribution of Y is estimated by $\hat{Y}_h$, as usual, and the variance of the distribution of Y is estimated by MSE. We cannot, however, simply use the prediction limits (2.34) with the parameters replaced by the corresponding point estimators. The reason is illustrated intuitively in Figure 2.5. Shown there are two probability distributions of Y, corresponding to the upper and lower limits of a confidence interval for $E\{Y_h\}$. In other words, the distribution of Y could be located as far left as the one shown, as far right as the other one shown, or anywhere in between. Since we do not know the mean $E\{Y_h\}$ and only estimate it by a confidence interval, we cannot be certain of the location of the distribution of Y.

Figure 2.5 also shows the prediction limits for each of the two probability distributions of Y presented there. Since we cannot be certain of the location of the distribution

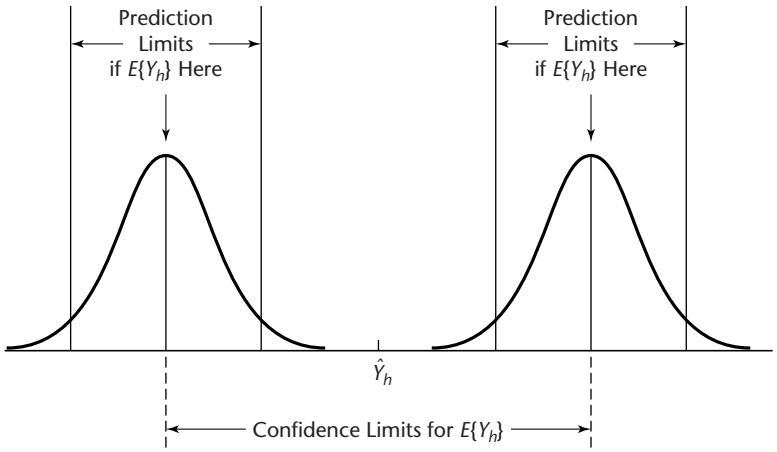


FIGURE 2.5
Prediction of $Y_{h(\text{new})}$ when Parameters Unknown.

of Y, prediction limits for $Y_{h(\text{new})}$ clearly must take account of two elements, as shown in Figure 2.5:

1. Variation in possible location of the distribution of Y.
2. Variation within the probability distribution of Y.

Prediction limits for a new observation $Y_{h(\text{new})}$ at a given level X_h are obtained by means of the following theorem:

$$\frac{Y_{h(\text{new})} - \hat{Y}_h}{s\{\text{pred}\}} \text{ is distributed as } t(n-2) \text{ for normal error regression model (2.1)} \quad \textbf{(2.35)}$$

Note that the studentized statistic (2.35) uses the point estimator $\hat{Y}_h$ in the numerator rather than the true mean $E\{Y_h\}$ because the true mean is unknown and cannot be used in making a prediction. The estimated standard deviation of the prediction, $s\{\text{pred}\}$, in the denominator of the studentized statistic will be defined shortly.

From theorem (2.35), it follows in the usual fashion that the $1-\alpha$ prediction limits for a new observation $Y_{h(\text{new})}$ are (for instance, compare (2.35) to (2.10) and relate $\hat{Y}_h$ to b_1 and $Y_{h(\text{new})}$ to β_1):

$$\hat{Y}_h \pm t(1-\alpha/2; n-2)s\{\text{pred}\} \quad \textbf{(2.36)}$$

Note that the numerator of the studentized statistic (2.35) represents how far the new observation $Y_{h(\text{new})}$ will deviate from the estimated mean $\hat{Y}_h$ based on the original n cases in the study. This difference may be viewed as the prediction error, with $\hat{Y}_h$ serving as the best point estimate of the value of the new observation $Y_{h(\text{new})}$. The variance of this prediction error can be readily obtained by utilizing the independence of the new observation $Y_{h(\text{new})}$ and the original n sample cases on which $\hat{Y}_h$ is based. We denote the variance of the prediction error by $\sigma^2\{\text{pred}\}$, and we obtain by (A.31b):

$$\sigma^2\{\text{pred}\} = \sigma^2\{Y_{h(\text{new})} - \hat{Y}_h\} = \sigma^2\{Y_{h(\text{new})}\} + \sigma^2\{\hat{Y}_h\} = \sigma^2 + \sigma^2\{\hat{Y}_h\} \quad \textbf{(2.37)}$$

Note that $\sigma^2\{\text{pred}\}$ has two components:

1. The variance of the distribution of Y at $X = X_h$, namely σ^2.
2. The variance of the sampling distribution of $\hat{Y}_h$, namely $\sigma^2\{\hat{Y}_h\}$.

An unbiased estimator of $\sigma^2\{\text{pred}\}$ is:

$$s^2\{\text{pred}\} = MSE + s^2\{\hat{Y}_h\} \tag{2.38}$$

which can be expressed as follows, using (2.30):

$$s^2\{\text{pred}\} = MSE\left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{2.38a}$$

Example

The Toluca Company studied the relationship between lot size and work hours primarily to obtain information on the mean work hours required for different lot sizes for use in determining the optimum lot size. The company was also interested, however, to see whether the regression relationship is useful for predicting the required work hours for individual lots. Suppose that the next lot to be produced consists of $X_h = 100$ units and that a 90 percent prediction interval is desired. We require $t(.95; 23) = 1.714$. From earlier work, we have:

$$\hat{Y}_h = 419.4 \qquad s^2\{\hat{Y}_h\} = 203.72 \qquad MSE = 2{,}384$$

Using (2.38), we obtain:

$$s^2\{\text{pred}\} = 2{,}384 + 203.72 = 2{,}587.72$$
$$s\{\text{pred}\} = 50.87$$

Hence, the 90 percent prediction interval for $Y_{h(\text{new})}$ is by (2.36):

$$419.4 - 1.714(50.87) \le Y_{h(\text{new})} \le 419.4 + 1.714(50.87)$$
$$332.2 \le Y_{h(\text{new})} \le 506.6$$

With confidence coefficient .90, we predict that the number of work hours for the next production run of 100 units will be somewhere between 332 and 507 hours.

This prediction interval is rather wide and may not be too useful for planning worker requirements for the next lot. The interval can still be useful for control purposes, though. For instance, suppose that the actual work hours on the next lot of 100 units were 550 hours. Since the actual work hours fall outside the prediction limits, management would have an indication that a change in the production process may have occurred and would be alerted to the possible need for remedial action.

Note that the primary reason for the wide prediction interval is the large lot-to-lot variability in work hours for any given lot size; $MSE = 2{,}384$ accounts for 92 percent of the estimated prediction variance $s^2\{\text{pred}\} = 2{,}587.72$. It may be that the large lot-to-lot variability reflects other factors that affect the required number of work hours besides lot size, such as the amount of experience of employees assigned to the lot production. If so, a multiple regression model incorporating these other factors might lead to much more precise predictions. Alternatively, a designed experiment could be conducted to determine the main factors leading to the large lot-to-lot variation. A quality improvement program would then use these findings to achieve more uniform performance, for example, by additional training of employees if inadequate training accounted for much of the variability.

Comments

1. The 90 percent prediction interval for $Y_{h(\text{new})}$ obtained in the Toluca Company example is wider than the 90 percent confidence interval for $E\{Y_h\}$ obtained in Example 2 on page 55. The reason is that when predicting the work hours required for a new lot, we encounter both the variability in $\hat{Y}_h$ from sample to sample as well as the lot-to-lot variation within the probability distribution of Y.

2. Formula (2.38a) indicates that the prediction interval is wider the further X_h is from $\bar{X}$. The reason for this is that the estimate of the mean $\hat{Y}_h$, as noted earlier, is less precise as X_h is located farther away from $\bar{X}$.

3. The prediction limits (2.36), unlike the confidence limits (2.33) for a mean response $E\{Y_h\}$, are sensitive to departures from normality of the error terms distribution. In Chapter 3, we discuss diagnostic procedures for examining the nature of the probability distribution of the error terms, and we describe remedial measures if the departure from normality is serious.

4. The confidence coefficient for the prediction limits (2.36) refers to the taking of repeated samples based on the same set of X values, and calculating prediction limits for $Y_{h(\text{new})}$ for each sample.

5. Prediction limits (2.36) apply for a single prediction based on the sample data. Next, we discuss how to predict the mean of several new observations at a given X_h, and in Chapter 4 we take up how to make several predictions at different X_h levels.

6. Prediction intervals resemble confidence intervals. However, they differ conceptually. A confidence interval represents an inference on a parameter and is an interval that is intended to cover the value of the parameter. A prediction interval, on the other hand, is a statement about the value to be taken by a random variable, the new observation $Y_{h(\text{new})}$. ■

Prediction of Mean of *m* New Observations for Given X_h

Occasionally, one would like to predict the mean of m new observations on Y for a given level of the predictor variable. Suppose the Toluca Company has been asked to bid on a contract that calls for $m = 3$ production runs of $X_h = 100$ units during the next few months. Management would like to predict the mean work hours per lot for these three runs and then convert this into a prediction of the total work hours required to fill the contract.

We denote the mean of the new Y observations to be predicted as $\bar{Y}_{h(\text{new})}$. It can be shown that the appropriate $1 - \alpha$ prediction limits are, assuming that the new Y observations are independent:

$$\hat{Y}_h \pm t(1 - \alpha/2; n - 2)s\{\text{predmean}\} \tag{2.39}$$

where:

$$s^2\{\text{predmean}\} = \frac{MSE}{m} + s^2\{\hat{Y}_h\} \tag{2.39a}$$

or equivalently:

$$s^2\{\text{predmean}\} = MSE\left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right] \tag{2.39b}$$

Note from (2.39a) that the variance $s^2\{\text{predmean}\}$ has two components:

1. The variance of the mean of m observations from the probability distribution of Y at $X = X_h$.
2. The variance of the sampling distribution of $\hat{Y}_h$.

Example

In the Toluca Company example, let us find the 90 percent prediction interval for the mean number of work hours $\bar{Y}_{h(\text{new})}$ in three new production runs, each for $X_h = 100$ units. From previous work, we have:

$$\hat{Y}_h = 419.4 \qquad s^2\{\hat{Y}_h\} = 203.72$$
$$MSE = 2{,}384 \qquad t(.95; 23) = 1.714$$

Hence, we obtain:

$$s^2\{\text{predmean}\} = \frac{2{,}384}{3} + 203.72 = 998.4$$
$$s\{\text{predmean}\} = 31.60$$

The prediction interval for the mean work hours per lot then is:

$$419.4 - 1.714(31.60) \le \bar{Y}_{h(\text{new})} \le 419.4 + 1.714(31.60)$$
$$365.2 \le \bar{Y}_{h(\text{new})} \le 473.6$$

Note that these prediction limits are narrower than those for predicting the work hours for a single lot of 100 units because they involve a prediction of the mean work hours for three lots.

We obtain the prediction interval for the total number of work hours for the three lots by multiplying the prediction limits for $\bar{Y}_{h(\text{new})}$ by 3:

$$1{,}095.6 = 3(365.2) \le \text{ Total work hours} \le 3(473.6) = 1{,}420.8$$

Thus, it can be predicted with 90 percent confidence that between 1,096 and 1,421 work hours will be needed to fill the contract for three lots of 100 units each.

Comment

The 90 percent prediction interval for $\bar{Y}_{h(\text{new})}$, obtained for the Toluca Company example above, is narrower than that obtained for $Y_{h(\text{new})}$ on page 59, as expected. Furthermore, both of the prediction intervals are wider than the 90 percent confidence interval for $E\{Y_h\}$ obtained in Example 2 on page 55—also as expected. ■

2.6 Confidence Band for Regression Line

At times we would like to obtain a confidence band for the entire regression line $E\{Y\} = \beta_0 + \beta_1 X$. This band enables us to see the region in which the entire regression line lies. It is particularly useful for determining the appropriateness of a fitted regression function, as we explain in Chapter 3.

The Working-Hotelling $1 - \alpha$ confidence band for the regression line for regression model (2.1) has the following two boundary values at any level X_h:

$$\hat{Y}_h \pm W s\{\hat{Y}_h\} \qquad \textbf{(2.40)}$$

where:

$$W^2 = 2F(1 - \alpha; 2, n - 2) \qquad \textbf{(2.40a)}$$

and $\hat{Y}_h$ and $s\{\hat{Y}_h\}$ are defined in (2.28) and (2.30), respectively. Note that the formula for the boundary values is of exactly the same form as formula (2.33) for the confidence limits for the mean response at X_h, except that the t multiple has been replaced by the W

multiple. Consequently, the boundary points of the confidence band for the regression line are wider apart the further X_h is from the mean $\bar{X}$ of the X observations. The W multiple will be larger than the t multiple in (2.33) because the confidence band must encompass the entire regression line, whereas the confidence limits for $E\{Y_h\}$ at X_h apply only at the single level X_h.

Example

We wish to determine how precisely we have been able to estimate the regression function for the Toluca Company example by obtaining the 90 percent confidence band for the regression line. We illustrate the calculations of the boundary values of the confidence band when $X_h = 100$. We found earlier for this case:

$$\hat{Y}_h = 419.4 \qquad s\{\hat{Y}_h\} = 14.27$$

We now require:

$$W^2 = 2F(1-\alpha; 2, n-2) = 2F(.90; 2, 23) = 2(2.549) = 5.098$$
$$W = 2.258$$

Hence, the boundary values of the confidence band for the regression line at $X_h = 100$ are $419.4 \pm 2.258(14.27)$, and the confidence band there is:

$$387.2 \leq \beta_0 + \beta_1 X_h \leq 451.6 \qquad \text{for } X_h = 100$$

In similar fashion, we can calculate the boundary values for other values of X_h by obtaining $\hat{Y}_h$ and $s\{\hat{Y}_h\}$ for each X_h level from (2.28) and (2.30) and then finding the boundary values by means of (2.40). Figure 2.6 contains a plot of the confidence band for the regression line. Note that at $X_h = 100$, the boundary values are 387.2 and 451.6, as we calculated earlier.

We see from Figure 2.6 that the regression line for the Toluca Company example has been estimated fairly precisely. The slope of the regression line is clearly positive, and the levels of the regression line at different levels of X are estimated fairly precisely except for small and large lot sizes.

