Supplementary Notes for Economics 520–522

Mark R. Montgomery

Updated: November 5, 2021

Contents

Part I

Introduction to Statistics and Econometrics

Chapter 1

Notes on Additional Reading

Our discussion in what follows assumes knowledge of the material covered in the first four chapters of Mittelhammer (2013), the textbook for Economics 520, which covers basic probability theory, univariate and multivariate probability distribution functions, expectations, moment-generating functions, and properties of the multivariate normal, χ^2, t, and $\mathcal{F}$ distributions. I've added a few supplementary notes to round out Mittelhammer's discussion in his first four chapters. I also include here some material drawn from other chapters of the Mittelhammer textbook, especially his excellent introduction to asymptotic theory.

Serious students of econometrics will want to have several books on hand that provide complementary perspectives on the material. In the 1980s, Amemiya (1985) published the first textbook treatment of modern asymptotic theory, and his book—which has had a remarkable influence on the field—is still very much worth reading. Amemiya's treatment of the issues has been taken up and expanded upon by several recent textbooks, notably Hayashi (2000), Ruud (2000), Mittelhammer, Judge, and Miller (2000), and more recently, Wooldridge (2010), Davidson and MacKinnon (2004), and Cameron and Trivedi (2005). Among these, Ruud (2000), Wooldridge (2010), and Cameron and Trivedi (2005) are aimed at micro-econometricians (although they give some attention to time-series models) and macro-econometricians with strong interests in time-series will probably find Hayashi (2000) and Davidson and MacKinnon (2004) more appealing. All these textbooks offer rigorous but accessible treatments of econometric theory and make good use of empirical applications to illustrate the theory. Although Greene (2003) is not written with comparable rigor, his book is valuable because of its astonishing breadth and his determination to keep up-to-date with the latest advances in applied econometrics.

In studying econometrics, I have found it essential to have a good advanced calculus text to consult—the Economics 590 syllabus with supplementary notes offers several recommendations. For calculus with attention to statistics, Khuri (1993) is good and so are Davidson (1994) and Bierens (2004) although the latter two books are written at a much higher mathematical level. If you can get hold of it, I would also recommend McFadden (2000), a remarkable set of handouts that he used for his graduate econometrics courses at Berkeley.

Chapter 2

Preliminaries

To understand the recent literature in econometrics, including modern classics such as Newey and McFadden (1994) that are aimed at non-specialists, some background in real analysis and linear algebra is needed to accompany the probability theory that is developed over our three-course econometrics sequence. Much of that material can now be found in the lecture notes for Economics 590. The idea of the present chapter is to collect in one place most of the remaining mathematical material you would require to be an informed consumer of the applied econometrics literature, which is becoming increasingly technical. To go deeper in econometrics, of course, even more mathematical background is necessary, but what's here should give you an adequate starting-point. You can either read the chapter from beginning to end or dip into it as needed.

2.1 Inequalities Involving Expectations

We will describe several inequalities involving expectations that appear often in econometric arguments. The Cauchy–Schwarz, Holder, Liapounov, and Minkowski inequalities are discussed, as is Jensen's inequality.

The lecture notes for Economics 590 present the non-stochastic version of the Cauchy-Schwarz inequality. The stochastic version pertains to random variables X and Y. It is

$$(\mathrm{E}\, YX)^2 \leq \mathrm{E}(X^2) \cdot \mathrm{E}(Y^2).$$

This is proven using the fact that $0 \leq \mathrm{E}(Y - aX)^2$ for a constant a and taking $a = \mathrm{E}(YX)/\mathrm{E}(X^2)$. When the two random variables are defined in terms of deviations from means, i.e., $X \equiv W - \mathrm{E}\, W$ and $Y \equiv Z - \mathrm{E}\, Z$, we obtain

$$(\mathrm{Cov}(Z, W))^2 \leq \mathrm{Var}(Z)\, \mathrm{Var}(W),$$

that is, the square of the covariance between Z and W is less than or equal to the product of their respective variances. This result is used to prove that correlation coefficients are bounded between -1 and 1.

Holder's inequality is useful in its own right, and yields the stochastic version of Cauchy–Schwarz as a special case (see Bierens 2004, p. 51). Consider two random variables X and Y,

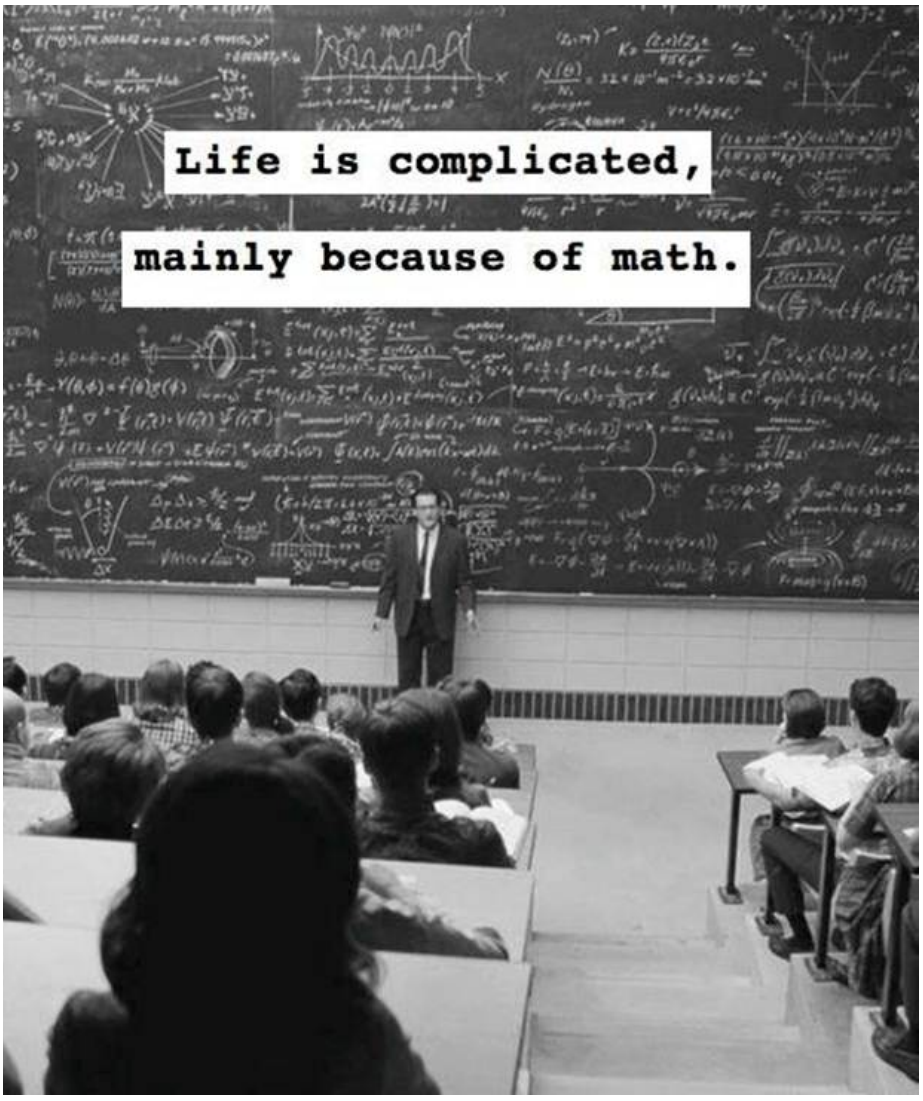


Figure 2.1: Eco 520, by Week 2

and constants p and q such that $p > 1$ and $p^{-1} + q^{-1} = 1$. The inequality is

$$\mathrm{E} \mid XY \mid \leq [\,\mathrm{E}(\mid X \mid^p)\,]^{1/p} \cdot [\,\mathrm{E}(\mid Y \mid^q)\,]^{1/q}.$$

With $p = q = 2$ this is the Cauchy–Schwarz result. The proof uses the fact that $\ln(.)$ is a concave function, so that for two positive scalars x and $y > x$ and a scalar $0 \leq \lambda \leq 1$, we have, by the definition of concave functions,

$$\ln(\lambda \cdot x + (1 - \lambda) \cdot y) \geq \lambda \cdot \ln(x) + (1 - \lambda) \cdot \ln(y).$$

Exponentiating yields

$$\lambda \cdot x + (1 - \lambda) \cdot y \geq x^{\lambda} \cdot y^{1-\lambda}.$$

Now substitute p^{-1} for λ and q^{-1} for $1 - \lambda$. Further substitute $\mid X \mid^p \,/\, \mathrm{E}(\mid X \mid^p)$ for x and do similarly for y. Making these substitutions yields

$$\begin{aligned} p^{-1}\frac{\mid X \mid^p}{\mathrm{E}(\mid X \mid^p)} + q^{-1}\frac{\mid Y \mid^q}{\mathrm{E}(\mid Y \mid^q)} &\geq \left(\frac{\mid X \mid^p}{\mathrm{E}(\mid X \mid^p)}\right)^{1/p} \left(\frac{\mid Y \mid^q}{\mathrm{E}(\mid Y \mid^q)}\right)^{1/q} \\ &= \frac{\mid XY \mid}{(\,\mathrm{E}(\mid X \mid^p)\,)^{1/p}\,(\,\mathrm{E}(\mid Y \mid^q)\,)^{1/q}} \end{aligned}$$

Because the expectation of the left-hand side is 1, the result follows from taking expectations of both sides. Bierens (2004) notes that with $Y \equiv 1$, we obtain *Liapounov's inequality*, which is, for $p \geq 1$,

$$\mathrm{E} \mid X \mid \leq (\,\mathrm{E}(\mid X \mid^p))^{1/p}.$$

Finally, *Minkowski's inequality* is related to the triangle inequality. For $p \geq 1$, it is

$$\mathrm{E} \mid X + Y \mid \leq (\,\mathrm{E}(\mid X \mid^p)\,)^{1/p} + (\,\mathrm{E}(\mid Y \mid^p)\,)^{1/p}$$

assuming of course that the expectations exist. Bierens (2004, p. 52) gives the proof. Note that at $p = 1$, we have $\mid X + Y \mid \leq \mid X \mid + \mid Y \mid$ by the triangle inequality, and the inequality is preserved when we take expectations.

Jensen's inequality for convex or concave functions is proven in Mittelhammer (1996, Chapter 3) and is therefore not reviewed here. See Khuri (1993, p. 233) for a proof of the inequality for twice-differentiable functions. As discussed by Khuri (1993, pp. 230–233), versions of the triangle, Cauchy–Schwarz, Holder, and Minkowski inequalities also apply to integrals.

2.2 The Multivariate Normal Distribution

In what follows, we use the symbol "$\sim$" to denote "distributed as" and represent the standard normal distribution with the notation $\mathcal{N}(0, 1)$. Let $(Y_1, \ldots, Y_n)$ be distributed as multivariate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. We denote this by $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The joint density is given by

$$f(Y_1, \ldots, Y_n) = (2\pi)^{-n/2} \left|\Sigma^{-1}\right|^{1/2} e^{-\frac{1}{2}(\mathbf{Y}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{Y}-\boldsymbol{\mu})}.$$

A special case is $\Sigma = \sigma^2\mathbf{I}$, for which the expression above reduces to

$$f(Y_1, \ldots, Y_n) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\boldsymbol{\mu})'(\mathbf{Y}-\boldsymbol{\mu})}.$$

Another special case, $\Sigma = \sigma^2\mathbf{V}$, will be seen in generalized least squares models. The density is

$$f(Y_1, \ldots, Y_n) = (2\pi)^{-n/2}(\sigma^2)^{-n/2}\left|\mathbf{V}^{-1}\right|^{1/2} e^{-\frac{1}{2\sigma^2}(\mathbf{Y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{Y}-\boldsymbol{\mu})}$$

and often this can be further simplified for certain $\mathbf{V}$ matrices, such as those produced by heteroskedasticity or serial correlation.

On occasion, such as in sample selection models, we will need the following result having to do with multivariate normal random variables. Let $\mathbf{Y}$ be distributed as multivariate normal $(\boldsymbol{\mu}, \Sigma)$ and let $\mathbf{Y}$ be partitioned as $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)'$. Then we can express the relationship of $\mathbf{Y}_1$ to $\mathbf{Y}_2$ in terms that resemble a regression model,

$$\mathbf{Y}_1 = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{Y}_2 - \boldsymbol{\mu}_2) + w_1,$$

where w_1 is itself normally distributed with mean zero and variance $\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. Furthermore, w_1 is independent of $\mathbf{Y}_2$.

2.3 The χ^2 Distribution

To the econometrician, the chi-square distribution stands second in importance only to the normal distribution. In combination with the normal distribution, it provides the basis for the t and $\mathcal{F}$ tests, and is linked as well to the so-called "classical trinity" of tests, the Likelihood Ratio, Wald, and Lagrange Multiplier tests. We should therefore review the relevant properties of the chi-square. Because the chi-square belongs to the larger family of gamma distributions, we begin by stating a key result for the gamma.

The Gamma Distribution

Let the random variable $X > 0$ be gamma-distributed with parameters p and λ. The standard form of the gamma density function is

$$g\,(\,x\,;\,p\,,\,\lambda\,)\;=\;\frac{\lambda^p\;x^{p-1}\;e^{-\lambda x}}{\Gamma(p)}$$

with $p, \lambda > 0$. Here $\Gamma(p)$ is the well-known gamma function (see Ramanathan (1993, p. 71) among others) defined by

$$\Gamma(p)\;=\;\int_0^\infty\;t^{p-1}\,e^{-t}\,dt$$

with the recursion property $\Gamma(p+1) = p\Gamma(p)$; hence for p a positive integer $\Gamma(p) = (p-1)!$. Also, $\Gamma(1/2) = \sqrt{\pi}$. The mean of the gamma is p/λ and the variance is p/λ^2.

Note that by a change-of-variables argument, if X is distributed as gamma (p, λ) then $Y = \lambda X$ is distributed as gamma $(p, 1)$. Hence, λ functions as a scale parameter or dispersion measure for the gamma.

If X_1 and X_2 are two independent gamma-distributed variables with parameters (p, λ) and (q, λ) respectively, then $Y = X_1 + X_2$ is distributed as gamma $(p + q, \lambda)$. It follows that if $X_1, X_2, ...X_n$ are n independent gamma-distributed variables with parameters $\{p_i\}$ and common scale parameter λ, then $Y = \sum X_i$ is distributed as gamma $(\sum p_i, \lambda)$. A proof can be found in Bickel and Doksum (1977, pp. 13–14).

The Central χ^2 Distribution

The central chi-square distribution is a special case of the gamma. With k being a positive integer, insert $p = k/2$ and $\lambda = 1/2$ into the gamma density above. This yields a central chi-square density with k degrees of freedom,

$$g(x; k) \;=\; \frac{x^{(k-2)/2}\, e^{-x/2}}{2^{k/2}\, \Gamma(k/2)}.$$

We say that X is distributed as χ^2_k.

The moment-generating function of the central chi-square is

$$\mathrm{E}\,[\, e^{tX}\,] \;=\; (1 - 2t)^{-k/2}$$

for $t < 1/2$. From this, the mean and variance are found to be

$$\mathrm{E}\;X = k\,, \quad \mathrm{Var}\,X = 2k$$

as shown in Johnson and Kotz (1970a, pp. 167–168). The result also follows from the mean and variance of the gamma distribution.

A relationship of fundamental importance in statistics is that between the chi-square distribution and the sum of squares of independent standard normal variables. If $Z_i \sim \mathcal{N}(0, 1)$ and $Z_1, \ldots, Z_k$ are mutually independent, then

$$Y = \sum_i^k Z_i^2 \sim \chi^2_k.$$

The proof (see Bickel and Doksum (1977, p. 16) or Larsen and Marx (1986, pp. 328–329)) proceeds as follows. First, one shows that the square of each Z_i is distributed as χ^2_1, or equivalently, as gamma$(1/2, 1/2)$. Then, using the result on sums of independent gamma variables with a common scale parameter, it follows that

$$\begin{aligned} Z_1^2 &\sim \chi^2_1 \\ Z_1^2 + Z_2^2 &\sim \chi^2_2 \end{aligned}$$

and so on. In an alternative proof, Taylor (1974, pp. 161–163) derives the moment-generating function of Y and shows it to be identical to that of the chi-square.

The Noncentral χ^2

The noncentral chi-square has a density function

$$g(x; p, \lambda) \;=\; e^{-\lambda} \sum_{i=0}^{\infty} \frac{\lambda^i \, x^{(p+2i-2)/2} \, e^{-x/2}}{i! \, 2^{(p+2i)/2} \, \Gamma((p+2i)/2)}.$$

Here p is the "degrees of freedom" and λ is the non-centrality parameter. Note that the distribution can be viewed as a *mixture* of a Poisson and a central chi-square. That is, for each i in the summation there is a Poisson term $\lambda^i \, e^{-\lambda}/i!$ which multiplies a term recognizable as the density of a central chi-square with $p + 2i$ degrees of freedom.

The noncentral chi-square distribution is of interest principally because of the following result having to do with the sums of squares of $\mathcal{N}(\mu_i, 1)$ random variables. If $Z_i \sim \mathcal{N}(\mu_i, 1)$ and $Z_1, ..., Z_k$ are mutually independent, then $Y = \sum Z_i^2$ is distributed as a noncentral chi-square with k degrees of freedom and noncentrality parameter $\lambda = (1/2) \sum \mu_i^2$. (STATA defines the non-centrality parameter in this way, but other authorities omit the $1/2$ in their definitions of λ, so you should be careful to verify the definition employed before using published tables or computer routines.) The proof is given in Graybill (1961, pp. 74–76), who shows that the moment-generating functions of Y and a noncentral chi-square with parameters $(k, (1/2) \sum \mu_i^2)$ are identical.

As will be seen later, the noncentral chi-square plays a very important role in the theory of the power of tests. Figures 2.2 and 2.3 depict the densities of the central and noncentral χ^2 distributions.

2.4 The t and $\mathcal{F}$ Distributions

If $X \sim \mathcal{N}(0, 1)$ and $Y \sim \chi^2_n$, and if X and Y are independent, then

$$T \;=\; \frac{X}{\sqrt{\frac{Y}{n}}}$$

is distributed as Student's t with n degrees of freedom. That is, the t distribution is produced by the ratio of two independent random variables: in the numerator is a standard normal variable, and in the denominator, the square root of a chi-square variable divided by its degrees of freedom. Clearly this distribution results from the case of $X \sim \mathcal{N}(0, \sigma^2)$ and $Y/\sigma^2 \sim \chi^2_n$, since σ will cancel when X and Y are transformed.

The density function for the t distribution (see Taylor (1974, pp. 163–164)) is

$$g(t; n) \;=\; \frac{\Gamma((n+1)/2)}{\sqrt{n\pi} \, \Gamma(n/2) \, (1 + t^2/n)^{(n+1)/2}}.$$

The central t distribution is symmetric about a zero mean, and approaches $\mathcal{N}(0, 1)$ as the degrees of freedom $n \to \infty$. For $n > 2$ the variance of the t distribution is $n/(n-2)$. To be precise, this is the *central* t distribution; as with the chi-square, there is also a noncentral version (Johnson and Kotz 1970b, p. 201) which is used to determine the power of a t-test.

