Chapter 4

Hypothesis Testing in Linear Regression Models

4.1 Introduction

As we saw in Section 3.2, 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 discuss hypothesis testing in this chapter and confidence intervals in the next one. We start with hypothesis testing because it typically plays a larger role than confidence intervals in applied econometrics and because it is essential to have a thorough grasp of hypothesis testing if the construction of confidence intervals is to be understood at anything more than a very superficial level.

In the next section, we develop the fundamental ideas of hypothesis testing in the context of a very simple special case. In Section 4.3, we review some of the properties of three important distributions which are related to the normal distribution and are commonly encountered in the context of hypothesis testing. This material is needed 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 develop the asymptotic theory of linear regression models. That theory is then used to study large-sample tests in Section 4.6.

The remainder of the chapter deals with more advanced topics. In Section 4.7, we discuss some of the rather tricky issues associated with performing two or more tests at the same time. In Section 4.8, we discuss the **power** of a test, that is, what determines the ability of a test to reject a hypothesis that is false. Finally, in Section 4.9, we introduce the important concept of **pretesting**, in which the results of a test are used to determine which of two or more estimators to use. We do not discuss bootstrap tests in this chapter. That very important topic will be dealt with in Chapter 6.

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 disturbance u_t. The least-squares estimator of β and its variance, for a sample of size n, are given by

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

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 $\text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ 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 the population mean 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

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

[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.

It turns out that, under the null hypothesis, z must be distributed as $\mathrm{N}(0,1)$. 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),

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

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 $\mathrm{N}(0,1)$ 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 $\hat{\gamma} \sim \mathrm{N}(0, \sigma^2/n)$. It follows that z is also normal (see Exercise 1.7 again), and we find from (4.03) that

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

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 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 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 generally simply called 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, then we reject the null hypothesis at level α. Otherwise, we do 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 rejects 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 depends 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 $\mathrm{N}(\lambda, 1)$ distribution away from the center of the $\mathrm{N}(0, 1)$ distribution and into its tails. This can be seen in Figure 4.1, which graphs the $\mathrm{N}(0, 1)$ density and the $\mathrm{N}(\lambda, 1)$

[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.

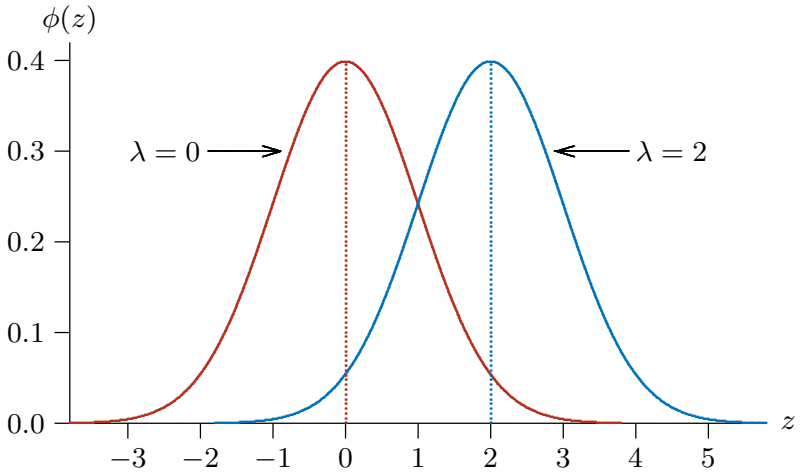

Figure 4.1 The normal distribution centered and uncentered

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 were 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 $\lambda = 0$.

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 increases 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 4.9.

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 $\mathrm{N}(0, 1)$, including the statistic z defined in (4.04), the critical value c_α is defined implicitly by the equation

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

Recall that Φ denotes the CDF of the standard normal distribution. Solving equation (4.05) for c_α in terms of the inverse function Φ^{-1}, we find that

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

According to (4.05), the probability that $z > c_\alpha$ is $1 - (1 - \alpha/2) = \alpha/2$. By symmetry, the probability that $z < -c_\alpha$ is also $\alpha/2$. 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, the critical value c_α increases as α

approaches 0. As an example, when $\alpha = .05$, we see from equation (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 equation (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 equation (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 $\mathrm{N}(0, 1)$ distribution into a new random variable $p(z)$ with the uniform $\mathrm{U}(0, 1)$ 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.

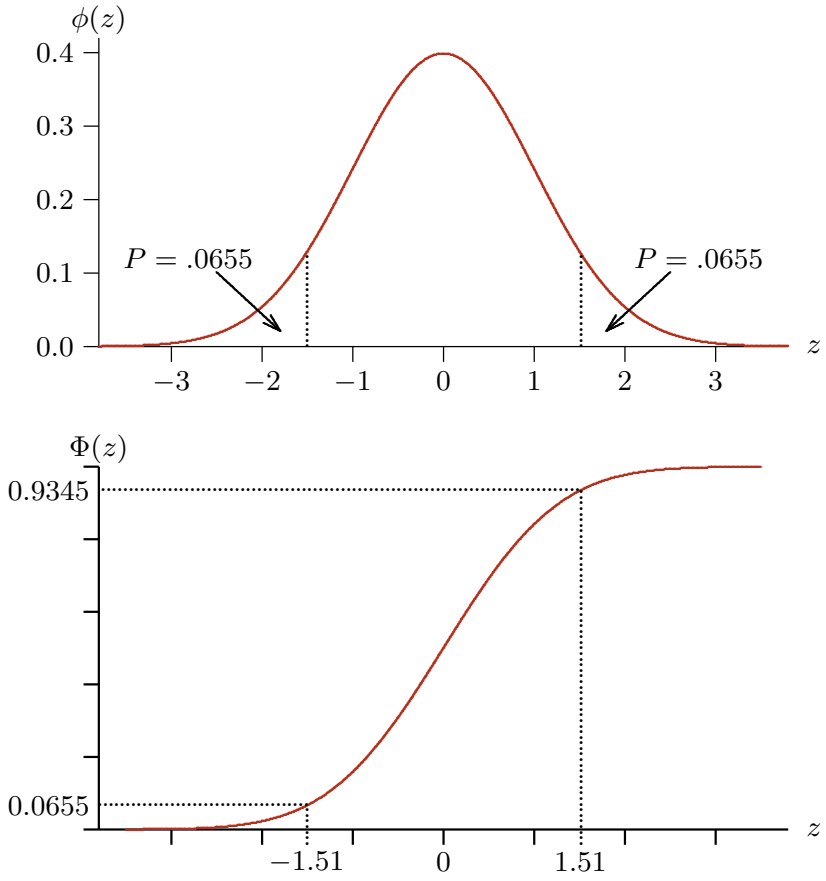

Figure 4.2 P values for a two-tailed test

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, but it is clear that a test based on $\hat{z}$ does 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 $\Pr(z < 1.51) = .9345$.

The P values discussed above, whether for one-tailed or two-tailed tests, are based on the symmetric $\mathrm{N}(0, 1)$ distribution. In Exercise 4.16, readers are asked to show how to compute P values for two-tailed tests based on an asymmetric distribution.

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 disturbances are normally distributed, and the second, which is grossly unrealistic, is that the variance of the disturbances 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 disturbance 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 4.9, 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 equation (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 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, as can be seen in the top panel of the figure, the density 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 $\mathrm{N}(\mu, \sigma^2)$. Thus the standard normal distribution is the $\mathrm{N}(0, 1)$ distribution. As readers were asked to show in Exercise 1.8, the PDF of the $\mathrm{N}(\mu, \sigma^2)$ 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 $\mathrm{N}(\mu, \sigma^2)$, we can invert equation (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 $\mathrm{N}(\mu, \sigma^2)$ 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 $w \equiv b_1 z_1 + b_2 z_2$. If we reason conditionally on z_1, we find that

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

[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.

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 $\mathrm{N}(0, 1)$.

The conditional variance of w is given by

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

where the last equality again follows because $z_2 \sim \mathrm{N}(0, 1)$. 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 $\mathrm{N}(b_1 z_1, b_2^2)$. 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 $z_1 \sim \mathrm{N}(0, 1)$, and so we see that the joint density is

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

For the moment, let us suppose that $b_1^2 + b_2^2 = 1$, although we will remove this restriction shortly. Then, if we use (1.06) to get an explicit expression for this joint density, we obtain

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

The right-hand side of equation (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 $w \sim \mathrm{N}(0, r^2)$. It is equally simple to extend the result to a linear combination of any number of mutually independent standard normal variables. Let w equal $b_1 z_1 + b_2 z_2 + b_3 z_3$, where z_1, z_2, and z_3 are mutually independent standard normal variables. Then $b_1 z_1 + b_2 z_2$ 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 $b_1 z_1 + b_2 z_2$ 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 $\boldsymbol{A}\boldsymbol{z}$, 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

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

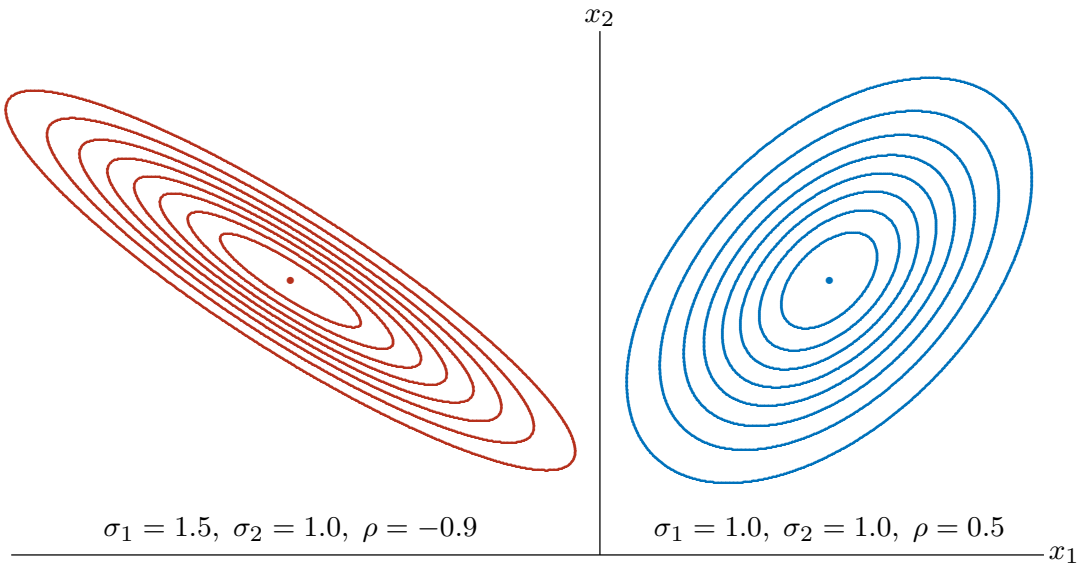

Figure 4.3 Contours of two bivariate normal densities

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 1.13.