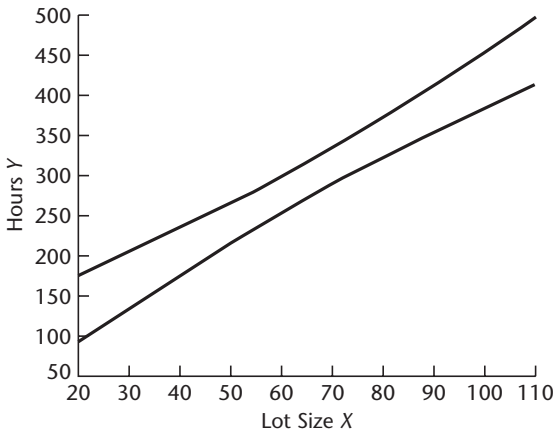


FIGURE 2.6 Confidence Band for Regression Line—Toluca Company Example.

Comments

1. The boundary values of the confidence band for the regression line in (2.40) define a hyperbola, as may be seen by replacing $\hat{Y}_h$ and $s\{\hat{Y}_h\}$ by their definitions in (2.28) and (2.30), respectively:

$$b_0 + b_1 X \pm W\sqrt{MSE}\left[\frac{1}{n} + \frac{(X - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]^{1/2} \quad \textbf{(2.41)}$$

2. The boundary values of the confidence band for the regression line at any value X_h often are not substantially wider than the confidence limits for the mean response at that single X_h level. In the Toluca Company example, the t multiple for estimating the mean response at $X_h = 100$ with a 90 percent confidence interval was $t(.95; 23) = 1.714$. This compares with the W multiple for the 90 percent confidence band for the entire regression line of $W = 2.258$. With the somewhat wider limits for the entire regression line, one is able to draw conclusions about any and all mean responses for the entire regression line and not just about the mean response at a given X level. Some uses of this broader base for inference will be explained in the next two chapters.

3. The confidence band (2.40) applies to the entire regression line over all real-numbered values of X from $-\infty$ to ∞. The confidence coefficient indicates the proportion of time that the estimating procedure will yield a band that covers the entire line, in a long series of samples in which the X observations are kept at the same level as in the actual study.

In applications, the confidence band is ignored for that part of the regression line which is not of interest in the problem at hand. In the Toluca Company example, for instance, negative lot sizes would be ignored. The confidence coefficient for a limited segment of the band of interest is somewhat higher than $1 - \alpha$, so $1 - \alpha$ serves then as a lower bound to the confidence coefficient.

4. Some alternative procedures for developing confidence bands for the regression line have been developed. The simplicity of the Working-Hotelling confidence band (2.40) arises from the fact that it is a direct extension of the confidence limits for a single mean response in (2.33). ■

2.7 Analysis of Variance Approach to Regression Analysis

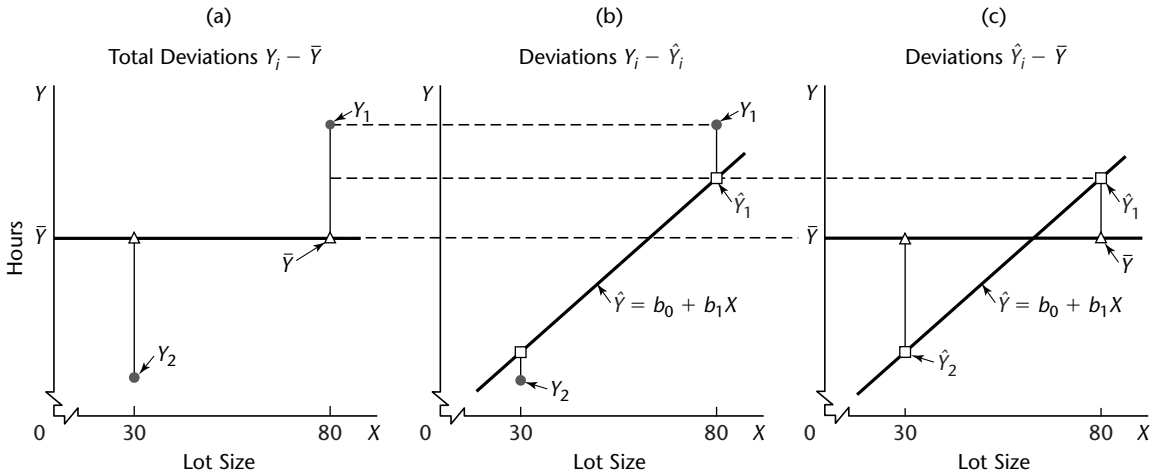
We now have developed the basic regression model and demonstrated its major uses. At this point, we consider the regression analysis from the perspective of analysis of variance. This new perspective will not enable us to do anything new, but the analysis of variance approach will come into its own when we take up multiple regression models and other types of linear statistical models.

Partitioning of Total Sum of Squares

Basic Notions. The analysis of variance approach is based on the partitioning of sums of squares and degrees of freedom associated with the response variable Y. To explain the motivation of this approach, consider again the Toluca Company example. Figure 2.7a shows the observations Y_i for the first two production runs presented in Table 1.1. Disregarding the lot sizes, we see that there is variation in the number of work hours Y_i, as in all statistical data. This variation is conventionally measured in terms of the deviations of the Y_i around their mean $\bar{Y}$:

$$Y_i - \bar{Y} \quad \textbf{(2.42)}$$

FIGURE 2.7 Illustration of Partitioning of Total Deviations $Y_i - \bar{Y}$—Toluca Company Example (not drawn to scale; only observations Y_1 and Y_2 are shown).



These deviations are shown by the vertical lines in Figure 2.7a. The measure of total variation, denoted by *SSTO*, is the sum of the squared deviations (2.42):

$$SSTO = \sum (Y_i - \bar{Y})^2 \tag{2.43}$$

Here *SSTO* stands for *total sum of squares*. If all Y_i observations are the same, $SSTO = 0$. The greater the variation among the Y_i observations, the larger is *SSTO*. Thus, *SSTO* for our example is a measure of the uncertainty pertaining to the work hours required for a lot, when the lot size is not taken into account.

When we utilize the predictor variable X, the variation reflecting the uncertainty concerning the variable Y is that of the Y_i observations around the fitted regression line:

$$Y_i - \hat{Y}_i \tag{2.44}$$

These deviations are shown by the vertical lines in Figure 2.7b. The measure of variation in the Y_i observations that is present when the predictor variable X is taken into account is the sum of the squared deviations (2.44), which is the familiar *SSE* of (1.21):

$$SSE = \sum (Y_i - \hat{Y}_i)^2 \tag{2.45}$$

Again, *SSE* denotes *error sum of squares*. If all Y_i observations fall on the fitted regression line, $SSE = 0$. The greater the variation of the Y_i observations around the fitted regression line, the larger is *SSE*.

For the Toluca Company example, we know from earlier work (Table 2.1) that:

$$SSTO = 307{,}203 \qquad SSE = 54{,}825$$

What accounts for the substantial difference between these two sums of squares? The difference, as we show shortly, is another sum of squares:

$$SSR = \sum (\hat{Y}_i - \bar{Y})^2 \tag{2.46}$$

where *SSR* stands for *regression sum of squares*. Note that *SSR* is a sum of squared deviations, the deviations being:

$$\hat{Y}_i - \bar{Y} \tag{2.47}$$

These deviations are shown by the vertical lines in Figure 2.7c. Each deviation is simply the difference between the fitted value on the regression line and the mean of the fitted values $\bar{Y}$. (Recall from (1.18) that the mean of the fitted values $\hat{Y}_i$ is $\bar{Y}$.) If the regression line is horizontal so that $\hat{Y}_i - \bar{Y} \equiv 0$, then $SSR = 0$. Otherwise, *SSR* is positive.

SSR may be considered a measure of that part of the variability of the Y_i which is associated with the regression line. The larger *SSR* is in relation to *SSTO*, the greater is the effect of the regression relation in accounting for the total variation in the Y_i observations.

For the Toluca Company example, we have:

$$SSR = SSTO - SSE = 307{,}203 - 54{,}825 = 252{,}378$$

which indicates that most of the total variability in work hours is accounted for by the relation between lot size and work hours.

Formal Development of Partitioning. The total deviation $Y_i - \bar{Y}$, used in the measure of the total variation of the observations Y_i without taking the predictor variable into account, can be decomposed into two components:

$$\underbrace{Y_i - \bar{Y}}_{\substack{\text{Total}\\ \text{deviation}}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\substack{\text{Deviation}\\ \text{of fitted}\\ \text{regression}\\ \text{value}\\ \text{around mean}}} + \underbrace{Y_i - \hat{Y}_i}_{\substack{\text{Deviation}\\ \text{around}\\ \text{fitted}\\ \text{regression}\\ \text{line}}} \tag{2.48}$$

The two components are:

1. The deviation of the fitted value $\hat{Y}_i$ around the mean $\bar{Y}$.
2. The deviation of the observation Y_i around the fitted regression line.

Figure 2.7 shows this decomposition for observation Y_1 by the broken lines.

It is a remarkable property that the sums of these squared deviations have the same relationship:

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 \tag{2.49}$$

or, using the notation in (2.43), (2.45), and (2.46):

$$SSTO = SSR + SSE \tag{2.50}$$

To prove this basic result in the analysis of variance, we proceed as follows:

$$\begin{aligned}\sum (Y_i - \bar{Y})^2 &= \sum [(\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)]^2 \\ &= \sum [(\hat{Y}_i - \bar{Y})^2 + (Y_i - \hat{Y}_i)^2 + 2(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)] \\ &= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)\end{aligned}$$

The last term on the right equals zero, as we can see by expanding it:

$$2\sum(\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = 2\sum \hat{Y}_i(Y_i - \hat{Y}_i) - 2\bar{Y}\sum(Y_i - \hat{Y}_i)$$

The first summation on the right equals zero by (1.20), and the second equals zero by (1.17). Hence, (2.49) follows.

Comment

The formulas for *SSTO*, *SSR*, and *SSE* given in (2.43), (2.45), and (2.46) are best for computational accuracy. Alternative formulas that are algebraically equivalent are available. One that is useful for deriving analytical results is:

$$SSR = b_1^2 \sum(X_i - \bar{X})^2 \tag{2.51}$$

■

Breakdown of Degrees of Freedom

Corresponding to the partitioning of the total sum of squares *SSTO*, there is a partitioning of the associated degrees of freedom (abbreviated *df*). We have $n - 1$ degrees of freedom associated with *SSTO*. One degree of freedom is lost because the deviations $Y_i - \bar{Y}$ are subject to one constraint: they must sum to zero. Equivalently, one degree of freedom is lost because the sample mean $\bar{Y}$ is used to estimate the population mean.

SSE, as noted earlier, has $n - 2$ degrees of freedom associated with it. Two degrees of freedom are lost because the two parameters β_0 and β_1 are estimated in obtaining the fitted values $\hat{Y}_i$.

SSR has one degree of freedom associated with it. Although there are n deviations $\hat{Y}_i - \bar{Y}$, all fitted values $\hat{Y}_i$ are calculated from the same estimated regression line. Two degrees of freedom are associated with a regression line, corresponding to the intercept and the slope of the line. One of the two degrees of freedom is lost because the deviations $\hat{Y}_i - \bar{Y}$ are subject to a constraint: they must sum to zero.

Note that the degrees of freedom are additive:

$$n - 1 = 1 + (n - 2)$$

For the Toluca Company example, these degrees of freedom are:

$$24 = 1 + 23$$

Mean Squares

A sum of squares divided by its associated degrees of freedom is called a *mean square* (abbreviated *MS*). For instance, an ordinary sample variance is a mean square since a sum of squares, $\sum(Y_i - \bar{Y})^2$, is divided by its associated degrees of freedom, $n - 1$. We are interested here in the *regression mean square,* denoted by *MSR*:

$$MSR = \frac{SSR}{1} = SSR \tag{2.52}$$

and in the *error mean square, MSE*, defined earlier in (1.22):

$$MSE = \frac{SSE}{n - 2} \tag{2.53}$$

For the Toluca Company example, we have $SSR = 252{,}378$ and $SSE = 54{,}825$. Hence:

$$MSR = \frac{252{,}378}{1} = 252{,}378$$

Also, we obtained earlier:

$$MSE = \frac{54{,}825}{23} = 2{,}384$$

Comment

The two mean squares *MSR* and *MSE* do not add to

$$\frac{SSTO}{(n-1)} = \frac{307{,}203}{24} = 12{,}800$$

Thus, mean squares are not additive. ■

Analysis of Variance Table

Basic Table. The breakdowns of the total sum of squares and associated degrees of freedom are displayed in the form of an analysis of variance table (ANOVA table) in Table 2.2. Mean squares of interest also are shown. In addition, the ANOVA table contains a column of expected mean squares that will be utilized shortly. The ANOVA table for the Toluca Company example is shown in Figure 2.2. The columns for degrees of freedom and sums of squares are reversed in the MINITAB output.

Modified Table. Sometimes an ANOVA table showing one additional element of decomposition is utilized. This modified table is based on the fact that the total sum of squares can be decomposed into two parts, as follows:

$$SSTO = \sum (Y_i - \bar{Y})^2 = \sum Y_i^2 - n\bar{Y}^2$$

In the modified ANOVA table, the *total uncorrected sum of squares,* denoted by *SSTOU*, is defined as:

$$SSTOU = \sum Y_i^2 \qquad \textbf{(2.54)}$$

and the *correction for the mean sum of squares*, denoted by *SS*(correction for mean), is defined as:

$$SS(\text{correction for mean}) = n\bar{Y}^2 \qquad \textbf{(2.55)}$$

Table 2.3 shows the general format of this modified ANOVA table. While both types of ANOVA tables are widely used, we shall usually utilize the basic type of table.

TABLE 2.2 ANOVA Table for Simple Linear Regression.

Source of Variation	*SS*	*df*	*MS*	*E*{*MS*}
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$	$\sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n-2$	$MSE = \frac{SSE}{n-2}$	σ^2
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n-1$		

TABLE 2.3 Modified ANOVA Table for Simple Linear Regression.

