Supplement

S.1 Introduction

This *Supplement* to *Estimation and Inference in Econometrics*, by Russell Davidson and James G. MacKinnon, is being written over a period of several years following the publication of the book. It contains a variety of material that was not included in the book at all, or that appeared originally in a different form. Eventually, if there is a second edition, most of this material will find a home there.

The following section is new. The material it discusses applies to all root-n consistent, asymptotically normal estimators. It might be logical to include this section in Chapter 5.

S.2 Functions of Parameter Estimates

In a great many cases, econometricians want to estimate, and make inferences about, functions of parameter estimates. As long as the estimator has the usual properties and asymptotic theory provides a good guide to them, this is very easy to do.

For simplicity, let us start with the single parameter case. Suppose that we have estimated a scalar parameter θ and that we are interested in $\gamma \equiv g(\theta)$, where $g(\cdot)$ is a monotonic function that is continuously differentiable. Assuming that the parameter estimate $\hat{\theta}$ is root-n consistent and asymptotically normal, as most of the estimators discussed in this book are under standard regularity conditions, we know that

$$n^{1/2}(\hat{\theta} - \theta_0) \stackrel{a}{\sim} N\big(0, V^\infty(\hat{\theta})\big), \tag{S.01}$$

where θ_0 denotes the true value of θ and $V^\infty(\hat{\theta})$ is a shorthand way of writing $V^\infty\big(n^{1/2}(\hat{\theta} - \theta_0)\big)$, that is, the asymptotic variance of the expression on the left-hand side of (S.01).

The obvious estimator of γ is $\hat{\gamma} \equiv g(\hat{\theta})$. To determine how $\hat{\gamma}$ is distributed asymptotically, we may Taylor expand $g(\hat{\theta})$ around θ_0 to obtain

$$\hat{\gamma} = g(\theta_0) + g'(\theta^*)(\hat{\theta} - \theta_0), \tag{S.02}$$

where g' is the first derivative of g and, as usual, θ^* is a convex combination of θ_0 and $\hat{\theta}$. Since the consistency of $\hat{\theta}$ implies that $\theta^* \to \theta_0$ as $n \to \infty$, we can replace θ^* by θ_0 without affecting the asymptotic validity of equation (S.02). Rearranging this equation and multiplying both sides by $n^{1/2}$, we conclude that

$$n^{1/2}(\hat{\gamma} - \gamma_0) \stackrel{a}{=} g_0' n^{1/2}(\hat{\theta} - \theta_0), \tag{S.03}$$

where $\gamma_0 \equiv g(\theta_0)$ and $g_0' \equiv g'(\theta_0)$.

From equation (S.03), it is obvious that $n^{1/2}(\hat{\gamma} - \gamma_0)$ is asymptotically normally distributed with mean zero, since it is just g_0' times a quantity that is asymptotically normal with mean zero; recall (S.01). It is also obvious that the variance of $n^{1/2}(\hat{\gamma} - \gamma_0)$ is $(g_0')^2 V^\infty(\hat{\theta})$. Thus we conclude that

$$n^{1/2}(\hat{\gamma} - \gamma_0) \stackrel{a}{\sim} N\big(0, (g_0')^2 V^\infty(\hat{\theta})\big). \tag{S.04}$$

This result is very simple, and it leads immediately to a practical procedure for making inferences about γ. If the estimated variance of $\hat{\theta}$ is $\hat{V}(\hat{\theta})$, then the estimated variance of $\hat{\gamma}$ will be

$$\hat{V}(\hat{\gamma}) = g'(\hat{\theta})^2 \hat{V}(\hat{\theta}). \tag{S.05}$$

This method of estimating the variance is sometimes called the **delta method**.

Although the result (S.04) is simple and practical, it reveals one of the problems with asymptotic theory. Whenever the relationship between $\hat{\theta}$ and $\hat{\gamma}$ is a nonlinear one, it is impossible that they should *both* be normally distributed in finite samples. Suppose that $\hat{\theta}$ really did happen to be normally distributed. Then, unless $g(\cdot)$ were linear, $\hat{\gamma}$ could not possibly be normally, or even symmetrically, distributed, and *vice versa*. This implies that confidence intervals or test statistics based on asymptotic theory may not always be reliable in finite samples.

There is more than one way to construct confidence intervals for θ and γ. Asymptotic theory suggests that we should use symmetric confidence intervals, based on the normal distribution, for both of them. However, that would not be a good thing to do if one of them had an asymmetric finite-sample distribution, which at least one of them must have when $g(\cdot)$ is sufficiently nonlinear. Suppose, for example, that for $\hat{\theta}$ the normality assumption is a good one, that $\hat{V}(\hat{\theta})$ provides an accurate estimate of the variance of $\hat{\theta}$, and that, in consequence, the level α confidence interval for θ is reasonably accurate. This interval is given by

$$\hat{\theta} - c_\alpha \hat{S}(\hat{\theta}) \quad \text{to} \quad \hat{\theta} + c_\alpha \hat{S}(\hat{\theta}), \tag{S.06}$$

where c_α is a two-tail critical value based on the $N(0,1)$ distribution (see Section 3.3), and $\hat{S}(\hat{\theta})$ is the square root of $\hat{V}(\hat{\theta})$. For example, if α were .05, c_α would be 1.96.

