Chapter 4

Hypothesis Testing in Linear Regression Models

4.1 Introduction

As we saw in Chapter 3, the vector of OLS parameter estimates $\hat{\boldsymbol{\beta}}$ is a random vector. Since it would be an astonishing coincidence if $\hat{\boldsymbol{\beta}}$ were equal to the true parameter vector $\boldsymbol{\beta}_0$ in any finite sample, we must take the randomness of $\hat{\boldsymbol{\beta}}$ into account if we are to make inferences about $\boldsymbol{\beta}$. In classical econometrics, the two principal ways of doing this are performing **hypothesis tests** and constructing **confidence intervals** or, more generally, **confidence regions**. We will discuss the first of these topics in this chapter, as the title implies, and the second in the next chapter. Hypothesis testing is easier to understand than the construction of confidence intervals, and it plays a larger role in applied econometrics.

In the next section, we develop the fundamental ideas of hypothesis testing in the context of a very simple special case. Then, in Section 4.3, we review some of the properties of several distributions which are related to the normal distribution and are commonly encountered in the context of hypothesis testing. We will need this material for Section 4.4, in which we develop a number of results about hypothesis tests in the classical normal linear model. In Section 4.5, we relax some of the assumptions of that model and introduce large-sample tests. An alternative approach to testing under relatively weak assumptions is bootstrap testing, which we introduce in Section 4.6. Finally, in Section 4.7, we discuss what determines the ability of a test to reject a hypothesis that is false.

4.2 Basic Ideas

The very simplest sort of hypothesis test concerns the (population) mean from which a random sample has been drawn. To test such a hypothesis, we may assume that the data are generated by the regression model

$$y_t = \beta + u_t, \quad u_t \sim \text{IID}(0, \sigma^2), \tag{4.01}$$

where y_t is an observation on the dependent variable, β is the population mean, which is the only parameter of the regression function, and σ^2 is the variance of the error term u_t. The least squares estimator of β and its variance, for a sample of size n, are given by

$$\hat{\beta} = \frac{1}{n}\sum_{t=1}^{n} y_t \qquad \text{and} \qquad \operatorname{Var}(\hat{\beta}) = \frac{1}{n}\sigma^2. \tag{4.02}$$

These formulas can either be obtained from first principles or as special cases of the general results for OLS estimation. In this case, $\boldsymbol{X}$ is just an n-vector of 1s. Thus, for the model (4.01), the standard formulas $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{y}$ and $\operatorname{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ yield the two formulas given in (4.02).

Now suppose that we wish to test the hypothesis that $\beta = \beta_0$, where β_0 is some specified value of β.[1] The hypothesis that we are testing is called the **null hypothesis**. It is often given the label H_0 for short. In order to test H_0, we must calculate a **test statistic**, which is a random variable that has a known distribution when the null hypothesis is true and some other distribution when the null hypothesis is false. If the value of this test statistic is one that might frequently be encountered by chance under the null hypothesis, then the test provides no evidence against the null. On the other hand, if the value of the test statistic is an extreme one that would rarely be encountered by chance under the null, then the test does provide evidence against the null. If this evidence is sufficiently convincing, we may decide to **reject** the null hypothesis that $\beta = \beta_0$.

For the moment, we will restrict the model (4.01) by making two very strong assumptions. The first is that u_t is normally distributed, and the second is that σ is known. Under these assumptions, a test of the hypothesis that $\beta = \beta_0$ can be based on the test statistic

$$z = \frac{\hat{\beta} - \beta_0}{\big(\operatorname{Var}(\hat{\beta})\big)^{1/2}} = \frac{n^{1/2}}{\sigma}(\hat{\beta} - \beta_0). \tag{4.03}$$

It turns out that, under the null hypothesis, z must be distributed as $N(0,1)$. It must have mean 0 because $\hat{\beta}$ is an unbiased estimator of β, and $\beta = \beta_0$ under the null. It must have variance unity because, by (4.02),

$$\mathrm{E}(z^2) = \frac{n}{\sigma^2}\mathrm{E}\big((\hat{\beta} - \beta_0)^2\big) = \frac{n}{\sigma^2}\frac{\sigma^2}{n} = 1.$$

[1] It may be slightly confusing that a 0 subscript is used here to denote the value of a parameter under the null hypothesis as well as its true value. So long as it is assumed that the null hypothesis is true, however, there should be no possible confusion.

Finally, to see that z must be normally distributed, note that $\hat{\beta}$ is just the average of the y_t, each of which must be normally distributed if the corresponding u_t is; see Exercise 1.7. As we will see in the next section, this implies that z is also normally distributed. Thus z has the first property that we would like a test statistic to possess: It has a known distribution under the null hypothesis.

For every null hypothesis there is, at least implicitly, an **alternative hypothesis**, which is often given the label H_1. The alternative hypothesis is what we are testing the null against, in this case the model (4.01) with $\beta \neq \beta_0$. Just as important as the fact that z follows the $N(0, 1)$ distribution under the null is the fact that z does *not* follow this distribution under the alternative. Suppose that β takes on some other value, say β_1. Then it is clear that $\hat{\beta} = \beta_1 + \hat{\gamma}$, where $\hat{\gamma}$ has mean 0 and variance σ^2/n; recall equation (3.05). In fact, $\hat{\gamma}$ is normal under our assumption that the u_t are normal, just like $\hat{\beta}$, and so $\hat{\gamma} \sim N(0, \sigma^2/n)$. It follows that z is also normal (see Exercise 1.7 again), and we find from (4.03) that

$$z \sim N(\lambda, 1), \quad \text{with} \quad \lambda = \frac{n^{1/2}}{\sigma}(\beta_1 - \beta_0). \tag{4.04}$$

Therefore, provided n is sufficiently large, we would expect the mean of z to be large and positive if $\beta_1 > \beta_0$ and large and negative if $\beta_1 < \beta_0$. Thus we will reject the null hypothesis whenever z is sufficiently far from 0. Just how we can decide what "sufficiently far" means will be discussed shortly.

Since we want to test the null that $\beta = \beta_0$ against the alternative that $\beta \neq \beta_0$, we must perform a **two-tailed test** and reject the null whenever the absolute value of z is sufficiently large. If instead we were interested in testing the null hypothesis that $\beta \leq \beta_0$ against the alternative that $\beta > \beta_0$, we would perform a **one-tailed test** and reject the null whenever z was sufficiently large and positive. In general, tests of equality restrictions are two-tailed tests, and tests of inequality restrictions are one-tailed tests.

Since z is a random variable that can, in principle, take on any value on the real line, no value of z is absolutely incompatible with the null hypothesis, and so we can never be absolutely certain that the null hypothesis is false. One way to deal with this situation is to decide in advance on a **rejection rule**, according to which we will choose to reject the null hypothesis if and only if the value of z falls into the **rejection region** of the rule. For two-tailed tests, the appropriate rejection region is the union of two sets, one containing all values of z greater than some positive value, the other all values of z less than some negative value. For a one-tailed test, the rejection region would consist of just one set, containing either sufficiently positive or sufficiently negative values of z, according to the sign of the inequality we wish to test.

A test statistic combined with a rejection rule is sometimes called simply a **test**. If the test incorrectly leads us to reject a null hypothesis that is true, we are said to make a **Type I error**. The probability of making such an error is, by construction, the probability, *under the null hypothesis*, that z falls into the rejection region. This probability is sometimes called the **level of significance**, or just the **level**, of the test. A common notation for this is α. Like all probabilities, α is a number between 0 and 1, although, in practice, it is generally much closer to 0 than 1. Popular values of α include .05 and .01. If the observed value of z, say $\hat{z}$, lies in a rejection region associated with a probability under the null of α, we will reject the null hypothesis at level α, otherwise we will not reject the null hypothesis. In this way, we ensure that the probability of making a Type I error is precisely α.

In the previous paragraph, we implicitly assumed that the distribution of the test statistic under the null hypothesis is known exactly, so that we have what is called an **exact test**. In econometrics, however, the distribution of a test statistic is often known only approximately. In this case, we need to draw a distinction between the **nominal level** of the test, that is, the probability of making a Type I error according to whatever approximate distribution we are using to determine the rejection region, and the actual **rejection probability**, which may differ greatly from the nominal level. The rejection probability is generally unknowable in practice, because it typically depends on unknown features of the DGP.[2]

The probability that a test will reject the null is called the **power** of the test. If the data are generated by a DGP that satisfies the null hypothesis, the power of an exact test is equal to its level. In general, power will depend on precisely how the data were generated and on the sample size. We can see from (4.04) that the distribution of z is entirely determined by the value of λ, with $\lambda = 0$ under the null, and that the value of λ depends on the parameters of the DGP. In this example, λ is proportional to $\beta_1 - \beta_0$ and to the square root of the sample size, and it is inversely proportional to σ.

Values of λ different from 0 move the probability mass of the $N(\lambda, 1)$ distribution away from the center of the $N(0, 1)$ distribution and into its tails. This can be seen in Figure 4.1, which graphs the $N(0, 1)$ density and the $N(\lambda, 1)$ density for $\lambda = 2$. The second density places much more probability than the first on values of z greater than 2. Thus, if the rejection region for our test was the interval from 2 to $+\infty$, there would be a much higher probability in that region for $\lambda = 2$ than for $\lambda = 0$. Therefore, we would reject the null hypothesis more often when the null hypothesis is false, with $\lambda = 2$, than when it is true, with $\lambda = 0$.

[2] Another term that often arises in the discussion of hypothesis testing is the **size** of a test. Technically, this is the supremum of the rejection probability over all DGPs that satisfy the null hypothesis. For an exact test, the size equals the level. For an approximate test, the size is typically difficult or impossible to calculate. It is often, but by no means always, greater than the nominal level of the test.

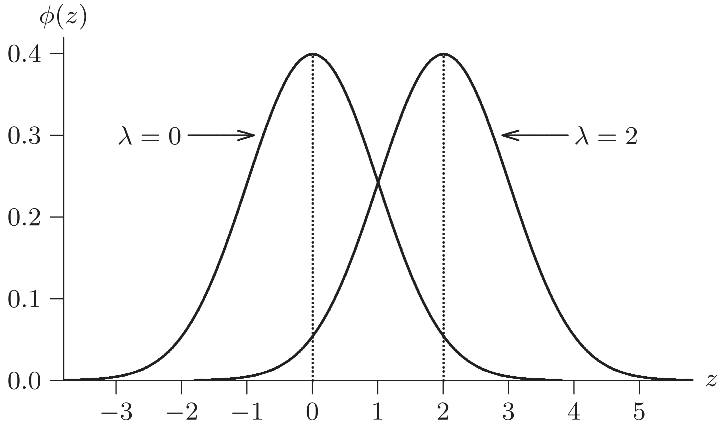

Figure 4.1 The normal distribution centered and uncentered

Mistakenly failing to reject a false null hypothesis is called making a **Type II error**. The probability of making such a mistake is equal to 1 minus the power of the test. It is not hard to see that, quite generally, the probability of rejecting the null with a two-tailed test based on z increases with the absolute value of λ. Consequently, the power of such a test will increase as $\beta_1 - \beta_0$ increases, as σ decreases, and as the sample size increases. We will discuss what determines the power of a test in more detail in Section 4.7.

In order to construct the rejection region for a test at level α, the first step is to calculate the **critical value** associated with the level α. For a two-tailed test based on any test statistic that is distributed as $N(0, 1)$, including the statistic z defined in (4.04), the critical value c_α is defined implicitly by

$$\Phi(c_\alpha) = 1 - \alpha/2. \tag{4.05}$$

Recall that Φ denotes the CDF of the standard normal distribution. In terms of the inverse function Φ^{-1}, c_α can be defined explicitly by the formula

$$c_\alpha = \Phi^{-1}(1 - \alpha/2). \tag{4.06}$$

According to (4.05), the probability that $z > c_\alpha$ is $1 - (1 - \alpha/2) = \alpha/2$, and the probability that $z < -c_\alpha$ is also $\alpha/2$, by symmetry. Thus the probability that $|z| > c_\alpha$ is α, and so an appropriate rejection region for a test at level α is the set defined by $|z| > c_\alpha$. Clearly, c_α increases as α approaches 0. As an example, when $\alpha = .05$, we see from (4.06) that the critical value for a two-tailed test is $\Phi^{-1}(.975) = 1.96$. We would reject the null at the .05 level whenever the observed absolute value of the test statistic exceeds 1.96.

P Values

As we have defined it, the result of a test is yes or no: Reject or do not reject. A more sophisticated approach to deciding whether or not to reject the null hypothesis is to calculate the ***P* value**, or **marginal significance level**, associated with the observed test statistic $\hat{z}$. The P value for $\hat{z}$ is defined as the greatest level for which a test based on $\hat{z}$ fails to reject the null. Equivalently, at least if the statistic z has a continuous distribution, it is the smallest level for which the test rejects. Thus, the test rejects for all levels greater than the P value, and it fails to reject for all levels smaller than the P value. Therefore, if the P value associated with $\hat{z}$ is denoted $p(\hat{z})$, we must be prepared to accept a probability $p(\hat{z})$ of Type I error if we choose to reject the null.

For a two-tailed test, in the special case we have been discussing,

$$p(\hat{z}) = 2\big(1 - \Phi(|\hat{z}|)\big). \tag{4.07}$$

To see this, note that the test based on $\hat{z}$ rejects at level α if and only if $|\hat{z}| > c_\alpha$. This inequality is equivalent to $\Phi(|\hat{z}|) > \Phi(c_\alpha)$, because $\Phi(\cdot)$ is a strictly increasing function. Further, $\Phi(c_\alpha) = 1 - \alpha/2$, by (4.05). The smallest value of α for which the inequality holds is thus obtained by solving the equation

$$\Phi(|\hat{z}|) = 1 - \alpha/2,$$

and the solution is easily seen to be the right-hand side of (4.07).

One advantage of using P values is that they preserve all the information conveyed by a test statistic, while presenting it in a way that is directly interpretable. For example, the test statistics 2.02 and 5.77 would both lead us to reject the null at the .05 level using a two-tailed test. The second of these obviously provides more evidence against the null than does the first, but it is only after they are converted to P values that the magnitude of the difference becomes apparent. The P value for the first test statistic is .0434, while the P value for the second is 7.93×10^{-9}, an extremely small number.

Computing a P value transforms z from a random variable with the $N(0, 1)$ distribution into a new random variable $p(z)$ with the uniform $U(0, 1)$ distribution. In Exercise 4.1, readers are invited to prove this fact. It is quite possible to think of $p(z)$ as a test statistic, of which the observed realization is $p(\hat{z})$. A test at level α rejects whenever $p(\hat{z}) < \alpha$. Note that the sign of this inequality is the opposite of that in the condition $|\hat{z}| > c_\alpha$. Generally, one rejects for *large* values of test statistics, but for *small* P values.

Figure 4.2 illustrates how the test statistic $\hat{z}$ is related to its P value $p(\hat{z})$. Suppose that the value of the test statistic is 1.51. Then

$$\Pr(z > 1.51) = \Pr(z < -1.51) = .0655. \tag{4.08}$$

This implies, by equation (4.07), that the P value for a two-tailed test based on $\hat{z}$ is .1310. The top panel of the figure illustrates (4.08) in terms of the PDF of the standard normal distribution, and the bottom panel illustrates it in terms of the CDF. To avoid clutter, no critical values are shown on the

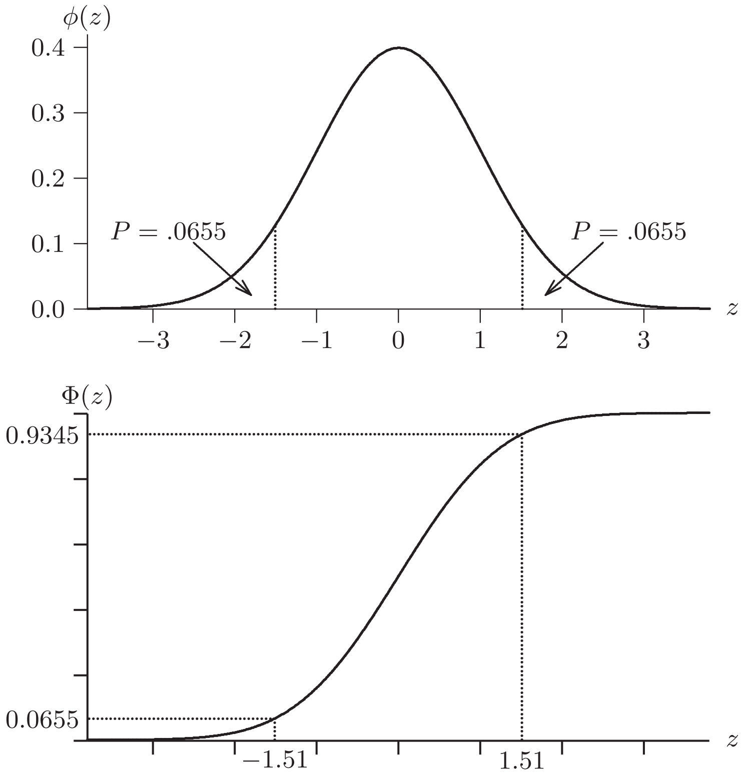

Figure 4.2 P values for a two-tailed test

figure, but it is clear that a test based on $\hat{z}$ will not reject at any level smaller than .131. From the figure, it is also easy to see that the P value for a one-tailed test of the hypothesis that $\beta \leq \beta_0$ is .0655. This is just $\Pr(z > 1.51)$. Similarly, the P value for a one-tailed test of the hypothesis that $\beta \geq \beta_0$ is $\Pr(z < 1.51) = .9345$.

In this section, we have introduced the basic ideas of hypothesis testing. However, we had to make two very restrictive assumptions. The first is that the error terms are normally distributed, and the second, which is grossly unrealistic, is that the variance of the error terms is known. In addition, we limited our attention to a single restriction on a single parameter. In Section 4.4, we will discuss the more general case of linear restrictions on the parameters of a linear regression model with unknown error variance. Before we can do so, however, we need to review the properties of the normal distribution and of several distributions that are closely related to it.

4.3 Some Common Distributions

Most test statistics in econometrics follow one of four well-known distributions, at least approximately. These are the standard normal distribution, the chi-squared (or χ^2) distribution, the Student's t distribution, and the F distribution. The most basic of these is the normal distribution, since the other three distributions can be derived from it. In this section, we discuss the standard, or **central**, versions of these distributions. Later, in Section 4.7, we will have occasion to introduce **noncentral** versions of all these distributions.

The Normal Distribution

The **normal distribution**, which is sometimes called the **Gaussian distribution** in honor of the celebrated German mathematician and astronomer Carl Friedrich Gauss (1777–1855), even though he did not invent it, is certainly the most famous distribution in statistics. As we saw in Section 1.2, there is a whole family of normal distributions, all based on the **standard normal distribution**, so called because it has mean 0 and variance 1. The PDF of the standard normal distribution, which is usually denoted by $\phi(\cdot)$, was defined in (1.06). No elementary closed-form expression exists for its CDF, which is usually denoted by $\Phi(\cdot)$. Although there is no closed form, it is perfectly easy to evaluate Φ numerically, and virtually every program for doing econometrics and statistics can do this. Thus it is straightforward to compute the P value for any test statistic that is distributed as standard normal. The graphs of the functions ϕ and Φ were first shown in Figure 1.1 and have just reappeared in Figure 4.2. In both tails, the PDF rapidly approaches 0. Thus, although a standard normal r.v. can, in principle, take on any value on the real line, values greater than about 4 in absolute value occur extremely rarely.