Source of Variation	*SS*	*df*	*MS*
Regression	$SSR = \sum(\hat{Y}_i - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum(Y_i - \hat{Y}_i)^2$	$n - 2$	$MSE = \frac{SSE}{n-2}$
Total	$SSTO = \sum(Y_i - \bar{Y})^2$	$n - 1$	
Correction for mean	$SS(\text{correction for mean}) = n\bar{Y}^2$	1	
Total, uncorrected	$SSTOU = \sum Y_i^2$	n	

Expected Mean Squares

In order to make inferences based on the analysis of variance approach, we need to know the expected value of each of the mean squares. The expected value of a mean square is the mean of its sampling distribution and tells us what is being estimated by the mean square. Statistical theory provides the following results:

$$E\{MSE\} = \sigma^2 \tag{2.56}$$

$$E\{MSR\} = \sigma^2 + \beta_1^2 \sum(X_i - \bar{X})^2 \tag{2.57}$$

The expected mean squares in (2.56) and (2.57) are shown in the analysis of variance table in Table 2.2. Note that result (2.56) is in accord with our earlier statement that *MSE* is an unbiased estimator of σ^2.

Two important implications of the expected mean squares in (2.56) and (2.57) are the following:

1. The mean of the sampling distribution of *MSE* is σ^2 whether or not X and Y are linearly related, i.e., whether or not $\beta_1 = 0$.
2. The mean of the sampling distribution of *MSR* is also σ^2 when $\beta_1 = 0$. Hence, when $\beta_1 = 0$, the sampling distributions of *MSR* and *MSE* are located identically and *MSR* and *MSE* will tend to be of the same order of magnitude.

   On the other hand, when $\beta_1 \neq 0$, the mean of the sampling distribution of *MSR* is greater than σ^2 since the term $\beta_1^2 \sum(X_i - \bar{X})^2$ in (2.57) then must be positive. Thus, when $\beta_1 \neq 0$, the mean of the sampling distribution of *MSR* is located to the right of that of *MSE* and, hence, *MSR* will tend to be larger than *MSE*.

This suggests that a comparison of *MSR* and *MSE* is useful for testing whether or not $\beta_1 = 0$. If *MSR* and *MSE* are of the same order of magnitude, this would suggest that $\beta_1 = 0$. On the other hand, if *MSR* is substantially greater than *MSE*, this would suggest that $\beta_1 \neq 0$. This indeed is the basic idea underlying the analysis of variance test to be discussed next.

Comment

The derivation of (2.56) follows from theorem (2.11), which states that $SSE/\sigma^2 \sim \chi^2(n-2)$ for regression model (2.1). Hence, it follows from property (A.42) of the chi-square distribution

that:

$$E\left\{\frac{SSE}{\sigma^2}\right\} = n - 2$$

or that:

$$E\left\{\frac{SSE}{n-2}\right\} = E\{MSE\} = \sigma^2$$

To find the expected value of *MSR*, we begin with (2.51):

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

Now by (A.15a), we have:

$$\sigma^2\{b_1\} = E\left\{b_1^2\right\} - (E\{b_1\})^2 \tag{2.58}$$

We know from (2.3a) that $E\{b_1\} = \beta_1$ and from (2.3b) that:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2}$$

Hence, substituting into (2.58), we obtain:

$$E\left\{b_1^2\right\} = \frac{\sigma^2}{\sum (X_i - \bar{X})^2} + \beta_1^2$$

It now follows that:

$$E\{SSR\} = E\left\{b_1^2\right\} \sum (X_i - \bar{X})^2 = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

Finally, $E\{MSR\}$ is:

$$E\{MSR\} = E\left\{\frac{SSR}{1}\right\} = \sigma^2 + \beta_1^2 \sum (X_i - \bar{X})^2$$

■

F Test of $\beta_1 = 0$ versus $\beta_1 \neq 0$

The analysis of variance approach provides us with a battery of highly useful tests for regression models (and other linear statistical models). For the simple linear regression case considered here, the analysis of variance provides us with a test for:

$$\begin{aligned} H_0&: \beta_1 = 0 \\ H_a&: \beta_1 \neq 0 \end{aligned} \tag{2.59}$$

Test Statistic. The test statistic for the analysis of variance approach is denoted by F^*. As just mentioned, it compares *MSR* and *MSE* in the following fashion:

$$F^* = \frac{MSR}{MSE} \tag{2.60}$$

The earlier motivation, based on the expected mean squares in Table 2.2, suggests that large values of F^* support H_a and values of F^* near 1 support H_0. In other words, the appropriate test is an upper-tail one.

Sampling Distribution of *F.** In order to be able to construct a statistical decision rule and examine its properties, we need to know the sampling distribution of F^*. We begin by considering the sampling distribution of F^* when H_0 ($\beta_1 = 0$) holds. *Cochran's theorem*

will be most helpful in this connection. For our purposes, this theorem can be stated as follows:

> If all n observations Y_i come from the same normal distribution with mean μ and variance σ^2, and *SSTO* is decomposed into k sums of squares SS_r, each with degrees of freedom df_r, then the SS_r/σ^2 terms are independent χ^2 variables with df_r degrees of freedom if: **(2.61)**

$$\sum_{r=1}^{k} df_r = n - 1$$

Note from Table 2.2 that we have decomposed *SSTO* into the two sums of squares *SSR* and *SSE* and that their degrees of freedom are additive. Hence:

> If $\beta_1 = 0$ so that all Y_i have the same mean $\mu = \beta_0$ and the same variance σ^2, SSE/σ^2 and SSR/σ^2 are independent χ^2 variables.

Now consider test statistic F^*, which we can write as follows:

$$F^* = \frac{\frac{SSR}{\sigma^2}}{1} \div \frac{\frac{SSE}{\sigma^2}}{n-2} = \frac{MSR}{MSE}$$

But by Cochran's theorem, we have when H_0 holds:

$$F^* \sim \frac{\chi^2(1)}{1} \div \frac{\chi^2(n-2)}{n-2} \qquad \text{when } H_0 \text{ holds}$$

where the χ^2 variables are independent. Thus, when H_0 holds, F^* is the ratio of two independent χ^2 variables, each divided by its degrees of freedom. But this is the definition of an F random variable in (A.47).

We have thus established that if H_0 holds, F^* follows the F distribution, specifically the $F(1, n-2)$ distribution.

When H_a holds, it can be shown that F^* follows the noncentral F distribution, a complex distribution that we need not consider further at this time.

Comment

Even if $\beta_1 \neq 0$, *SSR* and *SSE* are independent and $SSE/\sigma^2 \sim \chi^2$. However, the condition that both SSR/σ^2 and SSE/σ^2 are χ^2 random variables requires $\beta_1 = 0$. ■

Construction of Decision Rule. Since the test is upper-tail and F^* is distributed as $F(1, n-2)$ when H_0 holds, the decision rule is as follows when the risk of a Type I error is to be controlled at α:

$$\begin{aligned} &\text{If } F^* \leq F(1-\alpha; 1, n-2), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; 1, n-2), \text{ conclude } H_a \end{aligned} \qquad \textbf{(2.62)}$$

where $F(1-\alpha; 1, n-2)$ is the $(1-\alpha)100$ percentile of the appropriate F distribution.

Example

For the Toluca Company example, we shall repeat the earlier test on β_1, this time using the F test. The alternative conclusions are:

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

As before, let $\alpha = .05$. Since $n = 25$, we require $F(.95; 1, 23) = 4.28$. The decision rule is:

$$\text{If } F^* \leq 4.28, \text{ conclude } H_0$$
$$\text{If } F^* > 4.28, \text{ conclude } H_a$$

We have from earlier that $MSR = 252{,}378$ and $MSE = 2{,}384$. Hence, F^* is:

$$F^* = \frac{252{,}378}{2{,}384} = 105.9$$

Since $F^* = 105.9 > 4.28$, we conclude H_a, that $\beta_1 \neq 0$, or that there is a linear association between work hours and lot size. This is the same result as when the t test was employed, as it must be according to our discussion below.

The MINITAB output in Figure 2.2 on page 46 shows the F^* statistic in the column labeled F. Next to it is shown the P-value, $P\{F(1, 23) > 105.9\}$, namely, 0+, indicating that the data are not consistent with $\beta_1 = 0$.

Equivalence of *F* Test and *t* Test. For a given α level, the F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is equivalent algebraically to the two-tailed t test. To see this, recall from (2.51) that:

$$SSR = b_1^2 \sum (X_i - \bar{X})^2$$

Thus, we can write:

$$F^* = \frac{SSR \div 1}{SSE \div (n-2)} = \frac{b_1^2 \sum (X_i - \bar{X})^2}{MSE}$$

But since $s^2\{b_1\} = MSE / \sum (X_i - \bar{X})^2$, we obtain:

$$F^* = \frac{b_1^2}{s^2\{b_1\}} = \left(\frac{b_1}{s\{b_1\}}\right)^2 = (t^*)^2 \qquad \textbf{(2.63)}$$

The last step follows because the t^* statistic for testing whether or not $\beta_1 = 0$ is by (2.17):

$$t^* = \frac{b_1}{s\{b_1\}}$$

In the Toluca Company example, we just calculated that $F^* = 105.9$. From earlier work, we have $t^* = 10.29$ (see Figure 2.2). We thus see that $(10.29)^2 = 105.9$.

Corresponding to the relation between t^* and F^*, we have the following relation between the required percentiles of the t and F distributions for the tests: $[t(1 - \alpha/2; n - 2)]^2 = F(1 - \alpha; 1, n - 2)$. In our tests on β_1, these percentiles were $[t(.975; 23)]^2 = (2.069)^2 = 4.28 = F(.95; 1, 23)$. Remember that the t test is two-tailed whereas the F test is one-tailed.

Thus, at any given α level, we can use either the t test or the F test for testing $\beta_1 = 0$ versus $\beta_1 \neq 0$. Whenever one test leads to H_0, so will the other, and correspondingly for H_a. The t test, however, is more flexible since it can be used for one-sided alternatives involving $\beta_1 (\leq \geq) 0$ versus $\beta_1 (> <) 0$, while the F test cannot.

2.8 General Linear Test Approach

The analysis of variance test of $\beta_1 = 0$ versus $\beta_1 \neq 0$ is an example of the general test for a linear statistical model. We now explain this general test approach in terms of the simple linear regression model. We do so at this time because of the generality of the approach and the wide use we shall make of it, and because of the simplicity of understanding the approach in terms of simple linear regression.

The general linear test approach involves three basic steps, which we now describe in turn.

Full Model

We begin with the model considered to be appropriate for the data, which in this context is called the *full* or *unrestricted model*. For the simple linear regression case, the full model is the normal error regression model (2.1):

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad \text{Full model} \tag{2.64}$$

We fit this full model, either by the method of least squares or by the method of maximum likelihood, and obtain the error sum of squares. The error sum of squares is the sum of the squared deviations of each observation Y_i around its estimated expected value. In this context, we shall denote this sum of squares by $SSE(F)$ to indicate that it is the error sum of squares for the full model. Here, we have:

$$SSE(F) = \sum [Y_i - (b_0 + b_1 X_i)]^2 = \sum (Y_i - \hat{Y}_i)^2 = SSE \tag{2.65}$$

Thus, for the full model (2.64), the error sum of squares is simply *SSE*, which measures the variability of the Y_i observations around the fitted regression line.

Reduced Model

Next, we consider H_0. In this instance, we have:

$$\begin{aligned} H_0&: \beta_1 = 0 \\ H_a&: \beta_1 \neq 0 \end{aligned} \tag{2.66}$$

The model when H_0 holds is called the *reduced* or *restricted model*. When $\beta_1 = 0$, model (2.64) reduces to:

$$Y_i = \beta_0 + \varepsilon_i \qquad \text{Reduced model} \tag{2.67}$$

We fit this reduced model, by either the method of least squares or the method of maximum likelihood, and obtain the error sum of squares for this reduced model, denoted by $SSE(R)$. When we fit the particular reduced model (2.67), it can be shown that the least squares and maximum likelihood estimator of β_0 is $\bar{Y}$. Hence, the estimated expected value for each observation is $b_0 = \bar{Y}$, and the error sum of squares for this reduced model is:

$$SSE(R) = \sum (Y_i - b_0)^2 = \sum (Y_i - \bar{Y})^2 = SSTO \tag{2.68}$$

Test Statistic

The logic now is to compare the two error sums of squares $SSE(F)$ and $SSE(R)$. It can be shown that $SSE(F)$ never is greater than $SSE(R)$:

$$SSE(F) \leq SSE(R) \tag{2.69}$$

The reason is that the more parameters are in the model, the better one can fit the data and the smaller are the deviations around the fitted regression function. When $SSE(F)$ is not much less than $SSE(R)$, using the full model does not account for much more of the variability of the Y_i than does the reduced model, in which case the data suggest that the reduced model is adequate (i.e., that H_0 holds). To put this another way, when $SSE(F)$ is close to $SSE(R)$, the variation of the observations around the fitted regression function for the full model is almost as great as the variation around the fitted regression function for the reduced model. In this case, the added parameters in the full model really do not help to reduce the variation in the Y_i about the fitted regression function. Thus, a small difference $SSE(R) - SSE(F)$ suggests that H_0 holds. On the other hand, a large difference suggests that H_a holds because the additional parameters in the model do help to reduce substantially the variation of the observations Y_i around the fitted regression function.

The actual test statistic is a function of $SSE(R) - SSE(F)$, namely:

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F} \tag{2.70}$$

which follows the F distribution when H_0 holds. The degrees of freedom df_R and df_F are those associated with the reduced and full model error sums of squares, respectively. Large values of F^* lead to H_a because a large difference $SSE(R) - SSE(F)$ suggests that H_a holds. The decision rule therefore is:

$$\begin{aligned} &\text{If } F^* \leq F(1-\alpha; df_R - df_F,\ df_F), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; df_R - df_F,\ df_F), \text{ conclude } H_a \end{aligned} \tag{2.71}$$

For testing whether or not $\beta_1 = 0$, we therefore have:

$$\begin{aligned} SSE(R) &= SSTO & SSE(F) &= SSE \\ df_R &= n-1 & df_F &= n-2 \end{aligned}$$

so that we obtain when substituting into (2.70):

$$F^* = \frac{SSTO - SSE}{(n-1)-(n-2)} \div \frac{SSE}{n-2} = \frac{SSR}{1} \div \frac{SSE}{n-2} = \frac{MSR}{MSE}$$

which is identical to the analysis of variance test statistic (2.60).