A standard asymptotic confidence interval for γ is

$$\hat{\gamma} - c_\alpha \hat{S}(\hat{\gamma}) \quad \text{to} \quad \hat{\gamma} + c_\alpha \hat{S}(\hat{\gamma}), \tag{S.07}$$

where $\hat{S}(\hat{\gamma})$ is the square root of $\hat{V}(\hat{\gamma})$. Instead of using (S.07), however, we could transform the confidence interval (S.06) into a confidence interval for γ. Assuming, for concreteness, that $g' > 0$, the result would be

$$g\big(\hat{\theta} - c_\alpha \hat{S}(\hat{\theta})\big) \quad \text{to} \quad g\big(\hat{\theta} + c_\alpha \hat{S}(\hat{\theta})\big). \tag{S.08}$$

Similarly, we could transform the confidence interval (S.07) into a confidence interval for θ. If $g(\cdot)$ is nonlinear, confidence intervals like (S.08) will be asymmetric. Whether it is better to use a generally asymmetric confidence interval like (S.08) instead of a symmetric interval like (S.07) depends on the finite-sample distributions of both $\hat{\theta}$ and $\hat{\gamma}$. We need to know a good deal about both these distributions before we can make an informed decision about which approach to follow.

The result (S.04) can easily be extended to the case in which both $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\gamma}}$ are vectors. Suppose that the former is a k–vector and the latter is an l–vector, with $l \leq k$. The relation between $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ is $\boldsymbol{\gamma} = \boldsymbol{g}(\boldsymbol{\theta})$, where $\boldsymbol{g}(\cdot)$ is an l–vector of monotonic functions that are continuously differentiable. The vector equivalent of (S.01) is

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} N\big(\mathbf{0}, \boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}})\big), \tag{S.09}$$

where $\boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}})$ is the $k \times k$ asymptotic covariance matrix of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. It is a straightforward exercise to show that the vector equivalent of (S.04) is

$$n^{1/2}(\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \overset{a}{\sim} N\big(\mathbf{0}, \boldsymbol{G}_0 \boldsymbol{V}^\infty(\hat{\boldsymbol{\theta}}) \boldsymbol{G}_0^\top\big), \tag{S.10}$$

where $\boldsymbol{G}_0$ is an $l \times k$ matrix with typical element $\partial g_i(\theta)/\partial \theta_j$, evaluated at $\boldsymbol{\theta}_0$. The asymptotic covariance matrix that appears in (S.10) is $l \times l$, and it will generally have full rank l if the matrix of derivatives $\boldsymbol{G}_0$ has full rank l.

In practice, by analogy with (S.05), the covariance matrix of $\hat{\boldsymbol{\gamma}}$ may be estimated by

$$\hat{\boldsymbol{V}}(\hat{\boldsymbol{\gamma}}) = \hat{\boldsymbol{G}} \hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}}) \hat{\boldsymbol{G}}^\top, \tag{S.11}$$

where $\hat{\boldsymbol{V}}(\hat{\boldsymbol{\theta}})$ is the estimated covariance matrix of $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{G}} \equiv \boldsymbol{G}(\hat{\boldsymbol{\theta}})$. This can be a very useful result in many applications, but, like all results based on asymptotic theory, it should be used with caution.

The following section is new. The material it discusses applies to essentially all asymptotically efficient estimators, including OLS, NLS, ML, and efficient GMM.

S.3 Independence of Tests of Nested Hypotheses

In many cases, the hypotheses that we wish to test may be nested to a depth of more than two. For example, as we saw in Chapter 10, we may test a linear regression model with serially independent errors against a model with AR(1) errors, and we may in turn test the latter against a model with AR(2) errors. In a rather different context, we may use a DWH test to test the null hypothesis that consistent estimates can be obtained by OLS, and we may then test the overidentifying restrictions that are implicit in 2SLS estimation; see Chapter 7. In each of these cases, the models we are interested in form a **sequence of nested hypotheses**.

In the first example above, the nested hypotheses in the sequence can be written as:

$$
\begin{aligned}
H_0: &\quad y_t = \boldsymbol{X}_t\boldsymbol{\beta} + u_t, \\
H_1: &\quad y_t = \rho_1 y_{t-1} + \boldsymbol{X}_t\boldsymbol{\beta} - \rho_1\boldsymbol{X}_{t-1}\boldsymbol{\beta} + u_t, \text{ and} \\
H_2: &\quad y_t = \rho_1 y_{t-1} + \rho_2 y_{t-2} + \boldsymbol{X}_t\boldsymbol{\beta} - \rho_1\boldsymbol{X}_{t-1}\boldsymbol{\beta} - \rho_2\boldsymbol{X}_{t-2}\boldsymbol{\beta} + u_t,
\end{aligned}
$$

where in each case u_t is assumed to be $\text{IID}(0, \sigma^2)$. Under the most restrictive hypothesis, H_0, $\rho_1 = \rho_2 = 0$, while under the least restrictive hypothesis, H_2, there are no restrictions on any of the parameters. The hypothesis H_1 is more restrictive than H_2 but less restrictive than H_0. We may write the relationship among these hypotheses as: $H_0 \subset H_1 \subset H_2$.

