Example 3.2 Consumption Expenditures Suppose we wish to build a two-variable model that explains the dollar value of aggregate consumption expenditures C, measured in billions of dollars (seasonally adjusted).[6] As an explanatory variable we use aggregate personal disposable income Y, measured in billions of dollars (seasonally adjusted). When C is regressed on Y using quarterly data from the first quarter of 1959 to the second quarter of 1995, we obtain the following result (standard errors are in parentheses):

$$\underset{}{C} = \underset{(4.45)}{-27.53} + \underset{(.0018)}{.93Y}$$

In this case, the intercept of -27.53 is significant at the 5 percent level (the t statistic is -6.18 ($-27.53/4.45$). More important, the t statistic associated with the coefficient of disposable income is 517 (.93/.0018). We can clearly reject the null hypothesis of a zero slope in favor of the alternative hypothesis that the slope is nonzero. Rejection of the null hypothesis allows us to accept—at least provisionally—the two-variable regression model. Of course, further research might allow us to find a model of aggregate consumption expenditures that is more suitable than the one just described.

Suppose (for illustrative purposes) we replace Y as an explanatory variable by a *random* variable. (We chose a variable X that was drawn each time from a normal distribution with a mean of 50 and a variance of 25.) Then we would expect that approximately 1 time in 20 the coefficient on the X variable would be significantly different from zero (at the 5 percent significance level). We found that it took 22 trials before a significantly negative coefficient was obtained. This shows that no matter how reliable or unreliable a statistical estimator is, there is always a statistical chance that one will make incorrect inferences by relying on the regression results.

3.4 ANALYSIS OF VARIANCE AND CORRELATION

3.4.1 Goodness of Fit

Regression residuals can provide a useful measure of the fit between the estimated regression line and the data. A good regression equation is one which helps explain a large proportion of the variance of Y. Large residuals imply a poor fit, while small residuals imply a good fit. The problem with using the residual as a measure of goodness of fit is that its value depends on the units of the dependent variable. To find a measure of goodness of fit which is unit-

[6] This example uses data supplied by the Citibase database. The original data (GC and GYD) are seasonally adjusted *at annual rates.*

free, it seems reasonable to use the residual variance divided by the variation of Y.

$$\text{Variation } (Y) = \Sigma(Y_i - \bar{Y})^2$$

Our goal is to divide the variation of Y into two parts, the first accounted for by the regression equation and the second associated with the unexplained portion (the error term) of the model. Assume first that the slope of the linear regression model is known to be 0 and we fit a regression estimating only an intercept. Then the best prediction for Y_i associated with any X_i is given by the sample mean of Y:

$$\hat{Y}_i = \hat{\alpha} + 0 \cdot X_i = \hat{\alpha} = \bar{Y}$$

In this special case we can conclude that the variation of Y measures the square of the difference between the observed values Y_i and the predicted values $\hat{Y}_i = \bar{Y}$.

When the slope is nonzero we can improve our predictions by accounting for Y_i being dependent on X_i,

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$$

The additional information will reduce the unexplained portion of the variation in Y. To see this, consider the following identity, which holds for all observations:

$$Y_i - \bar{Y} = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \bar{Y}) \tag{3.24}$$

The term on the left of the equals sign denotes the difference between the sample value of Y and the mean of Y, the first right-hand term gives the residual $\hat{\varepsilon}_i$, and the second right-hand term gives the difference between the predicted value of Y and the mean of Y. This is shown in Fig. 3.4.