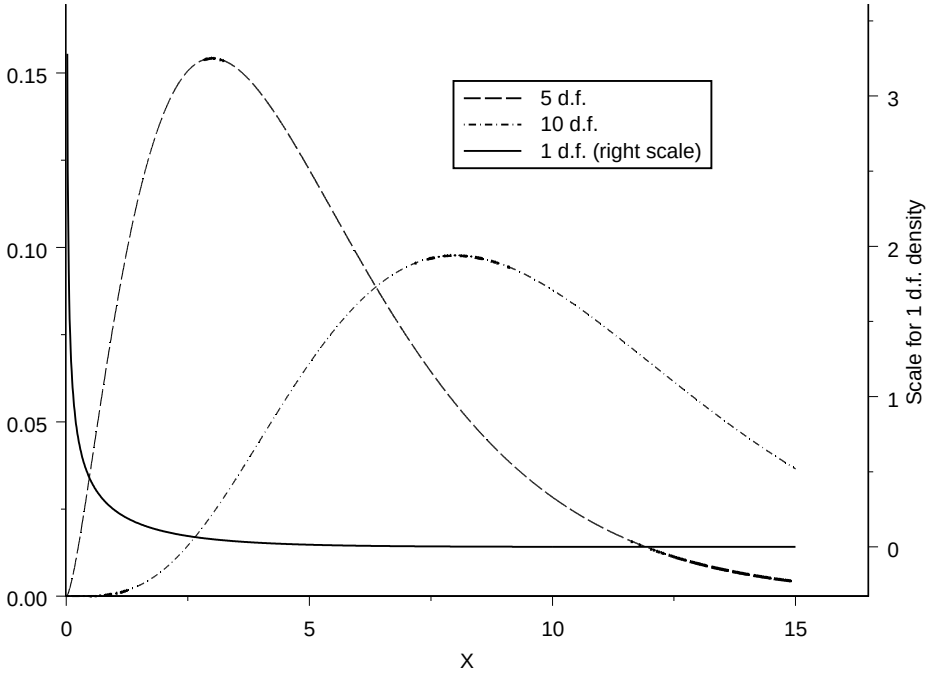


Figure 2.2: Central χ^2 densities

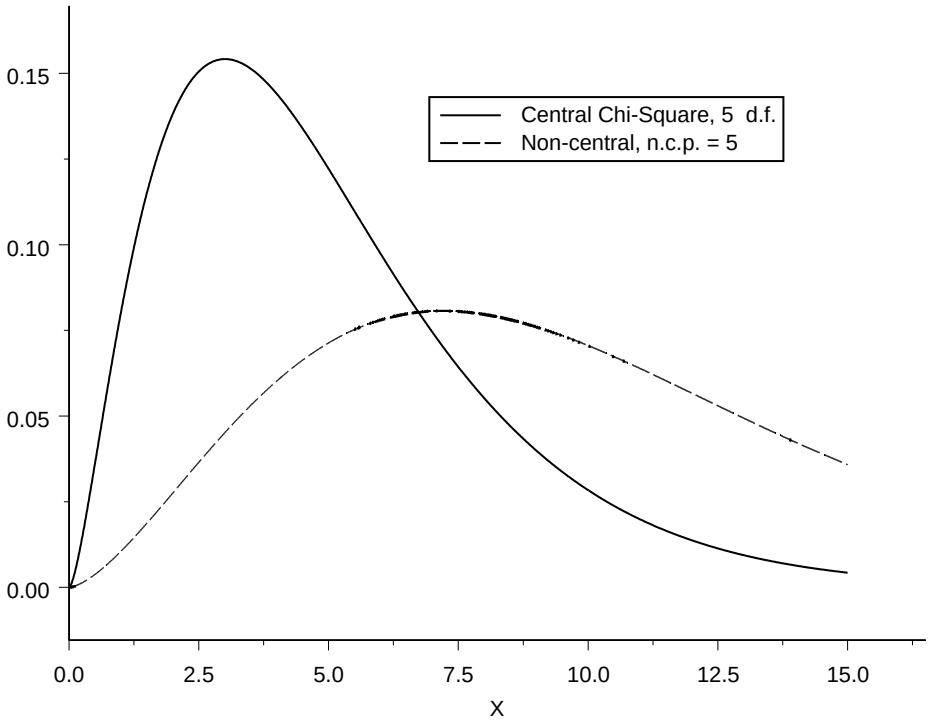


Figure 2.3: Central and non-central χ^2 densities

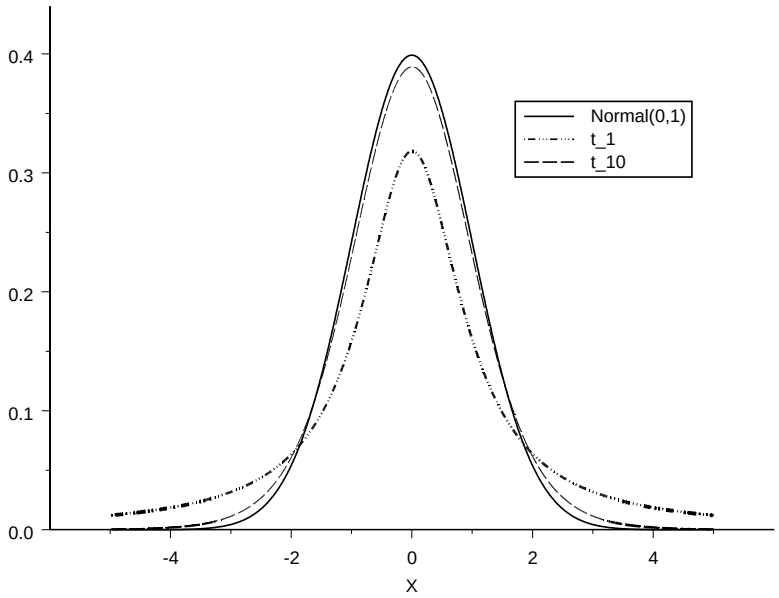


Figure 2.4: Standard Normal and t densities

A non-central t statistic is produced when the numerator is distributed not as standard normal but rather as $\mathcal{N}(\delta, \sigma^2)$, in which case $\lambda = \delta/\sigma$ is the non-centrality parameter.

If $W \sim \chi^2_m$ and $Y \sim \chi^2_n$, and if W and Y are independent, then

$$f = \frac{\frac{W}{m}}{\frac{Y}{n}}$$

is distributed according to the $\mathcal{F}$ distribution with (m, n) degrees of freedom. The density function for the $\mathcal{F}$ distribution (Taylor 1974, pp. 165–167) is

$$g(f; m, n) = \frac{\Gamma((m+n)/2)}{\Gamma(m/2)\Gamma(n/2)}(m/n)^{m/2} f^{(m-2)/2}(1 + (m/n)f)^{-(m+n)/2}.$$

This is the "central" version; there is also a noncentral $\mathcal{F}$ (Johnson and Kotz 1970b, p. 189) in which the numerator is a noncentral chi-square variable (this is useful in deriving the power functions for $\mathcal{F}$ tests), and there is a doubly noncentral $\mathcal{F}$ as well.

The central $\mathcal{F}$ distribution is right-skewed, taking a reversed J-shape for $m \leq 2$ and assuming a unimodal shape for $m > 2$. If $n > 2$ the mean of the distribution is $n/(n-2)$; otherwise the mean does not exist.

The standard normal, t and $\mathcal{F}$ densities are shown in Figures 2.4 and 2.5.

2.5 Quadratic Forms and the χ^2 distribution

We review three important results involving quadratic forms in multivariate normal vectors. These results establish the basis for hypothesis testing in econometrics. To begin, recall (see the Economics 590 notes) that a $n \times n$ matrix $\mathbf{V}$ whose entries are real numbers, and which

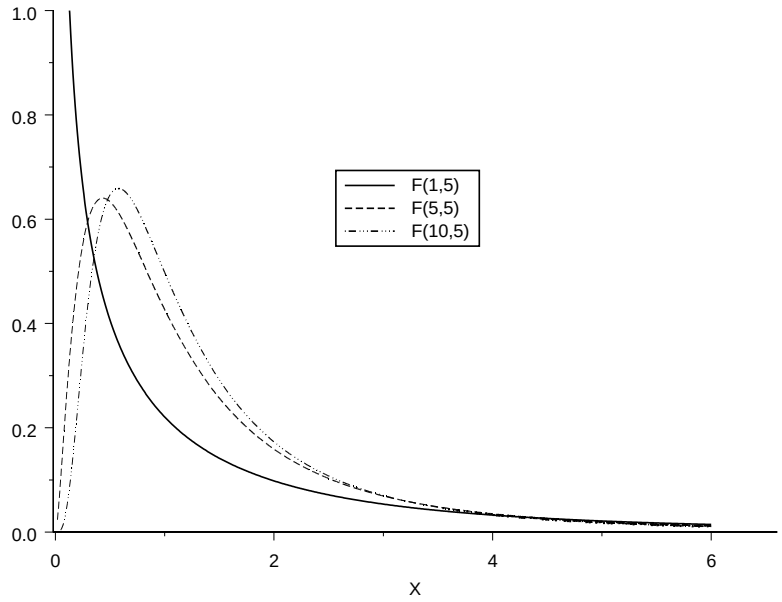


Figure 2.5: $\mathcal{F}$ densities

is symmetric, has eigenvalues that are real numbers. The eigenvalues are not necessarily distinct. The n eigenvectors of the matrix, $\{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_n\}$ can be selected and arranged in a full rank $n \times n$ matrix $\mathbf{P}$ in such a way that $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}$. A compact way of representing all eigenvector–eigenvalue pairings is via

$$\mathbf{V}\mathbf{P} = \mathbf{P}\mathbf{D}$$

in which $\mathbf{D}$ is a diagonal matrix with the eigenvalues of $\mathbf{V}$ on its diagonal. Hence

$$\mathbf{V} = \mathbf{P}\mathbf{D}\mathbf{P}'$$

and if $\mathbf{V}$ happens to be invertible, its inverse would be

$$\mathbf{V}^{-1} = \mathbf{P}\mathbf{D}^{-1}\mathbf{P}'$$

Clearly $\mathbf{D}^{-1}$ will not exist unless $\mathbf{V}$ is invertible. We'll come back to this issue in a moment.

Consider in more detail the matrix

$$\mathbf{D} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}$$

If $\mathbf{V}$ is positive semidefinite, then all of these eigenvalues are non-negative. If $\mathbf{V}$ is positive definite, then all eigenvalues are strictly positive. In either case, we can define a matrix $\mathbf{D}^{1/2}$

$$\mathbf{D}^{1/2} = \begin{bmatrix} \sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_2} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix}$$

such that $\mathbf{D}^{1/2}\mathbf{D}^{1/2} = \mathbf{D}$. Also, for a positive definite $\mathbf{V}$ whose eigenvalues are all positive, we can define

$$\mathbf{D}^{-1/2} = \begin{bmatrix} 1/\sqrt{\lambda_1} & 0 & \cdots & 0 \\ 0 & 1/\sqrt{\lambda_2} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 1/\sqrt{\lambda_n} \end{bmatrix}$$

and $\mathbf{D}^{-1/2}\mathbf{D}^{-1/2} = \mathbf{D}^{-1}$, the latter matrix being diagonal with the reciprocals of the eigenvalues on the diagonal.

From these results we get two nice decompositions. For a positive semidefinite or positive definite matrix $\mathbf{V}$, we can write

$$\mathbf{V} = \mathbf{V}^{1/2}\mathbf{V}^{1/2} = \mathbf{P}\mathbf{D}^{1/2}\mathbf{P}' \cdot \mathbf{P}\mathbf{D}^{1/2}\mathbf{P}'$$

If $\mathbf{V}$ is positive definite, we have this additional result for $\mathbf{V}^{-1}$,

$$\mathbf{V}^{-1} = \mathbf{V}^{-1/2}\mathbf{V}^{-1/2} = \mathbf{P}\mathbf{D}^{-1/2}\mathbf{P}' \cdot \mathbf{P}\mathbf{D}^{-1/2}\mathbf{P}'.$$

Let's now proceed to put these results to work.

1. Let the $n \times 1$ vector $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and let $\mathbf{M}$ be an $n \times n$ symmetric idempotent matrix of rank $k \leq n$. Then $\mathbf{Y}'\mathbf{M}\mathbf{Y} \sim \chi^2_k$.

   The proof is as follows. Recall that the eigenvalues of an idempotent matrix are zeroes and ones. Diagonalize $\mathbf{M}$ as above, such that

   $$\mathbf{P}'\mathbf{M}\mathbf{P} = \mathbf{D}_k = \begin{bmatrix} \mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

   and thus $\mathbf{M} = \mathbf{P}\mathbf{D}_k\mathbf{P}'$. Let $\mathbf{Z} = \mathbf{P}'\mathbf{Y}$. Obviously $\mathrm{E}\,\mathbf{Z} = \mathbf{0}$ and $\mathrm{Var}\,\mathbf{Z} = \mathbf{P}'\mathbf{I}\mathbf{P} = \mathbf{I}$. Hence, $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Using $\mathbf{M} = \mathbf{P}\mathbf{D}_k\mathbf{P}'$,

   $$\mathbf{Y}'\mathbf{M}\mathbf{Y} = \mathbf{Y}'\mathbf{P}\mathbf{D}_k\mathbf{P}'\mathbf{Y} = \mathbf{Z}'\mathbf{D}_k\mathbf{Z} = \sum_{i=1}^{k} z_i^2.$$

   We recognize this as the sum of the squares of k independent standard normal random variables, and such sums are distributed as central χ^2_k. Among other places, we'll make use of this result in establishing the properties of s^2, the estimator of the variance in a regression model.

2. Let the $n \times 1$ vector $\boldsymbol{\epsilon}$ be distributed $\mathcal{N}(\mathbf{0}, \mathbf{V})$, where $\mathbf{V}$ is positive definite (and symmetric, since it is a variance matrix). Then the quadratic form $\boldsymbol{\epsilon}'\mathbf{V}^{-1}\boldsymbol{\epsilon}$ is distributed as χ^2_n.

   Since $\mathbf{V}$ is symmetric positive definite, so is $\mathbf{V}^{-1}$, and we can represent $\mathbf{V}^{-1}$ in the form $\mathbf{V}^{-1} = \mathbf{V}^{-1/2} \cdot \mathbf{V}^{-1/2}$ using the "matrix square root" decomposition given earlier. Note that

   $$\mathbf{Z} \equiv \mathbf{V}^{-1/2}\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$$

which is multivariate standard normal. The quadratic form $\boldsymbol{\epsilon}'\mathbf{V}^{-1}\boldsymbol{\epsilon}$ is the same as $\mathbf{Z}'\mathbf{Z}$, which by the definition of $\mathbf{Z}$ is the sum of squares of n independent standard normals, hence distributed as a central chi-square with n degrees of freedom. This result provides the foundation for the Wald test.[1]

A very important extension is to the case of $\boldsymbol{\epsilon} \sim \mathcal{N}(\delta, \mathbf{V})$, for which the quadratic form $\boldsymbol{\epsilon}'\mathbf{V}^{-1}\boldsymbol{\epsilon}$ is distributed as non-central chi-square with non-centrality parameter $\lambda = (1/2)\delta'\mathbf{V}^{-1}\delta$. This result will be used in deriving the power function of Wald test statistics.

3. Independence of Linear and Quadratic Forms under Normality

   - Let the $n \times 1$ vector $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$, let the $n \times n$ matrix $\mathbf{A}$ be symmetric of rank $p \leq n$, and let $\mathbf{B}$ be a $q \times n$ matrix. Then, if $\mathbf{BA} = \mathbf{0}$, it follows that the $q \times n$ random matrix $\mathbf{BY}$ and the scalar random variable $\mathbf{Y}'\mathbf{AY}$ are independent.

     Greene (2003, Theorem B-12) gives an easy proof in the special case in which $\mathbf{A}$ is symmetric idempotent. In fact this is the case we encounter most often. Let $\mathbf{Y} = \mu + \epsilon$ with $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Let the $q \times 1$ vector $\tilde{\mathbf{Y}} = \mathbf{BY} \sim \mathcal{N}(\mathbf{B}\mu, \sigma^2\mathbf{BB}')$ and let the $n \times 1$ vector $\tilde{\mathbf{Q}} = \mathbf{AY} \sim \mathcal{N}(\mathbf{A}\mu, \sigma^2\mathbf{A})$ because $\mathbf{A}$ is symmetric idempotent. The quadratic form $\mathbf{Y}'\mathbf{AY} = \tilde{\mathbf{Q}}'\tilde{\mathbf{Q}}$ is a simple function of $\tilde{\mathbf{Q}}$. If we can prove that $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Q}}$ are independent, then it follows that $\tilde{\mathbf{Y}}$ and $\tilde{\mathbf{Q}}'\tilde{\mathbf{Q}}$ are also independent.

     With normal random vectors, all we need to do to establish independence is to prove zero covariance. Writing $\tilde{\mathbf{Y}} = \mathbf{B}\mu + \mathbf{B}\epsilon$ and $\tilde{\mathbf{Q}} = \mathbf{A}\mu + \mathbf{A}\epsilon$, we examine the $q \times n$ covariance matrix and find

     $$\mathrm{E}(\tilde{\mathbf{Y}} - \mathbf{B}\mu)(\tilde{\mathbf{Q}} - \mathbf{A}\mu)' = \mathrm{E}\,\mathbf{B}\epsilon\epsilon'\mathbf{A} = \sigma^2\mathbf{BA} = \mathbf{0}$$

     and the proof is done. We make use of this result in t-tests.

     To prove independence with $\mathbf{A}$ being symmetric but not necessarily idempotent, we need a more complicated approach, which is spelled out in Mittelhammer (1996, Theorem 6.11). Let

     $$\mathbf{P}'\mathbf{AP} = \mathbf{D} = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & \lambda_p \\ & & & & \mathbf{0} \end{bmatrix}$$

     and $\mathbf{A} = \mathbf{PDP}'$. Note that $\mathbf{BA} = \mathbf{0}_{q\times n}$ implies $\mathbf{BPP}'\mathbf{A} = \mathbf{0}_{q\times n}$ because $\mathbf{P}$ is orthonormal, and post-multiplying by $\mathbf{P}$ yields $\mathbf{BPP}'\mathbf{AP} = \mathbf{0}$, which we relabel

[1] There is a more general version of the result presented in Graybill (1961, pp. 82–84), which says that if $\boldsymbol{\epsilon}$ is distributed as $\mathcal{N}(\mathbf{0}, \mathbf{V})$, then $\boldsymbol{\epsilon}'\mathbf{B}\boldsymbol{\epsilon}$ is distributed as chi-square with k degrees of freedom if and only if the matrix product $\mathbf{BV}$ is idempotent of rank k. In the case considered above, $\mathbf{B} = \mathbf{V}^{-1}$ and $\mathbf{BV}$ is just the identity matrix.

as $\mathbf{CD} = \mathbf{0}$ with $\mathbf{C}_{q\times n} \equiv \mathbf{BP}$. Write $\mathbf{CD} = \mathbf{0}$ as

$$\begin{bmatrix} & \vdots & \\ \mathbf{C}_1 & \vdots & \mathbf{C}_2 \\ & \vdots & \end{bmatrix} \begin{bmatrix} \lambda & \vdots & \mathbf{0} \\ & \cdots & \\ \mathbf{0} & \vdots & \mathbf{0} \end{bmatrix} = \mathbf{0}$$

where $\mathbf{C}_1$ is of dimension $q \times p$ and $\mathbf{C}_2$ is $q \times (n-p)$. This can be simplified to

$$\mathbf{C}_1 \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{bmatrix} = \mathbf{0}$$

Since the diagonal matrix is invertible, it must be that $\mathbf{C}_1 = \mathbf{0}$. Hence,

$$\mathbf{C} = \begin{bmatrix} \mathbf{0} & \vdots & \mathbf{C}_2 \end{bmatrix}.$$

Let $\mathbf{Z} = \mathbf{P}'\mathbf{Y}$, which yields $\mathbf{Z} \sim \mathcal{N}(\mathbf{P}'\boldsymbol{\mu}, \sigma^2\mathbf{I})$, and because $\mathbf{P}$ is orthonormal, $\mathbf{Y} = \mathbf{PZ}$. Then the scalar random variable

$$\begin{aligned} \mathbf{Y}'\mathbf{AY} &= \mathbf{Z}'\mathbf{P}'\mathbf{APZ} \\ &= \mathbf{Z}'\mathbf{DZ} \\ &= \sum_{i=1}^{p} \lambda_i z_i^2, \end{aligned}$$

which is a function of $\mathbf{Z}_1$, a vector holding the first p elements of the full $\mathbf{Z}$ vector. Also, the random matrix

$$\mathbf{BY} = \mathbf{BPZ} = \mathbf{CZ} \tag{2.1}$$

$$= \begin{bmatrix} \mathbf{0} & \vdots & \mathbf{C}_2 \end{bmatrix} \begin{bmatrix} \mathbf{Z}_1 \\ \cdots \\ \mathbf{Z}_2 \end{bmatrix}, \tag{2.2}$$

in which the vector $\mathbf{Z}_2$ holds the remaining $n-p$ elements of the full $\mathbf{Z}$ vector. Clearly $\mathbf{BY}$ is a function only of $\mathbf{Z}_2$. Since Z is multivariate normal with $\text{Var}\, Z = \sigma^2\mathbf{I}$, we know that $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent. Hence, so are $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{BY}$.