Summary

The general linear test approach can be used for highly complex tests of linear statistical models, as well as for simple tests. The basic steps in summary form are:

1. Fit the full model and obtain the error sum of squares $SSE(F)$.
2. Fit the reduced model under H_0 and obtain the error sum of squares $SSE(R)$.
3. Use test statistic (2.70) and decision rule (2.71).

2.9 Descriptive Measures of Linear Association between X and Y

We have discussed the major uses of regression analysis—estimation of parameters and means and prediction of new observations—without mentioning the "degree of linear association" between X and Y, or similar terms. The reason is that the usefulness of estimates or predictions depends upon the width of the interval and the user's needs for precision, which vary from one application to another. Hence, no single descriptive measure of the "degree of linear association" can capture the essential information as to whether a given regression relation is useful in any particular application.

Nevertheless, there are times when the degree of linear association is of interest in its own right. We shall now briefly discuss two descriptive measures that are frequently used in practice to describe the degree of linear association between X and Y.

Coefficient of Determination

We saw earlier that *SSTO* measures the variation in the observations Y_i, or the uncertainty in predicting Y, when no account of the predictor variable X is taken. Thus, *SSTO* is a measure of the uncertainty in predicting Y when X is not considered. Similarly, *SSE* measures the variation in the Y_i when a regression model utilizing the predictor variable X is employed. A natural measure of the effect of X in reducing the variation in Y, i.e., in reducing the uncertainty in predicting Y, is to express the reduction in variation ($SSTO - SSE = SSR$) as a proportion of the total variation:

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \tag{2.72}$$

The measure R^2 is called the *coefficient of determination*. Since $0 \leq SSE \leq SSTO$, it follows that:

$$0 \leq R^2 \leq 1 \tag{2.72a}$$

We may interpret R^2 as the proportionate reduction of total variation associated with the use of the predictor variable X. Thus, the larger R^2 is, the more the total variation of Y is reduced by introducing the predictor variable X. The limiting values of R^2 occur as follows:

1. When all observations fall on the fitted regression line, then $SSE = 0$ and $R^2 = 1$. This case is shown in Figure 2.8a. Here, the predictor variable X accounts for all variation in the observations Y_i.

2. When the fitted regression line is horizontal so that $b_1 = 0$ and $\hat{Y}_i \equiv \bar{Y}$, then $SSE = SSTO$ and $R^2 = 0$. This case is shown in Figure 2.8b. Here, there is no linear association between X and Y in the sample data, and the predictor variable X is of no help in reducing the variation in the observations Y_i with linear regression.

In practice, R^2 is not likely to be 0 or 1 but somewhere between these limits. The closer it is to 1, the greater is said to be the degree of linear association between X and Y.

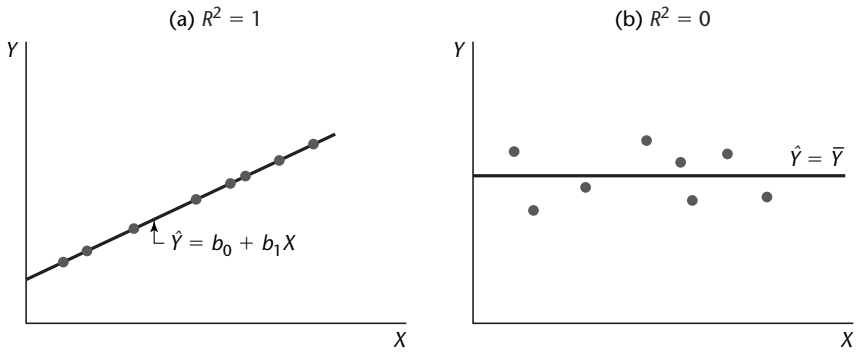


FIGURE 2.8
Scatter Plots when $R^2 = 1$ and $R^2 = 0$.

Example

For the Toluca Company example, we obtained $SSTO = 307{,}203$ and $SSR = 252{,}378$. Hence:

$$R^2 = \frac{252{,}378}{307{,}203} = .822$$

Thus, the variation in work hours is reduced by 82.2 percent when lot size is considered.

The MINITAB output in Figure 2.2 shows the coefficient of determination R^2 labeled as R–sq in percent form. The output also shows the coefficient R–sq(adj), which will be explained in Chapter 6.

Limitations of R^2

We noted that no single measure will be adequate for describing the usefulness of a regression model for different applications. Still, the coefficient of determination is widely used. Unfortunately, it is subject to serious misunderstandings. We consider now three common misunderstandings:

Misunderstanding 1. A high coefficient of determination indicates that useful predictions can be made. This is not necessarily correct. In the Toluca Company example, we saw that the coefficient of determination was high ($R^2 = .82$). Yet the 90 percent prediction interval for the next lot, consisting of 100 units, was wide (332 to 507 hours) and not precise enough to permit management to schedule workers effectively.

Misunderstanding 2. A high coefficient of determination indicates that the estimated regression line is a good fit. Again, this is not necessarily correct. Figure 2.9a shows a scatter plot where the coefficient of determination is high ($R^2 = .69$). Yet a linear regression function would not be a good fit since the regression relation is curvilinear.

Misunderstanding 3. A coefficient of determination near zero indicates that X and Y are not related. This also is not necessarily correct. Figure 2.9b shows a scatter plot where the coefficient of determination between X and Y is $R^2 = .02$. Yet X and Y are strongly related; however, the relationship between the two variables is curvilinear.

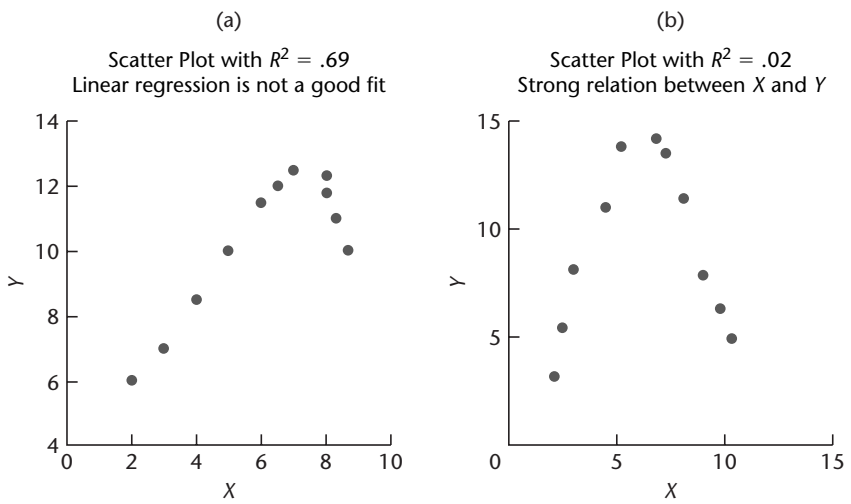


FIGURE 2.9 Illustrations of Two Misunderstandings about Coefficient of Determination.

Misunderstanding 1 arises because R^2 measures only a relative reduction from *SSTO* and provides no information about absolute precision for estimating a mean response or predicting a new observation. Misunderstandings 2 and 3 arise because R^2 measures the degree of *linear* association between X and Y, whereas the actual regression relation may be curvilinear.

Coefficient of Correlation

A measure of linear association between Y and X when both Y and X are random is the *coefficient of correlation*. This measure is the signed square root of R^2:

$$r = \pm\sqrt{R^2} \tag{2.73}$$

A plus or minus sign is attached to this measure according to whether the slope of the fitted regression line is positive or negative. Thus, the range of r is: $-1 \leq r \leq 1$.

Example

For the Toluca Company example, we obtained $R^2 = .822$. Treating X as a random variable, the correlation coefficient here is:

$$r = +\sqrt{.822} = .907$$

The plus sign is affixed since b_1 is positive. We take up the topic of correlation analysis in more detail in Section 2.11.

Comments

1. The value taken by R^2 in a given sample tends to be affected by the spacing of the X observations. This is implied in (2.72). *SSE* is not affected systematically by the spacing of the X_i since, for regression model (2.1), $\sigma^2\{Y_i\} = \sigma^2$ at all X levels. However, the wider the spacing of the X_i in the sample when $b_1 \neq 0$, the greater will tend to be the spread of the observed Y_i around $\bar{Y}$ and hence the greater *SSTO* will be. Consequently, the wider the X_i are spaced, the higher R^2 will tend to be.

2. The regression sum of squares *SSR* is often called the "explained variation" in Y, and the residual sum of squares *SSE* is called the "unexplained variation." The coefficient R^2 then is interpreted in terms of the proportion of the total variation in Y (*SSTO*) which has been "explained" by X. Unfortunately,

this terminology frequently is taken literally and, hence, misunderstood. Remember that in a regression model there is no implication that Y necessarily depends on X in a causal or explanatory sense.

3. Regression models do not contain a parameter to be estimated by R^2 or r. These are simply descriptive measures of the degree of linear association between X and Y in the sample observations that may, or may not, be useful in any instance. ■

2.10 Considerations in Applying Regression Analysis

We have now discussed the major uses of regression analysis—to make inferences about the regression parameters, to estimate the mean response for a given X, and to predict a new observation Y for a given X. It remains to make a few cautionary remarks about implementing applications of regression analysis.

1. Frequently, regression analysis is used to make inferences for the future. For instance, for planning staffing requirements, a school board may wish to predict future enrollments by using a regression model containing several demographic variables as predictor variables. In applications of this type, it is important to remember that the validity of the regression application depends upon whether basic causal conditions in the period ahead will be similar to those in existence during the period upon which the regression analysis is based. This caution applies whether mean responses are to be estimated, new observations predicted, or regression parameters estimated.

2. In predicting new observations on Y, the predictor variable X itself often has to be predicted. For instance, we mentioned earlier the prediction of company sales for next year from the demographic projection of the number of persons 16 years of age or older next year. A prediction of company sales under these circumstances is a conditional prediction, dependent upon the correctness of the population projection. It is easy to forget the conditional nature of this type of prediction.

3. Another caution deals with inferences pertaining to levels of the predictor variable that fall outside the range of observations. Unfortunately, this situation frequently occurs in practice. A company that predicts its sales from a regression relation of company sales to disposable personal income will often find the level of disposable personal income of interest (e.g., for the year ahead) to fall beyond the range of past data. If the X level does not fall far beyond this range, one may have reasonable confidence in the application of the regression analysis. On the other hand, if the X level falls far beyond the range of past data, extreme caution should be exercised since one cannot be sure that the regression function that fits the past data is appropriate over the wider range of the predictor variable.

4. A statistical test that leads to the conclusion that $\beta_1 \neq 0$ does not establish a cause-and-effect relation between the predictor and response variables. As we noted in Chapter 1, with nonexperimental data both the X and Y variables may be simultaneously influenced by other variables not in the regression model. On the other hand, the existence of a regression relation in controlled experiments is often good evidence of a cause-and-effect relation.

5. We should note again that frequently we wish to estimate several mean responses or predict several new observations for different levels of the predictor variable, and that special problems arise in this case. The confidence coefficients for the limits (2.33) for estimating a mean response and for the prediction limits (2.36) for a new observation apply

only for a single level of X for a given sample. In Chapter 4, we discuss how to make multiple inferences from a given sample.

6. Finally, when observations on the predictor variable X are subject to measurement errors, the resulting parameter estimates are generally no longer unbiased. In Chapter 4, we discuss several ways to handle this situation.

2.11 Normal Correlation Models

Distinction between Regression and Correlation Model

The normal error regression model (2.1), which has been used throughout this chapter and which will continue to be used, assumes that the X values are known constants. As a consequence of this, the confidence coefficients and risks of errors refer to repeated sampling when the X values are kept the same from sample to sample.

Frequently, it may not be appropriate to consider the X values as known constants. For instance, consider regressing daily bathing suit sales by a department store on mean daily temperature. Surely, the department store cannot control daily temperatures, so it would not be meaningful to think of repeated sampling where the temperature levels are the same from sample to sample. As a second example, an analyst may use a correlation model for the two variables "height of person" and "weight of person" in a study of a sample of persons, each variable being taken as random. The analyst might wish to study the relation between the two variables or might be interested in making inferences about weight of a person on the basis of the person's height, in making inferences about height on the basis of weight, or in both.

Other examples where a correlation model, rather than a regression model, may be appropriate are:

1. To study the relation between service station sales of gasoline, and sales of auxiliary products.
2. To study the relation between company net income determined by generally accepted accounting principles and net income according to tax regulations.
3. To study the relation between blood pressure and age in human subjects.

The correlation model most widely employed is the normal correlation model. We discuss it here for the case of two variables.

Bivariate Normal Distribution

The normal correlation model for the case of two variables is based on the *bivariate normal distribution*. Let us denote the two variables as Y_1 and Y_2. (We do not use the notation X and Y here because both variables play a symmetrical role in correlation analysis.) We say that Y_1 and Y_2 are *jointly normally distributed* if the density function of their joint distribution is that of the bivariate normal distribution.

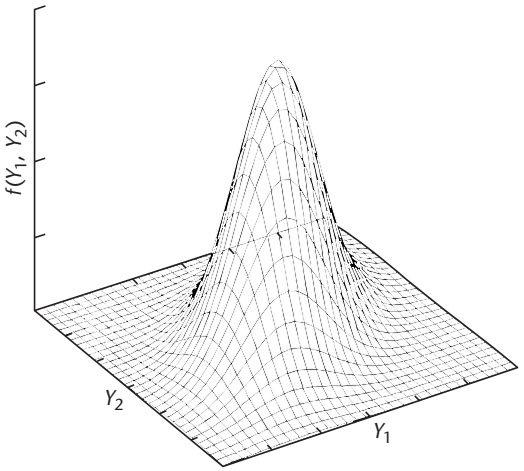


FIGURE 2.10
Example of Bivariate Normal Distribution.

Density Function. The density function of the bivariate normal distribution is as follows:

$$f(Y_1, Y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho_{12}^2}} \exp\left\{ -\frac{1}{2(1-\rho_{12}^2)}\left[\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho_{12}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)\left(\frac{Y_2-\mu_2}{\sigma_2}\right) + \left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]\right\} \quad \textbf{(2.74)}$$

Note that this density function involves five parameters: $\mu_1, \mu_2, \sigma_1, \sigma_2, \rho_{12}$. We shall explain the meaning of these parameters shortly. First, let us consider a graphic representation of the bivariate normal distribution.