There are many other examples of sequences of nested hypotheses. One is testing the hypothesis of serially independent errors against the hypothesis that the errors follow some AR process, and then testing the latter against a model that relaxes the common factor restrictions; see Chapter 10. A second is testing a restricted model estimated by instrumental variables or the generalized method of moments against an unrestricted model, and then testing the overidentifying restrictions on the latter; see Chapter 7 or Chapter 17. A third is testing a restricted simultaneous equations model estimated by FIML against an unrestricted model, and then testing the overidentifying restrictions on the entire system; see Chapter 18. A fourth is testing for structural change in a regression model and then testing whether the error variance is the same for the two parts of the sample in the unrestricted model; see Chapter 11 and Phillips and McCabe (1983). A fifth example is testing a VAR(p) model against a VAR($p+1$) model, and then testing the latter against a VAR($p+2$) model; see Chapter 19.

There may, of course, be more than three nested hypotheses in a sequence. In general, when there are $l + 1$ nested hypotheses, we can write

$$H_0 \subset H_1 \subset H_2 \subset \cdots \subset H_l.$$

Tests of the hypotheses in such a sequence have a very interesting property. Asymptotically, under H_0, the test of H_0 against H_1 is independent of the test of H_1 against H_2, both of these are independent of the test of H_2 against H_3, and so on. For simplicity, we shall henceforth assume that there are just three hypotheses. If the result is true for three hypotheses, then it must be true for any number.

The independence property of tests in a nested sequence has two useful implications. First of all, for test statistics that are asymptotically χ^2, the test of H_0 against H_2 either can be computed as the sum of the two component tests or is asymptotically equivalent to a test that can be computed in this way. This implies that, at least asymptotically, each of the component test statistics is bounded above by the test statistic for H_0 against H_2. Secondly, because the tests are independent, it is very easy to control their overall size. If we want the overall size to be α, the size of the two independent tests, say α^*, must be such that $\alpha = 1 - (1 - \alpha^*)^2$. This implies that

$$\alpha^* = 1 - (1 - \alpha)^{1/2}.$$

Thus if, for example, $\alpha = .05$, we find that $\alpha^* = .02532$. Mizon (1977) makes extensive use of the independence property in the context of model selection.

The result that tests of nested hypotheses are asymptotically independent is true for all of the efficient estimation methods discussed in this book: ordinary least squares (Chapter 3), nonlinear least squares (Chapter 5), generalized least squares (Chapter 9), maximum likelihood (Chapters 8 and 13), instrumental variables (Chapter 7), and efficient GMM estimation (Chapter 17). However, we shall prove it only for two cases: ordinary least squares and maximum likelihood.

The simplest case is that of the linear regression model

$$\boldsymbol{y} = \boldsymbol{X}_0\boldsymbol{\beta}_0 + \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}, \quad \boldsymbol{u} \sim \text{IID}(\boldsymbol{0}, \sigma^2\mathbf{I}).$$

Let H_2 denote the unrestricted model, H_1 denote the restricted model with $\boldsymbol{\beta}_2 = \boldsymbol{0}$, and H_0 denote the doubly restricted model with $\boldsymbol{\beta}_1 = \boldsymbol{0}$ and $\boldsymbol{\beta}_2 = \boldsymbol{0}$. Thus $H_0 \subset H_1 \subset H_2$, as required. Let k_0, k_1, and k_2 denote the number of parameters in $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, and $\boldsymbol{\beta}_2$, respectively.

Using the FWL Theorem, it is straightforward to show that the F statistic for H_0 against H_1 can be written as

$$F_{01} = \frac{\boldsymbol{y}^\top \boldsymbol{M}_0 \boldsymbol{X}_1 \big(\boldsymbol{X}_1{}^\top \boldsymbol{M}_0 \boldsymbol{X}_1\big)^{-1} \boldsymbol{X}_1{}^\top \boldsymbol{M}_0 \boldsymbol{y}/k_1}{\boldsymbol{y}^\top \boldsymbol{M}_{01} \boldsymbol{y}/(n - k_0 - k_1)}, \tag{S.12}$$

where $\boldsymbol{M}_{01}$ projects orthogonally on to $\mathcal{S}^\perp([\boldsymbol{X}_0 \ \ \boldsymbol{X}_1])$; see Section 3.5. Similarly, the F statistic for H_1 against H_2 can be written as

$$F_{12} = \frac{\boldsymbol{y}^\top \boldsymbol{M}_{01} \boldsymbol{X}_2 \big(\boldsymbol{X}_2{}^\top \boldsymbol{M}_{01} \boldsymbol{X}_2\big)^{-1} \boldsymbol{X}_2{}^\top \boldsymbol{M}_{01} \boldsymbol{y}/k_2}{\boldsymbol{y}^\top \boldsymbol{M}_{012} \boldsymbol{y}/(n - k_0 - k_1 - k_2)}, \tag{S.13}$$

where $\boldsymbol{M}_{012}$ projects orthogonally on to $\mathcal{S}^{\perp}([\boldsymbol{X}_0 \ \ \boldsymbol{X}_1 \ \ \boldsymbol{X}_2])$.

