1. Simple Regression Analysis

This chapter shows how a hypothetical linear relationship between two variables can be quantified using appropriate data. The principle of least squares regression analysis is explained, and expressions for the coefficients are derived.

1.1 The simple linear model

The correlation coefficient may indicate that two variables are associated with one another, but it does not give any idea of the kind of relationship involved. We will now take the investigation a step further in those cases for which we are willing to hypothesize that one variable, usually known as the dependent variable, is determined by other variables, usually known as explanatory variables, independent variables, or regressors. The hypothesized mathematical relationship linking them is known as the regression model. If there is only one regressor, as will be assumed in this chapter and the next, it is described as a simple regression model. If there are two or more regressors, it is described as a multiple regression model.

It must be stated immediately that one would not expect to find an exact relationship between any two economic variables, unless it is true as a matter of definition. In textbook expositions of economic theory, the usual way of dealing with this awkward fact is to write down the relationship as if it were exact and to warn the reader that it is really only an approximation. In statistical analysis, however, one generally acknowledges the fact that the relationship is not exact by explicitly including in it a random factor known as the disturbance term.

We shall start with the simple regression model:

$$Y_i = \beta_1 + \beta_2 X_i + u_i. \tag{1.1}$$

Y_i, the value of the dependent variable in observation i, has two components: (1) the nonrandom component $\beta_1 + \beta_2 X_i$, where β_1 and β_2 are fixed quantities known as the parameters of the equation and X_i is the value of the explanatory variable in observation i, and (2) the disturbance term, u_i.

Figure 1.1 illustrates how these two components combine to determine Y. X_1, X_2, X_3, and X_4 are four hypothetical values of the explanatory variable. If the relationship between Y and X were exact, the corresponding values of Y would

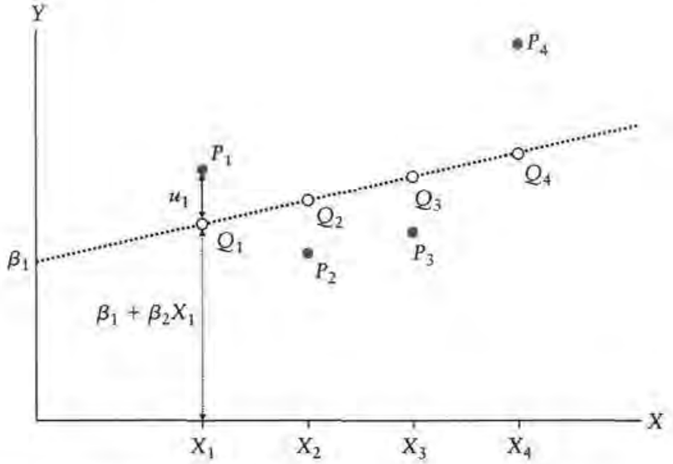


Figure 1.1 True relationship between Y and X

be represented by the points $Q_1 - Q_4$ on the line. The disturbance term causes the actual values of Y to be different. In the diagram, the disturbance term has been assumed to be positive in the first and fourth observations and negative in the other two, with the result that, if one plots the actual values of Y against the values of X, one obtains the points $P_1 - P_4$.

It must be emphasized that in practice the P points are all one can see of Figure 1.1. The actual values of β_1 and β_2, and hence the location of the Q points, are unknown, as are the values of the disturbance term in the observations. The task of regression analysis is to obtain estimates of β_1 and β_2, and hence an estimate of the location of the line, given the P points.

Why does the disturbance term exist? There are several reasons.

1. *Omission of explanatory variables*: The relationship between Y and X is almost certain to be a simplification. In reality, there will be other factors affecting Y that have been left out of equation (1.1), and their influence will cause the points to lie off the line. It often happens that there are variables that you would like to include in the regression equation but cannot because you are unable to measure them. For example, later on in this chapter we will fit an earnings function relating hourly earnings to years of schooling. We know very well that schooling is not the only determinant of earnings and eventually we will improve the model by including other variables, such as years of work experience. However, even the best specified earnings function accounts for at most half of the variation in earnings. Many other factors affect the chances of obtaining a good job, such as the unmeasurable attributes of an individual, and even pure luck in the sense of the individual finding a job that is a good match for his or her attributes. All of these other factors contribute to the disturbance term.

2. *Aggregation of variables*: In many cases, the relationship is an attempt to summarize in aggregate a number of microeconomic relationships. For example, the aggregate consumption function is an attempt to summarize a set of individual expenditure decisions. Since the individual relationships are likely to have different parameters, any attempt to relate aggregate expenditure to aggregate income can only be an approximation. The discrepancy is attributed to the disturbance term.
3. *Model misspecification*: The model may be misspecified in terms of its structure. Just to give one of the many possible examples, if the relationship refers to time series data, the value of Y may depend not on the actual value of X but on the value that had been anticipated in the previous period. If the anticipated and actual values are closely related, there will appear to be a relationship between Y and X, but it will only be an approximation, and again the disturbance term will pick up the discrepancy.
4. *Functional misspecification*: The functional relationship between Y and X may be misspecified mathematically. For example, the true relationship may be nonlinear instead of linear. We will consider the fitting of nonlinear relationships in Chapter 4. Obviously, one should try to avoid this problem by using an appropriate mathematical specification, but even the most sophisticated specification is likely to be only an approximation, and the discrepancy contributes to the disturbance term.
5. *Measurement error*: If the measurement of one or more of the variables in the relationship is subject to error, the observed values will not appear to conform to an exact relationship, and the discrepancy contributes to the disturbance term.

The disturbance term is the collective outcome of all these factors. Obviously, if you were concerned only with measuring the effect of X on Y, it would be much more convenient if the disturbance term did not exist. Were it not for its presence, the P points in Figure 1.1 would coincide with the Q points, you would know that every change in Y from observation to observation was due to a change in X, and you would be able to calculate β_1 and β_2 exactly. However, in fact, part of each change in Y is due to a change in u, and this makes life more difficult. For this reason, u is sometimes described as noise.

1.2 Least squares regression with one explanatory variable

Suppose that you are given the four observations on X and Y represented in Figure 1.1 and you are asked to obtain estimates of the values of β_1 and β_2 in equation (1.1). As a rough approximation, you could do this by plotting the four P points and drawing a line to fit them as best you can. This has been done in Figure 1.2. The intersection of the line with the Y-axis provides an estimate of the intercept β_1, which will be denoted b_1, and the slope provides an estimate of

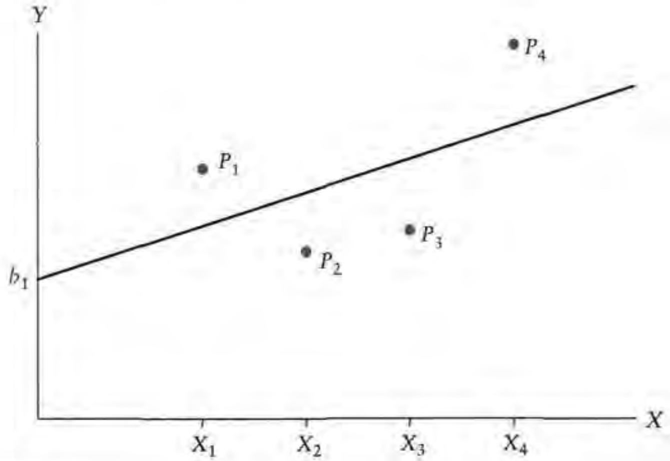


Figure 1.2 Fitted line

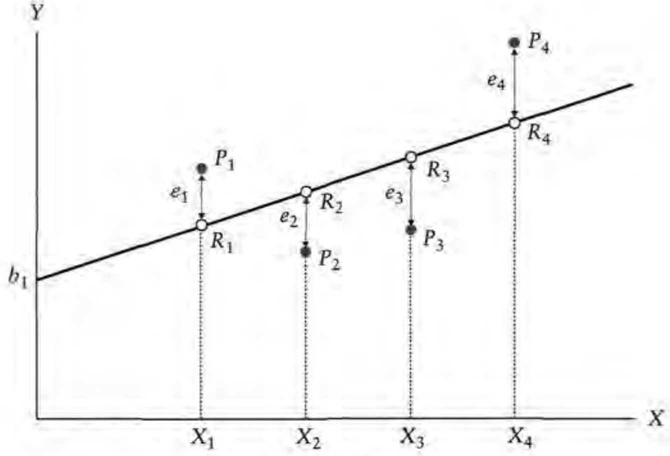


Figure 1.3 Fitted regression line showing residuals

the slope coefficient β_2, which will be denoted b_2. The line, known as the fitted model, will be written

$$\hat{Y}_i = b_1 + b_2 X_i, \tag{1.2}$$

the caret mark over Y indicating that it is the fitted value of Y corresponding to X, not the actual value. In Figure 1.3, the fitted points are represented by the points $R_1 - R_4$.