- Let $\mathbf{Y}$ be distributed $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Let $\mathbf{A}$ and $\mathbf{B}$ be symmetric $n \times n$ and positive semidefinite. Then the quadratic forms $\mathbf{Y}'\mathbf{AY}$ and $\mathbf{Y}'\mathbf{BY}$ are independent if $\mathbf{AB} = \mathbf{0}_{n\times n}$. Note that the condition $\mathbf{AB} = \mathbf{0}$ yields an interesting outcome only if both $\mathbf{A}$ and $\mathbf{B}$ are singular. Otherwise either $\mathbf{A}$ or $\mathbf{B}$ or both are $\mathbf{0}$ matrices.

  For the special case of symmetric idempotent $\mathbf{A}$ and $\mathbf{B}$, a variation on the proof above gives us the result (Greene 2003, Theorem B-9). This result is used in $\mathcal{F}$ tests.

Chapter 3

Supplement to Mittelhammer's Chapter 1

What follows are some notes to accompany Mittelhammer (2013, Chapter 1), mainly to explain a couple of points that might have escaped your notice.

First, I think it would have been nice if Mittelhammer had introduced some notation that helps to distinguish between the elementary events and sets in the sample space S that lists all possible outcomes of the experiment or process, on the one hand, from the single point that "nature chooses" as the actual outcome of that experiment or process, on the other. For instance, we might let $y = s_i$ indicate that of all points in S, the outcome that actually materializes is the point s_i. Also, Mittelhammer might have more strongly emphasized that by saying that a subset $A \subset S$ is the outcome, what is meant is that some point $s_i \in A$ is the elemental outcome. When any elementary event $s_i \in A$ occurs, we say that the event A has occurred. Definition 1.8 does in fact say this, to be fair, but the point could use extra emphasis.

In connection with sample spaces S that are uncountable, Mittelhammer tries to steer between an overly-technical approach focused on the possibility of subsets of S that cannot be assigned probability, a possibility that is of interest mainly to measure theorists and other mathematicians who are not really members of the audience for this textbook, and an alternative approach that stresses the features of the theory that have any substantive value, no matter how slight. He does this very well on the whole. As I understand the issues, if S is uncountable, not every conceivable subset of S can be assigned probability in a manner that is consistent with the three probability axioms. So there exist some subsets $A \subset S$ that cannot be assigned probability. Mittelhammer might have prefaced his definitions of the event space (Definition 1.13) and the non-negativity axiom (Axiom 1.1) by reminding the reader that *events* are defined to be subsets of S that *can* be assigned probability even if S is uncountable, and the *event space* is the set composed of all such event/subsets.

Mittelhammer might also have drawn your attention to an assumption that is implicit in the definition of the "additivity" axiom for probability in the case of sample spaces S that are *countably infinite*. Recall that if the sample space is *finite*, then we can represent S in terms of its n elementary events

$$S = \{s_1, s_2, \ldots, s_n\}$$

in which the s_i are the distinct points that compose S, or if you like, you may think of them as one-element sets that are mutually disjoint, that is $s_i \cap s_j = \emptyset \, \forall i \neq j$, whose union is S. For such finite S, the additivity axiom presents no technical difficulties. The condition $1 = P(S)$ is simply

$$P(S) = \sum_{i=1}^{n} P(s_i) = 1$$

and if $A \subseteq S$,

$$P(A) = \sum_{s_i \in A} P(s_i).$$

But things are not quite so clear when S is countably infinite. In this case $S = \{s_1, s_2, \dots\}$, which is an infinite collection of elementary events. The "probabilities sum to 1" requirement must be replaced by a condition on a limit,

$$\lim_{n \to \infty} \sum_{i=1}^{n} P(s_i) = 1,$$

in other words, the limit of this *series*—a sequence $\{S_1, S_2, S_3, \dots\}$ whose terms are the partial sums $S_n = \sum_{i=1}^{n} P(s_i)$—must both exist and be equal to 1.

This requirement puts some restrictions on the "functional form" of $P(s_i)$. As you may recall from your earlier calculus and real analysis courses, there are many tests that can be applied to determine if a series converges (e.g., the ratio test, the Cauchy test). You may also remember that if the series converges, convergence implies that $\lim_{i \to \infty} P(s_i) = 0$. The reverse implication does not hold: having the "tail" probabilities approach 0 does not guarantee that the series converges. Mittelhammer does not point out—-except in the problems at the end of Chapter 1, where you might not have fully appreciated the conceptual point he was making—that the third axiom rules out functional forms for P that do not produce a convergent series.

The case of S being *uncountable* is even trickier. If for example $S = [a, b] \subset \mathbb{R}$, an interval on the real line, then the elementary events are points $s_1 \in [a, b]$, $s_2 \in [a, b]$ and so on. The third axiom requires $P(S) = 1$. But we cannot apply the $\sum_i$ method of ordinary addition to sum up the probabilities of these elementary events. A different concept of addition is needed, one that involves integrals, as becomes evident in the end-of-chapter problems involving uncountable S sample spaces on the real line or in $\mathbb{R}^2$. But Mittelhammer does not explain until Chapter 2 what function would need to be integrated over the real line or in a subset of $\mathbb{R}^2$ to produce a result such as $P(S) = 1$. And the ordinary Reimann method of integration may not be adequate for uncountable sample spaces such as

$$S = [a, b] \cup c$$

in which the isolated point c has positive probability. (These are pretty common sample spaces in economics.) You have to wait for Chapter 2, therefore, to get a fuller accounting of the issues involved.

Chapter 4

Supplement to Mittelhammer's Chapter 2

Having defined probability as a function whose domain is the sample space S and whose range is the interval $[0, 1]$ in $\mathbb{R}$, in this chapter Mittelhammer begins by taking the same kind of approach to defining random variables, which are functions $X(w)$ connecting each elementary event $w \in S$ to a real number $x = X(w)$. Figure 2.1 of Mittelhammer shows how an elementary event $w \in S$ has a *direct image* $X(w) \in \mathbb{R}$. This approach should remind you of the material on direct images that we are currently studying in Economics 590. Likewise, Figure 2.2 links a set $A \subset \mathbb{R}$ back to its *inverse image* $B \subset S$ by way of an *inverse mapping* $X^{-1}(A)$, exactly the sort of thing we've also been looking at in Economics 590.

In the first edition of his textbook, Mittelhammer went to some lengths to demonstrate how the assignment of probability to the values taken on by random variables, which pushes the original sample space S and its event subsets somewhat into the background, is in fact fully consistent with the sample space approach. He has decided for the second edition that since this connection can only be made rigorous through an extensive discussion of direct and inverse images (see his Table 2.1 for a taste), accompanied by detours into measure theory to handle uncountable sample spaces S, this textbook is really not the place for a thorough and complete explication. In an advanced econometrics or mathematical statistics course, by contrast, you would undoubtedly delve into the technical issues. I agree with Mittelhammer's judgement, although I would also note that there are some econometric proofs that are actually made easier when you treat random variables as functions $X(w)$ of elementary events in the original sample space.

There are two things you need to know about this issue:

- The original structure of a sample space S and probability measure $P(A)$ for event $A \subseteq S$ does have to be fully consistent with an alternative structure based on the random variable $X(w)$ for $w \in S$, its range space $R_X = X(S)$ and probability measure $P_X(B_X)$ for $B_X \subseteq R_X$. Showing that the two structures are logically consistent requires a proof. For the purposes of our course, however, we needn't get into all the technical details of that proof.

- The most important component of the proof, which you *do* need to understand, is

in the answer to the following question. Given a set $B_X \subseteq R_X$, how do I assign probability to B_X in a way that is consistent with the original probabilistic structure?

The answer, which involves inverse images, is pretty straightforward. Let $B \subseteq S$ be defined as the inverse image of B_X, that is $B = \{w \in S : X(w) \in B_X\}$. To assign probability to B_X in a consistent manner, we simply set $P_X(B_X) = P(B)$.

4.1 Helpful Examples of Continuous and Discrete Distributions

Please note that in a rare departure from his usually impeccable judgment, Mittelhammer uses the phrase "probability density function" for discrete-valued random variables, to refer to what nearly everyone else in the world of statistics terms their "probability mass function". I'm told that German statisticians use this terminology. I will try to stick with the standard that has been adopted more broadly, and will reserve "density function" for the case of continuous random variables only.

Suppose the random variable Y is discrete, taking on integer values $y = 1$ and $y = 0$ with probabilities p and $1 - p$ respectively. Mittelhammer represents the *range space* for this case as $R_Y = \{0, 1\}$. A compact way to write Y's probability mass function is

$$f_Y(y) = (1 - p)^{1-y} \cdot p^y$$

This kind of random variable is often said to be "Bernoulli-distributed" or termed a "binary" random variable.

This notational approach extends nicely to the case of range spaces with 3 or more values, such as $R_Y = \{0, 1, 2\}$ with associated probabilities $\{1 - p_1 - p_2, p_1, p_2\}$. To represent the probability mass function compactly, it is helpful to define a pair of auxiliary random variables $Y_1 = 1$ if $Y = 1$ and $Y_1 = 0$ otherwise; also, $Y_2 = 1$ if $Y = 2$ and $Y_2 = 0$ otherwise. Then we may write the mass function in terms of the *bivariate vector* $\mathbf{Y} = (Y_1, Y_2)'$ as

$$f_{\mathbf{Y}}(\mathbf{y}) = (1 - p_1 - p_2)^{1-y_1-y_2} \cdot p_1^{y_1} \cdot p_2^{y_2},$$

with $\mathbf{y} = (y_1, y_2)$.

For continuous random variables, Mittelhammer likes to make use of the *exponential* distribution, which applies to strictly positive random variables $Y > 0$, with probability density function

$$f_Y(y) = r \cdot e^{-ry}.$$

whose parameter $r > 0$. The cdf of the exponential distribution is $F_Y(y) = 1 - e^{-ry}$. The median m of the distribution is found by solving

$$\frac{1}{2} = 1 - e^{-r \cdot m}$$

for m, yielding $m = \ln 2/r$. So the higher is the value of r, the smaller is the median. Exponential distributions are often used to represent the waiting time between events.

4.2 Example of a Mixed Discrete-continuous Distribution

These "hybrid" distributions are very important in economics. For an example, consider a study of the length of unemployment that focus on the time it takes for a set of workers, all of whom lose their jobs at time $t = 0$, to locate new work. Our research project is able to follow each of these people for as long as $t = 2$ years, but no longer than that. During that two-year window of observation, we take note of the exact time when a new job is obtained for each person who is lucky enough to find one. Some people continue to look for work, but sadly without success, all the way through the observation window of two years; after that, we do not know what happens to them.

The data-generating process may be specified as follows. Suppose that the *true* length of unemployment U is a continuous random variable that follows the exponential distribution with probability density function $f(u) = r \cdot e^{-r \cdot u}$ in which $r > 0$. But we do not necessarily get to observe this true length; if it happens that $U > 2$ (years), all we can say based on our research design is that whatever the true duration U turns out to be, it is greater than 2. Under the data-generating process we've specified, the probability $\Pr(U > 2) = 1 - \Pr(U \leq 2) = e^{-r \cdot 2}$. Let's agree to give such cases the label of 3, which simply indicates that for the person in question, no job was found while that person was being followed in our research. The value 3 is not really numerically meaningful: all it actually means is that the true length of unemployment (not known) exceeds 2 years.

The distribution of the *unemployment* variable $\tilde{U}$—this is the hybrid random variable whose value we register in our records—is therefore continuously distributed for values of $u \leq 2$, but has a "spike" at the point 3, indicating that no job was found while the person was being observed. The height of the spike equals $e^{-r \cdot 2}$. In other words, the height of the spike at $\tilde{U} = 3$ is the probability that the true $U > 2$. The range space of the hybrid random variable is thus $R_{\tilde{U}} = [0, 2] \cup 3$.

In short, the probability that the *observed* $\tilde{U}$ takes on any specific value of $u \leq 2$ is *zero*—this is a feature of all continuous random variables. But the probability that the observed length of unemployment $\tilde{U}$ equals the specific value 3, an event whose probability equals $\Pr(U > 2)$, is the decidedly non-zero quantity $e^{-r \cdot 2}$.

4.3 Change-of-variables Methods: Univariate

This material is addressed much later in Mittelhammer's textbook than I think is ideal: not until pages 332–339. *Please read these pages* and work through the examples 6.13, 6.14, and 6.15; we will also discuss them in class.

Here's an application. Suppose that X is a standard normal random variable (mean zero, variance 1, as we'll soon see) with density function

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

Let $Y = \sigma X + m$ with $\sigma > 0$ and m a constant. This is a positive monotonic transform of the orginal X. Hence, the cumulative distribution function of Y at point b is

$$F_Y(b) = F_X\left(\frac{b - m}{\sigma}\right)$$

and by differentiating with respect to b, we obtain

$$f_Y(b) = f_x\bigg((b-m)/\sigma\bigg)\,\frac{1}{\sigma}.$$

By substituting into the original density f_X,

$$\begin{aligned} f_Y(b) &= \frac{1}{\sqrt{2\pi}\sigma}\,\exp\left(-\frac{1}{2}\left(\frac{b-m}{\sigma}\right)^2\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma}\,\exp\left(-\frac{1}{2\sigma^2}(b-m)^2\right) \end{aligned}$$

which you may recognize from previous course-work as the density of a normal random variable with mean m and variance σ^2.

Here is an important application of the same ideas, which is a key result in econometric methods using simulations. Suppose that a random variable X has a continuous distribution with a strictly monotonic cumulative distribution function F_X. Then the random variable Y defined as $Y = F_X(X)$ has a uniform distribution. To see this,

$$\begin{aligned} F_Y(y) &= \Pr(Y \leq y) \\ &= \Pr(F_X(X) \leq y) \\ &= \Pr(X \leq F_X^{-1}(y)) \\ &= F_X(F_X^{-1}(y)) \\ &= y \end{aligned}$$

Hence, F_Y is a uniform random variable on the interval $[0,1]$.

4.4 Conditional Densities: Limits of Conditional Probabilities

For continuous random variables Y and X with joint density $f(y,x)$ and marginal densities $h(y)$ and $g(x)$, we often make use of the *conditional density*, a ratio of the form

$$\frac{f(y,x)}{g(x)},$$

when we are thinking of the distribution of Y given $X = x$. But since $\Pr(X = x) = 0$ in the continuous case, this casual practice needs something by way of formal justification.

I will present a lightly-edited version of Mittelhammer's argument here. Begin with a well-defined conditional probability,

$$\Pr(Y \in [a, a+\delta] | X \in [b, b+\Delta]) = \frac{\Pr(Y \in [a, a+\delta], X \in [b, b+\Delta])}{\Pr(X \in [b, b+\Delta])} = \frac{\int_a^{a+\delta}\int_b^{b+\Delta} f(y,x)\,dxdy}{\int_b^{b+\Delta} g(x)\,dx}.$$

Using the mean value theorem for integrals (Mittelhammer 2013, Lemma 2.2), the denominator can be written as

$$\int_b^{b+\Delta} g(x)dx = \Delta \cdot g(x^*)$$

in which $x^* \in [b, b+\Delta]$. Similarly, for a piece of the numerator,

$$\int_b^{b+\Delta} f(y,x)\,dx = \Delta \cdot f(y,\bar{x})$$

in which $\bar{x} \in [b, b+\Delta]$ too. Putting numerator over denominator causes Δ to cancel, leaving

$$\frac{\int_a^{a+\delta} f(y,\bar{x})\,dy}{g(x^*)}$$

assuming that $g(x^*) > 0$. If we now let $\Delta \to 0$, then both $x^* \to b$ and $\bar{x} \to b$. This yields

$$\frac{\int_a^{a+\delta} f(y,b)\,dy}{g(b)},$$

and from here, letting $\delta \to 0$ in

$$\frac{1}{\delta} \cdot \frac{\int_a^{a+\delta} f(y,b)\,dy}{g(b)}$$

gives

$$\frac{f(a,b)}{g(b)},$$

which is the conditional density expression we were striving to justify.

Chapter 5

Supplement to Mittelhammer's Chapter 3

In this chapter I will re-state some results already provided by Mittelhammer (2013) and develop some of the key ideas a bit more.

One thing to be aware of at the outset is that it may or may not make sense to give a substantive interpretation to the expected value of a random variable. If random variable X's values do not have numeric meaning—as is the case with a purely categorical variable whose values are merely numeric labels for its categories—then $\mathrm{E}\, X$ really has little-to-no substantive content. Possibly when X is ordered-categorical a case might be made for the meaning of $\mathrm{E}\, X$, since then the values of X at least indicate ordering, but that is stretching things a bit. Even with categorical variables, however, there are important *functions* of random variables whose expectations we will need. A leading example is in maximum likelihood estimation—here we anticipate material that you will learn next semester—where the expected value of the log-likelihood function $\mathrm{E}\, L(X, \theta)$ provides the fundamental motivation for this method of estimation. (The θ parameters of the distribution of X are what is estimated.) The expectation of the log-likelihood function (and its derivatives) is equally useful for numeric, categorical, and ordered-categorical X variables.

5.1 Existence of Expectations

For bounded random variables, whose range spaces are such that $|x| \leq B$ for all possible values of x, the expectation $\mathrm{E}\, X$ exists (by which we mean, *is finite*). This is because, for both discrete-valued random variables,

$$\mathrm{E}\, X = \sum_i x_i \cdot \Pr(X = x_i) \leq \sum_i |x_i| \cdot \Pr(X = x_i) \leq B \cdot \sum_i \Pr(X = x_i) = B$$

and for continuous variables,

$$\mathrm{E}\, X = \int_{-B}^{B} x \cdot f(x)\, dx \leq \int_{-B}^{B} |x| \cdot f(x)\, dx \leq B \cdot \int_{-B}^{B} f(x)\, dx = B,$$

the expectation is finite.

