

The coefficient of determination can be interpreted as *the proportion of the variation in Y that is explained by the regression relationship of Y with X*.

Recall that the correlation coefficient r can be between -1 and 1 . Its square, r^2 , can therefore be anywhere from 0 to 1 . This is in accordance with the interpretation of r^2 as the *percentage of the variation in Y explained by the regression*. The coefficient is a measure of how closely the regression line fits the data; it is a measure of how much the variation in the values of Y is reduced once we regress Y on the variable X . When $r^2 = 1$, we know that 100% of the variation in Y is explained by X . This means that the data all lie right on the regression line, and there are no resulting errors (because, from Equation 9–30, SSE must be equal to zero). Since r^2 cannot be negative, we do not know whether the line slopes upward or downward (the direction can be found from b_1 or r), but we know that the line gives a *perfect fit* to the data. Such cases do not occur in business or economics. In fact, when there are no errors, no natural variation, there is no need for statistics.

At the other extreme is the case where the regression line explains nothing. Here the errors account for everything, and SSR is zero. In this case, we see from Equation 9–30 that $r^2 = 0$. In such cases, there is no linear relationship between X and Y , and the true regression slope is probably zero (we say “probably” because r^2 is only an estimator, given to chance variation; it could possibly be estimating a nonzero ρ^2). Between the two cases, $r^2 = 0$ and $r^2 = 1$, are values of r^2 that give an indication of the *relative fit* of the regression model to the data. *The higher r^2 , the better the fit and the higher our confidence in the regression*. Be wary, however, of situations where the reported r^2 is exceptionally high, such as 0.99 or 0.999 . In such cases, something may be wrong. We will see an example of this in the next chapter. Incidentally, in the context of multiple regression, discussed in the next chapter, we will use the notation R^2 for the coefficient of determination to indicate that the relationship is based on several explanatory X variables.

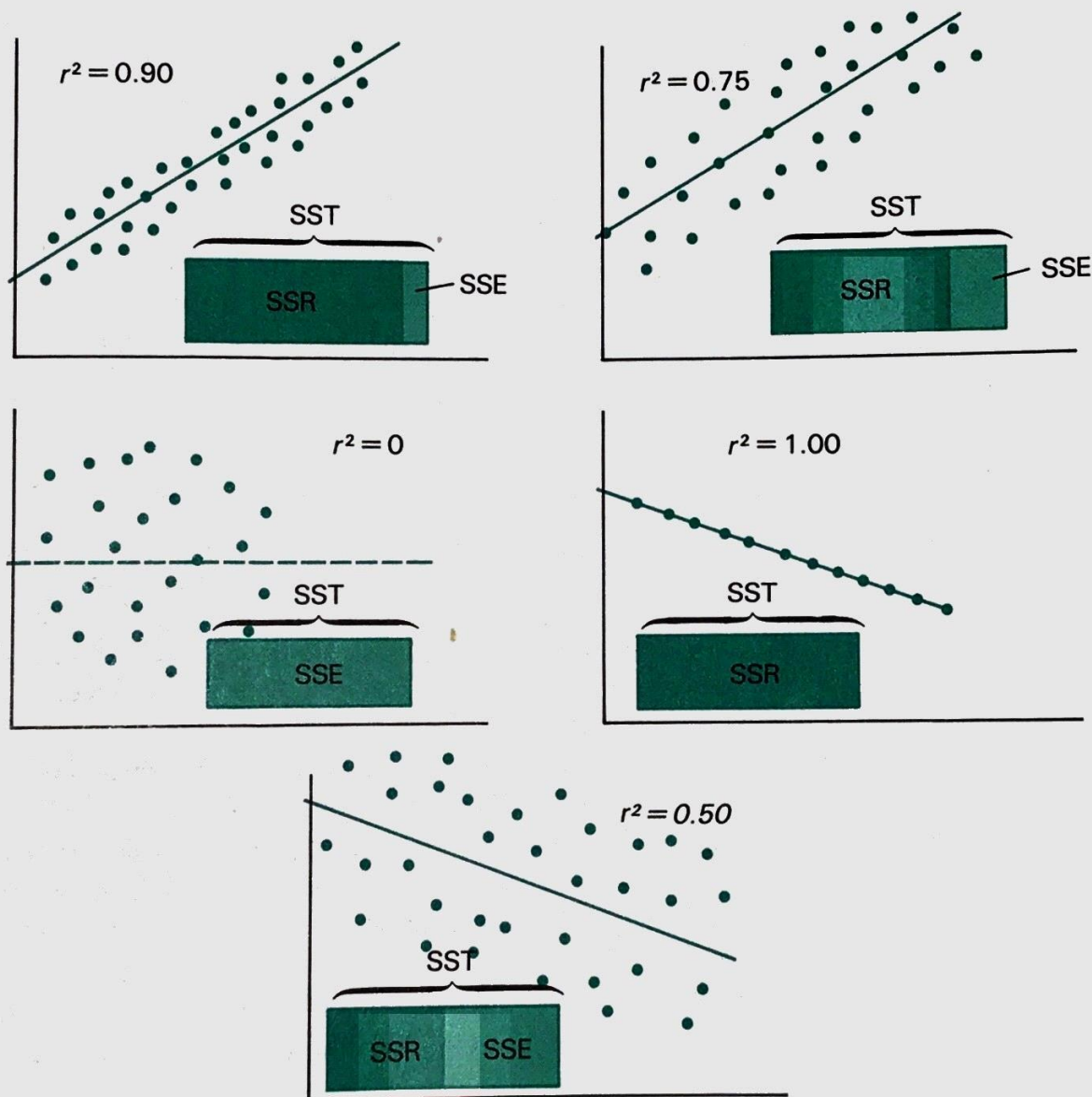
How high should the coefficient of determination be before we can conclude that a regression model fits the data well enough to use the regression with confidence? There is no clear-cut answer to this question. The answer depends on the intended use of the regression model. If we intend to use the regression for *prediction*, the higher the r^2 , the more accurate will be our predictions.

