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:

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

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

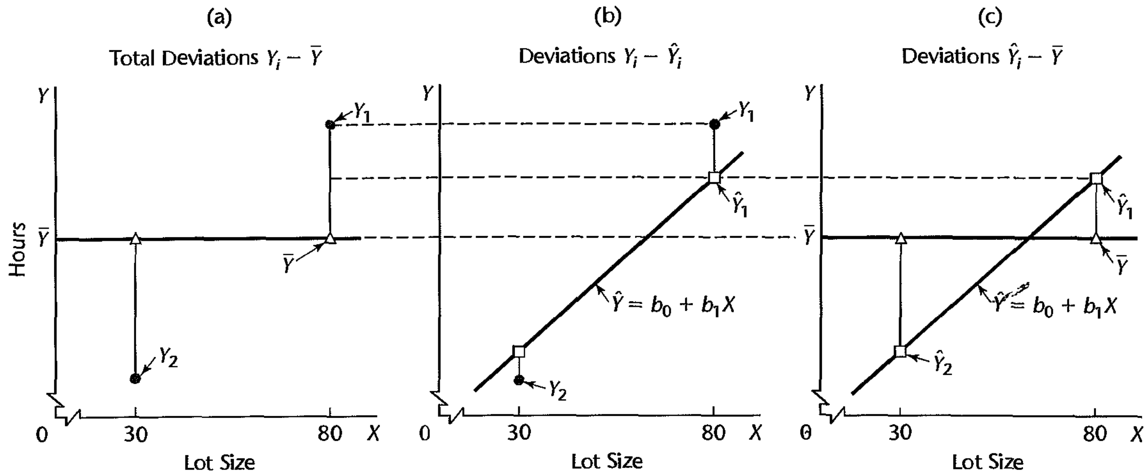
We now have developed the basic regression model and demonstrated its major uses. At this point, we consider the regression analysis from the perspective of analysis of variance. This new perspective will not enable us to do anything new, but the analysis of variance approach will come into its own when 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 $\bar{Y}$:

$$Y_i - \bar{Y} \tag{2.42}$$

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:

$$SSTOU = \sum Y_i^2 \tag{2.54}$$

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

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

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(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:

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

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:

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

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:

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

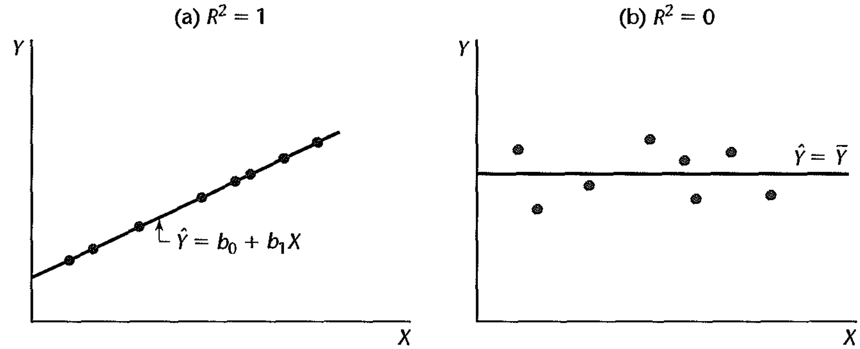
We may interpret R^2 as the proportionate reduction of total variation associated with the use of the predictor variable X. Thus, the larger R^2 is, the more the total variation of Y is reduced by introducing the predictor variable X. The limiting values of R^2 occur as follows:

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.

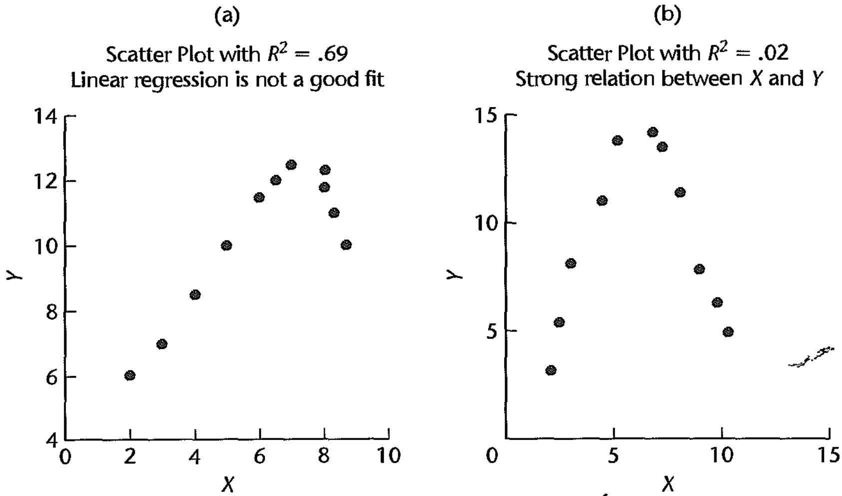


FIGURE 2.9 Illustrations of Two Misunderstandings about Coefficient of Determination.

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:

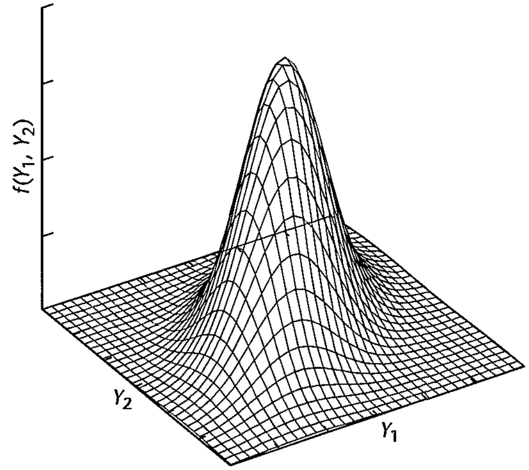
1. To study the relation between service station sales of gasoline, and sales of auxiliary products.
2. To study the relation between company net income determined by generally accepted accounting principles and net income according to tax regulations.
3. To study the relation between blood pressure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

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

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:

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

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

$$\sigma_{1|2}^2 = \sigma_1^2(1 - \rho_{12}^2) \tag{2.80c}$$

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:

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

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

$$\sigma_{2|1}^2 = \sigma_2^2(1 - \rho_{12}^2) \tag{2.82c}$$

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.

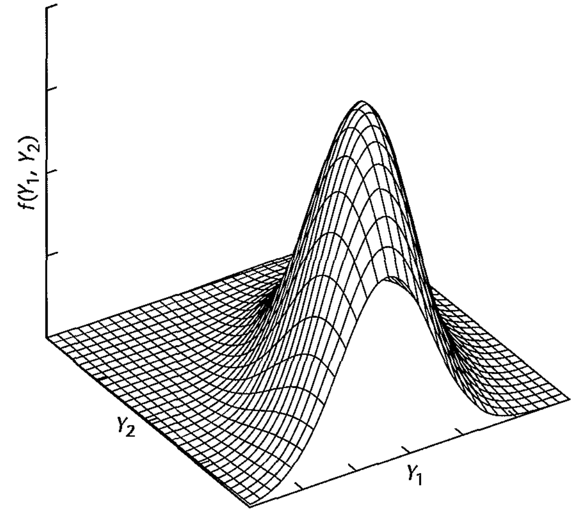
1. The conditional probability distribution of Y_1 for any given value of Y_2 is normal. Imagine that we slice a bivariate normal distribution vertically at a given value of Y_2, say, at Y_{h2}. That is, we slice it parallel to the Y_1 axis. This slicing is shown in Figure 2.11. The exposed cross section 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:

$$\begin{aligned} &H_0\text{: There is no association between } Y_1 \text{ and } Y_2 \\ &H_a\text{: There is an association between } Y_1 \text{ and } Y_2 \end{aligned} \tag{2.99}$$

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

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

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):

$$t^* = \frac{r_S\sqrt{n-2}}{\sqrt{1-r_S^2}} \tag{2.101}$$

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.

Test Market i	(1) Population (in thousands) Y_{i1}	(2) Per Capita Expenditure (dollars) Y_{i2}	(3) R_{i1}	(4) 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.

Chapter 3

Diagnostics and Remedial Measures

When a regression model, such as the simple linear regression model (2.1), is considered for an application, we can usually not be certain in advance that the model is appropriate for that application. Any one, or several, of the features of the model, such as linearity of the regression function or normality of the error terms, may not be appropriate for the particular data at hand. Hence, it is important to examine the aptness of the model for the data before inferences based on that model are undertaken. In this chapter, we discuss some simple graphic methods for studying the appropriateness of a model, as well as some formal statistical tests for doing so. We also consider some remedial techniques that can be helpful when the data are not in accordance with the conditions of regression model (2.1). We conclude the chapter with a case example that brings together the concepts and methods presented in this and the earlier chapters.

While the discussion in this chapter is in terms of the appropriateness of the simple linear regression model (2.1), the basic principles apply to all statistical models discussed in this book. In later chapters, additional methods useful for examining the appropriateness of statistical models and other remedial measures will be presented, as well as methods for validating the statistical model.

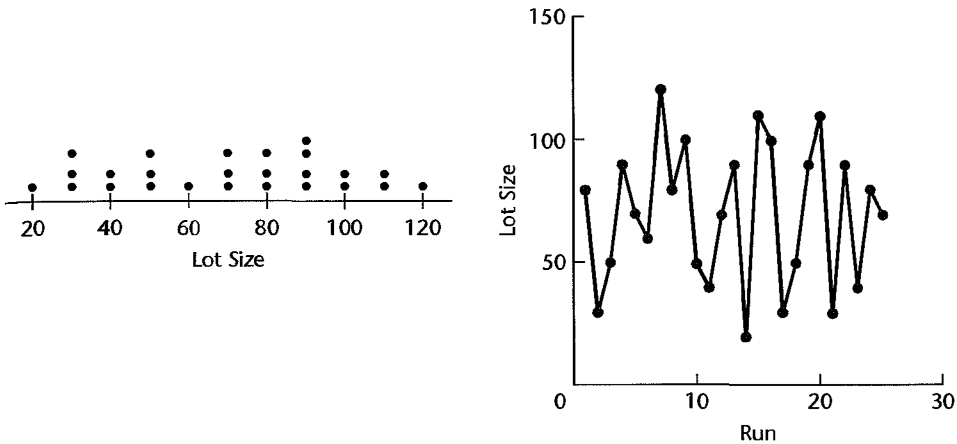
3.1 Diagnostics for Predictor Variable

We begin by considering some graphic diagnostics for the predictor variable. We need diagnostic information about the predictor variable to see if there are any outlying X values that could influence the appropriateness of the fitted regression function. We discuss the role of influential cases in detail in Chapter 10. Diagnostic information about the range and concentration of the X levels in the study is also useful for ascertaining the range of validity for the regression analysis.

Figure 3.1a contains a simple *dot plot* for the lot sizes in the Toluca Company example in Figure 1.10. A dot plot is helpful when the number of observations in the data set is not large. The dot plot in Figure 3.1a shows that the minimum and maximum lot sizes are 20 and 120, respectively, that the lot size levels are spread throughout this interval, and that

FIGURE 3.1 MINITAB and SYGRAPH Diagnostic Plots for Predictor Variable—Toluca Company Example.

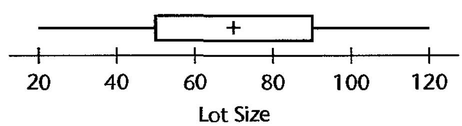
(a) Dot Plot (b) Sequence Plot



(c) Stem-and-Leaf Plot

```
 2   0
 3   000
 4   00
 5H  000
 6   0
 7M  000
 8   000
 9H  0000
10   00
11   00
12   0
```

(d) Box Plot



there are no lot sizes that are far outlying. The dot plot also shows that in a number of cases several runs were made for the same lot size.

A second useful diagnostic for the predictor variable is a *sequence plot.* Figure 3.1b contains a time sequence plot of the lot sizes for the Toluca Company example. Lot size is here plotted against production run (i.e., against time sequence). The points in the plot are connected to show more effectively the time sequence. Sequence plots should be utilized whenever data are obtained in a sequence, such as over time or for adjacent geographic areas. The sequence plot in Figure 3.1b contains no special pattern. If, say, the plot had shown that smaller lot sizes had been utilized early on and larger lot sizes later on, this information could be very helpful for subsequent diagnostic studies of the aptness of the fitted regression model.

Figures 3.1c and 3.1d contain two other diagnostic plots that present information similar to the dot plot in Figure 3.1a. The *stem-and-leaf plot* in Figure 3.1c provides information similar to a frequency histogram. By displaying the last digits, this plot also indicates here that all lot sizes in the Toluca Company example were multiples of 10. The letter M in the

SYGRAPH output denotes the stem where the median is located, and the letter H denotes the stems where the first and third quartiles (hinges) are located.

The *box plot* in Figure 3.1d shows the minimum and maximum lot sizes, the first and third quartiles, and the median lot size. We see that the middle half of the lot sizes range from 50 to 90, and that they are fairly symmetrically distributed because the median is located in the middle of the central box. A box plot is particularly helpful when there are many observations in the data set.

3.2 Residuals

Direct diagnostic plots for the response variable Y are ordinarily not too useful in regression analysis because the values of the observations on the response variable are a function of the level of the predictor variable. Instead, diagnostics for the response variable are usually carried out indirectly through an examination of the residuals.

The residual e_i, as defined in (1.16), is the difference between the observed value Y_i and the fitted value $\hat{Y}_i$:

$$e_i = Y_i - \hat{Y}_i \tag{3.1}$$

The residual may be regarded as the observed error, in distinction to the unknown true error ε_i in the regression model:

$$\varepsilon_i = Y_i - E\{Y_i\} \tag{3.2}$$

For regression model (2.1), the error terms ε_i are assumed to be independent normal random variables, with mean 0 and constant variance σ^2. If the model is appropriate for the data at hand, the observed residuals e_i should then reflect the properties assumed for the ε_i. This is the basic idea underlying *residual analysis*, a highly useful means of examining the aptness of a statistical model.

