2. Properties of the Regression Coefficients and Hypothesis Testing

With the aid of regression analysis, we can obtain estimates of the parameters of a relationship. However, they are only *estimates*. The next question to ask is, how reliable are they? What are their properties? We will investigate these questions in this chapter. Both the way that we ask these questions, and their answers, depend upon the assumptions that we are making relating to the regression model, and these in turn depend upon the nature of the data that we are using.

2.1 Types of data and regression model

We shall be applying our regression techniques to three kinds of data: cross-sectional, time series, and panel. Cross-sectional data consist of observations relating to units of observation at one moment in time. The units of observation may be individuals, households, enterprises, countries, or any set of elements that are sufficiently similar in nature to allow one reasonably to use them to explore hypothetical relationships. Time series data consist of repeated observations through time on the same entities, usually with fixed intervals between the observations. Examples within a macroeconomic context would be quarterly data on gross domestic product, consumption, the money supply, and interest rates. Panel data, which can be thought of as combining the features of cross-sectional data and time series data, consist of repeated observations on the same elements through time. An example is the US National Longitudinal Survey of Youth used to illustrate the interpretation of a regression in Section 1.4. This consists of observations on the same individuals from 1979 to the present, interviews having been conducted annually until 1994 and every two years since then.

Following the treatment in Davidson (2000), we will consider three types of regression model:

Model A (for regressions using cross-sectional data): the regressors (explanatory variables) are nonstochastic. This means that their values in the observations in a sample do not have stochastic (random) components. See Box 2.1 for a brief further discussion.

BOX 2.1 Nonstochastic regressors

For the first part of this text, until Chapter 8, we will assume that the regressors (explanatory variables) in the model do not have stochastic components. This is to simplify the analysis. In fact, it is not easy to think of truly nonstochastic variables, other than time, so the following example is a little artificial. Suppose that we are relating earnings to schooling, S, in terms of highest grade completed. Suppose that we know from the national census that 1 percent of the population have $S = 8$, 3 percent have $S = 9$, 5 percent have $S = 10$, 7 percent have $S = 11$, 43 percent have $S = 12$ (graduation from high school), and so on. Suppose that we have decided to undertake a survey with sample size 1,000 and we want the sample to match the population as closely as possible. We might then select what is known as a stratified random sample, designed so that it includes 10 individuals with $S = 8$, 30 individuals with $S = 9$, and so on. The values of S in the sample would then be predetermined and therefore nonstochastic. In large surveys drawn in such a way as to be representative of the population as a whole, such as the National Longitudinal Survey of Youth, schooling and other demographic variables probably approximate this condition quite well. In Chapter 8 we will acknowledge the restrictiveness of this assumption and replace it with the assumption that the values of the regressors are drawn from defined populations.

Model B (also for regressions using cross-sectional data): the values of the regressors are drawn randomly and independently from defined populations.

Model C (for regressions using time series data): the values of the regressors may exhibit persistence over time. The meaning of 'persistent over time' will be explained when we come to time series regressions in Chapters 11–13.

Regressions with panel data will be treated as an extension of Model B.

The first part of this text will be confined to regressions using cross-sectional data, that is, Models A and B. The reason for this is that regressions with time series data potentially involve complex technical issues that are best avoided initially.

We will start with Model A. We will do this purely for analytical convenience. It enables us to conduct the discussion of regression analysis within the relatively straightforward framework of what is known as the Classical Linear Regression Model. We will replace it in Chapter 8 by the weaker and more realistic assumption, appropriate for regressions with cross-sectional data, that the variables are randomly drawn from defined populations.

2.2 Assumptions for regression models with nonstochastic regressors

To examine the properties of the regression model we need to make some assumptions. In particular, for Model A, we will make the following six assumptions.

A.1 The model is linear in parameters and correctly specified.

$$Y = \beta_1 + \beta_2 X + u. \tag{2.1}$$

'Linear in parameters' means that each term on the right side includes a β as a simple factor and there is no built-in relationship among the βs. An example of a model that is not linear in parameters is

$$Y = \beta_1 X^{\beta_2} + u. \tag{2.2}$$

We will defer a discussion of issues relating to linearity and nonlinearity to Chapter 4.

A.2 There is some variation in the regressor in the sample.

Obviously, if X is constant in the sample, it cannot account for any of the variation in Y. If we tried to regress Y on X, when X is constant, we would find that we would not be able to compute the regression coefficients. X_i would be equal to $\bar{X}$ for all i and hence both the numerator and the denominator of

$$b_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \tag{2.3}$$

would be equal to zero. Since we would not be able to compute b_2, we would not be able to obtain b_1 either, except in the special (and unusual) case where $\bar{X}$ is zero.

A.3 The disturbance term has zero expectation.

$$E(u_i) = 0 \quad \text{for all } i. \tag{2.4}$$

We assume that the expected value of the disturbance term in any observation should be zero. Sometimes the disturbance term will be positive, sometimes negative, but it should not have a systematic tendency in either direction.

Actually, if an intercept is included in the regression equation, it is usually reasonable to assume that this condition is satisfied automatically since the role of the intercept is to pick up any systematic but constant tendency in Y not accounted for by the explanatory variables included in the regression equation. To put this mathematically, suppose that our regression model is

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2.5}$$

and

$$E(u_i) = \mu_u, \tag{2.6}$$

where $\mu_u \neq 0$. Define

$$v_i = u_i - \mu_u. \tag{2.7}$$

Then, using (2.7) to substitute for u_i in (2.5), one has

$$\begin{aligned} Y_i &= \beta_1 + \beta_2 X_i + v_i + \mu_u \\ &= \beta_1' + \beta_2 X_i + v_i, \end{aligned} \tag{2.8}$$

where $\beta_1' = \beta_1 + \mu_u$. The disturbance term in the respecified model now satisfies the condition because

$$E(v_i) = E(u_i - \mu_u) = E(u_i) - E(\mu_u) = \mu_u - \mu_u = 0. \tag{2.9}$$

The price that we pay is that the interpretation of the intercept has changed. It has absorbed the nonzero component of the disturbance term in addition to whatever had previously been responsible for it. Usually this does not matter because we are seldom interested in the intercept in a regression model.

A.4 The disturbance term is homoscedastic.

We assume that the disturbance term is homoscedastic, meaning that its value in each observation is drawn from a distribution with constant population variance. Here we are thinking about the *potential* distribution of the disturbance term *before* the sample is actually generated. Once we have generated the sample, the disturbance term will turn out to be greater in some observations, and smaller in others, but there should not be any reason for it to be more erratic in some observations than in others. Denoting the potential variance of the disturbance term $\sigma_{u_i}^2$ in observation i, the assumption is

$$\sigma_{u_i}^2 = \sigma_u^2 \quad \text{for all } i. \tag{2.10}$$

Since $E(u_i) = \mu_u = 0$ by virtue of Assumption A.3, the population variance of u_i, $E\left\{(u_i - \mu_u)^2\right\}$, is equal to $E(u_i^2)$, so the condition may also be written

$$E(u_i^2) = \sigma_u^2 \quad \text{for all } i. \tag{2.11}$$

Of course, σ_u^2 is unknown. One of the tasks of regression analysis is to estimate the variance of the disturbance term.

If Assumption A.4 is not satisfied, the OLS regression coefficients will be inefficient, and it should be possible to obtain more reliable results by using a modification of the regression technique. This will be discussed in Chapter 7.

A.5 The values of the disturbance term have independent distributions.

$$u_i \text{ is distributed independently of } u_j \quad \text{for all } j \neq i. \tag{2.12}$$

We assume that the disturbance term is not subject to autocorrelation, meaning that there should be no systematic association between its values in any two observations. For example, just because the disturbance term is large and positive in one observation, there should be no tendency for it to be large and

positive in the next (or large and negative, for that matter, or small and positive, or small and negative). The values of the disturbance term should be absolutely independent of one another.

The assumption implies that $\sigma_{u_i u_j}$, the population covariance between u_i and u_j, is zero, because

$$\begin{aligned}\sigma_{u_i u_j} &= E\{(u_i - \mu_u)(u_j - \mu_u)\} = E(u_i u_j) \\ &= E(u_i)E(u_j) = 0.\end{aligned} \tag{2.13}$$

(Note that $E(u_i u_j)$ can be decomposed as $E(u_i)E(u_j)$ if u_i and u_j are generated independently—see the Review chapter.)

If this assumption is not satisfied, OLS will again give inefficient estimates. Chapter 12 discusses the problems that arise and ways of treating them. Violations of this assumption are rare with cross-sectional data.

With these assumptions, we will show in this chapter that the OLS estimators of the coefficients are BLUE: best (most efficient) linear (function of the observations on Y) unbiased estimators and that the sum of the squares of the residuals divided by the number of degrees of freedom provides an unbiased estimator of σ_u^2.

A.6 The disturbance term has a normal distribution.

We usually assume that the disturbance term has a normal distribution. If u is normally distributed, so will be the regression coefficients, and this will be useful to us later in the chapter when we perform t tests and F tests of hypotheses and construct confidence intervals for β_1 and β_2 using the regression results.

The justification for the assumption depends on the Lindeberg–Feller central limit theorem. In essence, this states that, if a random variable is the composite result of the effects of a large number of other random variables, it will have an approximately normal distribution even if its components do not, provided that none of them is dominant. The disturbance term u is composed of a number of factors not appearing explicitly in the regression equation and so, even if we know nothing about the distributions of these factors (or even their identity), we are usually entitled to assume that the disturbance term is normally distributed. (The Lindeberg–Levy central limit theorem discussed in the Review chapter required all the random components to be drawn from the same distribution; the Lindeberg–Feller theorem is less restrictive.)

2.3 The random components and unbiasedness of the OLS regression coefficients

The random components of the OLS regression coefficients

A least squares regression coefficient is a special form of random variable whose properties depend on those of the disturbance term in the equation. This will be demonstrated first theoretically in this section and again in the next by means of a controlled experiment.

Throughout the discussion we shall continue to work with the simple regression model where Y depends on a nonstochastic variable X according to the relationship

$$Y_i = \beta_1 + \beta_2 X_i + u_i \tag{2.14}$$

and we are fitting the regression equation

$$\hat{Y}_i = b_1 + b_2 X_i \tag{2.15}$$

given a sample of n observations.

First, note that Y_i has two components. It has a nonrandom component ($\beta_1 + \beta_2 X_i$), which owes nothing to the laws of chance (β_1 and β_2 may be unknown, but nevertheless they are fixed constants), and it has the random component u_i.

This implies that, when we calculate b_2 according to the formula

$$b_2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}, \tag{2.16}$$

b_2 also has a random component. $\sum (X_i - \bar{X})(Y_i - \bar{Y})$ depends on the values of Y, and the values of Y depend on the values of u. If the values of the disturbance term had been different in the n observations, we would have obtained different values of Y, hence of $\sum (X_i - \bar{X})(Y_i - \bar{Y})$, and hence of b_2.