In Exercise 1.7, readers were asked to show that the full normal family can be generated by varying exactly two parameters, the mean and the variance. A random variable X that is normally distributed with mean μ and variance σ^2 can be generated by the formula

$$X = \mu + \sigma Z, \tag{4.09}$$

where Z is standard normal. The distribution of X, that is, the normal distribution with mean μ and variance σ^2, is denoted $N(\mu, \sigma^2)$. Thus the standard normal distribution is the $N(0, 1)$ distribution. As readers were asked to show in Exercise 1.8, the PDF of the $N(\mu, \sigma^2)$ distribution, evaluated at x, is

$$\frac{1}{\sigma}\phi\Big(\frac{x-\mu}{\sigma}\Big) = \frac{1}{\sigma\sqrt{2\pi}}\exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \tag{4.10}$$

In expression (4.10), as in Section 1.2, we have distinguished between the random variable X and a value x that it can take on. However, for the following discussion, this distinction is more confusing than illuminating. For the rest of this section, we therefore use lower-case letters to denote both random variables and the arguments of their PDFs or CDFs, depending on context. No confusion should result. Adopting this convention, then, we see that, if x is distributed as $N(\mu, \sigma^2)$, we can invert (4.09) and obtain $z = (x - \mu)/\sigma$, where z is standard normal. Note also that z is the argument of ϕ in the expression (4.10) of the PDF of x. In general, the PDF of a normal variable x with mean μ and variance σ^2 is $1/\sigma$ times ϕ evaluated at the corresponding standard normal variable, which is $z = (x - \mu)/\sigma$.

Although the normal distribution is fully characterized by its first two moments, the higher moments are also important. Because the distribution is symmetric around its mean, the third central moment, which measures the **skewness** of the distribution, is always zero.[3] This is true for all of the odd central moments. The fourth moment of a symmetric distribution provides a way to measure its **kurtosis**, which essentially means how thick the tails are. In the case of the $N(\mu, \sigma^2)$ distribution, the fourth central moment is $3\sigma^4$; see Exercise 4.2.

Linear Combinations of Normal Variables

An important property of the normal distribution, used in our discussion in the preceding section, is that any linear combination of independent normally distributed random variables is itself normally distributed. To see this, it is enough to show it for independent standard normal variables, because, by (4.09), all normal variables can be generated as linear combinations of standard normal ones plus constants. We will tackle the proof in several steps, each of which is important in its own right.

To begin with, let z_1 and z_2 be standard normal and mutually independent, and consider $w \equiv b_1 z_1 + b_2 z_2$. For the moment, we suppose that $b_1^2 + b_2^2 = 1$, although we will remove this restriction shortly. If we reason conditionally on z_1, then we find that

$$\mathrm{E}(w \,|\, z_1) = b_1 z_1 + b_2 \mathrm{E}(z_2 \,|\, z_1) = b_1 z_1 + b_2 \mathrm{E}(z_2) = b_1 z_1.$$

The first equality follows because $b_1 z_1$ is a deterministic function of the conditioning variable z_1, and so can be taken outside the conditional expectation. The second, in which the conditional expectation of z_2 is replaced by its unconditional expectation, follows because of the independence of z_1 and z_2 (see Exercise 1.9). Finally, $\mathrm{E}(z_2) = 0$ because z_2 is $N(0, 1)$.

The conditional variance of w is given by

$$\mathrm{E}\Big(\big(w - \mathrm{E}(w \,|\, z_1)\big)^2 \;\big|\; z_1\Big) = \mathrm{E}\big((b_2 z_2)^2 \,|\, z_1\big) = \mathrm{E}\big((b_2 z_2)^2\big) = b_2^2,$$

[3] A distribution is said to be skewed to the right if the third central moment is positive, and to the left if the third central moment is negative.

where the last equality again follows because $z_2 \sim N(0,1)$. Conditionally on z_1, w is the sum of the constant $b_1 z_1$ and b_2 times a standard normal variable z_2, and so the *conditional* distribution of w is normal. Given the conditional mean and variance we have just computed, we see that the conditional distribution must be $N(b_1 z_1, b_2^2)$. The PDF of this distribution is the density of w conditional on z_1, and, by (4.10), it is

$$f(w \,|\, z_1) = \frac{1}{b_2}\,\phi\Big(\frac{w - b_1 z_1}{b_2}\Big). \tag{4.11}$$

In accord with what we noted above, the argument of ϕ here is equal to z_2, which is the standard normal variable corresponding to w conditional on z_1.

The next step is to find the joint density of w and z_1. By (1.15), the density of w conditional on z_1 is the ratio of the joint density of w and z_1 to the marginal density of z_1. This marginal density is just $\phi(z_1)$, since $z_1 \sim N(0,1)$, and so we see that the joint density is

$$f(w, z_1) = f(z_1)\, f(w \,|\, z_1) = \phi(z_1)\,\frac{1}{b_2}\,\phi\Big(\frac{w - b_1 z_1}{b_2}\Big). \tag{4.12}$$

If we use (1.06) to get an explicit expression for this joint density, then we obtain

$$\begin{aligned} &\frac{1}{2\pi b_2}\exp\Big(-\frac{1}{2b_2^2}\big(b_2^2 z_1^2 + w^2 - 2b_1 z_1 w + b_1^2 z_1^2\big)\Big) \\ &= \frac{1}{2\pi b_2}\exp\Big(-\frac{1}{2b_2^2}\big(z_1^2 - 2b_1 z_1 w + w^2\big)\Big), \end{aligned} \tag{4.13}$$

since we assumed that $b_1^2 + b_2^2 = 1$. The right-hand side of (4.13) is symmetric with respect to z_1 and w. Thus the joint density can also be expressed as in (4.12), but with z_1 and w interchanged, as follows:

$$f(w, z_1) = \frac{1}{b_2}\,\phi(w)\,\phi\Big(\frac{z_1 - b_1 w}{b_2}\Big). \tag{4.14}$$

We are now ready to compute the unconditional, or marginal, density of w. To do so, we integrate the joint density (4.14) with respect to z_1; see (1.12). Note that z_1 occurs only in the last factor on the right-hand side of (4.14). Further, the expression $(1/b_2)\phi\big((z_1 - b_1 w)/b_2\big)$, like expression (4.11), is a probability density, and so it integrates to 1. Thus we conclude that the marginal density of w is $f(w) = \phi(w)$, and so it follows that w is standard normal, unconditionally, as we wished to show.

It is now simple to extend this argument to the case for which $b_1^2 + b_2^2 \neq 1$. We define $r^2 = b_1^2 + b_2^2$, and consider w/r. The argument above shows that w/r is standard normal, and so $w \sim N(0, r^2)$. It is equally simple to extend the result to a linear combination of any number of mutually independent standard normal variables. If we now let w be defined as $b_1 z_1 + b_2 z_2 + b_3 z_3$, where z_1, z_2, and z_3 are mutually independent standard normal variables, then $b_1z_1+b_2z_2$ is normal by the result for two variables, and it is independent of z_3. Thus, by applying the result for two variables again, this time to $b_1z_1 + b_2z_2$ and z_3, we see that w is normal. This reasoning can obviously be extended by induction to a linear combination of any number of independent standard normal variables. Finally, if we consider a linear combination of independent normal variables with nonzero means, the mean of the resulting variable is just the same linear combination of the means of the individual variables.

The Multivariate Normal Distribution

The results of the previous subsection can be extended to linear combinations of normal random variables that are not necessarily independent. In order to do so, we introduce the **multivariate normal distribution**. As the name suggests, this is a family of distributions for random *vectors*, with the scalar normal distributions being special cases of it. The pair of random variables z_1 and w considered above follow the **bivariate normal distribution**, another special case of the multivariate normal distribution. As we will see in a moment, all these distributions, like the scalar normal distribution, are completely characterized by their first two moments.

In order to construct the multivariate normal distribution, we begin with a set of m mutually independent standard normal variables, z_i, $i = 1, \ldots, m$, which we can assemble into a random m–vector $\boldsymbol{z}$. Then any m–vector $\boldsymbol{x}$ of linearly independent linear combinations of the components of $\boldsymbol{z}$ follows a multivariate normal distribution. Such a vector $\boldsymbol{x}$ can always be written as $\boldsymbol{Az}$, for some nonsingular $m \times m$ matrix $\boldsymbol{A}$. As we will see in a moment, the matrix $\boldsymbol{A}$ can always be chosen to be lower-triangular.

We denote the components of $\boldsymbol{x}$ as x_i, $i = 1, \ldots, m$. From what we have seen above, it is clear that each x_i is normally distributed, with (unconditional) mean zero. Therefore, from results proved in Section 3.4, it follows that the covariance matrix of $\boldsymbol{x}$ is

$$\operatorname{Var}(\boldsymbol{x}) = \mathrm{E}(\boldsymbol{x}\boldsymbol{x}^\top) = \boldsymbol{A}\mathrm{E}(\boldsymbol{z}\boldsymbol{z}^\top)\boldsymbol{A}^\top = \boldsymbol{A}\mathbf{I}\boldsymbol{A}^\top = \boldsymbol{A}\boldsymbol{A}^\top.$$

Here we have used the fact that the covariance matrix of $\boldsymbol{z}$ is the identity matrix $\mathbf{I}$. This is true because the variance of each component of $\boldsymbol{z}$ is 1, and, since the z_i are mutually independent, all the covariances are 0; see Exercise 1.11.

Let us denote the covariance matrix of $\boldsymbol{x}$ by $\boldsymbol{\Omega}$. Recall that, according to a result mentioned in Section 3.4 in connection with Crout's algorithm, for any positive definite matrix $\boldsymbol{\Omega}$, we can always find a lower-triangular $\boldsymbol{A}$ such that $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Omega}$. Thus the matrix $\boldsymbol{A}$ may always be chosen to be lower-triangular. The distribution of $\boldsymbol{x}$ is multivariate normal with mean vector $\mathbf{0}$ and covariance matrix $\boldsymbol{\Omega}$. We write this as $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{\Omega})$. If we add an m–vector $\boldsymbol{\mu}$ of constants to $\boldsymbol{x}$, the resulting vector must follow the $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$ distribution.

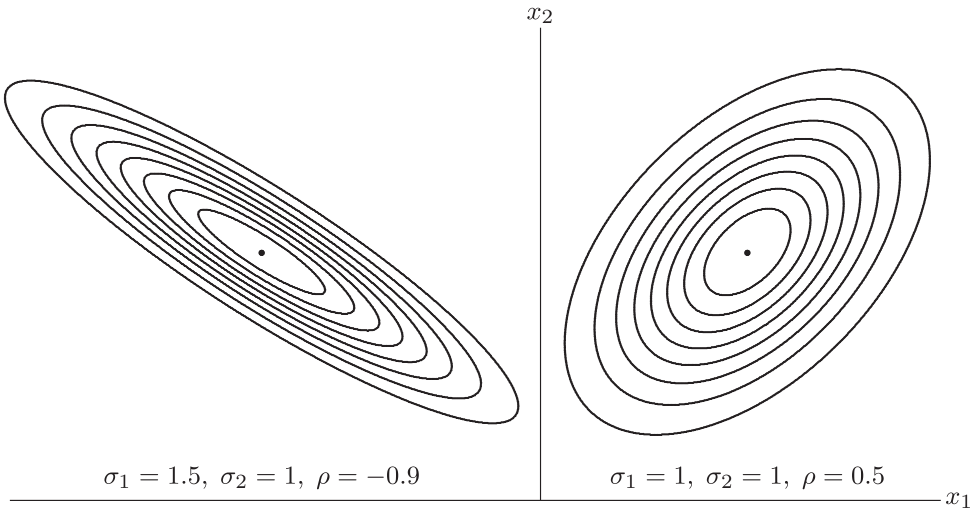

Figure 4.3 Contours of two bivariate normal densities

It is clear from this argument that any linear combination of random variables that are jointly multivariate normal is itself normally distributed. Thus, if $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, any scalar $\boldsymbol{a}^\top\boldsymbol{x}$, where $\boldsymbol{a}$ is an m–vector of fixed coefficients, is normally distributed with mean $\boldsymbol{a}^\top\boldsymbol{\mu}$ and variance $\boldsymbol{a}^\top\boldsymbol{\Omega}\boldsymbol{a}$.

We saw a moment ago that $\boldsymbol{z} \sim N(\mathbf{0}, \mathbf{I})$ whenever the components of the vector $\boldsymbol{z}$ are independent. Another crucial property of the multivariate normal distribution is that the converse of this result is also true: If $\boldsymbol{x}$ is any multivariate normal vector with zero covariances, the components of $\boldsymbol{x}$ are mutually independent. This is a very special property of the multivariate normal distribution, and readers are asked to prove it, for the bivariate case, in Exercise 4.5. In general, a zero covariance between two random variables does *not* imply that they are independent.

It is important to note that the results of the last two paragraphs do not hold unless the vector $\boldsymbol{x}$ is multivariate normal, that is, constructed as a set of linear combinations of *independent* normal variables. In most cases, when we have to deal with linear combinations of two or more normal random variables, it is reasonable to assume that they are jointly distributed as multivariate normal. However, as Exercise 1.12 illustrates, it is possible for two or more random variables not to be multivariate normal even though each one individually follows a normal distribution.

Figure 4.3 illustrates the bivariate normal distribution, of which the PDF is given in Exercise 4.5 in terms of the variances σ_1^2 and σ_2^2 of the two variables, and their correlation ρ. Contours of the density are plotted, on the right for $\sigma_1 = \sigma_2 = 1.0$ and $\rho = 0.5$, on the left for $\sigma_1 = 1.5$, $\sigma_2 = 1.0$, and $\rho = -0.9$. The contours of the bivariate normal density can be seen to be elliptical. The ellipses slope upward when $\rho > 0$ and downward when $\rho < 0$. They do so more steeply the larger is the ratio σ_2/σ_1. The closer $|\rho|$ is to 1, for given values of σ_1 and σ_2, the more elongated are the elliptical contours.

The Chi-Squared Distribution

Suppose, as in our discussion of the multivariate normal distribution, that the random vector $\boldsymbol{z}$ is such that its components $z_1, \ldots, z_m$ are mutually independent standard normal random variables. An easy way to express this is to write $\boldsymbol{z} \sim N(\mathbf{0}, \mathbf{I})$. Then the random variable

$$y \equiv \|\boldsymbol{z}\|^2 = \boldsymbol{z}^\top\boldsymbol{z} = \sum_{i=1}^{m} z_i^2 \tag{4.15}$$

is said to follow the **chi-squared distribution** with m **degrees of freedom**. A compact way of writing this is: $y \sim \chi^2(m)$. From (4.15), it is clear that m must be a positive integer. In the case of a test statistic, it will turn out to be equal to the number of restrictions being tested.

The mean and variance of the $\chi^2(m)$ distribution can easily be obtained from the definition (4.15). The mean is

$$\mathrm{E}(y) = \sum_{i=1}^{m} \mathrm{E}(z_i^2) = \sum_{i=1}^{m} 1 = m. \tag{4.16}$$

Since the z_i are independent, the variance of the sum of the z_i^2 is just the sum of the (identical) variances:

$$\begin{aligned}\mathrm{Var}(y) &= \sum_{i=1}^{m} \mathrm{Var}(z_i^2) = m\mathrm{E}\big((z_i^2-1)^2\big) \\ &= m\mathrm{E}(z_i^4 - 2z_i^2 + 1) = m(3-2+1) = 2m.\end{aligned} \tag{4.17}$$

The third equality here uses the fact that $\mathrm{E}(z_i^4) = 3$; see Exercise 4.2.

Another important property of the chi-squared distribution, which follows immediately from (4.15), is that, if $y_1 \sim \chi^2(m_1)$ and $y_2 \sim \chi^2(m_2)$, and y_1 and y_2 are independent, then $y_1 + y_2 \sim \chi^2(m_1 + m_2)$. To see this, rewrite (4.15) as

$$y = y_1 + y_2 = \sum_{i=1}^{m_1} z_i^2 + \sum_{i=m_1+1}^{m_1+m_2} z_i^2 = \sum_{i=1}^{m_1+m_2} z_i^2,$$

from which the result follows.

Figure 4.4 shows the PDF of the $\chi^2(m)$ distribution for $m = 1$, $m = 3$, $m = 5$, and $m = 7$. The changes in the location and height of the density function as m increases are what we should expect from the results (4.16) and (4.17) about its mean and variance. In addition, the PDF, which is extremely

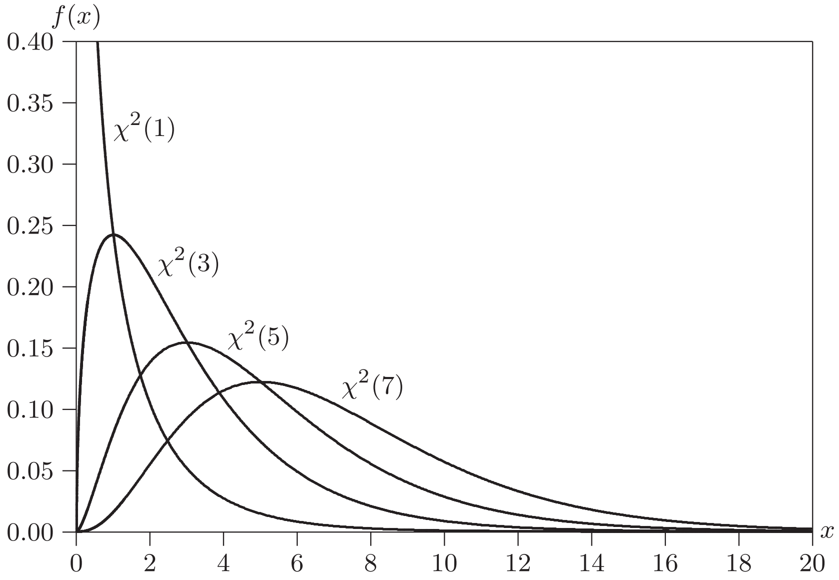

Figure 4.4 Various chi-squared PDFs

skewed to the right for $m = 1$, becomes less skewed as m increases. In fact, as we will see in Section 4.5, the $\chi^2(m)$ distribution approaches the $N(m, 2m)$ distribution as m becomes large.

In Section 3.4, we introduced quadratic forms. As we will see, many test statistics can be written as quadratic forms in normal vectors, or as functions of such quadratic forms. The following theorem states two results about quadratic forms in normal vectors that will prove to be extremely useful.

Theorem 4.1.

1. If the m–vector $\boldsymbol{x}$ is distributed as $N(\mathbf{0}, \boldsymbol{\Omega})$, then the quadratic form $\boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x}$ is distributed as $\chi^2(m)$;
2. If $\boldsymbol{P}$ is a projection matrix with rank r and $\boldsymbol{z}$ is an n–vector that is distributed as $N(\mathbf{0}, \mathbf{I})$, then the quadratic form $\boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{z}$ is distributed as $\chi^2(r)$.