Properties of Residuals

Mean. The mean of the n residuals e_i for the simple linear regression model (2.1) is, by (1.17):

$$\bar{e} = \frac{\sum e_i}{n} = 0 \tag{3.3}$$

where $\bar{e}$ denotes the mean of the residuals. Thus, since $\bar{e}$ is always 0, it provides no information as to whether the true errors ε_i have expected value $E\{\varepsilon_i\} = 0$.

Variance. The variance of the n residuals e_i is defined as follows for regression model (2.1):

$$s^2 = \frac{\sum(e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = \frac{SSE}{n-2} = MSE \tag{3.4}$$

If the model is appropriate, MSE is, as noted earlier, an unbiased estimator of the variance of the error terms σ^2.

Nonindependence. The residuals e_i are not independent random variables because they involve the fitted values $\hat{Y}_i$ which are based on the same fitted regression function. As a result, the residuals for regression model (2.1) are subject to two constraints. These are constraint (1.17)—that the sum of the e_i must be 0—and constraint (1.19)—that the products $X_i e_i$ must sum to 0.

When the sample size is large in comparison to the number of parameters in the regression model, the dependency effect among the residuals e_i is relatively unimportant and can be ignored for most purposes.

Semistudentized Residuals

At times, it is helpful to standardize the residuals for residual analysis. Since the standard deviation of the error terms ε_i is σ, which is estimated by $\sqrt{MSE}$, it is natural to consider the following form of standardization:

$$e_i^* = \frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}} \tag{3.5}$$

If $\sqrt{MSE}$ were an estimate of the standard deviation of the residual e_i, we would call e_i^* a studentized residual. However, the standard deviation of e_i is complex and varies for the different residuals e_i, and $\sqrt{MSE}$ is only an approximation of the standard deviation of e_i. Hence, we call the statistic e_i^* in (3.5) a *semistudentized residual*. We shall take up studentized residuals in Chapter 10. Both semistudentized residuals and studentized residuals can be very helpful in identifying outlying observations.

Departures from Model to Be Studied by Residuals

We shall consider the use of residuals for examining six important types of departures from the simple linear regression model (2.1) with normal errors:

1. The regression function is not linear.
2. The error terms do not have constant variance.
3. The error terms are not independent.
4. The model fits all but one or a few outlier observations.
5. The error terms are not normally distributed.
6. One or several important predictor variables have been omitted from the model.

3.3 Diagnostics for Residuals

We take up now some informal diagnostic plots of residuals to provide information on whether any of the six types of departures from the simple linear regression model (2.1) just mentioned are present. The following plots of residuals (or semistudentized residuals) will be utilized here for this purpose:

1. Plot of residuals against predictor variable.
2. Plot of absolute or squared residuals against predictor variable.
3. Plot of residuals against fitted values.
4. Plot of residuals against time or other sequence.
5. Plots of residuals against omitted predictor variables.
6. Box plot of residuals.
7. Normal probability plot of residuals.

FIGURE 3.2 MINITAB and SYGRAPH Diagnostic Residual Plots—Toluca Company Example.

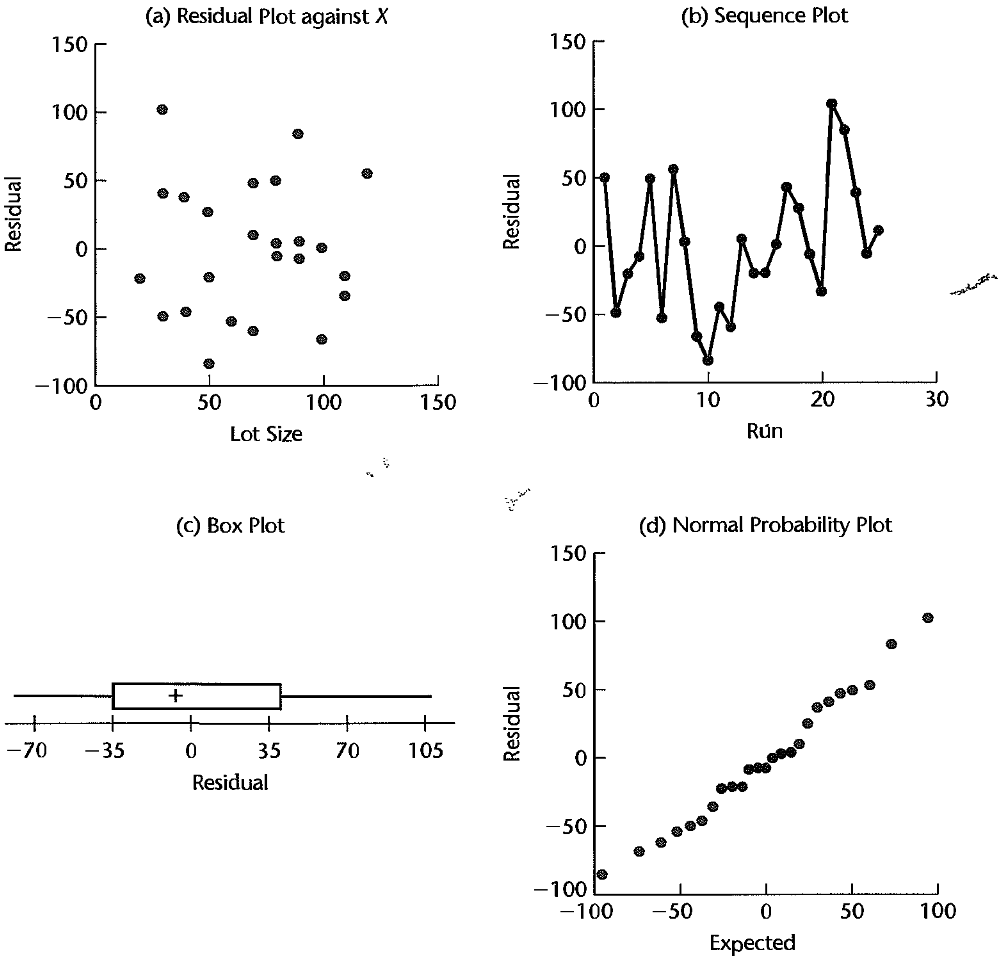


Figure 3.2 contains, for the Toluca Company example, MINITAB and SYGRAPH plots of the residuals in Table 1.2 against the predictor variable and against time, a box plot, and a normal probability plot. All of these plots, as we shall see, support the appropriateness of regression model (2.1) for the data.

We turn now to consider how residual analysis can be helpful in studying each of the six departures from regression model (2.1).

Nonlinearity of Regression Function

Whether a linear regression function is appropriate for the data being analyzed can be studied from a *residual plot against the predictor variable* or, equivalently, from a *residual plot against the fitted values.* Nonlinearity of the regression function can also be studied from a *scatter plot,* but this plot is not always as effective as a residual plot. Figure 3.3a

FIGURE 3.3 Scatter Plot and Residual Plot Illustrating Nonlinear Regression Function—Transit Example.

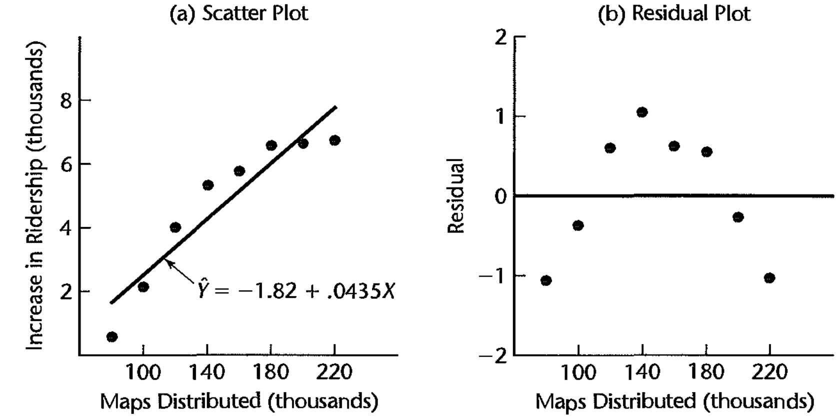


TABLE 3.1 Number of Maps Distributed and Increase in Ridership—Transit Example.

City i	(1) Increase in Ridership (thousands) Y_i	(2) Maps Distributed (thousands) X_i	(3) Fitted Value $\hat{Y}_i$	(4) Residual $Y_i - \hat{Y}_i = e_i$
1	.60	80	1.66	−1.06
2	6.70	220	7.75	−1.05
3	5.30	140	4.27	1.03
4	4.00	120	3.40	.60
5	6.55	180	6.01	.54
6	2.15	100	2.53	−.38
7	6.60	200	6.88	−.28
8	5.75	160	5.14	.61

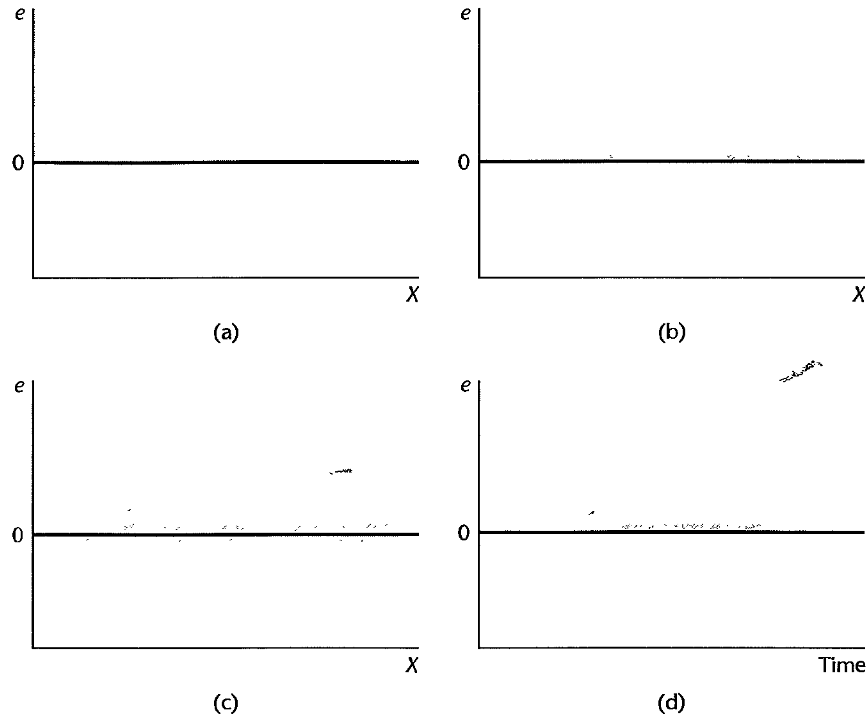
$$\hat{Y} = -1.82 + .0435X$$

contains a scatter plot of the data and the fitted regression line for a study of the relation between maps distributed and bus ridership in eight test cities. Here, X is the number of bus transit maps distributed free to residents of the city at the beginning of the test period and Y is the increase during the test period in average daily bus ridership during nonpeak hours. The original data and fitted values are given in Table 3.1, columns 1, 2, and 3. The plot suggests strongly that a linear regression function is not appropriate.

Figure 3.3b presents a plot of the residuals, shown in Table 3.1, column 4, against the predictor variable X. The lack of fit of the linear regression function is even more strongly suggested by the residual plot against X in Figure 3.3b than by the scatter plot. Note that the residuals depart from 0 in a systematic fashion; they are negative for smaller X values, positive for medium-size X values, and negative again for large X values.

In this case, both Figures 3.3a and 3.3b point out the lack of linearity of the regression function. In general, however, the residual plot is to be preferred, because it has some important advantages over the scatter plot. First, the residual plot can easily be used for examining other facets of the aptness of the model. Second, there are occasions when the

FIGURE 3.4 Prototype Residual Plots.



scaling of the scatter plot places the Y_i observations close to the fitted values $\hat{Y}_i$, for instance, when there is a steep slope. It then becomes more difficult to study the appropriateness of a linear regression function from the scatter plot. A residual plot, on the other hand, can clearly show any systematic pattern in the deviations around the fitted regression line under these conditions.

Figure 3.4a shows a prototype situation of the residual plot against X when a linear regression model is appropriate. The residuals then fall within a horizontal band centered around 0, displaying no systematic tendencies to be positive and negative. This is the case in Figure 3.2a for the Toluca Company example.

Figure 3.4b shows a prototype situation of a departure from the linear regression model that indicates the need for a curvilinear regression function. Here the residuals tend to vary in a systematic fashion between being positive and negative. This is the case in Figure 3.3b for the transit example. A different type of departure from linearity would, of course, lead to a picture different from the prototype pattern in Figure 3.4b.

Comment

A plot of residuals against the fitted values $\hat{Y}$ provides equivalent information as a plot of residuals against X for the simple linear regression model, and thus is not needed in addition to the residual plot against X. The two plots provide the same information because the fitted values $\hat{Y}_i$ are a linear function of the values X_i for the predictor variable. Thus, only the X scale values, not the basic pattern of the plotted points, are affected by whether the residual plot is against the X_i or the $\hat{Y}_i$. For curvilinear regression and multiple regression, on the other hand, separate plots of the residuals against the fitted values and against the predictor variable(s) are usually helpful. ■

Nonconstancy of Error Variance

Plots of the residuals against the predictor variable or against the fitted values are not only helpful to study whether a linear regression function is appropriate but also to examine whether the variance of the error terms is constant. Figure 3.5a shows a residual plot against age for a study of the relation between diastolic blood pressure of healthy, adult women (Y) and their age (X). The plot suggests that the older the woman is, the more spread out the residuals are. Since the relation between blood pressure and age is positive, this suggests that the error variance is larger for older women than for younger ones.