We can in theory decompose b_2 into its nonrandom and random components. In view of (2.14),

$$\begin{aligned}
\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) &= \sum_{i=1}^{n} (X_i - \bar{X})\left([\beta_1 + \beta_2 X_i + u_i] - [\beta_1 + \beta_2 \bar{X} + \bar{u}]\right) \\
&= \sum_{i=1}^{n} (X_i - \bar{X})\left(\beta_2 [X_i - \bar{X}] + [u_i - \bar{u}]\right) \\
&= \beta_2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u}).
\end{aligned} \tag{2.17}$$

Hence,

$$\begin{aligned}
b_2 &= \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{\beta_2 \sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \\
&= \beta_2 + \frac{\sum_{i=1}^{n} (X_i - \bar{X})(u_i - \bar{u})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}.
\end{aligned} \tag{2.18}$$

Thus, we have shown that the regression coefficient b_2 obtained from any sample consists of (1) a fixed component, equal to the true value, β_2, and (2) a random component dependent on the values of the disturbance term in the sample. The random component is responsible for the variations of b_2 around its fixed component β_2. If we wish, we can express this decomposition more tidily:

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})(u_i-\bar{u}) &= \sum_{i=1}^{n}(X_i-\bar{X})u_i-\bar{u}\sum_{i=1}^{n}(X_i-\bar{X}) \\ &= \sum_{i=1}^{n}(X_i-\bar{X})u_i-\bar{u}\left(\sum_{i=1}^{n}X_i-n\bar{X}\right) \\ &= \sum_{i=1}^{n}(X_i-\bar{X})u_i \end{aligned} \tag{2.19}$$

since $\sum_{i=1}^{n}X_i = n\bar{X}$. Hence,

$$b_2=\beta_2+\frac{\sum_{i=1}^{n}(X_i-\bar{X})u_i}{\sum_{i=1}^{n}(X_i-\bar{X})^2}=\beta_2+\sum_{i=1}^{n}\left\{\frac{(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}\right\}u_i=\beta_2+\sum_{i=1}^{n}a_iu_i, \tag{2.20}$$

where

$$a_i=\frac{(X_i-\bar{X})}{\sum_{i=1}^{n}(X_i-\bar{X})^2}. \tag{2.21}$$

Thus, we have shown that b_2 is equal to the true value, β_2, plus a linear combination of the values of the disturbance term in all the observations in the sample. There is a slight awkwardness in the definition of a_i and it is as well to deal with it before mathematicians start getting excited. The numerator $(X_i-\bar{X})$ changes as i changes and is different for each observation. However, the denominator is the sum of the squared deviations for the whole sample and is not dependent on i. So we are using i in two different senses in the definition. To avoid any ambiguity, we will use a different index for the summation and write the denominator $\sum(X_j-\bar{X})^2$. It still means the same thing. We could avoid the problem entirely by writing the denominator as $(X_1-\bar{X})^2+\cdots+(X_n-\bar{X})^2$, but this would be clumsy.

We will note for future reference three properties of the a_i coefficients:

$$\sum_{i=1}^{n}a_i=0,\quad \sum_{i=1}^{n}a_i^2=\frac{1}{\sum_{j=1}^{n}(X_j-\bar{X})^2},\quad \text{and}\quad \sum_{i=1}^{n}a_iX_i=1. \tag{2.22}$$

Proofs are supplied in Box 2.2.

BOX 2.2 Proofs of three properties of the a_i coefficients

Proof that $\sum_{i=1}^{n} a_i = 0$:

$$\sum_{i=1}^{n} a_i = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\right) = \frac{1}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\sum_{i=1}^{n}(X_i - \bar{X}) = 0$$

since

$$\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n} X_i - n\bar{X} = n\bar{X} - n\bar{X} = 0$$

using $\bar{X} = \frac{1}{n}\sum X_i$.

Proof that $\sum_{i=1}^{n} a_i^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$:

$$\sum_{i=1}^{n} a_i^2 = \sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\right)^2 = \frac{1}{\left(\sum_{j=1}^{n}(X_j - \bar{X})^2\right)^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Proof that $\sum_{i=1}^{n} a_i X_i = 1$:

First note that

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 &= \sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X}) = \sum_{i=1}^{n}(X_i - \bar{X})X_i - \sum_{i=1}^{n}(X_i - \bar{X})\bar{X} \\ &= \sum_{i=1}^{n}(X_i - \bar{X})X_i - \bar{X}\sum_{i=1}^{n}(X_i - \bar{X}) = \sum_{i=1}^{n}(X_i - \bar{X})X_i\end{aligned}$$

since $\sum(X_i - \bar{X}) = 0$ (see above). Then, using the above equation in reverse,

$$\sum_{i=1}^{n} a_i X_i = \sum_{i=1}^{n}\frac{(X_i - \bar{X})X_i}{\sum_{j=1}^{n}(X_j - \bar{X})^2} = \frac{1}{\sum_{j=1}^{n}(X_j - \bar{X})^2}\sum_{i=1}^{n}(X_i - \bar{X})X_i = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sum_{j=1}^{n}(X_j - \bar{X})^2} = 1.$$

In a similar manner, one may also show that b_1 has a fixed component equal to the true value, β_1, plus a random component that is a linear combination of the values of the disturbance term:

$$b_1 = \beta_1 + \sum_{i=1}^{n} c_i u_i, \tag{2.23}$$

where $c_i = 1/n - a_i \bar{X}$ and a_i is defined in equation (2.21). The proof is left as an exercise.

Note that you are not able to make these decompositions in practice because you do not know the true values of β_1 and β_2 or the actual values of u in the sample. We are interested in the decompositions because they enable us to analyze the theoretical properties of b_1 and b_2, given the regression model assumptions.

The unbiasedness of the OLS regression coefficients

From (2.20) it follows that b_2 is an unbiased estimator of β_2:

$$E(b_2) = E(\beta_2) + E\left\{\sum_{i=1}^{n} a_i u_i\right\} = \beta_2 + \sum_{i=1}^{n} E(a_i u_i) = \beta_2 + \sum_{i=1}^{n} a_i E(u_i) = \beta_2 \tag{2.24}$$

since $E(u_i) = 0$ for all i by Assumption A.3. The a_i coefficients are nonstochastic given the assumption that the values of X are nonstochastic. Hence, $E(a_i u_i) = a_i E(u_i)$. Unless the random factor in the n observations happens to cancel out exactly, which can happen only by coincidence, b_2 will be different from β_2 for any given sample, but in view of (2.24) there will be no systematic tendency for it to be either higher or lower.

Similarly, b_1 is an unbiased estimator of β_1:

$$E(b_1) = E(\beta_1) + E\left(\sum_{i=1}^{n} c_i u_i\right) = \beta_1 + \sum_{i=1}^{n} c_i E(u_i) = \beta_1. \tag{2.25}$$

OLS estimators of the parameters are not the only unbiased estimators. We will give an example of another. We continue to assume that the true relationship between Y and X is given by $Y_i = \beta_1 + \beta_2 X_i + u_i$. Someone who had never heard of regression analysis, on seeing a scatter diagram of a sample of observations, might be tempted to obtain an estimate of the slope merely by joining the first and the last observations, and by dividing the increase in the height by the horizontal distance between them, as in Figure 2.1. The estimator b_2 would then be given by

$$b_2 = \frac{Y_n - Y_1}{X_n - X_1}. \tag{2.26}$$

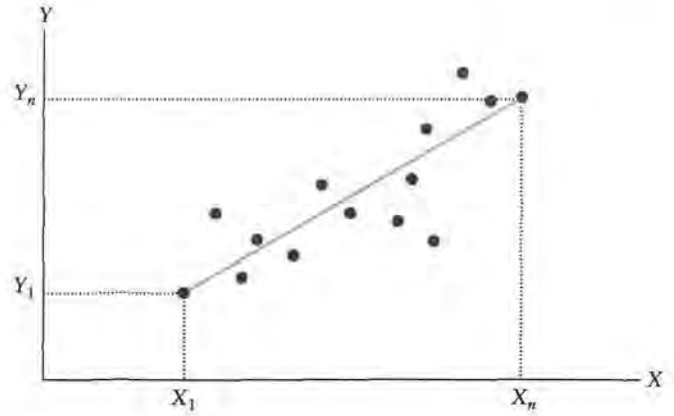

Figure 2.1 Naïve estimation of β_2

We will investigate whether it is biased or unbiased. For the first and last observations, we have

$$Y_1 = \beta_1 + \beta_2 X_1 + u_1 \tag{2.27}$$

and

$$Y_n = \beta_1 + \beta_2 X_n + u_n. \tag{2.28}$$

Hence,

$$b_2 = \frac{\beta_1 + \beta_2 X_n + u_n - \beta_1 - \beta_2 X_1 - u_1}{X_n - X_1}$$

$$= \beta_2 + \frac{u_n - u_1}{X_n - X_1}. \tag{2.29}$$

Thus, we have decomposed this naïve estimator into two components, the true value and an error term. This decomposition is parallel to that for the OLS estimator, but the error term is different. The expected value of the estimator is given by

$$E(b_2) = E(\beta_2) + E\left[\frac{u_n - u_1}{X_n - X_1}\right]$$

$$= \beta_2 + \frac{1}{X_n - X_1} E(u_n - u_1) \tag{2.30}$$

since β_2 is a constant and X_1 and X_n are nonstochastic. If Assumption A.3 is satisfied,

$$E(u_n - u_1) = E(u_n) - E(u_1) = 0. \tag{2.31}$$

Therefore, despite being naïve, this estimator is unbiased.

This is not by any means the only estimator besides OLS that is unbiased. You could derive one by joining any two arbitrarily selected observations, and in fact the possibilities are infinite if you are willing to consider less naïve procedures.

It is intuitively easy to see that we would not prefer a naïve estimator such as (2.26) to OLS. Unlike OLS, which takes account of every observation, it employs only the first and the last and is wasting most of the information in the sample. The naïve estimator will be sensitive to the value of the disturbance term u in those two observations, whereas the OLS estimator combines all the values of the disturbance term and takes greater advantage of the possibility that to some extent they cancel each other out. More rigorously, it can be shown that the population variance of the naïve estimator is greater than that of the OLS estimator, and that the naïve estimator is therefore less efficient. We will discuss efficiency in Section 2.5.

Normal distribution of the regression coefficients

A further implication of the decompositions (2.20) and (2.23) is that the regression coefficients will have normal distributions if the disturbance term in each observation has a normal distribution, as specified in Assumption A.6. This is because a linear combination of normal distributions is itself a normal distribution (a result that we shall take for granted).

Even if Assumption A.6 is invalid and the disturbance term has some other distribution, the distributions of the regression coefficients may be approximately normal. We may be able to invoke a central limit theorem that tells us that the linear combination of the values of the disturbance term may be approximately normal, even if each value has a non-normal distribution, provided that the sample size is large enough.

EXERCISES

2.1* Derive the decomposition of b_1 shown in equation (2.23).

2.2 For the model $Y_i = \beta_2 X_i + u_i$, the OLS estimator of β_2 is $b_2 = \sum_{i=1}^{n} X_i Y_i \Big/ \sum_{i=1}^{n} X_i^2$. Demonstrate that b_2 may be decomposed as

$$b_2 = \beta_2 + \sum_{i=1}^{n} d_i u_i,$$

where $d_i = \dfrac{X_i}{\sum_{j=1}^{n} X_j^2}$, and hence demonstrate that it is an unbiased estimator of β_2.

2.3 For the model $Y_i = \beta_1 + u_i$, the OLS estimator of β_1 is $b_1 = \bar{Y}$. Demonstrate that b_1 may be decomposed into the true value plus a linear combination of the

disturbance terms in the sample. Hence, demonstrate that b_1 is an unbiased estimator of β_1.

2.4 An investigator correctly believes that the relationship between two variables X and Y is given by $Y_i = \beta_1 + \beta_2 X_i + u_i$. Given a sample of n observations, the investigator estimates β_2 by calculating it as the average value of Y divided by the average value of X. Discuss the properties of this estimator. What difference would it make if it could be assumed that $\beta_1 = 0$?

2.5* An investigator correctly believes that the relationship between two variables X and Y is given by $Y_i = \beta_1 + \beta_2 X_i + u_i$. Given a sample of observations on Y, X, and a third variable Z (which is not a determinant of Y), the investigator estimates β_2 as

$$\frac{\sum_{i=1}^{n}(Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(Z_i - \bar{Z})(X_i - \bar{X})}.$$

Demonstrate that this estimator is unbiased.

2.4 A Monte Carlo experiment

It often happens in econometrics that we can establish the asymptotic properties of estimators analytically but we can say nothing analytically for finite samples. The reason for this is that asymptotically we may be in a position to use plims where we cannot use expectations, and we may be able to invoke a central limit theorem that cannot be applied to finite samples.

In reality, we deal with finite samples, but practitioners typically ignore the issue. If they can demonstrate that an estimator is consistent, they may then (sometimes with justification) assume that any element of bias in a finite sample can be ignored. Similarly, if they can demonstrate that asymptotically the distribution of the estimator is normal, then they may assume that it will be approximately normal for a finite sample and so the usual tests will be approximately valid.

Alternatively, practitioners may undertake a simulation to investigate the finite sample properties directly under controlled conditions. Such simulations are often described as Monte Carlo experiments. Nobody seems to know for certain how the term originated. Probably it has something to do with the famous casino, as a symbol of the laws of chance.

An example of a simulation was provided in Section R.14 of the Review chapter. We will undertake many simulations of this type later in the text, particularly in the context of regressions using time series data, where it is often impossible to establish finite-sample properties analytically.