One thing that should be accepted from the beginning is that you can never discover the true values of β_1 and β_2, however much care you take in drawing the line. b_1 and b_2 are only estimates, and they may be good or bad. Once in a while your estimates may be absolutely accurate, but this can only be by coincidence, and even then you will have no way of knowing that you have hit the target exactly.

This remains the case even when you use more sophisticated techniques. Drawing a regression line by eye is all very well, but it leaves a lot to subjective judgment. Furthermore, as will become obvious, it is not even possible when you have a variable Y depending on two or more explanatory variables instead of only one. The question arises, is there a way of calculating good estimates of β_1 and β_2 algebraically?

The first step is to define what is known as a residual for each observation. This is the difference between the actual value of Y in any observation and the fitted value given by the regression line: that is, the vertical distance between P_i and R_i in observation i. It will be denoted e_i:

$$e_i = Y_i - \hat{Y}_i. \tag{1.3}$$

The residuals for the four observations are shown in Figure 1.3. Substituting (1.2) into (1.3), we obtain

$$e_i = Y_i - b_1 - b_2X_i \tag{1.4}$$

and hence the residual in each observation depends on our choice of b_1 and b_2. Obviously, we wish to fit the regression line, that is, choose b_1 and b_2, in such a way as to make the residuals as small as possible. Equally obviously, a line that fits some observations well will fit others badly and vice versa. We need to devise a criterion of fit that takes account of the size of all the residuals simultaneously.

There are a number of possible criteria, some of which work better than others. It is useless minimizing the sum of the residuals, for example. The sum will automatically be equal to zero if you make $b_1 = \overline{Y}$ and $b_2 = 0$, obtaining the horizontal line $Y = \overline{Y}$. The positive residuals will then exactly balance the negative ones but, other than this, the line will not fit the observations.

One way of overcoming the problem is to minimize RSS, the residual sum of squares (sum of the squares of the residuals). For Figure 1.3,

$$RSS = e_1^2 + e_2^2 + e_3^2 + e_4^2. \tag{1.5}$$

The smaller one can make RSS, the better is the fit, according to this criterion. If one could reduce RSS to zero, one would have a perfect fit, for this would imply that all the residuals are equal to zero. The line would go through all the points, but of course in general the disturbance term makes this impossible.

There are other quite reasonable solutions, but the least squares criterion yields estimates of β_1 and β_2 that are unbiased and the most efficient of their type, provided that certain conditions are satisfied. For this reason, the least squares technique is far and away the most popular in uncomplicated applications of regression analysis. The form used here is usually referred to as ordinary least squares and abbreviated OLS. Variants designed to cope with particular problems will be discussed later in the text.

1.3 Derivation of the regression coefficients

We will begin with a very simple example with only three observations, just to show the mechanics working. Y is observed to be equal to 3 when X is equal to 1, 5 when X is equal to 2, and 6 when X is equal to 3, as shown in Figure 1.4.

We shall assume that the true model is

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1.6}$$

and we shall estimate the coefficients b_1 and b_2 of the equation

$$\hat{Y}_i = b_1 + b_2 X_i. \tag{1.7}$$

When X is equal to 1, $\hat{Y}$ is equal to $(b_1 + b_2)$, according to the regression line. When X is equal to 2, $\hat{Y}$ is equal to $(b_1 + 2b_2)$. When X is equal to 3, $\hat{Y}$ is equal to $(b_1 + 3b_2)$. Therefore, we can set up Table 1.1. So the residual for the first observation, e_1, which is given by $(Y_1 - \hat{Y}_1)$, is equal to $(3 - b_1 - b_2)$. Similarly, $e_2 = Y_2 - \hat{Y}_2 = 5 - b_1 - 2b_2$ and $e_3 = Y_3 - \hat{Y}_3 = 6 - b_1 - 3b_2$. Hence,

$$\begin{aligned} RSS &= (3 - b_1 - b_2)^2 + (5 - b_1 - 2b_2)^2 + (6 - b_1 - 3b_2)^2 \\ &= 9 + b_1^2 + b_2^2 - 6b_1 - 6b_2 + 2b_1b_2 \\ &\quad + 25 + b_1^2 + 4b_2^2 - 10b_1 - 20b_2 + 4b_1b_2 \\ &\quad + 36 + b_1^2 + 9b_2^2 - 12b_1 - 36b_2 + 6b_1b_2 \\ &= 70 + 3b_1^2 + 14b_2^2 - 28b_1 - 62b_2 + 12b_1b_2. \end{aligned} \tag{1.8}$$

Now we want to choose b_1 and b_2 to minimize RSS. To do this, we use the calculus and find the values of b_1 and b_2 that satisfy the first-order conditions

$$\frac{\partial RSS}{\partial b_1} = 0 \quad \text{and} \quad \frac{\partial RSS}{\partial b_2} = 0. \tag{1.9}$$

Taking partial differentials,

$$\frac{\partial RSS}{\partial b_1} = 6b_1 + 12b_2 - 28 \tag{1.10}$$

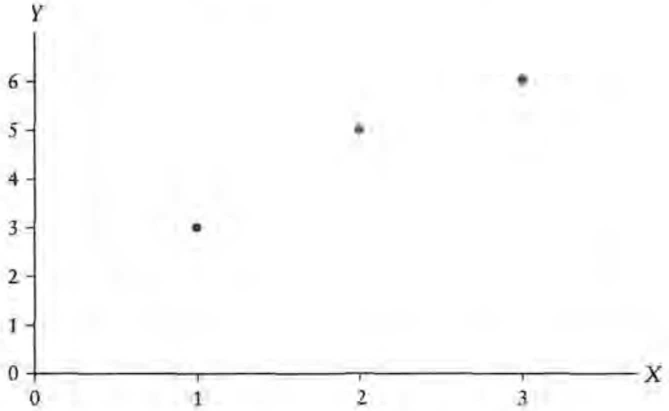


Figure 1.4 Three-observation example

Table 1.1 Three-observation example

X	Y	$\hat{Y}$	e
1	3	$b_1 + b_2$	$3 - b_1 - b_2$
2	5	$b_1 + 2b_2$	$5 - b_1 - 2b_2$
3	6	$b_1 + 3b_2$	$6 - b_1 - 3b_2$

and

$$\frac{\partial RSS}{\partial b_2} = 28b_2 + 12b_1 - 62 \qquad (1.11)$$

and so we have

$$3b_1 + 6b_2 - 14 = 0 \qquad (1.12)$$

and

$$6b_1 + 14b_2 - 31 = 0. \qquad (1.13)$$

Solving these two equations, one obtains $b_1 = 1.67$ and $b_2 = 1.50$. The regression equation is therefore

$$\hat{Y}_i = 1.67 + 1.50X_i. \qquad (1.14)$$

The three points and the regression line are shown in Figure 1.5.