The prototype plot in Figure 3.4a exemplifies residual plots when the error term variance is constant. The residual plot in Figure 3.2a for the Toluca Company example is of this type, suggesting that the error terms have constant variance here.

Figure 3.4c shows a prototype picture of residual plots when the error variance increases with X. In many business, social science, and biological science applications, departures from constancy of the error variance tend to be of the "megaphone" type shown in Figure 3.4c, as in the blood pressure example in Figure 3.5a. One can also encounter error variances decreasing with increasing levels of the predictor variable and occasionally varying in some more complex fashion.

Plots of the absolute values of the residuals or of the squared residuals against the predictor variable X or against the fitted values $\hat{Y}$ are also useful for diagnosing nonconstancy of the error variance since the signs of the residuals are not meaningful for examining the constancy of the error variance. These plots are especially useful when there are not many cases in the data set because plotting of either the absolute or squared residuals places all of the information on changing magnitudes of the residuals above the horizontal zero line so that one can more readily see whether the magnitude of the residuals (irrespective of sign) is changing with the level of X or $\hat{Y}$.

Figure 3.5b contains a plot of the absolute residuals against age for the blood pressure example. This plot shows more clearly that the residuals tend to be larger in absolute magnitude for older-aged women.

FIGURE 3.5 Residual Plots Illustrating Nonconstant Error Variance.

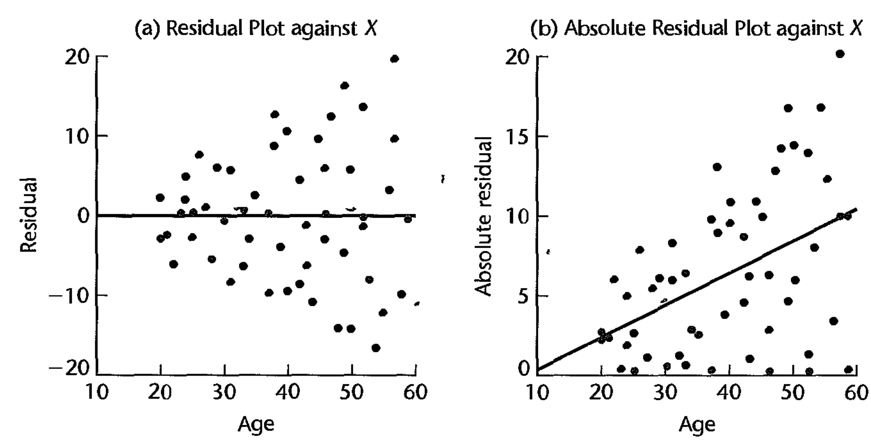
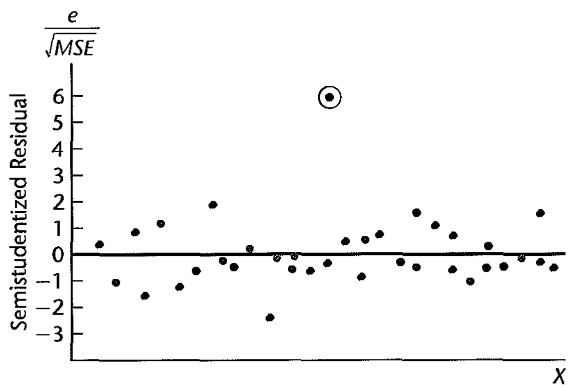


FIGURE 3.6 Residual Plot with Outlier.



Presence of Outliers

Outliers are extreme observations. Residual outliers can be identified from *residual plots against X or $\hat{Y}$*, as well as from *box plots, stem-and-leaf plots*, and *dot plots* of the residuals. Plotting of semistudentized residuals is particularly helpful for distinguishing outlying observations, since it then becomes easy to identify residuals that lie many standard deviations from zero. A rough rule of thumb when the number of cases is large is to consider semistudentized residuals with absolute value of four or more to be outliers. We shall take up more refined procedures for identifying outliers in Chapter 10.

The residual plot in Figure 3.6 presents semistudentized residuals and contains one outlier, which is circled. Note that this residual represents an observation almost six standard deviations from the fitted value.

Outliers can create great difficulty. When we encounter one, our first suspicion is that the observation resulted from a mistake or other extraneous effect, and hence should be discarded. A major reason for discarding it is that under the least squares method, a fitted line may be pulled disproportionately toward an outlying observation because the sum of the *squared* deviations is minimized. This could cause a misleading fit if indeed the outlying observation resulted from a mistake or other extraneous cause. On the other hand, outliers may convey significant information, as when an outlier occurs because of an interaction with another predictor variable omitted from the model. A safe rule frequently suggested is to discard an outlier only if there is direct evidence that it represents an error in recording, a miscalculation, a malfunctioning of equipment, or a similar type of circumstance.

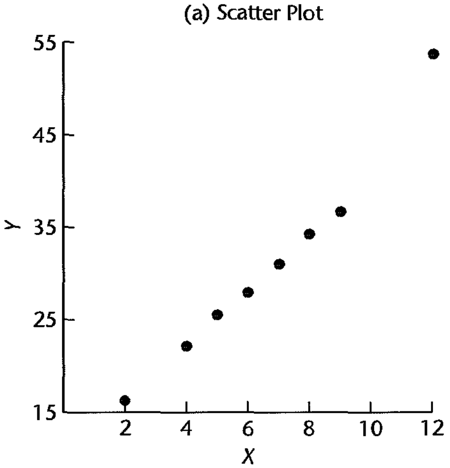
Comment

When a linear regression model is fitted to a data set with a small number of cases and an outlier is present, the fitted regression can be so distorted by the outlier that the residual plot may improperly suggest a lack of fit of the linear regression model, in addition to flagging the outlier. Figure 3.7 illustrates this situation. The scatter plot in Figure 3.7a presents a situation where all observations except the outlier fall around a straight-line statistical relationship. When a linear regression function is fitted to these data, the outlier causes such a shift in the fitted regression line as to lead to a systematic pattern of deviations from the fitted line for the other observations, suggesting a lack of fit of the linear regression function. This is shown by the residual plot in Figure 3.7b. ■

Nonindependence of Error Terms

Whenever data are obtained in a time sequence or some other type of sequence, such as for adjacent geographic areas, it is a good idea to prepare a *sequence plot of the residuals.*

FIGURE 3.7 Distorting Effect on Residuals Caused by an Outlier When Remaining Data Follow Linear Regression.



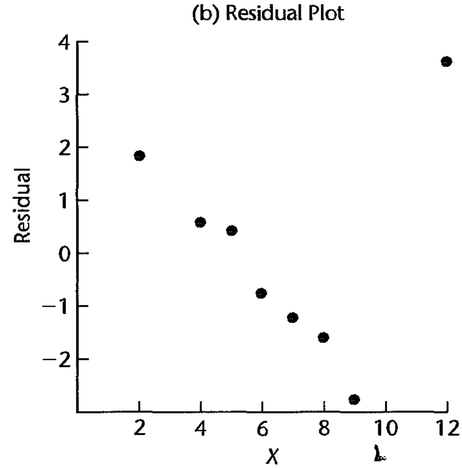
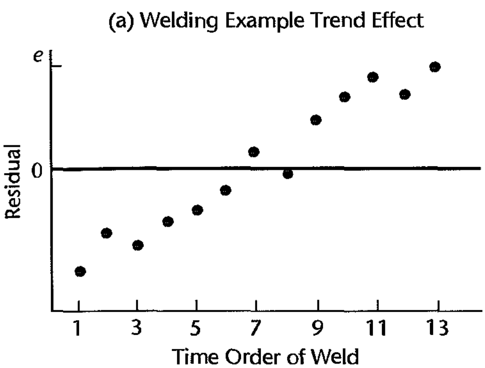
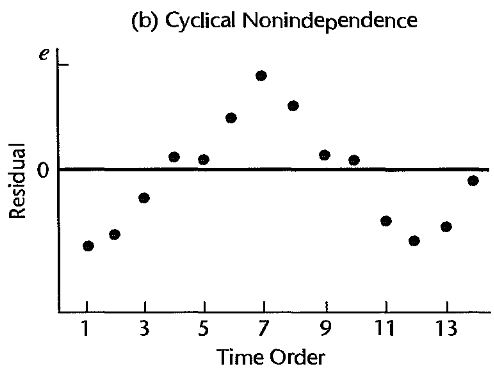


FIGURE 3.8 Residual Time Sequence Plots Illustrating Nonindependence of Error Terms.





The purpose of plotting the residuals against time or in some other type of sequence is to see if there is any correlation between error terms that are near each other in the sequence. Figure 3.8a contains a time sequence plot of the residuals in an experiment to study the relation between the diameter of a weld (X) and the shear strength of the weld (Y): An evident correlation between the error terms stands out. Negative residuals are associated mainly with the early trials, and positive residuals with the later trials. Apparently, some effect connected with time was present, such as learning by the welder or a gradual change in the welding equipment, so the shear strength tended to be greater in the later welds because of this effect.

A prototype residual plot showing a time-related trend effect is presented in Figure 3.4d, which portrays a linear time-related trend effect, as in the welding example. It is sometimes useful to view the problem of nonindependence of the error terms as one in which an important variable (in this case, time) has been omitted from the model. We shall discuss this type of problem shortly.

Another type of nonindependence of the error terms is illustrated in Figure 3.8b. Here the adjacent error terms are also related, but the resulting pattern is a cyclical one with no trend effect present.

When the error terms are independent, we expect the residuals in a sequence plot to fluctuate in a more or less random pattern around the base line 0, such as the scattering shown in Figure 3.2b for the Toluca Company example. Lack of randomness can take the form of too much or too little alternation of points around the zero line. In practice, there is little concern with the former because it does not arise frequently. Too little alternation, in contrast, frequently occurs, as in the welding example in Figure 3.8a.

Comment

When the residuals are plotted against X, as in Figure 3.3b for the transit example, the scatter may not appear to be random. For this plot, however, the basic problem is probably not lack of independence of the error terms but a poorly fitting regression function. This, indeed, is the situation portrayed in the scatter plot in Figure 3.3a. ■

Nonnormality of Error Terms

As we noted earlier, small departures from normality do not create any serious problems. Major departures, on the other hand, should be of concern. The normality of the error terms can be studied informally by examining the residuals in a variety of graphic ways.

Distribution Plots. A *box plot* of the residuals is helpful for obtaining summary information about the symmetry of the residuals and about possible outliers. Figure 3.2c contains a box plot of the residuals in the Toluca Company example. No serious departures from symmetry are suggested by this plot. A *histogram, dot plot,* or *stem-and-leaf plot* of the residuals can also be helpful for detecting gross departures from normality. However, the number of cases in the regression study must be reasonably large for any of these plots to convey reliable information about the shape of the distribution of the error terms.

Comparison of Frequencies. Another possibility when the number of cases is reasonably large is to compare actual frequencies of the residuals against expected frequencies under normality. For example, one can determine whether, say, about 68 percent of the residuals e_i fall between $\pm\sqrt{MSE}$ or about 90 percent fall between $\pm 1.645\sqrt{MSE}$. When the sample size is moderately large, corresponding t values may be used for the comparison.

To illustrate this procedure, we again consider the Toluca Company example of Chapter 1. Table 3.2, column 1, repeats the residuals from Table 1.2. We see from Figure 2.2 that $\sqrt{MSE} = 48.82$. Using the t distribution, we expect under normality about 90 percent of the residuals to fall between $\pm t(.95; 23)\sqrt{MSE} = \pm 1.714(48.82)$, or between -83.68 and 83.68. Actually, 22 residuals, or 88 percent, fall within these limits. Similarly, under normality, we expect about 60 percent of the residuals to fall between -41.89 and 41.89. The actual percentage here is 52 percent. Thus, the actual frequencies here are reasonably consistent with those expected under normality.

Normal Probability Plot. Still another possibility is to prepare a *normal probability plot of the residuals*. Here each residual is plotted against its expected value under normality. A plot that is nearly linear suggests agreement with normality, whereas a plot that departs substantially from linearity suggests that the error distribution is not normal.

Table 3.2, column 1, contains the residuals for the Toluca Company example. To find the expected values of the ordered residuals under normality, we utilize the facts that (1)

TABLE 3.2 Residuals and Expected Values under Normality—Toluca Company Example.

Run i	(1) Residual e_i	(2) Rank k	(3) Expected Value under Normality
1	51.02	22	51.95
2	−48.47	5	−44.10
3	−19.88	10	−14.76
...	...	...	...
23	38.83	19	31.05
24	−5.98	13	0
25	10.72	17	19.93

the expected value of the error terms for regression model (2.1) is zero and (2) the standard deviation of the error terms is estimated by $\sqrt{MSE}$. Statistical theory has shown that for a normal random variable with mean 0 and estimated standard deviation $\sqrt{MSE}$, a good approximation of the expected value of the kth smallest observation in a random sample of n is:

$$\sqrt{MSE}\left[z\left(\frac{k-.375}{n+.25}\right)\right] \tag{3.6}$$

where $z(A)$ as usual denotes the $(A)100$ percentile of the standard normal distribution.