Proof: Since the vector $\boldsymbol{x}$ is multivariate normal with mean vector $\mathbf{0}$, so is the vector $\boldsymbol{A}^{-1}\boldsymbol{x}$, where, as before, $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{\Omega}$. Moreover, the covariance matrix of $\boldsymbol{A}^{-1}\boldsymbol{x}$ is

$$\mathrm{E}\big(\boldsymbol{A}^{-1}\boldsymbol{x}\boldsymbol{x}^\top(\boldsymbol{A}^\top)^{-1}\big) = \boldsymbol{A}^{-1}\boldsymbol{\Omega}(\boldsymbol{A}^\top)^{-1} = \boldsymbol{A}^{-1}\boldsymbol{A}\boldsymbol{A}^\top(\boldsymbol{A}^\top)^{-1} = \mathbf{I}_m.$$

Thus we have shown that the vector $\boldsymbol{z} \equiv \boldsymbol{A}^{-1}\boldsymbol{x}$ is distributed as $N(\mathbf{0}, \mathbf{I})$.

The quadratic form $\boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x}$ is equal to $\boldsymbol{x}^\top(\boldsymbol{A}^\top)^{-1}\boldsymbol{A}^{-1}\boldsymbol{x} = \boldsymbol{z}^\top\boldsymbol{z}$. As we have just shown, this is equal to the sum of m independent, squared, standard normal random variables. From the definition of the chi-squared distribution, we know that such a sum is distributed as $\chi^2(m)$. This proves the first part of the theorem.

Since $\boldsymbol{P}$ is a projection matrix, it must project orthogonally on to some subspace of E^n. Suppose, then, that $\boldsymbol{P}$ projects on to the span of the columns of an $n \times r$ matrix $\boldsymbol{Z}$. This allows us to write

$$\boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{z} = \boldsymbol{z}^\top\boldsymbol{Z}(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top\boldsymbol{z}.$$

The r-vector $\boldsymbol{x} \equiv \boldsymbol{Z}^\top\boldsymbol{z}$ evidently follows the $N(\boldsymbol{0}, \boldsymbol{Z}^\top\boldsymbol{Z})$ distribution. Therefore, $\boldsymbol{z}^\top\boldsymbol{P}\boldsymbol{z}$ is seen to be a quadratic form in the multivariate normal r-vector $\boldsymbol{x}$ and $(\boldsymbol{Z}^\top\boldsymbol{Z})^{-1}$, which is the inverse of its covariance matrix. That this quadratic form is distributed as $\chi^2(r)$ follows immediately from the the first part of the theorem. ■

The Student's *t* Distribution

If $z \sim N(0,1)$ and $y \sim \chi^2(m)$, and z and y are independent, then the random variable

$$t \equiv \frac{z}{(y/m)^{1/2}} \tag{4.18}$$

is said to follow the **Student's *t* distribution** with m degrees of freedom. A compact way of writing this is: $t \sim t(m)$. The Student's t distribution looks very much like the standard normal distribution, since both are bell-shaped and symmetric around 0.

The moments of the t distribution depend on m, and only the first $m-1$ moments exist. Thus the $t(1)$ distribution, which is also called the **Cauchy distribution**, has no moments at all, and the $t(2)$ distribution has no variance. From (4.18), we see that, for the Cauchy distribution, the denominator of t is just the absolute value of a standard normal random variable. Whenever this denominator happens to be close to zero, the ratio is likely to be a very big number, even if the numerator is not particularly large. Thus the Cauchy distribution has very thick tails. As m increases, the chance that the denominator of (4.18) is close to zero diminishes (see Figure 4.4), and so the tails become thinner.

In general, if t is distributed as $t(m)$ with $m > 2$, then $\text{Var}(t) = m/(m-2)$. Thus, as $m \to \infty$, the variance tends to 1, the variance of the standard normal distribution. In fact, the entire $t(m)$ distribution tends to the standard normal distribution as $m \to \infty$. By (4.15), the chi-squared variable y can be expressed as $\sum_{i=1}^m z_i^2$, where the z_i are independent standard normal variables. Therefore, by a law of large numbers, such as (3.16), y/m, which is the average of the z_i^2, tends to its expectation as $m \to \infty$. By (4.16), this expectation is just $m/m = 1$. It follows that the denominator of (4.18), $(y/m)^{1/2}$, also tends to 1, and hence that $t \to z \sim N(0,1)$ as $m \to \infty$.

Figure 4.5 shows the PDFs of the standard normal, $t(1)$, $t(2)$, and $t(5)$ distributions. In order to make the differences among the various densities in the

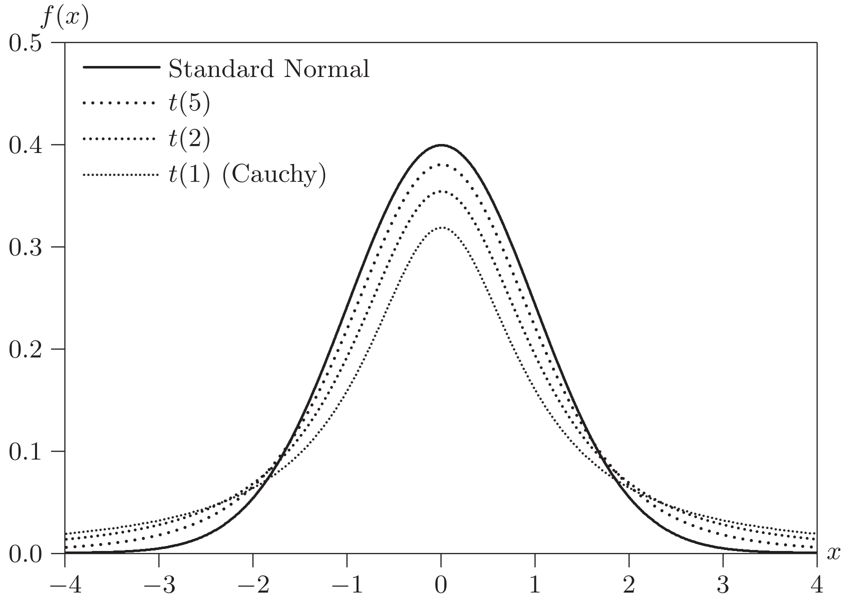

Figure 4.5 PDFs of the Student's t distribution

figure apparent, all the values of m are chosen to be very small. However, it is clear from the figure that, for larger values of m, the PDF of $t(m)$ will be very similar to the PDF of the standard normal distribution.

The *F* Distribution

If y_1 and y_2 are independent random variables distributed as $\chi^2(m_1)$ and $\chi^2(m_2)$, respectively, then the random variable

$$F \equiv \frac{y_1/m_1}{y_2/m_2} \tag{4.19}$$

is said to follow the ***F* distribution** with m_1 and m_2 degrees of freedom. A compact way of writing this is: $F \sim F(m_1, m_2)$. The notation F is used in honor of the well-known statistician R. A. Fisher. The $F(m_1, m_2)$ distribution looks a lot like a rescaled version of the $\chi^2(m_1)$ distribution. As for the t distribution, the denominator of (4.19) tends to unity as $m_2 \to \infty$, and so $m_1 F \to y_1 \sim \chi^2(m_1)$ as $m_2 \to \infty$. Therefore, for large values of m_2, a random variable that is distributed as $F(m_1, m_2)$ will behave very much like $1/m_1$ times a random variable that is distributed as $\chi^2(m_1)$.

The F distribution is very closely related to the Student's t distribution. It is evident from (4.19) and (4.18) that the square of a random variable which is distributed as $t(m_2)$ will be distributed as $F(1, m_2)$. In the next section, we will see how these two distributions arise in the context of hypothesis testing in linear regression models.

4.4 Exact Tests in the Classical Normal Linear Model

In the example of Section 4.2, we were able to obtain a test statistic z that was distributed as $N(0, 1)$. Tests based on this statistic are exact. Unfortunately, it is possible to perform exact tests only in certain special cases. One very important special case of this type arises when we test linear restrictions on the parameters of the classical normal linear model, which was introduced in Section 3.1. This model may be written as

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{4.20}$$

where $\boldsymbol{X}$ is an $n \times k$ matrix of regressors, so that there are n observations and k regressors, and it is assumed that the error vector $\boldsymbol{u}$ is statistically independent of the matrix $\boldsymbol{X}$. Notice that in (4.20) the assumption which in Section 3.1 was written as $u_t \sim \text{NID}(0, \sigma^2)$ is now expressed in matrix notation using the multivariate normal distribution. In addition, since the assumption that $\boldsymbol{u}$ and $\boldsymbol{X}$ are independent means that the generating process for $\boldsymbol{X}$ is independent of that for $\boldsymbol{y}$, we can express this independence assumption by saying that the regressors $\boldsymbol{X}$ are **exogenous** in the model (4.20); the concept of exogeneity[4] was introduced in Section 1.3 and discussed in Section 3.2.

Tests of a Single Restriction

We begin by considering a single, linear restriction on $\boldsymbol{\beta}$. This could, in principle, be any sort of linear restriction, for example, that $\beta_1 = 5$ or $\beta_3 = \beta_4$. However, it simplifies the analysis, and involves no loss of generality, if we confine our attention to a restriction that one of the coefficients should equal 0. If a restriction does not naturally have the form of a zero restriction, we can always apply suitable linear transformations to $\boldsymbol{y}$ and $\boldsymbol{X}$, of the sort considered in Sections 2.3 and 2.4, in order to rewrite the model so that it does; see Exercises 4.6 and 4.7.

Let us partition $\boldsymbol{\beta}$ as $[\boldsymbol{\beta}_1 \vdots \beta_2]$, where $\boldsymbol{\beta}_1$ is a $(k-1)$–vector and β_2 is a scalar, and consider a restriction of the form $\beta_2 = 0$. When $\boldsymbol{X}$ is partitioned conformably with $\boldsymbol{\beta}$, the model (4.20) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\mathbf{I}), \tag{4.21}$$

where $\boldsymbol{X}_1$ denotes an $n \times (k-1)$ matrix and $\boldsymbol{x}_2$ denotes an n–vector, with $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{x}_2]$.

By the FWL Theorem, the least squares estimate of β_2 from (4.21) is the same as the least squares estimate from the FWL regression

$$\boldsymbol{M}_1\boldsymbol{y} = \beta_2\boldsymbol{M}_1\boldsymbol{x}_2 + \text{residuals}, \tag{4.22}$$

[4] This assumption is usually called **strict exogeneity** in the literature, but, since we will not discuss any other sort of exogeneity in this book, it is convenient to drop the word "strict".

where $\boldsymbol{M}_1 \equiv \mathbf{I} - \boldsymbol{X}_1(\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^\top$ is the matrix that projects on to $\mathcal{S}^\perp(\boldsymbol{X}_1)$. By applying the standard formulas for the OLS estimator and covariance matrix to regression (4.22), under the assumption that the model (4.21) is correctly specified, we find that

$$\hat\beta_2 = \frac{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2} \quad \text{and} \quad \operatorname{Var}(\hat\beta_2) = \sigma^2(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{-1}.$$

In order to test the hypothesis that β_2 equals any specified value, say β_2^0, we have to subtract β_2^0 from $\hat\beta_2$ and divide by the square root of the variance. For the null hypothesis that $\beta_2 = 0$, this yields a test statistic analogous to (4.03),

$$z_{\beta_2} \equiv \frac{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{\sigma(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}}, \tag{4.23}$$

which can be computed only under the unrealistic assumption that σ is known.

If the data are actually generated by the model (4.21) with $\beta_2 = 0$, then

$$\boldsymbol{M}_1 \boldsymbol{y} = \boldsymbol{M}_1(\boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{u}) = \boldsymbol{M}_1 \boldsymbol{u}.$$

Therefore, the right-hand side of (4.23) becomes

$$\frac{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{u}}{\sigma(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}}. \tag{4.24}$$

It is now easy to see that z_{β_2} is distributed as $N(0, 1)$. Because we can condition on $\boldsymbol{X}$, the only thing left in (4.24) that is stochastic is $\boldsymbol{u}$. Since the numerator is just a linear combination of the components of $\boldsymbol{u}$, which is multivariate normal, the entire test statistic must be normally distributed. The variance of the numerator is

$$\begin{aligned} \operatorname{E}(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{M}_1 \boldsymbol{x}_2) &= \boldsymbol{x}_2^\top \boldsymbol{M}_1 \operatorname{E}(\boldsymbol{u}\boldsymbol{u}^\top)\boldsymbol{M}_1 \boldsymbol{x}_2 \\ &= \boldsymbol{x}_2^\top \boldsymbol{M}_1 \sigma^2 \mathbf{I} \boldsymbol{M}_1 \boldsymbol{x}_2 = \sigma^2 \boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2. \end{aligned}$$

Since the denominator of (4.24) is just the square root of the variance of the numerator, we conclude that z_{β_2} is distributed as $N(0, 1)$ under the null hypothesis.

The test statistic z_{β_2} defined in (4.23) has exactly the same distribution under the null hypothesis as the test statistic z defined in (4.03). The analysis of Section 4.2 therefore applies to it without any change. Thus we now know how to test the hypothesis that any coefficient in the classical normal linear model is equal to 0, or to any specified value, but only if we know the variance of the error terms.

In order to handle the more realistic case in which we do not know the variance of the error terms, we need to replace σ in (4.23) by s, the usual least squares standard error estimator for model (4.21), defined in (3.49). If, as usual, $\boldsymbol{M_X}$ is the orthogonal projection on to $\mathcal{S}^\perp(\boldsymbol{X})$, we have $s^2 = \boldsymbol{y}^\top\boldsymbol{M_X y}/(n-k)$, and so we obtain the test statistic

$$t_{\beta_2} \equiv \frac{\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{y}}{s(\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2)^{1/2}} = \left(\frac{\boldsymbol{y}^\top\boldsymbol{M_X y}}{n-k}\right)^{-1/2}\frac{\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{y}}{(\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2)^{1/2}}. \tag{4.25}$$

As we will now demonstrate, this test statistic is distributed as $t(n-k)$ under the null hypothesis. Not surprisingly, it is called a ***t* statistic**.

As we discussed in the last section, for a test statistic to have the $t(n-k)$ distribution, it must be possible to write it as the ratio of a standard normal variable z to the square root of $y/(n-k)$, where y is independent of z and distributed as $\chi^2(n-k)$. The t statistic defined in (4.25) can be rewritten as

$$t_{\beta_2} = \frac{z_{\beta_2}}{\big(\boldsymbol{y}^\top\boldsymbol{M_X y}/((n-k)\sigma^2)\big)^{1/2}}, \tag{4.26}$$

which has the form of such a ratio. We have already shown that $z_{\beta_2} \sim N(0,1)$. Thus it only remains to show that $\boldsymbol{y}^\top\boldsymbol{M_X y}/\sigma^2 \sim \chi^2(n-k)$ and that the random variables in the numerator and denominator of (4.26) are independent.

Under any DGP that belongs to (4.21),

$$\frac{\boldsymbol{y}^\top\boldsymbol{M_X y}}{\sigma^2} = \frac{\boldsymbol{u}^\top\boldsymbol{M_X u}}{\sigma^2} = \boldsymbol{\varepsilon}^\top\boldsymbol{M_X}\boldsymbol{\varepsilon}, \tag{4.27}$$

where $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}/\sigma$ is distributed as $N(\mathbf{0},\mathbf{I})$. Since $\boldsymbol{M_X}$ is a projection matrix with rank $n-k$, the second part of Theorem 4.1 shows that the rightmost expression in (4.27) is distributed as $\chi^2(n-k)$.

To see that the random variables z_{β_2} and $\boldsymbol{\varepsilon}^\top\boldsymbol{M_X}\boldsymbol{\varepsilon}$ are independent, we note first that $\boldsymbol{\varepsilon}^\top\boldsymbol{M_X}\boldsymbol{\varepsilon}$ depends on $\boldsymbol{y}$ only through $\boldsymbol{M_X y}$. Second, from (4.23), it is not hard to see that z_{β_2} depends on $\boldsymbol{y}$ only through $\boldsymbol{P_X y}$, since

$$\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{x}_2^\top\boldsymbol{P_X M}_1\boldsymbol{y} = \boldsymbol{x}_2^\top(\boldsymbol{P_X} - \boldsymbol{P_X P}_1)\boldsymbol{y} = \boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{P_X y};$$

the first equality here simply uses the fact that $\boldsymbol{x}_2 \in \mathcal{S}(\boldsymbol{X})$, and the third equality uses the result (2.36) that $\boldsymbol{P_X P}_1 = \boldsymbol{P}_1\boldsymbol{P_X}$. Independence now follows because, as we will see directly, $\boldsymbol{P_X y}$ and $\boldsymbol{M_X y}$ are independent.

We saw above that $\boldsymbol{M_X y} = \boldsymbol{M_X u}$. Further, from (4.20), $\boldsymbol{P_X y} = \boldsymbol{X\beta} + \boldsymbol{P_X u}$, from which it follows that the centered version of $\boldsymbol{P_X y}$ is $\boldsymbol{P_X u}$. The $n \times n$ matrix of covariances of components of $\boldsymbol{P_X u}$ and $\boldsymbol{M_X u}$ is thus

$$\mathrm{E}(\boldsymbol{P_X u u}^\top\boldsymbol{M_X}) = \sigma^2\boldsymbol{P_X M_X} = \mathbf{O},$$

by (2.26), because $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are complementary projections. These zero covariances imply that $\boldsymbol{P_X u}$ and $\boldsymbol{M_X u}$ are independent, since both are multivariate normal. Geometrically, these vectors have zero covariance because they lie in *orthogonal* subspaces, namely, the images of $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$. Thus, even though the numerator and denominator of (4.26) both depend on $\boldsymbol{y}$, this orthogonality implies that they are independent.

We therefore conclude that the t statistic (4.26) for $\beta_2 = 0$ in the model (4.21) has the $t(n-k)$ distribution. Performing one-tailed and two-tailed tests based on t_{β_2} is almost the same as performing them based on z_{β_2}. We just have to use the $t(n-k)$ distribution instead of the $N(0,1)$ distribution to compute P values or critical values. An interesting property of t statistics is explored in Exercise 14.8.

Tests of Several Restrictions

Economists frequently want to test more than one linear restriction. Let us suppose that there are r restrictions, with $r \leq k$, since there cannot be more equality restrictions than there are parameters in the unrestricted model. As before, there will be no loss of generality if we assume that the restrictions take the form $\boldsymbol{\beta}_2 = \mathbf{0}$. The alternative hypothesis is the model (4.20), which has been rewritten as

$$H_1: \quad \boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}). \tag{4.28}$$

Here $\boldsymbol{X}_1$ is an $n \times k_1$ matrix, $\boldsymbol{X}_2$ is an $n \times k_2$ matrix, $\boldsymbol{\beta}_1$ is a k_1–vector, $\boldsymbol{\beta}_2$ is a k_2–vector, $k = k_1 + k_2$, and the number of restrictions $r = k_2$. Unless $r = 1$, it is no longer possible to use a t test, because there will be one t statistic for each element of $\boldsymbol{\beta}_2$, and we want to compute a single test statistic for all the restrictions at once.

It is natural to base a test on a comparison of how well the model fits when the restrictions are imposed with how well it fits when they are not imposed. The null hypothesis is the regression model

$$H_0: \quad \boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}), \tag{4.29}$$

in which we impose the restriction that $\boldsymbol{\beta}_2 = \mathbf{0}$. As we saw in Section 3.8, the restricted model (4.29) must always fit worse than the unrestricted model (4.28), in the sense that the SSR from (4.29) cannot be smaller, and will almost always be larger, than the SSR from (4.28). However, if the restrictions are true, the reduction in SSR from adding $\boldsymbol{X}_2$ to the regression should be relatively small. Therefore, it seems natural to base a test statistic on the difference between these two SSRs. If USSR denotes the **unrestricted sum of squared residuals**, from (4.28), and RSSR denotes the **restricted sum of squared residuals**, from (4.29), the appropriate test statistic is

$$F_{\boldsymbol{\beta}_2} \equiv \frac{(\text{RSSR} - \text{USSR})/r}{\text{USSR}/(n-k)}. \tag{4.30}$$

Under the null hypothesis, as we will now demonstrate, this test statistic follows the F distribution with r and $n-k$ degrees of freedom. Not surprisingly, it is called an ***F* statistic**.