But when the range space is either countably infinite (in the case of a discrete-valued X) or unbounded, the question of existence needs to be addressed. It may not be obvious from your first pass through the reading, but careful treatments of the concept of expectations make an "if and only if" connection to the *absolutely convergent* criterion for countably infinite sums and unbounded integrals.

Let's take the discrete case. Let X be a discrete-valued random variable with a countably infinite range space ordered as $R_X = \{\ldots, x_{-3}, x_{-2}, x_{-1}, x_0, x_1, x_2, x_3, \ldots\}$ in which the negative subscripts are used to indicate the values of x that are negative. We say, informally, that the expected value of such a random variable exists—i.e., *is finite*—if the sum

$$\sum_{i=a}^{i=b} x_i \Pr(X = x_i)$$

converges (has a finite limit) as $a \to -\infty$ and $b \to +\infty$. If it converges, we write

$$\mathrm{E}\, X = \lim_{\substack{a \to -\infty \\ b \to +\infty}} \sum_{i=a}^{i=b} x_i \Pr(X = x_i)$$

A more precise and careful treatment, however, addresses the existence of expectations in this way:

$$\mathrm{E}\, X \text{ is said to exist if and only if } \mathrm{E}\,|X| \text{ exists .}$$

By this definition of "existence", we are allowed to equate the existence of expectations with satisfaction of the absolutely convergent criterion. That is, we say that $\mathrm{E}\, X$ exists if and only if the right-hand-side limit

$$\lim_{\substack{a \to -\infty \\ b \to +\infty}} \sum_{i=a}^{i=b} |x_i| \Pr(X = x_i)$$

exists. When the limit exists, then $\mathrm{E}\,|X|$ exists and it then follows by definition that $\mathrm{E}\, X$ exists. Note that this definition *does not imply* that the values of $\mathrm{E}\, X$ and $\mathrm{E}\,|X|$ are the same, only that if one of them is finite, then so is the other.

Mittelhammer slips this "if-and-only-if" part of the definition into footnote 2 on page 113—but many otherwise good books pass over the point entirely or touch on it so lightly that it would likely escape your notice. It turns out to be an important (if technical) point to keep in mind when the range space of the random variable includes negative values. If $X \geq 0$, there is of course no difference between ordinary and absolute convergence criteria. So the absolutely convergent aspect (or *absolutely summable* aspect for discrete random variables) of the definition only comes into play explicitly when the range space includes negative values. Nevertheless, it shapes Mittelhammer's strategy of proof for the very important Theorem 3.23 on page 139.

5.2 Expectations and Change-of-variables Issues

Another key point that may not be completely obvious from Mittelhammer's discussion (see Lemma 3.1 and the results that follow), is this: If $Y = \phi(X)$ is a function of a continuous

random variable X, and X has density $f(x)$, then to calculate the expectation of Y we simply integrate as follows,

$$\mathrm{E}\, Y = \int_{R_X} \phi(x) f(x) dx.$$

assuming that this integral is absolutely convergent. Some texts present this expression as if it were the definition of $\mathrm{E}\, Y$ and pass on without further comment. To his credit, Mittelhammer provides a formal proof in the Appendix to Chapter 3.

The proof of the counterpart result for discrete-valued random variables, that

$$\mathrm{E}\, Y = \sum_{R_X} \phi(x) f(x)$$

is trivial when the range space of X is finite, and is just about as easy when it is countably infinite, provided of course that $\sum_{R_X} |\phi(x)| f(x)$ converges so that the terms of the infinite sum $\sum_{R_X} \phi(x) f(x)$ can be handled in any order we wish without affecting the value of the sum. This freedom to re-arrange is an implication of absolute convergence.

For discrete random variables X with finite range space, the proof regarding $\mathrm{E}\, Y$ is a simple application of the "equivalent events" approach. To illustrate, let $R_X = \{-2, -1, 0, 1, 2\}$ and let $Y = \phi(X) = X^2$. Then $R_Y = \{0, 1, 4\}$. Consider $\sum_{R_X} \phi(x) f_X(x)$,

$$4 f_X(-2) + 1 f_X(-1) + 0 f_X(0) + 1 f_X(1) + 4 f_X(2)$$

and rearrange this sum into blocks that are specific to each value of y,

y	$f_Y(y)$
0	$f_X(0)$
1	$f_X(-1) + f_X(1)$
4	$f_X(-2) + f_X(2)$

Clearly if we multiply each value of Y in its range space by its associated f_Y mass, cumulating the results as we go down the rows of this table, we obtain $\mathrm{E}\, Y$.

If the range space of X is countably infinite, we will not necessarily be able to rearrange the sum $\sum_{R_X} \phi(x) f_X(x)$ in any order we like without altering the value of this (infinite) sum. We *can* do this, however, if the sum is *absolutely convergent*. That is, if

$$\sum_{R_X} |\phi(x)|\, f_X(x)$$

exists, then we can reorganize the infinite sum into blocks specific to each value y in the range space R_Y weighted by the sum of the $\phi(x_i)$ terms for which $y = \phi(x_i)$, exactly as we did in the finite range space case.

For continuous random variables, the proof that $\mathrm{E}\, Y = \int_{R_X} \phi(x) f_X(x) dx$ is quite a bit more involved. I will describe some of the machinery of the proof in the next section.

The main point is this: If *all we need to know* about Y is its *expected value*, there is no need to find the probability density or mass function of Y as such. This is a wonderful result, since as we know it can be very difficult to derive Y's density function analytically in the general case. We are equipped to handle *monotonic* ϕ functions, and know generally how to

approach *piecewise monotonic* cases, but the general case is awfully hard to deal with, and would be well-nigh impossible in many multivariate cases.

As must be obvious by now, the existence of $\mathrm{E}\, X$ does not in general imply that $\mathrm{E}\, \phi(X)$ exists. The properties of ϕ matter, as does the nature of the range space of X. As you will recall from calculus, if the range space is bounded, say $R_X = (a, b)$, and $\phi(X)$ is also *bounded*, then

$$\mathrm{E}\, \phi(X) = \int_a^b \phi(x) f(x)\, dx$$

exists. If ϕ is *continuous* and the range space is *closed and bounded*, say $R_X = [a, b]$, then $\mathrm{E}\, X$ exists. But consider the continuous function $\phi(x) = 1/x$ with bounded but half-open range space $R_X = (0, b]$. If the density $f(0) > 0$, then because $1/x \to \infty$ as $x \to 0$, the expectation $\mathrm{E}\, \phi(X)$ does not exist.

5.3 A Useful Alternative Expression for $\mathrm{E}\, X$

In some cases the following expression for the expected value comes in handy. Mittelhammer (2013, Lemma 3.1) presents it in the context of proving other results, where it might easily escape your notice.

To begin, suppose that a *non-negatively* valued random variable X has a mean $\mathrm{E}\, X = \int_0^\infty x f(x)\, dx$. An alternative expression for the mean is

$$\mathrm{E}\, X = \int_0^\infty (1 - F(x)) dx = \int_0^\infty \cdot \int_x^\infty f(z)\, dz \cdot dx$$

Why is this so? One method of proof starts from $\mathrm{E}\, X = \int_0^\infty x f(x)\, dx$ and applies integration by parts.

Recall that if you have an expression that can be viewed as the product of two functions $A(x) \cdot B(x)$, then since

$$\frac{d}{dx} A(x)B(x) = A'(x)B(x) + A(x)B'(x),$$

integrating both sides from $x = 0$ to $x = c$ yields

$$\int_0^c \frac{d}{dx} A(x)B(x) \cdot dx = \int_0^c A'(x)B(x) dx + \int_0^c A(x)B'(x) dx,$$

or

$$A(c)B(c) - A(0)B(0) = \int_0^c A'(x)B(x) dx + \int_0^c A(x)B'(x) dx.$$

This comes in handy when the function you actually want to integrate is in the form $A(x)B'(x)$ as in the integral on the far right. For the expected value, or $\int_0^\infty x f(x)\, dx$, think of $x \equiv A(x)$ and $f(x) \equiv B'(x)$ so that $F(x) \equiv B(x)$. Then making these substitutions,

$$c \cdot F(c) - 0 \cdot F(0) = \int_0^c F(x)\, dx + \int_0^c x f(x)\, dx,$$

or

$$\int_0^c x f(x)\, dx = c \cdot F(c) - \int_0^c F(x)\, dx.$$

Suppose that X is bounded with range space $R_X = [0, c]$. In this case, $F(c) = 1$ and

$$\mathrm{E}\, X = \int_0^c x f(x)\, dx = c \cdot 1 - \int_0^c F(x)\, dx = \int_0^c (1 - F(x))\, dx.$$

Alternatively, let X be unbounded with $R_X = [0, \infty)$ and consider limits as $c \to \infty$. Since $\mathrm{E}\, X$ exists, the limit as $c \to \infty$ of the integral on the left exists, and (with $F(c) \to 1$ as $c \to \infty$) it can be written in terms of what is on the right as

$$\mathrm{E}\, X = \int_0^\infty (1 - F(x))\, dx.$$

Another (and even quicker) route to this answer uses $A(x) \equiv x$ and $B(x) \equiv -(1 - F(x))$, which yields

$$\int_0^c x f(x)\, dx = -c(1 - F(c)) + \int_0^c (1 - F(x))\, dx$$

Then we show that the limit as $c \to \infty$ of the first term on the right is 0, using $0 \leq c(1 - F(c)) = c \int_c^\infty f(x)\, dx \leq \int_c^\infty x f(x)\, dx$, noting that as $c \to \infty$ the integral on the right goes to zero.

Another method of proof proceeds as follows. Consider again

$$\mathrm{E}\, X = \int_0^\infty (1 - F(x)) dx = \int_0^\infty \cdot \int_x^\infty f(z)\, dz \cdot dx$$

The inner of the integrals on the right can be re-expressed in words as "for each x from 0 to ∞, integrate $f(z)$ from $z = x$ to $z = \infty$". If you draw a picture, you can see that this same operation could be described as "for each z from 0 to ∞, integrate $f(z)$ from $x = 0$ to $x = z$, yielding $z \cdot f(z)$". Mathematically, that is, switching the ranges and order of integration

$$\int_0^\infty \cdot \int_x^\infty f(z)\, dz \cdot dx = \int_0^\infty \int_0^z f(z) dx\, dz = \int_0^\infty f(z) \cdot \int_0^z 1 dx \cdot dz = \int_0^\infty z f(z) dz = \mathrm{E}\, X.$$

In Lemma 3.1, Mittelhammer shows how to extend the result to handle random variables that can take on negative as well as non-negative values.

5.4 When $\mathrm{E}\, X^2$ Exists, this implies $\mathrm{E}\, X$ Exists

The proof we consider first is a special case of the more general proof presented by Mittelhammer (2013, Theorem 3.23). Note that the theorem is concerned with *integer powers* of the random variable X. There are additional theorems (not presented by Mittelhammer in Chapter 3) that have to do with $\mathrm{E}\, X^r$ and $\mathrm{E}\, X^s$ in which r and s are real numbers. But these additional theorems have to cope with the fact that if X can assume negative values, then (for example) $X^{1/2}$ is not a real number. The versions of these theorems that I have seen are therefore restricted to random variables whose range spaces are non-negative. Negative X values are not problematic, however, when we consider integer powers of X. As we'll see, in the integer case the negative values of X are fairly easily dealt with.

The strategy of Mittelhammer's proof is interesting. Since by assumption $\mathrm{E}\, X^2$ exists (that is, is finite), we can use this to show that $\mathrm{E}\, |X|$ also exists. That, in turn, implies the

existence of E X by the absolutely convergent criterion. Note that the proof does not address the *value* of E X relative to that of E X^2, but rather shows that E X must be finite.

The proof begins as follows, with an integral that would define the expected value of $|X|$ provided that the integral converges, that is, provided that this expectation exists. We now proceed to show that if E X^2 exists, then so must E $|X|$.

$$\int |x| f(x) dx = \int_{|x|<1} |x| f(x)\, dx + \int_{|x|\geq 1} |x| f(x)\, dx$$

For the first integral on the right,

$$\int_{|x|<1} |x| f(x)\, dx \leq \int_{|x|<1} f(x)\, dx = \Pr(|X| < 1),$$

which of course must be finite. The second piece, $\int_{|x|\geq 1} |x| f(x)\, dx$, obeys two inequalities

$$\int_{|x|\geq 1} |x| f(x)\, dx \leq \int_{|x|\geq 1} x^2 f(x)\, dx \leq \int x^2 f(x)\, dx = EX^2$$

So, taken together, the two pieces of $\int |x| f(x)\, dx$ are each proven to be finite, thus ensuring that E $|X|$ exists.

It is instructive to see how the proof just presented would need to be revised to allow for cases in which X can take on negative values. Suppose we had asserted that the existence of E X^3 implies the existence of E X and tried to replicate the proof above. Consider its last line, modified to put x^3 where x^2 appeared. The problem is that

$$\int_{|x|\geq 1} |x| f(x) dx \text{ is not neccessarily} \leq \int_{|x|\geq 1} x^3 f(x) dx.$$

because x^3 can be negative. The logic of the proof breaks down right here.

But to fix up the proof, remember what it means to say that E X^3 exists: by the (more careful) definition of existence, E X^3 exists *if and only if* E $|X^3|$ exists. So we would re-write the last line in this way:

$$\int_{|x|\geq 1} |x| f(x)\, dx \leq \int_{|x|\geq 1} |x^3| f(x)\, dx \leq \int |x^3| f(x)\, dx = E|X^3|$$

and since the existence of $E|X^3|$ implies (by definition) the existence (finite-ness) of EX^3, we are done.

You should now be able to recognize that essentially the same strategies are employed in Mittelhammer's more general version of the theorem, which has to do with proving that the existence of E X^s implies the existence of E X^r when $s > r > 0$. Unfortunately he does not remind you at the outset of the proof of the careful if-and-only-if definition of "existence" that he is employing. To keep this definition uppermost in mind, it might better to think of the content of Mittelhammer (2013, Theorem 3.23) as follows: For $s > r > 0$, the existence of E $|X^s|$ implies the existence of E $|X^r|$. From there, the definition of existence takes you the rest of the way.

5.5 Stochastic Dominance, Mean-preserving Spreads, and Variance

In studies of choice under uncertainty, economists make use of *expected-utility functions* $u(X)$, where X is a random variable, and both first- and second-order *stochastic dominance* to analyze the value of various choice alternatives.[1] To see the links to variance, we'll return to the integration by parts technique, applying it initially to $\mathrm{E}\,u(X)$, the expected utility derived from a random payoff X, sometimes termed a "lottery".

Expected-utility comparisons

Suppose that X is distributed according to the density function $f_1(x)$ with bounded range space $R_X = [0, c]$. Then the expected utility approach yields value V_1,

$$V_1 = \int_0^c u(x) f_1(x)\, dx$$

or, letting $A(x) \equiv u(x)$ and $B'(x) \equiv f_1(x)$ in an integration-by-parts expression,

$$V_1 = u(c) \cdot F_1(c) - u(0) \cdot F_1(0) - \int_0^c u'(x) F_1(x)\, dx = u(c) - \int_0^c u'(x) F_1(x)\, dx$$

since $F_1(c) = 1$ and $F_1(0) = 0$. We would naturally assume that $u'(x) > 0$. Examining a second distribution for X, which also has range space $[0, c]$, we obtain

$$V_2 = u(c) - \int_0^c u'(x) F_2(x)\, dx.$$

The difference in expected utility between these two distributions can therefore be written

$$V_1 - V_2 = \int_0^c u'(x) \Big(F_2(x) - F_1(x) \Big)\, dx$$

Evidently, so long as $u(x)$ is strictly increasing in x, Distribution 1 must be preferred to Distribution 2 when $F_2(x) - F_1(x) \geq x$ for all $x \in [0, c]$. The two cdfs meet at $x = 0$ (where they are both 0) and $x = c$ (where both are 1), but if elsewhere we have $F_2(x) > F_1(x)$, then distribution 1 will be preferred.

When $F_1(x) < F_2(x)$ for all $x \in (0, c)$, we say that F_1 is *first-order stochastic dominant* over distribution F_2. What we have seen is that in this case, any expected utility maximizer whose $u'(x) > 0$ would prefer Distribution 1.

The integration-by-parts technique can be carried a step further, producing results related to *second-order stochastic dominance*. Returning to V_1,

$$V_1 = u(c) - \int_0^c u'(x) F_1(x)\, dx$$

[1] The following was drawn from some excellent lecture notes for Economics 317, Princeton University, Fall 2007.

apply the technique once more to the integral, making the identifications $A(x) \equiv u'(x)$ and $B'(x) \equiv F_1(x)$ implying $B(x) = \int_0^x F_1(t)\,dt \equiv S_1(x)$, with $S_1(x)$ thus being the area under the cdf F_1 up to point x. Then

$$V_1 = u(c) - \Bigg(u'(c)S_1(c) - u'(0)S_1(0) \Bigg) + \int_0^c u''(x)S_1(x)\,dx$$

and

$$V_1 - V_2 = u'(c)\Bigg(S_2(c) - S_1(c) \Bigg) + \int_0^c u''(x)\Bigg(S_1(x) - S_2(x) \Bigg)\,dx$$

Since $S_1(c)$ and $S_2(c)$ are directly related to the expected values of the two distributions—which are $c - S_1(c)$ and $c - S_2(c)$, respectively—the utility difference has to do with the difference in expected values multiplied by the marginal utility. The second term also needs consideration. With *risk aversion* we have $u''(x) < 0$, that is, the expected-utility function is *concave*. This second derivative of the utility function multiplies the difference in areas under the respective cdfs.

Consider two distributions with the *same expected values* (implying $S_1(c) = S_2(c)$). Distribution 1 is defined to be *second-order stochastic dominant* over Distribution 2 if (and only if) $S_1(x) < S_2(x)$ for all $x \in (0, c)$ and also $S_1(c) = S_2(c)$. With $u''(x) < 0$, the overall expected-utility difference becomes

$$V_1 - V_2 = \int_0^c u''(x)\Bigg(S_1(x) - S_2(x) \Bigg)\,dx > 0$$

under second-order stochastic dominance. Distribution 2 is also said to be a *mean-preserving spread* of Distribution 1 when $S_1(x) < S_2(x)$ for all $x \in (0, c)$ and $S_1(c) = S_2(c)$. The implication is that a risk-averse expected utility maximizer prefers Distribution 1 to any mean-preserving spread of Distribution 1.

The implications for variance

Stepping away from expected-utility maximization as such, let's now suppose that we are examining how the expected value of any twice-differentiable *concave* function $\phi(x)$ differs between a Distribution 1 and a Distribution 2 that is constructed as a mean-preserving spread of Distribution 1. By the analysis above, $\mathrm{E}\,\phi(X)$ would be greater under Distribution 1 than Distribution 2.

Conversely, if the function $v(x)$ we are examining is *convex* with $v''(x) > 0$, then $\mathrm{E}\,v(X)$ would be greater under Distribution 2, which is the mean-preserving spread of Distribution 1. Indeed, one such convex function is $v(x) = (x - \mathrm{E}\,X)^2$ and its expected value is precisely the *variance* of the distribution.