Under H_0, the numerators of F_{01} and F_{12} are idempotent quadratic forms in $\boldsymbol{u}$. Asymptotically, their denominators do not matter, since they just tend to σ^2. In fact, we know that k_1 times F_{01} will be asymptotically distributed as $\chi^2(k_1)$, and that k_2 times F_{12} will be asymptotically distributed as $\chi^2(k_2)$. The product of the two idempotent matrices is

$$\boldsymbol{M}_0\boldsymbol{X}_1\big(\boldsymbol{X}_1{}^{\top}\boldsymbol{M}_0\boldsymbol{X}_1\big)^{-1}\boldsymbol{X}_1{}^{\top}\boldsymbol{M}_0\boldsymbol{M}_{01}\boldsymbol{X}_2\big(\boldsymbol{X}_2{}^{\top}\boldsymbol{M}_{01}\boldsymbol{X}_2\big)^{-1}\boldsymbol{X}_2{}^{\top}\boldsymbol{M}_{01} \tag{S.14}$$

Since $\boldsymbol{M}_0\boldsymbol{M}_{01} = \boldsymbol{M}_{01}$ and $\boldsymbol{X}_1\boldsymbol{M}_{01} = \boldsymbol{0}$, expression (S.14) is equal to zero. This implies that the numerators of (S.12) and (S.13) are independent, and so the two test statistics, when expressed in χ^2 form, must be asymptotically independent.

In the linear regression case with NID errors, F statistics for nested hypothesis tests are actually independent in finite samples. Phillips and McCabe (1983) cite the following result from Hogg and Tanis (1963):

If the random variables x_1, x_2, and x_3 are independently distributed as $\chi^2(d_1)$, $\chi^2(d_2)$, and $\chi^2(d_3)$, then

$$F_1 \equiv \frac{x_2/d_2}{x_1/d_1}$$

is independent of

$$F_2 \equiv \frac{x_3/d_3}{(x_1 + x_2)/(d_1 + d_2)}.$$

If we make the definitions

$$\begin{aligned}
x_1 &\equiv \boldsymbol{y}^{\top}\boldsymbol{M}_{012}\boldsymbol{y}/\sigma^2, \\
x_2 &\equiv \boldsymbol{y}^{\top}\boldsymbol{M}_{01}\boldsymbol{X}_2\big(\boldsymbol{X}_2{}^{\top}\boldsymbol{M}_{01}\boldsymbol{X}_2\big)^{-1}\boldsymbol{X}_2{}^{\top}\boldsymbol{M}_{01}\boldsymbol{y}/\sigma^2, \text{ and} \\
x_3 &\equiv \boldsymbol{y}^{\top}\boldsymbol{M}_0\boldsymbol{X}_1\big(\boldsymbol{X}_1{}^{\top}\boldsymbol{M}_0\boldsymbol{X}_1\big)^{-1}\boldsymbol{X}_1{}^{\top}\boldsymbol{M}_0\boldsymbol{y}/\sigma^2,
\end{aligned}$$

and let $d_1 = n - k_0 - k_1 - k_2$ and $d_2 = k_2$, then F_{12} plays the role of F_1 and F_{01} plays the role of F_2. Thus F_{01} and F_{12} are seen to be exactly independent.

Tests of nested hypotheses are exactly independent only in very special cases, notably the one just considered. However, such tests are asymptotically independent in many cases. Consider the case of the three classical tests (LR, LM, and Wald), which were introduced in Chapter 8 and discussed in depth in Chapter 13. Let the test statistic for H_i against H_j be denoted by τ_{ij}. Then, for the LR statistic, we have $\tau_{02} = \tau_{01} + \tau_{12}$ by the way the statistic is constructed. For the other statistics, the same equality must hold asymptotically. We know that τ_{01} is asymptotically distributed as $\chi^2(k_1)$, that τ_{12} is asymptotically distributed as $\chi^2(k_2)$, and that τ_{02} is asymptotically distributed as $\chi^2(k_1 + k_2)$. By a standard result, if we knew that τ_{01} and τ_{12} were asymptotically independent, we could assert that τ_{02} must be asymptotically distributed as $\chi^2(k_1 + k_2)$. What we need to do is to turn this standard

result around, in order to deduce the asymptotic independence of τ_{01} and τ_{12} from the fact that τ_{02} is asymptotically distributed as $\chi^2(k_1 + k_2)$.

Any random variable that is distributed as χ^2 can be represented as a sum of squares of standard normal variates. Let τ^*_{ij} represent the random variable that τ_{ij} tends to asymptotically. Then we have

$$\tau^*_{01} = \sum_{j=1}^{k_1} x_j^2, \qquad \tau^*_{12} = \sum_{j=k_1+1}^{k_1+k_2} x_j^2,$$

where

$$\boldsymbol{x}_1 \equiv [\, x_1 \;\; \ldots \;\; x_{k_1} \,]^\top \sim N(\mathbf{0}, \mathbf{I}_{k_1}), \quad \text{and}$$
$$\boldsymbol{x}_2 \equiv [\, x_{k_1+1} \;\; \ldots \;\; x_{k_1+k_2} \,]^\top \sim N(\mathbf{0}, \mathbf{I}_{k_2}).$$

Further, the two vectors $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are subvectors of a longer vector $\boldsymbol{x}$, which is also multivariate normal:

$$\boldsymbol{x} \equiv [\boldsymbol{x}_1 \vdots \boldsymbol{x}_2] \sim N\left(\mathbf{0}, \begin{bmatrix} \mathbf{I}_{k_1} & \boldsymbol{C} \\ \boldsymbol{C}^\top & \mathbf{I}_{k_2} \end{bmatrix}\right).$$

The assumption of multivariate normality is potentially restrictive, but it will be satisfied automatically by classical test statistics and in many other testing situations. Recall from Section 13.3 that all the classical tests can be written, asymptotically, as quadratic forms in the gradient vector.