The restricted SSR is $\boldsymbol{y}^\top\boldsymbol{M}_1\boldsymbol{y}$, and the unrestricted one is $\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$. One way to obtain a convenient expression for the difference between these two expressions is to use the FWL Theorem. By this theorem, the USSR is the SSR from the FWL regression

$$\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \text{residuals}. \tag{4.31}$$

The total sum of squares from (4.31) is $\boldsymbol{y}^\top\boldsymbol{M}_1\boldsymbol{y}$. The explained sum of squares can be expressed in terms of the orthogonal projection on to the r-dimensional subspace $\mathcal{S}(\boldsymbol{M}_1\boldsymbol{X}_2)$, and so the difference is

$$\text{USSR} = \boldsymbol{y}^\top\boldsymbol{M}_1\boldsymbol{y} - \boldsymbol{y}^\top\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y}. \tag{4.32}$$

Therefore,

$$\text{RSSR} - \text{USSR} = \boldsymbol{y}^\top\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y},$$

and the F statistic (4.30) can be written as

$$F_{\boldsymbol{\beta}_2} = \frac{\boldsymbol{y}^\top\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{y}/r}{\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}/(n-k)}. \tag{4.33}$$

Under the null hypothesis, $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$ and $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{u}$. Thus, under this hypothesis, the F statistic (4.33) reduces to

$$\frac{\boldsymbol{\varepsilon}^\top\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon}/r}{\boldsymbol{\varepsilon}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}/(n-k)}, \tag{4.34}$$

where, as before, $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}/\sigma$. We saw in the last subsection that the quadratic form in the denominator of (4.34) is distributed as $\chi^2(n-k)$. Since the quadratic form in the numerator can be written as $\boldsymbol{\varepsilon}^\top\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{\varepsilon}$, it is distributed as $\chi^2(r)$. Moreover, the random variables in the numerator and denominator are independent, because $\boldsymbol{M}_{\boldsymbol{X}}$ and $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}$ project on to mutually orthogonal subspaces: $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_1\boldsymbol{X}_2 = \boldsymbol{M}_{\boldsymbol{X}}(\boldsymbol{X}_2 - \boldsymbol{P}_1\boldsymbol{X}_2) = \mathbf{O}$. Thus it is apparent that the statistic (4.34) follows the $F(r, n-k)$ distribution under the null hypothesis.

A Threefold Orthogonal Decomposition

Each of the restricted and unrestricted models generates an orthogonal decomposition of the dependent variable $\boldsymbol{y}$. It is illuminating to see how these two decompositions interact to produce a threefold orthogonal decomposition. It turns out that all three components of this decomposition have useful interpretations. From the two models, we find that

$$\boldsymbol{y} = \boldsymbol{P}_1\boldsymbol{y} + \boldsymbol{M}_1\boldsymbol{y} \quad \text{and} \quad \boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} + \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}. \tag{4.35}$$

In Exercise 2.17, it was seen that $\boldsymbol{P_X} - \boldsymbol{P}_1$ is an orthogonal projection matrix, equal to $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}$. It follows that

$$\boldsymbol{P_X} = \boldsymbol{P}_1 + \boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}, \tag{4.36}$$

where the two projections on the right-hand side are obviously mutually orthogonal, since $\boldsymbol{P}_1$ annihilates $\boldsymbol{M}_1\boldsymbol{X}_2$. From (4.35) and (4.36), we obtain the threefold orthogonal decomposition

$$\boldsymbol{y} = \boldsymbol{P}_1\boldsymbol{y} + \boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{y} + \boldsymbol{M_X}\boldsymbol{y}. \tag{4.37}$$

The first term is the vector of fitted values from the restricted model, $\boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1$. In this and what follows, we use a tilde (˜) to denote the **restricted estimates**, and a hat (ˆ) to denote the **unrestricted estimates**. The second term is the vector of fitted values from the FWL regression (4.31). It equals $\boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$, where, by the FWL Theorem, $\hat{\boldsymbol{\beta}}_2$ is a subvector of estimates from the unrestricted model. Finally, $\boldsymbol{M_X}\boldsymbol{y}$ is the vector of residuals from the unrestricted model.

Since $\boldsymbol{P_X}\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$, the vector of fitted values from the unrestricted model, we see that

$$\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 = \boldsymbol{X}_1\tilde{\boldsymbol{\beta}}_1 + \boldsymbol{M}_1\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2. \tag{4.38}$$

In Exercise 4.9, this result is exploited to show how to obtain the restricted estimates in terms of the unrestricted estimates.

The F statistic (4.33) can be written as the ratio of the squared norm of the second component in (4.37) to the squared norm of the third, each normalized by the appropriate number of degrees of freedom. Under both hypotheses, the third component $\boldsymbol{M_X}\boldsymbol{y}$ equals $\boldsymbol{M_X}\boldsymbol{u}$, and so it consists of random noise. Its squared norm is a $\chi^2(n-k)$ variable times σ^2, which serves as the (unrestricted) estimate of σ^2 and can be thought of as a measure of the scale of the random noise. Since $\boldsymbol{u} \sim N(\boldsymbol{0}, \sigma^2\mathbf{I})$, every element of $\boldsymbol{u}$ has the same variance, and so every component of (4.37), if centered so as to leave only the random part, should have the same scale.

Under the null hypothesis, the second component is $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{y} = \boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{u}$, which just consists of random noise. But, under the alternative, $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{u}$, and it thus contains a systematic part related to $\boldsymbol{X}_2$. The length of the second component will be greater, on average, under the alternative than under the null, since the random part is there in all cases, but the systematic part is present only under the alternative. The F test compares the squared length of the second component with the squared length of the third. It thus serves to detect the possible presence of systematic variation, related to $\boldsymbol{X}_2$, in the second component of (4.37).

All this means that we want to reject the null whenever the numerator of the F statistic, RSSR − USSR, is relatively large. Consequently, the P value corresponding to a realized F statistic $\hat{F}$ is computed as $1 - F_{r,n-k}(\hat{F})$, where $F_{r,n-k}(\cdot)$ denotes the CDF of the F distribution with the appropriate numbers of degrees of freedom. Thus we compute the P value as if for a one-tailed test. However, F tests are really two-tailed tests, because they test equality restrictions, not inequality restrictions. An F test for $\boldsymbol{\beta}_2 = \mathbf{0}$ will reject the null hypothesis whenever $\hat{\boldsymbol{\beta}}_2$ is sufficiently far from $\mathbf{0}$, whether the individual elements of $\hat{\boldsymbol{\beta}}_2$ are positive or negative.

There is a very close relationship between F tests and t tests. In the previous section, we saw that the square of a random variable with the $t(n-k)$ distribution must have the $F(1, n-k)$ distribution. The square of the t statistic t_{β_2}, defined in (4.25), is

$$t_{\beta_2}^2 = \frac{\boldsymbol{y}^\top \boldsymbol{M}_1 \boldsymbol{x}_2 (\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{-1} \boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}/(n-k)}.$$

This test statistic is evidently a special case of (4.33), with the vector $\boldsymbol{x}_2$ replacing the matrix $\boldsymbol{X}_2$. Thus, when there is only one restriction, it makes no difference whether we use a two-tailed t test or an F test.

An Example of the F Test

The most familiar application of the F test is testing the hypothesis that all the coefficients in a classical normal linear model, except the constant term, are zero. The null hypothesis is that $\boldsymbol{\beta}_2 = \mathbf{0}$ in the model

$$\boldsymbol{y} = \beta_1 \boldsymbol{\iota} + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}), \tag{4.39}$$

where $\boldsymbol{\iota}$ is an n–vector of 1s and $\boldsymbol{X}_2$ is $n \times (k-1)$. In this case, using (4.32), the test statistic (4.33) can be written as

$$F_{\boldsymbol{\beta}_2} = \frac{\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{y}/(k-1)}{\big(\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{y} - \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{X}_2 (\boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{X}_2)^{-1} \boldsymbol{X}_2^\top \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{y}\big)/(n-k)}, \tag{4.40}$$

where $\boldsymbol{M}_{\boldsymbol{\iota}}$ is the projection matrix that takes deviations from the mean, which was defined in (2.32). Thus the matrix expression in the numerator of (4.40) is just the explained sum of squares, or ESS, from the FWL regression

$$\boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{\iota}} \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \text{residuals}.$$

Similarly, the matrix expression in the denominator is the total sum of squares, or TSS, from this regression, minus the ESS. Since the centered R^2 from (4.39) is just the ratio of this ESS to this TSS, it requires only a little algebra to show that

$$F_{\boldsymbol{\beta}_2} = \frac{n-k}{k-1} \times \frac{R_c^2}{1 - R_c^2}.$$

Therefore, the F statistic (4.40) depends on the data only through the centered R^2, of which it is a monotonically increasing function.

Testing the Equality of Two Parameter Vectors

It is often natural to divide a sample into two, or possibly more than two, subsamples. These might correspond to periods of fixed exchange rates and floating exchange rates, large firms and small firms, rich countries and poor countries, or men and women, to name just a few examples. We may then ask whether a linear regression model has the same coefficients for both the subsamples. It is natural to use an F test for this purpose. Because the classic treatment of this problem is found in Chow (1960), the test is often called a **Chow test**; later treatments include Fisher (1970) and Dufour (1982).

Let us suppose, for simplicity, that there are only two subsamples, of lengths n_1 and n_2, with $n = n_1 + n_2$. We will assume that both n_1 and n_2 are greater than k, the number of regressors. If we separate the subsamples by partitioning the variables, we can write

$$\boldsymbol{y} \equiv \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix}, \quad \text{and} \quad \boldsymbol{X} \equiv \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix},$$

where $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ are, respectively, an n_1–vector and an n_2–vector, while $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are $n_1 \times k$ and $n_2 \times k$ matrices. Even if we need different parameter vectors, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, for the two subsamples, we can nonetheless put the subsamples together in the following regression model:

$$\begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix} \boldsymbol{\beta}_1 + \begin{bmatrix} \mathbf{O} \\ \boldsymbol{X}_2 \end{bmatrix} \boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{4.41}$$

It can readily be seen that, in the first subsample, the regression functions are the components of $\boldsymbol{X}_1\boldsymbol{\beta}_1$, while, in the second, they are the components of $\boldsymbol{X}_2(\boldsymbol{\beta}_1 + \boldsymbol{\gamma})$. Thus $\boldsymbol{\gamma}$ is to be defined as $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$. If we define $\boldsymbol{Z}$ as an $n \times k$ matrix with $\mathbf{O}$ in its first n_1 rows and $\boldsymbol{X}_2$ in the remaining n_2 rows, then (4.41) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}_1 + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}). \tag{4.42}$$

This is a regression model with n observations and $2k$ regressors. It has been constructed in such a way that $\boldsymbol{\beta}_1$ is estimated directly, while $\boldsymbol{\beta}_2$ is estimated using the relation $\boldsymbol{\beta}_2 = \boldsymbol{\gamma} + \boldsymbol{\beta}_1$. Since the restriction that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$ is equivalent to the restriction that $\boldsymbol{\gamma} = \mathbf{0}$ in (4.42), the null hypothesis has been expressed as a set of k zero restrictions. Since (4.42) is just a classical normal linear model with k linear restrictions to be tested, the F test provides the appropriate way to test those restrictions.

The F statistic can perfectly well be computed as usual, by running (4.42) to get the USSR and then running the restricted model, which is just the regression of $\boldsymbol{y}$ on $\boldsymbol{X}$, to get the RSSR. However, there is another way to compute the USSR. In Exercise 4.10, readers are invited to show that it is simply the sum of the two SSRs obtained by running two independent regressions on the two subsamples. If SSR_1 and SSR_2 denote the sums of squared residuals from these two regressions, and RSSR denotes the sum of squared residuals from regressing $\boldsymbol{y}$ on $\boldsymbol{X}$, the F statistic becomes

$$F_\gamma = \frac{(\mathrm{RSSR} - \mathrm{SSR}_1 - \mathrm{SSR}_2)/k}{(\mathrm{SSR}_1 + \mathrm{SSR}_2)/(n-2k)}. \tag{4.43}$$

This **Chow statistic**, as it is often called, is distributed as $F(k, n-2k)$ under the null hypothesis that $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$.

4.5 Large-Sample Tests in Linear Regression Models

The t and F tests that we developed in the previous section are exact only under the strong assumptions of the classical normal linear model. If the error vector were not normally distributed or not independent of the matrix of regressors, we could still compute t and F statistics, but they would not actually follow their namesake distributions in finite samples. However, like a great many test statistics in econometrics which do not follow any known distribution exactly, they would in many cases approximately follow known distributions in large samples. In such cases, we can perform what are called **large-sample tests** or **asymptotic tests**, using the approximate distributions to compute P values or critical values.

Asymptotic theory is concerned with the distributions of estimators and test statistics as the sample size n tends to infinity. It often allows us to obtain simple results which provide useful approximations even when the sample size is far from infinite. In this book, we do not intend to discuss asymptotic theory at the advanced level of Davidson (1994) or White (1984). A rigorous introduction to the fundamental ideas may be found in Gallant (1997), and a less formal treatment is provided in Davidson and MacKinnon (1993). However, it is impossible to understand large parts of econometrics without having some idea of how asymptotic theory works and what we can learn from it. In this section, we will show that asymptotic theory gives us results about the distributions of t and F statistics under much weaker assumptions than those of the classical normal linear model.

Laws of Large Numbers

There are two types of fundamental results on which asymptotic theory is based. The first type, which we briefly discussed in Section 3.3, is called a **law of large numbers**, or **LLN**. A law of large numbers may apply to any quantity which can be written as an average of n random variables, that is, $1/n$ times their sum. Suppose, for example, that

$$\bar{x} \equiv \frac{1}{n}\sum_{t=1}^{n} x_t,$$

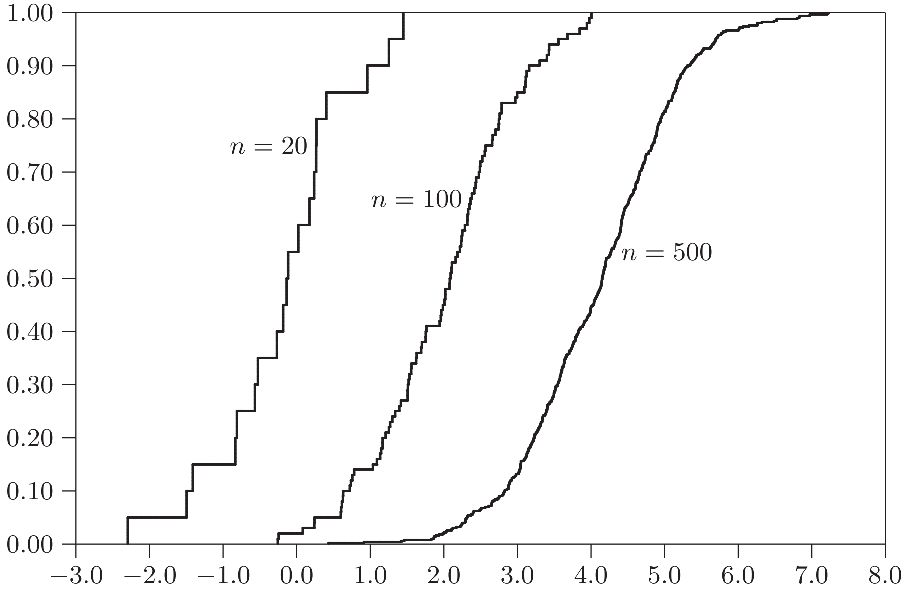

Figure 4.6 EDFs for several sample sizes

where the x_t are independent random variables, each with its own bounded finite variance σ_t^2 and with a common mean μ. Then a fairly simple LLN assures us that, as $n \to \infty$, $\bar{x}$ tends to μ.

An example of how useful a law of large numbers can be is the **Fundamental Theorem of Statistics**, which concerns the **empirical distribution function**, or **EDF**, of a random sample. The EDF was introduced in Exercises 1.1 and 3.4. Suppose that X is a random variable with CDF $F(X)$ and that we obtain a random sample of size n with typical element x_t, where each x_t is an independent realization of X. The **empirical distribution** defined by this sample is the discrete distribution that puts a weight of $1/n$ at each of the x_t, $t = 1, \ldots, n$. The EDF is the distribution function of the empirical distribution, and it can be expressed algebraically as

$$\hat{F}(x) \equiv \frac{1}{n} \sum_{t=1}^{n} I(x_t \leq x), \tag{4.44}$$

where $I(\cdot)$ is the **indicator function**, which takes the value 1 when its argument is true and takes the value 0 otherwise. Thus, for a given argument x, the sum on the right-hand side of (4.44) counts the number of realizations x_t that are smaller than or equal to x. The EDF has the form of a step function: The height of each step is $1/n$, and the width is equal to the difference between two successive values of x_t. According to the Fundamental Theorem of Statistics, the EDF consistently estimates the CDF of the random variable X.

Figure 4.6 shows the EDFs for three samples of sizes 20, 100, and 500 drawn from three normal distributions, each with variance 1 and with means 0, 2, and 4, respectively. These may be compared with the CDF of the standard normal distribution in the lower panel of Figure 4.2. There is not much resemblance between the EDF based on $n = 20$ and the normal CDF from which the sample was drawn, but the resemblance is somewhat stronger for $n = 100$ and very much stronger for $n = 500$. It is a simple matter to simulate data from an EDF, as we will see in the next section, and this type of simulation can be very useful.

It is very easy to prove the Fundamental Theorem of Statistics. For any real value of x, each term in the sum on the right-hand side of (4.44) depends only on x_t. The expectation of $I(x_t \leq x)$ can be found by using the fact that it can take on only two values, 1 and 0. The expectation is

$$\begin{aligned}\mathrm{E}\big(I(x_t \leq x)\big) &= 0 \cdot \Pr\big(I(x_t \leq x) = 0\big) + 1 \cdot \Pr\big(I(x_t \leq x) = 1\big) \\ &= \Pr\big(I(x_t \leq x) = 1\big) = \Pr(x_t \leq x) = F(x).\end{aligned}$$

Since the x_t are mutually independent, so too are the terms $I(x_t \leq x)$. Since the x_t all follow the same distribution, so too must these terms. Thus (4.44) is the mean of n IID random terms, each with finite expectation. The simplest of all LLNs (due to Khinchin) applies to such a mean, and we conclude that, for every x, $\hat{F}(x)$ is a consistent estimator of $F(x)$.

There are many different LLNs, some of which do not require that the individual random variables have a common mean or be independent, although the amount of dependence must be limited. If we can apply a LLN to any random average, we can treat it as a nonrandom quantity for the purpose of asymptotic analysis. In many cases, this means that we must divide the quantity of interest by n. For example, the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ that appears in the OLS estimator generally does not converge to anything as $n \to \infty$. In contrast, the matrix $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ will, under many plausible assumptions about how $\boldsymbol{X}$ is generated, tend to a nonstochastic limiting matrix $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ as $n \to \infty$.