What we have seen in a round-about way is that the variance of Distribution 2—which is a mean-preserving spread of Distribution 1—must be greater than the variance of Distribution 1. This is one of the fundamental reasons why we refer to the variance as a measure of *spread*.

Of course, the variance of one distribution can exceed that of another without second-order stochastic dominance being involved. (The areas under the respective cdfs might not be related in the way that is required for second-order dominance to hold.) But since second-order stochastic dominance implies an ordering of variances—if Distribution 1 is second-order stochastic dominant over Distribution 2, then the variance of Distribution 2 must be higher—for economists this stochastic dominance connection can be helpful when thinking of variance as an indicator of risk.

5.6 Bounds on the Value of a Correlation

The proof that the correlations are bounded by -1 and $+1$ is easier than suggested by Mittelhammer's slightly over-complicated approach. The direct route to the result involves a technique that you will recognize from the proof of the Cauchy–Schwarz inequality.

Consider $\mathrm{E}(Y - aX)^2$ with a a constant, which must be non-negative assuming that the expectation exists. We have

$$\mathrm{E}(Y - aX)^2 = \mathrm{E}\,Y^2 + a^2 E X^2 - 2a\,\mathrm{E}\,XY \geq 0.$$

Now let $a = \mathrm{E}\,XY / \mathrm{E}\,X^2$ (assuming that $\mathrm{E}\,X^2 > 0$) and substitute this into the expression. Simplifying a bit, this step yields

$$\mathrm{E}\,Y^2 \geq \frac{(\mathrm{E}\,XY)^2}{\mathrm{E}\,X^2}$$

or

$$\mathrm{E}\,Y^2 \cdot \mathrm{E}\,X^2 \geq (\mathrm{E}\,XY)^2.$$

Taking positive square roots of both sides and then dividing by $\sqrt{\mathrm{E}\,Y^2 \cdot \mathrm{E}\,X^2}$ (here we are assuming $\mathrm{E}\,Y^2 > 0$) brings us to the essential result

$$1 \geq \frac{|\mathrm{E}\,XY|}{\sqrt{\mathrm{E}\,Y^2 \cdot \mathrm{E}\,X^2}}.$$

Now if we let $X \equiv Z - \mathrm{E}\,Z$ and $Y \equiv W - \mathrm{E}\,W$, we see that the result implies $1 \geq |\rho_{Z,W}|$, with $\rho_{Z,W}$ being the correlation between random variables Z and W.

5.7 Jensen's Inequality

Mittelhammer has a nice proof of Jensen's inequality (see page 121 and the Chapter 3 Appendix), but he neglects to spell out the mathematical basis for the key result on convex functions that he exploits in the proof. The proof itself shows that if g is a convex function, then $\mathrm{E}\,g(\mathbf{X}) \geq g(\mathrm{E}\,\mathbf{X})$, with the random vector $\mathbf{X}$ taking values in $\mathbb{R}^k$. For concave functions $h(\mathbf{X})$, the result is $\mathrm{E}\,h(\mathbf{X}) \leq h(\mathrm{E}\,\mathbf{X})$, which comes from the fact that if h is concave, then $-h \equiv g$ is a convex function.

There are a couple of background facts that may help you to better understand the proof. First of all, if g is a convex function, then its *epigraph*

$$E = \{(\mathbf{x}, y) : y \geq g(\mathbf{x})\}$$

is a *convex set*. Assuming $g : \mathbb{R}^k \to \mathbb{R}$, the domain of g is in $\mathbb{R}^k$ and the epigraph is a convex set in $\mathbb{R}^{k+1}$. The boundary of this set is

$$\{(\mathbf{x}, y) : y = g(\mathbf{x})\} \text{ or simply } \{(\mathbf{x}, g(\mathbf{x}))\}$$

which we would normally describe as the *graph* of the function (picture a scalar x on the horizontal axis and $g(x)$ on the vertical).

Mittelhammer then uses an important result about the epigraph: at any point $(\bar{\mathbf{x}}, g(\bar{\mathbf{x}}))$ on its boundary, there exists (under mild conditions, with the proof involving the Supporting Hyperplane Theorem) what is termed a *subgradient*, a vector $\mathbf{s} \in \mathbb{R}^k$ such that

$$g(\mathbf{x}) \geq g(\bar{\mathbf{x}}) + \mathbf{s}'(\mathbf{x} - \bar{\mathbf{x}})$$

for all $\mathbf{x}$ in the interior of the domain of g. The terms of the subgradient can be reorganized into the form of a hyperplane that supports E at the point $(\bar{\mathbf{x}}, g(\bar{\mathbf{x}}))$. Fuente (2000, pp. 246–251) has a good discussion of these concepts for concave (rather than convex) functions. If you should ever take a course in convex optimization, you would find that subgradients and the like are discussed at great length.

We needn't be concerned with all the mathematical details right now. For us, the key is to cleverly select the boundary point such that $\bar{\mathbf{x}} = \mathrm{E}\,\mathbf{X}$, which then allows us to write the subgradient relationship as

$$g(\mathbf{x}) \geq g(\mathrm{E}\,\mathbf{X}) + \mathbf{s}'(\mathbf{x} - \mathrm{E}\,\mathbf{X})$$

for any $\mathbf{x}$ in the range space of the random vector $\mathbf{X}$. (Well, assuming that the range space of $\mathbf{X}$ can be regarded as the interior of some slightly larger space over which g is defined, since that is formally required for the vector $\mathbf{s}$ to exist.) Rewriting this relationship in terms of the random vector $\mathbf{X}$, we have

$$g(\mathbf{X}) \geq g(\mathrm{E}\,\mathbf{X}) + \mathbf{s}'(\mathbf{X} - \mathrm{E}\,\mathbf{X}).$$

and from this it follows that

$$\mathrm{E}\,g(\mathbf{X}) \geq g(\mathrm{E}\,\mathbf{X}) + \mathbf{s}'(\mathrm{E}\,\mathbf{X} - \mathrm{E}\,\mathbf{X}) = g(\mathrm{E}\,\mathbf{X}).$$

This proves Jensen's inequality for convex functions g, that $\mathrm{E}\,g(\mathbf{X}) \geq g(\mathrm{E}\,\mathbf{X})$ for such functions, assuming of course that the expectations exist. Multiplying g by -1 does the job for concave functions.

5.8 Iterated Expectations

Theorems 3.11 and 3.13 of Mittelhammer are extremely important in econometrics, in which they are known under the heading of *iterated expectations*. Before we embark on an analysis that makes use of Theorem 3.11, let's review this theorem.

Let Y and X be two continuous random variables, with marginal densities $h(y)$ and $g(x)$ and joint density $f(y, x)$. The expected value of Y is $\mathrm{E}\,Y = \int_Y yh(y)dy$ and we know that the marginal density $h(y) = \int_X f(y, x)dx$. Hence,

$$\mathrm{E}\,Y = \int_Y y \int_X f(y, x)\, dxdy = \int_Y \int_X yf(y, x)\, dxdy = \int_X \int_Y yf(y, x)\, dydx$$

with the order of integration having been switched in the last step. Now, within the integral, multiply and divide by the marginal density of X, yielding

$$\mathrm{E}\, Y = \int_X \left(\int_Y y \frac{f(y,x)}{g(x)} dy \right) g(x) dx$$

The expression in parentheses is $\mathrm{E}(Y|X=x)$, which is in general a function of x. Therefore,

$$\mathrm{E}\, Y = \int_X \mathrm{E}(Y|X=x) g(x) dx.$$

This is the enormously useful result that econometricians term *iterated expectations*. It applies to random vectors and matrices as well as to scalar random variables. We often express the result a bit more compactly, as

$$\mathrm{E}\, Y = \mathrm{E}_X \left(\mathrm{E}(Y|X) \right),$$

in which $\mathrm{E}(Y|X)$ is a function of the random variable X, with specific value $\mathrm{E}(Y|x)$ given $X = x$.

Note that using the same approach, we can show that $\mathrm{E}\, YX = \mathrm{E}_X \left(\mathrm{E}(YX|X) \right)$. The route we take for this case involves

$$\mathrm{E}\, YX = \int_Y \int_X yx\, f(y,x)\, dxdy = \int_X \left(\int_Y yx \frac{f(y,x)}{g(x)}\, dy \right) g(x)\, dx.$$

The expression in parentheses is $\mathrm{E}(YX|X=x) = \mathrm{E}(Y|X=x)\, x$. We then integrate over R_X using the marginal density of X to reach the final result. In econometric applications, we will often see cases in which $\mathrm{E}(Y|X=x) = 0$ for all values of $x \in R_X$, from which it is an immediate implication that $\mathrm{E}\, YX = 0$ as well.

One result that you will use many times links the conditional expectation of Y given X to their covariance, $\mathrm{Cov}(Y,X) = \mathrm{E}(Y - \mathrm{E}\, Y)(X - \mathrm{E}\, X) = \mathrm{E}\, YX - \mathrm{E}\, Y \cdot \mathrm{E}\, X$. If we happen to know that $\mathrm{E}(Y|X) = 0$, then it follows from iterated expectations that $\mathrm{E}\, Y = 0$ and also $\mathrm{E}\, YX = 0$, which establishes that the covariance must be zero.

A variation on this argument can be used to show that

$$\mathrm{E}(Y \mid Z) = \mathrm{E}_{W|Z} \left(E(Y \mid Z, W) \right).$$

Let $f(y,z)$, $g(y,z,w)$ and $h(z,w)$ be the various joint densities and $b(z)$ the relevant marginal density. Then

$$\mathrm{E}(Y \mid Z = z) = \int_Y y\, \frac{f(y,z)}{b(z)}\, dy = \frac{1}{b(z)} \int_Y y \cdot \int_W g(y,z,w)\, dw \cdot dy$$

and therefore, upon multiplying and dividing by $h(z,w)$,

$$\frac{1}{b(z)} \int_Y y \cdot \int_W g(y,z,w)\, dw \cdot dy = \frac{1}{b(z)} \int_Y \int_W y \cdot \frac{g(y,z,w)}{h(z,w)}\, h(z,w) dw \cdot dy.$$

From here, we switch the order of integration to obtain

$$\frac{1}{b(z)} \int_W \cdot \int_Y y \, \frac{g(y,z,w)}{h(z,w)} \, dy \cdot h(z,w) \, dw = \int_W E(Y \mid Z = z, W = w) \cdot \frac{h(z,w)}{b(z)} \, dw,$$

which is the result we set out to prove.

If you become interested in the econometrics of randomized experiments, and the tools used to understand how the average "treatment effects" can be estimated from such experiments, you will find yourself in the midst of extensive applications of the method of iterated expectations. In particular, our last result above is central to the proof of the properties of the propensity score method that is commonly used for evaluation when program participation involves self-selection on the part of individuals, but where the mechanism of self-selection can be assumed to be "strongly ignorable". See the discussion in Chapter 33 for the details.

5.9 Optimal Forecasts

An immediate application of iterated expectations is in showing that the conditional mean $h^*(\mathbf{X}_t) = \mathrm{E}(Y_{t+1}|\mathbf{X}_t)$ is the function that minimizes expected squared forecast error. Think of Y_{t+1} as a random variable that summarizes some aspect of the future state of the economy and think of $h(\mathbf{X}_t)$ as a function of a vector of current predictor variables, say, those entering the time-t information set on which the forecast is to be based. The problem is to search over all possible $h(\mathbf{X}_t)$ functions to

$$\min_{h(\mathbf{X}_t)} \mathrm{E}(Y_{t+1} - h(\mathbf{X}_t))^2$$

At first glance, this does not appear to be a mathematically well-posed problem. But consider re-writing what is in the parentheses as $Y_{t+1} - \mathrm{E}(Y_{t+1}|\mathbf{X}_t) - (h(\mathbf{X}_t) - \mathrm{E}(Y_{t+1}|\mathbf{X}_t))$, squaring the result and then using iterated expectations (conditioning first on $\mathbf{X}_t$) to find $\mathrm{E}(Y_{t+1} - h(\mathbf{X}_t))^2$. It is easy to see that the optimal $h(\mathbf{X}_t)$ function—the one that minimizes the expected squared forecast error—is $h^*(\mathbf{X}_t) = \mathrm{E}(Y_{t+1}|\mathbf{X}_t)$. This famous result about the conditional mean provides the foundation for much of economic forecasting. Mittelhammer (2013, page 130–131) explains the details of the proof, but you should be able to work it out without his help. Note that by iterated expectations, the expected value of the optimal forecast function is just the unconditional mean of Y_{t+1}, that is $\mathrm{E}\, Y_{t+1} = \mathrm{E}_{X_t}\left(\mathrm{E}(Y_{t+1}|X_t)\right)$. (Make sure that you understand this point—the forecast function depends on the random vector $\mathbf{X}_t$.)

Also, we can express the actual Y_{t+1} as the sum of its conditional mean given X_t and a forecast error ϵ_{t+1} that has conditional mean zero (by construction) and thus (by iterated expectations) also has an unconditional mean of zero:

$$Y_{t+1} = \mathrm{E}(Y_{t+1}|\mathbf{X}_t) + \epsilon_{t+1}.$$

Since the forecast error ϵ_{t+1} has mean zero, the expected value of ϵ_{t+1}^2 is the variance of the forecast error; that is

$$\mathrm{Var}\, \epsilon_{t+1} = \mathrm{E}\left(Y_{t+1} - \mathrm{E}(Y_{t+1}|\mathbf{X}_t)\right)^2$$

Another application of iterated expectations shows that the covariance of the optimal forecast $\mathrm{E}(Y_{t+1}|\mathbf{X}_t)$ and the forecast error ϵ_{t+1} is zero. (Confirm this.) Hence, from

$$Y_{t+1} = \mathrm{E}(Y_{t+1}|\mathbf{X}_t) + \epsilon_{t+1}$$

we find

$$\operatorname{Var} Y_{t+1} = \operatorname{Var}\left(\mathrm{E}(Y_{t+1}|\mathbf{X}_t)\right) + \operatorname{Var}\epsilon_{t+1}$$

because the covariance term disappears. Another way to express this result is to rearrange it as

$$\operatorname{Var}\epsilon_{t+1} = \operatorname{Var} Y_{t+1} - \operatorname{Var}\left(\mathrm{E}(Y_{t+1}|\mathbf{X}_t)\right),$$

which shows that the variance of the forecast error equals the difference between the variance of Y_{t+1} and the variance of the forecast function.

We needn't cast the optimal forecast problem literally in terms of predicting a future variable given a vector of current-period variables. The analysis above applies equally well to any problem that can be framed in terms of choosing an $h(\mathbf{X})$ function so as to minimize $\mathrm{E}(Y - h(\mathbf{X}))^2$. The conditional mean $\mathrm{E}(Y|\mathbf{X})$ is the optimal "predictor" for a problem of this kind.

It is sometimes interesting to examine a forecasting function $h(\mathbf{X})$ that is constrained to be *linear*, equalling $a + bX$ when X is a single random variable. It is easy to verify that the constrained minimization problem

$$\min_{a,b} \mathrm{E}\left(Y - (a + bX)\right)^2$$

has the solution $b^* = \operatorname{Cov}(Y, X)/\operatorname{Var} X$ and $a^* = \mathrm{E}\,Y - b^* \cdot \mathrm{E}\,X$. Hence the optimal linear forecast function is

$$\mathrm{E}\,Y + \frac{\operatorname{Cov}(Y, X)}{\operatorname{Var} X}(X - \mathrm{E}\,X).$$

as shown by Mittelhammer (2013, Theorem 3.37, page 156–157). This expression is very closely related to the quantities we'll examine in ordinary least squares regression models. It is sometimes described as the "best linear predictor" of Y given X.

The meaning of "best" in "best linear predictor" needs to be clearly understood. In general, the best linear predictor is decidedly sub-optimal when set against the conditional mean $\mathrm{E}(Y|X)$. In other words, the minimized mean squared error achieved by the linear predictor either exceeds or in the best case equals the level achieved by the conditional mean. The two predictors obviously coincide in the special case in which the conditional mean is linear in X, a case that we now investigate in more detail.

5.10 Linear Conditional Expectations

Suppose that $\mathrm{E}(Y|X) = a + b \cdot X$, that is, assume that the conditional expectation of Y given X is linear in X. (This might seem like an unusual special case, but in econometric models the linearity assumption is quite commonly invoked. When there are multiple X variables, linearity is not as restrictive an assumption as it might at first appear.) One

interesting feature of linear conditional expectations is that as the correlation between Y and X approaches either $+1$ or -1, the probability that the random variable Y departs from the conditional mean $a + bX$ approaches zero. Since both Y and X are random variables, we should think about this in terms of a "collapse" of their joint distribution along the line $a + bX$.

We will prove this using Markov's inequality. As you know,

$$\Pr\left((Y - (a + bX))^2 > c\right) \leq \frac{\mathrm{E}(Y - (a + bX))^2}{c}.$$

We want to re-express the numerator of the right-hand side in terms of the correlation between Y and X. To do so, we proceed as follows.

Since by assumption the conditional mean of Y is linear in X, then the values derived above, whereby $b = \mathrm{Cov}(Y, X)/\operatorname{Var} X$ and $a = \mathrm{E}\, Y - b \cdot \mathrm{E}\, X$, also describe the conditional mean. Inserting these values into $(Y - (a + bX))^2$, we have

$$(Y - (a + bX))^2 = (Y - \mathrm{E}\, Y)^2 + b^2(X - \mathrm{E}\, X)^2 - 2b(Y - \mathrm{E}\, Y)(X - \mathrm{E}\, X).$$

Taking expectations, using the expression for b, and simplifying, we obtain the following expression for the mean squared prediction error,

$$\mathrm{E}(Y - (a + bX))^2 = \operatorname{Var} Y - \frac{(\mathrm{Cov}(Y, X))^2}{\operatorname{Var} X} = \operatorname{Var} Y \cdot (1 - \rho_{Y,X}^2).$$

It equals the overall variance of Y scaled down by a factor that takes the correlation between Y and the predictor X into account. Therefore, returning to Markov's inequality,

$$\Pr\left((Y - (a + bX))^2 > c\right) \leq \frac{\operatorname{Var} Y \cdot (1 - \rho_{Y,X}^2)}{c}$$

for any $c > 0$. For any fixed value of c, no matter how small, as $\rho_{Y,X} \to \pm 1$ the numerator approaches zero. In other words, the probability of a squared prediction error larger than c approaches zero.

Note carefully that this analysis *assumes* that the conditional mean of Y given X is linear in X. In the linear case, as $\rho_{Y,X} \to \pm 1$, the joint distribution of Y and X essentially "collapses" on the line $a + bX$ that is the conditional mean function. We will see an illustration of this either in class or recitation.

5.11 Optimal Linear Forecasts and Their Errors

Mittelhammer (2013, p. 157) puts these same ideas to work in the context of forecasting, concentrating on the case in which the conditional expectation of Y is linear in X, and therefore the optimal forecast function and the optimal linear forecast function coincide. Given such an optimal linear forecast $Y^f \equiv a^* + b^* X$, with the coefficients determined as

above, we can write $Y = Y^f + \epsilon$ with ϵ being the forecast error. Using the solution for the optimal coefficients, we have

$$\epsilon = Y - Y^f = (Y - \mathrm{E}\, Y) - \frac{\mathrm{Cov}(Y, X)}{\mathrm{Var}\, X}(X - \mathrm{E}\, X)$$

and taking expectations of both sides, we see that $\mathrm{E}\,\epsilon = 0$, that is, the forecast error has mean zero. (We already knew this from our analysis of the forecast error using the conditional mean.) Since it has mean zero, the expected squared forecast error is the variance of the forecast error. Above we worked out an expression for this forecast error variance in the special case of linear conditional expectations,

$$\mathrm{E}\,\epsilon^2 = \mathrm{Var}\,\epsilon = \mathrm{Var}\, Y \cdot (1 - \rho^2_{Y,X}).$$

Mittelhammer goes on to show that the overall variance of Y can be decomposed into the variance of the forecast Y^f and the variance of the forecast error. (We have already seen this general result using the globally optimal forecast function, the conditional expectation.) The main point to take away from all this is that Y can be expressed as $Y = Y^f + \epsilon$, the sum of the optimal linear forecast Y^f and a forecast error ϵ that is uncorrelated with Y^f.