The characteristic function of a $\chi^2(k)$ random variable is $(1 - 2it)^{-k/2}$. More generally, if $\boldsymbol{x} \sim N(\mathbf{0}, \boldsymbol{V})$, the characteristic function of $\boldsymbol{x}^\top\boldsymbol{x}$ can be written as

$$\prod_{j=1}^{k} \big(1 + 2iv_j t\big)^{-1/2}, \tag{S.15}$$

where the v_j's are the eigenvalues of the covariance matrix $\boldsymbol{V}$, which are real and positive. For our problem, $k = k_1 + k_2$, and

$$\boldsymbol{V} = \begin{bmatrix} \mathbf{I}_{k_1} & \boldsymbol{C} \\ \boldsymbol{C}^\top & \mathbf{I}_{k_2} \end{bmatrix}. \tag{S.16}$$

Clearly, $\boldsymbol{x}^\top\boldsymbol{x} = \tau^*_{01} + \tau^*_{12}$, and so the characteristic function of $\tau^*_{01} + \tau^*_{12}$ is given by (S.15) with the v_j's being the eigenvalues of (S.16). Now suppose that the sum $\tau^*_{01}+\tau^*_{12}$ is known to have the $\chi^2(k_1+k_2)$ distribution. Therefore, its characteristic function must be

$$\big(1 + 2it\big)^{-k/2}. \tag{S.17}$$

In order for (S.17) and (S.15) to be equal for all real t, it is necessary that $v_j = 1$ for all $j = 1, \ldots, k$. But this means that $\boldsymbol{V} = \mathbf{I}_k$, and consequently that

$\boldsymbol{C} = \boldsymbol{0}$. Thus $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are uncorrelated, and, being multivariate normal, they are therefore stochastically independent. The independence of τ_{01}^* and τ_{12}^* now follows immediately.

What we have just proved is that any two test statistics τ_{01} and τ_{12} are asymptotically independent whenever they tend asymptotically to random variables τ_{01}^* and τ_{12}^* distributed as $\chi^2(k_1)$ and $\chi^2(k_2)$, respectively, and which sum to a random variable τ_{02}^* that is distributed as $\chi^2(k_1 + k_2)$. This result evidently applies to tests of nested hypotheses based on OLS, NLS, IV, and GMM estimation as well as to those based on ML estimation. Thus the independence result discussed in this section is a very general one.

The following section is new. Most of the material it discusses logically belongs in Chapters 5 and 8.

S.4 Sandwich Covariance Matrices

The asymptotic covariance matrices that are encountered in this book all have one of two general forms. Suppose that $\hat{\boldsymbol{\theta}}$ is a root-n consistent, asymptotically normal estimator of a k–vector of parameters. Then, much of the time, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ has the simple form

$$\boldsymbol{V}^{\infty}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = a\boldsymbol{A}^{-1}, \tag{S.18}$$

where a is a scalar, which may of course be equal to 1, and $\boldsymbol{A}$ is a $k \times k$ positive definite matrix. In quite a few other cases, however, the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is more complicated and can be written as

$$\boldsymbol{V}^{\infty}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = a\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{A}^{-1}, \tag{S.19}$$

where $\boldsymbol{B}$ is also a $k \times k$ positive definite matrix. This form of covariance matrix is often called a **sandwich covariance matrix**, for the obvious reason that $\boldsymbol{B}$ is sandwiched between the two instances of $\boldsymbol{A}^{-1}$.

Sandwich covariance matrices are discussed in Chapters 16 and 17, although not under that name. Section 16.3 deals with heteroskedasticity-consistent covariance matrices (HCCMEs) for linear and nonlinear regression models, and Section 17.5 deals with heteroskedasticity and autocorrelation consistent (HAC) covariance matrices for models estimated by GMM. Both the HCCME and HAC covariance matrix estimators have the sandwich form of (S.19), and they may therefore be referred to as **sandwich estimators**. Unfortunately, the book does not provide an adequate treatment of this type of covariance matrix estimator. In particular, it fails to make it clear that covariance matrices like (S.18) arise only in special cases, while ones like (S.19) arise much more generally. Also, although it derives the sandwich covariance matrix for models estimated by maximum likelihood, it fails to discuss the

corresponding sandwich estimator. In this section of the Supplement, we attempt to remedy these two deficiencies.

We first discuss the asymptotic covariance matrix of the NLS estimator $\hat{\boldsymbol{\beta}}$ for the univariate nonlinear regression model $\boldsymbol{y} = \boldsymbol{x}(\boldsymbol{\beta}) + \boldsymbol{u}$. The asymptotic covariance matrix for this model was derived in Section 5.4 under the standard assumption that $E(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma_0^2\mathbf{I}$. For the moment, however, we do not wish to make any assumption about $E(\boldsymbol{u}\boldsymbol{u}^\top)$.