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 matrix $\boldsymbol{A}$ 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 $\boldsymbol{x} \sim \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$. If we add an m–vector $\boldsymbol{\mu}$ of constants to $\boldsymbol{x}$, the resulting vector must follow the $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$ distribution.

It is clear from this argument that any linear combination of random variables that are jointly multivariate normal must itself be normally distributed. Thus, if $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, 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 above that $\boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$ 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 1.14 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 $\boldsymbol{z} \sim \mathrm{N}(\mathbf{0}, \mathbf{I})$. 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.

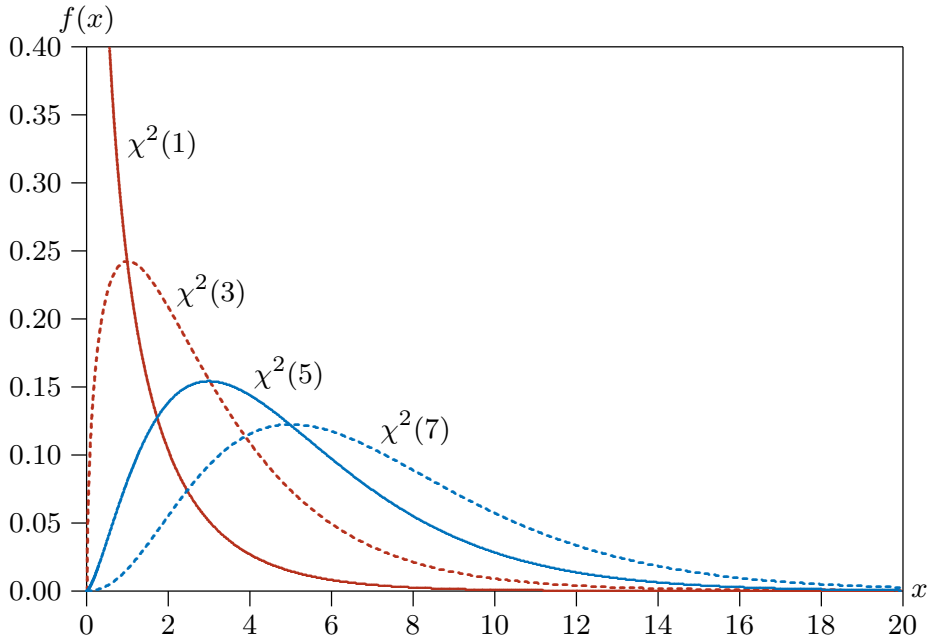

Figure 4.4 Various chi-squared PDFs

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 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 $\mathrm{N}(m, 2m)$ distribution as m becomes large.

In seclink34, 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 $\mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$, 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 $\mathrm{N}(\mathbf{0}, \mathbf{I})$, 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 $\mathrm{N}(\mathbf{0}, \mathbf{I})$.

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 $\mathrm{N}(\mathbf{0}, \boldsymbol{Z}^\top\boldsymbol{Z})$ 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 $z \sim \mathrm{N}(0,1)$ 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

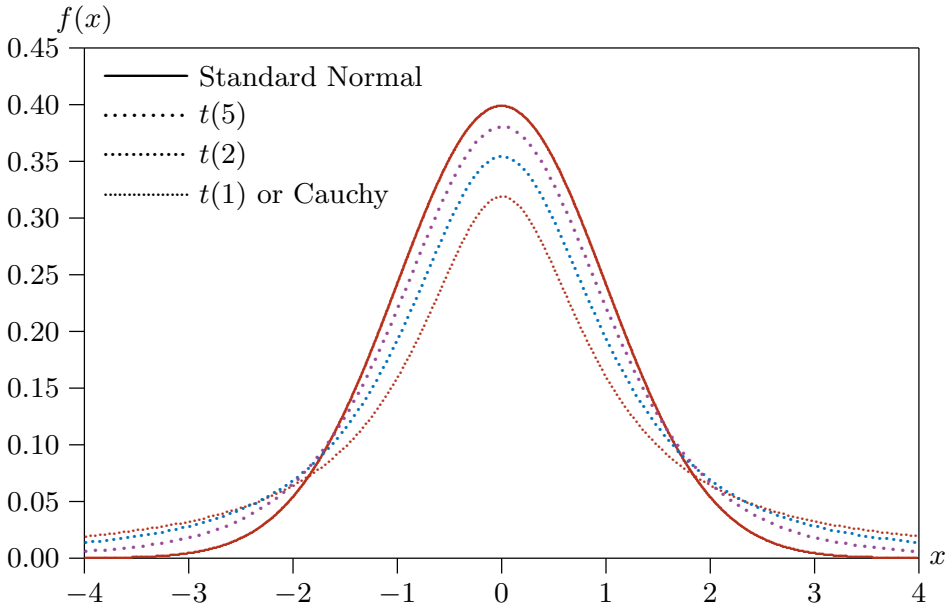

Figure 4.5 PDFs of the Student's t distribution

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 (3.22), 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 $t \to z \sim \text{N}(0,1)$ 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 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 $t(m)$ must 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: $F \sim F(m_1, m_2)$.[4] 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 $F(m_1, m_2)$ behaves 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 $t(m_2)$ is 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 is distributed as $\mathrm{N}(0, 1)$. 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

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

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 disturbance 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 $u_t \sim \mathrm{NID}(0, \sigma^2)$ 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[5] 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

[4] The F distribution was introduced by Snedecor (1934). The notation F is used in honor of the well-known statistician R. A. Fisher.

[5] 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."

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 4.7 and 4.8.

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

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

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

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

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

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

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 equation (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 $\mathrm{N}(0,1)$. 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

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

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 $\mathrm{N}(0,1)$ under the null hypothesis.

The test statistic z_{β_2} defined in equation (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 disturbances.

In order to handle the more realistic case in which the variance of the disturbances is unknown, we need to replace σ in equation (4.23) by s, the standard error of the regression (4.21), which was implicitly defined in equation (3.59). If, as usual, $\boldsymbol{M}_{\boldsymbol{X}}$ is the orthogonal projection on to $\mathcal{S}^\perp(\boldsymbol{X})$, then we have $s^2 = \boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}/(n-k)$, and so we obtain the test statistic

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

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 $\zeta/(n-k)$, where ζ is independent of z and distributed as $\chi^2(n-k)$. The t statistic defined in (4.25) can be rewritten as

$$t_{\beta_2} = \frac{z_{\beta_2}}{\left(\boldsymbol{y}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y} / ((n-k)\sigma^2) \right)^{1/2}}, \tag{4.26}$$

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

Under any DGP that belongs to (4.21),

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

where $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}/\sigma$ is distributed as $\mathrm{N}(\mathbf{0}, \mathbf{I})$. Since $\boldsymbol{M}_{\boldsymbol{X}}$ 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 $\boldsymbol{M_X} \boldsymbol{y}$. Second, from (4.23), it is not hard to see that z_{β_2} depends on $\boldsymbol{y}$ only through $\boldsymbol{P_X} \boldsymbol{y}$, since

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

the first equality here simply uses the fact that $\boldsymbol{x}_2 \in \mathcal{S}(\boldsymbol{X})$, and the third equality uses the result (2.35) that $\boldsymbol{P_X} \boldsymbol{P}_1 = \boldsymbol{P}_1 \boldsymbol{P_X}$. Independence now follows because, as we will see directly, $\boldsymbol{P_X} \boldsymbol{y}$ and $\boldsymbol{M_X} \boldsymbol{y}$ are independent.

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

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

by (2.25), because $\boldsymbol{P_X}$ and $\boldsymbol{M_X}$ are complementary projections. These zero covariances imply that the vectors $\boldsymbol{P_X} \boldsymbol{u}$ and $\boldsymbol{M_X} \boldsymbol{u}$ 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 $\mathrm{N}(0,1)$ distribution to compute P values or critical values. An interesting property of t statistics is explored in Exercise 4.9.

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 is 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

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

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 is 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

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

in which we impose the restriction that $\boldsymbol{\beta}_2 = \boldsymbol{0}$. 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 is almost always 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

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

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

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

Therefore,

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

and the F statistic (4.30) can be written as

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

In general, $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$. Under the null hypothesis, $\boldsymbol{M}_1\boldsymbol{y} = \boldsymbol{M}_1\boldsymbol{u}$. 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: $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_1\boldsymbol{X}_2 = \boldsymbol{M}_{\boldsymbol{X}}(\boldsymbol{X}_2 - \boldsymbol{P}_1\boldsymbol{X}_2) = \boldsymbol{O}$. 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 Exercises 2.18 and 2.19, $\boldsymbol{P}_{\boldsymbol{X}} - \boldsymbol{P}_1$ was seen to be an orthogonal projection matrix, equal to $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}$. It follows that

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

where the two projections on the right-hand side of this equation 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

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

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 ($\tilde{}$) to denote the **restricted estimates**, and a hat ($\hat{}$) 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, $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ is the vector of residuals from the unrestricted model.

Since $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2$, 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 4.10, 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, $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$, equals $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{u}$, and so it just 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 $\boldsymbol{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I})$, 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 must 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).