5.12 Forecasting Binary Variables: "Machine learning" Methods

Suppose that we are interested in forecasting Y_{t+1} in the special case in which Y_{t+1} takes on values 1 and 0 with conditional probabilities $p(\mathbf{X}_t) = \Pr(Y_{t+1} = 1|\mathbf{X}_t)$ and $1 - p(\mathbf{X}_t)$ respectively, $\mathbf{X}_t$ being a vector of time-t explanatory variables. Assume that for all values of the explanatory variables, we have $0 < p(\mathbf{X}_t) < 1$. As we've just seen, the optimal forecast function is the conditional mean of Y_{t+1} given $\mathbf{X}_t$, and in the case of a binary variable, the conditional mean is $h^*(\mathbf{X}_t) = p(\mathbf{X}_t)$. But this answer, while optimal in the sense that it minimizes the mean squared forecast error, nevertheless feels unsatisfactory for the problem at hand, because the support of Y_{t+1} is limited to the two integers $\{0, 1\}$ and the optimal forecast cannot equal either one of them!

Faced with this dilemma, statisticians have devised an alternative approach in which a threshold θ is chosen such that the forecast value $F_{t+1} = 1$ when $p(\mathbf{X}_t) \geq \theta$ and $F_{t+1} = 0$ when $p(\mathbf{X}_t) < \theta$. This kind of forecast at least "feels right" in being expressed in terms of the two support points $\{0, 1\}$, but how should we go about choosing the level of θ? What's the best way to proceed?

One method that economists should find appealing is to choose θ so as to minimize the expected costs of forecast error. The situation $Y_{t+1} = 1, F_{t+1} = 0$ gives a forecast error of $Y_{t+1} - F_{t+1} = 1$, whereas $Y_{t+1} = 0, F_{t+1} = 1$ gives an error of -1. Each type of error presumably comes with a cost; and the probability of making each error depends on the value of the θ threshold. We would seem to have the ingredients we need for a cost-minimization exercise.

To begin, express the marginal probability of $Y_{t+1} = 1$ as

$$\Pr(Y_{t+1} = 1) = \int p(\mathbf{x})h(\mathbf{x})d\mathbf{x},$$

in which $h(\mathbf{x})$ is the joint density of the time-t explanatory variables and $p(\mathbf{x})$ is the conditional probability of $Y_{t+1} = 1$ given $\mathbf{X}_t = \mathbf{x}$. Likewise, the marginal probability of $Y_{t+1} = 0$ is

$$\Pr(Y_{t+1} = 0) = \int (1 - p(\mathbf{x}))h(\mathbf{x})d\mathbf{x}.$$

To evaluate forecasts, we need to derive the four *joint* probabilities of Y_{t+1} and the forecast F_{t+1}. Since $p(\mathbf{x}) \geq \theta \Rightarrow F_{t+1} = 1$, the joint probabilities for the two zero-forecast-error cases are as follows:

$$\Pr(Y_{t+1} = 1, F_{t+1} = 1) = \int p(\mathbf{x}) \cdot I(p(\mathbf{x}) \geq \theta) \cdot h(\mathbf{x})d\mathbf{x} = \int_{\mathbf{x}:p(\mathbf{x})\geq\theta} p(\mathbf{x})h(\mathbf{x})d\mathbf{x},$$

in which $I(p(\mathbf{x}) \geq \theta)$ is an indicator function taking the value 1 when its argument $p(\mathbf{x}) \geq \theta$ and equalling 0 otherwise. Similarly,

$$\Pr(Y_{t+1} = 0, F_{t+1} = 0) = \int_{\mathbf{x}:p(\mathbf{x})<\theta} (1 - p(\mathbf{x}))h(\mathbf{x})d\mathbf{x}.$$

Considering the two cases of non-zero forecast error, the probability of an error of 1 is given by

$$\Pr(Y_{t+1} = 1, F_{t+1} = 0) = \int_{\mathbf{x}:p(\mathbf{x})<\theta} p(\mathbf{x})h(\mathbf{x})d\mathbf{x}.$$

Note that as θ goes up, the set of $\mathbf{x}$ values over which integration takes place expands (bringing more $h(\mathbf{x})$ into the integral) and the allowable values of $p(\mathbf{x})$ increase. The probability of a forecast error of 1 therefore *rises* as θ rises. In effect, increasing the θ threshold raises the bar on the event $F_{t+1} = 1$, making $F_{t+1} = 0$ more likely, while leaving completely unchanged the likelihood that $Y_{t+1} = 1$.

Analyzing the probability of a forecast error of -1, which is

$$\Pr(Y_{t+1} = 0, F_{t+1} = 1) = \int_{\mathbf{x}:p(\mathbf{x})\geq\theta} (1 - p(\mathbf{x}))h(\mathbf{x})d\mathbf{x},$$

we see that the allowable $1 - p(\mathbf{x})$ values decrease as θ goes up while the set of $\mathbf{x}$ over which integration occurs shrinks. This forecast error probability therefore *declines* as θ rises.

The behavior of the two error probabilities, which move in opposite directions as the θ threshold is varied, suggests that we might choose θ to minimize the expected costs of these errors, using

$$\min_{\theta} c_1 \cdot \Pr(Y_{t+1} = 1, F_{t+1} = 0 \mid \theta) + c_{-1} \cdot \Pr(Y_{t+1} = 0, F_{t+1} = 1 \mid \theta)$$

in which c_1 and c_{-1} are the costs associated with forecast errors of 1 and -1 respectively. This cost-minimization exercise could be implemented with knowledge of the conditional and marginal distributions.[2]

[2]Note that the exercise we've just undertaken could also be carried out on a conditional basis given $\mathbf{X}_t = \mathbf{x}$. In this approach, for a given θ threshold, $p(\mathbf{x}) \geq \theta$ implies that the forecast $F_{t+1} = 1$. Then the forecast error probability $\Pr(Y_{t+1} = 0, F_{t+1} = 1 \mid \theta) = 1 - p(\mathbf{x})$, which is a quantity that falls as θ rises. As for the other forecast error probability, we have $\Pr(Y_{t+1} = 1, F_{t+1} = 0 \mid \theta) = p(\mathbf{x})$, which rises as θ rises. So, just as in the general case considered above, the forecast error probabilities move in different directions. We might therefore proceed to consider the costs of the two kinds of forecast error conditional on $\mathbf{X}_t = \mathbf{x}$, and choose the cost-minimizing value of θ to minimize expected costs. The value-added by the conditional approach is that the optimal θ will generally be specific to $\mathbf{x}$, rather than being a constant.

A now-vast literature in what is termed *machine learning* has explored these and related ideas using somewhat different terminology. The problem of choosing θ is termed a *classification* rather than a forecasting problem, with much of the interest centering less on θ itself than on comparing the performance of alternative estimators of $p(\mathbf{x})$.

One important tool in this branch of the literature is the *Receiver Operating Characteristic* (ROC) curve, whose name comes from the world of radar detection methods. The curve is a way of visualizing the connections between the *true positive rate*—the conditional probability of a correct forecast that $F_{t+1} = 1$ given that $Y_{t+1} = 1$ occurs—and the *false positive rate*—the conditional probability of an incorrect forecast $F_{t+1} = 1$ given that $Y_{t+1} = 0$. Using our notation from above, these conditional probability concepts are

$$\text{TPR}(\theta) = \frac{\Pr(Y_{t+1} = 1, F_{t+1} = 1 \mid \theta)}{\Pr(Y_{t+1} = 1)}$$

and

$$\text{FPR}(\theta) = \frac{\Pr(Y_{t+1} = 0, F_{t+1} = 1 \mid \theta)}{\Pr(Y_{t+1} = 0)}$$

In an ROC curve, the TPR is shown on the vertical axis of a graph and the FPR on the horizontal axis. The curve linking the two is generated by varying the θ threshold.

Chapter 6

Supplement to Mittelhammer's Chapter 4

This is an excellent compendium of results on useful distributions. For our purposes, the most important discrete distributions are: the Bernoulli (Section 4.1.2), the multinomial (Section 4.1.4), and the geometric (Section 4.1.5). In other courses, you may also make use of the Poisson distribution and Poisson processes, which Mittelhammer describes in Section 4.1.7. As for continuous distributions, the most important for our purposes are: the uniform (Section 4.2.1), exponential (Section 4.2.3), and especially the chi-square distribution (discussed at appropriate length in Section 4.2.4) and of course the normal distribution (Section 4.3 in full).

We will soon meet up with two additional distributions that can be viewed as combinations of normal and chi-square distributions: the t and $\mathcal{F}$ distributions, which are explained in Mittelhammer (2013, Section 6.7).

Chapter 7

Sample Means and Variances

In this chapter we consider some of the statistical properties of sample means and variances. The discussion will employ notation that at first glance may appear to be needlessly complex. Once you master this notation, however, it becomes very easy to understand the statistical properties of estimators based on linear regressions.

Before we plunge into the details, we need to consider the larger context in which estimation takes place. From this point forward in our course, we will be giving sustained attention to *estimators*, which are expressed explicitly in terms of formulas in simple cases, and implicitly-defined formulas (such as we see in first-order conditions) in more complicated cases. An estimator is composed in the hope that it will prove informative about the value of an otherwise unknown *parameter* (or parameters) of the data-generating process, as we'll discuss in a moment. When it comes time to compute the value taken on by estimator, you will do so using the *realized values* of random variables that sit in a dataset on your computer. Apart from discussions of the relative ease or difficulty of this computation, there is not a great deal that we can say about the specific number (or vector) that emerges from it given the realized values in the data you happen to have in hand. *To evaluate estimators—that is, to assess whether and to what extent they are actually informative—we must study their behavior as a function of the random variables that are involved*.

Viewed properly, then, an estimator is not merely a formula; it is a function of random variables, and thus a random variable (or vector) in its own right. We will be asking whether the features of the probability distribution of an estimator give us information about the likely location of the parameter(s) the estimator is designed to estimate. To this end, we focus (at least initially, in simple textbook cases) on the expected value of an estimator, its variance, and other features of its distribution.

7.1 The Data-Generating Process

Suppose that we have a collection of n observed random variables $\{Y_i, i = 1, \ldots n\}$, which we term a *dataset*. To frame the estimation task, we need to specify the process by which these data were generated. Our specification can be full and complete, or it can be incomplete, but in either case we must make some assumptions here at the outset. Usually we begin with assumptions that apply to a given Y_i and then consider how the various $\{Y_i\}$ in the

dataset are associated. As we describe these random variables, we introduce one or more unknown *parameters* whose values are of interest to us. We hope to be able to estimate these parameters.

In the case developed in this chapter, Y_i is expressed in terms of a *simple linear model*, $Y_i = \mu + \epsilon_i$, with $\mathrm{E}\,\epsilon_i = 0\ \forall i$. In this model, μ is a scalar parameter whose value is unknown. The random variable ϵ_i, which is the difference between Y_i and the unknown scalar μ, is not itself listed in our dataset—and so, it too is an unknown. In econometrics we call ϵ_i an unobserved *disturbance term*. Obviously, given that $\mathrm{E}\,\epsilon_i = 0$, we have $\mathrm{E}\,Y_i = \mu$. We also want to specify the variance of Y_i, and do this by assuming that for all cases $i = 1, \ldots, n$, the variance of the disturbance term is the same: $\mathrm{Var}\,\epsilon_i = \sigma^2$. Since Y_i is the sum of the constant μ and a disturbance ϵ_i, its variance is also σ^2. The assumption of constant variances is termed *homoskedasticity*; at the end of the chapter we will investigate what happens if the variance of Y_i is allowed to depend on i (*heteroskedasticity*). Thus far, we have described a data-generating process (DGP for short) with two parameters, μ and σ^2, which we aim to estimate.

Now we need to address the question of how the n observed random variables $\{Y_i\}$, or equivalently, the n unobserved random variables $\{\epsilon_i\}$, are associated among themselves. Are they mutually independent? Simply uncorrelated? Are they identically distributed, with all moments higher than the variance being identical?

One assumption that we might make—although it is very strong and often more than we need—is that the $\{\epsilon_i\}$ are *independently and identically distributed*. In this case, each of the ϵ_i can be viewed as a draw from the same univariate density function f, and we envision n such draws being made independently to generate our dataset. Because the $\{Y_i\}$ are produced by adding a constant μ to the disturbances, the iid property also applies to them. Under the iid assumption the collection of random variables $\{Y_i\}$ is said to be a *simple random sample*. It is not often that one encounters such a sample in the real world, but it is a useful textbook case. For much of this chapter, however, the iid assumption is far stronger than what we really need, and we can make do with the weaker assumption that the disturbances $\{\epsilon_i\}$, and thus the $\{Y_i\}$, are *uncorrelated*.

Sometimes we will need to round out our description of the DGP with a further *distributional assumption*, that $Y_i \sim \mathcal{N}(\mu, \sigma^2)$. We will invoke this normality assumption only when we cannot easily obtain a result without it.[1] Until we make such distributional assumptions, our description of the data-generating process remains incomplete. This is not in itself a bad thing: Why make additional assumptions if the results we seek can be derived without them?

Having specified the main features of the data-generating process, and having said that we hope to estimate the parameters μ and σ^2, we should now clarify the notion of an *estimator*. As indicated above, when viewed as a formula, an estimator is simply a function of the observed data. When viewed as a function of random variables, however, the probability distribution of an estimator provides information about the likely location of one or more parameters.

Hence, we will be considering functions of the general form $\hat{\mu} = g(\mathbf{Y})$ and $\hat{\sigma}^2 = h(\mathbf{Y})$.

[1] A side-effect of invoking normality is that the assumption that the $\{Y_i\}$ are uncorrelated implies that they are fully independent. As you will recall, this property of the multivariate normal distribution was proved in Mittelhammer (2013, Chapter 4).

Because $\hat{\mu}$ and $\hat{\sigma}^2$ are functions of random variables, they are themselves random variables. We will evaluate their performance as estimators of the constants μ and σ^2 using two criteria. First, we consider their expectations, $\mathrm{E}\,\hat{\mu}$ and $\mathrm{E}\,\hat{\sigma}^2$. An estimator is termed *unbiased* if its expected value equals the true value of the parameter. In the case of $\hat{\mu}$, unbiasedness means that $\mathrm{E}\,\hat{\mu} = \mu$. Our first order of business is to discover functions (*estimators*) that have this desirable property. Focusing on all such estimators, we next examine their variances. We generally prefer estimators with smaller variances, that is, those having greater *efficiency*. Why? If there are two candidate estimators $\hat{\mu}$ and $\tilde{\mu}$, both of which are unbiased, the estimator with smaller variance will have a probability distribution that is more concentrated around the true value of the μ parameter. You can establish this fact for yourself using Markov's inequality.

7.2 The Sample Mean

For the i-th observation in the dataset we have $Y_i = \mu + \epsilon_i$, and the data vector $\mathbf{Y}$ can be written as $\mathbf{Y} = \boldsymbol{\iota} \cdot \mu + \boldsymbol{\epsilon}$, where $\boldsymbol{\iota}$ is an n-vector of ones. Using this notation, we explore the properties of the sample average $\hat{\mu}$ as an estimator of the μ parameter. The sample average is simply

$$\begin{aligned} \hat{\mu} &= \frac{1}{n} \sum_{i=1}^{n} Y_i \\ &= (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \boldsymbol{\iota}' \mathbf{Y} \\ &= (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \boldsymbol{\iota}' (\boldsymbol{\iota} \cdot \mu + \boldsymbol{\epsilon}) \\ &= \mu + (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \boldsymbol{\iota}' \boldsymbol{\epsilon}. \end{aligned} \tag{7.1}$$

Because $\mathrm{E}\,\boldsymbol{\epsilon} = \mathbf{0}$, we see that $\mathrm{E}\,\hat{\mu} = \mu$, or, to put the result in different words, the sample mean is an *unbiased* estimator of the true mean.

But the sample mean $\hat{\mu}$ is itself a random variable—it is a linear function of the n random variables $\mathbf{Y}$—and its variance also needs consideration. Recall that the elements of $\mathbf{Y}$ have been assumed to be uncorrelated in our specification of the DGP. Hence,

$$\mathrm{Var}(\hat{\mu}) = \mathrm{Var}\left(\frac{1}{n} \cdot \sum_{i=1}^{n} Y_i \right) = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \mathrm{Var}\, Y_i = \frac{1}{n^2} \cdot \sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}.$$

To show the same thing in matrix notation, we write

$$\begin{aligned} \mathrm{E}(\hat{\mu} - \mu)(\hat{\mu} - \mu)' &= \mathrm{E}\,(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \boldsymbol{\iota}' \boldsymbol{\epsilon}\boldsymbol{\epsilon}' \boldsymbol{\iota} (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \\ &= (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \boldsymbol{\iota}'\, \sigma^2 \mathbf{I}\, \boldsymbol{\iota} (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \\ &= \sigma^2 (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \boldsymbol{\iota}'\boldsymbol{\iota} (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} \\ &= \sigma^2 (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}. \end{aligned} \tag{7.2}$$

As we will see later, this matrix–vector notation, although admittedly cumbersome in the present case, is easily extended to linear regression models with multiple explanatory variables.

We have found the mean and variance of $\hat{\mu}$ without too much difficulty. If we now invoke the normality assumption, then we can say that

$$\hat{\mu} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right),$$

because linear combinations of normally-distributed random variables are themselves normally distributed, and $\hat{\mu}$ is one such linear combination (Mittelhammer 2013, Chapter 4).

7.3 The Sample Variance

Let us examine the properties of the sample variance s^2, which is defined as

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \hat{\mu})^2$$

This expression needs some explanation—in particular, why is it that $n-1$ appears in the denominator?

Write the summation using vector notation,

$$\sum_{i=1}^{n}(Y_i - \hat{\mu})^2 = (\mathbf{Y} - \boldsymbol{\iota}\cdot\hat{\mu})'(\mathbf{Y} - \boldsymbol{\iota}\cdot\hat{\mu}).$$

Letting $\mathbf{Y} = \boldsymbol{\iota}\cdot\mu + \boldsymbol{\epsilon}$ and recalling that $\hat{\mu} = \mu + (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{\epsilon}$, we note that

$$\mathbf{e} = \mathbf{Y} - \boldsymbol{\iota}\cdot\hat{\mu} = \left(\mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\right)\boldsymbol{\epsilon} = \mathbf{M}\boldsymbol{\epsilon}. \tag{7.3}$$

Later, when we talk about multivariate linear models, we will be describing $\mathbf{e}$ as a vector of "residuals," these being the difference between $\mathbf{Y}$ and its fitted value $\boldsymbol{\iota}\cdot\hat{\mu}$. For now, however, we set that point aside and simply note that the $n \times n$ matrix $\mathbf{M}$ is symmetric and idempotent. Its rank therefore equals its trace, and the trace can be found from

$$\begin{aligned}\operatorname{trace}\mathbf{M} &= \operatorname{trace}(\mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}') \\ &= \operatorname{trace}\mathbf{I} - \operatorname{trace}(\boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}') \\ &= n - 1,\end{aligned}$$

again using the result that $\operatorname{trace}\mathbf{AB} = \operatorname{trace}\mathbf{BA}$ when the matrices are conformable.