Using this approximation, let us calculate the expected values of the residuals under normality for the Toluca Company example. Column 2 of Table 3.2 shows the ranks of the residuals, with the smallest residual being assigned rank 1. We see that the rank of the residual for run 1, $e_1 = 51.02$, is 22, which indicates that this residual is the 22nd smallest among the 25 residuals. Hence, for this residual $k = 22$. We found earlier (Table 2.1) that $MSE = 2{,}384$. Hence:

$$\frac{k-.375}{n+.25} = \frac{22-.375}{25+.25} = \frac{21.625}{25.25} = .8564$$

so that the expected value of this residual under normality is:

$$\sqrt{2{,}384}[z(.8564)] = \sqrt{2{,}384}(1.064) = 51.95$$

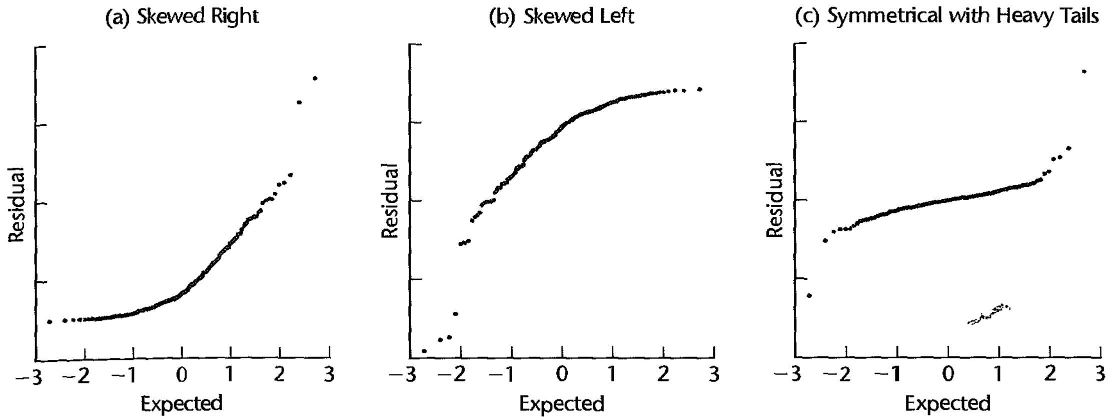
Similarly, the expected value of the residual for run 2, $e_2 = -48.47$, is obtained by noting that the rank of this residual is $k = 5$; in other words, this residual is the fifth smallest one among the 25 residuals. Hence, we require $(k - .375)/(n + .25) = (5 - .375)/(25 + .25) = .1832$, so that the expected value of this residual under normality is:

$$\sqrt{2{,}384}[z(.1832)] = \sqrt{2{,}384}(-.9032) = -44.10$$

Table 3.2, column 3, contains the expected values under the assumption of normality for a portion of the 25 residuals. Figure 3.2d presents a plot of the residuals against their expected values under normality. Note that the points in Figure 3.2d fall reasonably close to a straight line, suggesting that the distribution of the error terms does not depart substantially from a normal distribution.

Figure 3.9 shows three normal probability plots when the distribution of the error terms departs substantially from normality. Figure 3.9a shows a normal probability plot when the error term distribution is highly skewed to the right. Note the concave-upward shape

FIGURE 3.9 Normal Probability Plots when Error Term Distribution Is Not Normal.



of the plot. Figure 3.9b shows a normal probability plot when the error term distribution is highly skewed to the left. Here, the pattern is concave downward. Finally, Figure 3.9c shows a normal probability plot when the distribution of the error terms is symmetrical but has heavy tails; in other words, the distribution has higher probabilities in the tails than a normal distribution. Note the concave-downward curvature in the plot at the left end, corresponding to the plot for a left-skewed distribution, and the concave-upward plot at the right end, corresponding to a right-skewed distribution.

Comments

1. Many computer packages will prepare normal probability plots, either automatically or at the option of the user. Some of these plots utilize semistudentized residuals, others omit the factor $\sqrt{MSE}$ in (3.6), but neither of these variations affect the nature of the plot.

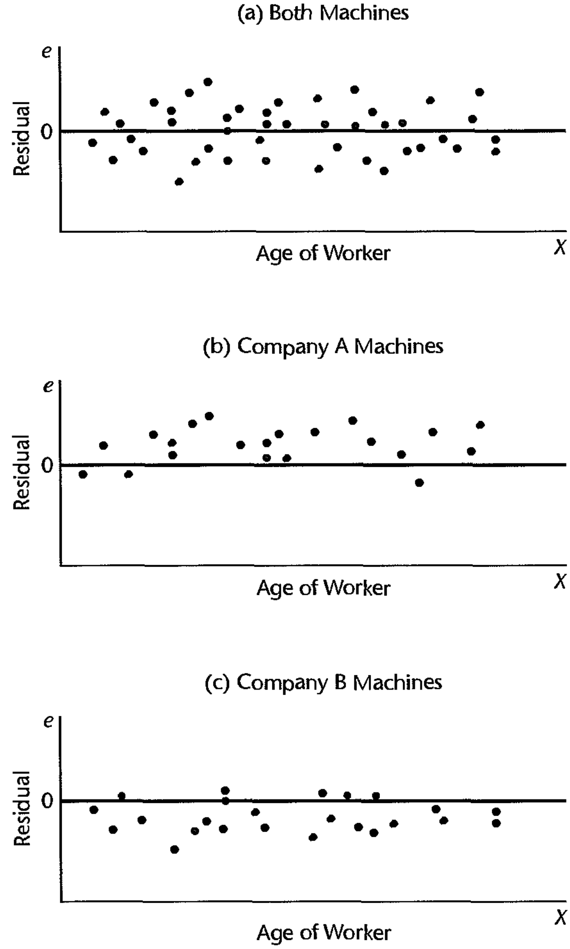
2. For continuous data, ties among the residuals should occur only rarely. If two residuals do have the same value, a simple procedure is to use the average rank for the tied residuals for calculating the corresponding expected values. ■

Difficulties in Assessing Normality. The analysis for model departures with respect to normality is, in many respects, more difficult than that for other types of departures. In the first place, random variation can be particularly mischievous when studying the nature of a probability distribution unless the sample size is quite large. Even worse, other types of departures can and do affect the distribution of the residuals. For instance, residuals may appear to be not normally distributed because an inappropriate regression function is used or because the error variance is not constant. Hence, it is usually a good strategy to investigate these other types of departures first, before concerning oneself with the normality of the error terms.

Omission of Important Predictor Variables

Residuals should also be plotted against variables omitted from the model that might have important effects on the response. The time variable cited earlier in the welding example is

FIGURE 3.10 **Residual Plots for Possible Omission of Important Predictor Variable—Productivity Example.**



an illustration. The purpose of this additional analysis is to determine whether there are any other key variables that could provide important additional descriptive and predictive power to the model.

As another example, in a study to predict output by piece-rate workers in an assembling operation, the relation between output (Y) and age (X) of worker was studied for a sample of employees. The plot of the residuals against X, shown in Figure 3.10a, indicates no ground for suspecting the appropriateness of the linearity of the regression function or the constancy of the error variance. Since machines produced by two companies (A and B) are used in the assembling operation and could have an effect on output, residual plots against X by type of machine were undertaken and are shown in Figures 3.10b and 3.10c. Note that the residuals for Company A machines tend to be positive, while those for Company B machines tend to be negative. Thus, type of machine appears to have a definite effect on productivity, and output predictions may turn out to be far superior when this variable is added to the model.

While this second example dealt with a qualitative variable (type of machine), the residual analysis for an additional quantitative variable is analogous. The residuals are plotted against the additional predictor variable to see whether or not the residuals tend to vary systematically with the level of the additional predictor variable.

Comment

We do not say that the original model is "wrong" when it can be improved materially by adding one or more predictor variables. Only a few of the factors operating on any response variable Y in real-world situations can be included explicitly in a regression model. The chief purpose of residual analysis in identifying other important predictor variables is therefore to test the adequacy of the model and see whether it could be improved materially by adding one or more predictor variables. ■

Some Final Comments

1. We discussed model departures one at a time. In actuality, several types of departures may occur together. For instance, a linear regression function may be a poor fit and the variance of the error terms may not be constant. In these cases, the prototype patterns of Figure 3.4 can still be useful, but they would need to be combined into composite patterns.

2. Although graphic analysis of residuals is only an informal method of analysis, in many cases it suffices for examining the aptness of a model.

3. The basic approach to residual analysis explained here applies not only to simple linear regression but also to more complex regression and other types of statistical models.

4. Several types of departures from the simple linear regression model have been identified by diagnostic tests of the residuals. Model misspecification due to either nonlinearity or the omission of important predictor variables tends to be serious, leading to biased estimates of the regression parameters and error variance. These problems are discussed further in Section 3.9 and Chapter 10. Nonconstancy of error variance tends to be less serious, leading to less efficient estimates and invalid error variance estimates. The problem is discussed in depth in Section 11.1. The presence of outliers can be serious for smaller data sets when their influence is large. Influential outliers are discussed further in Section 10.4. Finally, the nonindependence of error terms results in estimators that are unbiased but whose variances are seriously biased. Alternative estimation methods for correlated errors are discussed in Chapter 12.

3.4 Overview of Tests Involving Residuals

Graphic analysis of residuals is inherently subjective. Nevertheless, subjective analysis of a variety of interrelated residual plots will frequently reveal difficulties with the model more clearly than particular formal tests. There are occasions, however, when one wishes to put specific questions to a test. We now briefly review some of the relevant tests.

Most statistical tests require independent observations. As we have seen, however, the residuals are dependent. Fortunately, the dependencies become quite small for large samples, so that one can usually then ignore them.

Tests for Randomness

A runs test is frequently used to test for lack of randomness in the residuals arranged in time order. Another test, specifically designed for lack of randomness in least squares residuals, is the Durbin-Watson test. This test is discussed in Chapter 12.

Tests for Constancy of Variance

When a residual plot gives the impression that the variance may be increasing or decreasing in a systematic manner related to X or $E\{Y\}$, a simple test is based on the rank correlation between the absolute values of the residuals and the corresponding values of the predictor variable. Two other simple tests for constancy of the error variance—the Brown-Forsythe test and the Breusch-Pagan test—are discussed in Section 3.6.

Tests for Outliers

A simple test for identifying an outlier observation involves fitting a new regression line to the other $n - 1$ observations. The suspect observation, which was not used in fitting the new line, can now be regarded as a new observation. One can calculate the probability that in n observations, a deviation from the fitted line as great as that of the outlier will be obtained by chance. If this probability is sufficiently small, the outlier can be rejected as not having come from the same population as the other $n - 1$ observations. Otherwise, the outlier is retained. We discuss this approach in detail in Chapter 10.

Many other tests to aid in evaluating outliers have been developed. These are discussed in specialized references, such as Reference 3.1.

Tests for Normality

Goodness of fit tests can be used for examining the normality of the error terms. For instance, the chi-square test or the Kolmogorov-Smirnov test and its modification, the Lilliefors test, can be employed for testing the normality of the error terms by analyzing the residuals. A simple test based on the normal probability plot of the residuals will be taken up in Section 3.5.

Comment

The runs test, rank correlation, and goodness of fit tests are commonly used statistical procedures and are discussed in many basic statistics texts. ■

3.5 Correlation Test for Normality

In addition to visually assessing the approximate linearity of the points plotted in a normal probability plot, a formal test for normality of the error terms can be conducted by calculating the coefficient of correlation (2.74) between the residuals e_i and their expected values under normality. A high value of the correlation coefficient is indicative of normality. Table B.6, prepared by Looney and Gulledge (Ref. 3.2), contains critical values (percentiles) for various sample sizes for the distribution of the coefficient of correlation between the ordered residuals and their expected values under normality when the error terms are normally distributed. If the observed coefficient of correlation is at least as large as the tabled value, for a given α level, one can conclude that the error terms are reasonably normally distributed.

Example

For the Toluca Company example in Table 3.2, the coefficient of correlation between the ordered residuals and their expected values under normality is .991. Controlling the α risk at .05, we find from Table B.6 that the critical value for $n = 25$ is .959. Since the observed coefficient exceeds this level, we have support for our earlier conclusion that the distribution of the error terms does not depart substantially from a normal distribution.

Comment

The correlation test for normality presented here is simpler than the Shapiro-Wilk test (Ref. 3.3), which can be viewed as being based approximately also on the coefficient of correlation between the ordered residuals and their expected values under normality. ■

3.6 Tests for Constancy of Error Variance

We present two formal tests for ascertaining whether the error terms have constant variance: the Brown-Forsythe test and the Breusch-Pagan test.

Brown-Forsythe Test

The Brown-Forsythe test, a modification of the Levene test (Ref. 3.4), does not depend on normality of the error terms. Indeed, this test is robust against serious departures from normality, in the sense that the nominal significance level remains approximately correct when the error terms have equal variances even if the distribution of the error terms is far from normal. Yet the test is still relatively efficient when the error terms are normally distributed. The Brown-Forsythe test as described is applicable to simple linear regression when the variance of the error terms either increases or decreases with X, as illustrated in the prototype megaphone plot in Figure 3.4c. The sample size needs to be large enough so that the dependencies among the residuals can be ignored.

The test is based on the variability of the residuals. The larger the error variance, the larger the variability of the residuals will tend to be. To conduct the Brown-Forsythe test, we divide the data set into two groups, according to the level of X, so that one group consists of cases where the X level is comparatively low and the other group consists of cases where the X level is comparatively high. If the error variance is either increasing or decreasing with X, the residuals in one group will tend to be more variable than those in the other group. Equivalently, the absolute deviations of the residuals around their group mean will tend to be larger for one group than for the other group. In order to make the test more robust, we utilize the absolute deviations of the residuals around the median for the group (Ref. 3.5). The Brown-Forsythe test then consists simply of the two-sample t test based on test statistic (A.67) to determine whether the mean of the absolute deviations for one group differs significantly from the mean absolute deviation for the second group.

Although the distribution of the absolute deviations of the residuals is usually not normal, it has been shown that the t^* test statistic still follows approximately the t distribution when the variance of the error terms is constant and the sample sizes of the two groups are not extremely small.