We want to reject the null whenever RSSR − USSR, the numerator of the F statistic, 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 r and $n-k$ degrees of freedom. Although we compute the P value as if for a one-tailed test, F tests are really two-tailed tests, because they test equality restrictions, not inequality restrictions. An F test for $\boldsymbol{\beta}_2 = \mathbf{0}$ rejects 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

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

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

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

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

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

where $\boldsymbol{M}_\iota$ is the projection matrix that takes deviations from the mean, which was defined in (2.31). Thus the matrix expression in the numerator of (4.40) is just the explained sum of squares, or ESS, from the FWL regression

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

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:

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

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

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

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 4.11, 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

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

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 Asymptotic Theory for 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 disturbance 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 that do not follow any known distribution exactly, they would in many cases approximately follow known distributions whenever the sample size were large enough. 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. In this section, we introduce several key results of asymptotic theory for linear regression models. These are then applied to large-sample tests in Section 4.6.

In general, asymptotic theory is concerned with the distributions of estimators and test statistics as the sample size n tends to infinity. Nevertheless, it often allows us to obtain simple results which provide useful approximations even when the sample size is far from infinite. Some of the basic ideas of asymptotic theory, in particular the concept of consistency, were introduced in Section 3.3. In this section, we investigate the asymptotic properties of the linear regression model. We show that, under much weaker assumptions than those of the classical normal linear model, the OLS estimator is asymptotically normally distributed with a familiar-looking covariance matrix.

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

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

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 3.8. Let X be a random variable with CDF $F(X)$, and suppose 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 gives a weight of $1/n$ to each of the x_t for $t = 1, \ldots, n$. The EDF is the distribution function of the empirical distribution. It is defined algebraically as

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

where $\mathbb{I}(\cdot)$ 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. As an illustration, 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 Chapter 6, and this type of simulation can be very useful.

The Fundamental Theorem of Statistics tells us that the EDF consistently estimates the CDF of the random variable X. More formally, the theorem can be stated as

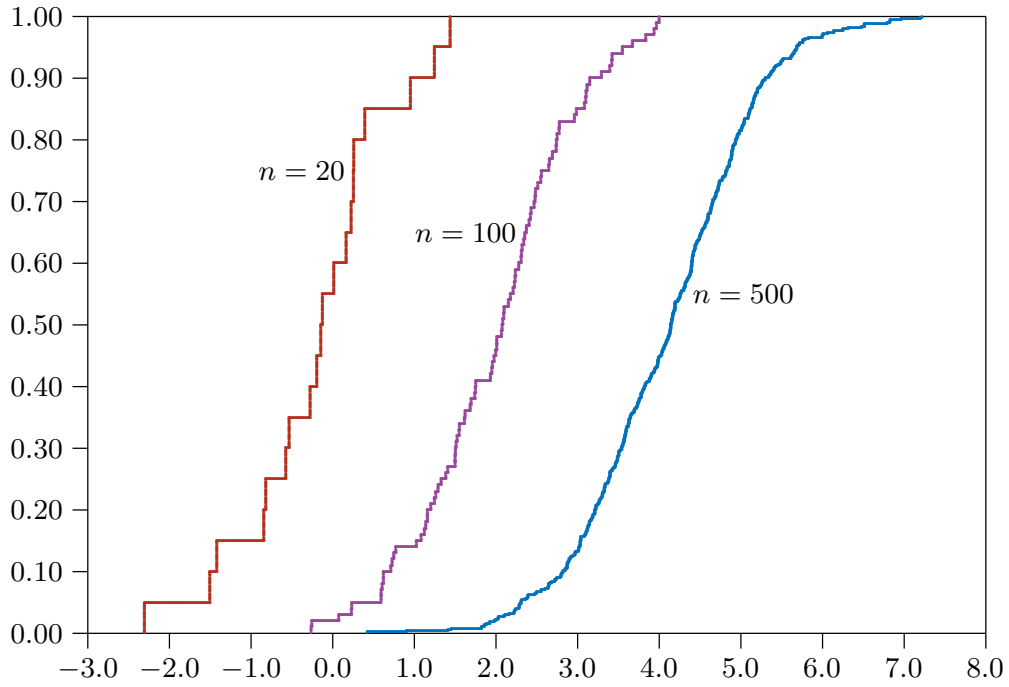

Figure 4.6 EDFs for several sample sizes

Theorem 4.2. (Fundamental Theorem of Statistics)

For the EDF $\hat{F}(x)$ defined in (4.44), for any x,

$$\operatorname*{plim}_{n\to\infty} \hat{F}(x) = F(x).$$

Proof: For any real value of x, each term in the sum on the right-hand side of equation (4.44) depends only on x_t. The expectation of $\mathbb{I}(x_t \leq x)$ can be found by using the fact that it can take on only two values, 1 and 0. The expectation is

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

Since the x_t are mutually independent, so too are the terms $\mathbb{I}(x_t \leq x)$. Since the x_t all follow the same distribution, so too must these terms. Thus $\hat{F}(x)$ 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. Thus 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, as we saw in Section 3.3, 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 $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$, under many plausible assumptions about how the rows of the matrix $\boldsymbol{X}$ are generated, tends 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 approximately follows 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

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

is **asymptotically distributed** as $\mathrm{N}(0, 1)$. This means that, as $n \to \infty$, the sequence of random variables z_n converges in distribution to the $\mathrm{N}(0, 1)$ distribution; recall the discussion of convergence in distribution in Section 3.3. We can write this result compactly as $z_n \xrightarrow{d} \mathrm{N}(0, 1)$.

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, let us calculate the variance of z_n. Since the terms in the sum in (4.45) are independent, the variance of z_n is just the sum of the variances of the n terms:

$$\mathrm{Var}(z_n) = n\,\mathrm{Var}\left(\frac{1}{\sqrt{n}}\frac{x_t - \mu}{\sigma}\right) = \frac{n}{n} = 1.$$

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 Davidson and MacKinnon (1993, Section 4.7). Davidson (1994) provides a more advanced treatment.

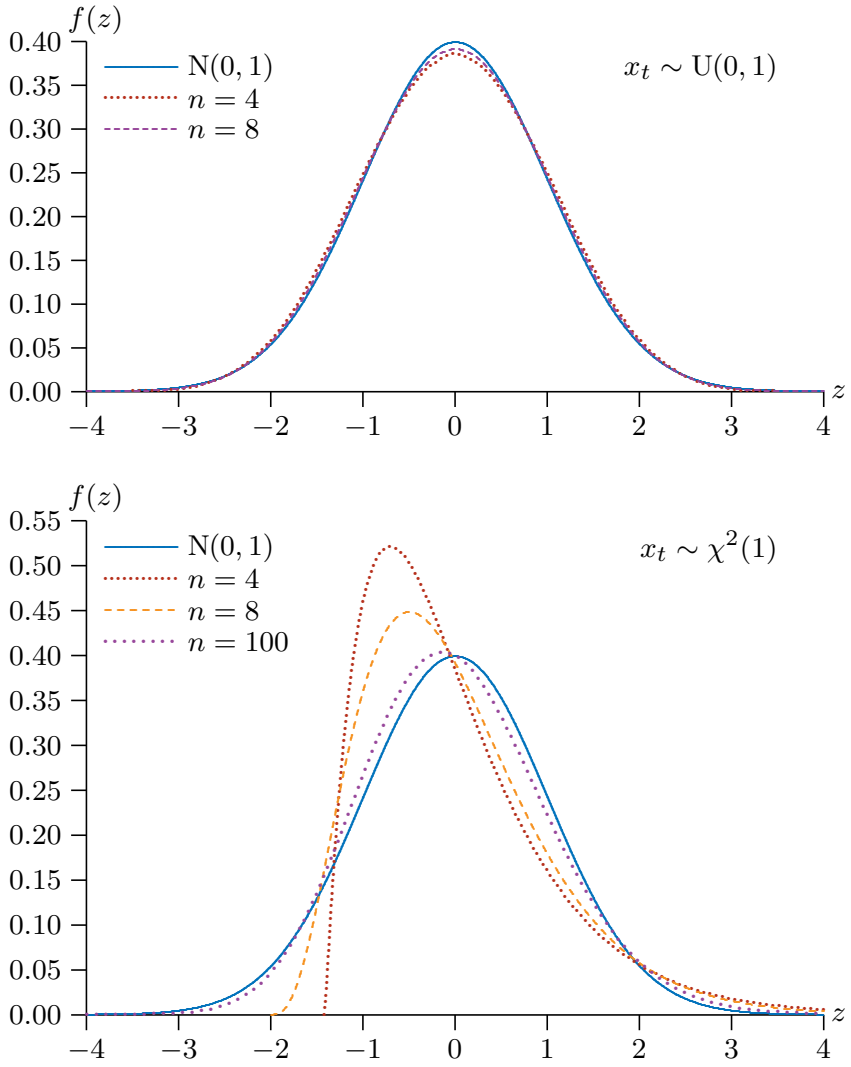

Figure 4.7 The normal approximation for different values of n

In all cases of interest to us, the CLT says that, for a sequence of uncorrelated random variables x_t, $t = 1, \ldots, \infty$, with $\mathrm{E}(x_t) = 0$,

$$n^{-1/2} \sum_{t=1}^{n} x_t = x_n^0 \xrightarrow{d} \mathrm{N}\Big(0, \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathrm{Var}(x_t)\Big).$$

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

$$n^{-1/2} \sum_{t=1}^{n} \boldsymbol{x}_t = \boldsymbol{x}_n^0 \xrightarrow{d} \mathrm{N}\Big(\mathbf{0}, \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathrm{Var}(\boldsymbol{x}_t)\Big), \tag{4.46}$$

where $\boldsymbol{x}_n^0$ 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_n defined as in (4.45). In the top panel, the x_t are uniformly distributed, and we see that z_n 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[6] of the distribution is 0, there are no values less than 0, and there is a very long right-hand tail. For $n = 4$ and $n = 8$, the standard normal provides a poor approximation to the actual distribution of z_n. For $n = 100$, on the other hand, the approximation is not bad at all, although it is still noticeably skewed to the right.