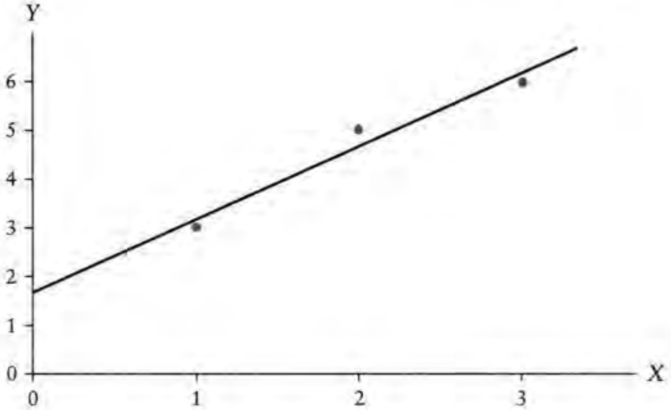


Figure 1.5 Three-observation example with regression line

Least squares regression with one explanatory variable: the general case

We shall now consider the general case where there are n observations on two variables X and Y and, supposing Y to depend on X, we will fit the equation

$$\hat{Y}_i = b_1 + b_2 X_i. \tag{1.15}$$

The fitted value of the dependent variable in observation i, $\hat{Y}_i$, will be $(b_1 + b_2X_i)$, and the residual e_i will be $(Y_i - b_1 - b_2X_i)$. We wish to choose b_1 and b_2 so as to minimize the residual sum of the squares, RSS, given by

$$RSS = e_1^2 + \cdots + e_n^2 = \sum_{i=1}^{n} e_i^2. \tag{1.16}$$

We will find that RSS is minimized when

$$b_2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{1.17}$$

and

$$b_1 = \bar{Y} - b_2\bar{X}. \tag{1.18}$$

The derivation of the expressions for b_1 and b_2 will follow the same procedure as the derivation in the two preceding examples, and you can compare the general version with the examples at each step. We will begin by expressing the square of the residual in observation i in terms of b_1, b_2, and the data on X and Y:

$$\begin{aligned} e_i^2 &= \left(Y_i - \hat{Y}_i\right)^2 = \left(Y_i - b_1 - b_2X_i\right)^2 \\ &= Y_i^2 + b_1^2 + b_2^2X_i^2 - 2b_1Y_i - 2b_2X_iY_i + 2b_1b_2X_i. \end{aligned} \tag{1.19}$$

Summing over all the n observations, we can write RSS as

$$\begin{aligned} RSS &= \left(Y_1 - b_1 - b_2X_1\right)^2 + \cdots + \left(Y_n - b_1 - b_2X_n\right)^2 \\ &= Y_1^2 + b_1^2 + b_2^2X_1^2 - 2b_1Y_1 - 2b_2X_1Y_1 + 2b_1b_2X_1 \\ &\quad + \cdots \\ &\quad + Y_n^2 + b_1^2 + b_2^2X_n^2 - 2b_1Y_n - 2b_2X_nY_n + 2b_1b_2X_n \\ &= \sum_{i=1}^{n} Y_i^2 + nb_1^2 + b_2^2\sum_{i=1}^{n} X_i^2 - 2b_1\sum_{i=1}^{n} Y_i - 2b_2\sum_{i=1}^{n} X_iY_i + 2b_1b_2\sum_{i=1}^{n} X_i. \end{aligned} \tag{1.20}$$

Note that *RSS* is effectively a quadratic expression in b_1 and b_2, with numerical coefficients determined by the data on X and Y in the sample. We can influence the size of *RSS* only through our choice of b_1 and b_2. The data on X and Y, which determine the locations of the observations in the scatter diagram, are fixed once we have taken the sample. The equation is the generalized version of equation (1.8) in the three-observation example.

The first-order conditions for a minimum, $\partial RSS/\partial b_1 = 0$ and $\partial RSS/\partial b_2 = 0$, yield the following equations:

$$2nb_1 - 2\sum_{i=1}^{n} Y_i + 2b_2\sum_{i=1}^{n} X_i = 0 \tag{1.21}$$

$$2b_2\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_iY_i + 2b_1\sum_{i=1}^{n} X_i = 0. \tag{1.22}$$

These equations are known as the normal equations for the regression coefficients and are the generalized versions of (1.12) and (1.13) in the three-observation example. Equation (1.21) allows us to write b_1 in terms of $\overline{Y}$, $\overline{X}$, and the as yet unknown b_2. Noting that $\overline{X} = \frac{1}{n}\sum X_i$ and $\overline{Y} = \frac{1}{n}\sum Y_i$, (1.21) may be rewritten

$$2nb_1 - 2n\overline{Y} + 2b_2n\overline{X} = 0 \tag{1.23}$$

and hence,

$$b_1 = \overline{Y} - b_2\overline{X}. \tag{1.24}$$

Substituting for b_1 in (1.22), and again noting that $\sum X_i = n\overline{X}$, we obtain

$$2b_2\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_iY_i + 2\left(\overline{Y} - b_2\overline{X}\right)n\overline{X} = 0. \tag{1.25}$$

Separating the terms involving b_2 and not involving b_2 on opposite sides of the equation, we have

$$2b_2\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) = 2\sum_{i=1}^{n} X_iY_i - 2n\overline{X}\overline{Y}. \tag{1.26}$$

Hence,

$$b_2 = \frac{\sum_{i=1}^{n} X_iY_i - n\overline{X}\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2}. \tag{1.27}$$

An alternative form that we shall prefer is

$$b_2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}. \tag{1.28}$$

To see the equivalence, note that

$$\begin{aligned}
\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) &= \sum_{i=1}^{n} X_iY_i - \sum_{i=1}^{n} X_i\overline{Y} - \sum_{i=1}^{n} \overline{X}Y_i + \sum_{i=1}^{n} \overline{X}\overline{Y} \\
&= \sum_{i=1}^{n} X_iY_i - \overline{Y}\sum_{i=1}^{n} X_i \ - \overline{X}\sum_{i=1}^{n} Y_i + n\overline{X}\overline{Y} \\
&= \sum_{i=1}^{n} X_iY_i - \overline{Y}\left(n\overline{X}\right) - \overline{X}\left(n\overline{Y}\right) + n\overline{X}\overline{Y} \\
&= \sum_{i=1}^{n} X_iY_i - n\overline{X}\overline{Y}. \qquad (1.29)
\end{aligned}$$

Similarly,

$$\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 = \sum_{i=1}^{n} X_i^2 - n\overline{X}^2. \tag{1.30}$$

Put X instead of Y in (1.29). Having found b_2 from (1.28), you find b_1 from (1.24). A check of the second-order conditions would confirm that we have minimized *RSS*.

Example

In the three-observation example, $\overline{Y} = 4.67$, $\overline{X} = 2.00$, $\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = 3.00$, and $\sum\left(X_i - \overline{X}\right)^2 = 2.00$, so

$$b_2 = 3.00/2.00 = 1.50 \tag{1.31}$$

and

$$b_1 = \overline{Y} - b_2\overline{X} = 4.67 - 1.50 \times 2.00 = 1.67, \tag{1.32}$$

which confirms the original calculation.

Two decompositions of the dependent variable

In the preceding analysis we have encountered two ways of decomposing the value of the dependent variable in a regression model. They are going to be used throughout the text, so it is important that they be distinguished conceptually.

The first decomposition relates to the process by which the values of Y are generated:

$$Y_i = \beta_1 + \beta_2 X_i + u_i. \tag{1.33}$$

In observation i, Y_i is generated as the sum of two components, the nonstochastic component, $\beta_1 + \beta_2 X_i$, and the disturbance term u_i. This decomposition is purely theoretical. We will use it in the analysis of the properties of the regression estimators. It is illustrated in Figure 1.6a, where QT is the nonstochastic component of Y and PQ is the disturbance term.