We shall now use e_{i1} to denote the ith residual for group 1 and e_{i2} to denote the ith residual for group 2. Also we shall use n_1 and n_2 to denote the sample sizes of the two groups, where:

$$n = n_1 + n_2 \tag{3.7}$$

Further, we shall use $\tilde{e}_1$ and $\tilde{e}_2$ to denote the medians of the residuals in the two groups. The Brown-Forsythe test uses the absolute deviations of the residuals around their group median, to be denoted by d_{i1} and d_{i2}:

$$d_{i1} = |e_{i1} - \tilde{e}_1| \qquad d_{i2} = |e_{i2} - \tilde{e}_2| \tag{3.8}$$

With this notation, the two-sample t test statistic (A.67) becomes:

$$t^*_{BF} = \frac{\bar{d}_1 - \bar{d}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}} \tag{3.9}$$

where $\bar{d}_1$ and $\bar{d}_2$ are the sample means of the d_{i1} and d_{i2}, respectively, and the pooled variance s^2 in (A.63) becomes:

$$s^2 = \frac{\sum(d_{i1} - \bar{d}_1)^2 + \sum(d_{i2} - \bar{d}_2)^2}{n - 2} \tag{3.9a}$$

We denote the test statistic for the Brown-Forsythe test by t^*_{BF}.

If the error terms have constant variance and n_1 and n_2 are not extremely small, t^*_{BF} follows approximately the t distribution with $n - 2$ degrees of freedom. Large absolute values of t^*_{BF} indicate that the error terms do not have constant variance.

Example

We wish to use the Brown-Forsythe test for the Toluca Company example to determine whether or not the error term variance varies with the level of X. Since the X levels are spread fairly uniformly (see Figure 3.1a), we divide the 25 cases into two groups with approximately equal X ranges. The first group consists of the 13 runs with lot sizes from 20 to 70. The second group consists of the 12 runs with lot sizes from 80 to 120. Table 3.3

TABLE 3.3 Calculations for Brown-Forsythe Test for Constancy of Error Variance—Toluca Company Example.

Group 1

i	Run	(1) Lot Size	(2) Residual e_{i1}	(3) d_{i1}	(4) $(d_{i1} - \bar{d}_1)^2$
1	14	20	−20.77	.89	1,929.41
2	2	30	−48.47	28.59	263.25
...	...	...	...	...	...
12	12	70	−60.28	40.40	19.49
13	25	70	10.72	30.60	202.07
	Total			582.60	12,566.6

$\tilde{e}_1 = -19.88 \qquad \bar{d}_1 = 44.815$

Group 2

i	Run	(1) Lot Size	(2) Residual e_{i2}	(3) d_{i2}	(4) $(d_{i2} - \bar{d}_2)^2$
1	1	80	−51.02	53.70	637.56
2	8	80	4.02	6.70	473.06
...	...	...	...	...	...
11	20	110	−34.09	31.41	8.76
12	7	120	55.21	57.89	866.71
	Total			341.40	9,610.2

$\tilde{e}_2 = -2.68 \qquad \bar{d}_2 = 28.450$

presents a portion of the data for each group. In columns 1 and 2 are repeated the lot sizes and residuals from Table 1.2. We see from Table 3.3 that the median residual is $\tilde{e}_1 = -19.88$ for group 1 and $\tilde{e}_2 = -2.68$ for group 2. Column 3 contains the absolute deviations of the residuals around their respective group medians. For instance, we obtain:

$$d_{11} = |e_{11} - \tilde{e}_1| = |-20.77 - (-19.88)| = .89$$
$$d_{12} = |e_{12} - \tilde{e}_2| = |51.02 - (-2.68)| = 53.70$$

The means of the absolute deviations are obtained in the usual fashion:

$$\bar{d}_1 = \frac{582.60}{13} = 44.815 \qquad \bar{d}_2 = \frac{341.40}{12} = 28.450$$

Finally, column 4 contains the squares of the deviations of the d_{i1} and d_{i2} around their respective group means. For instance, we have:

$$(d_{11} - \bar{d}_1)^2 = (.89 - 44.815)^2 = 1,929.41$$
$$(d_{12} - \bar{d}_2)^2 = (53.70 - 28.450)^2 = 637.56$$

We are now ready to calculate test statistic (3.9):

$$s^2 = \frac{12,566.6 + 9,610.2}{25 - 2} = 964.21$$
$$s = 31.05$$
$$t_{BF}^* = \frac{44.815 - 28.450}{31.05\sqrt{\frac{1}{13} + \frac{1}{12}}} = 1.32$$

To control the α risk at .05, we require $t(.975; 23) = 2.069$. The decision rule therefore is:

If $|t_{BF}^*| \leq 2.069$, conclude the error variance is constant

If $|t_{BF}^*| > 2.069$, conclude the error variance is not constant

Since $|t_{BF}^*| = 1.32 \leq 2.069$, we conclude that the error variance is constant and does not vary with the level of X. The two-sided P-value of this test is .20.

Comments

1. If the data set contains many cases, the two-sample t test for constancy of error variance can be conducted after dividing the cases into three or four groups, according to the level of X, and using the two extreme groups.

2. A robust test for constancy of the error variance is desirable because nonnormality and lack of constant variance often go hand in hand. For example, the distribution of the error terms may become increasingly skewed and hence more variable with increasing levels of X. ■

Breusch-Pagan Test

A second test for the constancy of the error variance is the Breusch-Pagan test (Ref. 3.6). This test, a large-sample test, assumes that the error terms are independent and normally distributed and that the variance of the error term ε_i, denoted by σ_i^2, is related to the level of X in the following way:

$$\log_e \sigma_i^2 = \gamma_0 + \gamma_1 X_i \tag{3.10}$$

Note that (3.10) implies that σ_i^2 either increases or decreases with the level of X, depending on the sign of γ_1. Constancy of error variance corresponds to $\gamma_1 = 0$. The test of H_0: $\gamma_1 = 0$ versus H_a: $\gamma_1 \neq 0$ is carried out by means of regressing the squared residuals e_i^2 against X_i in the usual manner and obtaining the regression sum of squares, to be denoted by SSR^*. The test statistic X_{BP}^2 is as follows:

$$X_{BP}^2 = \frac{SSR^*}{2} \div \left(\frac{SSE}{n}\right)^2 \tag{3.11}$$

where SSR^* is the regression sum of squares when regressing e^2 on X and SSE is the error sum of squares when regressing Y on X. If H_0: $\gamma_1 = 0$ holds and n is reasonably large, X_{BP}^2 follows approximately the chi-square distribution with one degree of freedom. Large values of X_{BP}^2 lead to conclusion H_a, that the error variance is not constant.

Example

To conduct the Breusch-Pagan test for the Toluca Company example, we regress the squared residuals in Table 1.2, column 5, against X and obtain $SSR^* = 7{,}896{,}128$. We know from Figure 2.2 that $SSE = 54{,}825$. Hence, test statistic (3.11) is:

$$X_{BP}^2 = \frac{7{,}896{,}128}{2} \div \left(\frac{54{,}825}{25}\right)^2 = .821$$

To control the α risk at .05, we require $\chi^2(.95;\, 1) = 3.84$. Since $X_{BP}^2 = .821 \leq 3.84$, we conclude H_0, that the error variance is constant. The P-value of this test is .64 so that the data are quite consistent with constancy of the error variance.

Comments

1. The Breusch-Pagan test can be modified to allow for different relationships between the error variance and the level of X than the one in (3.10).
2. Test statistic (3.11) was developed independently by Cook and Weisberg (Ref. 3.7), and the test is sometimes referred to as the Cook-Weisberg test. ■

3.7 *F* Test for Lack of Fit

We next take up a formal test for determining whether a specific type of regression function adequately fits the data. We illustrate this test for ascertaining whether a linear regression function is a good fit for the data.

Assumptions

The lack of fit test assumes that the observations Y for given X are (1) independent and (2) normally distributed, and that (3) the distributions of Y have the same variance σ^2.

The lack of fit test requires repeat observations at one or more X levels. In nonexperimental data, these may occur fortuitously, as when in a productivity study relating workers' output and age, several workers of the same age happen to be included in the study. In an experiment, one can assure by design that there are repeat observations. For instance, in an experiment on the effect of size of salesperson bonus on sales, three salespersons can be offered a particular size of bonus, for each of six bonus sizes, and their sales then observed.

Repeat trials for the same level of the predictor variable, of the type described, are called *replications*. The resulting observations are called *replicates*.

Example

In an experiment involving 12 similar but scattered suburban branch offices of a commercial bank, holders of checking accounts at the offices were offered gifts for setting up money market accounts. Minimum initial deposits in the new money market account were specified to qualify for the gift. The value of the gift was directly proportional to the specified minimum deposit. Various levels of minimum deposit and related gift values were used in the experiment in order to ascertain the relation between the specified minimum deposit and gift value, on the one hand, and number of accounts opened at the office, on the other. Altogether, six levels of minimum deposit and proportional gift value were used, with two of the branch offices assigned at random to each level. One branch office had a fire during the period and was dropped from the study. Table 3.4a contains the results, where X is the amount of minimum deposit and Y is the number of new money market accounts that were opened and qualified for the gift during the test period.

A linear regression function was fitted in the usual fashion; it is:

$$\hat{Y} = 50.72251 + .48670X$$

The analysis of variance table also was obtained and is shown in Table 3.4b. A scatter plot, together with the fitted regression line, is shown in Figure 3.11. The indications are strong that a linear regression function is inappropriate. To test this formally, we shall use the general linear test approach described in Section 2.8.

TABLE 3.4 Data and Analysis of Variance Table—Bank Example.

(a) Data

Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i	Branch i	Size of Minimum Deposit (dollars) X_i	Number of New Accounts Y_i
1	125	160	7	75	42
2	100	112	8	175	124
3	200	124	9	125	150
4	75	28	10	200	104
5	150	152	11	100	136
6	175	156			

(b) ANOVA Table

Source of Variation	SS	df	MS
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Total	19,882.9	10	

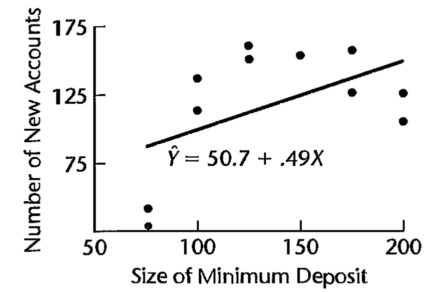


FIGURE 3.11 Scatter Plot and Fitted Regression Line—Bank Example.

TABLE 3.5 Data Arranged by Replicate Number and Minimum Deposit—Bank Example.

	Size of Minimum Deposit (dollars)					
Replicate	$j=1$ $X_1 = 75$	$j=2$ $X_2 = 100$	$j=3$ $X_3 = 125$	$j=4$ $X_4 = 150$	$j=5$ $X_5 = 175$	$j=6$ $X_6 = 200$
$i=1$	28	112	160	152	156	124
$i=2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

Notation

First, we need to modify our notation to recognize the existence of replications at some levels of X. Table 3.5 presents the same data as Table 3.4a, but in an arrangement that recognizes the replicates. We shall denote the different X levels in the study, whether or not replicated observations are present, as $X_1, \ldots, X_c$. For the bank example, $c = 6$ since there are six minimum deposit size levels in the study, for five of which there are two observations and for one there is a single observation. We shall let $X_1 = 75$ (the smallest minimum deposit level), $X_2 = 100, \ldots, X_6 = 200$. Further, we shall denote the number of replicates for the jth level of X as n_j; for our example, $n_1 = n_2 = n_3 = n_5 = n_6 = 2$ and $n_4 = 1$. Thus, the total number of observations n is given by:

$$n = \sum_{j=1}^{c} n_j \tag{3.12}$$

We shall denote the observed value of the response variable for the ith replicate for the jth level of X by Y_{ij}, where $i = 1, \ldots, n_j$, $j = 1, \ldots, c$. For the bank example (Table 3.5), $Y_{11} = 28$, $Y_{21} = 42$, $Y_{12} = 112$, and so on. Finally, we shall denote the mean of the Y observations at the level $X = X_j$ by $\bar{Y}_j$. Thus, $\bar{Y}_1 = (28 + 42)/2 = 35$ and $\bar{Y}_4 = 152/1 = 152$.

Full Model

The general linear test approach begins with the specification of the full model. The full model used for the lack of fit test makes the same assumptions as the simple linear regression model (2.1) except for assuming a linear regression relation, the subject of the test. This full model is:

$$Y_{ij} = \mu_j + \varepsilon_{ij} \qquad \text{Full model} \tag{3.13}$$

where:

μ_j are parameters $j = 1, \ldots, c$
ε_{ij} are independent $N(0, \sigma^2)$

Since the error terms have expectation zero, it follows that:

$$E\{Y_{ij}\} = \mu_j \tag{3.14}$$

Thus, the parameter μ_j $(j = 1, \ldots, c)$ is the mean response when $X = X_j$.

The full model (3.13) is like the regression model (2.1) in stating that each response Y is made up of two components: the mean response when $X = X_j$ and a random error term. The difference between the two models is that in the full model (3.13) there are no restrictions on the means μ_j, whereas in the regression model (2.1) the mean responses are linearly related to X (i.e., $E\{Y\} = \beta_0 + \beta_1 X$).

To fit the full model to the data, we require the least squares or maximum likelihood estimators for the parameters μ_j. It can be shown that these estimators of μ_j are simply the sample means $\bar{Y}_j$:

$$\hat{\mu}_j = \bar{Y}_j \tag{3.15}$$

Thus, the estimated expected value for observation Y_{ij} is $\bar{Y}_j$, and the error sum of squares for the full model therefore is:

$$SSE(F) = \sum_j \sum_i (Y_{ij} - \bar{Y}_j)^2 = SSPE \tag{3.16}$$

In the context of the test for lack of fit, the full model error sum of squares (3.16) is called the *pure error sum of squares* and is denoted by *SSPE*.