Simulations are often also useful for expository purposes. Even if it is possible to establish the finite-sample properties of an estimator mathematically, it

may be helpful to illustrate them graphically. We will do this for the OLS estimators of the parameters of the simple regression model

$$Y_i = \beta_1 + \beta_2 X_i + u_i. \tag{2.32}$$

To perform the simulation, first (step 1),

1. you choose the true values of β_1 and β_2,
2. you choose the value of X in each observation, and
3. you use some random number generating process to provide the random element u in each observation.

Next (step 2), you *generate* the value of Y in each observation, using the relationship (2.32) and the values of β_1, β_2, X, and u. Then (step 3), using only the values of Y thus generated and the data for X, you use regression analysis to obtain estimates b_1 and b_2.

In the first two steps you are preparing a challenge for the regression technique. You are in complete control of the model that you are constructing and you *know* the true values of the parameters because you yourself have determined them. In the third step you see how the regression technique provides estimates of β_1 and β_2 using only the data for Y and X. Note that the inclusion of a stochastic term in the generation of Y is responsible for the element of challenge. If you did not include it, the observations would lie exactly on the straight line $Y_i = \beta_1 + \beta_2 X_i$, and it would be a trivial matter to determine the exact values of β_1 and β_2 from the data for Y and X.

You repeat the process a large number of times, keeping the *same* values of β_1 and β_2, and the *same* values of X, but using a *new* set of random numbers for the random element u each time. Finally, you plot the distributions of b_1 and b_2. If you have been able to determine these analytically, you can check that they conform to what you anticipated. If analysis was not possible, you have gained information otherwise unavailable to you.

Obviously, the distributions of b_1 and b_2 will depend on your choice of β_1 and β_2 and on your choice of the data for X. This may even be a focus of interest, as in the simulations in Chapters 11 and 13. However, as in the present case, we are interested in the qualitative nature of the results and they do not depend on our choices.

Quite arbitrarily, let us put β_1 equal to 2 and β_2 equal to 0.5, so the true relationship is

$$Y_i = 2 + 0.5X_i + u_i. \tag{2.33}$$

To keep things simple, we will assume that we have 20 observations and that the values of X go from 1 to 20. For u, the disturbance term, we will use random numbers drawn from a normally distributed population with zero mean and unit variance. We will need a set of 20 and will denote them rn_1 to rn_{20}. u_1, the disturbance term in the first observation, is simply equal to rn_1, u_2 to rn_2, etc.

Given the values of X_i and u_i in each observation, the values of Y_i are determined by (2.33), and this is done in Table 2.1.

Table 2.1

X	u	Y	X	u	Y
1	−0.59	1.91	11	1.59	9.09
2	−0.24	2.76	12	−0.92	7.08
3	−0.83	2.67	13	−0.71	7.79
4	0.03	4.03	14	−0.25	8.75
5	−0.38	4.12	15	1.69	11.19
6	−2.19	2.81	16	0.15	10.15
7	1.03	6.53	17	0.02	10.52
8	0.24	6.24	18	−0.11	10.89
9	2.53	9.03	19	−0.91	10.59
10	−0.13	6.87	20	1.42	13.42

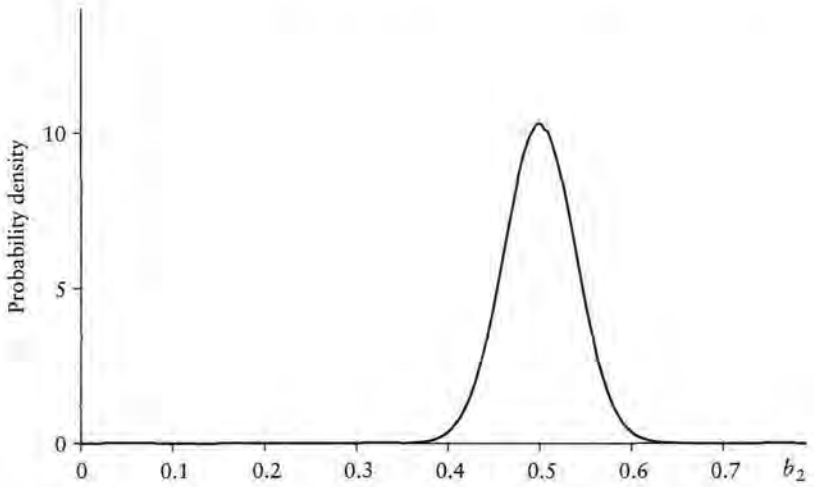

Figure 2.2 Distribution of b_2 in the Monte Carlo experiment

If you now regress Y on X, you obtain

$$\hat{Y}_i = 1.63 + 0.54X_i. \tag{2.34}$$

In this case, b_1 is an underestimate of β_1 (1.63 as opposed to 2.00) and b_2 is a slight overestimate of β_2 (0.54 as opposed to 0.50). The discrepancies are caused by the collective effects of the values of the disturbance term in the 20 observations.

We perform the regression with 10 million samples, each time keeping the same values of β_1, β_2, and the same observations on X, but using a fresh set of 20 random values for the disturbance term each time. The distribution of b_2 is shown in Figure 2.2.

b_2 is symmetrically distributed around the true value, 0.5, confirming that it is an unbiased estimator. We can see this visually, and we can also compute the

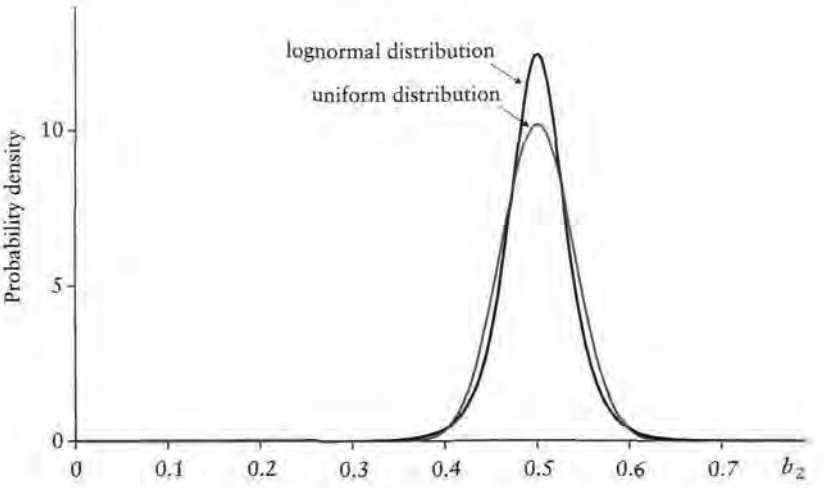

Figure 2.3 Distributions of b_2 with uniform and lognormal distributions for the disturbance term

mean of the 10 million estimates, which is exactly 0.5 to four decimal places. Of course, the distribution we see in Figure 2.2 depends on the value that we have given to β_2. However, qualitatively, the finding does not depend on this choice. b_2 would have been distributed around the true value, whatever it was.

Figure 2.2 also confirms the assertion that, if Assumption A.6 is satisfied and the disturbance term has a normal distribution, b_2, being a linear combination of the values of the disturbance term, must also have a normal distribution.

However, suppose that Assumption A.6 is not valid. What can one say then? This issue is explored with two variations on the Monte Carlo experiment whose results are shown in Figure 2.3. In the first, the disturbance term was generated from a uniform distribution with mean zero and variance 1. We know from the Lindeberg–Levy central limit theorem that the mean of a sample of observations from any distribution with finite mean and variance will have a distribution that converges on the normal distribution as the sample size becomes large. This was demonstrated in the case of the uniform and lognormal distributions in Section R.15 of the Review chapter. For the uniform distribution, the convergence was close to perfect for a sample size as small as 10. In the present context, we are not dealing with the straight mean of a set of observations, but the weighted linear combination $\sum_{i=1}^{n} a_i u_i$ in equation (2.20). However, another central limit theorem, the Lindeberg–Feller CLT, covers this case. Subject to certain conditions, the weighted linear combination will also have a normal distribution. The distribution of b_2 with the disturbance term generated from a uniform distribution is shown in Figure 2.3. It is identical to the normal distribution in Figure 2.2.

A more severe test is provided by the case where the disturbance term has a lognormal distribution, because the lognormal distribution is highly skewed.

When the application of the Lindeberg–Levy CLT to the lognormal distribution was investigated in Section R.15, it was found that the sample size had to be quite large for the distribution of the mean to approach normality. Even for $n = 100$, the shape was noticeably non-normal. Naturally, the outcome is similar in the present context. The value of the disturbance term in each observation was generated by drawing a value from a lognormal distribution with mean 1 and variance 1, and then subtracting 1 to make the mean zero. The distribution of the estimates of β_2 is shown in Figure 2.3. Unsurprisingly, it is distinctly non-normal, with excessively high mode and excessively large tails. However, it should be noted that the sample size is very small and that the lognormal distribution may be considered to be an extreme case.

Returning to Figure 2.2 where the disturbance term had a normal distribution, the standard deviation of the distribution is 0.0388. We have not yet said anything analytically about the determinants of the dispersion of the distribution. To this we now turn.

2.5 Precision of the regression coefficients

Variances of the regression coefficients

The population variances of b_1 and b_2 about their population means, $\sigma_{b_1}^2$ and $\sigma_{b_2}^2$, are given by the following expressions:

$$\sigma_{b_1}^2 = \sigma_u^2 \left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right) \quad \text{and} \quad \sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}. \tag{2.35}$$

Box 2.3 provides a proof of the expression for $\sigma_{b_2}^2$. The proof of the expression for $\sigma_{b_1}^2$ follows similar lines and is left as an exercise.

We will focus on the implications of the expression for $\sigma_{b_2}^2$. Clearly, the larger is $\sum (X_i - \bar{X})^2$, the smaller is the variance of b_2. However, the size of $\sum (X_i - \bar{X})^2$ depends on two factors: the number of observations and the size of the deviations of X_i about its sample mean. To discriminate between them, it is convenient to define the mean square deviation of X, MSD(X):

$$\text{MSD}(X) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2. \tag{2.36}$$

Using this to rewrite $\sigma_{b_2}^2$ as

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{n \, \text{MSD}(X)}, \tag{2.37}$$

it is then obvious that the variance of b_2 is inversely proportional to the number of observations in the sample, holding the mean square deviation constant.

BOX 2.3 Proof of the expression for the population variance of b_2

By definition,

$$\sigma_{b_2}^2 = E\left\{(b_2 - E(b_2))^2\right\} = E\left\{(b_2 - \beta_2)^2\right\}$$

since we have shown that $E(b_2) = \beta_2$. We have seen that

$$b_2 = \beta_2 + \sum_{i=1}^{n} a_i u_i,$$

where

$$a_i = \frac{(X_i - \bar{X})}{\sum_{j=1}^{n} (X_j - \bar{X})^2}.$$

Hence,

$$\sigma_{b_2}^2 = E\left\{\left(\sum_{i=1}^{n} a_i u_i\right)^2\right\}.$$

Expanding the quadratic,

$$\begin{aligned}\sigma_{b_2}^2 &= E\left\{\sum_{i=1}^{n} a_i^2 u_i^2 + \sum_{i=1}^{n}\sum_{j\neq i}^{n} a_i a_j u_i u_j\right\} \\ &= \sum_{i=1}^{n} a_i^2 E(u_i^2) + \sum_{i=1}^{n}\sum_{j\neq i}^{n} a_i a_j E(u_i u_j).\end{aligned}$$

Now, by virtue of Assumption A.4, $E(u_i^2) = \sigma_u^2$, and by virtue of Assumption A.5, $E(u_i u_j) = 0$, for $j \neq i$, so

$$\sigma_{b_2}^2 = \sum_{i=1}^{n} a_i^2 \sigma_u^2 = \sigma_u^2 \sum_{i=1}^{n} a_i^2 = \frac{\sigma_u^2}{\sum_{j=1}^{n} (X_j - \bar{X})^2},$$

using the second of the properties of the a_i coefficients proved in Box 2.2.

This makes good sense. The larger the number of observations, the more closely will the sample resemble the population from which it is drawn, and the more accurate b_2 should be as an estimator of β_2.