Central Limit Theorems

The second type of fundamental result on which asymptotic theory is based is called a **central limit theorem**, or **CLT**. Central limit theorems are crucial in establishing the asymptotic distributions of estimators and test statistics. They tell us that, in many circumstances, $1/\sqrt{n}$ times the sum of n centered random variables will approximately follow a normal distribution when n is sufficiently large.

Suppose that the random variables x_t, $t = 1, \ldots, n$, are independently and identically distributed with mean μ and variance σ^2. Then, according to the Lindeberg-Lévy central limit theorem, the quantity

$$z \equiv \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{x_t - \mu}{\sigma} \tag{4.45}$$

is **asymptotically distributed** as $N(0,1)$. This means that, as $n \to \infty$, the random variable z tends to a random variable which follows the $N(0,1)$ distribution. It may seem curious that we divide by $\sqrt{n}$ instead of by n in (4.45), but this is an essential feature of every CLT. To see why, we calculate the variance of z. Since the terms in the sum in (4.45) are independent, the variance of z is just the sum of the variances of the n terms:

$$\mathrm{Var}(z) = n\,\mathrm{Var}\Big(\frac{1}{\sqrt{n}}\frac{x_t-\mu}{\sigma}\Big) = \frac{n}{n} = 1.$$

If we had divided by n, we would, by a law of large numbers, have obtained a random variable with a plim of 0 instead of a random variable with a limiting standard normal distribution. Thus, whenever we want to use a CLT, we must ensure that a factor of $n^{-1/2} = 1/\sqrt{n}$ is present.

Just as there are many different LLNs, so too are there many different CLTs, almost all of which impose weaker conditions on the x_t than those imposed by the Lindeberg-Lévy CLT. The assumption that the x_t are identically distributed is easily relaxed, as is the assumption that they are independent. However, if there is either too much dependence or too much heterogeneity, a CLT may not apply. Several CLTs are discussed in Section 4.7 of Davidson and MacKinnon (1993), and Davidson (1994) provides a more advanced treatment. In all cases of interest to us, the CLT says that, for a sequence of random variables x_t, $t = 1, \ldots, \infty$, with $\mathrm{E}(x_t) = 0$,

$$\operatorname*{plim}_{n\to\infty} n^{-1/2}\sum_{t=1}^{n} x_t = x_0 \sim N\Big(0, \lim_{n\to\infty}\tfrac{1}{n}\sum_{t=1}^{n}\mathrm{Var}(x_t)\Big).$$

We sometimes need vector, or **multivariate**, versions of CLTs. Suppose that we have a sequence of random m–vectors $\boldsymbol{x}_t$, for some fixed m, with $\mathrm{E}(\boldsymbol{x}_t) = \mathbf{0}$. Then the appropriate multivariate version of a CLT tells us that

$$\operatorname*{plim}_{n\to\infty} n^{-1/2}\sum_{t=1}^{n} \boldsymbol{x}_t = \boldsymbol{x}_0 \sim N\Big(\mathbf{0}, \lim_{n\to\infty}\tfrac{1}{n}\sum_{t=1}^{n}\mathrm{Var}(\boldsymbol{x}_t)\Big), \qquad (4.46)$$

where $\boldsymbol{x}_0$ is multivariate normal, and each $\mathrm{Var}(\boldsymbol{x}_t)$ is an $m \times m$ matrix.

Figure 4.7 illustrates the fact that CLTs often provide good approximations even when n is not very large. Both panels of the figure show the densities of various random variables z defined as in (4.45). In the top panel, the x_t are uniformly distributed, and we see that z is remarkably close to being distributed as standard normal even when n is as small as 8. This panel does not show results for larger values of n because they would have made it too hard to read. In the bottom panel, the x_t follow the $\chi^2(1)$ distribution, which exhibits extreme right skewness. The mode of the distribution is 0, there are no values less than 0, and there is a very long right-hand tail. For $n = 4$

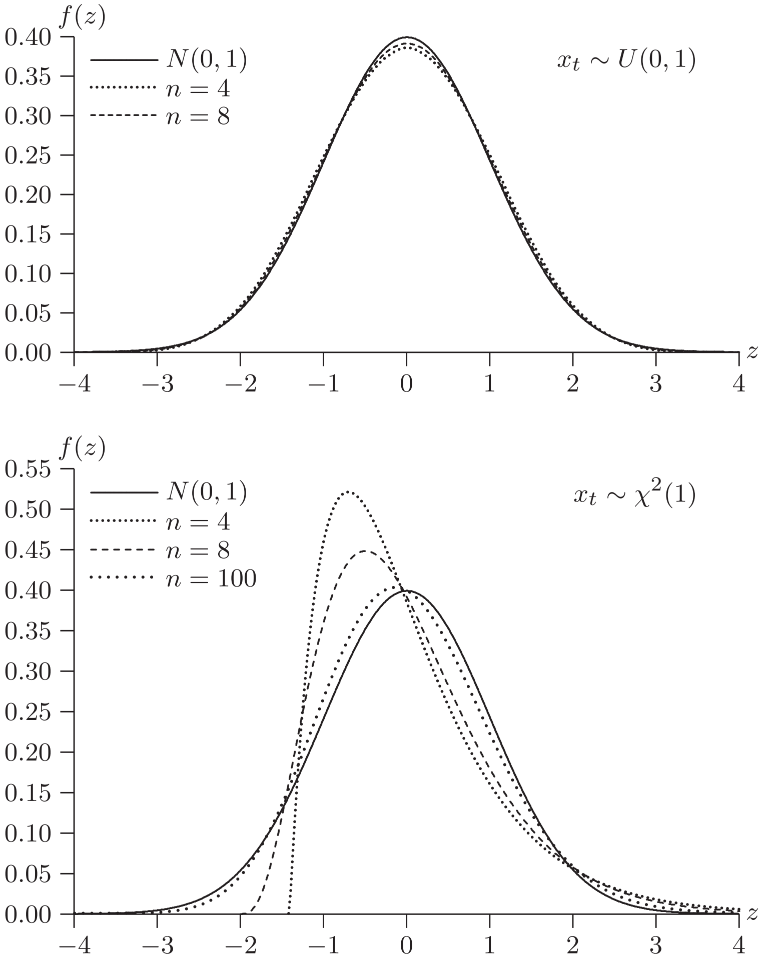

Figure 4.7 The normal approximation for different values of n

and $n = 8$, the standard normal provides a poor approximation to the actual distribution of z. For $n = 100$, on the other hand, the approximation is not bad at all, although it is still noticeably skewed to the right.

Asymptotic Tests

The t and F tests that we discussed in the previous section are asymptotically valid under much weaker conditions than those needed to prove that they actually have their namesake distributions in finite samples. Suppose that the DGP is

$$\boldsymbol{y} = \boldsymbol{X\beta}_0 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\mathbf{0}, \sigma_0^2\mathbf{I}), \tag{4.47}$$

where $\boldsymbol{\beta}_0$ satisfies whatever hypothesis is being tested, and the error terms are drawn from some specific but unknown distribution with mean 0 and variance σ_0^2. We allow $\boldsymbol{X}_t$ to contain lagged dependent variables, and so we abandon the assumption of exogenous regressors and replace it with assumption (3.10) from Section 3.2, plus an analogous assumption about the variance. These two assumptions can be written as

$$\mathrm{E}(u_t \,|\, \boldsymbol{X}_t) = 0 \quad \text{and} \quad \mathrm{E}(u_t^2 \,|\, \boldsymbol{X}_t) = \sigma_0^2. \tag{4.48}$$

The first of these assumptions, which is assumption (3.10), can be referred to in two ways. From the point of view of the error terms, it says that they are **innovations**. An innovation is a random variable of which the mean is 0 conditional on the information in the explanatory variables, and so knowledge of the values taken by the latter is of no use in predicting the mean of the innovation. From the point of view of the explanatory variables $\boldsymbol{X}_t$, assumption (3.10) says that they are **predetermined** with respect to the error terms. We thus have two different ways of saying the same thing. Both can be useful, depending on the circumstances.

Although we have greatly weakened the assumptions of the classical normal linear model, we now need to make an additional assumption in order to be able to use asymptotic results. We therefore assume that the data-generating process for the explanatory variables is such that

$$\operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}, \tag{4.49}$$

where $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ is a finite, deterministic, positive definite matrix. We made this assumption previously, in Section 3.3, when we proved that the OLS estimator is consistent. Although it is often reasonable, condition (4.49) is violated in many cases. For example, it cannot hold if one of the columns of the $\boldsymbol{X}$ matrix is a linear time trend, because $\sum_{t=1}^n t^2$ grows at a rate faster than n.

Now consider the t statistic (4.25) for testing the hypothesis that $\beta_2 = 0$ in the model (4.21). The key to proving that (4.25), or any test statistic, has a certain **asymptotic distribution** is to write it as a function of quantities to which we can apply either a LLN or a CLT. Therefore, we rewrite (4.25) as

$$t_{\beta_2} = \left(\frac{\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}}{n-k}\right)^{-1/2} \frac{n^{-1/2}\boldsymbol{x}_2{}^\top\boldsymbol{M}_1\boldsymbol{y}}{(n^{-1}\boldsymbol{x}_2{}^\top\boldsymbol{M}_1\boldsymbol{x}_2)^{1/2}}, \tag{4.50}$$

where the numerator and denominator of the second factor have both been multiplied by $n^{-1/2}$. Under the DGP (4.47), $s^2 \equiv \boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}/(n-k)$ tends to σ_0^2 as $n \to \infty$. This statement, which is equivalent to saying that the OLS error variance estimator s^2 is consistent under our weaker assumptions, follows from a LLN, because s^2 has the form of an average, and the calculations leading to (3.49) showed that the mean of s^2 is σ_0^2. It follows from the consistency of s^2 that the first factor in (4.50) tends to $1/\sigma_0$ as $n \to \infty$. When the data are generated by (4.47) with $\beta_2 = 0$, we have that $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{u}$, and so (4.50) is asymptotically equivalent to

$$\frac{n^{-1/2}\boldsymbol{x}_2^{\top}\boldsymbol{M}_1\boldsymbol{u}}{\sigma_0(n^{-1}\boldsymbol{x}_2^{\top}\boldsymbol{M}_1\boldsymbol{x}_2)^{1/2}}. \tag{4.51}$$

It is now easy to derive the asymptotic distribution of t_{β_2} if for a moment we reinstate the assumption that the regressors are exogenous. In that case, we can work conditionally on $\boldsymbol{X}$, which means that the only part of (4.51) that is treated as random is $\boldsymbol{u}$. The numerator of (4.51) is $n^{-1/2}$ times a weighted sum of the u_t, each of which has mean 0, and the conditional variance of this weighted sum is

$$\mathrm{E}(\boldsymbol{x}_2^{\top}\boldsymbol{M}_1\boldsymbol{u}\boldsymbol{u}^{\top}\boldsymbol{M}_1\boldsymbol{x}_2 \,|\, \boldsymbol{X}) = \sigma_0^2\boldsymbol{x}_2^{\top}\boldsymbol{M}_1\boldsymbol{x}_2.$$

Thus (4.51) evidently has mean 0 and variance 1, conditional on $\boldsymbol{X}$. But since 0 and 1 do not depend on $\boldsymbol{X}$, these are also the unconditional mean and variance of (4.51). Provided that we can apply a CLT to the numerator of (4.51), the numerator of t_{β_2} must be asymptotically normally distributed, and we conclude that, under the null hypothesis, with exogenous regressors,

$$t_{\beta_2} \overset{a}{\sim} N(0,1). \tag{4.52}$$

The notation "$\overset{a}{\sim}$" means that t_{β_2} is **asymptotically distributed** as $N(0,1)$. Since the DGP is assumed to be (4.47), this result does *not* require that the error terms be normally distributed.

The *t* Test with Predetermined Regressors

If we relax the assumption of exogenous regressors, the analysis becomes more complicated. Readers not interested in the algebraic details may well wish to skip to next section, since what follows is not essential for understanding the rest of this chapter. However, this subsection provides an excellent example of how asymptotic theory works, and it illustrates clearly just why we can relax some assumptions but not others.

We begin by applying a CLT to the k–vector

$$\boldsymbol{v} \equiv n^{-1/2}\boldsymbol{X}^{\top}\boldsymbol{u} = n^{-1/2}\sum_{t=1}^{n} u_t\boldsymbol{X}_t^{\top}. \tag{4.53}$$

By (3.10), $\mathrm{E}(u_t \,|\, \boldsymbol{X}_t) = 0$. This implies that $\mathrm{E}(u_t\boldsymbol{X}_t^{\top}) = \boldsymbol{0}$, as required for the CLT, which then tells us that

$$\boldsymbol{v} \overset{a}{\sim} N\Big(\boldsymbol{0}, \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{Var}(u_t\boldsymbol{X}_t^{\top})\Big) = N\Big(\boldsymbol{0}, \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(u_t^2\boldsymbol{X}_t^{\top}\boldsymbol{X}_t)\Big);$$

recall (4.46). Notice that, because $\boldsymbol{X}_t$ is a $1 \times k$ row vector, the covariance matrix here is $k \times k$, as it must be. The second assumption in (4.48) allows us to simplify the limiting covariance matrix:

$$\begin{aligned}\lim_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(u_t^2 \boldsymbol{X}_t^\top \boldsymbol{X}_t) &= \lim_{n\to\infty} \sigma_0^2 \frac{1}{n}\sum_{t=1}^{n} \mathrm{E}(\boldsymbol{X}_t^\top \boldsymbol{X}_t)\\ &= \sigma_0^2 \operatorname*{plim}_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} \boldsymbol{X}_t^\top \boldsymbol{X}_t \qquad (4.54)\\ &= \sigma_0^2 \operatorname*{plim}_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top \boldsymbol{X} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}.\end{aligned}$$

We applied a LLN in reverse to go from the first line to the second, and the last equality follows from (4.49).

Now consider the numerator of (4.51). It can be written as

$$n^{-1/2}\boldsymbol{x}_2^\top \boldsymbol{u} - n^{-1/2}\boldsymbol{x}_2^\top \boldsymbol{P}_1 \boldsymbol{u}. \qquad (4.55)$$

The first term of this expression is just the last, or k^{th}, component of $\boldsymbol{v}$, which we can denote by v_2. By writing out the projection matrix $\boldsymbol{P}_1$ explicitly, and dividing various expressions by n in a way that cancels out, the second term can be rewritten as

$$n^{-1}\boldsymbol{x}_2^\top \boldsymbol{X}_1 (n^{-1}\boldsymbol{X}_1^\top \boldsymbol{X}_1)^{-1} n^{-1/2}\boldsymbol{X}_1^\top \boldsymbol{u}. \qquad (4.56)$$

By assumption (4.49), the first and second factors of (4.56) tend to deterministic limits. In obvious notation, the first tends to $\boldsymbol{S}_{21}$, which is a submatrix of $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$, and the second tends to $\boldsymbol{S}_{11}^{-1}$, which is the inverse of a submatrix of $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$. Thus only the last factor remains random when $n \to \infty$. It is just the subvector of $\boldsymbol{v}$ consisting of the first $k-1$ components, which we denote by $\boldsymbol{v}_1$. Asymptotically, in partitioned matrix notation, (4.55) becomes

$$v_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1 = \begin{bmatrix} -\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{v}_1 \\ v_2 \end{bmatrix}.$$

Since $\boldsymbol{v}$ is asymptotically multivariate normal, this scalar expression is asymptotically normal, with mean zero and variance

$$\sigma_0^2 \begin{bmatrix} -\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1} & 1 \end{bmatrix} \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}} \begin{bmatrix} -\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} \\ 1 \end{bmatrix},$$

where, since $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ is symmetric, $\boldsymbol{S}_{12}$ is just the transpose of $\boldsymbol{S}_{21}$. If we now express $\boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}$ as a partitioned matrix, the variance of (4.55) is seen to be

$$\sigma_0^2 \begin{bmatrix} -\boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} \\ \boldsymbol{S}_{21} & \boldsymbol{S}_{22} \end{bmatrix} \begin{bmatrix} -\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} \\ 1 \end{bmatrix} = \sigma_0^2 \big(\boldsymbol{S}_{22} - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}\big). \qquad (4.57)$$

The denominator of (4.51) is, thankfully, easier to analyze. The square of the second factor is

$$\begin{aligned} n^{-1}\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2 &= n^{-1}\boldsymbol{x}_2^\top\boldsymbol{x}_2 - n^{-1}\boldsymbol{x}_2^\top\boldsymbol{P}_1\boldsymbol{x}_2 \\ &= n^{-1}\boldsymbol{x}_2^\top\boldsymbol{x}_2 - n^{-1}\boldsymbol{x}_2^\top\boldsymbol{X}_1\big(n^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_1\big)^{-1}n^{-1}\boldsymbol{X}_1^\top\boldsymbol{x}_2. \end{aligned}$$

In the limit, all the pieces of this expression become submatrices of $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$, and so we find that

$$n^{-1}\boldsymbol{x}_2^\top\boldsymbol{M}_1\boldsymbol{x}_2 \to \boldsymbol{S}_{22} - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}.$$

When it is multiplied by σ_0^2, this is just (4.57), the variance of the numerator of (4.51). Thus, asymptotically, we have shown that t_{β_2} is the ratio of a normal random variable with mean zero to its standard deviation. Consequently, we have established that, under the null hypothesis, with regressors that are not necessarily exogenous but merely predetermined, $t_{\beta_2} \overset{a}{\sim} N(0,1)$. This result is what we previously obtained as (4.52) when we assumed that the regressors were exogenous.

Asymptotic $\boldsymbol{F}$ Tests

A similar analysis can be performed for the F statistic (4.33) for the null hypothesis that $\boldsymbol{\beta}_2 = \boldsymbol{0}$ in the model (4.28). Under the null, $F_{\boldsymbol{\beta}_2}$ is equal to expression (4.34), which can be rewritten as

$$\frac{n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{M}_1\boldsymbol{X}_2(n^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon}/r}{\boldsymbol{\varepsilon}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}/(n-k)}, \tag{4.58}$$

where $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}/\sigma_0$. It is not hard to use the results we obtained for the t statistic to show that, as $n \to \infty$,

$$rF_{\boldsymbol{\beta}_2} \overset{a}{\sim} \chi^2(r) \tag{4.59}$$

under the null hypothesis; see Exercise 4.12. Since $1/r$ times a random variable that follows the $\chi^2(r)$ distribution is distributed as $F(r,\infty)$, we can also conclude that $F_{\boldsymbol{\beta}_2} \overset{a}{\sim} F(r, n-k)$.

The results (4.52) and (4.59) justify the use of t and F tests outside the confines of the classical normal linear model. We can compute P values using either the standard normal or t distributions in the case of t statistics, and either the χ^2 or F distributions in the case of F statistics. Of course, if we use the χ^2 distribution, we have to multiply the F statistic by r.

Whatever distribution we use, these P values will be approximate, and tests based on them will not be exact in finite samples. In addition, our theoretical results do not tell us just how accurate they will be. If we decide to use a nominal level of α for a test, we will reject if the approximate P value is less than α. In many cases, but certainly not all, such tests will probably be quite accurate, committing Type I errors with probability reasonably close to α. They may either **overreject**, that is, reject the null hypothesis more than $100\alpha\%$ of the time when it is true, or **underreject**, that is, reject the null hypothesis less than $100\alpha\%$ of the time. Whether they will overreject or underreject, and how severely, will depend on many things, including the sample size, the distribution of the error terms, the number of regressors and their properties, and the relationship between the error terms and the regressors.