Note that *SSPE* is made up of the sums of squared deviations at each X level. At level $X = X_j$, this sum of squared deviations is:

$$\sum_i (Y_{ij} - \bar{Y}_j)^2 \tag{3.17}$$

These sums of squares are then added over all of the X levels $(j = 1, \ldots, c)$. For the bank example, we have:

$$\begin{aligned} SSPE = {} & (28 - 35)^2 + (42 - 35)^2 + (112 - 124)^2 + (136 - 124)^2 + (160 - 155)^2 \\ & + (150 - 155)^2 + (152 - 152)^2 + (156 - 140)^2 + (124 - 140)^2 \\ & + (124 - 114)^2 + (104 - 114)^2 \\ = {} & 1{,}148 \end{aligned}$$

Note that any X level with no replications makes no contribution to *SSPE* because $\bar{Y}_j = Y_{1j}$ then. Thus, $(152 - 152)^2 = 0$ for $j = 4$ in the bank example.

The degrees of freedom associated with *SSPE* can be obtained by recognizing that the sum of squared deviations (3.17) at a given level of X is like an ordinary total sum of squares based on n observations, which has $n - 1$ degrees of freedom associated with it. Here, there are n_j observations when $X = X_j$; hence the degrees of freedom are $n_j - 1$. Just as *SSPE* is the sum of the sums of squares (3.17), so the number of degrees of freedom associated with $SSPE$ is the sum of the component degrees of freedom:

$$df_F = \sum_j (n_j - 1) = \sum_j n_j - c = n - c \tag{3.18}$$

For the bank example, we have $df_F = 11 - 6 = 5$. Note that any X level with no replications makes no contribution to df_F because $n_j - 1 = 1 - 1 = 0$ then, just as such an X level makes no contribution to $SSPE$.

Reduced Model

The general linear test approach next requires consideration of the reduced model under H_0. For testing the appropriateness of a linear regression relation, the alternatives are:

$$\begin{aligned} H_0&: E\{Y\} = \beta_0 + \beta_1 X \\ H_a&: E\{Y\} \neq \beta_0 + \beta_1 X \end{aligned} \tag{3.19}$$

Thus, H_0 postulates that μ_j in the full model (3.13) is linearly related to X_j:

$$\mu_j = \beta_0 + \beta_1 X_j$$

The reduced model under H_0 therefore is:

$$Y_{ij} = \beta_0 + \beta_1 X_j + \varepsilon_{ij} \qquad \text{Reduced model} \tag{3.20}$$

Note that the reduced model is the ordinary simple linear regression model (2.1), with the subscripts modified to recognize the existence of replications. We know that the estimated expected value for observation Y_{ij} with regression model (2.1) is the fitted value $\hat{Y}_{ij}$:

$$\hat{Y}_{ij} = b_0 + b_1 X_j \tag{3.21}$$

Hence, the error sum of squares for the reduced model is the usual error sum of squares SSE:

$$\begin{aligned} SSE(R) &= \sum\sum[Y_{ij} - (b_0 + b_1 X_j)]^2 \\ &= \sum\sum(Y_{ij} - \hat{Y}_{ij})^2 = SSE \end{aligned} \tag{3.22}$$

We also know that the degrees of freedom associated with $SSE(R)$ are:

$$df_R = n - 2$$

For the bank example, we have from Table 3.4b:

$$SSE(R) = SSE = 14{,}741.6$$
$$df_R = 9$$

Test Statistic

The general linear test statistic (2.70):

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

here becomes:

$$F^* = \frac{SSE - SSPE}{(n-2) - (n-c)} \div \frac{SSPE}{n-c} \tag{3.23}$$

The difference between the two error sums of squares is called the *lack of fit sum of squares* here and is denoted by *SSLF*:

$$SSLF = SSE - SSPE \tag{3.24}$$

We can then express the test statistic as follows:

$$\begin{aligned} F^* &= \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} \\ &= \frac{MSLF}{MSPE} \end{aligned} \tag{3.25}$$

where *MSLF* denotes the *lack of fit mean square* and *MSPE* denotes the *pure error mean square*.

We know that large values of F^* lead to conclusion H_a in the general linear test. Decision rule (2.71) here becomes:

$$\begin{aligned} &\text{If } F^* \leq F(1-\alpha; c-2, n-c), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; c-2, n-c), \text{ conclude } H_a \end{aligned} \tag{3.26}$$

For the bank example, the test statistic can be constructed easily from our earlier results:

$$\begin{aligned} SSPE &= 1{,}148.0 & n - c &= 11 - 6 = 5 \\ SSE &= 14{,}741.6 \\ SSLF &= 14{,}741.6 - 1{,}148.0 = 13{,}593.6 & c - 2 &= 6 - 2 = 4 \end{aligned}$$

$$\begin{aligned} F^* &= \frac{13{,}593.6}{4} \div \frac{1{,}148.0}{5} \\ &= \frac{3{,}398.4}{229.6} = 14.80 \end{aligned}$$

If the level of significance is to be $\alpha = .01$, we require $F(.99; 4, 5) = 11.4$. Since $F^* = 14.80 > 11.4$, we conclude H_a, that the regression function is not linear. This, of course, accords with our visual impression from Figure 3.11. The P-value for the test is .006.

ANOVA Table

The definition of the lack of fit sum of squares *SSLF* in (3.24) indicates that we have, in fact, decomposed the error sum of squares *SSE* into two components:

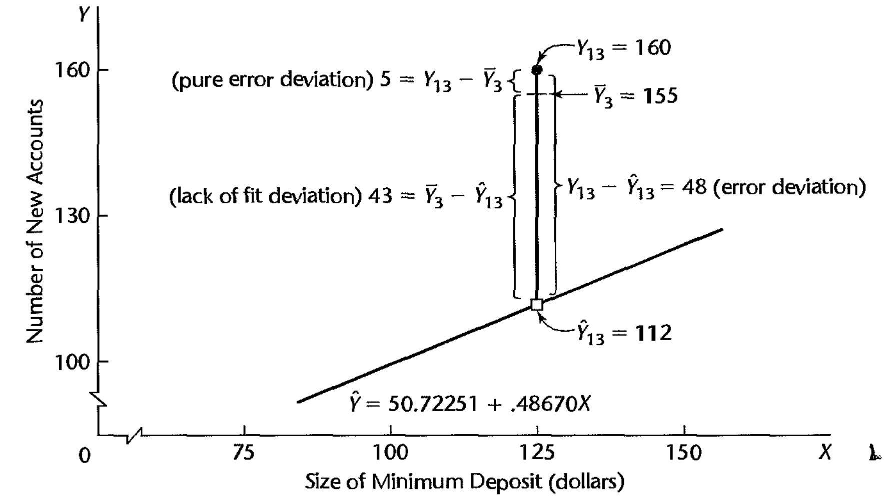
$$SSE = SSPE + SSLF \tag{3.27}$$

This decomposition follows from the identity:

$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\substack{\text{Error}\\\text{deviation}}} = \underbrace{Y_{ij} - \bar{Y}_j}_{\substack{\text{Pure error}\\\text{deviation}}} + \underbrace{\bar{Y}_j - \hat{Y}_{ij}}_{\substack{\text{Lack of fit}\\\text{deviation}}} \tag{3.28}$$

This identity shows that the error deviations in *SSE* are made up of a pure error component and a lack of fit component. Figure 3.12 illustrates this partitioning for the case $Y_{13} = 160$, $X_3 = 125$ in the bank example.

FIGURE 3.12 Illustration of Decomposition of Error Deviation $Y_{ij} - \hat{Y}_{ij}$—Bank Example.



When (3.28) is squared and summed over all observations, we obtain (3.27) since the cross-product sum equals zero:

$$\underset{SSE}{\sum\sum(Y_{ij}-\hat{Y}_{ij})^2} = \underset{SSPE}{\sum\sum(Y_{ij}-\bar{Y}_j)^2} + \underset{SSLF}{\sum\sum(\bar{Y}_j-\hat{Y}_{ij})^2} \tag{3.29}$$

Note from (3.29) that we can define the lack of fit sum of squares directly as follows:

$$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2 \tag{3.30}$$

Since all Y_{ij} observations at the level X_j have the same fitted value, which we can denote by $\hat{Y}_j$, we can express (3.30) equivalently as:

$$SSLF = \sum_j n_j(\bar{Y}_j - \hat{Y}_j)^2 \tag{3.30a}$$

Formula (3.30a) indicates clearly why *SSLF* measures lack of fit. If the linear regression function is appropriate, then the means $\bar{Y}_j$ will be near the fitted values $\hat{Y}_j$ calculated from the estimated linear regression function and *SSLF* will be small. On the other hand, if the linear regression function is not appropriate, the means $\bar{Y}_j$ will not be near the fitted values calculated from the estimated linear regression function, as in Figure 3.11 for the bank example, and *SSLF* will be large.

Formula (3.30a) also indicates why $c - 2$ degrees of freedom are associated with *SSLF*. There are c means $\bar{Y}_j$ in the sum of squares, and two degrees of freedom are lost in estimating the parameters β_0 and β_1 of the linear regression function to obtain the fitted values $\hat{Y}_j$.

An ANOVA table can be constructed for the decomposition of *SSE*. Table 3.6a contains the general ANOVA table, including the decomposition of *SSE* just explained and the mean squares of interest, and Table 3.6b contains the ANOVA decomposition for the bank example.

TABLE 3.6 General ANOVA Table for Testing Lack of Fit of Simple Linear Regression Function and ANOVA Table—Bank Example.

(a) General

Source of Variation	SS	df	MS
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

(b) Bank Example

Source of Variation	SS	df	MS
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Lack of fit	13,593.6	4	3,398.4
Pure error	1,148.0	5	229.6
Total	19,882.9	10	

Comments

1. As shown by the bank example, not all levels of X need have repeat observations for the F test for lack of fit to be applicable. Repeat observations at only one or some levels of X are sufficient.

2. It can be shown that the mean squares $MSPE$ and $MSLF$ have the following expectations when testing whether the regression function is linear:

$$E\{MSPE\} = \sigma^2 \tag{3.31}$$

$$E\{MSLF\} = \sigma^2 + \frac{\sum n_j[\mu_j - (\beta_0 + \beta_1 X_j)]^2}{c-2} \tag{3.32}$$

The reason for the term "pure error" is that $MSPE$ is always an unbiased estimator of the error term variance σ^2, no matter what is the true regression function. The expected value of $MSLF$ also is σ^2 if the regression function is linear, because $\mu_j = \beta_0 + \beta_1 X_j$ then and the second term in (3.32) becomes zero. On the other hand, if the regression function is not linear, $\mu_j \neq \beta_0 + \beta_1 X_j$ and $E\{MSLF\}$ will be greater than σ^2. Hence, a value of F^* near 1 accords with a linear regression function; large values of F^* indicate that the regression function is not linear.

3. The terminology "error sum of squares" and "error mean square" is not precise when the regression function under test in H_0 is not the true function since the error sum of squares and error mean square then reflect the effects of both the lack of fit and the variability of the error terms. We continue to use the terminology for consistency and now use the term "pure error" to identify the variability associated with the error term only.

4. Suppose that prior to any analysis of the appropriateness of the model, we had fitted a linear regression model and wished to test whether or not $\beta_1 = 0$ for the bank example (Table 3.4b). Test statistic (2.60) would be:

$$F^* = \frac{MSR}{MSE} = \frac{5{,}141.3}{1{,}638.0} = 3.14$$

For $\alpha = .10$, $F(.90; 1, 9) = 3.36$, and we would conclude H_0, that $\beta_1 = 0$ or that there is no *linear association* between minimum deposit size (and value of gift) and number of new accounts. A conclusion that there is no *relation* between these variables would be improper, however. Such an inference requires that regression model (2.1) be appropriate. Here, there is a definite relationship, but the regression function is not linear. This illustrates the importance of always examining the appropriateness of a model before any inferences are drawn.

5. The general linear test approach just explained can be used to test the appropriateness of other regression functions. Only the degrees of freedom for $SSLF$ will need be modified. In general, $c - p$ degrees of freedom are associated with $SSLF$, where p is the number of parameters in the regression function. For the test of a simple linear regression function, $p = 2$ because there are two parameters, β_0 and β_1, in the regression function.

6. The alternative H_a in (3.19) includes all regression functions other than a linear one. For instance, it includes a quadratic regression function or a logarithmic one. If H_a is concluded, a study of residuals can be helpful in identifying an appropriate function.

7. When we conclude that the employed model in H_0 is appropriate, the usual practice is to use the error mean square MSE as an estimator of σ^2 in preference to the pure error mean square $MSPE$, since the former contains more degrees of freedom.

8. Observations at the same level of X are genuine repeats only if they involve independent trials with respect to the error term. Suppose that in a regression analysis of the relation between hardness (Y) and amount of carbon (X) in specimens of an alloy, the error term in the model covers, among other things, random errors in the measurement of hardness by the analyst and effects of uncontrolled production factors, which vary at random from specimen to specimen and affect hardness. If the analyst takes two readings on the hardness of a specimen, this will not provide a genuine replication because the effects of random variation in the production factors are fixed in any given specimen. For genuine replications, different specimens with the same carbon content (X) would have to be measured by the analyst so that *all* the effects covered in the error term could vary at random from one repeated observation to the next.