Asymptotic Normality and Root-n Consistency

Although the notion of **asymptotic normality** is very general, for now we will introduce it for linear regression models only. Suppose that the data are generated by the DGP

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

instead of the classical normal linear model (4.20). The disturbances here are drawn from some specific but unknown distribution with mean 0 and variance σ_0^2. Although some or all of the regressors may be exogenous, that assumption is stronger than we need. Instead, we allow $\boldsymbol{X}_t$ to contain lagged dependent variables, replacing the exogeneity assumption with assumption (3.13) from Section 3.2, plus an analogous assumption about the variance. These two assumptions can be written as

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

The first equation here, which is assumption (3.13), can be referred to in two ways. From the point of view of the explanatory variables $\boldsymbol{X}_t$, it says that they are **predetermined** with respect to the disturbances, a terminology that was introduced in Section 3.2. From the point of view of the disturbances, however, 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. 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 in equations (4.47) and (4.48), we now need to make an additional

[6] A **mode** of a distribution is a point at which the density achieves a local maximum. If there is just one such point, a density is said to be **unimodal**.

assumption in order to be able to use asymptotic results. We 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 k–vector

$$\boldsymbol{v} \equiv n^{-1/2} \boldsymbol{X}^\top \boldsymbol{u} = n^{-1/2} \sum_{t=1}^{n} u_t \boldsymbol{X}_t^\top. \tag{4.50}$$

We wish to apply a multivariate CLT to this vector. By the first assumption in (4.48), $\mathrm{E}(u_t \,|\, \boldsymbol{X}_t) = 0$. This implies that $\mathrm{E}(u_t \boldsymbol{X}_t^\top) = \mathbf{0}$, as required for the CLT. Thus, assuming that the vectors $u_t \boldsymbol{X}_t^\top$ satisfy the technical assumptions for an appropriate multivariate CLT to apply, we have from (4.46) that

$$\boldsymbol{v} \xrightarrow{d} \mathrm{N}\Big(\mathbf{0}, \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathrm{Var}(u_t \boldsymbol{X}_t^\top)\Big) = \mathrm{N}\Big(\mathbf{0}, \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathrm{E}(u_t^2 \boldsymbol{X}_t^\top \boldsymbol{X}_t)\Big).$$

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 equations (4.48) says that the errors are conditionally homoskedastic. It allows us to simplify the limiting covariance matrix:

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \mathrm{E}(u_t^2 \boldsymbol{X}_t^\top \boldsymbol{X}_t) &= \lim_{n\to\infty} \sigma_0^2 \frac{1}{n} \sum_{t=1}^{n} \mathrm{E}(\boldsymbol{X}_t^\top \boldsymbol{X}_t) \\ &= \sigma_0^2 \operatorname*{plim}_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} \boldsymbol{X}_t^\top \boldsymbol{X}_t \\ &= \sigma_0^2 \operatorname*{plim}_{n\to\infty} \frac{1}{n} \boldsymbol{X}^\top \boldsymbol{X} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}. \end{aligned} \tag{4.51}$$

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

$$\boldsymbol{v} \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top \boldsymbol{X}}). \tag{4.52}$$

Consider now the estimation error of the vector of OLS estimates. For the DGP (4.47), this is

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{u}. \tag{4.53}$$

As we saw in Section 3.3, $\hat{\boldsymbol{\beta}}$ is consistent under fairly weak conditions. If it is, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ must tend to a limit of $\mathbf{0}$ as the sample size $n \to \infty$. Therefore, its limiting covariance matrix is a zero matrix. Thus it would appear that asymptotic theory has nothing to say about limiting variances for consistent estimators. However, this is easily corrected by the usual device of introducing a few well-chosen powers of n. If we rewrite equation (4.53) as

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = (n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1} n^{-1/2}\boldsymbol{X}^\top\boldsymbol{u},$$

then the first factor on the right-hand side tends to $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}^{-1}$ as $n \to \infty$, and the second factor, which is just $\boldsymbol{v}$, tends to a random vector distributed as $\mathrm{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})$. Because $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ is deterministic, we find that, asymptotically,

$$\mathrm{Var}\big(n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big) = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}^{-1} \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}} \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}^{-1} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}^{-1}.$$

Moreover, since $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is, asymptotically, just a deterministic linear combination of the components of the multivariate normal random vector $\boldsymbol{v}$, we conclude that

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}^{-1}).$$

Informally, we may say that the vector $\hat{\boldsymbol{\beta}}$ is **asymptotically normal**, or exhibits **asymptotic normality**.

It is convenient to collect the key results above into a theorem.

Theorem 4.3.

For the correctly specified linear regression model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{IID}(\mathbf{0}, \sigma^2\mathbf{I}), \tag{4.54}$$

where the data are generated by the DGP (4.47), the regressors and disturbances satisfy assumptions (4.48), and the regressors satisfy assumption (4.49),

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}^{-1}), \tag{4.55}$$

and

$$\operatorname*{plim}_{n\to\infty} s^2(n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1} = \sigma_0^2 \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}^{-1}. \tag{4.56}$$

Remark: The first part of the theorem allows us to pretend that $\hat{\boldsymbol{\beta}}$ is normally distributed with mean $\mathbf{0}$, and the second part allows us to use $s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ to estimate $\mathrm{Var}(\hat{\boldsymbol{\beta}})$. Of course, these are both just approximations. The theorem does not tell us that **asymptotic inference** based on these approximations will necessarily be reliable.

The result (4.56) tells us that the **asymptotic covariance matrix** of the vector $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is the limit of the matrix $\sigma_0^2(n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1}$ as $n \to \infty$. This result follows from (4.51) and the consistency of s^2, which is just $n/(n-k)$ times the average of the $\hat{u}_t^2$ and must tend to σ_0^2 by a law of large numbers; see Section 3.7 and Exercise 3.18.

It is important to remember that, whenever the matrix $n^{-1}\boldsymbol{X}^\top\boldsymbol{X}$ tends to $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$ as $n \to \infty$, the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$, without the factor of n, simply tends to a zero matrix. As we saw just below equation (4.53), this is simply a consequence of the fact that $\hat{\boldsymbol{\beta}}$ is consistent. Thus, although it would be convenient if we could dispense with powers of n when working out asymptotic approximations to covariance matrices, it would be mathematically incorrect and very risky to do so.

The result (4.55) gives us the **rate of convergence** of $\hat{\boldsymbol{\beta}}$ to its probability limit of $\boldsymbol{\beta}_0$. Since multiplying the estimation error by $n^{1/2}$ gives rise to an expression of zero mean and finite covariance matrix, it follows that the estimation error itself tends to zero at the same rate as $n^{-1/2}$. This property is expressed by saying that the estimator $\hat{\boldsymbol{\beta}}$ is **root-n consistent**.

Quite generally, suppose that $\hat{\boldsymbol{\theta}}$ is a root-n consistent, asymptotically normal, estimator of a parameter vector $\boldsymbol{\theta}$. Any estimator of the covariance matrix of $\hat{\boldsymbol{\theta}}$ must tend to zero as $n \to \infty$. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$, and let $\boldsymbol{V}(\boldsymbol{\theta})$ denote the limiting covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Then an estimator $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})$ is said to be consistent for the covariance matrix of $\hat{\boldsymbol{\theta}}$ if

$$\operatorname*{plim}_{n\to\infty}\left(n\,\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}})\right) = \boldsymbol{V}(\boldsymbol{\theta}). \tag{4.57}$$

For every root-n consistent estimator, there is generally at least one such covariance matrix estimator.

4.6 Large-Sample Tests

Theorem 4.3 implies that the t test discussed in Section 4.4 is asymptotically valid under weaker conditions than those needed to prove that the t statistic actually follows the Student's t distribution in finite samples. Consider the linear regression model (4.21), but with IID errors and regressors that may be predetermined rather than exogenous. As before, we wish to test the hypothesis that $\beta_2 = \beta_2^0$. The t statistic for this hypothesis is simply

$$t_{\beta_2} = \frac{\hat{\beta}_2 - \beta_2^0}{\sqrt{s^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}}} = \frac{n^{1/2}(\hat{\beta}_2 - \beta_2^0)}{\sqrt{s^2(n^{-1}\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}}}, \tag{4.58}$$

where $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}$ denotes the second element on the main diagonal of the matrix $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}$. The result (4.55) tells us that $n^{1/2}(\hat{\beta}_2 - \beta_2^0)$ is asymptotically

normally distributed with variance the second element on the main diagonal of $\sigma_0^2(\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})^{-1}$. Equation (4.56) tells us that s^2 times $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{22}$ consistently estimates this variance. Therefore, t_{β_2} must follow the standard normal distribution asymptotically under the null hypothesis. We may write

$$t_{\beta_2} \overset{a}{\sim} \mathrm{N}(0,1). \tag{4.59}$$

The notation "$\overset{a}{\sim}$" means that t_{β_2} is **asymptotically distributed** as $\mathrm{N}(0,1)$. This is just a different way of saying that t_{β_2} converges in distribution to $\mathrm{N}(0,1)$.

The result (4.59) justifies the use of t tests outside the confines of the classical normal linear model. We can compute asymptotic P values or critical values using either the standard normal or t distributions. Of course, these **asymptotic *t* tests** are not exact in finite samples, and they may or may not be reliable. It is often possible to perform more reliable tests by using the bootstrap, which will be introduced in Chapter 6.