To measure variation, we square both sides of Eq. (3.24) and then sum over all observations $i = 1, 2, \ldots, N$:

$$\Sigma(Y_i - \bar{Y})^2 = \Sigma(Y_i - \hat{Y}_i)^2 + \Sigma(\hat{Y}_i - \bar{Y})^2 + 2\Sigma(Y_i - \hat{Y}_i)(\hat{Y}_i - \bar{Y}) \tag{3.25}$$

The last term in Eq. (3.25) can be shown to be identically 0 by using two properties of the least-squares residuals, $\Sigma\hat{\varepsilon}_i = 0$ and $\Sigma\hat{\varepsilon}_i X_i = 0$. All the derivations appear in Appendix 3.2. It follows that

$$\begin{array}{ccccc} \Sigma(Y_i - \bar{Y})^2 & = & \Sigma(Y_i - \hat{Y}_i)^2 & + & \Sigma(\hat{Y}_i - \bar{Y})^2 \\ \text{total variation of } Y \text{ (or total sum of squares)} & & \text{residual variation of } Y \text{ (or error sum of squares)} & & \text{explained variation of } Y \text{ (or regression sum of squares)} \\ \text{TSS} & = & \text{ESS} & + & \text{RSS} \end{array} \tag{3.26}$$

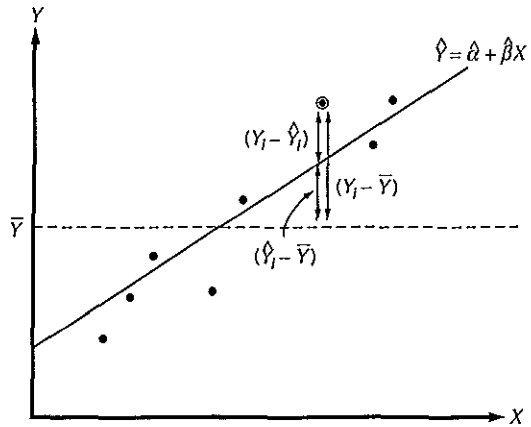


FIGURE 3.4
Decomposition of Y_i.

To normalize, we divide both sides of Eq. (3.26) by the total sum of squares to get

$$1 = \frac{\text{ESS}}{\text{TSS}} + \frac{\text{RSS}}{\text{TSS}}$$

We define the *R-squared* (R^2) of the regression equation as

$$R^2 = 1 - \frac{\text{ESS}}{\text{TSS}} = \frac{\text{RSS}}{\text{TSS}} \tag{3.27}$$

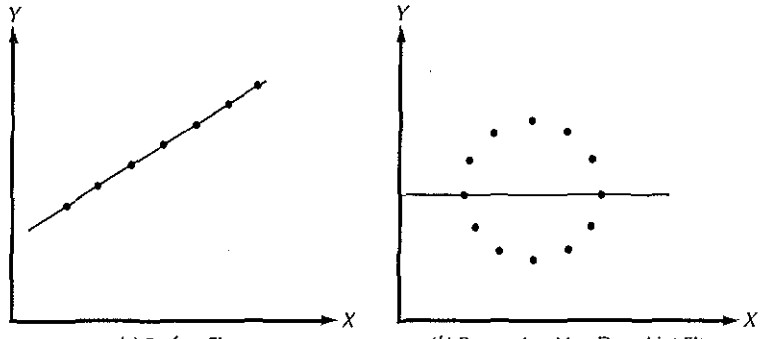
R^2 is the proportion of the total variation in Y explained by the regression of Y on X. Since the error sum of squares ranges in value between 0 and the total sum of squares, it is easy to see that R^2 ranges in value between 0 and 1. An R^2 of 0 occurs when the *linear* regression model does nothing to help explain the variation in Y. This may occur when the values of Y lie randomly around the horizontal line $Y = \overline{Y}$ or when the sample points lie on a circle (Fig. 3.5*b*). An R^2 of 1 can occur only when all sample points lie on the estimated regression line (Fig. 3.5*a*).