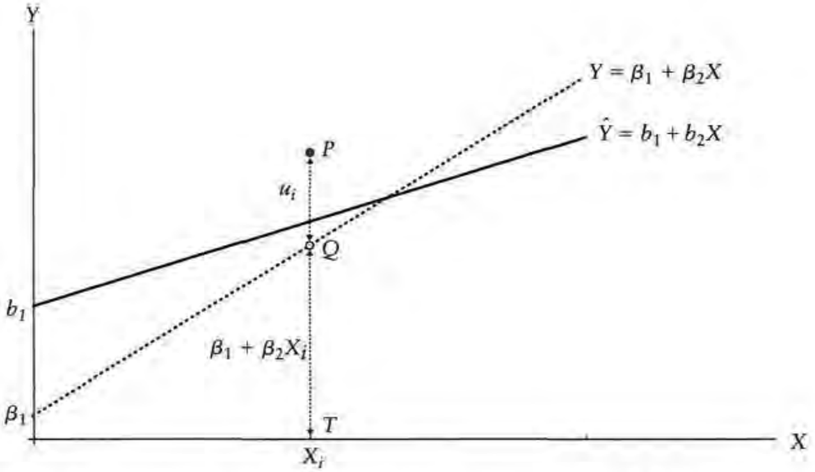


Figure 1.6a Decomposition of Y into nonstochastic component and disturbance term

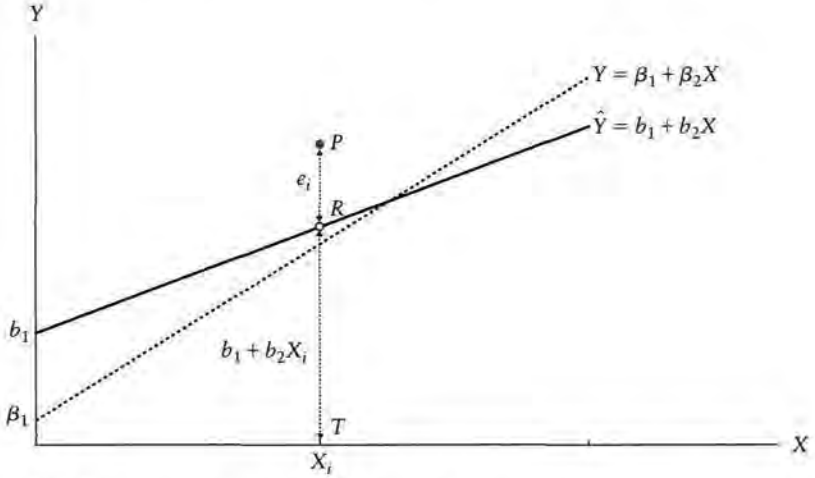


Figure 1.6b Decomposition of Y into fitted value and residual

The other decomposition relates to the regression line:

$$\begin{aligned} Y_i &= \hat{Y}_i + e_i \\ &= b_1 + b_2 X_i + e_i. \end{aligned} \tag{1.34}$$

Once we have chosen the values of b_1 and b_2, each value of Y is split into the fitted value, $\hat{Y}_i$, and the residual, e_i. This decomposition is operational, but it is to some extent arbitrary because it depends on our criterion for determining b_1 and b_2 and it will inevitably be affected by the particular values taken by the disturbance term in the observations in the sample. It is illustrated in Figure 1.6b, where RT is the fitted value and PR is the residual.

Regression model without an intercept

Typically, an intercept should be included in the regression specification. Occasionally, however, one may have reason to fit the regression without an intercept. In the case of a simple regression model, the specification becomes

$$Y_i = \beta_2 X_i + u_i \tag{1.35}$$

and the fitted model is

$$\hat{Y}_i = b_2 X_i. \tag{1.36}$$

We will derive the expression for b_2 from first principles using the least squares criterion. The residual in observation i is

$$e_i = Y_i - \hat{Y}_i = Y_i - b_2 X_i. \tag{1.37}$$

The sum of the squares of the residuals is

$$RSS = \sum_{i=1}^{n} (Y_i - b_2 X_i)^2 = \sum_{i=1}^{n} Y_i^2 - 2b_2 \sum_{i=1}^{n} X_i Y_i + b_2^2 \sum_{i=1}^{n} X_i^2. \tag{1.38}$$

The first-order condition for a minimum, $\frac{\mathrm{d}RSS}{\mathrm{d}b_2} = 0$, yields:

$$2b_2 \sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i Y_i = 0 \tag{1.39}$$

and this gives us

$$b_2 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}. \tag{1.40}$$

The second derivative, $2\sum_{i=1}^{n} X_i^2$, is positive, confirming that we have minimized RSS.

EXERCISES

1.1 Suppose that the fitted line is $\hat{Y} = b_1 + b_2 X$, with b_1 and b_2 defined as in equations (1.24) and (1.28). Demonstrate that the fitted line must pass through the point $\{\overline{X}, \overline{Y}\}$ representing the mean of the observations in the sample.

1.2 Using the normal equations (1.21) and (1.22), show that b_1 is defined, but b_2 is not, if $X_i = 0$ for all i. Give an intuitive explanation of this result.

1.3 Demonstrate from first principles that the least squares estimator of β_1 in the primitive model where Y consists simply of a constant plus a disturbance term,

$$Y_i = \beta_1 + u_i$$

is $b_1 = \overline{Y}$. (First define *RSS* and then differentiate.)

1.4 Interpretation of a regression equation

There are two stages in the interpretation of a regression equation. The first is to turn the equation into words so that it can be understood by a noneconometrician. The second is to decide whether this literal interpretation should be taken at face value or whether the relationship should be investigated further.

Both stages are important. We will leave the second until later and concentrate for the time being on the first. It will be illustrated with an earnings function, hourly earnings in 2002, *EARNINGS*, measured in dollars, being regressed on schooling, *S*, measured as highest grade completed, for 540 respondents from the United States National Longitudinal Survey of Youth 1979, the data set that is used for many of the practical illustrations and exercises in this text. See Appendix B for a description of it. This regression uses *EAEF* Data Set 21. The Stata output for the regression is shown in Table 1.2. The scatter diagram and regression line are shown in Figure 1.7.

Table 1.2

```
. reg EARNINGS S
    Source  |        SS          df          MS           Number of obs =     540
------------------------------------------------------    F(1, 538)     =  112.15
     Model  |    19321.5589       1      19321.5589       Prob > F      =  0.0000
  Residual  |    92688.6722     538      172.283777       R-squared     =  0.1725
------------------------------------------------------    Adj R-squared =  0.1710
     Total  |    112010.231     539      207.811189       Root MSE      =  13.126

---------------------------------------------------------------------------------
  EARNINGS  |         Coef.   Std. Err.         t   P>|t|     [95% Conf. Interval]
---------------------------------------------------------------------------------
         S  |      2.455321    .2318512     10.59   0.000     1.999876     2.910765
     _cons  |     -13.93347    3.219851     -4.33   0.000    -20.25849    -7.608444
---------------------------------------------------------------------------------
```

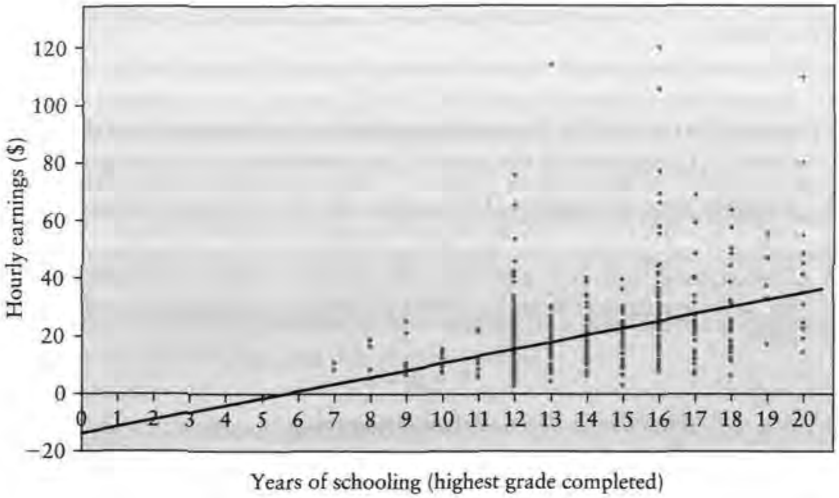


Figure 1.7 A simple earnings function

For the time being, ignore everything except the column headed 'coef.' in the bottom half of the table. This gives the estimates of the coefficient of S and the constant, and thus the following fitted equation:

$$\widehat{EARNINGS} = -13.93 + 2.46\,S. \tag{1.41}$$

Interpreting it literally, the slope coefficient indicates that, as S increases by one unit (of S), *EARNINGS* increases by 2.46 units (of *EARNINGS*). Since S is measured in years, and *EARNINGS* is measured in dollars per hour, the coefficient of S implies that hourly earnings increase by \$2.46 for every extra year of schooling.