Recall that $ssr^n(\boldsymbol{y}, \boldsymbol{\beta})$ denotes $n^{-1}\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta})\big)$, which is n^{-1} times the sum of squared residuals, written as a function of $\boldsymbol{y}$ and $\boldsymbol{\beta}$. The key equation in Section 5.4 is (5.32), which writes $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ as a function of the first and second derivatives of $ssr^n(\boldsymbol{y}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$. This equation implies that

$$n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \stackrel{a}{=} -\boldsymbol{H}^{-1}(\boldsymbol{y}, \boldsymbol{\beta}_0)\, n^{1/2}\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\beta}_0), \tag{S.20}$$

where $\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\beta})$ denotes the k–vector of first derivatives of $ssr^n(\boldsymbol{y}, \boldsymbol{\beta})$ with respect to $\boldsymbol{\beta}$, and $\boldsymbol{H}(\boldsymbol{y}, \boldsymbol{\beta})$ denotes the $k \times k$ matrix of second derivatives. The notation emphasizes the fact that $\boldsymbol{g}$ is the gradient and $\boldsymbol{H}$ is the Hessian of ssr^n. Equation (S.20) is obtained from (5.32) by evaluating $\boldsymbol{H}$ at the true parameter vector $\boldsymbol{\beta}_0$ instead of at $\boldsymbol{\beta}^*$, a vector that lies between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. The consistency of $\hat{\boldsymbol{\beta}}$ implies that (S.20) holds asymptotically, but it does not hold as an equality in finite samples.

The covariance matrix of the vector $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is the expectation of the vector times itself transposed. Asymptotically, this is equal to the expectation of the vector on the right-hand side of (S.20) times itself transposed. Thus

$$\boldsymbol{V}^\infty\big(n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big) = E\big(\boldsymbol{H}_0^{-1}(n\boldsymbol{g}_0\boldsymbol{g}_0^\top)\boldsymbol{H}_0^{-1}\big), \tag{S.21}$$

where $\boldsymbol{g}_0 \equiv \boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\beta}_0)$ and $\boldsymbol{H}_0 \equiv \boldsymbol{H}(\boldsymbol{y}, \boldsymbol{\beta}_0)$. It is easy to see that, under the DGP characterized by $\boldsymbol{\beta}_0$,

$$\boldsymbol{g}(\boldsymbol{y}, \boldsymbol{\beta}_0) = -\tfrac{2}{n}\boldsymbol{X}_0^\top\big(\boldsymbol{y} - \boldsymbol{x}(\boldsymbol{\beta}_0)\big) = -\tfrac{2}{n}\boldsymbol{X}_0^\top\boldsymbol{u}. \tag{S.22}$$

We saw in Section 5.4 that

$$\operatorname*{plim}_{n\to\infty 0}\boldsymbol{H}(\boldsymbol{y}, \boldsymbol{\beta}_0) = 2\operatorname*{plim}_{n\to\infty 0}\Big(\tfrac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big), \tag{S.23}$$

where plim_0 means that we are taking the probability limit under the DGP characterized by $\boldsymbol{\beta}_0$. The three factors inside the expectations operator in (S.21) are all $O(1)$, and, under reasonable assumptions, they all tend to nonstochastic probability limits. Therefore, we can substitute (S.23) and the plim of (S.22) into (S.21), dropping the expectations operator, so as to obtain

$$\begin{aligned}\boldsymbol{V}^\infty\big(n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\big) = \\ \operatorname*{plim}_{n\to\infty 0}\Big(\tfrac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big)^{-1}\operatorname*{plim}_{n\to\infty 0}\Big(\tfrac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{u}\boldsymbol{u}^\top\boldsymbol{X}_0\Big)\operatorname*{plim}_{n\to\infty 0}\Big(\tfrac{1}{n}\boldsymbol{X}_0^\top\boldsymbol{X}_0\Big)^{-1}.\end{aligned} \tag{S.24}$$

Thus the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is of the sandwich form.

Of course, as we saw in Section 5.4, when $E(\boldsymbol{u}\boldsymbol{u}^\top) = \sigma_0^2\mathbf{I}$, expression (S.24) simplifies to the more familiar result (5.25), which does not have the sandwich form. It is not a coincidence that this simplification is available only in the case for which NLS is asymptotically efficient; see Section 5.5. In general, covariance matrices like (S.18) are available only for estimators that are asymptotically efficient within some class of estimators.

The theoretical result (S.24) can be made to yield operational covariance matrix estimators if we can find ways to estimate the middle matrix consistently. That is precisely what the HCCMEs discussed in Section 16.3 do in the case of errors that are heteroskedastic but serially uncorrelated and what the HAC estimators discussed in Section 17.5 do in the more general case where there is both heteroskedasticity and serial correlation. The former is probably the best-known example of a sandwich estimator in econometrics.

We now turn our attention to maximum likelihood estimation. The asymptotic covariance matrix of the ML estimator $\hat{\boldsymbol{\theta}}$ was derived in Section 8.5. The key equation in this section is (8.38). It can be rewritten in slightly simpler notation as

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{a}{=} -\mathcal{H}_0^{-1} n^{-1/2} \boldsymbol{g}_0, \tag{S.25}$$

where $\mathcal{H}_0$ denotes the expectation of $1/n$ times the matrix of second derivatives of the loglikelihood function with respect to the parameter values, evaluated at $\boldsymbol{\theta}_0$, and $\boldsymbol{g}_0$ denotes the gradient of the loglikelihood function, also evaluated at $\boldsymbol{\theta}_0$. As we showed in Section 8.5, equation (S.25) implies that

$$\boldsymbol{V}^\infty\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big) = \mathcal{H}_0^{-1}\mathcal{I}_0\mathcal{H}_0^{-1}, \tag{S.26}$$

where $\mathcal{I}_0$ is the limiting information matrix evaluated at $\boldsymbol{\theta}_0$. Equation (S.26), which is just equation (8.42) rewritten, shows that the asymptotic covariance matrix of $\hat{\boldsymbol{\theta}}$ is of the sandwich form.