Figure 2.10 contains a SYSTAT three-dimensional plot of a bivariate normal probability distribution. The probability distribution is a surface in three-dimensional space. For every pair of (Y_1, Y_2) values, the density $f(Y_1, Y_2)$ represents the height of the surface at that point. The surface is continuous, and probability corresponds to volume under the surface.

Marginal Distributions. If Y_1 and Y_2 are jointly normally distributed, it can be shown that their marginal distributions have the following characteristics:

The marginal distribution of Y_1 is normal with mean μ_1 and standard deviation σ_1: **(2.75a)**

$$f_1(Y_1) = \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left[-\frac{1}{2}\left(\frac{Y_1-\mu_1}{\sigma_1}\right)^2\right]$$

The marginal distribution of Y_2 is normal with mean μ_2 and standard deviation σ_2: **(2.75b)**

$$f_2(Y_2) = \frac{1}{\sqrt{2\pi}\sigma_2} \exp\left[-\frac{1}{2}\left(\frac{Y_2-\mu_2}{\sigma_2}\right)^2\right]$$

Thus, when Y_1 and Y_2 are jointly normally distributed, each of the two variables by itself is normally distributed. The converse, however, is not generally true; if Y_1 and Y_2 are each normally distributed, they need not be jointly normally distributed in accord with (2.74).

Meaning of Parameters. The five parameters of the bivariate normal density function (2.74) have the following meaning:

1. μ_1 and σ_1 are the mean and standard deviation of the marginal distribution of Y_1.
2. μ_2 and σ_2 are the mean and standard deviation of the marginal distribution of Y_2.
3. ρ_{12} is the *coefficient of correlation* between the random variables Y_1 and Y_2. This coefficient is denoted by $\rho\{Y_1, Y_2\}$ in Appendix A, using the correlation operator notation, and defined in (A.25a):

$$\rho_{12} = \rho\{Y_1, Y_2\} = \frac{\sigma_{12}}{\sigma_1\sigma_2} \tag{2.76}$$

Here, σ_1 and σ_2, as just mentioned, denote the standard deviations of Y_1 and Y_2, and σ_{12} denotes the covariance $\sigma\{Y_1, Y_2\}$ between Y_1 and Y_2 as defined in (A.21):

$$\sigma_{12} = \sigma\{Y_1, Y_2\} = E\{(Y_1 - \mu_1)(Y_2 - \mu_2)\} \tag{2.77}$$

Note that $\sigma_{12} \equiv \sigma_{21}$ and $\rho_{12} \equiv \rho_{21}$.

If Y_1 and Y_2 are independent, $\sigma_{12} = 0$ according to (A.28) so that $\rho_{12} = 0$. If Y_1 and Y_2 are positively related—that is, Y_1 tends to be large when Y_2 is large, or small when Y_2 is small—σ_{12} is positive and so is ρ_{12}. On the other hand, if Y_1 and Y_2 are negatively related—that is, Y_1 tends to be large when Y_2 is small, or vice versa—σ_{12} is negative and so is ρ_{12}. The coefficient of correlation ρ_{12} can take on any value between -1 and 1 inclusive. It assumes 1 if the linear relation between Y_1 and Y_2 is perfectly positive (direct) and -1 if it is perfectly negative (inverse).

Conditional Inferences

As noted, one principal use of a bivariate correlation model is to make conditional inferences regarding one variable, given the other variable. Suppose Y_1 represents a service station's gasoline sales and Y_2 its sales of auxiliary products. We may then wish to predict a service station's sales of auxiliary products Y_2, given that its gasoline sales are $Y_1 = \$5{,}500$.

Such conditional inferences require the use of conditional probability distributions, which we discuss next.

Conditional Probability Distribution of Y_1. The density function of the conditional probability distribution of Y_1 for any given value of Y_2 is denoted by $f(Y_1|Y_2)$ and defined as follows:

$$f(Y_1|Y_2) = \frac{f(Y_1, Y_2)}{f_2(Y_2)} \tag{2.78}$$

where $f(Y_1, Y_2)$ is the joint density function of Y_1 and Y_2, and $f_2(Y_2)$ is the marginal density function of Y_2. When Y_1 and Y_2 are jointly normally distributed according to (2.74) so that the marginal density function $f_2(Y_2)$ is given by (2.75b), it can be shown that:

The conditional probability distribution of Y_1 for any given value of Y_2 is normal with mean $\alpha_{1|2} + \beta_{12}Y_2$ and standard deviation $\sigma_{1|2}$ and its density function is: **(2.79)**

$$f(Y_1|Y_2) = \frac{1}{\sqrt{2\pi}\sigma_{1|2}} \exp\left[-\frac{1}{2}\left(\frac{Y_1 - \alpha_{1|2} - \beta_{12}Y_2}{\sigma_{1|2}}\right)^2\right]$$

The parameters $\alpha_{1|2}$, β_{12}, and $\sigma_{1|2}$ of the conditional probability distributions of Y_1 are functions of the parameters of the joint probability distribution (2.74), as follows:

$$\alpha_{1|2} = \mu_1 - \mu_2\rho_{12}\frac{\sigma_1}{\sigma_2} \tag{2.80a}$$

$$\beta_{12} = \rho_{12}\frac{\sigma_1}{\sigma_2} \tag{2.80b}$$

$$\sigma^2_{1|2} = \sigma^2_1\left(1 - \rho^2_{12}\right) \tag{2.80c}$$

The parameter $\alpha_{1|2}$ is the intercept of the line of regression of Y_1 on Y_2, and the parameter β_{12} is the slope of this line. Thus we find that the conditional distribution of Y_1, given Y_2, is equivalent to the normal error regression model (1.24).

Conditional Probability Distributions of Y_2. The random variables Y_1 and Y_2 play symmetrical roles in the bivariate normal probability distribution (2.74). Hence, it follows:

The conditional probability distribution of Y_2 for any given value of Y_1 is normal with mean $\alpha_{2|1} + \beta_{21}Y_1$ and standard deviation $\sigma_{2|1}$ and its density function is: **(2.81)**

$$f(Y_2|Y_1) = \frac{1}{\sqrt{2\pi}\sigma_{2|1}}\exp\left[-\frac{1}{2}\left(\frac{Y_2 - \alpha_{2|1} - \beta_{21}Y_1}{\sigma_{2|1}}\right)^2\right]$$

The parameters $\alpha_{2|1}$, β_{21}, and $\sigma_{2|1}$ of the conditional probability distributions of Y_2 are functions of the parameters of the joint probability distribution (2.74), as follows:

$$\alpha_{2|1} = \mu_2 - \mu_1\rho_{12}\frac{\sigma_2}{\sigma_1} \tag{2.82a}$$

$$\beta_{21} = \rho_{12}\frac{\sigma_2}{\sigma_1} \tag{2.82b}$$

$$\sigma^2_{2|1} = \sigma^2_2\left(1 - \rho^2_{12}\right) \tag{2.82c}$$

Important Characteristics of Conditional Distributions. Three important characteristics of the conditional probability distributions of Y_1 are normality, linear regression, and constant variance. We take up each of these in turn.

1. The conditional probability distribution of Y_1 for any given value of Y_2 is normal. Imagine that we slice a bivariate normal distribution vertically at a given value of Y_2, say, at Y_{h2}. That is, we slice it parallel to the Y_1 axis. This slicing is shown in Figure 2.11. The exposed cross section has the shape of a normal distribution, and after being scaled so that its area is 1, it portrays the conditional probability distribution of Y_1, given that $Y_2 = Y_{h2}$.

This property of normality holds no matter what the value Y_{h2} is. Thus, whenever we slice the bivariate normal distribution parallel to the Y_1 axis, we obtain (after proper scaling) a normal conditional probability distribution.

2. The means of the conditional probability distributions of Y_1 fall on a straight line, and hence are a linear function of Y_2:

$$E\{Y_1|Y_2\} = \alpha_{1|2} + \beta_{12}Y_2 \tag{2.83}$$

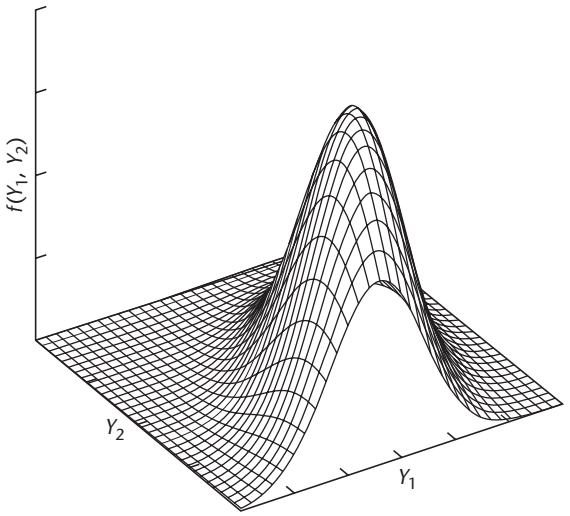


FIGURE 2.11
Cross Section of Bivariate Normal Distribution at Y_{h2}.

Here, $\alpha_{1|2}$ is the intercept parameter and β_{12} the slope parameter. Thus, the relation between the conditional means and Y_2 is given by a linear regression function.

3. All conditional probability distributions of Y_1 have the same standard deviation $\sigma_{1|2}$. Thus, no matter where we slice the bivariate normal distribution parallel to the Y_1 axis, the resulting conditional probability distribution (after scaling to have an area of 1) has the same standard deviation. Hence, constant variances characterize the conditional probability distributions of Y_1.

Equivalence to Normal Error Regression Model. Suppose that we select a random sample of observations (Y_1, Y_2) from a bivariate normal population and wish to make conditional inferences about Y_1, given Y_2. The preceding discussion makes it clear that the normal error regression model (1.24) is entirely applicable because:

1. The Y_1 observations are independent.
2. The Y_1 observations when Y_2 is considered given or fixed are normally distributed with mean $E\{Y_1|Y_2\} = \alpha_{1|2} + \beta_{12}Y_2$ and constant variance $\sigma^2_{1|2}$.

Use of Regression Analysis. In view of the equivalence of each of the conditional bivariate normal correlation models (2.81) and (2.79) with the normal error regression model (1.24), all conditional inferences with these correlation models can be made by means of the usual regression methods. For instance, if a researcher has data that can be appropriately described as having been generated from a bivariate normal distribution and wishes to make inferences about Y_2, given a particular value of Y_1, the ordinary regression techniques will be applicable. Thus, the regression function of Y_2 on Y_1 can be estimated by means of (1.12), the slope of the regression line can be estimated by means of the interval estimate (2.15), a new observation Y_2, given the value of Y_1, can be predicted by means of (2.36), and so on. Computer regression packages can be used in the usual manner. To avoid notational problems, it may be helpful to relabel the variables according to regression usage: $Y = Y_2$, $X = Y_1$. Of course, if conditional inferences on Y_1 for given values of Y_2 are desired, the notation correspondences would be: $Y = Y_1$, $X = Y_2$.

Can we still use regression model (2.1) if Y_1 and Y_2 are not bivariate normal? It can be shown that all results on estimation, testing, and prediction obtained from regression model (2.1) apply if $Y_1 = Y$ and $Y_2 = X$ are random variables, and if the following conditions hold:

1. The conditional distributions of the Y_i, given X_i, are normal and independent, with conditional means $\beta_0 + \beta_1 X_i$ and conditional variance σ^2.
2. The X_i are independent random variables whose probability distribution $g(X_i)$ does not involve the parameters β_0, β_1, σ^2.

These conditions require only that regression model (2.1) is appropriate for each *conditional* distribution of Y_i, and that the probability distribution of the X_i does not involve the regression parameters. If these conditions are met, all earlier results on estimation, testing, and prediction still hold even though the X_i are now random variables. The major modification occurs in the interpretation of confidence coefficients and specified risks of error. When X is random, these refer to repeated sampling of pairs of (X_i, Y_i) values, where the X_i values as well as the Y_i values change from sample to sample. Thus, in our bathing suit sales illustration, a confidence coefficient would refer to the proportion of correct interval estimates if repeated samples of n days' sales and temperatures were obtained and the confidence interval calculated for each sample. Another modification occurs in the test's power, which is different when X is a random variable.

Comments

1. The notation for the parameters of the conditional correlation models departs somewhat from our previous notation for regression models. The symbol α is now used to denote the regression intercept. The subscript 1|2 to α indicates that Y_1 is regressed on Y_2. Similarly, the subscript 2|1 to α indicates that Y_2 is regressed on Y_1. The symbol β_{12} indicates that it is the slope in the regression of Y_1 on Y_2, while β_{21} is the slope in the regression of Y_2 on Y_1. Finally, $\sigma_{2|1}$ is the standard deviation of the conditional probability distributions of Y_2 for any given Y_1, while $\sigma_{1|2}$ is the standard deviation of the conditional probability distributions of Y_1 for any given Y_2.

2. Two distinct regressions are involved in a bivariate normal model, that of Y_1 on Y_2 when Y_2 is fixed and that of Y_2 on Y_1 when Y_1 is fixed. In general, the two regression lines are not the same. For instance, the two slopes β_{12} and β_{21} are the same only if $\sigma_1 = \sigma_2$, as can be seen from (2.80b) and (2.82b).

3. When interval estimates for the conditional correlation models are obtained, the confidence coefficient refers to repeated samples where pairs of observations (Y_1, Y_2) are obtained from the bivariate normal distribution. ■

Inferences on Correlation Coefficients

A principal use of the bivariate normal correlation model is to study the relationship between two variables. In a bivariate normal model, the parameter ρ_{12} provides information about the degree of the linear relationship between the two variables Y_1 and Y_2.

Point Estimator of ρ_{12}. The maximum likelihood estimator of ρ_{12}, denoted by r_{12}, is given by:

$$r_{12} = \frac{\sum(Y_{i1} - \bar{Y}_1)(Y_{i2} - \bar{Y}_2)}{\left[\sum(Y_{i1} - \bar{Y}_1)^2 \sum(Y_{i2} - \bar{Y}_2)^2\right]^{1/2}} \tag{2.84}$$

This estimator is often called the *Pearson product-moment correlation coefficient*. It is a biased estimator of ρ_{12} (unless $\rho_{12} = 0$ or 1), but the bias is small when n is large.