What about the constant term? Strictly speaking, it indicates the predicted level of *EARNINGS* when $S = 0$. Sometimes the constant will have a clear meaning, but sometimes not. If the sample values of the explanatory variable are a long way from zero, extrapolating the regression line back to zero may be dangerous. Even if the regression line gives a good fit for the sample of observations, there is no guarantee that it will continue to do so when extrapolated to the left or to the right.

In this case, a literal interpretation of the constant would lead to the nonsensical conclusion that an individual with no schooling would have hourly earnings of −\$13.93. In this data set, no individual had less than seven years of schooling, so it is not surprising that extrapolation to zero leads to trouble.

Box 1.1 gives a general guide to interpreting regression equations when the variables are measured in natural units.

It is important to keep three things in mind when interpreting a regression equation. First, b_1 is only an estimate of β_1 and b_2 is only an estimate of β_2, so the interpretation is really only an estimate. Second, the regression equation refers only to the general tendency for the sample. Any individual case will be further affected by the random factor. Third, the interpretation is conditional on the equation being correctly specified.

BOX 1.1 Interpretation of a linear regression equation

This is a foolproof way of interpreting the coefficients of a linear regression

$$\hat{Y}_i = b_1 + b_2 X_i$$

when Y and X are variables with straightforward natural units (not logarithms or other functions).

The first step is to say that a one-unit increase in X (measured in units of X) will cause a b_2 unit increase in Y (measured in units of Y). The second step is to check to see what the units of X and Y actually are, and to replace the word 'unit' with the actual unit of measurement. The third step is to see whether the result could be expressed in a better way, without altering its substance.

The constant, b_1, gives the predicted value of Y (in units of Y) for X equal to 0. It may or may not have a plausible meaning, depending on the context.

In fact, this is actually a naïve specification of an earnings function. We will reconsider it several times in later chapters. You should be undertaking parallel experiments using one of the other *EAEF* data sets described in Appendix B.

Having fitted a regression, it is natural to ask whether we have any means of telling how accurate our estimates are. This very important issue will be discussed in the next chapter.

Changes in the units of measurement

Suppose that the units of measurement of Y or X are changed. How will this affect the regression results? Intuitively, we would anticipate that nothing of substance will be changed, and this is correct. We will demonstrate this for the estimates of the regression coefficients in this section, and we will trace the implications for the rest of the regression output in due course. We begin by supposing that the true model is

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{1.42}$$

and that the fitted model is

$$\hat{Y}_i = b_1 + b_2 X_i. \tag{1.43}$$

We now suppose that the units of measurement of Y are changed, with the new measure, Y^*, being related to the old one by

$$Y_i^* = \lambda_1 + \lambda_2 Y_i. \tag{1.44}$$

Typically, a change of measurement involves a simple multiplicative scaling, such as when we convert pounds into grams. However, one occasionally encounters a full linear transformation. Conversion of temperatures from degrees Celsius to degrees Fahrenheit is an example. Regressing Y^* on X, we have

$$b_2^* = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i^* - \overline{Y}^*)}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})\left(\left[\lambda_1 + \lambda_2 Y_i\right] - \left[\lambda_1 + \lambda_2 \overline{Y}\right]\right)}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(\lambda_2 Y_i - \lambda_2 \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \lambda_2 b_2. \qquad \textbf{(1.45)}$$

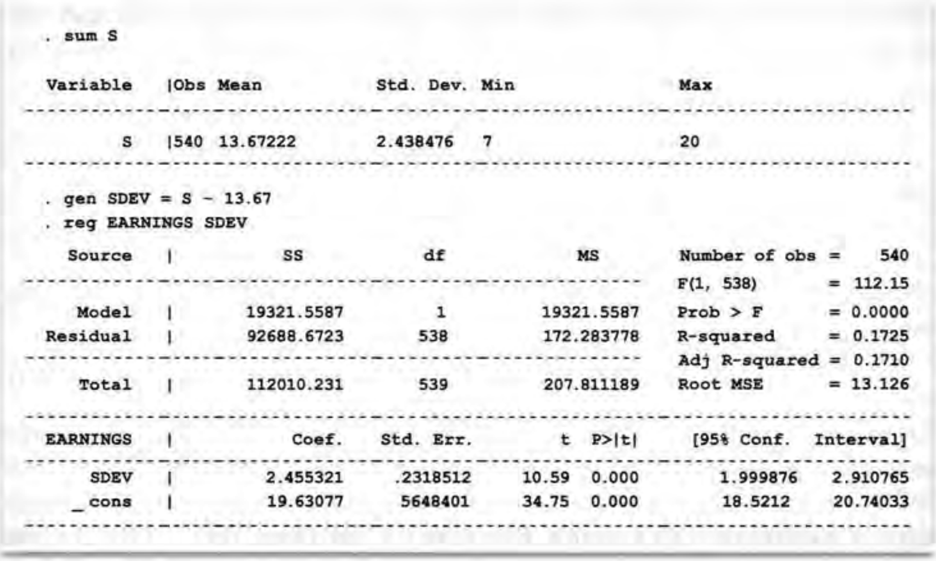
The change of measurement has caused the slope coefficient to be multiplied by λ_2. This is logical. A unit change in Y is the same as a change of λ_2 units in Y^*. According to the regression equation, a unit change in X leads to a change of b_2 units in Y, so it should lead to a change of $\lambda_2 b_2$ units in Y^*. The effect on the intercept will be left as an exercise. The effect of a change in the units of measurement of X will also be left as an exercise.

Demeaning

Often the intercept in a regression equation has no sensible interpretation because $X = 0$ is distant from the data range. The earnings function illustrated in Figure 1.7 is an example, with the intercept actually being negative. Sometimes it is useful to deal with the problem by defining X^* as the deviation of X about its sample mean

$$X_i^* = X_i - \overline{X} \qquad \textbf{(1.46)}$$

Table 1.3



```
. sum S

Variable   |Obs Mean         Std. Dev. Min                Max
------------------------------------------------------------------------------
         S  |540 13.67222     2.438476  7                 20
------------------------------------------------------------------------------

. gen SDEV = S - 13.67
. reg EARNINGS SDEV
   Source   |        SS           df          MS        Number of obs =    540
-----------------------------------------------------   F(1, 538)     = 112.15
    Model   |    19321.5587        1     19321.5587     Prob > F      = 0.0000
 Residual   |    92688.6723      538     172.283778     R-squared     = 0.1725
-----------------------------------------------------   Adj R-squared = 0.1710
    Total   |    112010.231      539     207.811189     Root MSE      = 13.126

------------------------------------------------------------------------------
 EARNINGS   |         Coef.   Std. Err.        t  P>|t|     [95% Conf.  Interval]
------------------------------------------------------------------------------
     SDEV   |      2.455321   .2318512     10.59  0.000      1.999876   2.910765
    _cons   |      19.63077   5648401      34.75  0.000       18.5212   20.74033
------------------------------------------------------------------------------
```

and regressing Y on X^*. The slope coefficient will not be affected, but the intercept will now be the fitted value of Y at the sample mean of X. Since this is, by construction, in the middle of the sample, it may be more useful for analytical purposes.

Table 1.3 shows the revised output from Table 1.2 when the schooling variable is demeaned. The sum (summarize) command yielded descriptive statistics for S and the gen (generate) command generated the demeaned variable $SDEV$. These are Stata commands but similar commands will be found in all regression applications.

Mean years of schooling was 13.67, that is nearly two years of college. For an individual at the mean, the intercept predicts hourly earnings of $19.63.

EXERCISES

Note: Some of the exercises in this and later chapters require you to fit regressions using one of the *EAEF* data sets. See Appendix B for details.

1.4 The table shows the average annual percentage rates of growth of employment, e, and real GDP, g, for 25 OECD countries for the period 1988–97. The regression output shows the result of regressing e on g. Provide an interpretation of the coefficients.