4.6 Simulation-Based Tests

When we introduced the concept of a test statistic in Section 4.2, we specified that it should have a known distribution under the null hypothesis. In the previous section, we relaxed this requirement and developed large-sample test statistics for which the distribution is known only approximately. In all the cases we have studied, the distribution of the statistic under the null hypothesis was not only (approximately) known, but also the *same* for all DGPs contained in the null hypothesis. This is a very important property, and it is useful to introduce some terminology that will allow us to formalize it.

We begin with a simple remark. A hypothesis, null or alternative, can always be represented by a *model*, that is, a set of DGPs. For instance, the null and alternative hypotheses (4.29) and (4.28) associated with an F test of several restrictions are both classical normal linear models. The most fundamental sort of null hypothesis that we can test is a **simple hypothesis**. Such a hypothesis is represented by a model that contains one and only one DGP. Simple hypotheses are very rare in econometrics. The usual case is that of a **compound hypothesis**, which is represented by a model that contains more than one DGP. This can cause serious problems. Except in certain special cases, such as the exact tests in the classical normal linear model that we investigated in Section 4.4, a test statistic will have different distributions under the different DGPs contained in the model. In such a case, if we do not know just which DGP in the model generated our data, then we cannot know the distribution of the test statistic.

If a test statistic is to have a known distribution under some given null hypothesis, then it must have the same distribution for each and every DGP contained in that null hypothesis. A random variable with the property that its distribution is the same for all DGPs in a model $\mathbb{M}$ is said to be **pivotal**, or to be a **pivot**, for the model $\mathbb{M}$. The distribution is allowed to depend on the sample size, and perhaps on the observed values of exogenous variables. However, for any given sample size and set of exogenous variables, it must be invariant across all DGPs in $\mathbb{M}$. Note that *all* test statistics are pivotal for a simple null hypothesis.

The large sample tests considered in the last section allow for null hypotheses that do not respect the rigid constraints of the classical normal linear model.

The price they pay for this added generality is that t and F statistics now have distributions that depend on things like the error distribution: They are therefore not pivotal statistics. However, their *asymptotic* distributions are independent of such things, and are thus invariant across all the DGPs of the model that represents the null hypothesis. Such statistics are said to be **asymptotically pivotal**, or **asymptotic pivots**, for that model.

Simulated *P* Values

The distributions of the test statistics studied in Section 4.3 are all thoroughly known, and their CDFs can easily be evaluated by computer programs. The computation of P values is therefore straightforward. Even if it were not, we could always estimate them by simulation. For any pivotal test statistic, the P value can be estimated by simulation to any desired level of accuracy. Since a pivotal statistic has the same distribution for all DGPs in the model under test, we can arbitrarily choose any such DGP for generating simulated samples and simulated test statistics.

The theoretical justification for using simulation to estimate P values is the Fundamental Theorem of Statistics, which we discussed in Section 4.5. It tells us that the empirical distribution of a set of independent drawings of a random variable generated by some DGP converges to the true CDF of the random variable under that DGP. This is just as true of simulated drawings generated by the computer as for random variables generated by a natural random mechanism. Thus, if we knew that a certain test statistic was pivotal but did not know how it was distributed, we could select any DGP in the null model and generate simulated samples from it. For each of these, we could then compute the test statistic. If the simulated samples are mutually independent, the set of simulated test statistics thus generated constitutes a set of independent drawings from the distribution of the test statistic, and their EDF is a consistent estimate of the CDF of that distribution.

Suppose that we have computed a test statistic $\hat\tau$, which could be a t statistic, an F statistic, or some other type of test statistic, using some data set with n observations. We can think of $\hat\tau$ as being a realization of a random variable τ. We wish to test a null hypothesis represented by a model $\mathbb{M}$ for which τ is pivotal, and we want to reject the null whenever $\hat\tau$ is sufficiently large, as in the cases of an F statistic, a t statistic when the rejection region is in the upper tail, or a squared t statistic. If we denote by F the CDF of the distribution of τ under the null hypothesis, the P value for a test based on $\hat\tau$ is

$$p(\hat\tau) \equiv 1 - F(\hat\tau). \tag{4.60}$$

Since $\hat\tau$ is computed directly from our original data, this P value can be estimated if we can estimate the CDF F evaluated at $\hat\tau$.

The procedure we are about to describe is very general in its application, and so we describe it in detail. In order to estimate a P value by simulation, we choose any DGP in $\mathbb{M}$, and draw B samples of size n from it. How to choose B will be discussed shortly; it will typically be rather large, and $B = 999$ may often be a reasonable choice. We denote the simulated samples as $\boldsymbol{y}_j^*$, $j = 1, \ldots, B$. The star (*) notation will be used systematically to denote quantities generated by simulation. B is used to denote the number of simulations in order to emphasize the connection with the bootstrap, which we will discuss below.

Using the simulated sample, for each j we compute a simulated test statistic, say τ_j^*, in exactly the same way that $\hat\tau$ was computed from the original data $\boldsymbol{y}$. We can then construct the EDF of the τ_j^* analogously to (4.44):

$$\hat F^*(x) = \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* \le x).$$

Our estimate of the true P value (4.60) is therefore

$$\hat p^*(\hat\tau) = 1 - \hat F^*(\hat\tau) = 1 - \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* \le \hat\tau) = \frac{1}{B}\sum_{j=1}^{B} I(\tau_j^* > \hat\tau). \tag{4.61}$$

The third equality in (4.61) can be understood by noting that the rightmost expression is the proportion of simulations for which τ_j^* is greater than $\hat\tau$, while the second expression from the right is 1 minus the proportion for which τ_j^* is less than or equal to $\hat\tau$. These proportions are obviously the same.

We can see that $\hat p^*(\hat\tau)$ must lie between 0 and 1, as any P value must. For example, if $B = 999$, and 36 of the τ_j^* were greater than $\hat\tau$, we would have $\hat p^*(\hat\tau) = 36/999 = 0.036$. In this case, since $\hat p^*(\hat\tau)$ is less than 0.05, we would reject the null hypothesis at the .05 level. Since the EDF converges to the true CDF, it follows that, if B were infinitely large, this procedure would yield an exact test, and the outcome of the test would be the same as if we computed the P value analytically using the CDF of τ. In fact, as we will see shortly, this procedure will yield an exact test even for certain finite values of B.

The sort of test we have just described, based on simulating a pivotal statistic, is called a **Monte Carlo test**. Simulation experiments in general are often referred to as **Monte Carlo experiments**, because they involve generating random numbers, as do the games played in casinos. Around the time that computer simulations first became possible, the most famous casino was the one in Monte Carlo. If computers had been developed just a little later, we would probably be talking now of Las Vegas tests and Las Vegas experiments.

Random Number Generators

Drawing a simulated sample of size n requires us to generate at least n random, or pseudo-random, numbers. As we mentioned in Section 1.3, a **random number generator**, or **RNG**, is a program for generating random numbers.

Most such programs generate numbers that appear to be drawings from the uniform $U(0,1)$ distribution, which can then be transformed into drawings from other distributions. There is a large literature on RNGs, to which Press *et al.* (1992a, 1992b, Chapter 7) provides an accessible introduction. See also Knuth (1998, Chapter 3) and Gentle (1998).

Although there are many types of RNG, the most common are variants of the **linear congruential generator**,

$$z_i = \lambda z_{i-1} + c \;\, [\mathrm{mod}\; m], \quad \eta_i = \frac{z_i}{m}, \quad i = 1, 2, \ldots, \tag{4.62}$$

where η_i is the i^{th} random number generated, and m, λ, c, and so also the z_i, are positive integers. The notation [mod m] means that we divide what precedes it by m and retain the remainder. This generator starts with a (generally large) positive integer z_0 called the **seed**, multiplies it by λ, and then adds c to obtain an integer that may well be bigger than m. It then obtains z_1 as the remainder from division by m. To generate the next random number, the process is repeated with z_1 replacing z_0, and so on. At each stage, the actual random number output by the generator is z_i/m, which, since $0 \leq z_i \leq m$, lies in the interval $[0,1]$. For a given generator defined by λ, m, and c, the sequence of random numbers depends entirely on the seed. If we provide the generator with the same seed, we will get the same sequence of numbers.

How well or badly this procedure works depends on how λ, m, and c are chosen. On 32-bit computers, many commonly used generators set $c = 0$ and use for m a prime number that is either a little less than 2^{32} or a little less than 2^{31}. When $c = 0$, the generator is said to be **multiplicative congruential**. The parameter λ, which will be large but substantially smaller than m, must be chosen so as to satisfy some technical conditions. When λ and m are chosen properly with $c = 0$, the RNG will have a **period** of $m - 1$. This means that it will generate every rational number with denominator m between $1/m$ and $(m-1)/m$ precisely once until, after $m - 1$ steps, z_0 comes up again. After that, the generator repeats itself, producing the same $m - 1$ numbers in the same order each time.

Unfortunately, many random number generators, whether or not they are of the linear congruential variety, perform poorly. The random numbers they generate may fail to be independent in all sorts of ways, and the period may be relatively short. In the case of multiplicative congruential generators, this means that λ and m have not been chosen properly. See Gentle (1998) and the other references cited above for discussion of bad random number generators. Toy examples of multiplicative congruential generators are examined in Exercise 4.13, where the choice of λ and m is seen to matter.

There are several ways to generate drawings from a normal distribution if we can generate random numbers from the $U(0,1)$ distribution. The simplest, but not the fastest, is to use the fact that, if η_i is distributed as $U(0,1)$, then $\Phi^{-1}(\eta_i)$ is distributed as $N(0,1)$; this follows from the result of Exercise 4.14.

Most of the random number generators available in econometrics software packages use faster algorithms to generate drawings from the standard normal distribution, usually in a way entirely transparent to the user, who merely has to ask for so many independent drawings from $N(0, 1)$. Drawings from $N(\mu, \sigma^2)$ can then be obtained by use of the formula (4.09).

Bootstrap Tests

Although pivotal test statistics do arise from time to time, most test statistics in econometrics are not pivotal. The vast majority of them are, however, asymptotically pivotal. If a test statistic has a known asymptotic distribution that does not depend on anything unobservable, as do t and F statistics under the relatively weak assumptions of Section 4.5, then it is certainly asymptotically pivotal. Even if it does not follow a known asymptotic distribution, a test statistic may be asymptotically pivotal.

A statistic that is not an exact pivot cannot be used for a Monte Carlo test. However, approximate P values for statistics that are only asymptotically pivotal, or even nonpivotal, can be obtained by a simulation method called the **bootstrap**. This method can be a valuable alternative to the large sample tests based on asymptotic theory that we discussed in the previous section. The term **bootstrap**, which was introduced to statistics by Efron (1979), is taken from the phrase "to pull oneself up by one's own bootstraps." Although the link between this improbable activity and simulated P values is tenuous at best, the term is by now firmly established. We will speak of **bootstrapping** in order to obtain **bootstrap samples**, from which we compute **bootstrap test statistics** that we use to perform **bootstrap tests** on the basis of **bootstrap *P* values**, and so on.

The difference between a Monte Carlo test and a bootstrap test is that for the former, the DGP is assumed to be known, whereas, for the latter, it is necessary to estimate a **bootstrap DGP** from which to draw the simulated samples. Unless the null hypothesis under test is a simple hypothesis, the DGP that generated the original data is unknown, and so it cannot be used to generate simulated data. The bootstrap DGP is an estimate of the unknown true DGP. The hope is that, if the bootstrap DGP is close, in some sense, to the true one, then data generated by the bootstrap DGP will be similar to data that would have been generated by the true DGP, if it were known. If so, then a simulated P value obtained by use of the bootstrap DGP will be close enough to the true P value to allow accurate inference.

Even for models as simple as the linear regression model, there are many ways to specify the bootstrap DGP. The key requirement is that it should satisfy the restrictions of the null hypothesis. If this is assured, then how well a bootstrap test performs in finite samples depends on how good an estimate the bootstrap DGP is of the process that would have generated the test statistic if the null hypothesis were true. In the next subsection, we discuss bootstrap DGPs for regression models.

Bootstrap DGPs for Regression Models

If the null and alternative hypotheses are regression models, the simplest approach is to estimate the model that corresponds to the null hypothesis and then use the estimates to generate the bootstrap samples, under the assumption that the error terms are normally distributed. We considered examples of such procedures in Section 1.3 and in Exercise 1.22.

Since bootstrapping is quite unnecessary in the context of the classical normal linear model, we will take for our example a linear regression model with normal errors, but with a lagged dependent variable among the regressors:

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{Z}_t\boldsymbol{\gamma} + \delta y_{t-1} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2), \tag{4.63}$$

where $\boldsymbol{X}_t$ and $\boldsymbol{\beta}$ each have $k_1 - 1$ elements, $\boldsymbol{Z}_t$ and $\boldsymbol{\gamma}$ each have k_2 elements, and the null hypothesis is that $\boldsymbol{\gamma} = \mathbf{0}$. Thus the model that represents the null is

$$y_t = \boldsymbol{X}_t\boldsymbol{\beta} + \delta y_{t-1} + u_t, \quad u_t \sim \text{NID}(0, \sigma^2). \tag{4.64}$$

The observations are assumed to be indexed in such a way that y_0 is observed, along with n observations on y_t, $\boldsymbol{X}_t$, and $\boldsymbol{Z}_t$ for $t = 1, \ldots, n$. By estimating the models (4.63) and (4.64) by OLS, we can compute the F statistic for $\boldsymbol{\gamma} = \mathbf{0}$, which we will call $\hat{\tau}$. Because the regression function contains a lagged dependent variable, however, the F test based on $\hat{\tau}$ will not be exact.

The model (4.64) is a fully specified parametric model, which means that each set of parameter values for $\boldsymbol{\beta}$, δ, and σ^2 defines just one DGP. The simplest type of bootstrap DGP for fully specified models is given by the **parametric bootstrap**. The first step in constructing a parametric bootstrap DGP is to estimate (4.64) by OLS, yielding the restricted estimates $\tilde{\boldsymbol{\beta}}$, $\tilde{\delta}$, and $\tilde{s}^2 \equiv \text{SSR}(\tilde{\boldsymbol{\beta}}, \tilde{\delta})/(n - k_1)$. Then the bootstrap DGP is given by

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + \tilde{\delta} y_{t-1}^* + u_t^*, \quad u_t^* \sim \text{NID}(0, \tilde{s}^2), \tag{4.65}$$

which is just the element of the model (4.64) characterized by the parameter estimates under the null, with stars to indicate that the data are simulated.

In order to draw a bootstrap sample from the bootstrap DGP (4.65), we first draw an n–vector $\boldsymbol{u}^*$ from the $N(\mathbf{0}, \tilde{s}^2\mathbf{I})$ distribution. The presence of a lagged dependent variable implies that the bootstrap samples must be constructed **recursively**. This is necessary because y_t^*, the t^{th} element of the bootstrap sample, must depend on y_{t-1}^* and not on y_{t-1} from the original data. The recursive rule for generating a bootstrap sample is

$$\begin{aligned}
y_1^* &= \boldsymbol{X}_1\tilde{\boldsymbol{\beta}} + \tilde{\delta} y_0 + u_1^* \\
y_2^* &= \boldsymbol{X}_2\tilde{\boldsymbol{\beta}} + \tilde{\delta} y_1^* + u_2^* \\
\vdots \quad & \quad \vdots \qquad \vdots \qquad \vdots \\
y_n^* &= \boldsymbol{X}_n\tilde{\boldsymbol{\beta}} + \tilde{\delta} y_{n-1}^* + u_n^*.
\end{aligned} \tag{4.66}$$

Notice that every bootstrap sample is conditional on the observed value of y_0. There are other ways of dealing with pre-sample values of the dependent variable, but this is certainly the most convenient, and it may, in many circumstances, be the only method that is feasible.

The rest of the procedure for computing a bootstrap P value is identical to the one for computing a simulated P value for a Monte Carlo test. For each of the B bootstrap samples, $\boldsymbol{y}_j^*$, a bootstrap test statistic τ_j^* is computed from $\boldsymbol{y}_j^*$ in just the same way as $\hat{\tau}$ was computed from the original data, $\boldsymbol{y}$. The bootstrap P value $\hat{p}^*(\hat{\tau})$ is then computed by formula (4.61).

A Nonparametric Bootstrap DGP

The parametric bootstrap procedure that we have just described, based on the DGP (4.65), does not allow us to relax the strong assumption that the error terms are normally distributed. How can we construct a satisfactory bootstrap DGP if we extend the models (4.63) and (4.64) to admit nonnormal errors? If we knew the true error distribution, whether or not it was normal, we could always generate the $\boldsymbol{u}^*$ from it. Since we do not know it, we will have to find some way to estimate this distribution.

Under the null hypothesis, the OLS residual vector $\tilde{\boldsymbol{u}}$ for the restricted model is a consistent estimator of the error vector $\boldsymbol{u}$. This is an immediate consequence of the consistency of the OLS estimator itself. In the particular case of model (4.64), we have for each t that

$$\operatorname*{plim}_{n\to\infty} \tilde{u}_t = \operatorname*{plim}_{n\to\infty} \big(y_t - \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} - \tilde{\delta}y_{t-1}\big) = y_t - \boldsymbol{X}_t\boldsymbol{\beta}_0 - \delta_0 y_{t-1} = u_t,$$

where $\boldsymbol{\beta}_0$ and δ_0 are the parameter values for the true DGP. This means that, if the u_t are mutually independent drawings from the error distribution, then so are the residuals $\tilde{u}_t$, asymptotically.

From the Fundamental Theorem of Statistics, we know that the empirical distribution function of the error terms is a consistent estimator of the unknown CDF of the error distribution. Because the residuals consistently estimate the errors, it follows that the EDF of the residuals is also a consistent estimator of the CDF of the error distribution. Thus, if we draw bootstrap error terms from the empirical distribution of the residuals, we are drawing them from a distribution that tends to the true error distribution as $n \to \infty$. This is completely analogous to using estimated parameters in the bootstrap DGP that tend to the true parameters as $n \to \infty$.

Drawing simulated error terms from the empirical distribution of the residuals is called **resampling**. In order to **resample the residuals**, all the residuals are, metaphorically speaking, thrown into a hat and then randomly pulled out one at a time, with replacement. Thus each bootstrap sample will contain some of the residuals exactly once, some of them more than once, and some of them not at all. Therefore, the value of each drawing must be the value of one of the residuals, with equal probability for each residual. This is precisely what we mean by the empirical distribution of the residuals.

To resample concretely rather than metaphorically, we can proceed as follows. First, we draw a random number η from the $U(0,1)$ distribution. Then we divide the interval $[0,1]$ into n subintervals of length $1/n$ and associate each of these subintervals with one of the integers between 1 and n. When η falls into the l^{th} subinterval, we choose the index l, and our random drawing is the l^{th} residual. Repeating this procedure n times yields a single set of bootstrap error terms drawn from the empirical distribution of the residuals.

As an example of how resampling works, suppose that $n = 10$, and the ten residuals are

$$6.45,\ 1.28,\ -3.48,\ 2.44,\ -5.17,\ -1.67,\ -2.03,\ 3.58,\ 0.74,\ -2.14.$$

Notice that these numbers sum to zero. Now suppose that, when forming one of the bootstrap samples, the ten drawings from the $U(0,1)$ distribution happen to be

$$0.631,\ 0.277,\ 0.745,\ 0.202,\ 0.914,\ 0.136,\ 0.851,\ 0.878,\ 0.120,\ 0.259.$$

This implies that the ten index values will be

$$7,\ 3,\ 8,\ 3,\ 10,\ 2,\ 9,\ 9,\ 2,\ 3.$$

Therefore, the error terms for this bootstrap sample will be

$$-2.03,\ -3.48,\ 3.58,\ -3.48,\ -2.14,\ 1.28,\ 0.74,\ 0.74,\ 1.28,\ -3.48.$$

Some of the residuals appear just once in this particular sample, some of them (numbers 2, 3, and 9) appear more than once, and some of them (numbers 1, 4, 5, and 6) do not appear at all. On average, however, each of the residuals will appear once in each of the bootstrap samples.