Asymptotic *F* Tests

In view of the result (4.59) for the asymptotic t statistic, it should not be surprising that the F statistic (4.33) for the null hypothesis that $\boldsymbol{\beta}_2 = \mathbf{0}$ in the model (4.28) is also valid asymptotically when the DGP is (4.47) and the disturbances satisfy assumptions (4.48). Under the null, $F_{\boldsymbol{\beta}_2}$ is equal to expression (4.34), which can be rewritten as

$$F_{\boldsymbol{\beta}_2} = \frac{n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{M}_1\boldsymbol{X}_2(n^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon}/r}{\boldsymbol{\varepsilon}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}/(n-k)}, \tag{4.60}$$

where $\boldsymbol{\varepsilon} \equiv \boldsymbol{u}/\sigma_0$ and $r = k_2$, the dimension of $\boldsymbol{\beta}_2$. We now show that $rF_{\boldsymbol{\beta}_2}$ is asymptotically distributed as $\chi^2(r)$. This result follows from Theorem 4.3, but it is not entirely obvious.

The denominator of the F statistic (4.60) is $\boldsymbol{\varepsilon}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{\varepsilon}/(n-k)$, which is just s^2 times $1/\sigma_0^2$. Since s^2 is consistent for σ_0^2, it is evident that the denominator of expression (4.60) must tend to 1 asymptotically.

The numerator of the F statistic, multiplied by r, is

$$n^{-1/2}\boldsymbol{\varepsilon}^\top\boldsymbol{M}_1\boldsymbol{X}_2(n^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\,n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon}. \tag{4.61}$$

Let $\boldsymbol{v} = n^{-1/2}\boldsymbol{X}^\top\boldsymbol{\varepsilon}$. Then a central limit theorem shows that $\boldsymbol{v} \overset{a}{\sim} \mathrm{N}(\mathbf{0}, \boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}})$, as in the previous section. If we partition $\boldsymbol{v}$, conformably with the partition of $\boldsymbol{X}$, into two subvectors $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, we have

$$n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{\varepsilon} = n^{-1/2}\boldsymbol{X}_2^\top\boldsymbol{\varepsilon} - n^{-1}\boldsymbol{X}_2^\top\boldsymbol{X}_1(n^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_1)^{-1}\,n^{-1/2}\boldsymbol{X}_1^\top\boldsymbol{\varepsilon}. \tag{4.62}$$

This expression evidently tends to the vector $\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1$ as $n \to \infty$. Here $\boldsymbol{S}_{11}$ and $\boldsymbol{S}_{21}$ are submatrices of $\boldsymbol{S}_{\boldsymbol{X}^\top\boldsymbol{X}}$, so that $n^{-1}\boldsymbol{X}_1^\top\boldsymbol{X}_1$ tends to $\boldsymbol{S}_{11}$ and

$n^{-1}\boldsymbol{X}_2^\top\boldsymbol{X}_1$ tends to $\boldsymbol{S}_{21}$. Since $\boldsymbol{v}$ is asymptotically multivariate normal, and $\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1$ is just a linear combination of the elements of $\boldsymbol{v}$, this vector must itself be asymptotically multivariate normal.

The vector (4.62) evidently has mean $\boldsymbol{0}$. Thus its covariance matrix is the expectation of

$$n^{-1}\boldsymbol{X}_2^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}_2 + \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}n^{-1}\boldsymbol{X}_1^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}_1\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12} - 2n^{-1}\boldsymbol{X}_2^\top\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top\boldsymbol{X}_1\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}.$$

We can replace $\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top$ by its expectation, which is $\sigma_0^2\mathbf{I}$. Then, by the second part of Theorem 4.3, we can replace $n^{-1}\boldsymbol{X}_i^\top\boldsymbol{X}_j$ by $\boldsymbol{S}_{ij}$, for $i, j = 1, 2$, which is what each of those submatrices tends to asymptotically. This yields an expression that can be simplified, allowing us to conclude that

$$\mathrm{Var}(\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1) = \boldsymbol{S}_{22} - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12}.$$

Thus the numerator of the F statistic, expression (4.61), is asymptotically equal to

$$(\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1)^\top(\boldsymbol{S}_{22} - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{S}_{12})^{-1}(\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1). \tag{4.63}$$

This is simply a quadratic form in the r–vector $\boldsymbol{v}_2 - \boldsymbol{S}_{21}\boldsymbol{S}_{11}^{-1}\boldsymbol{v}_1$, which is asymptotically multivariate normal, and the inverse of its covariance matrix. By Theorem 4.1, it follows that expression (4.63) is asymptotically distributed as $\chi^2(r)$. Because the denominator of the F statistic tends to 1 asymptotically, we conclude that

$$rF_{\boldsymbol{\beta}_2} \overset{a}{\sim} \chi^2(r) \tag{4.64}$$

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

The result (4.64) justifies the use of **asymptotic $\boldsymbol{F}$ tests** when the disturbances are not normally distributed and some of the regressors are predetermined rather than exogenous. We can the compute P value associated with an F statistic using either the χ^2 or F distributions. Of course, if we use the χ^2 distribution, we have to multiply the F statistic by r.

Wald Tests

A vector of r linear restrictions on a parameter vector $\boldsymbol{\beta}$ can always be written in the form

$$\boldsymbol{R}\boldsymbol{\beta} = \boldsymbol{r}, \tag{4.65}$$

where $\boldsymbol{R}$ is an $r \times k$ matrix and $\boldsymbol{r}$ is an r–vector. For example, if $k = 3$ and the restrictions were that $\beta_1 = 0$ and $\beta_2 = -1$, equations (4.65) would be

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}\boldsymbol{\beta} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}.$$

The elements of the matrix $\boldsymbol{R}$ and the vector $\boldsymbol{r}$ must be known. They are not functions of the data, and, as in this example, they are very often integers.

Now suppose that we have obtained a k–vector of unrestricted parameter estimates $\hat{\boldsymbol{\beta}}$, of which the covariance matrix is $\mathrm{Var}(\hat{\boldsymbol{\beta}})$. By a slight generalization of the result (3.40), the covariance matrix of the vector $\boldsymbol{R}\boldsymbol{\beta}$ must be $\boldsymbol{R}\,\mathrm{Var}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top$. Then the simplest way to test the restrictions (4.65) is to calculate the **Wald statistic**

$$W(\hat{\boldsymbol{\beta}}) = (\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})^\top \big(\boldsymbol{R}\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top\big)^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r}), \tag{4.66}$$

where $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ estimates $\mathrm{Var}(\hat{\boldsymbol{\beta}})$ consistently. Inserting appropriate factors of n, equation (4.66) can be rewritten as

$$W(\hat{\boldsymbol{\beta}}) = \big(n^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})\big)^\top \big(\boldsymbol{R}n\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\boldsymbol{R}^\top\big)^{-1}\big(n^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}} - \boldsymbol{r})\big). \tag{4.67}$$

Theorem 4.3 implies that the vector $n^{1/2}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$ is asymptotically multivariate normal. Therefore, the right-hand side of equation (4.67) is asymptotically a quadratic form in an r–vector that is multivariate normal and the inverse of its covariance matrix. It follows that, by Theorem 4.1, the Wald statistic must be asymptotically distributed as $\chi^2(r)$ under the null hypothesis.

These results are much more general than the ones for asymptotic t tests and F tests. Equation (4.66) would still define a Wald statistic for the hypothesis (4.65) if $\hat{\boldsymbol{\beta}}$ were any root-n consistent estimator and $\widehat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$ were any consistent estimator of its covariance matrix. Thus we will encounter Wald tests many times throughout this book. For the specific case of a linear regression model with zero restrictions on some of the parameters, Wald tests turn out to be very closely related to t tests and F tests. In fact, the square of the t statistic (4.25) is a Wald statistic, and r times the F statistic (4.33) is a Wald statistic. Readers are asked to demonstrate these results in Exercise 4.13 and Exercise 4.14, respectively.

Asymptotic tests cannot be exact in finite samples, because they are necessarily based on P values, or critical values, that are approximate. By itself, asymptotic theory cannot tell us just how accurate such tests are. If we decide to use a nominal level of α for a test, we reject if the asymptotic P value is less than α. In many cases, but certainly not all, asymptotic tests are probably 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 overreject or underreject, and how severely, depends on many things, including the sample size, the distribution of the disturbances, the number of regressors and their properties, the number of restrictions, and the relationship between the disturbances and the regressors.

4.7 Performing Several Hypothesis Tests

Up to this point, we have implicitly assumed that just one hypothesis test is performed at a time. This allows test statistics and P values to be interpreted in the usual way. In practice, however, investigators almost always perform several tests. For example, whenever an econometrics package is used to run an OLS regression, the package will normally report a t statistic for every coefficient. Unless the investigator consciously chooses to ignore all but one of these t statistics, which most people would find it almost impossible to do, he or she is implicitly (and often explicitly) engaged in **multiple testing**. This simply means performing two or more hypothesis tests as part of the same investigation. It is not to be confused with testing two or more restrictions via a single test, such as an F test or a Wald test.

The problem with multiple testing is that an unusually large test statistic is much more likely to be obtained by accident when several tests are performed rather than just one. This is easiest to see if the test statistics are independent. Suppose that we perform m exact tests at level α. Let α_m denote the **familywise error rate**, which is the probability that at least one of the tests rejects. Because the tests are independent, the familywise error rate is simply one minus the probability that none of the tests rejects:

$$\alpha_m = 1 - (1 - \alpha)^m. \tag{4.68}$$

When m is large, α_m can be much larger than α. For example, if $\alpha = 0.05$, then $\alpha_2 = 0.0975$, $\alpha_4 = 0.18549$, $\alpha_8 = 0.33658$, and $\alpha_{16} = 0.55987$. It is evident from (4.68) that the familywise error rate can be very much larger than the level of each individual test when the number of tests is large.

The simplest method for controlling the familywise error rate is known as the **Bonferroni procedure**. Instead of rejecting the joint null hypothesis whenever the smallest P value is less than α, it rejects whenever the smallest P value is less than α/m. This procedure is based on the Bonferroni inequality

$$\Pr\bigl(\cup_{i=1}^{m}(P_i \leq \alpha/m)\bigr) \leq \alpha,$$

where P_i is the P value for the i^{th} test. The Bonferroni procedure is very easy to implement, but it can be extremely conservative. For large m, α/m is very much smaller than α. It is even smaller than the value α that solves equation (4.68) for a given α_m, which would be appropriate if all the tests were independent. When the P values are positively correlated, as is often the case in practice, α/m can be much too small. Consider the extreme case in which there is perfect dependence and all the tests yield identical P values. In that case, the familywise error rate for individual tests at level α is just α, and no correction is needed.