After obtaining the theoretical result (S.26) in Section 8.5, we went on in Section 8.6 to prove the information matrix equality. This famous result tells us that, for a correctly specified model, $\mathcal{I}_0 = -\mathcal{H}_0$. Obviously, if this equality holds, there is no reason to use a sandwich estimator. However, as White (1982) and Gouriéroux, Monfort, and Trognon (1984) showed, the information matrix equality generally will not hold when the DGP is not a special case of the model being estimated, even in cases for which maximum likelihood yields a quasi-ML, or QML, estimator that is consistent. In such cases, the sandwich estimator

$$\boldsymbol{H}^{-1}(\hat{\boldsymbol{\theta}})\,\boldsymbol{G}^\top(\hat{\boldsymbol{\theta}})\boldsymbol{G}(\hat{\boldsymbol{\theta}})\,\boldsymbol{H}^{-1}(\hat{\boldsymbol{\theta}}) \tag{S.27}$$

should be used instead of the estimators (8.49), (8.50), or (8.51) that are discussed in Section 8.6. At present, little seems to be known about the

performance, in finite samples and for particular classes of models, of the sandwich estimator (S.27) relative to the performance of more conventional covariance matrix estimators. However, if (8.49), (8.50), and (S.27) all yield substantially different estimates, it would seem prudent to rely on the last of these. Of course, in this circumstance, it would also seem prudent to investigate the possibility that the model may be misspecified.

The following section is new. The material it discusses applies to all root-n consistent estimators. It might logically be included in Chapter 4.

S.5 Properties of Root-n Consistent Estimators

Although almost all of the estimators we study in this book are root-n consistent under standard regularity conditions (which, however, are not always applicable; see Chapter 20), the properties of root-n consistent estimators are never discussed. In fact, the concept is not even mentioned in Chapter 4, where other types of consistency are discussed in some detail. In this section, we therefore discuss some of the properties of root-n consistent estimators.

Suppose that $\hat{\boldsymbol{\theta}}$ denotes a k–vector of parameter estimates and $\boldsymbol{\theta}_0$ denotes the vector of true parameter values. Then $\hat{\boldsymbol{\theta}}$ is root-n consistent if

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 = O(n^{-1/2}). \tag{S.28}$$

In words, an estimator is root-n consistent if the difference between the estimator and the true value is (stochastically) proportional to $n^{-1/2}$; see Section 4.3. This implies that the covariance matrix of $\hat{\boldsymbol{\theta}}$ must be $O(n^{-1})$, as can be seen by taking the expectation of $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0$ times itself transposed. From (S.28), each element of the resulting matrix must be the expectation of the product of two things that are $O(n^{-1/2})$. Unless these expectations happen to be zero, they must be $O(n^{-1})$.

Although root-n consistency does not imply asymptotic normality, the vast majority of root-n consistent estimators that we will encounter are asymptotically normally distributed. That is,

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \overset{a}{\sim} N\Big(\mathbf{0}, \boldsymbol{V}^{\infty}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big)\Big), \tag{S.29}$$

where $\boldsymbol{V}^{\infty}\big(n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big)$ denotes the asymptotic covariance matrix of the vector $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Since this asymptotic covariance matrix is $O(1)$, it is obvious in this case that the covariance matrix of $\hat{\boldsymbol{\theta}}$ itself must be $O(n^{-1})$.

In Section 4.5, we showed that a consistent estimator might not be asymptotically unbiased. That is not the case for root-n consistent estimators that are asymptotically normal. In fact, for such estimators, we can be sure that, if they are not unbiased, then their bias is at most $O(n^{-1})$. From (S.29), we observe that the mean of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ must be equal to $\mathbf{0}$, asymptotically.

But that could not be the case if $\hat{\boldsymbol{\theta}}$ were biased at $O(n^{-1/2})$ or greater, since then $n^{1/2}$ times the bias would be $O(1)$ or greater. Thus any bias in $\hat{\boldsymbol{\theta}}$ must be $o(n^{-1/2})$.

The above argument does not quite establish what we set out to show. To do this, we must suppose that $\hat{\boldsymbol{\theta}}$ admits a stochastic expansion in powers of $n^{-1/2}$. Under standard regularity conditions, this will be the case. This expansion can be written as

$$\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + n^{-1/2}\boldsymbol{w}_1 + n^{-1}\boldsymbol{w}_2 + O(n^{-3/2}), \tag{S.30}$$

where $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are random k–vectors that are $O(1)$ and independent of n. Multiplying both sides of (S.30) by $n^{1/2}$ and rearranging yields

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \boldsymbol{w}_1 + n^{-1/2}\boldsymbol{w}_2 + O(n^{-1}). \tag{S.31}$$

It is clear from (S.31) that $\boldsymbol{w}_1$ is the random vector to which $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ tends asymptotically. By (S.29), this random vector must have mean vector $\mathbf{0}$. Therefore, the $O(n^{-1/2})$ term in (S.30) cannot contribute to any bias in $\hat{\boldsymbol{\theta}}$. The first term that can do so is $n^{-1}\boldsymbol{w}_2$. Thus the bias is at most $O(n^{-1})$.