With these results in hand, let us reconsider

$$(\mathbf{Y} - \boldsymbol{\iota}\cdot\hat{\mu})'(\mathbf{Y} - \boldsymbol{\iota}\cdot\hat{\mu}) = \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}$$

in which we have used equation (7.3) and the idempotence and symmetry of $\mathbf{M}$ have helped us to simplify things.

We want to find the expected value of $\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}$ and since this is a scalar random variable, nothing is changed if we first take its trace. The properties of traces allow us to say that

$$\operatorname{E}\operatorname{trace}(\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}) = \operatorname{E}\operatorname{trace}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{M}),$$

and, switching the order of operations,

$$\mathrm{E}\,\mathrm{trace}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{M}) = \mathrm{trace}(\mathrm{E}\,\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{M}) = \sigma^2\,\mathrm{trace}\,\mathbf{M} = \sigma^2 \cdot (n-1).$$

In this way we find that the expected value of s^2 is σ^2, the true variance. Like the sample mean, the sample variance is an unbiased estimator.

I know of no easy way to derive the variance of s^2 without the normality assumption.[2] But if we are willing to assume that the disturbance terms are distributed as multivariate normal, the derivation is simple enough. Dividing each of the $\boldsymbol{\epsilon}$ vectors by the scalar σ, and multiplying by σ^2, we obtain

$$\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon} = \sigma^2 \cdot \mathbf{u}'\mathbf{M}\mathbf{u}.$$

The second factor on the right is a quadratic form in $\mathbf{u}$, a multivariate standard normal random vector. This quadratic form is distributed as central chi-square with degrees of freedom equal to the rank of $\mathbf{M}$, which we have already found to be $n-1$. Recalling that the variance of a central chi-square variable is twice its degrees of freedom, we obtain

$$\mathrm{Var}(s^2) = \frac{\sigma^4}{(n-1)^2} \cdot \mathrm{Var}\left(\mathbf{u}'\mathbf{M}\mathbf{u}\right) = \frac{\sigma^4}{(n-1)^2} \cdot 2(n-1) = \frac{2\sigma^4}{(n-1)}.$$

7.4 Confidence Intervals

Since the sample mean $\hat{\mu}$ is a random variable, and the population parameter μ is unknown, we cannot say with certainty how far any particular realization of $\hat{\mu}$ is from the unknown μ. However, we can make probabilistic statements about the range in which the unknown μ must fall, by constructing what are termed *confidence intervals*.

This approach is grounded on Chebyshev's elaboration of Markov's inequality. For any random variable Z possessing a mean and a variance, we can write

$$\Pr\left(|Z - \mathrm{E}\,Z| > k \cdot \sigma_Z\right) \leq \frac{1}{k^2}.$$

Hence, for the $k = 2$ case, the probability that Z lies beyond two standard deviations of its mean is less than or equal to 0.25. To put this differently, the probability that it lies *within* 2 standard deviations of its mean is at least 0.75. That is, $\Pr(\mu - 2 \cdot \sigma_z \leq Z \leq \mu + 2 \cdot \sigma_z) > 0.75$.

If we apply this Chebyshev result to the sample mean $\hat{\mu}$, the equivalent expression is

$$\Pr\left(\mu - 2 \cdot \frac{\sigma}{\sqrt{n}} \leq \hat{\mu} \leq \mu + 2 \cdot \frac{\sigma}{\sqrt{n}}\right) > .75$$

Note that the event $\mu - 2 \cdot \frac{\sigma}{\sqrt{n}} \leq \hat{\mu} \leq \mu + 2 \cdot \frac{\sigma}{\sqrt{n}}$ is the same arithmetically as the event $\hat{\mu} - 2 \cdot \frac{\sigma}{\sqrt{n}} \leq \mu \leq \hat{\mu} + 2 \cdot \frac{\sigma}{\sqrt{n}}$ as you can easily check by subtracting $\mu + \hat{\mu}$ and then multiplying through by -1. So, if we wish, we can center the interval on $\hat{\mu}$, writing it in the form

[2] Mittelhammer (2013, p. 317) gives a link to the derivation, which he describes as involving much tedious but straightforward algebra.

$\hat{\mu} \pm 2 \cdot \frac{\sigma}{\sqrt{n}}$, and be assured that with probability greater than 0.75 the true mean μ will be found in that interval. This is the conventional definition of a "confidence interval".

Two final points: First, if we assume a distribution (say, normal) for the sample mean, then the probability statement can be made tighter. (Under the normal distribution assumption, for example, the 2-standard-deviation probability is approximately 0.954, far above the 0.75 given by the Chebyschev inequality.) Second, note that the confidence interval depends on σ^2, an unknown parameter, and the simple theory laid out above does not take this fact into account. The theory can only be viewed as approximately correct if the estimator s^2 is used in place of the σ^2 parameter.

Developing a proper theory for confidence intervals would seem to require inspection of the joint distribution of $\hat{\mu}$ and s^2. However, if you are willing to make a normal distribution assumption, then hypothesis-testing via the t-distribution—which, as we will see in our next chapter, effectively dispenses with the problem of the unknown σ^2—does essentially what you would hope to do with confidence intervals, and is probably the more sensible way to go.

7.5 Specification Errors

Students who are seeing econometrics for the first time often view the representation of the data-generating process

$$Y_i = \mu + \epsilon_i$$

as a simple equation, and when asked for proofs of one kind or another, sometimes give in to the temptation to blithely move things from the right-hand to the left-hand side of the equation as they have been doing for years in math courses. This is not always the wrong approach, but it can lead you astray. In referring to the "data-generating process" we mean that the right-hand side *produces* or generates the left-hand Y_i, and ideally we would use notation such as $Y_i := \mu + \epsilon_i$ or $Y_i \leftarrow \mu + \epsilon_i$ to emphasize this point. But that kind of notation has not been used very often in econometrics. Therefore, to keep consistent with the literature, we too will represent the data-generating process in terms of equations.

Even for the simple linear models considered in this chapter, there is ample room for the researcher to make mistakes in specifying the data-generating process. One common mistake is to ignore heterogeneity in Y_i by which $\mathrm{E}\, Y_i = \mu_i$. The consequences for the sample mean are easy to see:

$$\mathrm{E}\, \hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mu_i = \bar{\mu}_n$$

but with the other assumptions of the data-generating process left as they are, we still have

$$\mathrm{Var}\, \hat{\mu} = \frac{\sigma^2}{n}.$$

In this instance, the sample mean is *not unbiased* because its expectation does not equal a parameter. Instead, its expectation coincides with the average of n different parameters. When we generalize our approach to consider multiple right-hand side "explanatory variables", with attached coefficients estimated by the method of ordinary least squares regression, we will spend a good deal of time on the consequences of this kind of specification error.

For the case at hand, specification mistakes regarding the variances and covariances of the disturbance terms are more interesting to analyze. Let's consider the implications of assuming homoskedasticity (all observations have the same variance) when in fact the data are heteroskedastic (different variances). In this case, with $Y_i = \mu + \epsilon_i$, we have $\mathrm{E}\, Y_i = \mu$ but $\mathrm{Var}\, Y_i = \mathrm{Var}\, \epsilon_i = \sigma_{ii}$. For the moment, we'll continue to assume that the $\{Y_i\}$ sequence is uncorrelated.

What are the consequences of this misspecification for the properties of our estimators $\hat{\mu}$ and s^2? Given $\hat{\mu} = n^{-1} \sum_i Y_i$, we still have $\mathrm{E}\, \hat{\mu} = \mu$, but

$$\mathrm{Var}\, \hat{\mu} = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}\, Y_i = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_{ii} = \frac{1}{n} \bar{\sigma}^2,$$

with $\bar{\sigma}^2$ being the average variance. This doesn't look too bad; at least the expression is loosely analogous to what we obtained with homoskedasticity.

But does the expected value of s^2 equalf $\bar{\sigma}^2$? Consider

$$s^2 = \frac{1}{n-1} \boldsymbol{\epsilon}' \mathbf{M} \boldsymbol{\epsilon} = \frac{1}{n-1} \left(\boldsymbol{\epsilon}' \boldsymbol{\epsilon} - \boldsymbol{\epsilon}' \boldsymbol{\iota} (\boldsymbol{\iota}' \boldsymbol{\iota})^{-1} \boldsymbol{\iota}' \boldsymbol{\epsilon} \right).$$

Now, $\mathrm{E}\, \boldsymbol{\epsilon}' \boldsymbol{\epsilon} = \mathrm{E} \sum_{i=1}^{n} \epsilon_i^2 = n \bar{\sigma}^2$. To deal with the other term, consider

$$\mathrm{E}\, \boldsymbol{\epsilon}' \boldsymbol{\iota} (\boldsymbol{\iota}' \boldsymbol{\iota})^{-1} \boldsymbol{\iota}' \boldsymbol{\epsilon} = \mathrm{E}\, \boldsymbol{\iota}' \boldsymbol{\epsilon} \boldsymbol{\epsilon}' \boldsymbol{\iota} (\boldsymbol{\iota}' \boldsymbol{\iota})^{-1},$$

and, taking the expectation of the $n \times n$ matrix $\boldsymbol{\epsilon} \boldsymbol{\epsilon}'$ yields

$$\mathrm{E}\, \boldsymbol{\iota}' \boldsymbol{\epsilon} \boldsymbol{\epsilon}' \boldsymbol{\iota} (\boldsymbol{\iota}' \boldsymbol{\iota})^{-1} = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \begin{bmatrix} \sigma_{11} & & & 0 \\ & \sigma_{22} & & \\ & & \ddots & \\ 0 & & & \sigma_{nn} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \cdot n^{-1} = \bar{\sigma}^2.$$

Hence,

$$\mathrm{E}\, s^2 = \frac{1}{n-1} \left(n \bar{\sigma}^2 - \bar{\sigma}^2 \right) = \bar{\sigma}^2.$$

We would not say that s^2 is "unbiased," but at least its expectation coincides with a potentially interesting average of parameters.

A more sophisticated method of proof, which makes better use of the properties of $\mathbf{M}$, proceeds as follows. Note that when the $n \times n$ matrix $\mathbf{M}$ is post-multiplied by any n-vector $\mathbf{z}$, it transforms the original $\mathbf{z}$ vector into deviations from the mean of $\mathbf{z}$'s elements; that is,

$$\mathbf{M}\, \mathbf{z} = \tilde{\mathbf{z}} = \mathbf{z} - \boldsymbol{\iota} \cdot \frac{1}{n} \sum_{i=1}^{n} z_i.$$

Now consider $\mathrm{trace}(\epsilon' \mathbf{M} \epsilon) = \mathrm{trace}(\mathbf{M} \epsilon \epsilon')$ and then take expectations, yielding $\mathrm{trace}(\mathbf{M} \mathbf{V})$ in which $\mathbf{V}$ is diagonal. Since post-multiplying $\mathbf{M}$ by any $n \times 1$ vector produces the difference between that vector and the average of its elements, we can think of $\mathbf{M}$ acting upon each of the columns of $\mathbf{V}$ in this way.

Let $\mathbf{v}_i$ be the i-th column of $\mathbf{V}$; its only non-zero element is σ_{ii}, which will be replaced in the $\mathbf{M}$-transformed version $\tilde{\mathbf{v}}_i$ by $\sigma_{ii} - \frac{\sigma_{ii}}{n} = \frac{1}{n}\sigma_{ii}(n-1)$. The other transformed elements of $\tilde{\mathbf{v}}_i$ will lie off the diagonal of $\tilde{\mathbf{V}} = \mathbf{MV}$ and so will not figure into the trace. The trace of $\mathbf{MV}$ is therefore $\bar{\sigma}^2(n-1)$, which implies $\mathrm{E}\, s^2 = \bar{\sigma}^2$ as we found in the alternative proof above.

What about cross-observation correlation in the disturbances? To study the implications, again let $\mathrm{E}\, \epsilon\epsilon' = \mathbf{V}$ but now allow $\mathbf{V}$ to have non-zero terms off the diagonal. Then when we come to analyze trace$(\mathbf{MV})$, we see that each column of $\mathbf{V}$ is transformed into deviations-from-means form as before, but the result is more complicated than in the merely heteroskedastic case:

$$\mathbf{M}\mathbf{v}_i = \mathbf{v}_i - \boldsymbol{\iota} \cdot \frac{1}{n} \sum_{j=1}^{n} \sigma_{ji}$$

and the term on the diagonal (on which the trace focuses) will be $\sigma_{ii} - \frac{1}{n}\sum_{j=1}^{n} \sigma_{ji}$. Summing up these diagonal terms using a trace yields a complicated expression that has no clear substantive interpretation. In other words, the expected value of s^2 is of no obvious use to us in this case—-it corresponds to a peculiar combination of averages of variance and covariance parameters.

Chapter 8

Testing Hypotheses About the Mean and Variance

8.1 The Classical Approach

The testing procedures we'll describe below involve several steps.

- First, the researcher specifies the *data-generating process* or DGP, making explicit the model by which the data are assumed to have been generated. As we discussed earlier, the DGP can be fully specified, as in the case of $Y_i = \mu + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and the disturbances in the sequence $\{\epsilon_i, i = 1, \ldots, n\}$ assumed mutually independent. A similar but incompletely-specified DGP would be $Y_i = \mu + \epsilon_i$ with $\mathrm{E}\,\epsilon_i = 0$, $\mathrm{E}\,\epsilon_i^2 = \sigma^2$, and $\{\epsilon_i, i = 1, \ldots, n\}$ assumed to be uncorrelated. In the latter specification, no distributional assumption is made and nothing is said about the higher moments of ϵ_i. In either case, a key element in the specification of the DGP is a listing of the model's parameters (here, μ and σ^2).

- Next, the researcher lists the *maintained hypotheses*, the features of the DGP that will not be subjected to testing but rather taken to be correct for the duration of the test procedure. For instance, normality might be a maintained hypothesis in the first of the models above.

- The *null hypothesis* is then specified, this being a proposition to be tested about one or more of the model's parameters. For the model above, a simple null hypothesis is

  $$H_0 : \mu = \mu_0,$$

  by which the true value of the mean μ is posited to equal the number μ_0. This is the hypothesis on which the test will focus. The researcher also indicates the *alternative hypothesis* under consideration, such as $H_A : \mu \neq \mu_0$ or $H_A : \mu > \mu_0$.

- A *test statistic* T is formulated to shed light on the truth of the null hypothesis. This statistic must have two properties: it must be wholly composed of quantities that can be calculated from the data and from parameter values given in the null hypothesis; and *its distribution (i.e., pdf or cdf) must be known when the null hypothesis is true*. We

choose a functional form for T so that the test statistic summarizes the evidence *against* the null hypothesis.

- We also choose the *size of the test*, which is the probability of mistakenly rejecting the null hypothesis when that hypothesis is correct. Usually we choose the size of the test to be small, so that the chance of mistakenly rejecting the null is on the order of 0.05, 0.01, or 0.001. A *rejection region* R is specified that corresponds to the test size, such that when $T \in R$ we say that we "reject" the null hypothesis. By construction, $\Pr(T \in R)$ equals the test size when the null hypothesis is correct.

- To evaluate the testing procedure, we must also calculate the *power function* of the test. The power of the test is (loosely speaking) the probability of rejecting the null hypothesis when the null hypothesis is false. If the null hypothesis is stated as $H_0 : \mu = \mu_0$, then to assess power we must calculate the probability of rejecting the null (i.e., $\Pr(T \in R)$) for each value of $\mu \neq \mu_0$ considered in the alternative hypothesis. Obviously, values of the true mean μ that are very close to the μ_0 value specified in the null will tend to yield rejection probabilities that are small, not much larger than the size of the test. Indeed, as $\mu \to \mu_0$, we expect the test's power to approach its size. By contrast, true μ values that are very far from what was specified in the null should yield high probabilities of rejecting the null.

  A useful test—one that illuminates the truth or falsity of the null hypothesis—must have acceptable power. It is not enough to know the distribution of T under the null hypothesis; we also need to verify that when $T \in R$ this event is interpretable as casting doubt on the null relative to the alternative.

8.2 The Role of the Chi-squared Distribution

As we will see later, the χ^2 distribution is at the heart of testing procedures in econometrics. We are not yet in a position to fully appreciate its role, but the following may be useful by way of illustration. Let us consider the model $Y_i = \mu + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ and $\{\epsilon_i\}$ mutually independent. As you'll recall, for this DGP the sample mean $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)$ and let the null hypothesis be specified as $H_0 : \mu = \mu_0$.

By the theorem given in Chapter 2, the quadratic form

$$\tilde{T} = (\hat{\mu} - \mu_0)' \left[\frac{\sigma^2}{n}\right]^{-1} (\hat{\mu} - \mu_0)$$

is distributed as central χ^2_1 when the null hypothesis is true. If we knew σ^2, then $\tilde{T}$ could serve as our test statistic. Examining $\tilde{T}$, we see that the greater is the difference between the sample mean $\hat{\mu}$ and the hypothesized true mean μ_0, the greater is the value of the quadratic form. If we wished to specify a test of size 0.01, then we should determine a critical value t_H such that $\tilde{T} > t_H$ with 1 percent probability when $\tilde{T} \sim \chi^2_1$. This would be easy to do in R or STATA, which have built-in functions for the cdf of central χ^2 distributions.[1] Of

[1] To determine the power function of this test, however, we would need to examine non-central χ^2_1 distributions (also built into R and STATA) as we will go on to discuss later in this chapter.

course we do not actually know σ^2, and so $\tilde{T}$ fails to meet one of the basic requirements for a test statistic. To test the null hypothesis $H_0 : \mu = \mu_0$, we must search for an alternative formulation that contains no such unknown parameters. It is for this reason that tests of hypotheses about the mean are a bit more difficult to develop in full than are tests focusing on the variance.

8.3 Testing Hypotheses About σ^2

To begin with the easier case, suppose that the null hypothesis addresses the value of the variance and is expressed as $H_0 : \sigma^2 = \sigma_0^2$. How would we test this proposition? We seek a test statistic whose value can provide evidence *against* the null hypothesis, and whose distribution is *known* under H_0, that is, known if H_0 is true.

Consider $s^2 = \sigma^2 \cdot \mathbf{u}'\mathbf{M}\mathbf{u}/(n-1)$, in which $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Now if H_0 is true, so that the true $\sigma^2 = \sigma_0^2$, then statistic

$$T = \frac{(n-1)s^2}{\sigma_0^2} = \mathbf{u}'\mathbf{M}\mathbf{u} \sim \chi^2_{n-1}.$$

If the true $\sigma^2 = \sigma_0^2$ as hypothesized, then the mean of this test statistic is $\mathrm{E}\, T = n - 1$.