9. When no replications are present in a data set, an approximate test for lack of fit can be conducted if there are some cases at adjacent X levels for which the mean responses are quite close to each other. Such adjacent cases are grouped together and treated as pseudoreplicates, and the test for lack of fit is then carried out using these groupings of adjacent cases. A useful summary of this and related procedures for conducting a test for lack of fit when no replicates are present may be found in Reference 3.8. ■

3.8 Overview of Remedial Measures

If the simple linear regression model (2.1) is not appropriate for a data set, there are two basic choices:

1. Abandon regression model (2.1) and develop and use a more appropriate model.
2. Employ some transformation on the data so that regression model (2.1) is appropriate for the transformed data.

Each approach has advantages and disadvantages. The first approach may entail a more complex model that could yield better insights, but may also lead to more complex procedures for estimating the parameters. Successful use of transformations, on the other hand, leads to relatively simple methods of estimation and may involve fewer parameters than a complex model, an advantage when the sample size is small. Yet transformations may obscure the fundamental interconnections between the variables, though at other times they may illuminate them.

We consider the use of transformations in this chapter and the use of more complex models in later chapters. First, we provide a brief overview of remedial measures.

Nonlinearity of Regression Function

When the regression function is not linear, a direct approach is to modify regression model (2.1) by altering the nature of the regression function. For instance, a quadratic regression function might be used:

$$E\{Y\} = \beta_0 + \beta_1 X + \beta_2 X^2$$

or an exponential regression function:

$$E\{Y\} = \beta_0 \beta_1^X$$

In Chapter 7, we discuss polynomial regression functions, and in Part III we take up nonlinear regression functions, such as an exponential regression function.

The transformation approach employs a transformation to linearize, at least approximately, a nonlinear regression function. We discuss the use of transformations to linearize regression functions in Section 3.9.

When the nature of the regression function is not known, exploratory analysis that does not require specifying a particular type of function is often useful. We discuss exploratory regression analysis in Section 3.10.

Nonconstancy of Error Variance

When the error variance is not constant but varies in a systematic fashion, a direct approach is to modify the model to allow for this and use the method of *weighted least squares* to obtain the estimators of the parameters. We discuss the use of weighted least squares for this purpose in Chapter 11.

Transformations can also be effective in stabilizing the variance. Some of these are discussed in Section 3.9.

Nonindependence of Error Terms

When the error terms are correlated, a direct remedial measure is to work with a model that calls for correlated error terms. We discuss such a model in Chapter 12. A simple remedial transformation that is often helpful is to work with first differences, a topic also discussed in Chapter 12.

Nonnormality of Error Terms

Lack of normality and nonconstant error variances frequently go hand in hand. Fortunately, it is often the case that the same transformation that helps stabilize the variance is also helpful in approximately normalizing the error terms. It is therefore desirable that the transformation for stabilizing the error variance be utilized first, and then the residuals studied to see if serious departures from normality are still present. We discuss transformations to achieve approximate normality in Section 3.9.

Omission of Important Predictor Variables

When residual analysis indicates that an important predictor variable has been omitted from the model, the solution is to modify the model. In Chapter 6 and later chapters, we discuss multiple regression analysis in which two or more predictor variables are utilized.

Outlying Observations

When outlying observations are present, as in Figure 3.7a, use of the least squares and maximum likelihood estimators (1.10) for regression model (2.1) may lead to serious distortions in the estimated regression function. When the outlying observations do not represent recording errors and should not be discarded, it may be desirable to use an estimation procedure that places less emphasis on such outlying observations. We discuss one such robust estimation procedure in Chapter 11.

3.9 Transformations

We now consider in more detail the use of transformations of one or both of the original variables before carrying out the regression analysis. Simple transformations of either the response variable Y or the predictor variable X, or of both, are often sufficient to make the simple linear regression model appropriate for the transformed data.

Transformations for Nonlinear Relation Only

We first consider transformations for linearizing a nonlinear regression relation when the distribution of the error terms is reasonably close to a normal distribution and the error terms have approximately constant variance. In this situation, transformations on X should be attempted. The reason why transformations on Y may not be desirable here is that a transformation on Y, such as $Y' = \sqrt{Y}$, may materially change the shape of the distribution of the error terms from the normal distribution and may also lead to substantially differing error term variances.

Figure 3.13 contains some prototype nonlinear regression relations with constant error variance and also presents some simple transformations on X that may be helpful to linearize the regression relationship without affecting the distributions of Y. Several alternative transformations may be tried. Scatter plots and residual plots based on each transformation should then be prepared and analyzed to decide which transformation is most effective.

Example

Data from an experiment on the effect of number of days of training received (X) on performance (Y) in a battery of simulated sales situations are presented in Table 3.7, columns 1 and 2, for the 10 participants in the study. A scatter plot of these data is shown in Figure 3.14a. Clearly the regression relation appears to be curvilinear, so the simple linear regression model (2.1) does not seem to be appropriate. Since the variability at the different X levels appears to be fairly constant, we shall consider a transformation on X. Based on the prototype plot in Figure 3.13a, we shall consider initially the square root transformation $X' = \sqrt{X}$. The transformed values are shown in column 3 of Table 3.7.

FIGURE 3.13 Prototype Nonlinear Regression Patterns with Constant Error Variance and Simple Transformations of X.

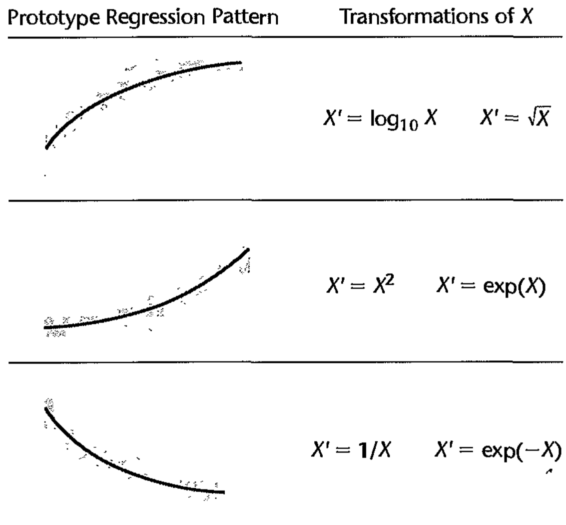


TABLE 3.7 Use of Square Root Transformation of X to Linearize Regression Relation—Sales Training Example.

Sales Trainee i	(1) Days of Training X_i	(2) Performance Score Y_i	(3) $X'_i = \sqrt{X_i}$
1	.5	42.5	.70711
2	.5	50.6	.70711
3	1.0	68.5	1.00000
4	1.0	80.7	1.00000
5	1.5	89.0	1.22474
6	1.5	99.6	1.22474
7	2.0	105.3	1.41421
8	2.0	111.8	1.41421
9	2.5	112.3	1.58114
10	2.5	125.7	1.58114

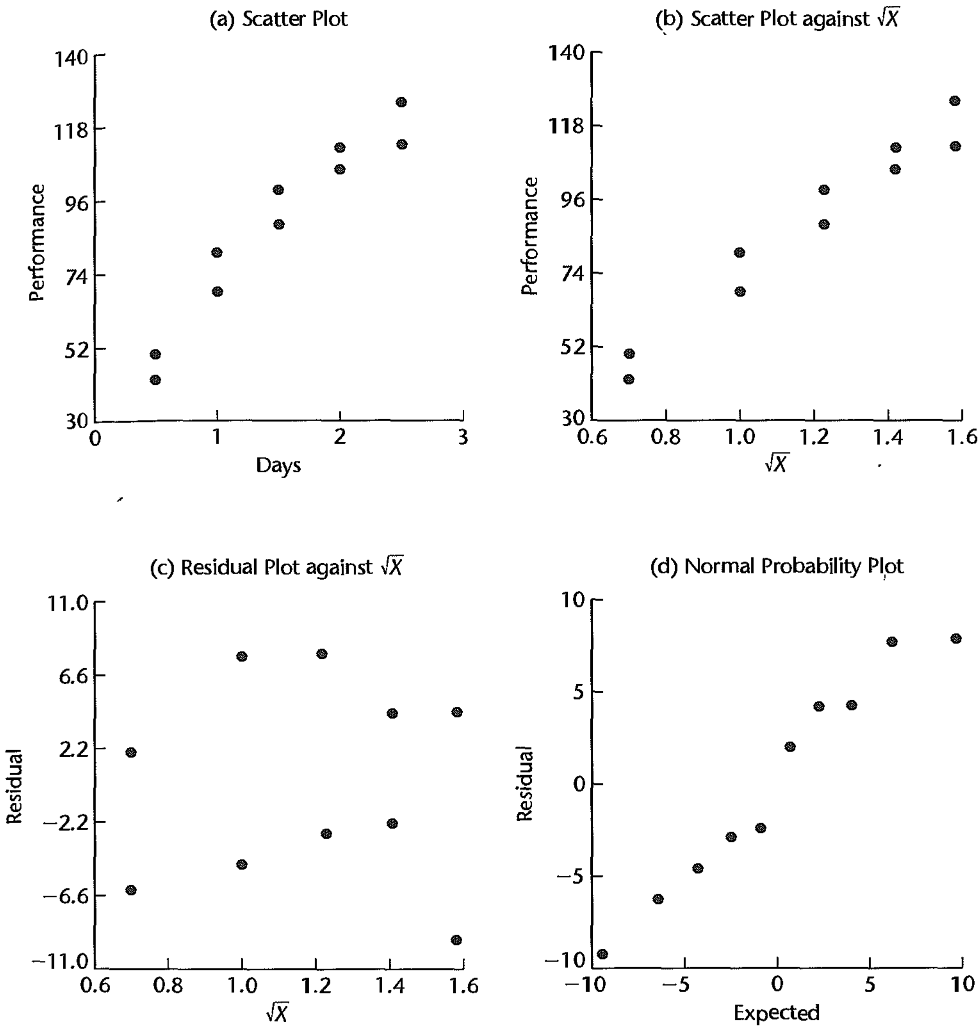
In Figure 3.14b, the same data are plotted with the predictor variable transformed to $X' = \sqrt{X}$. Note that the scatter plot now shows a reasonably linear relation. The variability of the scatter at the different X levels is the same as before, since we did not make a transformation on Y.

To examine further whether the simple linear regression model (2.1) is appropriate now, we fit it to the transformed X data. The regression calculations with the transformed X data are carried out in the usual fashion, except that the predictor variable now is X'. We obtain the following fitted regression function:

$$\hat{Y} = -10.33 + 83.45X'$$

Figure 3.14c contains a plot of the residuals against X'. There is no evidence of lack of fit or of strongly unequal error variances. Figure 3.14d contains a normal probability plot of

FIGURE 3.14 **Scatter Plots and Residual Plots—Sales Training Example.**



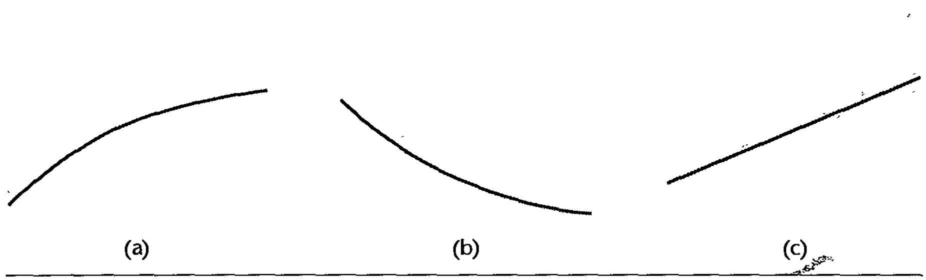
the residuals. No strong indications of substantial departures from normality are indicated by this plot. This conclusion is supported by the high coefficient of correlation between the ordered residuals and their expected values under normality, .979. For $\alpha = .01$, Table B.6 shows that the critical value is .879, so the observed coefficient is substantially larger and supports the reasonableness of normal error terms. Thus, the simple linear regression model (2.1) appears to be appropriate here for the transformed data.

The fitted regression function in the original units of X can easily be obtained, if desired:

$$\hat{Y} = -10.33 + 83.45\sqrt{X}$$

FIGURE 3.15 Prototype Regression Patterns with Unequal Error Variances and Simple Transformations of Y.

Prototype Regression Pattern



Transformations on Y

$Y' = \sqrt{Y}$

$Y' = \log_{10} Y$

$Y' = 1/Y$

Note: A simultaneous transformation on X may also be helpful or necessary.

Comment

At times, it may be helpful to introduce a constant into the transformation. For example, if some of the X data are near zero and the reciprocal transformation is desired, we can shift the origin by using the transformation $X' = 1/(X + k)$, where k is an appropriately chosen constant. ■

Transformations for Nonnormality and Unequal Error Variances

Unequal error variances and nonnormality of the error terms frequently appear together. To remedy these departures from the simple linear regression model (2.1), we need a transformation on Y, since the shapes and spreads of the distributions of Y need to be changed. Such a transformation on Y may also at the same time help to linearize a curvilinear regression relation. At other times, a simultaneous transformation on X may be needed to obtain or maintain a linear regression relation.

Frequently, the nonnormality and unequal variances departures from regression model (2.1) take the form of increasing skewness and increasing variability of the distributions of the error terms as the mean response $E\{Y\}$ increases. For example, in a regression of yearly household expenditures for vacations (Y) on household income (X), there will tend to be more variation and greater positive skewness (i.e., some very high yearly vacation expenditures) for high-income households than for low-income households, who tend to consistently spend much less for vacations. Figure 3.15 contains some prototype regression relations where the skewness and the error variance increase with the mean response $E\{Y\}$. This figure also presents some simple transformations on Y that may be helpful for these cases. Several alternative transformations on Y may be tried, as well as some simultaneous transformations on X. Scatter plots and residual plots should be prepared to determine the most effective transformation(s).

Example

Data on age (X) and plasma level of a polyamine (Y) for a portion of the 25 healthy children in a study are presented in columns 1 and 2 of Table 3.8. These data are plotted in Figure 3.16a as a scatter plot. Note the distinct curvilinear regression relationship, as well as the greater variability for younger children than for older ones.