It is also obvious that the variance of b_2 is proportional to the variance of the disturbance term. The bigger the variance of the random factor in the relationship, the worse the estimates of the parameters are likely to be, other things being equal. This is illustrated graphically in Figures 2.4a and 2.4b. We will use the same model as in the Monte Carlo experiment in Section 2.4. In both diagrams, the nonstochastic component of the relationship between Y and X, depicted by the dotted line, is given by

$$Y_i = 2.0 + 0.5X_i. \tag{2.38}$$

There are 20 observations, with the values of X being the integers from 1 to 20. In the two figures, the same random numbers are used to generate the values of the disturbance term, but those in Figure 2.4b have been multiplied by a factor of 5. As a consequence, the regression line, depicted by the solid line, is a much poorer approximation to the nonstochastic relationship in Figure 2.4b than in Figure 2.4a.

From (2.37) it can be seen mathematically that the variance of b_2 is inversely related to the mean square deviation of X. What is the reason for this? The regression coefficients are calculated on the assumption that the observed variations in Y are attributable to variations in X, but in reality the observed variations in Y are *partly* attributable to variations in X and *partly* to variations in u.

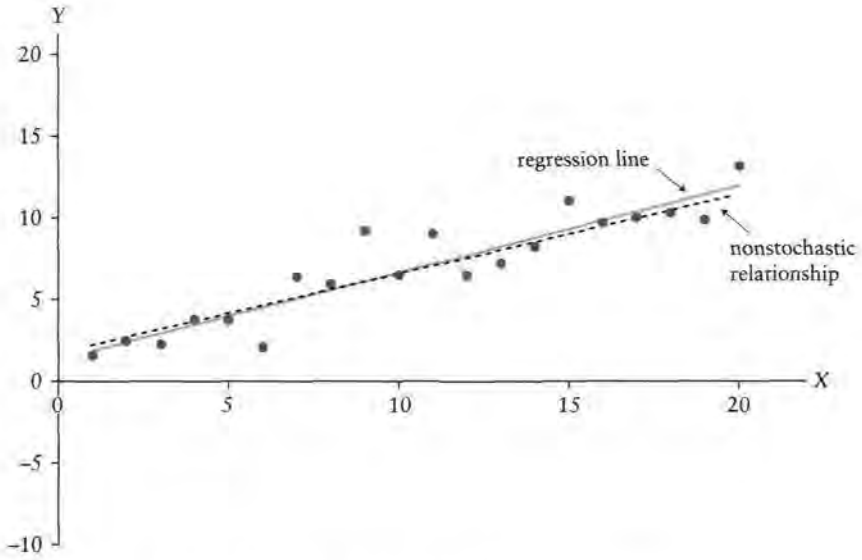

Figure 2.4a Disturbance term with relatively small variance

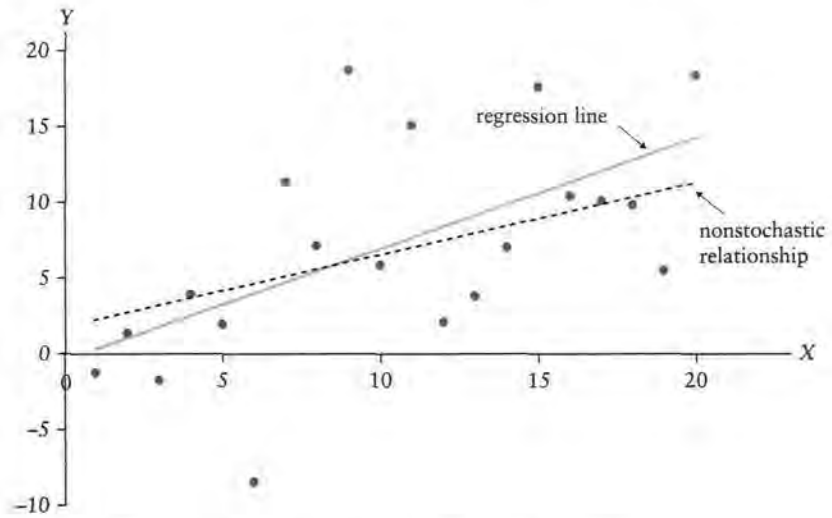

Figure 2.4b Disturbance term with relatively large variance

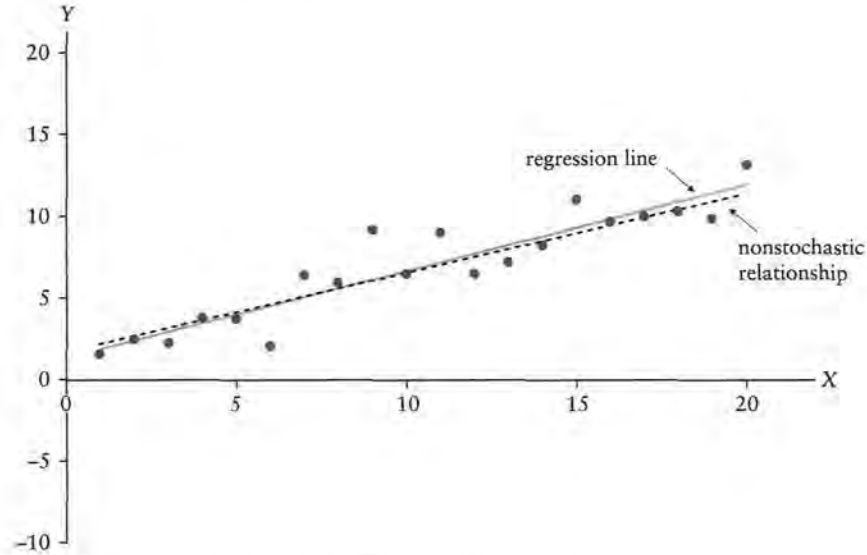

Figure 2.5a X with relatively large mean square deviation

The smaller the variations in X, as summarized by its mean square deviation, the greater is likely to be the relative influence of the random factor in determining the variations in Y and the more likely is regression analysis to give inaccurate estimates.

This is illustrated by Figures 2.5a and 2.5b. The nonstochastic component of the relationship is given by (2.38), and the disturbance terms in the two figures are identical. In Figure 2.5a, the values of X are the integers from 1

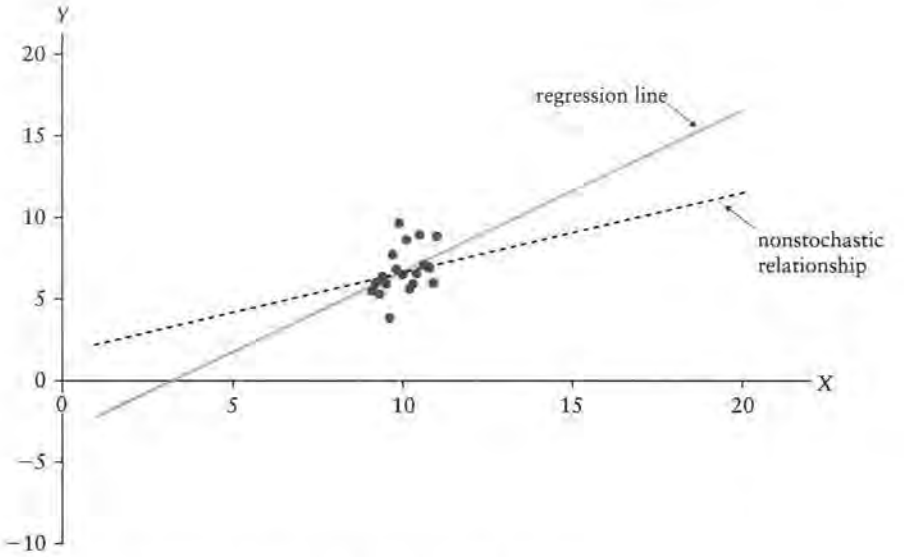

Figure 2.5b X with relatively small mean square deviation

to 20. In Figure 2.5b, the values of X are the numbers 9.1, 9.2, ..., 10.9, 11. In Figure 2.5a, the variations in X are responsible for most of the variations in Y and the relationship between the two variables can be determined relatively accurately. However, in Figure 2.5b, the variations in X are so small that their influence is overwhelmed by the effect of the variance of u. As a consequence, the effect of X is difficult to pick out and the estimates of the regression coefficients are likely to be relatively inaccurate.

Figure 2.5b is sufficient to demonstrate that the slope of the regression line is likely to be sensitive to the actual values of the disturbance term in the sample. We can make the same point more systematically with a simulation. Figure 2.6 shows the distributions of the slope coefficients when regressions similar to those shown in Figures 2.5a and 2.5b are performed with 10 million samples. The variance of the distribution is much greater in the case where X is limited to the range 9–11. Indeed, sometimes it yields estimates with the wrong sign.

Of course, Figures 2.4 and 2.5 make the same point in different ways. As can be seen from (2.37), it is the *relative* size of σ_u^2 and MSD(X) that is important, rather than the *actual* size of either.

Standard errors of the regression coefficients

In practice, one cannot calculate the population variances of either b_1 or b_2 because σ_u^2 is unknown. However, we can derive an estimator of σ_u^2 from the residuals. Clearly the scatter of the residuals around the regression line will reflect the unseen scatter of u about the line $Y_i = \beta_1 + \beta_2 X_i$, although, in general,

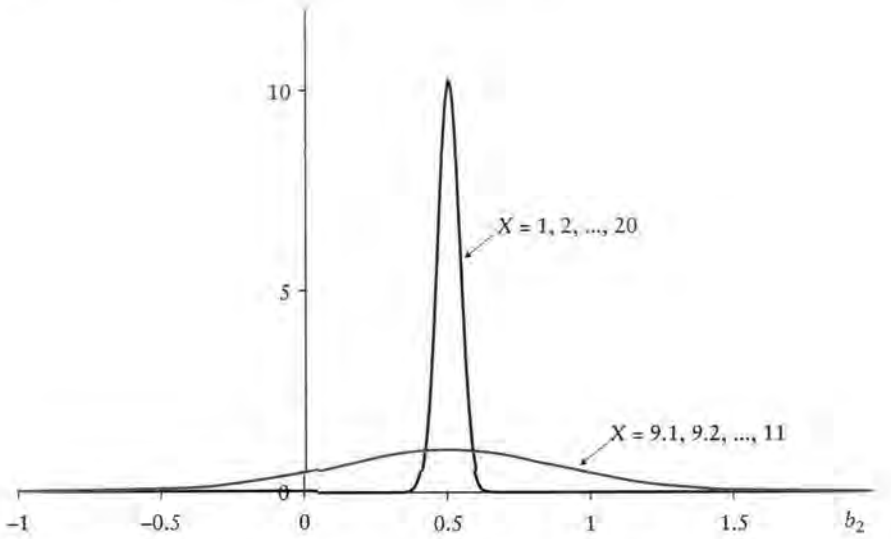

Figure 2.6 Distributions of b_2 with high and low dispersions of X

the residual and the value of the disturbance term in any given observation are not equal to one another. One measure of the scatter of the residuals is their mean square deviation, MSD(e), defined by

$$\text{MSD}(e) = \frac{1}{n}\sum_{i=1}^{n}(e_i - \bar{e})^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \tag{2.39}$$

($\bar{e} = 0$; see Section 1.5). Intuitively MSD(e) should provide a guide to σ_u^2.

Before going any further, one should consider the following question. Which line is likely to be closer to the points representing the sample of observations on X and Y, the true line $Y_i = \beta_1 + \beta_2 X_i$ or the regression line $\hat{Y}_i = b_1 + b_2 X_i$? The answer is the regression line, because by definition it is drawn in such a way as to minimize the sum of the squares of the distances between it and the observations. Hence the spread of the residuals will tend to be smaller than the spread of the values of u, and MSD(e) will tend to underestimate σ_u^2. Indeed, it can be shown that the expected value of MSD(e), when there is just one explanatory variable, is given by

$$E\{\text{MSD}(e)\} = \frac{n-2}{n}\sigma_u^2. \tag{2.40}$$

However, it follows that, if one defines s_u^2 by

$$s_u^2 = \frac{n}{n-2}\text{MSD}(e) = \frac{n}{n-2}\frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2, \tag{2.41}$$

then s_u^2 will be an unbiased estimator of σ_u^2.