It can be shown that the range of r_{12} is:

$$-1 \leq r_{12} \leq 1 \tag{2.85}$$

Generally, values of r_{12} near 1 indicate a strong positive (direct) linear association between Y_1 and Y_2 whereas values of r_{12} near -1 indicate a strong negative (indirect) linear association. Values of r_{12} near 0 indicate little or no linear association between Y_1 and Y_2.

Test whether $\rho_{12} = 0$. When the population is bivariate normal, it is frequently desired to test whether the coefficient of correlation is zero:

$$\begin{aligned} H_0&: \rho_{12} = 0 \\ H_a&: \rho_{12} \neq 0 \end{aligned} \tag{2.86}$$

The reason for interest in this test is that in the case where Y_1 and Y_2 are jointly normally distributed, $\rho_{12} = 0$ implies that Y_1 and Y_2 are independent.

We can use regression procedures for the test since (2.80b) implies that the following alternatives are equivalent to those in (2.86):

$$\begin{aligned} H_0&: \beta_{12} = 0 \\ H_a&: \beta_{12} \neq 0 \end{aligned} \tag{2.86a}$$

and (2.82b) implies that the following alternatives are also equivalent to the ones in (2.86):

$$\begin{aligned} H_0&: \beta_{21} = 0 \\ H_a&: \beta_{21} \neq 0 \end{aligned} \tag{2.86b}$$

It can be shown that the test statistics for testing either (2.86a) or (2.86b) are the same and can be expressed directly in terms of r_{12}:

$$t^* = \frac{r_{12}\sqrt{n-2}}{\sqrt{1-r_{12}^2}} \tag{2.87}$$

If H_0 holds, t^* follows the $t(n-2)$ distribution. The appropriate decision rule to control the Type I error at α is:

$$\begin{aligned} &\text{If } |t^*| \leq t(1-\alpha/2; n-2), \text{ conclude } H_0 \\ &\text{If } |t^*| > t(1-\alpha/2; n-2), \text{ conclude } H_a \end{aligned} \tag{2.88}$$

Test statistic (2.87) is identical to the regression t^* test statistic (2.17).

Example

A national oil company was interested in the relationship between its service station gasoline sales and its sales of auxiliary products. A company analyst obtained a random sample of 23 of its service stations and obtained average monthly sales data on gasoline sales (Y_1) and comparable sales of its auxiliary products and services (Y_2). These data (not shown) resulted in an estimated correlation coefficient $r_{12} = .52$. Suppose the analyst wished to test whether or not the association was positive, controlling the level of significance at $\alpha = .05$. The alternatives would then be:

$$\begin{aligned} H_0&: \rho_{12} \leq 0 \\ H_a&: \rho_{12} > 0 \end{aligned}$$

and the decision rule based on test statistic (2.87) would be:

$$\text{If } t^* \leq t(1-\alpha; n-2), \text{ conclude } H_0$$
$$\text{If } t^* > t(1-\alpha; n-2), \text{ conclude } H_a$$

For $\alpha = .05$, we require $t(.95; 21) = 1.721$. Since:

$$t^* = \frac{.52\sqrt{21}}{\sqrt{1-(.52)^2}} = 2.79$$

is greater than 1.721, we would conclude H_a, that $\rho_{12} > 0$. The P-value for this test is .006.

Interval Estimation of ρ_{12} Using the z' Transformation. Because the sampling distribution of r_{12} is complicated when $\rho_{12} \neq 0$, interval estimation of ρ_{12} is usually carried out by means of an approximate procedure based on a transformation. This transformation, known as the *Fisher z transformation,* is as follows:

$$z' = \frac{1}{2}\log_e\left(\frac{1+r_{12}}{1-r_{12}}\right) \tag{2.89}$$

When n is large (25 or more is a useful rule of thumb), the distribution of z' is approximately normal with approximate mean and variance:

$$E\{z'\} = \zeta = \frac{1}{2}\log_e\left(\frac{1+\rho_{12}}{1-\rho_{12}}\right) \tag{2.90}$$

$$\sigma^2\{z'\} = \frac{1}{n-3} \tag{2.91}$$

Note that the transformation from r_{12} to z' in (2.89) is the same as the relation in (2.90) between ρ_{12} and $E\{z'\} = \zeta$. Also note that the approximate variance of z' is a known constant, depending only on the sample size n.

Table B.8 gives paired values for the left and right sides of (2.89) and (2.90), thus eliminating the need for calculations. For instance, if r_{12} or ρ_{12} equals .25, Table B.8 indicates that z' or ζ equals .2554, and vice versa. The values on the two sides of the transformation always have the same sign. Thus, if r_{12} or ρ_{12} is negative, a minus sign is attached to the value in Table B.8. For instance, if $r_{12} = -.25$, $z' = -.2554$.

Interval Estimate. When the sample size is large ($n \geq 25$), the standardized statistic:

$$\frac{z'-\zeta}{\sigma\{z'\}} \tag{2.92}$$

is approximately a standard normal variable. Therefore, approximate $1-\alpha$ confidence limits for ζ are:

$$z' \pm z(1-\alpha/2)\sigma\{z'\} \tag{2.93}$$

where $z(1-\alpha/2)$ is the $(1-\alpha/2)100$ percentile of the standard normal distribution. The $1-\alpha$ confidence limits for ρ_{12} are then obtained by transforming the limits on ζ by means of (2.90).

Example

An economist investigated food purchasing patterns by households in a midwestern city. Two hundred households with family incomes between $40,000 and $60,000 were selected to ascertain, among other things, the proportions of the food budget expended for beef and poultry, respectively. The economist expected these to be negatively related, and wished to estimate the coefficient of correlation with a 95 percent confidence interval. Some supporting evidence suggested that the joint distribution of the two variables does not depart markedly from a bivariate normal one.

The point estimate of ρ_{12} was $r_{12} = -.61$ (data and calculations not shown). To obtain an approximate 95 percent confidence interval estimate, we require:

$$z' = -.7089 \quad \text{when } r_{12} = -.61 \qquad \text{(from Table B.8)}$$
$$\sigma\{z'\} = \frac{1}{\sqrt{200-3}} = .07125$$
$$z(.975) = 1.960$$

Hence, the confidence limits for ζ, by (2.93), are $-.7089 \pm 1.960(.07125)$, and the approximate 95 percent confidence interval is:

$$-.849 \leq \zeta \leq -.569$$

Using Table B.8 to transform back to ρ_{12}, we obtain:

$$-.69 \leq \rho_{12} \leq -.51$$

This confidence interval was sufficiently precise to be useful to the economist, confirming the negative relation and indicating that the degree of linear association is moderately high.

Comments

1. As usual, a confidence interval for ρ_{12} can be employed to test whether or not ρ_{12} has a specified value—say, .5—by noting whether or not the specified value falls within the confidence limits.

2. It can be shown that the square of the coefficient of correlation, namely ρ_{12}^2, measures the relative reduction in the variability of Y_2 associated with the use of variable Y_1. To see this, we noted earlier in (2.80c) and (2.82c) that:

$$\sigma_{1|2}^2 = \sigma_1^2\left(1 - \rho_{12}^2\right) \tag{2.94a}$$
$$\sigma_{2|1}^2 = \sigma_2^2\left(1 - \rho_{12}^2\right) \tag{2.94b}$$

We can rewrite these expressions as follows:

$$\rho_{12}^2 = \frac{\sigma_1^2 - \sigma_{1|2}^2}{\sigma_1^2} \tag{2.95a}$$
$$\rho_{12}^2 = \frac{\sigma_2^2 - \sigma_{2|1}^2}{\sigma_2^2} \tag{2.95b}$$

The meaning of ρ_{12}^2 is now clear. Consider first (2.95a). ρ_{12}^2 measures how much smaller relatively is the variability in the conditional distributions of Y_1, for any given level of Y_2, than is the variability in the marginal distribution of Y_1. Thus, ρ_{12}^2 measures the relative reduction in the variability of Y_1 associated with the use of variable Y_2. Correspondingly, (2.95b) shows that ρ_{12}^2 also measures the relative reduction in the variability of Y_2 associated with the use of variable Y_1.

It can be shown that:

$$0 \leq \rho_{12}^2 \leq 1 \tag{2.96}$$

The limiting value $\rho_{12}^2 = 0$ occurs when Y_1 and Y_2 are independent, so that the variances of each variable in the conditional probability distributions are then no smaller than the variance in the marginal distribution. The limiting value $\rho_{12}^2 = 1$ occurs when there is no variability in the conditional probability distributions for each variable, so perfect predictions of either variable can be made from the other.

3. The interpretation of ρ_{12}^2 as measuring the relative reduction in the conditional variances as compared with the marginal variance is valid for the case of a bivariate normal population, but not for many other bivariate populations. Of course, the interpretation implies nothing in a causal sense.

4. Confidence limits for ρ_{12}^2 can be obtained by squaring the respective confidence limits for ρ_{12}, provided the latter limits do not differ in sign. ■

Spearman Rank Correlation Coefficient

At times the joint distribution of two random variables Y_1 and Y_2 differs considerably from the bivariate normal distribution (2.74). In those cases, transformations of the variables Y_1 and Y_2 may be sought to make the joint distribution of the transformed variables approximately bivariate normal and thus permit the use of the inference procedures about ρ_{12} described earlier.

When no appropriate transformations can be found, a nonparametric *rank correlation* procedure may be useful for making inferences about the association between Y_1 and Y_2. The *Spearman rank correlation coefficient* is widely used for this purpose. First, the observations on Y_1 (i.e., $Y_{11}, \ldots, Y_{n1}$) are expressed in ranks from 1 to n. We denote the rank of Y_{i1} by R_{i1}. Similarly, the observations on Y_2 (i.e., $Y_{12}, \ldots, Y_{n2}$) are ranked, with the rank of Y_{i2} denoted by R_{i2}. The Spearman rank correlation coefficient, to be denoted by r_S, is then defined as the ordinary Pearson product-moment correlation coefficient in (2.84) based on the rank data:

$$r_S = \frac{\sum(R_{i1} - \bar{R}_1)(R_{i2} - \bar{R}_2)}{\left[\sum(R_{i1} - \bar{R}_1)^2 \sum(R_{i2} - \bar{R}_2)^2\right]^{1/2}} \tag{2.97}$$

Here $\bar{R}_1$ is the mean of the ranks R_{i1} and $\bar{R}_2$ is the mean of the ranks R_{i2}. Of course, since the ranks R_{i1} and R_{i2} are the integers $1, \ldots, n$, it follows that $\bar{R}_1 = \bar{R}_2 = (n+1)/2$.

Like an ordinary correlation coefficient, the Spearman rank correlation coefficient takes on values between -1 and 1 inclusive:

$$-1 \leq r_S \leq 1 \tag{2.98}$$

The coefficient r_S equals 1 when the ranks for Y_1 are identical to those for Y_2, that is, when the case with rank 1 for Y_1 also has rank 1 for Y_2, and so on. In that case, there is perfect association between the ranks for the two variables. The coefficient r_S equals -1 when the case with rank 1 for Y_1 has rank n for Y_2, the case with rank 2 for Y_1 has rank $n-1$ for Y_2, and so on. In that event, there is perfect inverse association between the ranks for the two variables. When there is little, if any, association between the ranks of Y_1 and Y_2, the Spearman rank correlation coefficient tends to have a value near zero.

The Spearman rank correlation coefficient can be used to test the alternatives:

$$H_0: \text{There is no association between } Y_1 \text{and } Y_2$$
$$H_a: \text{There is an association between } Y_1 \text{ and } Y_2 \quad \textbf{(2.99)}$$

A two-sided test is conducted here since H_a includes either positive or negative association. When the alternative H_a is:

$$H_a: \text{There is positive (negative) association between } Y_1 \text{ and } Y_2 \quad \textbf{(2.100)}$$

an upper-tail (lower-tail) one-sided test is conducted.

The probability distribution of r_S under H_0 is not difficult to obtain. It is based on the condition that, for any ranking of Y_1, all rankings of Y_2 are equally likely when there is no association between Y_1 and Y_2. Tables have been prepared and are presented in specialized texts such as Reference 2.1. Computer packages generally do not present the probability distribution of r_S under H_0 but give only the two-sided P-value. When the sample size n exceeds 10, the test can be carried out approximately by using test statistic (2.87):

$$t^* = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} \quad \textbf{(2.101)}$$

based on the t distribution with $n - 2$ degrees of freedom.

Example

A market researcher wished to examine whether an association exists between population size (Y_1) and per capita expenditures for a new food product (Y_2). The data for a random sample of 12 test markets are given in Table 2.4, columns 1 and 2. Because the distributions of the variables do not appear to be approximately normal, a nonparametric test of association is desired. The ranks for the variables are given in Table 2.4, columns 3 and 4. A computer package found that the coefficient of simple correlation between the ranked data in columns 3 and 4 is $r_S = .895$. The alternatives of interest are the two-sided ones in (2.99). Since n

TABLE 2.4
Data on Population and Expenditures and Their Ranks—Sales Marketing Example.

	(1)	(2)	(3)	(4)
Test Market i	Population (in thousands) Y_{i1}	Per Capita Expenditure (dollars) Y_{i2}	R_{i1}	R_{i2}
1	29	127	1	2
2	435	214	8	11
3	86	133	3	4
4	1,090	208	11	10
5	219	153	7	6
6	503	184	9	8
7	47	130	2	3
8	3,524	217	12	12
9	185	141	6	5
10	98	154	5	7
11	952	194	10	9
12	89	103	4	1

exceeds 10 here, we use test statistic (2.101):

$$t^* = \frac{.895\sqrt{12-2}}{\sqrt{1-(.895)^2}} = 6.34$$

For $\alpha = .01$, we require $t(.995; 10) = 3.169$. Since $|t^*| = 6.34 > 3.169$, we conclude H_a, that there is an association between population size and per capita expenditures for the food product. The two-sided P-value of the test is .00008.