To relate R^2 to the regression parameters estimated earlier in this chapter, we write the predicted values of y_i as

$$\hat{y}_i = \hat{\beta}x_i$$

Then, each dependent variable observation can be subdivided as

$$y_i = \hat{y}_i + \hat{\varepsilon}_i$$



(a) Perfect Fit

(b) Regression Line Does Not Fit the Scatter of Points

FIGURE 3.5
Measuring *R*-squared.

where $\hat{\varepsilon}_i$ is the regression residual. Now

$$\begin{aligned} \Sigma y_i^2 &= \Sigma \hat{y}_i^2 + \Sigma \hat{\varepsilon}_i^2 \qquad \text{since } \Sigma \hat{y}_i \hat{\varepsilon}_i = \hat{\beta} \Sigma x_i \hat{\varepsilon}_i = 0 \\ &= \hat{\beta}^2 \Sigma x_i^2 + \Sigma \hat{\varepsilon}_i^2 \end{aligned}$$

from which it follows that

$$R^2 = \frac{\text{RSS}}{\text{TSS}} = \frac{\Sigma \hat{y}_i^2}{\Sigma y_i^2} = \hat{\beta}^2 \frac{\Sigma x_i^2}{\Sigma y_i^2}$$

or

$$R^2 = 1 - \frac{\Sigma \hat{\varepsilon}_i^2}{\Sigma y_i^2} \tag{3.28}$$

Equation (3.28) provides a simple formula for calculating R^2.

Note that R^2 is only a descriptive statistic. Roughly speaking, we associate a high value of R^2 with a good fit of the regression line and associate a low value of R^2 with a poor fit. We must realize, however, that a low value of R^2 can occur for several related reasons. In certain cases X may not be a good explanatory variable. Even though there is reason to believe that X does help in the prediction of Y, unexplained variation in Y may remain even after X has appeared in the equation. In time-series studies, however, one often obtains high values of R^2 simply because any variable that grows over time is likely to do a good job of explaining the variation of any other variable that grows over time. In cross-section studies, by contrast, a lower R^2 may occur even if the model is

a satisfactory one because of the large variation across individual units of observation.[7]

It is occasionally useful to summarize the breakdown of the variation in Y in terms of an *analysis of variance*. In such a case the total unexplained and explained variations in Y are converted into *variances* by dividing by the appropriate number of degrees of freedom.[8] Thus, the variance in Y is the total variation divided by $N - 1$, the explained variance is equal to the explained variation (since the regression involves only one additional constraint above the one used to estimate the mean of Y), and the residual variance is the residual variation divided by $N - 2$.

3.4.2 Correlation

Because R^2 is of value in analyzing a model with a causal relationship between the dependent variable Y and the independent variable X, R^2 is interpreted as more than a measure of correlation between two variables. Correlation techniques do not involve an implicit assumption of causality, while regression techniques do. We saw in Chapter 1 that the choice of dependent and independent variables in a regression model is crucial. The dependent variable is the variable to be explained, while the independent variable is the moving force. The least-squares technique is appropriate only if the causal structure of the model can be determined before the data are examined. If a model $Y = \alpha + \beta X$ is specified, one may interpret a significant t statistic on the regression slope parameter as evidence tending to *validate* the model. By contrast, an insignificant statistic would *invalidate* it.

As an example of correlation without causality, consider a series of observations over time that might have been obtained in a nineteenth-century study of medicine in Africa. One might find a high correlation between the number of doctors present in a region and the prevalence of disease in that region, but it would be wrong to infer that the presence of doctors is a cause of spreading disease.

Thus, high correlations do not provide for an inference of causality. One must specify *a priori* (based on previous information) that the number of doctors in a region is a function of the prevalence of disease and test statistically whether such a relationship holds if one is to use regression correctly. Correlation techniques are often used to suggest hypotheses or to confirm previously held

[7] This suggests that R^2 alone may not be a suitable measure of the extent to which a model is satisfactory. A better overall measure might be a statistic which describes the predictive power of the model in the face of new data.

[8] The number of degrees of freedom is the number of observations minus the number of constraints placed on the data by the calculation procedure. Thus, an estimate of the variation in Y involves $N - 1$ degrees of freedom because one constraint is placed on the data when deviations are measured about the sample mean (which must in itself be calculated). An additional degree of freedom is used up in the calculation of the slope parameter, leaving $N - 2$ degrees of freedom associated with the unexplained variation in the problem.