Using (2.35) and (2.41), one can obtain estimates of the population variances of b_1 and b_2 and, by taking square roots, estimates of their standard deviations. Rather than talk about the estimate of the standard deviation of the probability density function of a regression coefficient, which is a bit cumbersome, one uses the term standard error of a regression coefficient, which in this text will frequently be abbreviated to s.e. For simple regression analysis, therefore, one has

$$\text{s.e.}(b_1) = \sqrt{s_u^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right)} \quad \text{and} \quad \text{s.e.}(b_2) = \sqrt{\frac{s_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}. \tag{2.42}$$

The standard errors of the regression coefficients are automatically calculated as part of the computer output.

Example

These relationships will be illustrated with the Monte Carlo experiment described in Section 2.4. The disturbance term was determined by random numbers drawn from a population with zero mean and unit variance, so $\sigma_u^2 = 1$. X was the set of numbers from 1 to 20. $\bar{X} = 10.5$ and $\sum(X_i - \bar{X})^2 = 665$. Hence,

$$\sigma_{b_1}^2 = \sigma_u^2\left(\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right) = \frac{1}{20} + \frac{10.5^2}{665} = 0.2158 \tag{2.43}$$

and

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{1}{665} = 0.001504. \tag{2.44}$$

Therefore, the true standard deviation of the distribution of b_2 shown in Figure 2.2 (and again as the narrower distribution in Figure 2.6) is $\sqrt{0.001504} = 0.0388$. The distribution of the standard error for the 10 million samples is shown in Figure 2.7, with the true standard deviation also marked.

We know that b_2 has a potential distribution around β_2. We would like the standard error to be as accurate as possible as an estimator of the standard deviation of this distribution since it is our main, indeed usually only, guide to the reliability of b_2 as an estimator of β_2. Figure 2.7 suggests that, in this case, the standard error, as a reliability measure, is itself not particularly reliable. But

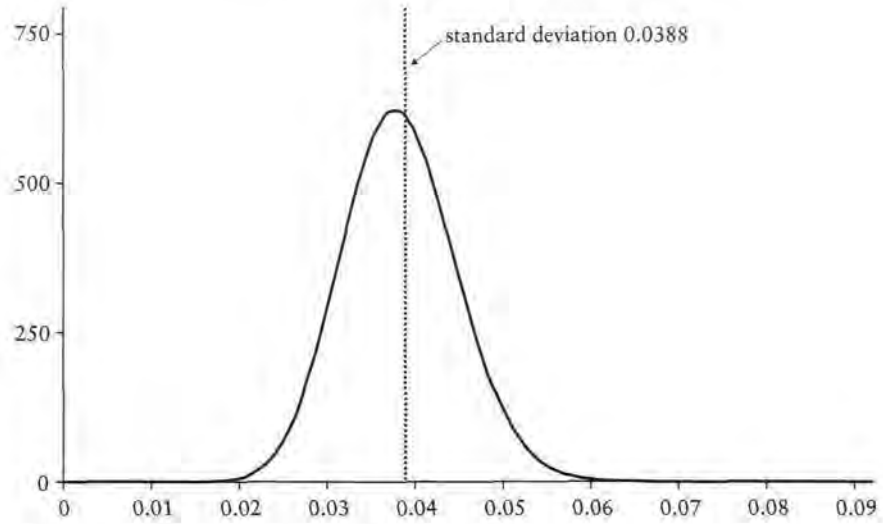

Figure 2.7 Distribution of the standard error of b_2

then, with only 20 observations, the sample is very small. With a larger sample, its distribution would be closer to the actual standard deviation.

One fundamental point must be emphasized. The standard error gives only a general guide to the likely accuracy of a regression coefficient. It enables you to obtain some idea of the width, or narrowness, of its probability density function as represented in Figure 2.2, but it does *not* tell you whether your regression estimate comes from the middle of the function, and is therefore accurate, or from the tails, and is therefore relatively inaccurate.

The higher the variance of the disturbance term, the higher the sample variance of the residuals is likely to be, and hence the higher will be the standard errors of the coefficients in the regression equation, reflecting the risk that the coefficients are inaccurate. However, it is only a *risk*. It is possible that in any particular sample the effects of the disturbance term in the different observations will cancel each other out and the regression coefficients will be accurate after all. The trouble is that in general there is no way of telling whether you happen to be in this fortunate position or not.

The Gauss–Markov theorem

In the Review chapter, we considered estimators of the unknown population mean μ_X of a random variable X, given a sample of observations. Although we instinctively use the sample mean $\bar{X}$ as our estimator, we saw that it was only one of an infinite number of possible unbiased estimators of μ_X. The reason that the sample mean is preferred to any other estimator is that, under certain assumptions, it is the most efficient. Of all unbiased estimators, it has the smallest variance.

Similar considerations apply to regression coefficients. The Gauss–Markov theorem states that, provided that the assumptions in Section 2.2 are satisfied, the OLS estimators are efficient. Sometimes, they are described as BLUE: best (smallest variance) linear (combinations of the Y_i) unbiased estimators of the regression parameters. This is demonstrated for the slope coefficient b_2 in Appendix 2.1.

EXERCISES

2.6* Using the decomposition of b_1 obtained in Exercise 2.1, derive the expression for $\sigma_{b_1}^2$ given in equation (2.35).

2.7* Given the decomposition in Exercise 2.2 of the OLS estimator of β_2 in the model $Y_i = \beta_2 X_i + u_i$, demonstrate that the variance of the slope coefficient is given by

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n} X_i^2}.$$

2.8 Given the decomposition in Exercise 2.3 of the OLS estimator of β_1 in the model $Y_i = \beta_1 + u_i$, demonstrate that the variance of the slope coefficient is given by

$$\sigma_{b_1}^2 = \frac{\sigma_u^2}{n}.$$

2.9 Assuming that the true model is $Y_i = \beta_1 + \beta_2 X_i + u_i$, it was demonstrated in Section 2.3 that the naïve estimator of the slope coefficient,

$$b_2 = \frac{Y_n - Y_1}{X_n - X_1},$$

is unbiased. It can be shown that its variance is given by

$$\frac{\sigma_u^2}{(X_1 - \bar{X})^2 + (X_n - \bar{X})^2 - 0.5(X_1 + X_n - 2\bar{X})^2}.$$

Use this information to verify that the estimator is less efficient than the OLS estimator.

2.10* It can be shown that the variance of the estimator of the slope coefficient in Exercise 2.5,

$$\frac{\sum_{i=1}^{n} (Z_i - \bar{Z})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (Z_i - \bar{Z})(X_i - \bar{X})},$$

is given by

$$\sigma_{b_2}^2 = \frac{\sigma_u^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \times \frac{1}{r_{XZ}^2},$$

where r_{XZ} is the correlation between X and Z. What are the implications for the efficiency of the estimator?

2.11 Can one come to any conclusions concerning the efficiency of the estimator in Exercise 2.4, for the case $\beta_1 = 0$?

2.12 Suppose that the true relationship between Y and X is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. In Section 1.4, it was shown that if $Y_i^* = \lambda_1 + \lambda_2 Y_i$, and Y^* is regressed on X, the slope coefficient $b_2^* = \lambda_2 b_2$. How will the standard error of b_2^* be related to the standard error of b_2?

2.13* Suppose that the true relationship between Y and X is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. In Exercise 1.12, it was shown that if $X_i^* = \mu_1 + \mu_2 X_i$, and Y is regressed on X^*, the slope coefficient $b_2^* = b_2/\mu_2$. How will the standard error of b_2^* be related to the standard error of b_2?

2.6 Testing hypotheses relating to the regression coefficients

The principles relating to tests of hypotheses and the construction of confidence intervals have been discussed in Sections R.9–R.12 of the Review chapter. There the context was the estimation of the unknown mean of a random variable X with a normal distribution with mean μ and variance σ^2. We will begin by summarizing what we did there.

The estimator of μ was $\bar{X}$. If a null hypothesis H_0: $\mu = \mu_0$ is true, the potential distribution of $\bar{X}$ is as shown in Figure 2.8, which reproduces Figure R.15. The figure supposes that we know the standard deviation of the distribution. We decided to reject H_0 if the discrepancy between μ_0 and $\bar{X}$ was 'too great'. 'Too great' is, of course, a subjective matter. In the case of a 5 percent significance test, we decided to reject H_0 if $\bar{X}$ fell in either the upper or the lower 2.5 percent tails of the distribution, conditional on H_0: $\mu = \mu_0$ being true. These are the rejection regions in Figure 2.8. They start 1.96 standard deviations above and below μ_0. Thus, a value of $\bar{X}$ represented by the point A would not lead to a rejection of H_0, while a value represented by the point B would lead to rejection. In general, we would reject H_0 if $\bar{X} > \mu_0 + 1.96$ s.d. or if $\bar{X} < \mu_0 - 1.96$ s.d. Equivalently, we could say that we would reject H_0 if $z > 1.96$ or if $z < -1.96$, where $z = (\bar{X} - \mu_0)/\text{s.d.}$

All this supposes that we know the standard deviation of the distribution. In practice, it has to be estimated, and we call the estimate the standard error. As a consequence of using the estimated standard error instead of the true standard deviation, the test statistic has a t distribution instead of a normal distribution, and the decision rule is to reject H_0 if $t > t_{crit}$ or if $t < -t_{crit}$,

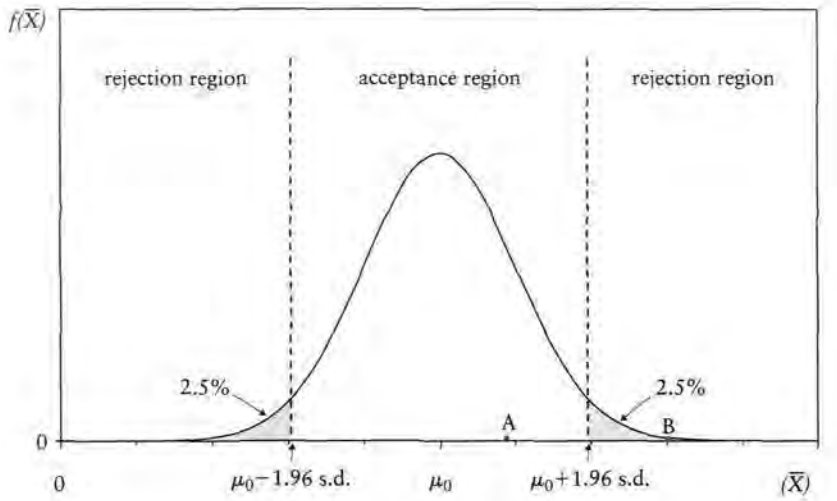

Figure 2.8 Rejection regions, conditional on H_0: $\mu = \mu_0$, 5 percent test

where $t = (\bar{X} - \mu_0)/\text{s.e.}$ and t_{crit} is the critical value of t, given the significance level and the number of degrees of freedom.

Performing tests of hypotheses relating to regression coefficients follows the same pattern in a straightforward manner. Suppose that the true model, as usual, is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and we are fitting the regression equation $\hat{Y}_i = b_1 + b_2 X_i$. Suppose that we have a hypothesis relating to the slope coefficient H_0: $\beta_2 = \beta_2^0$. We will reject H_0 if the discrepancy between b_2 and β_2^0 is too great, measured in terms of standard errors. We define the t statistic

$$t = \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} \tag{2.45}$$

and we reject H_0 if $|t| > t_{crit}$, that is, if $t > t_{crit}$ or if $t < -t_{crit}$.

There is one important difference. When we performed t tests of H_0: $\mu = \mu_0$ in Section R.11, the number of degrees of freedom was equal to $n - 1$, where n is the number of observations in the sample. In a regression equation, the estimation of each parameter consumes one degree of freedom in the sample. Hence, the number of degrees of freedom is equal to the number of observations in the sample minus the number of parameters estimated. The parameters are the constant (assuming that this is specified in the regression model) and the coefficients of the explanatory variables. In the present case of simple regression analysis, two parameters, β_1 and β_2, are estimated and hence the number of degrees of freedom is $n - 2$. It should be emphasized that a more general expression will be required when we come to multiple regression analysis.