There is a large literature on multiple testing in statistics. For example, Simes (1986) and Hochberg (1988) proposed improved Bonferroni procedures that

are less conservative. They both use all the P values, not just the smallest one. The Simes procedure is quite simple. We first order the P values from the smallest, $P_{(1)}$, to the largest, $P_{(m)}$. Then we reject the joint null hypothesis whenever

$$P_{(j)} \leq j\alpha/m \text{ for any } j = 1, \ldots, m, \tag{4.69}$$

where α is the desired familywise error rate. If the smallest P value is less than α/m, both this procedure and the Bonferroni procedure reject. But the Simes procedure can also reject when the second-smallest P value is less than $2/m$, the third-smallest is less than $3/m$, and so on. Thus it is always less conservative than the Bonferroni procedure. Because it is based on an inequality that may not always hold, the Simes procedure can conceivably yield misleading results, but it seems to work well in practice.

A more recent approach is to control the **false discovery rate** instead of the familywise error rate; see Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). The false discovery rate is the expected proportion of erroneous rejections among all rejections. The idea is that some of the tested hypotheses may be true and others may be false. Instead of either accepting or rejecting the joint null hypothesis, we want to reject the false nulls but not the true ones. As in (4.69), we order the P values and compare them to $j\alpha/m$. Let J denote the largest j for which the inequality holds. Then we reject the first J hypotheses and do not reject the remaining ones. If the inequality is never satisfied, then the Benjamini-Hochberg and Simes procedures yield the same results, namely, that the joint null hypothesis is not rejected.

In this section, we have given a very brief introduction to testing multiple hypotheses. We have assumed that little or nothing is known about the joint distribution of the test statistics. If we knew that distribution, then we could in principle do better than any of the procedures discussed above. This suggests that bootstrap-based procedures may be attractive, and these will be discussed in Chapter 6.

4.8 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 has 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. The problem with the Bonferroni procedure discussed in the previous section is that it can have very low power when there are many hypotheses to be tested. 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.

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 $\boldsymbol{P}_{\boldsymbol{M}_1\boldsymbol{X}_2}\boldsymbol{y}$ for the numerator and $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}$ for the denominator. 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; see expression (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 equation (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 $\mathrm{N}(\boldsymbol{0}, \mathbf{I})$, then $\|\boldsymbol{z}\|^2 = \boldsymbol{z}^\top\boldsymbol{z}$ is distributed as (central) chi-squared with m degrees of freedom. Similarly, if $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega})$, then $\boldsymbol{x}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{x} \sim \chi^2(m)$. If instead $\boldsymbol{z} \sim \mathrm{N}(\boldsymbol{\mu}, \mathbf{I})$, 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 4.19. Likewise, if $\boldsymbol{x} \sim \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Omega})$, 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

$$\begin{aligned}\Lambda &\equiv \frac{1}{\sigma^2}\,\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2(\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2 \\ &= \frac{1}{\sigma^2}\,\boldsymbol{\beta}_2^\top\boldsymbol{X}_2^\top\boldsymbol{M}_1\boldsymbol{X}_2\boldsymbol{\beta}_2. \end{aligned}\tag{4.71}$$

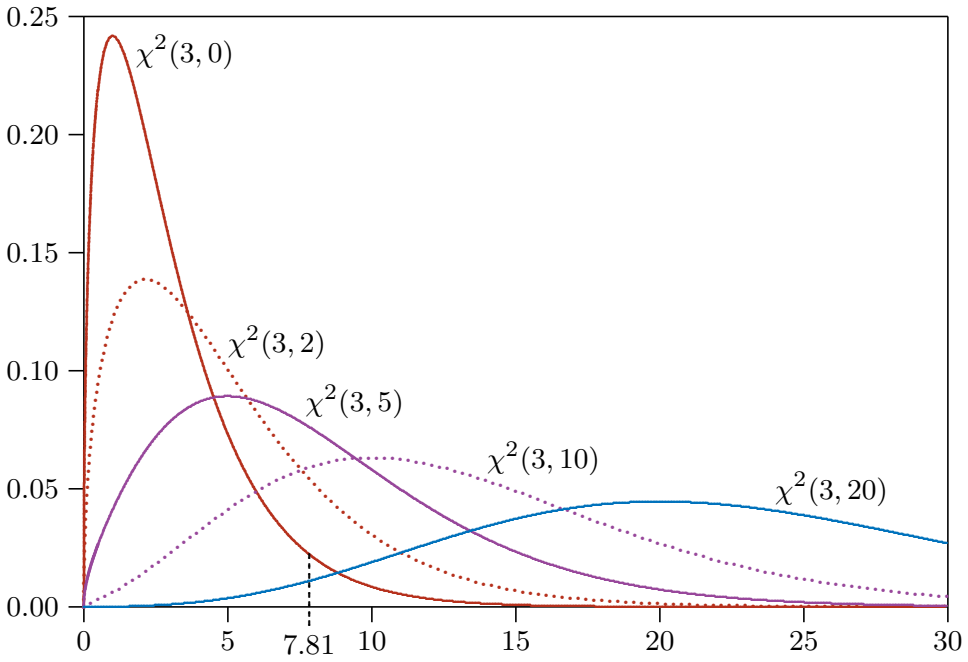

Figure 4.8 Densities of noncentral χ^2 distributions

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 expression (4.70), r/σ^2 times the F statistic, 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 4.8, 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 $\chi^2(3,\Lambda)$ distribution, the probability that the null hypothesis is rejected at the .05 level is the probability mass to the right of 7.81. It is evident from the figure that this probability is small for small values of the NCP and large for large ones.

In Figure 4.8, 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 increases 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 equation (4.25), the t statistic t_{β_2} is $1/s$ times

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

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 expression (4.72) is normal with variance 1 and mean

$$\lambda \equiv \frac{1}{\sigma}(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2} \beta_2. \tag{4.73}$$

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

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

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 $\mathrm{N}(\lambda, 1)$ distribution. It is also very much like an ordinary, or **central**, t distribution with its mean shifted from the origin to (4.73), 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 at any 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 (4.73). 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 $(\boldsymbol{x}_2^\top \boldsymbol{M}_1 \boldsymbol{x}_2)^{1/2}$ is roughly proportional to $n^{1/2}$, and so λ tends to infinity with n at a rate similar to that of $n^{1/2}$.

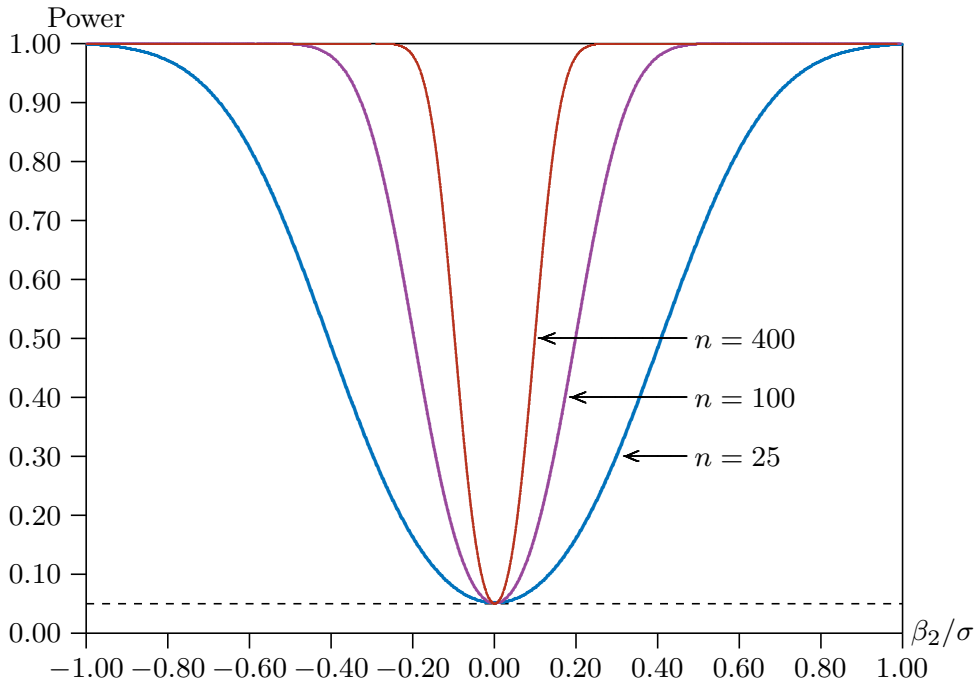

Figure 4.9 Power functions for t tests at the .05 level

Figure 4.9 shows power functions for tests at the .05 level 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 $\beta_2 = 0$. Power then increases as β_2 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 $\beta_2 \neq 0$. It is clear that, as $n \to \infty$, the power function converges 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 rejects 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 $n^{1/2}\beta_2/\sigma$ is sufficiently small.

Because t tests in the classical normal linear regression model are exact, the case shown in Figure 4.9 is an ideal one. Tests that are only valid asymptotically may have power functions that look quite different from the ones in the figure. Power may be greater or less than .05 when the null hypothesis holds, depending on whether the test overrejects or underrejects, and it may well be minimized at a parameter value that does not correspond to the null. Instead of being a symmetric inverted bell shape, the power function may be quite asymmetrical, and in some cases power may not even tend to unity as the parameter under test becomes infinitely far from the null hypothesis. Readers are asked to investigate a less than ideal case in Exercise 4.20.