If we adopt this resampling procedure, we can write the bootstrap DGP as

$$y_t^* = \boldsymbol{X}_t\tilde{\boldsymbol{\beta}} + \tilde{\delta}y_{t-1}^* + u_t^*, \quad u_t^* \sim \text{EDF}(\tilde{\boldsymbol{u}}), \tag{4.67}$$

where $\text{EDF}(\tilde{\boldsymbol{u}})$ denotes the distribution that assigns probability $1/n$ to each of the elements of the residual vector $\tilde{\boldsymbol{u}}$. The DGP (4.67) is one form of what is usually called a **nonparametric bootstrap**, although, since it still uses the parameter estimates $\tilde{\boldsymbol{\beta}}$ and $\tilde{\delta}$, it should really be called **semiparametric** rather than nonparametric. Once bootstrap error terms have been drawn by resampling, bootstrap samples can be created by the recursive procedure (4.66).

The empirical distribution of the residuals may fail to satisfy some of the properties that the null hypothesis imposes on the true error distribution, and so the DGP (4.67) may fail to belong to the null hypothesis. One case in which this failure has grave consequences arises when the regression (4.64) does not contain a constant term, because then the sample mean of the residuals is not, in general, equal to 0. The expectation of the EDF of the residuals is simply their sample mean; recall Exercise 1.1. Thus, if the bootstrap error terms are drawn from a distribution with nonzero mean, the bootstrap DGP lies outside the null hypothesis. It is, of course, simple to correct this problem. We just need to *center* the residuals before throwing them into the hat, by subtracting their mean $\bar{u}$. When we do this, the bootstrap errors are drawn from $\text{EDF}(\tilde{\boldsymbol{u}} - \bar{u}\boldsymbol{\iota})$, a distribution that does indeed have mean 0.

A somewhat similar argument gives rise to an improved bootstrap DGP. If the sample mean of the restricted residuals is 0, then the variance of their empirical distribution is the second moment $n^{-1}\sum_{t=1}^{n}\tilde{u}_t^2$. Thus, by using the definition (3.49) of $\tilde{s}^2$ in Section 3.6, we see that the variance of the empirical distribution of the residuals is $\tilde{s}^2(n-k_1)/n$. Since we do not know the value of σ_0^2, we cannot draw from a distribution with exactly that variance. However, as with the parametric bootstrap (4.65), we can at least draw from a distribution with variance $\tilde{s}^2$. This is easy to do by drawing from the EDF of the **rescaled residuals**, which are obtained by multiplying the OLS residuals by $(n/(n-k_1))^{1/2}$. If we resample these rescaled residuals, the bootstrap error distribution is

$$\text{EDF}\left(\left(\frac{n}{n-k_1}\right)^{1/2}\tilde{\boldsymbol{u}}\right), \tag{4.68}$$

which has variance $\tilde{s}^2$. A somewhat more complicated approach, based on the result (3.44), is explored in Exercise 4.15.

Although they may seem strange, these resampling procedures often work astonishingly well, except perhaps when the sample size is very small or the distribution of the error terms is very unusual; see Exercise 4.18. If the distribution of the error terms displays substantial skewness (that is, a nonzero third moment) or excess kurtosis (that is, a fourth moment greater than $3\sigma_0^4$), then there is a good chance that the EDF of the recentered and rescaled residuals will do so as well.

Other methods for bootstrapping regression models nonparametrically and semiparametrically are discussed by Efron and Tibshirani (1993), Davison and Hinkley (1997), and Horowitz (2001), which also discuss many other aspects of the bootstrap. A more advanced book, which deals primarily with the relationship between asymptotic theory and the bootstrap, is Hall (1992).

How Many Bootstraps?

Suppose that we wish to perform a bootstrap test at level α. Then B should be chosen to satisfy the condition that $\alpha(B+1)$ is an integer. If $\alpha = .05$, the values of B that satisfy this condition are 19, 39, 59, and so on. If $\alpha = .01$, they are 99, 199, 299, and so on. It is illuminating to see why B should be chosen in this way.

Imagine that we sort the original test statistic $\hat{\tau}$ and the B bootstrap statistics τ_j^*, $j = 1, \ldots, B$, in decreasing order. If τ is pivotal, then, under the null hypothesis, these are all independent drawings from the same distribution. Thus the rank r of $\hat{\tau}$ in the sorted set can have $B + 1$ possible values, $r = 0, 1, \ldots, B$, all of them equally likely under the null hypothesis if τ is pivotal. Here, r is defined in such a way that there are exactly r simulations for which $\tau_j^* > \hat{\tau}$. Thus, if $r = 0$, $\hat{\tau}$ is the largest value in the set, and if $r = B$, it is the smallest. The estimated P value $\hat{p}^*(\hat{\tau})$ is just r/B.

The bootstrap test rejects if $r/B < \alpha$, that is, if $r < \alpha B$. Under the null, the probability that this inequality will be satisfied is the proportion of the $B + 1$ possible values of r that satisfy it. If we denote by $[\alpha B]$ the largest integer that is smaller than αB, it is easy to see that there are exactly $[\alpha B]+1$ such values of r, namely, $0, 1, \ldots, [\alpha B]$. Thus the probability of rejection is $([\alpha B] + 1)/(B + 1)$. If we equate this probability to α, we find that

$$\alpha(B + 1) = [\alpha B] + 1.$$

Since the right-hand side of this equality is the sum of two integers, this equality can hold only if $\alpha(B+1)$ is an integer. Moreover, it will hold whenever $\alpha(B + 1)$ is an integer. Therefore, the Type I error will be precisely α if and only if $\alpha(B + 1)$ is an integer. Although this reasoning is rigorous only if τ is an exact pivot, experience shows that bootstrap P values based on nonpivotal statistics are less misleading if $\alpha(B + 1)$ is an integer.

As a concrete example, suppose that $\alpha = .05$ and $B = 99$. Then there are 5 out of 100 values of r, namely, $r = 0, 1, \ldots, 4$, that would lead us to reject the null hypothesis. Since these are equally likely if the test statistic is pivotal, we will make a Type I error precisely 5% of the time, and the test will be exact. But suppose instead that $B = 89$. Since the same 5 values of r would still lead us to reject the null, we would now do so with probability $5/90 = 0.0556$.

It is important that B be sufficiently large, since two problems can arise if it is not. The first problem is that the outcome of the test will depend on the sequence of random numbers used to generate the bootstrap samples. Different investigators may therefore obtain different results, even though they are using the same data and testing the same hypothesis. The second problem, which we will discuss in the next section, is that the ability of a bootstrap test to reject a false null hypothesis declines as B becomes smaller. As a rule of thumb, we suggest choosing $B = 999$. If calculating the τ_j^* is inexpensive and the outcome of the test is at all ambiguous, it may be desirable to use a larger value, like 9999. On the other hand, if calculating the τ_j^* is very expensive and the outcome of the test is unambiguous, because $\hat{p}^*$ is far from α, it may be safe to use a value as small as 99.

It is not actually necessary to choose B in advance. An alternative approach, which is a bit more complicated but can save a lot of computer time, has been proposed by Davidson and MacKinnon (2000). The idea is to calculate a sequence of estimated P values, based on increasing values of B, and to stop as soon as the estimate $\hat{p}^*$ allows us to be very confident that p^* is either greater or less than α. For example, we might start with $B = 99$, then perform an additional 100 simulations if we cannot be sure whether or not to reject the null hypothesis, then perform an additional 200 simulations if we still cannot be sure, and so on. Eventually, we either stop when we are confident that the null hypothesis should or should not be rejected, or when B has become so large that we cannot afford to continue.

Bootstrap versus Asymptotic Tests

Although bootstrap tests based on test statistics that are merely asymptotically pivotal are not exact, there are strong theoretical reasons to believe that they will generally perform better than tests based on approximate asymptotic distributions. The errors committed by both asymptotic and bootstrap tests diminish as n increases, but those committed by bootstrap tests diminish more rapidly. The fundamental theoretical result on this point is due to Beran (1988). The results of a number of Monte Carlo experiments have provided strong support for this proposition. References include Horowitz (1994), Godfrey (1998), and Davidson and MacKinnon (1999a, 1999b, 2002a).

We can illustrate this by means of an example. Consider the following simple special case of the linear regression model (4.63)

$$y_t = \beta_1 + \beta_2 X_t + \beta_3 y_{t-1} + u_t, \quad u_t \sim N(0, \sigma^2), \tag{4.69}$$

where the null hypothesis is that $\beta_3 = 0.9$. A Monte Carlo experiment to investigate the properties of tests of this hypothesis would work as follows. First, we fix a DGP in the model (4.69) by choosing values for the parameters. Here $\beta_3 = 0.9$, and so we investigate only what happens under the null hypothesis. For each **replication**, we generate an artificial data set from our chosen DGP and compute the ordinary t statistic for $\beta_3 = 0.9$. We then compute three P values. The first of these, for the asymptotic test, is computed using the Student's t distribution with $n - 3$ degrees of freedom, and the other two are bootstrap P values from the parametric and semiparametric bootstraps, with residuals rescaled using (4.68), for $B = 199$.[5] We perform many replications and record the frequencies with which tests based on the three P values reject at the .05 level. Figure 4.8 shows the rejection frequencies based on 500,000 replications for each of 31 sample sizes: $n = 10, 12, 14, \ldots, 60$.

The results of this experiment are striking. The asymptotic test overrejects quite noticeably, although it gradually improves as n increases. In contrast,

[5] We used $B = 199$, a smaller value than we would ever recommend using in practice, in order to reduce the costs of doing the Monte Carlo experiments. Because experimental errors tend to cancel out across replications, this does not materially affect the results of the experiments.

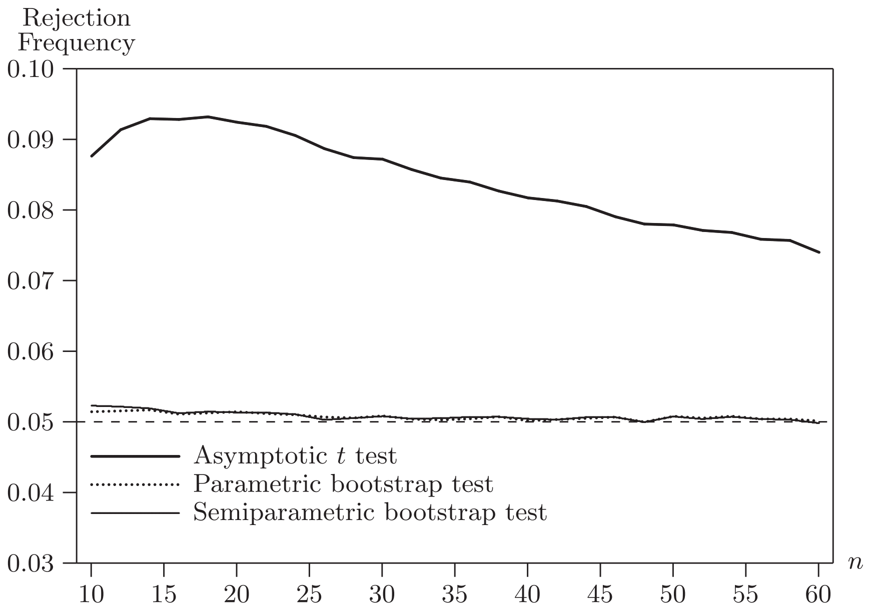

Figure 4.8 Rejection frequencies for bootstrap and asymptotic tests

the two bootstrap tests overreject only very slightly. Their rejection frequencies are always very close to the nominal level of .05, and they approach that level quite quickly as n increases. For the very smallest sample sizes, the parametric bootstrap seems to outperform the semiparametric one, but, for most sample sizes, there is nothing to choose between them.

This example is, perhaps, misleading in one respect. For linear regression models, asymptotic t and F tests generally do not perform as badly as the asymptotic t test does here. For example, the t test for $\beta_3 = 0$ in (4.69) performs much better than the t test for $\beta_3 = 0.9$; it actually underrejects moderately in small samples. However, the example is not at all misleading in suggesting that bootstrap tests will often perform extraordinarily well, even when the corresponding asymptotic test does not perform well at all.

4.7 The Power of Hypothesis Tests

To be useful, hypothesis tests must be able to discriminate between the null hypothesis and the alternative. Thus, as we saw in Section 4.2, the distribution of a useful test statistic under the null is different from its distribution when the DGP does not belong to the null. Whenever a DGP places most of the probability mass of the test statistic in the rejection region of a test, the test will have high **power**, that is, a high probability of rejecting the null.

For a variety of reasons, it is important to know something about the power of the tests we employ. If a test with high power fails to reject the null, this tells us more than if a test with lower power fails to do so. In practice, more than one test of a given null hypothesis is usually available. Of two equally reliable tests, if one has more power than the other against the alternatives in which we are interested, then we would surely prefer to employ the more powerful one.

The Power of Exact Tests

In Section 4.4, we saw that an F statistic is a ratio of the squared norms of two vectors, each divided by its appropriate number of degrees of freedom. In the notation of that section, these vectors are, for the numerator, $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{y}$, and, for the denominator, $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$. If the null and alternative hypotheses are classical normal linear models, as we assume throughout this subsection, then, under the null, both the numerator and the denominator of this ratio are independent χ^2 variables, divided by their respective degrees of freedom; recall (4.34). Under the alternative hypothesis, the distribution of the denominator is unchanged, because, under either hypothesis, $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$. Consequently, the difference in distribution under the null and the alternative that gives the test its power must come from the numerator alone.

From (4.33), r/σ^2 times the numerator of the F statistic $F_{\boldsymbol{\beta}_2}$ is

$$\frac{1}{\sigma^2}\boldsymbol{y}^{\top}\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{y}. \tag{4.70}$$

The vector $\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{y}$ is normal under both the null and the alternative. Its mean is $\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2$, which vanishes under the null when $\boldsymbol{\beta}_2 = \boldsymbol{0}$, and its covariance matrix is $\sigma^2\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{X}_2$. We can use these facts to determine the distribution of the quadratic form (4.70). To do so, we must introduce the **noncentral chi-squared distribution**, which is a generalization of the ordinary, or **central**, chi-squared distribution.

We saw in Section 4.3 that, if the m–vector $\boldsymbol{z}$ is distributed as $N(\boldsymbol{0}, \mathbf{I})$, then $\|\boldsymbol{z}\|^2 = \boldsymbol{z}^{\top}\boldsymbol{z}$ is distributed as (central) chi-squared with m degrees of freedom. Similarly, if $\boldsymbol{x} \sim N(\boldsymbol{0}, \boldsymbol{\Omega})$, then $\boldsymbol{x}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{x} \sim \chi^2(m)$. If instead $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \mathbf{I})$, then $\boldsymbol{z}^{\top}\boldsymbol{z}$ follows the noncentral chi-squared distribution with m degrees of freedom and **noncentrality parameter**, or **NCP**, $\Lambda \equiv \boldsymbol{\mu}^{\top}\boldsymbol{\mu}$. This distribution is written as $\chi^2(m, \Lambda)$. It is easy to see that its expectation is $m + \Lambda$; see Exercise 4.17. Likewise, if $\boldsymbol{x} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$, then $\boldsymbol{x}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{x} \sim \chi^2(m, \boldsymbol{\mu}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{\mu})$. Although we will not prove it, the distribution depends on $\boldsymbol{\mu}$ and $\boldsymbol{\Omega}$ only through the quadratic form $\boldsymbol{\mu}^{\top}\boldsymbol{\Omega}^{-1}\boldsymbol{\mu}$. If we set $\boldsymbol{\mu} = \boldsymbol{0}$, we see that the $\chi^2(m, 0)$ distribution is just the central $\chi^2(m)$ distribution.

Under either the null or the alternative hypothesis, therefore, the distribution of expression (4.70) is noncentral chi-squared, with r degrees of freedom, and with noncentrality parameter given by

$$\Lambda \equiv \frac{1}{\sigma^2}\boldsymbol{\beta}_2^{\top}\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 = \frac{1}{\sigma^2}\boldsymbol{\beta}_2^{\top}\boldsymbol{X}_2^{\top}\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2.$$

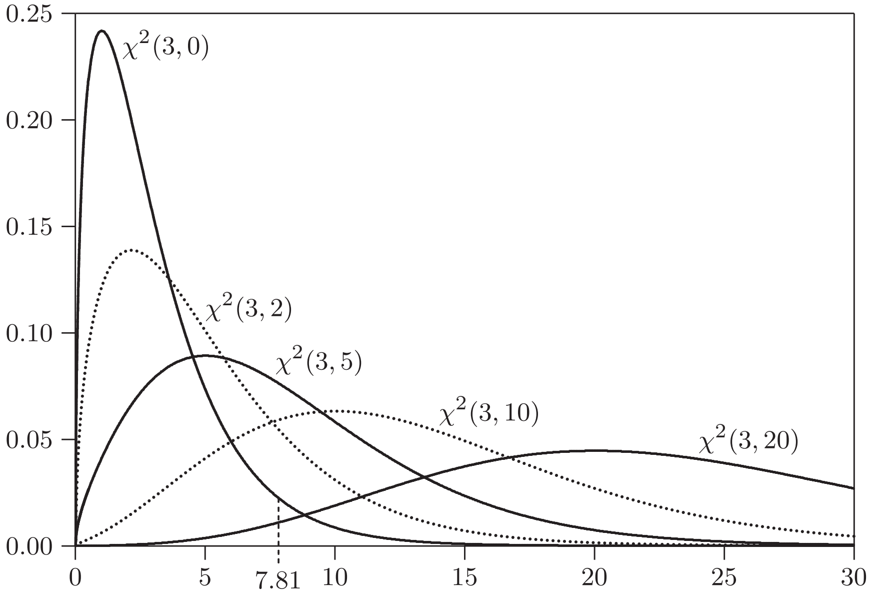

Figure 4.9 Densities of noncentral χ^2 distributions

Under the null, $\Lambda = 0$. Under either hypothesis, the distribution of the denominator of the F statistic, divided by σ^2, is central chi-squared with $n-k$ degrees of freedom, and it is independent of the numerator. The F statistic therefore has a distribution that we can write as

$$\frac{\chi^2(r, \Lambda)/r}{\chi^2(n-k)/(n-k)},$$

with numerator and denominator mutually independent. This distribution is called the **noncentral *F* distribution**, with r and $n-k$ degrees of freedom and noncentrality parameter Λ. In any given testing situation, r and $n-k$ are given, and so the difference between the distributions of the F statistic under the null and under the alternative depends only on the NCP Λ.

To illustrate this, we limit our attention to the expression (4.70), which is distributed as $\chi^2(r, \Lambda)$. As Λ increases, the distribution moves to the right and becomes more spread out. This is illustrated in Figure 4.9, which shows the density of the noncentral χ^2 distribution with 3 degrees of freedom for noncentrality parameters of 0, 2, 5, 10, and 20. The .05 critical value for the central $\chi^2(3)$ distribution, which is 7.81, is also shown. If a test statistic has the noncentral $\chi^2(3)$ distribution, the probability that the null hypothesis will be rejected at the .05 level is the probability mass to the right of 7.81. It is evident from the figure that this probability will be small for small values of the NCP and large for large ones.