For instance, suppose that we hypothesize that the percentage rate of price inflation in an economy, p, depends on the percentage rate of wage inflation, w,

according to the linear equation

$$p = \beta_1 + \beta_2 w + u, \tag{2.46}$$

where β_1 and β_2 are parameters and u is a disturbance term. We might further hypothesize that, apart from the effects of the disturbance term, the rate of price inflation is equal to the rate of wage inflation. The idea is that if wages increase by a certain proportion, the increase in costs is likely to give rise to a similar proportional increase in prices. The null hypothesis is then $H_0: \beta_2 = 1$ and the alternative hypothesis is $H_1: \beta_2 \neq 1$. Suppose that we take actual observations on average rates of price inflation and wage inflation over the past five years for a sample of 20 countries and the fitted model is

$$\hat{p} = -1.21 + 0.82w, \tag{2.47}$$
$$(0.05)\ (0.10)$$

where the numbers in parentheses are standard errors. The t statistic for testing $H_0: \beta_2 = 1$ is

$$t = \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} = \frac{0.82 - 1.00}{0.10} = -1.80. \tag{2.48}$$

Since there are 20 observations in the sample, the number of degrees of freedom is 18 and the critical value of t at the 5 percent significance level is 2.101. The absolute value of the t statistic is less than this, so on this occasion we do not reject the null hypothesis. The estimate 0.82 is below the hypothesized value 1.00, but not so far below as to exclude the possibility that the null hypothesis is correct.

In this example, we tested a particular null hypothesis $H_0: \beta_2 = 1$ that was suggested by theory. In practice, given a theoretical model $Y = \beta_1 + \beta_2 X + u$, we are seldom able to hypothesize that a variable X has a *specific* effect on another variable Y. Usually, the aim of a t test is less ambitious. The aim is to determine whether X has *some* effect on Y. We believe that X does influence Y and that β_2 is therefore nonzero, but we are not able to anticipate the value of the parameter. So we adopt an inverse strategy. We set up the null hypothesis $H_0: \beta_2 = 0$. We then hope to demonstrate that H_0 should be rejected. If we are successful, our interest next turns to the magnitude of the effect and the margin of error of the estimate.

For example, consider the simple earnings function

$$EARNINGS = \beta_1 + \beta_2 S + u, \tag{2.49}$$

where *EARNINGS* is hourly earnings in dollars and S is years of schooling. On very reasonable theoretical grounds, you expect earnings to be influenced by years of schooling, but your theory is not strong enough to enable you to specify a particular value for β_2. You can nevertheless establish the dependence of earnings on schooling by the inverse procedure in which you take as your

Table 2.2

```
. 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
------------------------------------------------------------------------------
```

null hypothesis the assertion that earnings do *not* depend on schooling, that is, $H_0: \beta_2 = 0$. Your alternative hypothesis is $H_1: \beta_2 \neq 0$, that is, that schooling *does* affect earnings. If you can reject the null hypothesis, you have established the relationship, at least in general terms.

Table 2.2 reproduces the regression of hourly earnings on years of schooling using data from the United States National Longitudinal Survey of Youth shown in Table 1.2. The first two columns of the lower part of the output give the names of the variables and the estimates of their coefficients. The third column gives the corresponding standard errors. The t statistic for the null hypothesis $H_0: \beta_2 = 0$, using (2.45), is simply the estimate of the coefficient divided by its standard error:

$$t = \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} = \frac{b_2 - 0}{\text{s.e.}(b_2)} = \frac{2.4553}{0.2319} = 10.59. \tag{2.50}$$

Since there are 540 observations in the sample and we have estimated two parameters, the number of degrees of freedom is 538. Table A.2 does not give the critical values of t for 538 degrees of freedom, but we know that they must be lower than the corresponding critical values for 500, since the critical value is inversely related to the number of degrees of freedom. The critical value with 500 degrees of freedom at the 5 percent level is 1.965. Hence, we can be sure that we would reject H_0 at the 5 percent level with 538 degrees of freedom and we conclude that schooling does affect earnings.

Of course, since we are using the 5 percent significance level as the basis for the test, there is in principle a 5 percent risk of a Type I error, if the null hypothesis of no effect is true. We could reduce the risk to 1 percent by using the 1 percent significance level instead. The critical value of t at the 1 percent significance level with 500 degrees of freedom is 2.586. Since the t statistic is greater than this, we see that we can easily reject the null hypothesis at this level as well.

Note that when the 5 percent and 1 percent tests lead to the same conclusion, there is no need to report both, and indeed you would look ignorant if you did. See Box 2.4 on reporting test results.

BOX 2.4 *Reporting the results of t tests*

Suppose you have a theoretical relationship

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

and your null and alternative hypotheses are H_0: $\beta_2 = \beta_2^0$, H_1: $\beta_2 \neq \beta_2^0$. Given an experimental estimate b_2 of β_2, the acceptance and rejection regions for the hypothesis for the 5 percent and 1 percent significance levels can be represented in general terms by the left part of Figure 2.9.

The right side of the figure gives the same regions for a specific example, the price inflation/wage inflation model, the null hypothesis being that β_2 is equal to 1. The null hypothesis will not be rejected at the 5 percent level if b_2 lies within 2.101 standard errors of 1, that is, in the range 0.79 to 1.21, and it will not be rejected at the 1 percent level if b_2 lies within 2.878 standard errors of 1, that is, in the range 0.71 to 1.29.

From Figure 2.9 it can be seen that there are three types of decision zone:

1. where b_2 is so far from the hypothetical β_2 that the null hypothesis is rejected at both the 5 percent and the 1 percent levels,
2. where b_2 is far enough from the hypothetical β_2 for the null hypothesis to be rejected at the 5 percent but not the 1 percent level,
3. where b_2 is close enough to the hypothetical β_2 for the null hypothesis not to be rejected at either level.

From the diagram it can be verified that if the null hypothesis is rejected at the 1 percent level, it is automatically rejected at the 5 percent level. Hence, in case (1) it is only necessary to report the rejection of the hypothesis at the 1 percent level. To report that it is rejected also at the 5 percent level is superfluous and suggests that you are not aware of this. It would be a bit like reporting that a certain high jumper can clear two meters, and then adding that the athlete can also clear one and a half meters.

In case (3), likewise, you only need to make one statement, in this case that the hypothesis is not rejected at the 5 percent level. It automatically follows that it is not rejected at the 1 percent level, and to add a statement to this effect as well would be like saying that the high jumper cannot clear one and a half meters, and also reporting that the athlete cannot clear two meters either.

Only in case (2) is it necessary (and desirable) to report the results of both tests.

Note that if you find that you can reject the null hypothesis at the 5 percent level, you should not stop there. You have established that the null hypothesis can be rejected at that level, but there remains a 5 percent chance of a Type I error. You should also perform the test at the 1 percent level. If you find that you can reject the null hypothesis at this level, this is the outcome that you should report. The risk of a Type I error is now only 1 percent and your conclusion is much more convincing. This is case (1) above. If you cannot reject at the 1 percent level, you have reached case (2) and you should report the results of both tests.

This procedure of establishing a relationship between a dependent and an explanatory variable by setting up, and then refuting, a null hypothesis H_0: $\beta_2 = 0$ is used very frequently indeed. Consequently, all serious regression applications automatically print out the t statistic for this special case: that is,

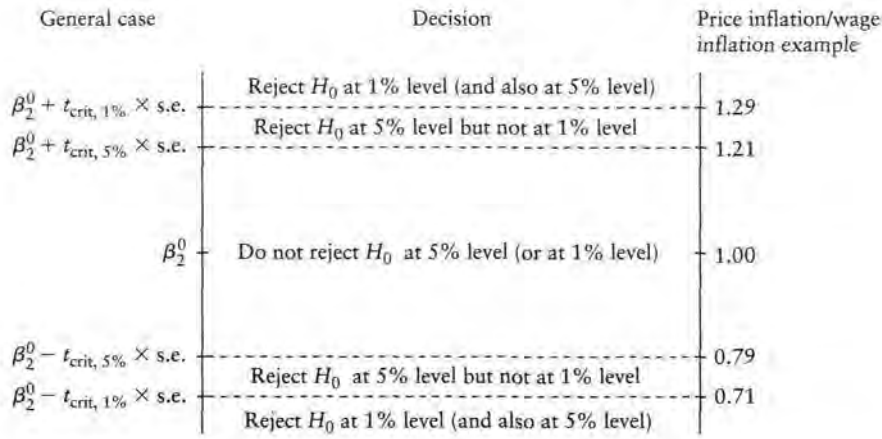

Figure 2.9 Reporting the results of a t test (no need to report conclusions in parentheses)

the coefficient divided by its standard error. The ratio is often denoted 'the' t statistic. In Table 2.2, the t statistics for the constant and slope coefficient appear in the middle column of the regression output.

However, if the null hypothesis specifies some nonzero value of β_2, the more general expression (2.71) has to be used and the t statistic has to be calculated by hand, as in the price inflation/wage inflation example.

0.1 percent tests

If the t statistic is very high, you should check whether you can reject the null hypothesis at the 0.1 percent level. If you can, you should always report the result of the 0.1 percent test in preference to that of the 1 percent test because it demonstrates that you are able to reject the null hypothesis of no effect with an even smaller risk of a Type I error.

p values

The fifth column of the lower half of the output in Table 2.2, headed $P > |t|$, provides an alternative approach to reporting the significance of regression coefficients. The figures in this column give the *p value* for each coefficient. This is the probability of obtaining the corresponding t statistic as a matter of chance, if the null hypothesis H_0: $\beta_2 = 0$ were true. A p value of less than 0.01 means that the probability is less than 1 percent, which in turn means that the null hypothesis would be rejected at the 1 percent level; a p value between 0.01 and 0.05 means that the null hypothesis would be rejected at the 5 percent, but not the 1 percent level; and a p value of 0.05 or more means that it would not be rejected at the 5 percent level.

The p value approach is more informative than the 5 percent/1 percent approach, in that it gives the exact probability of a Type I error, if the null hypothesis is true. For example, in the earnings function output in Table 2.2, the p value for the slope coefficient is 0.000, meaning that the probability of obtaining a t statistic as large as 10.59, or larger, as a matter of chance is less than 0.0005 percent. Hence, we would reject the null hypothesis that the slope coefficient is zero at the 1 percent level. Indeed, we would reject it at the 0.1 percent level.

The choice between using the p value approach and the 5 percent/1 percent approach appears to be entirely conventional. The medical literature uses p values, but the economics literature generally uses 5 percent/1 percent.

One-sided tests

The logic underlying one-sided tests, and the potential benefits from performing them, have been discussed at length in Section R.13 of the Review chapter. Application to regression analysis is straightforward. For example, in the case of the price inflation/wage inflation model

$$p = \beta_1 + \beta_2 w + u, \tag{2.51}$$

the null hypothesis was that price inflation is equal to wage inflation: H_0: $\beta_2 = 1$ since increases in wages give rise to increases in costs, and subsequently prices. In practice, there is another important element in the relationship, the rate of improvement in productivity. The rates of inflation of wages and prices may differ because improvements in productivity may cause cost inflation, and hence price inflation, to be lower than wage inflation. Certainly, improvements in productivity will not cause price inflation to be greater than wage inflation and so in this case we are justified in ruling out $\beta_2 > 1$. We are left with H_0: $\beta_2 = 1$ and H_1: $\beta_2 < 1$. Given the regression result

$$\begin{aligned} \hat{p} &= -1.21 + 0.82w \\ &\quad (0.05)\ (0.10) \end{aligned} \tag{2.52}$$

for a sample of 20 countries, the t statistic for the null hypothesis is

$$t = \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} = \frac{0.82 - 1}{0.10} = -1.80. \tag{2.53}$$

This is not high enough, in absolute terms, to cause H_0 to be rejected at the 5 percent level using a two-sided test (critical value 2.10 for 18 degrees of freedom). However, if we use a one-sided test, as we are entitled to, the critical value falls to 1.73 and we *can* reject the null hypothesis. In other words, we can conclude that price inflation is significantly lower than wage inflation.

One-sided tests are particularly useful in the very common case where we have adopted the inverse approach to demonstrating that a variable Y is influenced by another variable X. We have set up the model $Y_i = \beta_1 + \beta_2 X_i + u_i$ and

we hope to show that H_0: $\beta_2 = 0$ can be rejected. Often, we are in a position to argue either that β_2 cannot be negative or that it cannot be positive. If we can argue that it cannot be negative, the alternative hypothesis is H_1: $\beta_2 > 0$. If we can argue that it cannot be positive, the alternative hypothesis is H_1: $\beta_2 < 0$. In either case, the test statistic is

$$t = \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} = \frac{b_2 - 0}{\text{s.e.}(b_2)} = \frac{b_2}{\text{s.e.}(b_2)} \tag{2.54}$$

and we compare it with the critical value of t. At any given significance level, the critical value of t for a one-sided test is smaller in magnitude than that for a two-sided test. This will sometimes make it possible to reject H_0 and establish the relationship with a one-sided test when we could not with a two-sided test.