Comments

1. In case of ties among some data values, each of the tied values is given the average of the ranks involved.

2. It is interesting to note that had the data in Table 2.4 been analyzed by assuming the bivariate normal distribution assumption (2.74) and test statistic (2.87), then the strength of the association would have been somewhat weaker. In particular, the Pearson product-moment correlation coefficient is $r_{12} = .674$, with $t^* = .674\sqrt{10}/\sqrt{1-(.674)^2} = 2.885$. Our conclusion would have been to conclude H_0, that there is no association between population size and per capita expenditures for the food product. The two-sided P-value of the test is .016.

3. Another nonparametric rank procedure similar to Spearman's r_S is *Kendall's* τ. This statistic also measures how far the rankings of Y_1 and Y_2 differ from each other, but in a somewhat different way than the Spearman rank correlation coefficient. A discussion of Kendall's τ may be found in Reference 2.2. ■

Cited References

2.1. Gibbons, J. D. *Nonparametric Methods for Quantitative Analysis*. 2nd ed. Columbus, Ohio: American Sciences Press, 1985.

2.2. Kendall, M. G., and J. D. Gibbons. *Rank Correlation Methods*. 5th ed. London: Oxford University Press, 1990.

Problems

2.1. A student working on a summer internship in the economic research department of a large corporation studied the relation between sales of a product (Y, in million dollars) and population (X, in million persons) in the firm's 50 marketing districts. The normal error regression model (2.1) was employed. The student first wished to test whether or not a linear association between Y and X existed. The student accessed a simple linear regression program and obtained the following information on the regression coefficients:

Parameter	Estimated Value	95 Percent Confidence Limits	
Intercept	7.43119	−1.18518	16.0476
Slope	.755048	.452886	1.05721

a. The student concluded from these results that there is a linear association between Y and X. Is the conclusion warranted? What is the implied level of significance?

b. Someone questioned the negative lower confidence limit for the intercept, pointing out that dollar sales cannot be negative even if the population in a district is zero. Discuss.

2.2. In a test of the alternatives H_0: $\beta_1 \leq 0$ versus H_a: $\beta_1 > 0$, an analyst concluded H_0. Does this conclusion imply that there is no linear association between X and Y? Explain.

2.3. A member of a student team playing an interactive marketing game received the following computer output when studying the relation between advertising expenditures (X) and sales (Y) for one of the team's products:

Estimated regression equation: $\hat{Y} = 350.7 - .18X$

Two-sided P-value for estimated slope: .91

The student stated: "The message I get here is that the more we spend on advertising this product, the fewer units we sell!" Comment.

2.4. Refer to **Grade point average** Problem 1.19.

a. Obtain a 99 percent confidence interval for β_1. Interpret your confidence interval. Does it include zero? Why might the director of admissions be interested in whether the confidence interval includes zero?

b. Test, using the test statistic t^*, whether or not a linear association exists between student's ACT score (X) and GPA at the end of the freshman year (Y). Use a level of significance of .01. State the alternatives, decision rule, and conclusion.

c. What is the P-value of your test in part (b)? How does it support the conclusion reached in part (b)?

*2.5. Refer to **Copier maintenance** Problem 1.20.

a. Estimate the change in the mean service time when the number of copiers serviced increases by one. Use a 90 percent confidence interval. Interpret your confidence interval.

b. Conduct a t test to determine whether or not there is a linear association between X and Y here; control the α risk at .10. State the alternatives, decision rule, and conclusion. What is the P-value of your test?

c. Are your results in parts (a) and (b) consistent? Explain.

d. The manufacturer has suggested that the mean required time should not increase by more than 14 minutes for each additional copier that is serviced on a service call. Conduct a test to decide whether this standard is being satisfied by Tri-City. Control the risk of a Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

e. Does b_0 give any relevant information here about the "start-up" time on calls—i.e., about the time required before service work is begun on the copiers at a customer location?

*2.6. Refer to **Airfreight breakage** Problem 1.21.

a. Estimate β_1 with a 95 percent confidence interval. Interpret your interval estimate.

b. Conduct a t test to decide whether or not there is a linear association between number of times a carton is transferred (X) and number of broken ampules (Y). Use a level of significance of .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. β_0 represents here the mean number of ampules broken when no transfers of the shipment are made—i.e., when $X = 0$. Obtain a 95 percent confidence interval for β_0 and interpret it.

d. A consultant has suggested, on the basis of previous experience, that the mean number of broken ampules should not exceed 9.0 when no transfers are made. Conduct an appropriate test, using $\alpha = .025$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

e. Obtain the power of your test in part (b) if actually $\beta_1 = 2.0$. Assume $\sigma\{b_1\} = .50$. Also obtain the power of your test in part (d) if actually $\beta_0 = 11$. Assume $\sigma\{b_0\} = .75$.

2.7. Refer to **Plastic hardness** Problem 1.22.

a. Estimate the change in the mean hardness when the elapsed time increases by one hour. Use a 99 percent confidence interval. Interpret your interval estimate.

b. The plastic manufacturer has stated that the mean hardness should increase by 2 Brinell units per hour. Conduct a two-sided test to decide whether this standard is being satisfied; use $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

c. Obtain the power of your test in part (b) if the standard actually is being exceeded by .3 Brinell units per hour. Assume $\sigma\{b_1\} = .1$.

2.8. Refer to Figure 2.2 for the Toluca Company example. A consultant has advised that an increase of one unit in lot size should require an increase of 3.0 in the expected number of work hours for the given production item.

a. Conduct a test to decide whether or not the increase in the expected number of work hours in the Toluca Company equals this standard. Use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

b. Obtain the power of your test in part (a) if the consultant's standard actually is being exceeded by .5 hour. Assume $\sigma\{b_1\} = .35$.

c. Why is $F^* = 105.88$, given in the printout, not relevant for the test in part (a)?

2.9. Refer to Figure 2.2. A student, noting that $s\{b_1\}$ is furnished in the printout, asks why $s\{\hat{Y}_h\}$ is not also given. Discuss.

2.10. For each of the following questions, explain whether a confidence interval for a mean response or a prediction interval for a new observation is appropriate.

a. What will be the humidity level in this greenhouse tomorrow when we set the temperature level at 31°C?

b. How much do families whose disposable income is $23,500 spend, on the average, for meals away from home?

c. How many kilowatt-hours of electricity will be consumed next month by commercial and industrial users in the Twin Cities service area, given that the index of business activity for the area remains at its present level?

2.11. A person asks if there is a difference between the "mean response at $X = X_h$" and the "mean of m new observations at $X = X_h$." Reply.

2.12. Can $\sigma^2\{\text{pred}\}$ in (2.37) be brought increasingly close to 0 as n becomes large? Is this also the case for $\sigma^2\{\hat{Y}_h\}$ in (2.29b)? What is the implication of this difference?

2.13. Refer to **Grade point average** Problem 1.19.

a. Obtain a 95 percent interval estimate of the mean freshman GPA for students whose ACT test score is 28. Interpret your confidence interval.

b. Mary Jones obtained a score of 28 on the entrance test. Predict her freshman GPA using a 95 percent prediction interval. Interpret your prediction interval.

c. Is the prediction interval in part (b) wider than the confidence interval in part (a)? Should it be?

d. Determine the boundary values of the 95 percent confidence band for the regression line when $X_h = 28$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

*2.14. Refer to **Copier maintenance** Problem 1.20.

a. Obtain a 90 percent confidence interval for the mean service time on calls in which six copiers are serviced. Interpret your confidence interval.

b. Obtain a 90 percent prediction interval for the service time on the next call in which six copiers are serviced. Is your prediction interval wider than the corresponding confidence interval in part (a)? Should it be?

c. Management wishes to estimate the expected service time *per copier* on calls in which six copiers are serviced. Obtain an appropriate 90 percent confidence interval by converting the interval obtained in part (a). Interpret the converted confidence interval.

d. Determine the boundary values of the 90 percent confidence band for the regression line when $X_h = 6$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

*2.15. Refer to **Airfreight breakage** Problem 1.21.

a. Because of changes in airline routes, shipments may have to be transferred more frequently than in the past. Estimate the mean breakage for the following numbers of transfers: $X = 2$, 4. Use separate 99 percent confidence intervals. Interpret your results.

b. The next shipment will entail two transfers. Obtain a 99 percent prediction interval for the number of broken ampules for this shipment. Interpret your prediction interval.

c. In the next several days, three independent shipments will be made, each entailing two transfers. Obtain a 99 percent prediction interval for the mean number of ampules broken in the three shipments. Convert this interval into a 99 percent prediction interval for the total number of ampules broken in the three shipments.

d. Determine the boundary values of the 99 percent confidence band for the regression line when $X_h = 2$ and when $X_h = 4$. Is your confidence band wider at these two points than the corresponding confidence intervals in part (a)? Should it be?

2.16. Refer to **Plastic hardness** Problem 1.22.

a. Obtain a 98 percent confidence interval for the mean hardness of molded items with an elapsed time of 30 hours. Interpret your confidence interval.

b. Obtain a 98 percent prediction interval for the hardness of a newly molded test item with an elapsed time of 30 hours.

c. Obtain a 98 percent prediction interval for the mean hardness of 10 newly molded test items, each with an elapsed time of 30 hours.

d. Is the prediction interval in part (c) narrower than the one in part (b)? Should it be?

e. Determine the boundary values of the 98 percent confidence band for the regression line when $X_h = 30$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

2.17. An analyst fitted normal error regression model (2.1) and conducted an F test of $\beta_1 = 0$ versus $\beta_1 \neq 0$. The P-value of the test was .033, and the analyst concluded H_a: $\beta_1 \neq 0$. Was the α level used by the analyst greater than or smaller than .033? If the α level had been .01, what would have been the appropriate conclusion?

2.18. For conducting statistical tests concerning the parameter β_1, why is the t test more versatile than the F test?

2.19. When testing whether or not $\beta_1 = 0$, why is the F test a one-sided test even though H_a includes both $\beta_1 < 0$ and $\beta_1 > 0$? [*Hint*: Refer to (2.57).]

2.20. A student asks whether R^2 is a point estimator of any parameter in the normal error regression model (2.1). Respond.

2.21. A value of R^2 near 1 is sometimes interpreted to imply that the relation between Y and X is sufficiently close so that suitably precise predictions of Y can be made from knowledge of X. Is this implication a necessary consequence of the definition of R^2?

2.22. Using the normal error regression model (2.1) in an engineering safety experiment, a researcher found for the first 10 cases that R^2 was zero. Is it possible that for the complete set of 30 cases R^2 will not be zero? Could R^2 not be zero for the first 10 cases, yet equal zero for all 30 cases? Explain.

2.23. Refer to **Grade point average** Problem 1.19.

a. Set up the ANOVA table.

b. What is estimated by *MSR* in your ANOVA table? by *MSE*? Under what condition do *MSR* and *MSE* estimate the same quantity?

c. Conduct an F test of whether or not $\beta_1 = 0$. Control the α risk at .01. State the alternatives, decision rule, and conclusion.

d. What is the absolute magnitude of the reduction in the variation of Y when X is introduced into the regression model? What is the relative reduction? What is the name of the latter measure?

e. Obtain r and attach the appropriate sign.

f. Which measure, R^2 or r, has the more clear-cut operational interpretation? Explain.

*2.24. Refer to **Copier maintenance** Problem 1.20.

a. Set up the basic ANOVA table in the format of Table 2.2. Which elements of your table are additive? Also set up the ANOVA table in the format of Table 2.3. How do the two tables differ?

b. Conduct an F test to determine whether or not there is a linear association between time spent and number of copiers serviced; use $\alpha = .10$. State the alternatives, decision rule, and conclusion.

c. By how much, relatively, is the total variation in number of minutes spent on a call reduced when the number of copiers serviced is introduced into the analysis? Is this a relatively small or large reduction? What is the name of this measure?

d. Calculate r and attach the appropriate sign.

e. Which measure, r or R^2, has the more clear-cut operational interpretation?

*2.25. Refer to **Airfreight breakage** Problem 1.21.

a. Set up the ANOVA table. Which elements are additive?

b. Conduct an F test to decide whether or not there is a linear association between the number of times a carton is transferred and the number of broken ampules; control the α risk at .05. State the alternatives, decision rule, and conclusion.

c. Obtain the t^* statistic for the test in part (b) and demonstrate numerically its equivalence to the F^* statistic obtained in part (b).

d. Calculate R^2 and r. What proportion of the variation in Y is accounted for by introducing X into the regression model?

2.26. Refer to **Plastic hardness** Problem 1.22.

a. Set up the ANOVA table.

b. Test by means of an F test whether or not there is a linear association between the hardness of the plastic and the elapsed time. Use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. Plot the deviations $Y_i - \hat{Y}_i$ against X_i on a graph. Plot the deviations $\hat{Y}_i - \bar{Y}$ against X_i on another graph, using the same scales as for the first graph. From your two graphs, does *SSE* or *SSR* appear to be the larger component of *SSTO*? What does this imply about the magnitude of R^2?

d. Calculate R^2 and r.

*2.27. Refer to **Muscle mass** Problem 1.27.

a. Conduct a test to decide whether or not there is a negative linear association between amount of muscle mass and age. Control the risk of Type I error at .05. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

b. The two-sided P-value for the test whether $\beta_0 = 0$ is 0+. Can it now be concluded that b_0 provides relevant information on the amount of muscle mass at birth for a female child?

c. Estimate with a 95 percent confidence interval the difference in expected muscle mass for women whose ages differ by one year. Why is it not necessary to know the specific ages to make this estimate?

*2.28. Refer to **Muscle mass** Problem 1.27.

a. Obtain a 95 percent confidence interval for the mean muscle mass for women of age 60. Interpret your confidence interval.

b. Obtain a 95 percent prediction interval for the muscle mass of a woman whose age is 60. Is the prediction interval relatively precise?

c. Determine the boundary values of the 95 percent confidence band for the regression line when $X_h = 60$. Is your confidence band wider at this point than the confidence interval in part (a)? Should it be?