TABLE 3.8 Use of Logarithmic Transformation of Y to Linearize Regression Relation and Stabilize Error Variance—Plasma Levels Example.

Child i	(1) Age X_i	(2) Plasma Level Y_i	(3) $Y_i' = \log_{10} Y_i$
1	0 (newborn)	13.44	1.1284
2	0 (newborn)	12.84	1.1086
3	0 (newborn)	11.91	1.0759
4	0 (newborn)	20.09	1.3030
5	0 (newborn)	15.60	1.1931
6	1.0	10.11	1.0048
7	1.0	11.38	1.0561
...	...	...	...
19	3.0	6.90	.8388
20	3.0	6.77	.8306
21	4.0	4.86	.6866
22	4.0	5.10	.7076
23	4.0	5.67	.7536
24	4.0	5.75	.7597
25	4.0	6.23	.7945

On the basis of the prototype regression pattern in Figure 3.15b, we shall first try the logarithmic transformation $Y' = \log_{10} Y$. The transformed Y values are shown in column 3 of Table 3.8. Figure 3.16b contains the scatter plot with this transformation. Note that the transformation not only has led to a reasonably linear regression relation, but the variability at the different levels of X also has become reasonably constant.

To further examine the reasonableness of the transformation $Y' = \log_{10} Y$, we fitted the simple linear regression model (2.1) to the transformed Y data and obtained:

$$\hat{Y}' = 1.135 - .1023X$$

A plot of the residuals against X is shown in Figure 3.16c, and a normal probability plot of the residuals is shown in Figure 3.16d. The coefficient of correlation between the ordered residuals and their expected values under normality is .981. For $\alpha = .05$, Table B.6 indicates that the critical value is .959 so that the observed coefficient supports the assumption of normality of the error terms. All of this evidence supports the appropriateness of regression model (2.1) for the transformed Y data.

Comments

1. At times it may be desirable to introduce a constant into a transformation of Y, such as when Y may be negative. For instance, the logarithmic transformation to shift the origin in Y and make all Y observations positive would be $Y' = \log_{10}(Y + k)$, where k is an appropriately chosen constant.

2. When unequal error variances are present but the regression relation is linear, a transformation on Y may not be sufficient. While such a transformation may stabilize the error variance, it will also change the linear relationship to a curvilinear one. A transformation on X may therefore also be required. This case can also be handled by using weighted least squares, a procedure explained in Chapter 11. ■

The difference between the two error sums of squares is called the *lack of fit sum of squares* here and is denoted by *SSLF*:

$$SSLF = SSE - SSPE \tag{3.24}$$

We can then express the test statistic as follows:

$$\begin{aligned} F^* &= \frac{SSLF}{c-2} \div \frac{SSPE}{n-c} \\ &= \frac{MSLF}{MSPE} \end{aligned} \tag{3.25}$$

where *MSLF* denotes the *lack of fit mean square* and *MSPE* denotes the *pure error mean square.*

We know that large values of F^* lead to conclusion H_a in the general linear test. Decision rule (2.71) here becomes:

$$\begin{aligned} &\text{If } F^* \leq F(1-\alpha; c-2, n-c), \text{ conclude } H_0 \\ &\text{If } F^* > F(1-\alpha; c-2, n-c), \text{ conclude } H_a \end{aligned} \tag{3.26}$$

For the bank example, the test statistic can be constructed easily from our earlier results:

$$SSPE = 1{,}148.0 \qquad n - c = 11 - 6 = 5$$

$$SSE = 14{,}741.6$$

$$SSLF = 14{,}741.6 - 1{,}148.0 = 13{,}593.6 \qquad c - 2 = 6 - 2 = 4$$

$$\begin{aligned} F^* &= \frac{13{,}593.6}{4} \div \frac{1{,}148.0}{5} \\ &= \frac{3{,}398.4}{229.6} = 14.80 \end{aligned}$$

If the level of significance is to be $\alpha = .01$, we require $F(.99; 4, 5) = 11.4$. Since $F^* = 14.80 > 11.4$, we conclude H_a, that the regression function is not linear. This, of course, accords with our visual impression from Figure 3.11. The P-value for the test is .006.

ANOVA Table

The definition of the lack of fit sum of squares *SSLF* in (3.24) indicates that we have, in fact, decomposed the error sum of squares *SSE* into two components:

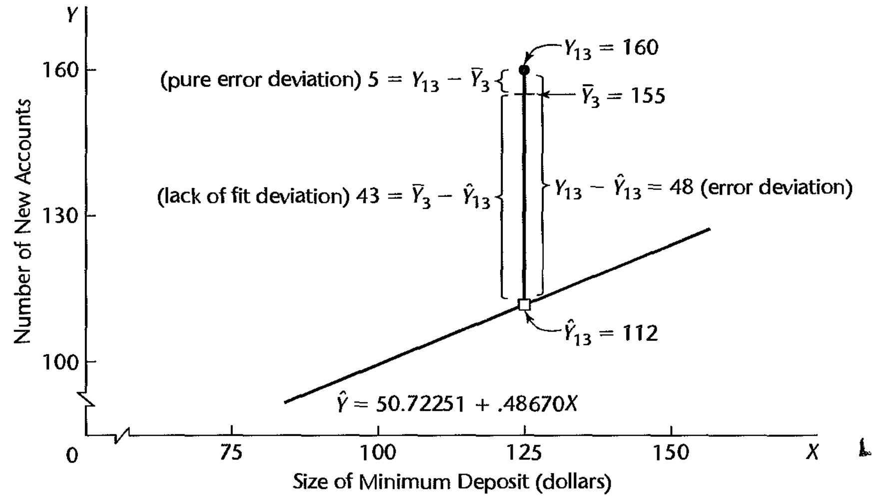
$$SSE = SSPE + SSLF \tag{3.27}$$

This decomposition follows from the identity:

$$\underbrace{Y_{ij} - \hat{Y}_{ij}}_{\text{Error deviation}} = \underbrace{Y_{ij} - \bar{Y}_j}_{\text{Pure error deviation}} + \underbrace{\bar{Y}_j - \hat{Y}_{ij}}_{\text{Lack of fit deviation}} \tag{3.28}$$

This identity shows that the error deviations in *SSE* are made up of a pure error component and a lack of fit component. Figure 3.12 illustrates this partitioning for the case $Y_{13} = 160$, $X_3 = 125$ in the bank example.

FIGURE 3.12 **Illustration of Decomposition of Error Deviation $Y_{ij} - \hat{Y}_{ij}$—Bank Example.**



When (3.28) is squared and summed over all observations, we obtain (3.27) since the cross-product sum equals zero:

$$\underset{SSE}{\sum\sum(Y_{ij} - \hat{Y}_{ij})^2} = \underset{SSPE}{\sum\sum(Y_{ij} - \bar{Y}_j)^2} + \underset{SSLF}{\sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2} \tag{3.29}$$

Note from (3.29) that we can define the lack of fit sum of squares directly as follows:

$$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2 \tag{3.30}$$

Since all Y_{ij} observations at the level X_j have the same fitted value, which we can denote by $\hat{Y}_j$, we can express (3.30) equivalently as:

$$SSLF = \sum_j n_j(\bar{Y}_j - \hat{Y}_j)^2 \tag{3.30a}$$

Formula (3.30a) indicates clearly why *SSLF* measures lack of fit. If the linear regression function is appropriate, then the means $\bar{Y}_j$ will be near the fitted values $\hat{Y}_j$ calculated from the estimated linear regression function and *SSLF* will be small. On the other hand, if the linear regression function is not appropriate, the means $\bar{Y}_j$ will not be near the fitted values calculated from the estimated linear regression function, as in Figure 3.11 for the bank example, and *SSLF* will be large.

Formula (3.30a) also indicates why $c - 2$ degrees of freedom are associated with *SSLF*. There are c means $\bar{Y}_j$ in the sum of squares, and two degrees of freedom are lost in estimating the parameters β_0 and β_1 of the linear regression function to obtain the fitted values $\hat{Y}_j$.

An ANOVA table can be constructed for the decomposition of *SSE*. Table 3.6a contains the general ANOVA table, including the decomposition of *SSE* just explained and the mean squares of interest, and Table 3.6b contains the ANOVA decomposition for the bank example.

TABLE 3.6 General ANOVA Table for Testing Lack of Fit of Simple Linear Regression Function and ANOVA Table—Bank Example.

(a) General

Source of Variation	SS	df	MS
Regression	$SSR = \sum\sum(\hat{Y}_{ij} - \bar{Y})^2$	1	$MSR = \frac{SSR}{1}$
Error	$SSE = \sum\sum(Y_{ij} - \hat{Y}_{ij})^2$	$n-2$	$MSE = \frac{SSE}{n-2}$
Lack of fit	$SSLF = \sum\sum(\bar{Y}_j - \hat{Y}_{ij})^2$	$c-2$	$MSLF = \frac{SSLF}{c-2}$
Pure error	$SSPE = \sum\sum(Y_{ij} - \bar{Y}_j)^2$	$n-c$	$MSPE = \frac{SSPE}{n-c}$
Total	$SSTO = \sum\sum(Y_{ij} - \bar{Y})^2$	$n-1$	

(b) Bank Example

Source of Variation	SS	df	MS
Regression	5,141.3	1	5,141.3
Error	14,741.6	9	1,638.0
Lack of fit	13,593.6	4	3,398.4
Pure error	1,148.0	5	229.6
Total	19,882.9	10	

Comments

1. As shown by the bank example, not all levels of X need have repeat observations for the F test for lack of fit to be applicable. Repeat observations at only one or some levels of X are sufficient.

2. It can be shown that the mean squares $MSPE$ and $MSLF$ have the following expectations when testing whether the regression function is linear:

$$E\{MSPE\} = \sigma^2 \tag{3.31}$$

$$E\{MSLF\} = \sigma^2 + \frac{\sum n_j[\mu_j - (\beta_0 + \beta_1 X_j)]^2}{c-2} \tag{3.32}$$

The reason for the term "pure error" is that $MSPE$ is always an unbiased estimator of the error term variance σ^2, no matter what is the true regression function. The expected value of $MSLF$ also is σ^2 if the regression function is linear, because $\mu_j = \beta_0 + \beta_1 X_j$ then and the second term in (3.32) becomes zero. On the other hand, if the regression function is not linear, $\mu_j \neq \beta_0 + \beta_1 X_j$ and $E\{MSLF\}$ will be greater than σ^2. Hence, a value of F^* near 1 accords with a linear regression function; large values of F^* indicate that the regression function is not linear.

3. The terminology "error sum of squares" and "error mean square" is not precise when the regression function under test in H_0 is not the true function since the error sum of squares and error mean square then reflect the effects of both the lack of fit and the variability of the error terms. We continue to use the terminology for consistency and now use the term "pure error" to identify the variability associated with the error term only.

4. Suppose that prior to any analysis of the appropriateness of the model, we had fitted a linear regression model and wished to test whether or not $\beta_1 = 0$ for the bank example (Table 3.4b). Test statistic (2.60) would be:

$$F^* = \frac{MSR}{MSE} = \frac{5{,}141.3}{1{,}638.0} = 3.14$$

For $\alpha = .10$, $F(.90; 1, 9) = 3.36$, and we would conclude H_0, that $\beta_1 = 0$ or that there is no *linear association* between minimum deposit size (and value of gift) and number of new accounts. A conclusion that there is no *relation* between these variables would be improper, however. Such an inference requires that regression model (2.1) be appropriate. Here, there is a definite relationship, but the regression function is not linear. This illustrates the importance of always examining the appropriateness of a model before any inferences are drawn.

5. The general linear test approach just explained can be used to test the appropriateness of other regression functions. Only the degrees of freedom for *SSLF* will need be modified. In general, $c - p$ degrees of freedom are associated with *SSLF*, where p is the number of parameters in the regression function. For the test of a simple linear regression function, $p = 2$ because there are two parameters, β_0 and β_1, in the regression function.

6. The alternative H_a in (3.19) includes all regression functions other than a linear one. For instance, it includes a quadratic regression function or a logarithmic one. If H_a is concluded, a study of residuals can be helpful in identifying an appropriate function.

7. When we conclude that the employed model in H_0 is appropriate, the usual practice is to use the error mean square *MSE* as an estimator of σ^2 in preference to the pure error mean square *MSPE*, since the former contains more degrees of freedom.

8. Observations at the same level of X are genuine repeats only if they involve independent trials with respect to the error term. Suppose that in a regression analysis of the relation between hardness (Y) and amount of carbon (X) in specimens of an alloy, the error term in the model covers, among other things, random errors in the measurement of hardness by the analyst and effects of uncontrolled production factors, which vary at random from specimen to specimen and affect hardness. If the analyst takes two readings on the hardness of a specimen, this will not provide a genuine replication because the effects of random variation in the production factors are fixed in any given specimen. For genuine replications, different specimens with the same carbon content (X) would have to be measured by the analyst so that *all* the effects covered in the error term could vary at random from one repeated observation to the next.

9. When no replications are present in a data set, an approximate test for lack of fit can be conducted if there are some cases at adjacent X levels for which the mean responses are quite close to each other. Such adjacent cases are grouped together and treated as pseudoreplicates, and the test for lack of fit is then carried out using these groupings of adjacent cases. A useful summary of this and related procedures for conducting a test for lack of fit when no replicates are present may be found in Reference 3.8. ■

3.8 Overview of Remedial Measures

If the simple linear regression model (2.1) is not appropriate for a data set, there are two basic choices:

1. Abandon regression model (2.1) and develop and use a more appropriate model.
2. Employ some transformation on the data so that regression model (2.1) is appropriate for the transformed data.