There are three possibilities. One is that the effect of X on Y is very strong, the t statistic is large, and we reject H_0 at a high significance level, even if we use a two-sided test. Obviously, we would also reject H_0 using a one-sided test. The second is that the effect is weak or non-existent, and we do not reject H_0, even using a one-sided test. Obviously, we would not reject H_0 using a two-sided test. The third possibility is that the t statistic lies between the critical values of t for the one-sided and two-sided tests. Only in this case is there any actual benefit from using a one-sided test.

The earnings–schooling regression in Table 2.2 provides an example. There are 538 degrees of freedom and the critical value of t, using the 0.1 percent significance level and a two-sided test, is 3.31. However, one may reasonably rule out the possibility that, in general, extra schooling will be responsible for a fall in earnings. Hence, one may perform a one-sided test with H_0: $\beta_2 = 0$ and H_1: $\beta_2 > 0$. For a one-sided test, the critical value is reduced to 3.11. The t statistic is in fact equal to 10.59, so in this case the refinement makes no difference. The estimated coefficient is so large relative to its standard error that we reject the null hypothesis at the 0.1 percent significance level regardless of whether we use a two-sided or a one-sided test.

A final comment on the justification of the use of a one-sided test is in order. In the case of the earnings–schooling example, it is tempting to say that we are justified in using a one-sided test because schooling can be expected to have a positive effect on earnings. However, this is too strong. By assumption, we have excluded the null hypothesis H_0: $\beta_2 = 0$ before we start. There is nothing left to test. The correct justification is that we can exclude the possibility of a negative effect. This leaves us with the null hypothesis H_0: $\beta_2 = 0$ and the alternative hypothesis H_1: $\beta_2 > 0$, and we use the test to discriminate between them.

Similarly, in the case of the price inflation/wage inflation example, it is tempting to say that we can use a one-sided test because improvements in productivity will cause the rate of price inflation to be lower than the rate of wage inflation. Again, this is too strong, because we have excluded H_0: $\beta_2 = 1$ before we start.

We should say that we exclude the possibility that $\beta_2 > 1$, which leaves us to test the null hypothesis against the alternative hypothesis H_1: $\beta_2 < 1$.

Confidence intervals

Confidence intervals were also treated at length in the Review chapter and their application to regression analysis presents no problems. We will briefly provide the mathematical derivation in the context of a regression. For a further graphical explanation, see Section R.12.

From the initial discussion in this section, we saw that, given a theoretical model $Y_i = \beta_1 + \beta_2 X_i + u_i$ and a fitted model $\hat{Y}_i = b_1 + b_2 X_i$, the regression coefficient b_2 and a hypothetical value $\beta_2 = \beta_2^0$ are incompatible if either

$$\frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} > t_{\text{crit}} \quad \text{or} \quad \frac{b_2 - \beta_2^0}{\text{s.e.}(b_2)} < -t_{\text{crit}}, \tag{2.55}$$

that is, if either

$$b_2 - \beta_2^0 > \text{s.e.}(b_2) \times t_{\text{crit}} \quad \text{or} \quad b_2 - \beta_2^0 < -\text{s.e.}(b_2) \times t_{\text{crit}}, \tag{2.56}$$

that is, if either

$$b_2 - \text{s.e.}(b_2) \times t_{\text{crit}} > \beta_2^0 \quad \text{or} \quad b_2 + \text{s.e.}(b_2) \times t_{\text{crit}} < \beta_2^0. \tag{2.57}$$

It therefore follows that a hypothetical β_2 *is* compatible with the regression result if both

$$b_2 - \text{s.e.}(b_2) \times t_{\text{crit}} \le \beta_2 \quad \text{and} \quad b_2 + \text{s.e.}(b_2) \times t_{\text{crit}} \ge \beta_2, \tag{2.58}$$

that is, if β_2 satisfies the double inequality

$$b_2 - \text{s.e.}(b_2) \times t_{\text{crit}} \le \beta_2 \le b_2 + \text{s.e.}(b_2) \times t_{\text{crit}}. \tag{2.59}$$

This is the confidence interval, in abstract. Any hypothetical value of β_2 that satisfies (2.59) will be compatible with the estimate b_2, that is, will not be rejected by it. To make the confidence interval operational, we need to select a significance level and determine the corresponding critical value of t.

Example

In the earnings function output in Table 2.2, the coefficient of S was 2.455, its standard error was 0.232, and the critical value of t at the 5 percent significance level was about 1.965. The corresponding 95 percent confidence interval is therefore

$$2.455 - 0.232 \times 1.965 \le \beta_2 \le 2.455 + 0.232 \times 1.965, \tag{2.60}$$

that is,

$$1.999 \leq \beta_2 \leq 2.911. \tag{2.61}$$

We would therefore reject hypothetical values below 1.999 and above 2.911. Any hypotheses within these limits would not be rejected, given the regression result. This confidence interval actually appears as the final column in the Stata output. However, this is not a standard feature of a regression application, so you usually have to calculate the interval yourself.

EXERCISES

2.14 A researcher hypothesizes that years of schooling, S, may be related to the number of siblings (brothers and sisters), *SIBLINGS*, according to the relationship

$$S = \beta_1 + \beta_2 SIBLINGS + u.$$

She is prepared to test the null hypothesis H_0: $\beta_2 = 0$ against the alternative hypothesis H_1: $\beta_2 \neq 0$ at the 5 percent and 1 percent levels. She has a sample of 60 observations. What should she report:

1. if $b_2 = -0.20$, s.e.$(b_2) = 0.07$?
2. if $b_2 = -0.12$, s.e.$(b_2) = 0.07$?
3. if $b_2 = 0.06$, s.e.$(b_2) = 0.07$?
4. if $b_2 = 0.20$, s.e.$(b_2) = 0.07$?

2.15* A researcher with a sample of 50 individuals with similar education but differing amounts of training hypothesizes that hourly earnings, *EARNINGS*, may be related to hours of training, *TRAINING*, according to the relationship

$$EARNINGS = \beta_1 + \beta_2 TRAINING + u.$$

He is prepared to test the null hypothesis H_0: $\beta_2 = 0$ against the alternative hypothesis H_1: $\beta_2 \neq 0$ at the 5 percent and 1 percent levels. What should he report:

1. if $b_2 = 0.30$, s.e.$(b_2) = 0.12$?
2. if $b_2 = 0.55$, s.e.$(b_2) = 0.12$?
3. if $b_2 = 0.10$, s.e.$(b_2) = 0.12$?
4. if $b_2 = -0.27$, s.e.$(b_2) = 0.12$?

2.16 Perform a t test on the slope coefficient and the intercept of the educational attainment function fitted using your *EAEF* data set in Exercise 1.6, and state your conclusions.

2.17 Perform a t test on the slope coefficient and the intercept of the earnings function fitted using your *EAEF* data set in Exercise 1.7, and state your conclusions.

2.18 In Exercise 1.4, the growth rate of employment was regressed on the growth rate of GDP for a sample of 25 OECD countries. Perform t tests on the slope coefficient and the intercept and state your conclusions.

2.19 Explain whether it would have been justifiable to perform one-sided tests instead of two-sided tests in Exercise 2.14. If you think that one-sided tests are justified, perform them and state whether the use of a one-sided test makes any difference.

2.20* Explain whether it would have been justifiable to perform one-sided tests instead of two-sided tests in Exercise 2.15. If you think that one-sided tests are justified, perform them and state whether the use of a one-sided test makes any difference.

2.21 Explain whether it would have been justifiable to perform one-sided tests instead of two-sided tests in Exercise 2.16. If you think that one-sided tests are justified, perform them and state whether the use of a one-sided test makes any difference.

2.22 Explain whether it would have been justifiable to perform one-sided tests instead of two-sided tests in Exercise 2.17. If you think that one-sided tests are justified, perform them and state whether the use of a one-sided test makes any difference.

2.23 In Exercise 1.9, the number of children in the family was regressed on the years of schooling of the mother for a sample of 540 NLSY respondents. Explain whether it would be justifiable to perform a one-sided test instead of a two-sided test on the slope coefficient.

2.24 Suppose that the true relationship between Y and X is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. In Section 1.4, it was shown that if $Y_i^* = \lambda_1 + \lambda_2 Y_i$, and Y^* is regressed on X, the slope coefficient $b_2^* = \lambda_2 b_2$. How will the t statistic for b_2^* be related to the t statistic for b_2? (See also Exercise 2.12.)

2.25* Suppose that the true relationship between Y and X is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. In Exercise 1.12, it was shown that if $X_i^* = \mu_1 + \mu_2 X_i$, and Y is regressed on X^*, the slope coefficient $b_2^* = b_2/\mu_2$. How will the t statistic for b_2^* be related to the t statistic for b_2? (See also Exercise 2.13.)

2.26 Calculate the 99 percent confidence interval for b_2 in the earnings function example in Table 2.2 ($b_2 = 2.455$, s.e.$(b_2) = 0.232$), and explain why it includes some values not included in the 95 percent confidence interval calculated in inequality (2.61).

2.27 Calculate the 95 percent confidence interval for the slope coefficient in the earnings function fitted with your *EAEF* data set in Exercise 1.7.

2.28* Calculate the 95 percent confidence interval for b_2 in the price inflation/wage inflation example:

$$\hat{p} = \underset{(0.05)}{-1.21} + \underset{(0.10)}{0.82} w.$$

What can you conclude from this calculation?

2.7 The F test of goodness of fit

Even if there is no relationship between Y and X, in any given sample of observations there may appear to be one, if only a faint one. Only by coincidence will the R^2 from a regression of Y on X be exactly equal to zero. So how do we know if the R^2 for the regression reflects a true relationship or if it has arisen as a matter of chance?

We could in principle adopt the following procedure. Suppose that the regression model is

$$Y_i = \beta_1 + \beta_2 X_i + u_i. \tag{2.62}$$

We take as our null hypothesis H_0: $\beta_2 = 0$, that there is no relationship between Y and X. We calculate the value that R^2 would exceed 5 percent of the time as a matter of chance if H_0 is true. We then take this figure as the critical level of R^2 for a 5 percent significance test. If it is exceeded, we reject the null hypothesis and conclude that $\beta_2 \neq 0$.

Such a test, like the t test on a coefficient, would not be error-free. Indeed, at the 5 percent significance level, one would risk making a Type I error (rejecting the null hypothesis when it is in fact true) 5 percent of the time. Of course you could cut down on this risk by using a higher significance level, for example, the 1 percent level. The critical level of R^2 would then be that which would be exceeded by chance only 1 percent of the time if H_0 is true, so it would be higher than the critical level for the 5 percent test.

How does one find the critical level of R^2 at either significance level? There is a problem. There is no such thing as a table of critical levels of R^2. The traditional procedure is to use an indirect approach and perform what is known as an F test based on analysis of variance.

We saw in Chapter 1 that the variations in the dependent variable may be decomposed into 'explained' and 'unexplained' components using (1.58):

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} e_i^2. \tag{2.63}$$

The left side is *TSS*, the total sum of squares of the values of the dependent variable about its sample mean. The first term on the right side is *ESS*, the explained sum of squares, and the second term is *RSS*, the unexplained, residual sum of squares:

$$TSS = ESS + RSS. \tag{2.64}$$

R^2 was then defined as the ratio of the explained sum of squares to the total sum of squares:

$$R^2 = \frac{ESS}{TSS} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n}\left(Y_i - \bar{Y}\right)^2}. \tag{2.65}$$

The F statistic for the goodness of fit of a regression is written as the explained sum of squares, per explanatory variable, divided by the residual sum of squares, per degree of freedom remaining:

$$F = \frac{ESS/(k-1)}{RSS/(n-k)}, \tag{2.66}$$

where k is the number of parameters in the regression equation (intercept and $k-1$ slope coefficients).