*2.29. Refer to **Muscle mass** Problem 1.27.

a. Plot the deviations $Y_i - \hat{Y}_i$ against X_i on one graph. Plot the deviations $\hat{Y}_i - \bar{Y}$ against X_i on another graph, using the same scales as in the first graph. From your two graphs, does *SSE* or *SSR* appear to be the larger component of *SSTO*? What does this imply about the magnitude of R^2?

b. Set up the ANOVA table.

c. Test whether or not $\beta_1 = 0$ using an F test with $\alpha = .05$. State the alternatives, decision rule, and conclusion.

d. What proportion of the total variation in muscle mass remains "unexplained" when age is introduced into the analysis? Is this proportion relatively small or large?

e. Obtain R^2 and r.

2.30. Refer to **Crime rate** Problem 1.28.

a. Test whether or not there is a linear association between crime rate and percentage of high school graduates, using a t test with $\alpha = .01$. State the alternatives, decision rule, and conclusion. What is the P-value of the test?

b. Estimate β_1 with a 99 percent confidence interval. Interpret your interval estimate.

2.31. Refer to **Crime rate** Problem 1.28

a. Set up the ANOVA table.

b. Carry out the test in Problem 2.30a by means of the F test. Show the numerical equivalence of the two test statistics and decision rules. Is the P-value for the F test the same as that for the t test?

c. By how much is the total variation in crime rate reduced when percentage of high school graduates is introduced into the analysis? Is this a relatively large or small reduction?

d. Obtain r.

2.32. Refer to **Crime rate** Problems 1.28 and 2.30. Suppose that the test in Problem 2.30a is to be carried out by means of a general linear test.

a. State the full and reduced models.

b. Obtain (1) $SSE(F)$, (2) $SSE(R)$, (3) df_F, (4) df_R, (5) test statistic F^* for the general linear test, (6) decision rule.

c. Are the test statistic F^* and the decision rule for the general linear test numerically equivalent to those in Problem 2.30a?

2.33. In developing empirically a cost function from observed data on a complex chemical experiment, an analyst employed normal error regression model (2.1). β_0 was interpreted here as the cost of setting up the experiment. The analyst hypothesized that this cost should be \$7.5 thousand and wished to test the hypothesis by means of a general linear test.

a. Indicate the alternative conclusions for the test.

b. Specify the full and reduced models.

c. Without additional information, can you tell what the quantity $df_R - df_F$ in test statistic (2.70) will equal in the analyst's test? Explain.

2.34. Refer to **Grade point average** Problem 1.19.

a. Would it be more reasonable to consider the X_i as known constants or as random variables here? Explain.

b. If the X_i were considered to be random variables, would this have any effect on prediction intervals for new applicants? Explain.

2.35. Refer to **Copier maintenance** Problems 1.20 and 2.5. How would the meaning of the confidence coefficient in Problem 2.5a change if the predictor variable were considered a random variable and the conditions on page 83 were applicable?

2.36. A management trainee in a production department wished to study the relation between weight of rough casting and machining time to produce the finished block. The trainee selected castings so that the weights would be spaced equally apart in the sample and then observed the corresponding machining times. Would you recommend that a regression or a correlation model be used? Explain.

2.37. A social scientist stated: "The conditions for the bivariate normal distribution are so rarely met in my experience that I feel much safer using a regression model." Comment.

2.38. A student was investigating from a large sample whether variables Y_1 and Y_2 follow a bivariate normal distribution. The student obtained the residuals when regressing Y_1 on Y_2, and also obtained the residuals when regressing Y_2 on Y_1, and then prepared a normal probability plot for each set of residuals. Do these two normal probability plots provide sufficient information for determining whether the two variables follow a bivariate normal distribution? Explain.

2.39. For the bivariate normal distribution with parameters $\mu_1 = 50$, $\mu_2 = 100$, $\sigma_1 = 3$, $\sigma_2 = 4$, and $\rho_{12} = .80$.

a. State the characteristics of the marginal distribution of Y_1.

b. State the characteristics of the conditional distribution of Y_2 when $Y_1 = 55$.

c. State the characteristics of the conditional distribution of Y_1 when $Y_2 = 95$.

2.40. Explain whether any of the following would be affected if the bivariate normal model (2.74) were employed instead of the normal error regression model (2.1) with fixed levels of the predictor variable: (1) point estimates of the regression coefficients, (2) confidence limits for the regression coefficients, (3) interpretation of the confidence coefficient.

2.41. Refer to **Plastic hardness** Problem 1.22. A student was analyzing these data and received the following standard query from the interactive regression and correlation computer package: CALCULATE CONFIDENCE INTERVAL FOR POPULATION CORRELATION COEFFICIENT RHO? ANSWER Y OR N. Would a "yes" response lead to meaningful information here? Explain.

*2.42. **Property assessments.** The data that follow show assessed value for property tax purposes (Y_1, in thousand dollars) and sales price (Y_2, in thousand dollars) for a sample of 15 parcels of land for industrial development sold recently in "arm's length" transactions in a tax district. Assume that bivariate normal model (2.74) is appropriate here.

i:	1	2	3	...	13	14	15
Y_{i1}:	13.9	16.0	10.3	...	14.9	12.9	15.8
Y_{i2}:	28.6	34.7	21.0	...	35.1	30.0	36.2

a. Plot the data in a scatter diagram. Does the bivariate normal model appear to be appropriate here? Discuss.

b. Calculate r_{12}. What parameter is estimated by r_{12}? What is the interpretation of this parameter?

c. Test whether or not Y_1 and Y_2 are statistically independent in the population, using test statistic (2.87) and level of significance .01. State the alternatives, decision rule, and conclusion.

d. To test $\rho_{12} = .6$ versus $\rho_{12} \neq .6$, would it be appropriate to use test statistic (2.87)?

2.43. **Contract profitability.** A cost analyst for a drilling and blasting contractor examined 84 contracts handled in the last two years and found that the coefficient of correlation between value of contract (Y_1) and profit contribution generated by the contract (Y_2) is $r_{12} = .61$. Assume that bivariate normal model (2.74) applies.

a. Test whether or not Y_1 and Y_2 are statistically independent in the population; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

b. Estimate ρ_{12} with a 95 percent confidence interval.

c. Convert the confidence interval in part (b) to a 95 percent confidence interval for ρ_{12}^2. Interpret this interval estimate.

*2.44. **Bid preparation.** A building construction consultant studied the relationship between cost of bid preparation (Y_1) and amount of bid (Y_2) for the consulting firm's clients. In a sample of 103 bids prepared by clients, $r_{12} = .87$. Assume that bivariate normal model (2.74) applies.

a. Test whether or not $\rho_{12} = 0$; control the risk of Type I error at .10. State the alternatives, decision rule, and conclusion. What would be the implication if $\rho_{12} = 0$?

b. Obtain a 90 percent confidence interval for ρ_{12}. Interpret this interval estimate.

c. Convert the confidence interval in part (b) to a 90 percent confidence interval for ρ_{12}^2.

2.45. **Water flow.** An engineer, desiring to estimate the coefficient of correlation ρ_{12} between rate of water flow at point A in a stream (Y_1) and concurrent rate of flow at point B (Y_2), obtained $r_{12} = .83$ in a sample of 147 cases. Assume that bivariate normal model (2.74) is appropriate.

a. Obtain a 99 percent confidence interval for ρ_{12}.

b. Convert the confidence interval in part (a) to a 99 percent confidence interval for ρ_{12}^2.

2.46. Refer to **Property assessments** Problem 2.42. There is some question as to whether or not bivariate model (2.74) is appropriate.

a. Obtain the Spearman rank correlation coefficient r_S.

b. Test by means of the Spearman rank correlation coefficient whether an association exists between property assessments and sales prices using test statistic (2.101) with $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in Problem 2.42?

*2.47. Refer to **Muscle mass** Problem 1.27. Assume that the normal bivariate model (2.74) is appropriate.

a. Compute the Pearson product-moment correlation coefficient r_{12}.

b. Test whether muscle mass and age are statistically independent in the population; use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

c. The bivariate normal model (2.74) assumption is possibly inappropriate here. Compute the Spearman rank correlation coefficient, r_S.

d. Repeat part (b), this time basing the test of independence on the Spearman rank correlation computed in part (c) and test statistic (2.101). Use $\alpha = .05$. State the alternatives, decision rule, and conclusion.

e. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in parts (c) and (d)?

2.48. Refer to **Crime rate** Problems 1.28, 2.30, and 2.31. Assume that the normal bivariate model (2.74) is appropriate.

a. Compute the Pearson product-moment correlation coefficient r_{12}.

b. Test whether crime rate and percentage of high school graduates are statistically independent in the population; use $\alpha = .01$. State the alternatives, decision rule, and conclusion.

c. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in 2.31b and 2.30a, respectively?

2.49. Refer to **Crime rate** Problems 1.28 and 2.48. The bivariate normal model (2.74) assumption is possibly inappropriate here.

a. Compute the Spearman rank correlation coefficient r_S.

b. Test by means of the Spearman rank correlation coefficient whether an association exists between crime rate and percentage of high school graduates using test statistic (2.101) and a level of significance .01. State the alternatives, decision rule, and conclusion.

c. How do your estimates and conclusions in parts (a) and (b) compare to those obtained in Problems 2.48a and 2.48b, respectively?

Exercises

2.50. Derive the property in (2.6) for the k_i.

2.51. Show that b_0 as defined in (2.21) is an unbiased estimator of β_0.

2.52. Derive the expression in (2.22b) for the variance of b_0, making use of (2.31). Also explain how variance (2.22b) is a special case of variance (2.29b).

2.53. (Calculus needed.)

a. Obtain the likelihood function for the sample observations $Y_1, \ldots, Y_n$ given $X_1, \ldots, X_n$, if the conditions on page 83 apply.

b. Obtain the maximum likelihood estimators of β_0, β_1, and σ^2. Are the estimators of β_0 and β_1 the same as those in (1.27) when the X_i are fixed?

2.54. Suppose that normal error regression model (2.1) is applicable except that the error variance is not constant; rather the variance is larger, the larger is X. Does $\beta_1 = 0$ still imply that there is no linear association between X and Y? That there is no association between X and Y? Explain.

2.55. Derive the expression for *SSR* in (2.51).

2.56. In a small-scale regression study, five observations on Y were obtained corresponding to $X = 1$, 4, 10, 11, and 14. Assume that $\sigma = .6$, $\beta_0 = 5$, and $\beta_1 = 3$.

a. What are the expected values of *MSR* and *MSE* here?

b. For determining whether or not a regression relation exists, would it have been better or worse to have made the five observations at $X = 6$, 7, 8, 9, and 10? Why? Would the same answer apply if the principal purpose were to estimate the mean response for $X = 8$? Discuss.

2.57. The normal error regression model (2.1) is assumed to be applicable.

a. When testing H_0: $\beta_1 = 5$ versus H_a: $\beta_1 \neq 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

b. When testing H_0: $\beta_0 = 2$, $\beta_1 = 5$ versus H_a: not both $\beta_0 = 2$ and $\beta_1 = 5$ by means of a general linear test, what is the reduced model? What are the degrees of freedom df_R?

2.58. The random variables Y_1 and Y_2 follow the bivariate normal distribution in (2.74). Show that if $\rho_{12} = 0$, Y_1 and Y_2 are independent random variables.

2.59. (Calculus needed.)

a. Obtain the maximum likelihood estimators of the parameters of the bivariate normal distribution in (2.74).

b. Using the results in part (a), obtain the maximum likelihood estimators of the parameters of the conditional probability distribution of Y_1 for any value of Y_2 in (2.80).

c. Show that the maximum likelihood estimators of $\alpha_{1|2}$ and β_{12} obtained in part (b) are the same as the least squares estimators (1.10) for the regression coefficients in the simple linear regression model.

2.60. Show that test statistics (2.17) and (2.87) are equivalent.

2.61. Show that the ratio $SSR/SSTO$ is the same whether Y_1 is regressed on Y_2 or Y_2 is regressed on Y_1. [*Hint*: Use (1.10a) and (2.51).]

Projects

2.62. Refer to the **CDI** data set in Appendix C.2 and Project 1.43. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability in the number of active physicians?

2.63. Refer to the **CDI** data set in Appendix C.2 and Project 1.44. Obtain a separate interval estimate of β_1 for each region. Use a 90 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes?

2.64. Refer to the **SENIC** data set in Appendix C.1 and Project 1.45. Using R^2 as the criterion, which predictor variable accounts for the largest reduction in the variability of the average length of stay?

2.65. Refer to the **SENIC** data set in Appendix C.1 and Project 1.46. Obtain a separate interval estimate of β_1 for each region. Use a 95 percent confidence coefficient in each case. Do the regression lines for the different regions appear to have similar slopes?

2.66. Five observations on Y are to be taken when $X = 4$, 8, 12, 16, and 20, respectively. The true regression function is $E\{Y\} = 20 + 4X$, and the ε_i are independent $N(0, 25)$.

a. Generate five normal random numbers, with mean 0 and variance 25. Consider these random numbers as the error terms for the five Y observations at $X = 4, 8, 12, 16$, and 20 and calculate Y_1, Y_2, Y_3, Y_4, and Y_5. Obtain the least squares estimates b_0 and b_1 when fitting a straight line to the five cases. Also calculate $\hat{Y}_h$ when $X_h = 10$ and obtain a 95 percent confidence interval for $E\{Y_h\}$ when $X_h = 10$.

b. Repeat part (a) 200 times, generating new random numbers each time.

c. Make a frequency distribution of the 200 estimates b_1. Calculate the mean and standard deviation of the 200 estimates b_1. Are the results consistent with theoretical expectations?

d. What proportion of the 200 confidence intervals for $E\{Y_h\}$ when $X_h = 10$ include $E\{Y_h\}$? Is this result consistent with theoretical expectations?

2.67. Refer to **Grade point average** Problem 1.19.

a. Plot the data, with the least squares regression line for ACT scores between 20 and 30 superimposed.

b. On the plot in part (a), superimpose a plot of the 95 percent confidence band for the true regression line for ACT scores between 20 and 30. Does the confidence band suggest that the true regression relation has been precisely estimated? Discuss.

2.68. Refer to **Copier maintenance** Problem 1.20.

a. Plot the data, with the least squares regression line for numbers of copiers serviced between 1 and 8 superimposed.

b. On the plot in part (a), superimpose a plot of the 90 percent confidence band for the true regression line for numbers of copiers serviced between 1 and 8. Does the confidence band suggest that the true regression relation has been precisely estimated? Discuss.