4.9 Pretesting

In regression analysis, interest often centers only on certain explanatory variables. The other explanatory variables are generally included solely to avoid possible misspecification. Consider the linear regression model (3.61), which was discussed in Section 3.8 and is rewritten here for convenience:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2 \mathbf{I}). \tag{4.74}$$

Here $\boldsymbol{\beta}$ is a k–vector, $\boldsymbol{\gamma}$ is an r–vector, and the regressors in $\boldsymbol{X}$ and $\boldsymbol{Z}$ are assumed, for simplicity, to be exogenous. The parameters of interest are the k elements of $\boldsymbol{\beta}$. We would like to estimate them as well as possible, but we do not care about $\boldsymbol{\gamma}$.

As we saw in Section 3.8, the unrestricted OLS estimator of $\boldsymbol{\beta}$ from (4.74) and its covariance matrix are

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{y} \quad \text{and} \quad \text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{M}_{\boldsymbol{Z}} \boldsymbol{X})^{-1}. \tag{4.75}$$

Similarly, the restricted OLS estimator and its covariance matrix are

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y} \quad \text{and} \quad \text{Var}(\tilde{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1}. \tag{4.76}$$

Except in the very special case in which the matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ are orthogonal, the restricted estimator $\tilde{\boldsymbol{\beta}}$ is more efficient than the unrestricted estimator $\hat{\boldsymbol{\beta}}$. However, because the estimator $\tilde{\boldsymbol{\beta}}$ is biased if $\boldsymbol{\gamma} \neq \boldsymbol{0}$, its mean squared error matrix is larger than its covariance matrix in that case. As we showed in equation (3.72), that matrix is

$$\text{MSE}(\tilde{\boldsymbol{\beta}}) = \sigma^2 (\boldsymbol{X}^\top \boldsymbol{X})^{-1} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Z} \boldsymbol{\gamma} \boldsymbol{\gamma}^\top \boldsymbol{Z}^\top \boldsymbol{X} (\boldsymbol{X}^\top \boldsymbol{X})^{-1}, \tag{4.77}$$

which is the sum of the covariance matrix from (4.76) and the bias vector times itself transposed.

Since $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$ when $\boldsymbol{\gamma}$ is zero, it seems natural to test the hypothesis that $\boldsymbol{\gamma} = \boldsymbol{0}$ and use the latter estimator when the test rejects and the former when it does not. This test is called a **preliminary test**, or **pretest** for short. Such a procedure implicitly defines a new estimator, which is called a **pretest estimator**. Formally, we can write

$$\acute{\boldsymbol{\beta}} = \mathbb{I}(F_{\boldsymbol{\gamma}=\boldsymbol{0}} > c_\alpha) \hat{\boldsymbol{\beta}} + \mathbb{I}(F_{\boldsymbol{\gamma}=\boldsymbol{0}} \leq c_\alpha) \tilde{\boldsymbol{\beta}}, \tag{4.78}$$

where $F_{\boldsymbol{\gamma}=\boldsymbol{0}}$ is the F statistic for $\boldsymbol{\gamma} = \boldsymbol{0}$, and c_α is the critical value for an F test with r and $n - k - r$ degrees of freedom at level α. Thus $\acute{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ when the pretest rejects, and $\acute{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}$ when the pretest does not reject.

Equation (4.78) for the pretest estimator can be written in a simpler form as

$$\acute{\boldsymbol{\beta}} = \hat{\lambda} \hat{\boldsymbol{\beta}} + (1 - \hat{\lambda}) \tilde{\boldsymbol{\beta}}, \tag{4.79}$$

where $\hat{\lambda} = 1$ when $F_{\boldsymbol{\gamma}=\mathbf{0}} > c_\alpha$ and $\hat{\lambda} = 0$ when $F_{\boldsymbol{\gamma}=\mathbf{0}} \leq c_\alpha$. In this form, $\hat{\boldsymbol{\beta}}$ looks like a weighted average of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$, but with random weights that can only equal 0 or 1. Thus the pretest estimator defined in (4.79) is actually a special case of a **model average estimator**. As the name implies, such an estimator is formed by taking a weighted average of other estimators. Model average estimators will be discussed in Chapter 17.

The MSE Matrix of the Pretest Estimator

Because the outcome of the pretest is random, the MSE matrix of the pretest estimator is not simply a weighted average of $\text{Var}(\hat{\boldsymbol{\beta}})$, from (4.75), and the MSE matrix (4.77). In fact, it takes some effort to obtain this matrix. The analysis that follows is based on results in Magnus and Durbin (1999) and Danilov and Magnus (2004). It requires us to make the additional assumption that the disturbances in regression (4.74) are normally distributed. Thus we are now dealing with the classical normal linear model.

The probability that the pretest rejects at level α depends on the $\boldsymbol{X}$ and $\boldsymbol{Z}$ matrices and the unknown parameters $\boldsymbol{\gamma}$ and σ^2 through the noncentrality parameter

$$\Lambda = \frac{1}{\sigma^2}\boldsymbol{\gamma}^\top \boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z}\boldsymbol{\gamma}, \tag{4.80}$$

which is a special case of expression (4.71). Now make the definition

$$\boldsymbol{\theta} \equiv (\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z})^{1/2}\boldsymbol{\gamma}, \tag{4.81}$$

and observe from (4.80) that $\boldsymbol{\theta}^\top\boldsymbol{\theta}$ is the numerator of the NCP Λ. If we then make the definition

$$\boldsymbol{Q} \equiv (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z})^{-1/2}, \tag{4.82}$$

it is easy to see that the second term in expression (4.77), which is the contribution of the bias to $\text{MSE}(\tilde{\boldsymbol{\beta}})$, can be written as $\boldsymbol{Q}\boldsymbol{\theta}\boldsymbol{\theta}^\top\boldsymbol{Q}^\top$.

Replacing $\boldsymbol{\gamma}$ by its OLS estimator $\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$ in expression (4.81) yields the least squares estimate of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}} = (\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z})^{-1/2}\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}. \tag{4.83}$$

The next step is to show that

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Z}\hat{\boldsymbol{\gamma}} = \boldsymbol{Q}\hat{\boldsymbol{\theta}}. \tag{4.84}$$

The second equation in (4.84) follows immediately from equations (4.82) and (4.83), and the fact that $\hat{\boldsymbol{\gamma}} = (\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{M}_{\boldsymbol{X}} \boldsymbol{y}$, when we postmultiply $\boldsymbol{Q}$ by $\hat{\boldsymbol{\theta}}$. The first one is not quite so obvious. By definition, the fitted values and

the residuals must sum to the regressand. This is true for both the restricted and unrestricted models. Therefore,

$$\boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{Z}\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{u}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{u}},$$

which can be rearranged to yield

$$\boldsymbol{X}\tilde{\boldsymbol{\beta}} - \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{Z}\hat{\boldsymbol{\gamma}} + \hat{\boldsymbol{u}} - \tilde{\boldsymbol{u}}. \tag{4.85}$$

If we premultiply both sides of equation (4.85) by $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top$ and use the fact that $(\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{X} = \mathbf{I}$, we obtain

$$\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top\boldsymbol{Z}\hat{\boldsymbol{\gamma}} + (\boldsymbol{X}^\top\boldsymbol{X})^{-1}\boldsymbol{X}^\top(\hat{\boldsymbol{u}} - \tilde{\boldsymbol{u}}). \tag{4.86}$$

The last term on the right-hand side here must be zero, because the residuals, both restricted and unrestricted, are orthogonal to $\boldsymbol{X}$. This establishes the first equation in (4.84).

We are now ready to derive the MSE matrix of the pretest estimator $\acute{\boldsymbol{\beta}}$. By (4.79) and (4.85), this estimator can be written as

$$\acute{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}} + \hat{\lambda}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = \tilde{\boldsymbol{\beta}} - \hat{\lambda}\boldsymbol{Q}\hat{\boldsymbol{\theta}}. \tag{4.87}$$

Because $\tilde{\boldsymbol{\beta}}$ is an efficient estimator, the vector $\boldsymbol{Q}\hat{\boldsymbol{\theta}}$ must be uncorrelated with it; see the proof of the Gauss-Markov Theorem, especially equations (3.50). Because the disturbances are assumed to be normally distributed, this implies that the vectors $\tilde{\boldsymbol{\beta}}$ and $\boldsymbol{Q}\hat{\boldsymbol{\theta}}$ are independent. It must also be the case that $\tilde{\boldsymbol{\beta}}$ and $\hat{\lambda}$ are independent, because $\hat{\lambda}$ is simply a function of the F statistic for $\boldsymbol{\gamma} = \mathbf{0}$, and, under the normality assumption, the vector $\tilde{\boldsymbol{\beta}}$ is independent of the F statistic; see Exercise 4.21.

Since $\tilde{\boldsymbol{\beta}}$ is independent of both $\boldsymbol{Q}\hat{\boldsymbol{\theta}}$ and $\hat{\lambda}$, it is easy to see from (4.87) that

$$\operatorname{Var}(\acute{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \boldsymbol{Q}\operatorname{Var}(\hat{\lambda}\hat{\boldsymbol{\theta}})\boldsymbol{Q}^\top.$$

To obtain the MSE, we use the fact that the bias in the pretest estimator arises entirely from $\tilde{\boldsymbol{\beta}}$. This bias is $\boldsymbol{Q}\big(\mathrm{E}(\hat{\lambda}\hat{\boldsymbol{\theta}}) - \lambda\boldsymbol{\theta}\big)$. Thus we conclude that

$$\operatorname{MSE}(\acute{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^\top\boldsymbol{X})^{-1} + \boldsymbol{Q}\operatorname{MSE}(\hat{\lambda}\hat{\boldsymbol{\theta}})\boldsymbol{Q}^\top, \tag{4.88}$$

where $\operatorname{MSE}(\hat{\lambda}\hat{\boldsymbol{\theta}}) = \mathrm{E}\big((\hat{\lambda}\hat{\boldsymbol{\theta}} - \lambda\boldsymbol{\theta})(\hat{\lambda}\hat{\boldsymbol{\theta}} - \lambda\boldsymbol{\theta})\big)^\top$ is the MSE matrix of the vector $\hat{\lambda}\hat{\boldsymbol{\theta}}$, and λ is the probability that the F test for $\boldsymbol{\gamma} = \mathbf{0}$ will reject the null hypothesis according to the noncentral F distribution with NCP $\varLambda$ given in equation (4.80).