Average annual percentage rates of growth of employment and real GDP, 1988–97

	Employment	*GDP*		*Employment*	*GDP*
Australia	1.68	3.04	Korea	2.57	7.73
Austria	0.65	2.55	Luxembourg	3.02	5.64
Belgium	0.34	2.16	Netherlands	1.88	2.86
Canada	1.17	2.03	New Zealand	0.91	2.01
Denmark	0.02	2.02	Norway	0.36	2.98
Finland	–1.06	1.78	Portugal	0.33	2.79
France	0.28	2.08	Spain	0.89	2.60
Germany	0.08	2.71	Sweden	–0.94	1.17
Greece	0.87	2.08	Switzerland	0.79	1.15
Iceland	–0.13	1.54	Turkey	2.02	4.18
Ireland	2.16	6.40	United Kingdom	0.66	1.97
Italy	–0.30	1.68	United States	1.53	2.46
Japan	1.06	2.81			

```
. reg e g
  Source    |      SS           df        MS            Number of obs =     25
------------+-----------------------------------        F(1, 23)      =  33.10
   Model    |  14.5753023        1     14.5753023        Prob > F      = 0.0000
Residual    |  10.1266731       23     .440290135        R-squared     = 0.5900
------------+-----------------------------------        Adj R-squared = 0.5722
   Total    |  24.7019754       24     1.02924898        Root MSE      = .66354
------------------------------------------------------------------------------
       e    |      Coef.   Std. Err.      t   P>|t|      [95% Conf. Interval]
------------+-----------------------------------------------------------------
       g    |    .489737   .0851184     5.75  0.000      .3136561     .6658179
   _cons    |  -.5458912   .2740387    -1.99  0.058     -1.112784     .0210011
------------------------------------------------------------------------------
```

1.5 In Exercise 1.4, $\bar{e} = 0.83$, $\bar{g} = 2.82$, $\sum(e_i - \bar{e})(g_i - \bar{g}) = 29.76$, and $\sum(g_i - \bar{g})^2 = 60.77$. Calculate the regression coefficients and check that they are the same as in the regression output.

1.6 Does educational attainment depend on intellectual ability? In the United States, as in most countries, there is a positive correlation between educational attainment and cognitive ability. *S* (highest grade completed by 2002) is the number of years of schooling of the respondent. *ASVABC* is a composite measure of numerical and verbal ability with mean 50 and standard deviation 10 (both approximately; for further details of the measure, see Appendix B). Perform a regression of *S* on *ASVABC* and interpret the regression results.

1.7 Do earnings depend on education? Using your *EAEF* data set, fit an earnings function parallel to that in Table 1.2, regressing *EARNINGS* on *S*, and give an interpretation of the coefficients.

1.8* The output shows the result of regressing the weight of the respondent in 1985, measured in pounds, on his or her height, measured in inches, using *EAEF* Data Set 21. Provide an interpretation of the coefficients.

```
. reg WEIGHT85 HEIGHT
   Source   |      SS           df        MS            Number of obs =    540
------------+-----------------------------------        F(1, 538)     = 355.97
    Model   |  261111.383        1     261111.383       Prob > F      = 0.0000
 Residual   |  394632.365      538     733.517407       R-squared     = 0.3982
------------+-----------------------------------        Adj R-squared = 0.3971
    Total   |  655743.748      539     1216.59322       Root MSE      = 27.084
------------------------------------------------------------------------------
 WEIGHT85   |      Coef.   Std. Err.        t   P>|t|    [95% Conf. Interval]
------------+-----------------------------------------------------------------
   HEIGHT   |   5.192973    .275238     18.87  0.000      4.6523     5.733646
    _cons   |  -194.6815    18.6629    -10.43  0.000   -231.3426    -158.0204
------------------------------------------------------------------------------
```

1.9 The output shows the result of regressing the number of children in the family on the years of schooling of the mother, using *EAEF* Data Set 21. Provide an interpretation of the coefficients. (The data set contains data on siblings, the number of brothers and sisters of the respondent. Therefore the total number of children in the family is the number of siblings plus one.)

```
. g CHILDREN = SIBLINGS + 1
. reg CHILDREN SM
   Source  |       SS         df        MS          Number of obs =    540
-----------+------------------------------------    F(1, 538)     =  63.60
    Model  |    272.69684      1     272.69684      Prob > F      = 0.0000
 Residual  |    2306.7402    538    4.28762118      R-squared     = 0.1057
-----------+------------------------------------    Adj R-squared = 0.1041
    Total  |   2579.43704    539    4.78559747      Root MSE      = 2.0707

------------------------------------------------------------------------------
 CHILDREN  |       Coef.   Std. Err.       t   P>|t|      [95% Conf. Interval]
-----------+------------------------------------------------------------------
       SM  |   -.2525473   .0316673    -7.98   0.000      -.314754   -.1903406
    _cons  |    7.198478   .3773667    19.08   0.000      6.457186    7.939771
------------------------------------------------------------------------------
```

1.10* A researcher has international cross-sectional data on aggregate wages, W, aggregate profits, P, and aggregate income, Y, for a sample of n countries. By definition,

$$Y_i = W_i + P_i.$$

The regressions

$$\hat{W}_i = a_1 + a_2 Y_i$$

$$\hat{P}_i = b_1 + b_2 Y_i$$

are fitted using OLS regression analysis. Show that the regression coefficients will automatically satisfy the following equations:

$$a_2 + b_2 = 1$$

$$a_1 + b_1 = 0.$$

Explain intuitively why this should be so.

1.11 Demonstrate that, if the units of Y are changed so that $Y_i^* = \lambda_1 + \lambda_2 Y_i$, the new intercept b_1^* will be given by $b_1^* = \lambda_1 + \lambda_2 b_1$, where b_1 is the intercept in a regression of Y on X.

1.12* Suppose that the units of measurement of X are changed so that the new measure, X^*, is related to the original one by $X_i^* = \mu_1 + \mu_2 X_i$. Show that the new estimate of the slope coefficient is b_2/μ_2, where b_2 is the slope coefficient in the original regression.

1.13* Demonstrate that if X is demeaned but Y is left in its original units, the intercept in a regression of Y on demeaned X will be equal to $\overline{Y}$.

1.14* The regression output shows the result of regressing weight on height using the same sample as in Exercise 1.8, but with weight and height measured in kilos and centimetres: *WMETRIC* = 0.454* *WEIGHT85* and *HMETRIC* = 2.54* *HEIGHT*.

Confirm that the estimates of the intercept and slope coefficient are as should be expected from the change in units of measurement.

```
. gen HMETRIC = 2.54*HEIGHT
. gen WMETRIC = WEIGHT85*0.454
. reg WMETRIC HMETRIC

  Source   |       SS          df        MS              Number of obs =    540
-----------+------------------------------------         F(1, 538)     = 355.97
   Model   |   53819.2324       1     53819.2324         Prob > F      = 0.0000
 Residual  |    81340.044     538     151.189673         R-squared     = 0.3982
-----------+------------------------------------         Adj R-squared = 0.3971
   Total   |   135159.276     539     250.759325         Root MSE      = 12.296
-------------------------------------------------------------------------------
  WMETRIC  |      Coef.    Std. Err.      t  P>|t|       [95% Conf.  Interval]
-----------+-------------------------------------------------------------------
  HMETRIC  |    .9281928   .0491961   18.87  0.000        .8315528    1.024833
   _cons   |   -88.38539   8.472958  -10.43  0.000       -105.0295   -71.74125
-------------------------------------------------------------------------------
```

1.15* Consider the regression model

$$Y_i = \beta_1 + \beta_2 X_i + u_i.$$

It implies

$$\overline{Y} = \beta_1 + \beta_2 \overline{X} + \overline{u}$$

and hence that

$$Y_i^* = \beta_2 X_i^* + v_i,$$

where $Y_i^* = Y_i - \overline{Y}$, $X_i^* = X_i - \overline{X}$, and $v_i = u_i - \overline{u}$.

Demonstrate that a regression of Y^* on X^* using (1.40) will yield the same estimate of the slope coefficient as a regression of Y on X. (*Note:* (1.40) should be used instead of (1.28) because there is no intercept in this model.)

Evaluate the outcome if the slope coefficient were estimated using (1.28), despite the fact that there is no intercept in the model.

Determine the estimate of the intercept if Y^* were regressed on X^* with an intercept included in the regression specification.