An r^2 value of 0.9 or above is very good, a value above 0.8 is good, and a value of 0.6 or above may be satisfactory in some applications, although we must be aware of the fact that, in such cases, errors in prediction may be relatively high. When the r^2 value is 0.5 or below, the regression explains only 50% or less of the variation in the data; therefore, predictions may be poor. If we are interested only in understanding the relationship between the variables, lower values of r^2 may be acceptable, as long as we realize that the model does not explain much.

Figure 9–20 shows several regressions and their corresponding r^2 values. If you think of the total sum of squared deviations as being in a box, then the r^2 is the proportion of the box that is filled with the explained sum of squares, the remaining part being the squared errors. This is shown for each regression in the figure.

It is easy to compute r^2 if we express SSR, SSE, and SST in terms of the computational sums of squares and cross products (Equations 9–9):

FIGURE 9-20 The Value of the Coefficient of Determination in Different Regressions



$$SST = SS_Y \quad SSR = b_1 SS_{XY} \quad SSE = SS_Y - b_1 SS_{XY} \quad (9-31)$$

We will now use Equation 9-31 in computing the coefficient of determination for Example (a). For this example, we have:

$$SST = SS_Y = 66,855,898$$

$$SSR = b_1 SS_{XY} = (1.255333776)(51,402,852.4) = 64,527,736.8$$

and

$$SSE = SST - SSR = 2,328,161.2$$

(These were computed when we found the MSE for this example.) We now compute r^2 as:

$$r^2 = \frac{SSR}{SST} = \frac{64,527,736.8}{66,855,898} = 0.96518$$

The r^2 in this example is very high. The interpretation is that over 96.5% of the variation in charges on the American Express card can be explained by the relationship between charges on the card and extent of travel (miles). Again we note that while the computational formulas are easy to use, r^2 is always reported in a prominent place in regression computer output. We will see this in Section 9–11.

In the next section, we will see how the sums of squares, along with the corresponding degrees of freedom, lead to mean squares—and to an analysis of variance in the context of regression. In closing this section, we note that in Chapter 10, we will introduce an adjusted coefficient of determination that accounts for degrees of freedom.