Many estimators that are commonly encountered in econometrics are in fact biased at $O(n^{-1})$. Consider, for example, the maximum likelihood estimator of the error variance σ^2 in a linear regression model. As we saw in Section 3.2, dividing *SSR* by $n - k$ yields an unbiased estimator of σ^2. However, the ML estimator divides *SSR* by n instead of by $n - k$, and it is easy to see that this induces a bias that is $O(n^{-1})$:

$$E\Big(\frac{1}{n}SSR\Big) - \sigma_0^2 = \frac{(n-k)\sigma_0^2}{n} - \sigma_0^2 = -\frac{k}{n}\sigma_0^2.$$

There are many other examples of estimators that are biased at $O(n^{-1})$. These include least squares estimators of dynamic regression models (Section 19.4), least squares estimators of time series models (Shaman and Stine, 1988), and maximum likelihood estimators of probit and logit models (Amemiya, 1980b).

Suppose, as is often the case, that the bias of a root-n consistent, asymptotically normal estimator is $O(n^{-1})$. As we have seen, its covariance matrix is also $O(n^{-1})$. Therefore, its mean squared error matrix must be dominated by the latter. Recall that the mean squared error matrix of $\hat{\boldsymbol{\theta}}$ is

$$E\big((\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)^\top\big) = \boldsymbol{V}(\hat{\boldsymbol{\theta}}) + \big(E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big)\big(E(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\big)^\top. \tag{S.32}$$

The first matrix on the right-hand side of (S.32) is $O(n^{-1})$. In contrast, the second is $O(n^{-2})$, since it is the product of two vectors, each of which is $O(n^{-1})$. Therefore, for large sample sizes, we can be confident that the mean squared error which arises from the bias of $\hat{\boldsymbol{\theta}}$ will be small relative to the mean squared error which arises from its variance.

Several readers have suggested that the material on the noncentral chi-squared distribution, which is discussed in Sections 12.4 and B.4, should be supplemented by a figure. That is done in this very short section.

S.6 The Noncentral Chi-squared Distribution

Figure S.1 shows the density of the noncentral χ^2 distribution with 3 degrees of freedom for noncentrality parameters of 0, 2, 5, 10, and 20. As the NCP increases, both the mean and the variance increase, and the distribution becomes more symmetrical. The .05 critical value for the central $\chi^2(3)$ distribution, which is 7.81, is shown in the figure. If a test statistic has the noncentral $\chi^2(3)$ distribution, the probability that the null hypothesis will be rejected at the .05 level is the probability mass to the right of 7.81. It is evident from the figure that this probability will be quite small for small values of the NCP. In contrast, for an NCP of 20, it is .975.

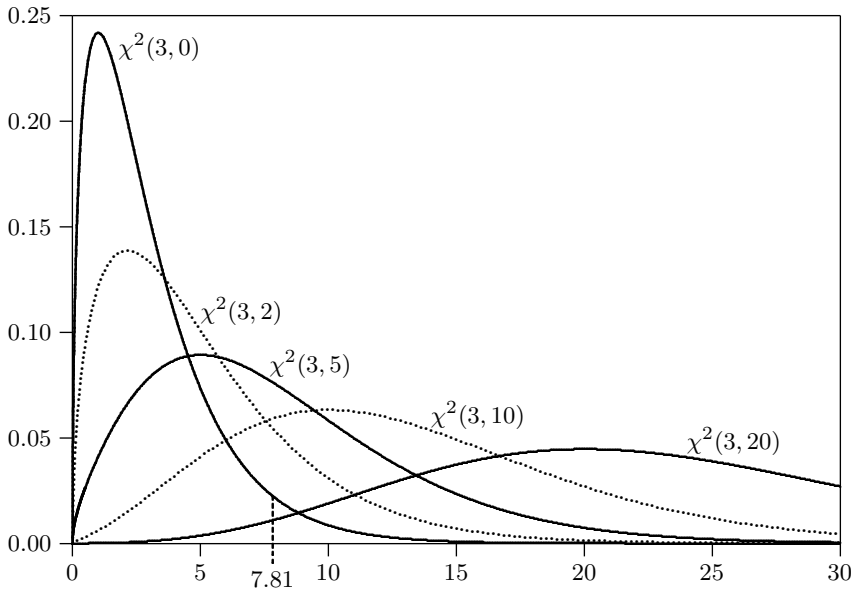

Figure S.1 Densities of noncentral χ^2 distributions

Additional References

Amemiya, T. (1980b). "The n^{-2}-order mean squared errors of the maximum likelihood and the minimum logit chi-square estimator," *Annals of Statistics*, **8**, 488–505.

Hogg, R. V., and E. A. Tanis (1963). "An iterated procedure for testing the equality of several exponential distributions," *Journal of the American Statistical Association*, **58**, 435–43.

Mizon, G. E. (1977). "Inferential procedures in nonlinear models: an application in a UK industrial cross section study of factor substitution and returns to scale," *Econometrica*, **45**, 1221–42.

Phillips, G. D. A., and B. P. McCabe (1983). "The independence of tests for structural change in regression models," *Economics Letters*, **12**, 283–87.

Shaman, P., and R. P. Stine (1988). "The bias of autoregressive coefficient estimators," *Journal of the American Statistical Association*, **83**, 842–48.