In Figure 4.9, the number of degrees of freedom r is held constant as Λ is increased. If, instead, we held Λ constant, the density functions would move to the right as r was increased, as they do in Figure 4.4 for the special case with $\Lambda = 0$. Thus, at any given level, the critical value of a χ^2 or F test will increase as r increases. It has been shown by Das Gupta and Perlman (1974) that this rightward shift of the critical value has a greater effect than the rightward shift of the density for any positive Λ. Specifically, Das Gupta and Perlman show that, for a given NCP, the power of a χ^2 or F test at any given level is strictly decreasing in r, as well as being strictly increasing in Λ, as we indicated in the previous paragraph.

The square of a t statistic for a single restriction is just the F test for that restriction, and so the above analysis applies equally well to t tests. Things can be made a little simpler, however. From (4.25), the t statistic t_{β_2} is $1/s$ times

$$\frac{\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}}{(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}}. \tag{4.71}$$

The numerator of this expression, $\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{y}$, is normally distributed under both the null and the alternative, with variance $\sigma^2 \boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2$ and mean $\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2 \beta_2$. Thus $1/\sigma$ times (4.71) is normal with variance 1 and mean

$$\lambda \equiv \frac{1}{\sigma}(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2} \beta_2. \tag{4.72}$$

It follows that t_{β_2} has a distribution which can be written as

$$\frac{N(\lambda, 1)}{\big(\chi^2(n-k)/(n-k)\big)^{1/2}},$$

with independent numerator and denominator. This distribution is known as the **noncentral t distribution**, with $n-k$ degrees of freedom and noncentrality parameter λ; it is written as $t(n-k, \lambda)$. Note that $\lambda^2 = \Lambda$, where Λ is the NCP of the corresponding F test. Except for very small sample sizes, the $t(n-k, \lambda)$ distribution is quite similar to the $N(\lambda, 1)$ distribution. It is also very much like an ordinary, or **central**, t distribution with its mean shifted from the origin to (4.72), but it has a bit more variance, because of the stochastic denominator.

When we know the distribution of a test statistic under the alternative hypothesis, we can determine the power of a test of given level as a function of the parameters of that hypothesis. This function is called the **power function** of the test. The distribution of t_{β_2} under the alternative depends only on the NCP λ. For a given regressor matrix $\boldsymbol{X}$ and sample size n, λ in turn depends on the parameters only through the ratio β_2/σ; see (4.72). Therefore, the power of the t test depends only on this ratio. According to assumption (4.49), as $n \to \infty$, $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ tends to a nonstochastic limiting matrix $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$. Thus, as n increases, the factor $(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}$ will be roughly proportional to $n^{1/2}$, and so λ will tend to infinity with n at a rate similar to that of $n^{1/2}$.

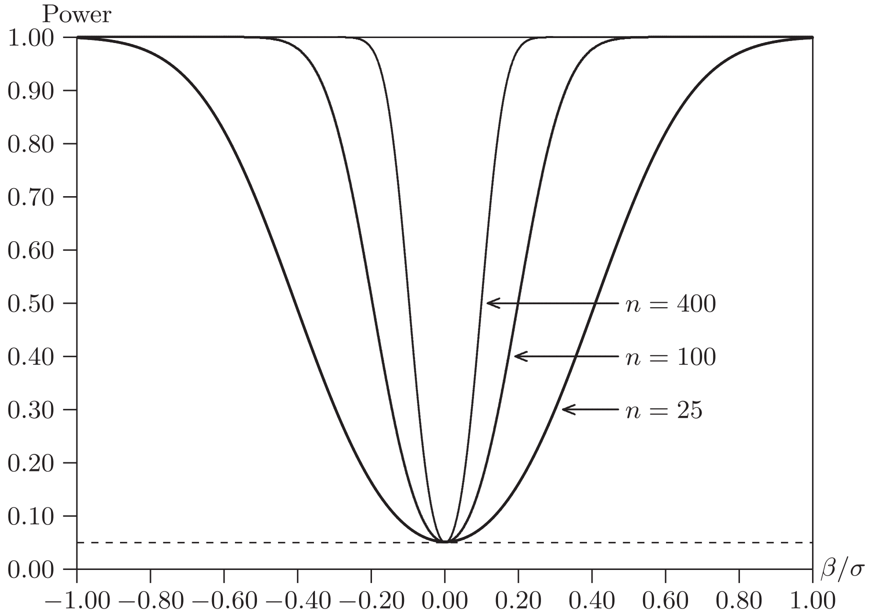

Figure 4.10 Power functions for t tests at the .05 level

Figure 4.10 shows power functions for a very simple model, in which $\boldsymbol{x}_2$, the only regressor, is a constant. Power is plotted as a function of β_2/σ for three sample sizes: $n = 25$, $n = 100$, and $n = 400$. Since the test is exact, all the power functions are equal to .05 when $\beta = 0$. Power then increases as β moves away from 0. As we would expect, the power when $n = 400$ exceeds the power when $n = 100$, which in turn exceeds the power when $n = 25$, for every value of $\beta \neq 0$. It is clear that, as $n \to \infty$, the power function will converge to the shape of a T, with the foot of the vertical segment at .05 and the horizontal segment at 1.0. Thus, asymptotically, the test will reject the null with probability 1 whenever it is false. In finite samples, however, we can see from the figure that a false hypothesis is very unlikely to be rejected if $n^{1/2}\beta/\sigma$ is sufficiently small.

The Power of Bootstrap Tests

As we remarked in Section 4.6, the power of a bootstrap test depends on B, the number of bootstrap samples. The reason why it does so is illuminating. If, to any test statistic, we add random noise independent of the statistic, we inevitably reduce the power of tests based on that statistic. The bootstrap P value $\hat{p}^*(\hat{\tau})$ defined in (4.61) is simply an estimate of the **ideal bootstrap *P* value**

$$p^*(\hat{\tau}) \equiv \Pr(\tau > \hat{\tau}) = \operatorname*{plim}_{B\to\infty} \hat{p}^*(\hat{\tau}),$$

where $\Pr(\tau > \hat{\tau})$ is evaluated under the bootstrap DGP. When B is finite, $\hat{p}^*$ will differ from p^* because of random variation in the bootstrap samples. This

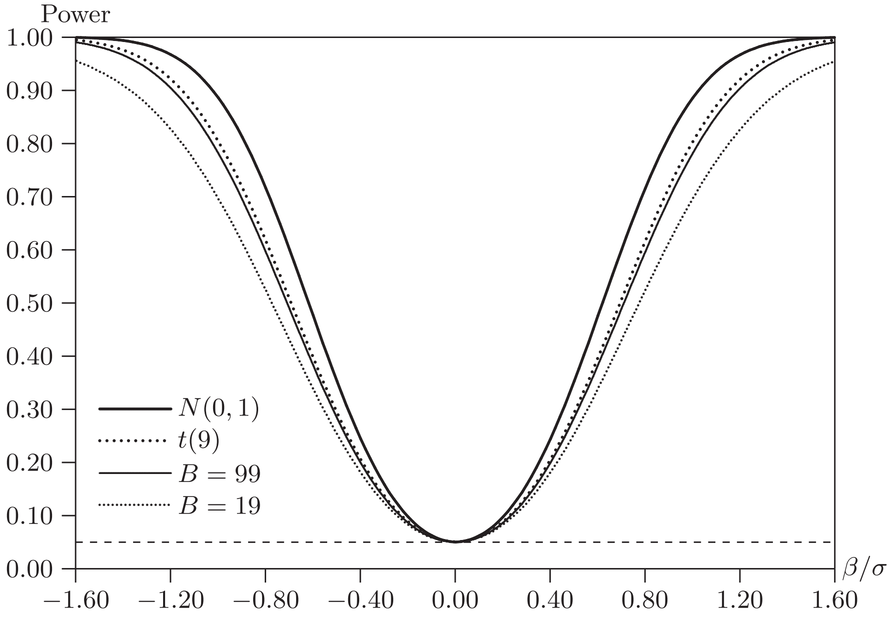

Figure 4.11 Power functions for tests at the .05 level

random variation is generated in the computer, and is therefore completely independent of the random variable τ. The bootstrap testing procedure discussed in Section 4.6 incorporates this random variation, and in so doing it reduces the power of the test.

Another example of how randomness affects test power is provided by the tests z_{β_2} and t_{β_2}, which were discussed in Section 4.4. Recall that z_{β_2} follows the $N(0,1)$ distribution, because σ is known, and t_{β_2} follows the $t(n-k)$ distribution, because σ has to be estimated. As equation (4.26) shows, t_{β_2} is equal to z_{β_2} times the random variable σ/s, which has the same distribution under the null and alternative hypotheses, and is independent of z_{β_2}. Therefore, multiplying z_{β_2} by σ/s simply adds independent random noise to the test statistic. This additional randomness requires us to use a larger critical value, and that in turn causes the test based on t_{β_2} to be less powerful than the test based on z_{β_2}.

Both types of power loss are illustrated in Figure 4.11. It shows power functions for four tests at the .05 level of the null hypothesis that $\beta = 0$ in the model (4.01) with normally distributed error terms and 10 observations. All four tests are exact, as can be seen from the fact that, in all cases, power equals .05 when $\beta = 0$. For all values of $\beta \neq 0$, there is a clear ordering of the four curves in Figure 4.11. The highest curve is for the test based on z_{β_2}, which uses the $N(0,1)$ distribution and is available only when σ is known. The next three curves are for tests based on t_{β_2}. The loss of power from using t_{β_2} with the $t(9)$ distribution, instead of z_{β_2} with the $N(0,1)$ distribution, is quite noticeable. Of course, 10 is a very small sample size; the loss of power from not knowing σ would be very much less for more reasonable sample sizes. There is a further loss of power from using a bootstrap test with finite B. This further loss is quite modest when $B = 99$, but it is substantial when $B = 19$.

Figure 4.11 suggests that the loss of power from using bootstrap tests is generally modest, except when B is very small. However, readers should be warned that the loss can be more substantial in other cases. A reasonable rule of thumb is that power loss will very rarely be a problem when $B = 999$, and that it will never be a problem when $B = 9999$.

4.8 Final Remarks

This chapter has introduced a number of important concepts, which we will encounter again and again throughout this book. In particular, we will encounter many types of hypothesis test, sometimes exact but more commonly asymptotic. Some of the asymptotic tests work well in finite samples, but others do not. Many of them can easily be bootstrapped, and they will perform much better when bootstrapped, but others are difficult to bootstrap or do not perform particularly well.

Although hypothesis testing plays a central role in classical econometrics, it is not the only method by which econometricians attempt to make inferences from parameter estimates about the true values of parameters. In the next chapter, we turn our attention to the other principal method, namely, the construction of confidence intervals and confidence regions.

4.9 Exercises

4.1 Suppose that the random variable z follows the $N(0,1)$ density. If z is a test statistic used in a two-tailed test, the corresponding P value, according to (4.07), is $p(z) \equiv 2(1 - \Phi(|z|))$. Show that $F_p(\cdot)$, the CDF of $p(z)$, is the CDF of the uniform distribution on $[0,1]$. In other words, show that

$$F_p(x) = x \quad \text{for all } x \in [0,1].$$

4.2 Extend Exercise 1.6 to show that the third and fourth moments of the standard normal distribution are 0 and 3, respectively. Use these results in order to calculate the centered and uncentered third and fourth moments of the $N(\mu, \sigma^2)$ distribution.

4.3 Let the density of the random variable x be $f(x)$. Show that the density of the random variable $w \equiv tx$, where $t > 0$, is $t^{-1}f(w/t)$. Next let the joint density of the set of random variables x_i, $i = 1, \ldots, m$, be $f(x_1, \ldots, x_m)$. For $i = 1, \ldots, m$, let $w_i = t_i x_i$, $t_i > 0$. Show that the joint density of the w_i is

$$f(w_1, \ldots, w_m) = \frac{1}{\prod_{i=1}^m t_i} f\Big(\frac{w_1}{t_1}, \ldots, \frac{w_m}{t_m}\Big).$$

4.4 Consider the random variables x_1 and x_2, which are bivariate normal with $x_1 \sim N(0, \sigma_1^2)$, $x_2 \sim N(0, \sigma_2^2)$, and correlation ρ. Show that the expectation of x_1 conditional on x_2 is $\rho(\sigma_1/\sigma_2)x_2$ and that the variance of x_1 conditional on x_2 is $\sigma_1^2(1-\rho^2)$. How are these results modified if the means of x_1 and x_2 are μ_1 and μ_2, respectively?

4.5 Suppose that, as in the previous question, the random variables x_1 and x_2 are bivariate normal, with means 0, variances σ_1^2 and σ_2^2, and correlation ρ. Starting from (4.13), show that $f(x_1, x_2)$, the joint density of x_1 and x_2, is given by

$$\frac{1}{2\pi}\frac{1}{(1-\rho^2)^{1/2}\sigma_1\sigma_2}\exp\biggl(\frac{-1}{2(1-\rho^2)}\Bigl(\frac{x_1^2}{\sigma_1^2} - 2\rho\frac{x_1x_2}{\sigma_1\sigma_2} + \frac{x_2^2}{\sigma_2^2}\Bigr)\biggr).$$

Then use this result to show that x_1 and x_2 are statistically independent if $\rho = 0$.

4.6 Consider the linear regression model

$$y_t = \beta_1 + \beta_2 X_{t1} + \beta_3 X_{t2} + u_t.$$

Rewrite this model so that the restriction $\beta_2 - \beta_3 = 1$ becomes a single zero restriction.

4.7 Consider the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, where there are n observations and k regressors. Suppose that this model is potentially subject to r restrictions which can be written as $\boldsymbol{R\beta} = \boldsymbol{r}$, where $\boldsymbol{R}$ is an $r \times k$ matrix and $\boldsymbol{r}$ is an r–vector. Rewrite the model so that the restrictions become r zero restrictions.

4.8 Show that the t statistic (4.25) is $(n-k)^{1/2}$ times the cotangent of the angle between the n–vectors $\boldsymbol{M}_1\boldsymbol{y}$ and $\boldsymbol{M}_1\boldsymbol{x}_2$.

Now consider the regressions

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}, \text{ and} \\ \boldsymbol{x}_2 &= \boldsymbol{X}_1\boldsymbol{\gamma}_1 + \gamma_2\boldsymbol{y} + \boldsymbol{v}. \end{aligned} \tag{4.73}$$

What is the relationship between the t statistic for $\beta_2 = 0$ in the first of these regressions and the t statistic for $\gamma_2 = 0$ in the second?

4.9 Show that the OLS estimates $\tilde{\boldsymbol{\beta}}_1$ from the model (4.29) can be obtained from those of model (4.28) by the formula

$$\tilde{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_1 + (\boldsymbol{X}_1^{\top}\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^{\top}\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2.$$

Formula (4.38) is useful for this exercise.

4.10 Show that the SSR from regression (4.42), or equivalently, regression (4.41), is equal to the sum of the SSRs from the two subsample regressions:

$$\begin{aligned} \boldsymbol{y}_1 &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{u}_1, \quad \boldsymbol{u}_1 \sim N(\boldsymbol{0}, \sigma^2\mathbf{I}), \text{ and} \\ \boldsymbol{y}_2 &= \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}_2, \quad \boldsymbol{u}_2 \sim N(\boldsymbol{0}, \sigma^2\mathbf{I}). \end{aligned}$$

4.11 When performing a Chow test, one may find that one of the subsamples is smaller than k, the number of regressors. Without loss of generality, assume that $n_2 < k$. Show that, in this case, the F statistic becomes

$$\frac{(\text{RSSR} - \text{SSR}_1)/n_2}{\text{SSR}_1/(n_1 - k)},$$

and that the numerator and denominator really have the degrees of freedom used in this formula.

4.12 Show, using the results of Section 4.5, that r times the F statistic (4.58) is asymptotically distributed as $\chi^2(r)$.

4.13 Consider a multiplicative congruential generator with modulus $m = 7$, and with all reasonable possible values of λ, that is, $\lambda = 2, 3, 4, 5, 6$. Show that, for any integer seed between 1 and 6, the generator generates each number of the form $i/7$, $i = 1, \ldots, 6$, exactly once before cycling for $\lambda = 3$ and $\lambda = 5$, but that it repeats itself more quickly for the other choices of λ. Repeat the exercise for $m = 11$, and determine which choices of λ yield generators that return to their starting point before covering the full range of possibilities.

4.14 If F is a strictly increasing CDF defined on an interval $[a, b]$ of the real line, where either or both of a and b may be infinite, then the inverse function F^{-1} is a well-defined mapping from $[0, 1]$ on to $[a, b]$. Show that, if the random variable X is a drawing from the $U(0, 1)$ distribution, then $F^{-1}(X)$ is a drawing from the distribution of which F is the CDF.

4.15 In Section 3.6, we saw that $\text{Var}(\hat{u}_t) = (1 - h_t)\sigma_0^2$, where $\hat{u}_t$ is the t^{th} residual from the linear regression model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$, and h_t is the t^{th} diagonal element of the "hat matrix" $\boldsymbol{P_X}$; this was the result (3.44). Use this result to derive an alternative to (4.68) as a method of rescaling the residuals prior to resampling. Remember that the rescaled residuals must have mean 0.

4.16 Suppose that z is a test statistic distributed as $N(0, 1)$ under the null hypothesis, and as $N(\lambda, 1)$ under the alternative, where λ depends on the DGP that generates the data. If c_α is defined by (4.06), show that the power of the two-tailed test at level α based on z is equal to

$$\Phi(\lambda - c_\alpha) + \Phi(-c_\alpha - \lambda).$$

Plot this power function for λ in the interval $[-5, 5]$ for $\alpha = .05$ and $\alpha = .01$.

4.17 Show that, if the m–vector $\boldsymbol{z} \sim N(\boldsymbol{\mu}, \mathbf{I})$, the expectation of the noncentral chi-squared variable $\boldsymbol{z}^\top\boldsymbol{z}$ is $m + \boldsymbol{\mu}^\top\boldsymbol{\mu}$.

4.18 The file **classical.data** contains 50 observations on three variables: $\boldsymbol{y}$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$. These are artificial data generated from the classical linear regression model

$$\boldsymbol{y} = \beta_1\boldsymbol{\iota} + \beta_2\boldsymbol{x}_2 + \beta_3\boldsymbol{x}_3 + \boldsymbol{u}, \quad \boldsymbol{u} \sim N(\mathbf{0}, \sigma^2\mathbf{I}).$$

Compute a t statistic for the null hypothesis that $\beta_3 = 0$. On the basis of this test statistic, perform an exact test. Then perform parametric and semiparametric bootstrap tests using 99, 999, and 9999 simulations. How do the two types of bootstrap P values correspond with the exact P value? How does this correspondence change as B increases?

4.19 Consider again the data in the file **consumption.data** and the ADL model studied in Exercise 3.22, which is reproduced here for convenience:

$$c_t = \alpha + \beta c_{t-1} + \gamma_0 y_t + \gamma_1 y_{t-1} + u_t. \tag{3.70}$$

Compute a t statistic for the hypothesis that $\gamma_0 + \gamma_1 = 0$. On the basis of this test statistic, perform an asymptotic test, a parametric bootstrap test, and a semiparametric bootstrap test using residuals rescaled according to (4.68).