1.16 Two individuals fit earnings functions relating *EARNINGS* to *S* using *EAEF* Data Set 21. The first individual does it correctly and obtains the result found in Table 1.2:

$$\widehat{EARNINGS} = -13.93 + 2.46\,S.$$

The second individual makes a mistake and regresses *S* on *EARNINGS*, obtaining the following result:

$$\hat{S} = 12.29 + 0.070 EARNINGS.$$

From this result the second individual derives

$$\widehat{EARNINGS} = -175.57 + 14.29S.$$

Explain why this equation is different from that fitted by the first individual.

1.5 Two important results relating to OLS regressions

It is convenient at this point to establish two important results relating to OLS regressions. Both of them are purely mechanical. They are valid only if the model includes an intercept (as is usually the case). They hold automatically, irrespective of whether the model is well or poorly specified. Both generalize to the multiple regression case where we have more than one explanatory variable. We will use them and their corollaries immediately in the next section, where we consider goodness of fit. As usual, the true model is assumed to be $Y_i = \beta_1 + \beta_2 X_i + u_i$ and the fitted model to be $\hat{Y}_i = b_1 + b_2 X_i$.

The mean value of the residuals is zero

To demonstrate this, we start with the definition of the residual e_i in observation i:

$$e_i = Y_i - \hat{Y}_i = Y_i - b_1 - b_2 X_i. \tag{1.47}$$

Summing over all the observations in the sample,

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} Y_i - nb_1 - b_2 \sum_{i=1}^{n} X_i. \tag{1.48}$$

Dividing by n,

$$\begin{aligned} \bar{e} &= \bar{Y} - b_1 - b_2 \bar{X} \\ &= \bar{Y} - (\bar{Y} - b_2 \bar{X}) - b_2 \bar{X} = 0. \end{aligned} \tag{1.49}$$

As a corollary, we can immediately demonstrate that the mean of the fitted values of Y is equal to the mean of the actual values of Y. We start again with the definition of the residual,

$$e_i = Y_i - \hat{Y}_i. \tag{1.50}$$

Summing over all the observations in the sample and dividing by n, we have

$$\bar{e} = \bar{Y} - \bar{\hat{Y}}. \tag{1.51}$$

But $\bar{e} = 0$, so $\bar{\hat{Y}} = \bar{Y}$.

The sample correlation between the observations on *X* and the residuals is zero

Intuitively, we would expect this to be the case since the residuals are, by definition, the part of Y not explained by X in the model. We will first demonstrate that $\sum X_i e_i = 0$.

$$\sum_{i=1}^{n} X_i e_i = \sum_{i=1}^{n} X_i \left(Y_i - b_1 - b_2 X_i\right) = \sum_{i=1}^{n} X_i Y_i - b_1 \sum_{i=1}^{n} X_i - b_2 \sum_{i=1}^{n} X_i^2 = 0. \qquad \textbf{(1.52)}$$

The final step uses the normal equation (1.22). The numerator of the sample correlation coefficient for X and e can be decomposed as follows, using the fact that $\bar{e} = 0$:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(e_i - \bar{e}\right) &= \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)e_i \\ &= \frac{1}{n}\sum_{i=1}^{n} X_i e_i - \frac{1}{n}\sum_{i=1}^{n} \bar{X} e_i \\ &= 0 - \bar{X}\frac{1}{n}\sum_{i=1}^{n} e_i = 0. \qquad \textbf{(1.53)}\end{aligned}$$

Hence, the correlation coefficient is zero, assuming that it is defined. (This requires that the denominator is nonzero. which in turn requires that the sample variances of X and e are both nonzero.)

As a corollary of (1.52), we have $\sum \hat{Y}_i e_i = 0$:

$$\sum_{i=1}^{n} \hat{Y}_i e_i = \sum_{i=1}^{n} \left(b_1 + b_2 X_i\right) e_i = b_1 \sum_{i=1}^{n} e_i + b_2 \sum_{i=1}^{n} X_i e_i = 0, \qquad \textbf{(1.54)}$$

since $\sum e_i = n\bar{e} = 0$. Hence, we may demonstrate that the sample correlation between the fitted values of Y and the residuals is zero. This is left as an exercise.

EXERCISE

1.17* Demonstrate that the fitted values of the dependent variable are uncorrelated with the residuals in a simple regression model. (This result generalizes to the multiple regression case.)

1.6 Goodness of fit: R^2

The aim of regression analysis is to explain the behavior of the dependent variable Y. In any given sample, Y is relatively low in some observations and relatively high in others. We want to know why. The variations in Y in any sample can be summarized by $\sum \left(Y_i - \bar{Y}\right)^2$, the sum of the squared deviations about its sample mean. We should like to be able to account for the size of this statistic.

We have seen that we can split the value of Y_i in each observation into two components, $\hat{Y}_i$ and e_i, after running a regression:

$$Y_i = \hat{Y}_i + e_i. \tag{1.55}$$

We can use this to decompose $\sum\left(Y_i - \overline{Y}\right)^2$:

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\left[\hat{Y}_i + e_i\right] - \left[\overline{\hat{Y}} + \overline{e}\right]\right)^2 = \sum_{i=1}^{n}\left(\left[\hat{Y}_i - \overline{Y}\right] + e_i\right)^2. \tag{1.56}$$

In the second step, we have used the results $\overline{e} = 0$ and $\overline{\hat{Y}} = \overline{Y}$ from Section 1.5. Hence,

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 &= \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n}\left(\left[\hat{Y}_i - \overline{Y}\right]e_i\right)\\ &= \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} e_i^2 + 2\sum_{i=1}^{n}\hat{Y}_i e_i - 2\overline{Y}\sum_{i=1}^{n} e_i.\end{aligned} \tag{1.57}$$

Now $\sum \hat{Y}_i e_i = 0$, as demonstrated in Section 1.5, and $\sum e_i = n\overline{e} = 0$. Hence,

$$\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2 = \sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2 + \sum_{i=1}^{n} e_i^2. \tag{1.58}$$

Thus, we have the decomposition

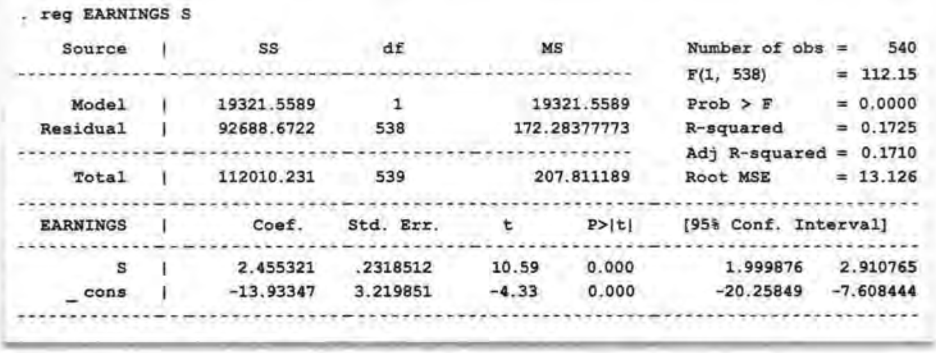
$$TSS = ESS + RSS, \tag{1.59}$$

where *TSS*, the total sum of squares, is given by the left side of the equation and *ESS*, the 'explained' sum of squares, and *RSS*, the residual ('unexplained') sum of squares, are the two terms on the right side. (*Note:* The words *explained* and *unexplained* have been put in quotation marks because the explanation may in fact be false. *Y* might really depend on some other variable *Z*, and *X* might be acting as a proxy for *Z* (more about this later). It would be safer to use the expression *apparently explained* instead of *explained*.)