The F statistic is an increasing function of R^2. Dividing both the numerator and the denominator of the ratio by TSS, we have

$$F = \frac{(ESS/TSS)/(k-1)}{(RSS/TSS)/(n-k)} = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}. \tag{2.67}$$

An increase in R^2 leads to an increase in the numerator and a decrease in the denominator, and hence unambiguously to an increase in F. In the present context, $k = 2$, so (2.67) becomes

$$F = \frac{R^2}{(1-R^2)/(n-2)}. \tag{2.68}$$

F can be calculated using (2.66) and ESS and RSS or, equivalently, (2.68) and R^2. It is then compared with F_{crit}, the critical level of F, in the appropriate table. If F is greater than F_{crit}, you reject the null hypothesis H_0: $\beta_2 = 0$ and conclude that the 'explanation' of Y is better than is likely to have arisen by chance. In any serious regression application, F is automatically presented as part of the output.

Why do we take this indirect approach? Why not publish a table of critical levels of R^2? The answer is that the F table is useful for testing many forms of analysis of variance, of which R^2 is only one. Rather than have a specialized table for each application, it is more convenient (or, at least, it saves a lot of paper) to have just one general table, and make transformations such as (2.66) when necessary.

Table A.3 gives the critical levels of F at the 5 percent, 1 percent, and 0.1 percent significance levels. In each case, the critical level depends on the number of explanatory variables, $k-1$, which is read from along the top of the table, and the number of degrees of freedom, $n-k$, which is read off down the side. In the present context, we are concerned with simple regression analysis, k is 2, and we should use the first column of the table.

Example

Table 2.2 presents the output for the earnings function example regression. In the top left corner of the output, there is a column headed 'SS', meaning sums of squares. In that column one sees that ESS = 19,322 (Stata refers to ESS as

the 'model' sum of squares), $RSS = 92{,}689$, and $TSS = 112{,}010$. The number of observations is 540, $k = 2$, and so the number of degrees of freedom is 538. Hence,

$$F = \frac{ESS/(k-1)}{RSS/(n-k)} = \frac{19{,}322/1}{92{,}689/538} = \frac{19{,}322}{172.28} = 112.15. \tag{2.69}$$

The F statistic is printed in the top right corner of the output, together with its p value: the probability of obtaining an F statistic as high as that as a matter of chance if H_0: $\beta_2 = 0$ is true and there is no real relationship. (In this case, no chance, at least to four decimal places, so the F statistic is very strong evidence of a genuine relationship.) Looking at Table A.3, at the 0.1 percent significance level, the critical level of F for 1 and 500 degrees of freedom (first column, row 500) is 10.96. The critical value for 1 and 538 degrees of freedom must be lower, so we have no hesitation in rejecting the null hypothesis in this example.

We will verify that we obtain the same value of F using the expression for it as a function of R^2:

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)} = \frac{0.1725/1}{0.8275/538} = \frac{0.1725}{0.001538} = 112.16, \tag{2.70}$$

which is the same, apart from rounding error on the last digit.

Relationship between the F test and the t test on the slope coefficient in simple regression analysis

In the context of simple regression analysis (and *only* simple regression analysis), the F test and the two-sided t test on the slope coefficient have the same null hypothesis H_0: $\beta_2 = 0$ and the same alternative hypothesis H_1: $\beta_2 \neq 0$. This gives rise to the possibility that they might lead to different conclusions. Fortunately, they are in fact equivalent. The F statistic is equal to the square of the t statistic, and, at any given significance level, the critical value of F is equal to the square of the critical value of t. Starting with the definition of F in (2.66), and putting $k = 2$,

$$F = \frac{ESS}{RSS/(n-2)} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum_{i=1}^{n} e_i^2 \Big/ (n-2)} = \frac{\sum_{i=1}^{n}\left([b_1 + b_2 X_i] - [b_1 + b_2 \bar{X}]\right)^2}{s_u^2}$$

$$= \frac{1}{s_u^2}\sum_{i=1}^{n} b_2^2 \left(X_i - \bar{X}\right)^2 = \frac{b_2^2}{s_u^2 \Big/ \sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2} = \frac{b_2^2}{(\text{s.e.}(b_2))^2} = t^2. \tag{2.71}$$

The proof that, at any given significance level, the critical value of F is equal to the critical value of t for a two-sided t test is more complicated and will be omitted. When we come to multiple regression analysis, we will see that the F test and the t tests have different roles and different null hypotheses. However, in simple regression analysis, the fact that they are equivalent means that there is no point in performing both. Indeed, you would look ignorant if you did. Obviously, provided that it is justifiable, a one-sided t test would be preferable to either.

Key terms

- autocorrelation
- cross-sectional data
- F statistic
- F test of goodness of fit
- Gauss–Markov theorem
- homoscedastic disturbance term
- Monte Carlo experiment
- nonstochastic regressor
- p value
- panel data
- standard error of a regression coefficient
- stochastic regressor
- time series data

EXERCISES

2.29 In Exercise 1.4, in the regression of the rate of growth of employment on the rate of growth of real GDP using a sample of 25 OECD countries, *ESS* = 14.58 and *RSS* = 10.13. Calculate the corresponding F statistic and check that it is equal to 33.1, the value printed in the output. Also calculate the F statistic using $R^2 = 0.5900$ and verify that it is the same. Perform the F test at the 5 percent, 1 percent, and 0.1 percent significance levels. Is it necessary to report the results of the tests at all three levels?

2.30 Calculate the F statistic from *ESS* and *RSS* obtained in the earnings function fitted using your *EAEF* data set and check that it is equal to the value printed in the output. Check that the F statistic derived from R^2 is the same. Perform an appropriate F test.

2.31 Verify that the F statistic in the earnings function regression run by you using your *EAEF* data set is equal to the square of the t statistic for the slope coefficient, and that the critical value of F at the 1 percent significance level is equal to the square of the critical value of t.

2.32 In Exercise 1.16, both researchers obtained a t statistic of 10.59 for the slope coefficient in their regressions. Was this a coincidence?

2.33 Suppose that the true relationship between Y and X is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. Suppose that $Y_i^* = \lambda_1 + \lambda_2 Y_i$, and Y^* is regressed on X. How will the F statistic for this regression be related to the F statistic for the original regression? (See also Exercises 1.21, 2.12, and 2.24.)

2.34* Suppose that the true relationship between Y and X is $Y_i = \beta_1 + \beta_2 X_i + u_i$ and that the fitted model is $\hat{Y}_i = b_1 + b_2 X_i$. Suppose that $X_i^* = \mu_1 + \mu_2 X_i$, and Y is regressed on X^*. How will the F statistic for this regression be related to the F statistic for the original regression? (See also Exercises 1.22, 2.13, and 2.25.)

Appendix 2.1 The Gauss–Markov theorem

At the end of Section 2.5, it was asserted that the OLS estimators of the parameters are BLUE (best linear unbiased estimator), if the assumptions for Model A are satisfied. This appendix provides a demonstration for the OLS slope coefficient.

To see the linearity property, note that

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) &= \sum_{i=1}^{n}(X_i-\bar{X})Y_i-\sum_{i=1}^{n}(X_i-\bar{X})\bar{Y}\\ &= \sum_{i=1}^{n}(X_i-\bar{X})Y_i-\bar{Y}\sum_{i=1}^{n}(X_i-\bar{X})\\ &= \sum_{i=1}^{n}(X_i-\bar{X})Y_i-\bar{Y}\left\{\sum_{i=1}^{n}X_i-n\bar{X}\right\}\\ &= \sum_{i=1}^{n}(X_i-\bar{X})Y_i. \end{aligned} \tag{A2.1}$$

Then

$$b_2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{j=1}^{n}(X_j-\bar{X})^2}=\frac{\sum_{i=1}^{n}(X_i-\bar{X})Y_i}{\sum_{j=1}^{n}(X_j-\bar{X})^2}=\sum_{i=1}^{n}\frac{(X_i-\bar{X})}{\sum_{j=1}^{n}(X_j-\bar{X})^2}Y_i=\sum_{i=1}^{n}a_iY_i, \tag{A2.2}$$

where the a_i are defined as before.

We will demonstrate the efficiency property. Consider any other unbiased estimator

$$\tilde{b}_2=\sum_{i=1}^{n}g_iY_i \tag{A2.3}$$

that is a linear function of the Y_i. We will show that it has a larger variance unless $g_i = a_i$ for all i. For $\tilde{b}_2$ to be unbiased, we need $E(\tilde{b}_2)=\beta_2$.

$$\tilde{b}_2=\sum_{i=1}^{n}g_iY_i=\sum_{i=1}^{n}g_i(\beta_1+\beta_2X_i+u_i)=\sum_{i=1}^{n}\beta_1g_i+\sum_{i=1}^{n}\beta_2g_iX_i+\sum_{i=1}^{n}g_iu_i. \tag{A2.4}$$

Hence,

$$E(\tilde{b}_2)=\beta_1\sum_{i=1}^{n}g_i+\beta_2\sum_{i=1}^{n}g_iX_i+E\left\{\sum_{i=1}^{n}g_iu_i\right\}. \tag{A2.5}$$

The first two terms on the right side are nonstochastic and are therefore unaffected by taking expectations. Now

$$E\left\{\sum_{i=1}^{n}g_iu_i\right\}=\sum_{i=1}^{n}E(g_iu_i)=\sum_{i=1}^{n}g_iE(u_i)=0. \tag{A2.6}$$

The first step used the first expected value rule. Thus,

$$E(\tilde{b}_2) = \beta_1 \sum_{i=1}^{n} g_i + \beta_2 \sum_{i=1}^{n} g_i X_i. \tag{A2.7}$$

Hence for $E(\tilde{b}_2) = \beta_2$, the g_i must satisfy $\sum g_i = 0$ and $\sum g_i X_i = 1$. The variance of $\tilde{b}_2$ is given by

$$\sigma^2_{\tilde{b}_2} = E\left\{\left(\tilde{b}_2 - E(\tilde{b}_2)\right)^2\right\} = E\left\{\sum_{i=1}^{n} (g_i u_i)^2\right\} = \sigma_u^2 \sum_{i=1}^{n} g_i^2. \tag{A2.8}$$

The last step is exactly parallel to that in the proof that $E\left\{\sum (a_i u_i)^2\right\} = \sigma_u^2 \sum a_i^2$ in Box 2.3. Let

$$h_i = g_i - a_i. \tag{A2.9}$$

Writing $g_i = a_i + h_i$, the first condition for the unbiasedness of $\tilde{b}_2$ becomes

$$\sum_{i=1}^{n} g_i = \sum_{i=1}^{n} (a_i + h_i) = 0. \tag{A2.10}$$

Since $\sum a_i = 0$ (see Box 2.2), this implies $\sum h_i = 0$. The second condition for the unbiasedness of $\tilde{b}_2$ becomes

$$\sum_{i=1}^{n} g_i X_i = \sum_{i=1}^{n} (a_i + h_i) X_i = \sum_{i=1}^{n} a_i X_i + \sum_{i=1}^{n} h_i X_i = 1. \tag{A2.11}$$

Since $\sum a_i X_i = 1$ (see Box 2.2 again), this implies $\sum h_i X_i = 0$. The variance of $\tilde{b}_2$ becomes

$$\sigma^2_{\tilde{b}_2} = \sigma_u^2 \sum_{i=1}^{n} g_i^2 = \sigma_u^2 \sum_{i=1}^{n} (a_i + h_i)^2 = \sigma_u^2 \left\{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} h_i^2 + 2\sum_{i=1}^{n} a_i h_i\right\}. \tag{A2.12}$$

Now

$$\sum_{i=1}^{n} a_i h_i = \sum_{i=1}^{n} \frac{(X_i - \bar{X}) h_i}{\sum_{i=1}^{n} (X_i - \bar{X})^2} = \frac{1}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \left\{\sum_{i=1}^{n} h_i X_i - \bar{X} \sum_{i=1}^{n} h_i\right\}. \tag{A2.13}$$

This is zero because, as we have seen, the conditions for unbiasedness of $\tilde{b}_2$ require $\sum h_i = 0$ and $\sum h_i X_i = 0$. Hence,

$$\sigma^2_{\tilde{b}_2} = \sigma_u^2 \left\{\sum_{i=1}^{n} a_i^2 + \sum_{i=1}^{n} h_i^2\right\}. \tag{A2.14}$$

This must be greater than $\sigma_u^2 \sum a_i^2$, the variance of the OLS estimator b_2, unless $h_i = 0$ for all i, in which case $\tilde{b}_2$ is the same as b_2.