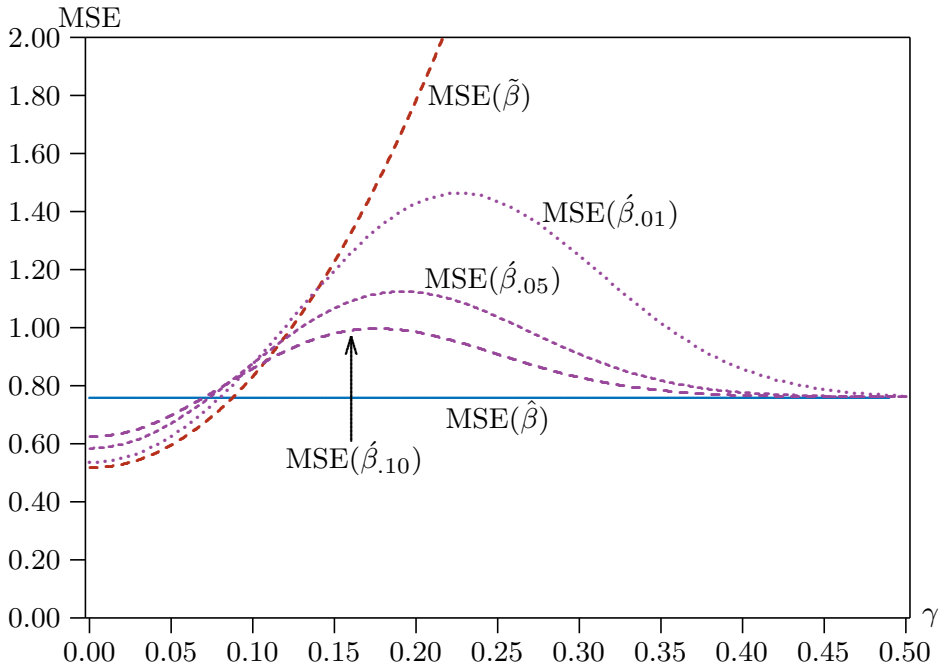

Figure 4.10 MSEs of Several Estimators

Properties of Pretest Estimators

The result (4.88) allows us to compare the MSE of the pretest estimator $\acute{\boldsymbol{\beta}}$ with the MSEs of the restricted and unrestricted estimators. This comparison turns out to be quite illuminating. For simplicity, we confine our attention to the model

$$\boldsymbol{y} = \beta \boldsymbol{x} + \gamma \boldsymbol{z} + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{NID}(\boldsymbol{0}, \sigma^2 \mathbf{I}), \tag{4.89}$$

in which there is just one parameter of interest and one restriction, and MSE is therefore a scalar rather than a matrix. We would get very similar results if there were several parameters of interest and/or several restrictions. We assume that the two regressors are bivariate normal with correlation $\rho = 0.5$. The potential reduction in variance from using the restricted estimator $\tilde{\beta}$ or the pretest estimator $\acute{\beta}$ rather than the unrestricted estimator $\hat{\beta}$ is evidently increasing in $|\rho|$, but so is the potential bias.

Figure 4.10 shows the MSE for five different estimators of β as functions of γ in the model (4.89). The figure would look different if the NCP were on the horizontal axis, but since the NCP is proportional to γ^2, both figures would actually contain the same information. The horizontal line is the MSE of the unrestricted OLS estimator, $\hat{\beta}$. It is the only unbiased estimator here, and therefore it is the only one for which the MSE does not depend on γ.

The MSE of the restricted estimator $\tilde{\beta}$ is lower than MSE($\hat{\beta}$) when γ is sufficiently small. However, as γ increases, MSE($\tilde{\beta}$) increases in proportion to γ^2 (in other words, in proportion to the NCP), rapidly becoming so large that

it is impossible to show it on the figure. If ρ had been larger, $\text{MSE}(\tilde{\beta})$ would have increased even more rapidly.

The other three estimators are all pretest estimators. They differ only in the level of the pretest, which is .10, .05, or .01, and they are therefore denoted $\acute{\beta}_{.10}$, $\acute{\beta}_{.05}$, and $\acute{\beta}_{.01}$. The three MSE functions have similar shapes, but they become substantially more extreme as the level of the pretest becomes smaller. For small values of γ, the pretest estimators are more efficient than $\hat{\beta}$ but less efficient than $\tilde{\beta}$. For very large values of γ, the pretest estimators perform essentially the same as $\hat{\beta}$, presumably because the pretests always reject.

There is a large region in the middle of the figure where the pretest estimators are more efficient than $\tilde{\beta}$ but less efficient than $\hat{\beta}$. The loss in efficiency, especially for $\acute{\beta}_{.01}$, is very substantial over a wide range of values of γ. For each of the pretest estimators, there is also a fairly small region near the point where $\text{MSE}(\tilde{\beta})$ crosses $\text{MSE}(\hat{\beta})$ for which that pretest estimator is less efficient than either $\tilde{\beta}$ or $\hat{\beta}$.

Figure 4.10 makes it clear that the level of the pretest is important. When the level is relatively high, the potential gain in efficiency for small values of γ is smaller, but the potential loss for intermediate values is very much smaller. There is absolutely no reason to use a "conventional" level like .05 for when pretesting.

4.10 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 emphatically do not. In Chapter 6, we will introduce the concept of bootstrap tests, which often work very much better than asymptotic tests when exact tests are not available.

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.

4.11 Exercises

4.1 Suppose that the random variable z follows the $\text{N}(0, 1)$ density. If z is a test statistic used in a two-tailed test, the corresponding P value, according to (4.07), is $p(z) \equiv 2\big(1 - \Phi(|z|)\big)$. 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 $(1/t)f(w/t)$. 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

$$\frac{1}{2\pi} \frac{1}{(1 - \rho^2)^{1/2} \sigma_1 \sigma_2} \exp\left(\frac{-1}{2(1 - \rho^2)} \left(\frac{x_1^2}{\sigma_1^2} - 2\rho \frac{x_1 x_2}{\sigma_1 \sigma_2} + \frac{x_2^2}{\sigma_2^2} \right) \right). \tag{4.90}$$

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

⋆**4.6** Let the random variables x_1 and x_2 be distributed as bivariate normal, with means μ_1 and μ_2, variances σ_1^2 and σ_2^2, and covariance σ_{12}. Using the result of Exercise 4.5, write down the joint density of x_1 and x_2 in terms of the parameters just specified. Then find the marginal density of x_1.

What is the density of x_2 conditional on x_1? Show that the mean of x_2 conditional on x_1 can be written as $E(x_2 \,|\, x_1) = \beta_1 + \beta_2 x_1$, and solve for the parameters β_1 and β_2 as functions of the parameters of the bivariate distribution. How are these parameters related to the least-squares estimates that would be obtained if we regressed realizations of x_2 on a constant and realizations of x_1?

4.7 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.

⋆**4.8** 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.

⋆**4.9** 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

$$\begin{aligned} \boldsymbol{y} &= \boldsymbol{X}_1\boldsymbol{\beta}_1 + \beta_2\boldsymbol{x}_2 + \boldsymbol{u}, \text{ and} \\ \boldsymbol{x}_2 &= \boldsymbol{X}_1\boldsymbol{\gamma}_1 + \gamma_2\boldsymbol{y} + \boldsymbol{v}. \end{aligned} \tag{4.91}$$

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?

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

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

Hint: Equation (4.38) is useful for this exercise.

4.11 Consider regressions (4.42) and (4.41), which are numerically equivalent. Drop the normality assumption and assume that the disturbances are merely IID. Show that the SSR from these regressions is equal to the sum of the SSRs from the two subsample regressions:

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

4.12 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.

4.13 Wald and t. [We could relax implicit restrictions on the house-price regression. However, because sample size is not small, Wald and t tests should yield extremely similar results.]

4.14 Wald and F. [Same comment. Not clear whether we need two questions, or just one.]

4.15 Consider the linear regression model

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

where $\boldsymbol{X}$ is an $n \times k$ matrix. If σ_0 denotes the true value of σ, how is the quantity $\boldsymbol{y}^\top\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{y}/\sigma_0^2$ distributed? Use this result to derive a test of the null hypothesis that $\sigma = \sigma_0$. Is this a one-tailed test or a two-tailed test?

⋆**4.16** P values for two-tailed tests based on statistics that have asymmetric distributions are not calculated as in Section 4.2. Let the CDF of the statistic τ be denoted as F, where $F(-x) \neq 1 - F(x)$ for general x. Suppose that, for

any level α, the critical values c_α^- and c_α^+ are defined, analogously to (4.05), by the equations

$$F(c_\alpha^-) = \alpha/2 \quad \text{and} \quad F(c_\alpha^+) = 1 - \alpha/2.$$

Show that the marginal significance level, or P value, associated with a realized statistic $\hat{\tau}$ is $2\min\big(F(\hat{\tau}), 1 - F(\hat{\tau})\big)$.

4.17 The file **house-price-data.txt** contains 546 observations. Regress the logarithm of the house price on a constant, the logarithm of lot size, and the other ten explanatory variables, as in Exercise 1.23. What is s, the standard error of the regression? Test the hypothesis that $\sigma = 0.20$ at the .05 level. Also compute a P value for the test. **Hint:** See Exercises 4.15 and 4.16.

4.18 Suppose that z is a test statistic distributed as $\mathrm{N}(0, 1)$ under the null hypothesis, and as $\mathrm{N}(\lambda, 1)$ 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 $\alpha = .01$.

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

4.20 Get them to plot a power function that looks weird. Lagged dependent variable?

4.21 Prove that that the F statistic (4.33) is independent of $\tilde{\boldsymbol{\beta}}_1$ under normality. This is true because both s^2 and numerator are independent of it.