In view of (1.58), $\sum\left(\hat{Y}_i - \overline{Y}\right)^2 \Big/ \sum\left(Y_i - \overline{Y}\right)^2$ is the proportion of the total sum of squares explained by the regression line. This proportion is known as the coefficient of determination or, more usually, R^2:

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}. \tag{1.60}$$

Table 1.4



```
. reg EARNINGS S
   Source   |        SS          df          MS               Number of obs =     540
------------------------------------------------------        F(1, 538)     =  112.15
    Model   |    19321.5589       1          19321.5589       Prob > F      =  0.0000
 Residual   |    92688.6722     538        172.28377773       R-squared     =  0.1725
------------------------------------------------------        Adj R-squared =  0.1710
    Total   |    112010.231     539          207.811189       Root MSE      =  13.126
------------------------------------------------------------------------------------
 EARNINGS   |       Coef.    Std. Err.       t       P>|t|       [95% Conf. Interval]
------------------------------------------------------------------------------------
        S   |     2.455321    .2318512     10.59     0.000        1.999876   2.910765
    _ cons  |    -13.93347    3.219851     -4.33     0.000       -20.25849  -7.608444
------------------------------------------------------------------------------------
```

Regression output always includes R^2 and may also present the underlying analysis of variance. Table 1.4 reproduces the Stata earnings function output in Table 1.2. The column heading 'SS' stands for sums of squares. *ESS*, here described as the 'model' sum of squares, is 19,322. *TSS* is 112,010. Dividing *ESS* by *TSS*, we have $R^2 = 19{,}322/112{,}010 = 0.1725$, as stated in the top right quarter of the output. The low R^2 is partly attributable to the fact that important variables, such as work experience, are missing from the model. It is also partly attributable to the fact that unobservable characteristics are important in determining earnings, R^2 seldom being much above 0.5 even in a well-specified model.

The maximum value of R^2 is 1. This occurs when the regression line fits the observations exactly, so that $\hat{Y}_i = Y_i$ in all observations and all the residuals are zero. Then $\sum\left(\hat{Y}_i - \bar{Y}\right)^2 = \sum\left(Y_i - \bar{Y}\right)^2$, $\sum e_i^2 = 0$, and one has a perfect fit. If there is no apparent relationship between the values of Y and X in the sample, R^2 will be close to zero.

Other things being equal, one would like R^2 to be as high as possible. In particular, we would like the coefficients b_1 and b_2 to be chosen in such a way as to maximize R^2. Does this conflict with our criterion that b_1 and b_2 should be chosen to minimize the sum of the squares of the residuals? No, they are easily shown to be equivalent criteria. In view of (1.58), we can rewrite R^2 as

$$R^2 = 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2} \qquad (1.61)$$

and so the values of b_1 and b_2 that minimize the residual sum of squares automatically maximize R^2.

Note that the results in Section 1.5 depend on the model including an intercept. If there is no intercept, the decomposition (1.58) is invalid and the two definitions of R^2 in equations (1.60) and (1.61) are no longer equivalent. Any definition of R^2 in this case may be misleading and should be treated with caution.

Table 1.5 Analysis of variance in the three-observation example

Observation	X	Y	$\hat{Y}$	e	$Y-\bar{Y}$	$\hat{Y}-\bar{\hat{Y}}$	$(Y-\bar{Y})^2$	$(\hat{Y}-\bar{\hat{Y}})^2$	e^2
1	1	3	3.1667	−0.1667	−1.6667	−1.5	2.7778	2.25	0.0278
2	2	5	4.6667	0.3333	0.3333	0.0	0.1111	0.00	0.1111
3	3	6	6.1667	−0.1667	1.3333	1.5	1.7778	2.25	0.0278
Total	6	14	14				4.6667	4.50	0.1667
Mean	2	4.6667	4.6667						

Example of how R^2 is calculated

R^2 is always calculated by the computer as part of the regression output, so this example is for illustration only. We shall use the primitive three-observation example described in Section 1.3, where the regression line

$$\hat{Y}_i = 1.6667 + 1.5000X_i \tag{1.62}$$

was fitted to the observations on X and Y in Table 1.5. The table also shows $\hat{Y}_i$ and e_i for each observation. $\sum\left(Y_i-\bar{Y}\right)^2 = 4.6667$, $\sum\left(\hat{Y}_i-\bar{Y}\right)^2 = 4.5000$, and $\sum e_i^2 = 0.1667$. From these figures, we can calculate R^2 using either (1.60) or (1.61):

$$R^2 = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2} = \frac{4.5000}{4.6667} = 0.96 \tag{1.63}$$

$$R^2 = 1-\frac{\sum_{i=1}^{n}e_i^2}{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2} = 1-\frac{0.1667}{4.6667} = 0.96. \tag{1.64}$$

Alternative interpretation of R^2

It should be intuitively obvious that, the better is the fit achieved by the regression equation, the higher should be the correlation coefficient for the actual and predicted values of Y. We will show that R^2 is in fact equal to the square of this correlation coefficient, which we will denote $r_{Y,\hat{Y}}$:

$$r_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)\left(\hat{Y}_i-\bar{Y}\right)}{\sqrt{\sum_{i=1}^{n}\left(Y_i-\bar{Y}\right)^2\sum_{i=1}^{n}\left(\hat{Y}_i-\bar{Y}\right)^2}}. \tag{1.65}$$

Now

$$\begin{aligned}\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)\left(\hat{Y}_i-\overline{Y}\right) &= \sum_{i=1}^{n}\left(\left[\hat{Y}_i+e_i\right]-\left[\overline{Y}+\overline{e}\right]\right)\left(\hat{Y}_i-\overline{Y}\right)\\ &= \sum_{i=1}^{n}\left(\left[\hat{Y}_i-\overline{Y}\right]+e_i\right)\left(\hat{Y}_i-\overline{Y}\right)\\ &= \sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2+\sum_{i=1}^{n}e_i\hat{Y}_i-\overline{Y}\sum_{i=1}^{n}e_i\\ &= \sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2. \qquad (1.66)\end{aligned}$$

In the second line we have used $\overline{e}=0$ and in the fourth we have used $\sum \hat{Y}_i e_i = 0$, as demonstrated in Section 1.5. In the fourth line we have also used $\sum e_i = n\overline{e} = 0$. Hence,

$$r_{Y,\hat{Y}} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2}{\sqrt{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2}} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{Y}_i-\overline{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i-\overline{Y}\right)^2}} = \sqrt{R^2}. \qquad (1.67)$$

Key terms

- coefficient of determination
- dependent variable
- disturbance term
- explained sum of squares (*ESS*)
- explanatory variable
- fitted model
- fitted value
- independent variable
- least squares criterion
- multiple regression model
- ordinary least squares (OLS)
- parameter
- R^2
- regression model
- regressor
- residual
- residual sum of squares (*RSS*)
- simple regression model
- total sum of squares (*TSS*)

EXERCISES

1.18 Using the data in Table 1.5, calculate the correlation between Y and $\hat{Y}$ and verify that its square is equal to the value of R^2.

1.19 What was the value of R^2 in the educational attainment regression fitted by you in Exercise 1.6? Comment on it.

1.20 What was the value of R^2 in the earnings function fitted by you in Exercise 1.7? Comment on it.

1.21 Demonstrate that, in a regression with an intercept, a regression of Y^* on X must have the same R^2 as a regression of Y on X, where $Y^* = \lambda_1 + \lambda_2 Y$.

1.22* Demonstrate that, in a regression with an intercept, a regression of Y on X^* must have the same R^2 as a regression of Y on X, where $X^* = \mu_1 + \mu_2 X$.

1.23 In a regression with an intercept, show that R^2 is zero if the estimated slope coefficient is zero.

1.24* The output shows the result of regressing weight in 2002 on height, using *EAEF* Data Set 21. In 2002 the respondents were aged 37–44. Explain why R^2 is lower than in the regression reported in Exercise 1.8.

```
. reg WEIGHT02 HEIGHT

  Source  |       SS          df          MS          Number of obs =    540
----------+-----------------------------------        F(1, 538)     = 216.95
   Model  |   311260.383       1     311260.383       Prob > F      = 0.0000
Residual  |   771880.527     538     1434.72217       R-squared     = 0.2874
----------+-----------------------------------        Adj R-squared = 0.2860
   Total  |   1083140.91     539     2009.53787       Root MSE      = 37.878

------------------------------------------------------------------------------
WEIGHT02  |      Coef.    Std. Err.      t   P>|t|       [95% Conf. Interval]
----------+-------------------------------------------------------------------
  HEIGHT  |    5.669766   .3849347   14.73   0.000       4.913606    6.425925
   _cons  |   -199.6832   26.10105   -7.65   0.000      -250.9556   -148.4107
------------------------------------------------------------------------------
```

1.25 In Exercise 1.16 both researchers obtained values of R^2 equal to 0.17 in their regressions. Was this a coincidence?