Suppose, however, that the true $\sigma^2 \neq \sigma_0^2$. Then the denominator of the test statistic T is incorrect—instead of being a χ^2_{n-1} random variable, the test statistic T is *proportional* to such a variable. Let σ^2 denote the true value of the parameter and let σ_0^2 be the hypothesized value; then

$$T = \frac{\sigma^2}{\sigma_0^2}\left(\frac{(n-1)s^2}{\sigma^2}\right) = \frac{\sigma^2}{\sigma_0^2} \cdot \chi^2_{n-1}.$$

Thus if the true $\sigma^2 > \sigma_0^2$, $\mathrm{E}\, T > n - 1$ and if $\sigma^2 < \sigma_0^2$, $\mathrm{E}\, T < n - 1$. Evidently, small and large values of T relative to $n - 1$ cast doubt on the null hypothesis. Clearly we will need to specify a two-tailed rejection region $R = (-\infty, t_L] \cup [t_H, \infty)$ to take into account both directions of departure from the $n - 1$ benchmark.

The values of the cut-points t_L and t_H are chosen such that the size $\Pr(T \leq t_L) + \Pr(T \geq t_H) = 0.05$ or some probability even smaller, such as 0.01 or 0.001. The idea is that there should be only a low probability of mistakenly rejecting the null when the null is actually true; we want to guard against that sort of mistake. If we decide that the size of the test should be 0.05, then we should choose values for t_L and t_H to define the rejection region such that $\Pr(T \leq t_L) + \Pr(T \geq t_H) = 0.05$. It is conventional to choose these values to ensure that the events $T \leq t_L$ and $T \geq t_H$ each occur with 0.025 probability. For a sample of size $n = 50$, we refer to the distribution of the χ^2_{49} random variable and find that the values $t_L \approx 31.55$ and $t_H \approx 70.22$ yield an appropriate rejection region.

The power of the test

Continuing with the null hypothesis $\sigma^2 = \sigma_0^2$, we ask how likely we are to *reject* this hypothesis *when the null is not true*. Given

$$T = \frac{(n-1)s^2}{\sigma_0^2}$$

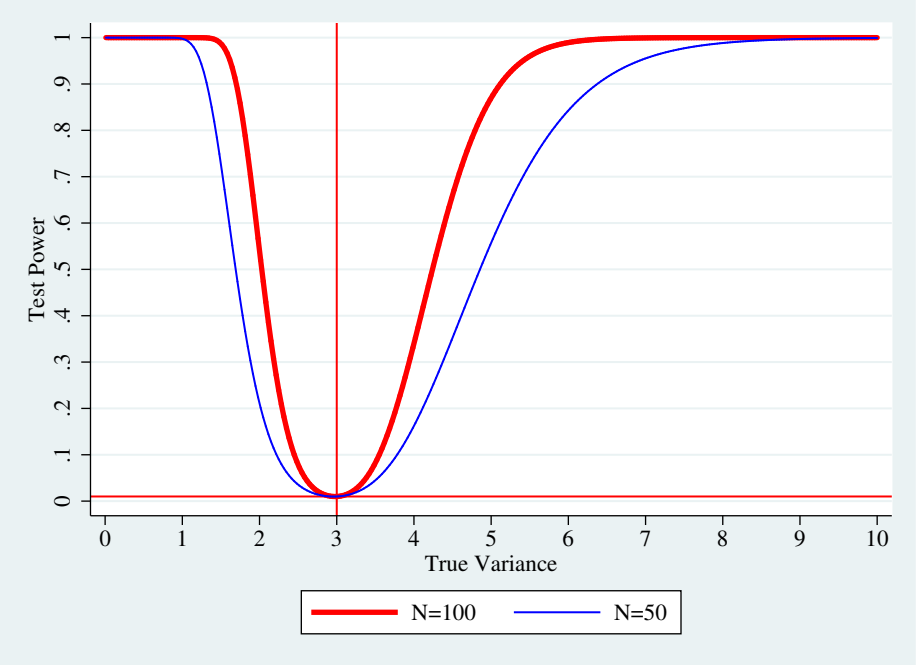


Figure 8.1: Power functions for the test of $H_0 : \sigma^2 = 3$ for sample sizes $n = 50$ and $n = 100$ and test size = 0.01.

we can multiply T as follows,

$$\frac{\sigma_0^2}{\sigma^2}\, T = \frac{\sigma_0^2}{\sigma^2}\frac{(n-1)s^2}{\sigma_0^2} = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$$

With this result, we can determine the power of the test for any given true value of σ^2.

To see how, study the following reasoning for the $n = 50$ case in which the low-end cut-point is $t_L = 31.55$. The event $T \leq 31.55$, which might be termed a "low-end" rejection, can be re-expressed as

$$\begin{aligned} T &\leq 31.55 \\ \frac{\sigma_0^2}{\sigma^2} T &\leq \frac{\sigma_0^2}{\sigma^2}\, 31.55 \\ \chi^2_{49} &\leq \frac{\sigma_0^2}{\sigma^2}\, 31.55 \end{aligned}$$

So for any given true value of σ^2, we can have R calculate the probability of this event using its built-in routine for the cdf of a central χ^2_k variable. Likewise, the event $T \geq 70.22$ (a high-end rejection) is equivalent to $\chi^2_{49} \geq (70.22)(\sigma_0^2/\sigma^2)$. For any given value of $\sigma^2 \neq \sigma_0^2$, we can easily find the probabilities attached to these two mutually exclusive events. The power function of the test evaluated at σ^2 is the sum of the two probabilities.

For instance, suppose that the true $\sigma^2 = 2$ but the null hypothesis specifies $\sigma_0^2 = 1$. Then in a test using $n = 50$ observations and a size of 0.05,

$$\Pr\left(\chi^2_{49} \leq (31.55)(1/2)\right) + \Pr\left(\chi^2_{49} \geq (70.22)(1/2)\right) \approx 1.6^{-6} + 0.9324 \approx 0.9324.$$

We're clearly very likely to reject the null when the true σ^2 departs to this extent from the value specified in the null hypothesis. Still, a rejection is not guaranteed; there is about a 7 percent chance of mistakenly failing to reject.

This procedure is easily automated to produce a graph of the power function for all values of $\sigma^2 \neq \sigma_0^2$. Power functions for sample sizes of $n = 50$ and $n = 100$ and a test size of 0.01 are shown in Figure 8.1. As can be seen, the probability of rejecting the null for any given $\sigma^2 \neq \sigma_0^2$ increases with the sample size. (Here, $\sigma_0^2 = 3$.) Note that these functions are slightly asymmetric.

An R illustration of the power function

You can get a deeper understanding of the power function by running the R program which follows:

```
rm(list=ls())

#-------------------------------------------------#
# Power function of a hypothesis test on the
#  variance in an iid normal sample of size nobs.
#-------------------------------------------------#

#-------------------------------------------------#
# ?dchisq will show you the 4 functions available
# for the chi-squared distribution. The "d" prefix
# stands for "density", the "p" prefix for the cdf,
# the "q" prefix for quantiles (essentially the
# inverse of p), and the function prefixed by "r"
# draws random numbers from the distribution.
#-------------------------------------------------#

# For a chi-squared test statistic, consider
# T = (n-1) s^2/sigma_0^2 distributed as
# central chi-squared with n-1 degrees of freedom
# if the null hypothesis H_0: sigma^2 = sigma_0^2
# is true.

nobs <- 100  # Number of observations
size <- 0.05 # Size of the test
sigma_sq_0 = 5.0 # Null hypothesis

# Rejection region:
t_L <- qchisq(size/2.0, df=nobs-1, lower.tail=TRUE)
t_H <- qchisq(size/2.0, df=nobs-1, lower.tail=FALSE)

# Deriving the power function of the test
Power <- function(v){
  p <- pchisq(t_L*sigma_sq_0/v, df=nobs-1, lower.tail=TRUE) +
       pchisq(t_H*sigma_sq_0/v, df=nobs-1, lower.tail=FALSE)
  return(p)
}

# Plotting the power function:
plot(function(v){Power(v)}, from=0.01, to=10.0,
```

```
    ylab="Power of the Test", xlab="True Variance")
```

8.4 Testing Hypotheses about the Mean

To develop a testing procedure for hypotheses about the mean μ, we need to introduce the t_k distribution, and a full treatment requires understanding of both the "central" and "non-central" versions of this distribution. As you may remember, a random variable is distributed as t_k if it can be expressed as the ratio of a $\mathcal{N}(0,1)$ variable to the square root of a χ^2_k variable divided by its degrees of freedom,

$$t_k = \frac{\mathcal{N}(0,1)}{\sqrt{\chi^2_k/k}},$$

and if the $\mathcal{N}(0,1)$ and χ^2_k variables are independent. This is the "central" t_k distribution. The non-central version arises from the ratio

$$t_k = \frac{\mathcal{N}(\delta,1)}{\sqrt{\chi^2_k/k}},$$

with $\delta \neq 0$ termed the "non-centrality parameter". When $\delta > 0$ the distribution of the non-central $t_{k,\delta}$ distribution is shifted to the right relative to the central version, and when $\delta < 0$ the distribution is shifted to the left. We will need the non-central version to calculate the power function.

Let's consider the sample mean and variance in relation to the t_k distribution. The sample mean $\hat{\mu} = \mu + (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{\epsilon}$, and $\hat{\mu} - \mu = \mathbf{B}\boldsymbol{\epsilon}$, with $\mathbf{B} \equiv (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'$, a row vector of dimension $1 \times n$. The variance of $\hat{\mu}$ is σ^2/n, so that

$$\frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} = \frac{1}{\sqrt{\sigma^2/n}}\mathbf{B}\boldsymbol{\epsilon} \sim \mathcal{N}(0,1).$$

The sample variance is

$$s^2 = \frac{\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}}{n-1} \Rightarrow \frac{(n-1)s^2}{\sigma^2} = \mathbf{u}'\mathbf{M}\mathbf{u} \sim \chi^2_{n-1}$$

and s^2/σ^2 is therefore a χ^2_{n-1} variable divided by its degrees of freedom.

We have written the sample mean and sample variance using $\mathbf{B}$ and $\mathbf{M}$ so that we can invoke the theorem of Chapter 2 concerning independence of linear and quadratic forms in standard normal random vectors. Using that theorem, we establish independence by showing that the $1 \times n$ matrix $\mathbf{BM} = \mathbf{0}$. The other elements of the expressions above are constants—we could redefine $\mathbf{B}$ and $\mathbf{M}$ to include these constants, but that isn't really necessary. The essence of the problem is to show that

$$\begin{aligned}\mathbf{BM} &= (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\left[\mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\right] \\ &= (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}' - (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}' = \mathbf{0}_{1\times n},\end{aligned}$$

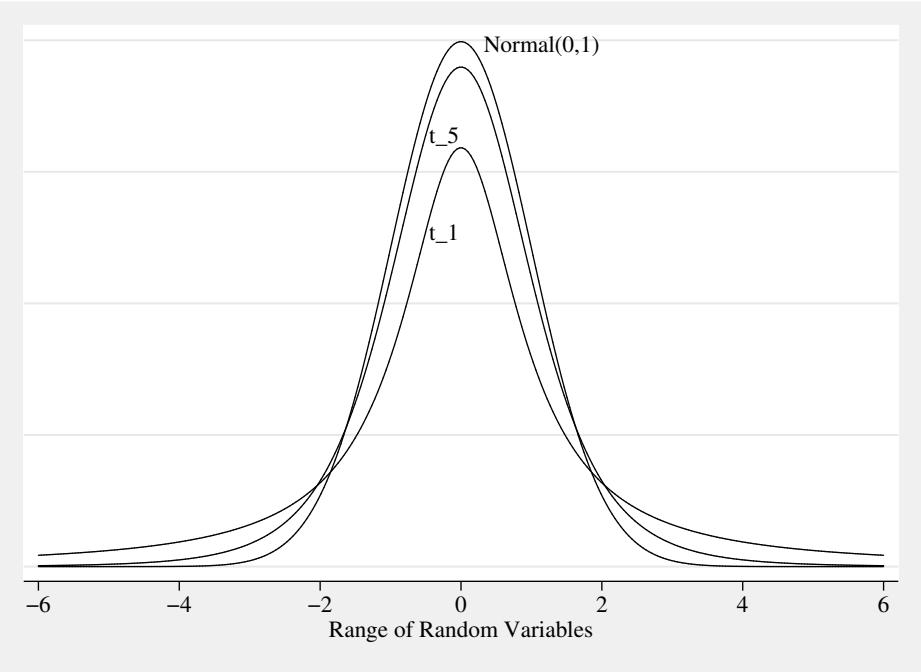


Figure 8.2: Densities of the t and standard normal distributions compared

and in this way independence is proved.

In summary, then, we find that

$$T = \frac{\dfrac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}}}{\sqrt{s^2/\sigma^2}} = \frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} \, \frac{1}{\sqrt{s^2/\sigma^2}} = \frac{\hat{\mu} - \mu}{\sqrt{s^2/n}},$$

from which the unknown σ^2 has cancelled out, has a t_{n-1} distribution. Note that the denominator is the estimated standard deviation of the sample mean. Also note that $n - 1$ is the number of degrees of freedom, but n appears in the formula. The density function of t_{n-1} is illustrated in the figure for degrees of freedom of 1 and 5.

With this as background, we can now proceed to develop our test of the null hypothesis $H_0 : \mu = \mu_0$ against the alternative hypothesis $H_A : \mu \neq \mu_0$. As the development above has suggested, the value of the sample mean $\hat{\mu}$ will provide the key element of the test statistic. Because $\mathrm{E}\,\hat{\mu} = \mu$, values of $\hat{\mu}$ that depart in either direction from μ_0 cast doubt on the null, and we would therefore want to choose values t_L and t_H such that the events $T \leq t_L$ and $T \geq t_H$ cause the null to be rejected. It is conventional to set t_L and t_H so that the probability of a low-end rejection equals the probability of a high-end rejection. Since the t distribution is symmetric about zero, symmetric rejection points t_L and t_H can be chosen that satisfy $\Pr(t_{n-1} \leq t_L) = \Pr(t_{n-1} \geq t_H)$. For a test of size 0.05, we would have $t_L \approx -2.01$ and $t_H \approx 2.01$ if the dataset contains $n = 50$ observations.

The power of the test

The analysis thus far has focused on the behavior of the test statistic T when the null hypothesis is true. To understand the power function of the test, we should now proceed to explore its behavior when the null is false. That task requires us to examine the rather exotic *noncentral t* distribution, which is available in R but not (at least as I recall) in STATA. The null hypothesis is that the true $\mu = \mu_0$. The alternative hypothesis is that the true μ is $\mu = \mu_0 + d$, for any d not equal to 0.

The numerator of the t-test is $(\hat{\mu} - \mu_0)/\sqrt{\sigma^2/n}$. Under the null this is distributed as $\mathcal{N}(0,1)$. Under the alternative, though, with $\mu_0 = \mu - d$, we have $\hat{\mu} - \mu_0 = \hat{\mu} - \mu + d$. Under the alternative, therefore,

$$\frac{\hat{\mu} - \mu_0}{\sqrt{\sigma^2/n}} \sim \mathcal{N}(\delta, 1)$$

with δ being defined as $d/\sqrt{\sigma^2/n}$. Think of any given δ as measuring the difference between the null and an alternative value, but with the difference expressed in units that are standard deviations of the sample mean. That is, $\delta = 1$ corresponds to $d = \sqrt{(\sigma^2/n)}$; and $\delta = 2$ is $2\sqrt{(\sigma^2/n)}$; and so on. When this $\mathcal{N}(\delta, 1)$ is set atop the square root of an independent χ^2_{n-k} variable divided by its degrees of freedom, the result is distributed as non-central $t_{n-k,\delta}$.

In R, the non-central t distribution is already coded for you and can be used to determine the power function of the t-test. The following program illustrates how to proceed.

```
rm(list=ls())
library(ggplot2)
#-------------------------------------------------------------------------
#  We'll consider a range of possible d values from -5.0 to 5.0. These
#  values will appear on the horizontal axis, and the probability of
#  rejecting the null (mu=0) will appear on the vertical axis.
#
#-------------------------------------------------------------------------

size <- 0.05 # Choose the test size

# If you want to change the sample size, do so here
# *before* defining t_L and t_H.
n <- 50

# Rejection region for a two-tailed test:
t_L <- qt(size/2.0, df=n-1, lower.tail=TRUE)
t_H <- qt(size/2.0, df=n-1, lower.tail=FALSE)

# Define the power function in terms of its dummy arguments.
# Note that we will pass in a *vector* of d values, but R will
# handle the vector properly:
Power <- function(d, nobs, sigma_sq, low, high){
 delta <- d/(sqrt(sigma_sq/nobs)) # convert from d to delta
 power <- pt(low,  df=nobs-1, ncp=delta, lower.tail=TRUE) +
          pt(high, df=nobs-1, ncp=delta, lower.tail=FALSE)
 return(power)
}
```

```
# Here is the vector of d values for the absolute difference between the
# true and the hypothesized values of the mu parameter: d = mu - mu_0:
d <- seq(from=-5.0, to=5.0, by=0.05)

# We pass *actual arguments* into the Power function by indicating that
# the *dummy argument* in the function should be given the value of the
# corresponding actual argument defined outside the function.
# Thus nobs (the dummy argument by which the sample size is known within
#  the function) is assigned the actual value n (defined above, outside the
#  function), the dummy argument low is assigned the actual argument value
#  t_L, and so on.
# The results are placed inside a data.frame so that they can be plotted
#  for three values of the variance parameter.

T_Power <- data.frame(D=d,
            Power1=Power(d=d, nobs=n, sigma_sq = 1.0, low=t_L, high=t_H),
            Power5=Power(d=d, nobs=n, sigma_sq = 5.0, low=t_L, high=t_H)
            Power10=Power(d=d, nobs=n, sigma_sq = 10.0, low=t_L, high=t_H),
            )

P <- ggplot(T_Power, aes(x=D,y=Power1)) + geom_line(color="steelblue") +
      ylab("Rejection Probability") + xlab("Value of d = mu - mu_0") +
      theme_bw()
print(P)
```

8.5 One-sided hypotheses

We often want to explore null and alternative hypotheses such as $H_0 : \mu \leq \mu_0$ versus $H_A : \mu > \mu_0$. The conceptual difficulty here has to do with the fact that we need to know how the test statistic is distributed when the null hypothesis is true, so that we can establish the rejection region. In the case at hand, however, for *any* value of μ that is less than or equal to μ_0 the null is true. There is an infinity of ways in which the null could be true. Where, then, do we situate the test statistic so that it best summarizes evidence against the null?

The convention is to situate the test statistic at (or with specific reference to) the boundary between the regions given in the null and alternative hypotheses, which in this instance would be exactly at μ_0. To understand the logic, consider Figure 8.3 which refers to $H_0 : \mu \leq 1$ as against $H_A : \mu > 1$. To focus on the essentials, let's simply take the sample mean $\hat{\mu}$ to be the test statistic on the (artificial) assumption that we know the value of σ^2. For any given true value of μ, we therefore know all the particulars of the distribution of this test statistic: It is distributed as $T = \hat{\mu} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

If we take the true $\mu = \mu_0 = 1$, which is at the boundary between H_0 and H_A, and choose the size of the test—say, 0.05—we obtain the distribution shown on the right. The test statistic is distributed as $\mathcal{N}(1, \frac{\sigma^2}{n})$, hence centered on the boundary point $\mu = \mu_0 = 1$, with this boundary point being represented in the thick red vertical line. The null is rejected for values of the test statistic greater than t_H. The area under the curve to the right of the high-end rejection point t_H, which is shaded in red, represents the size of the test. This is the conventional approach adopted for a one-sided null hypothesis. The rejection region begins well to the right of 1, so that decisive evidence against the null is required before we

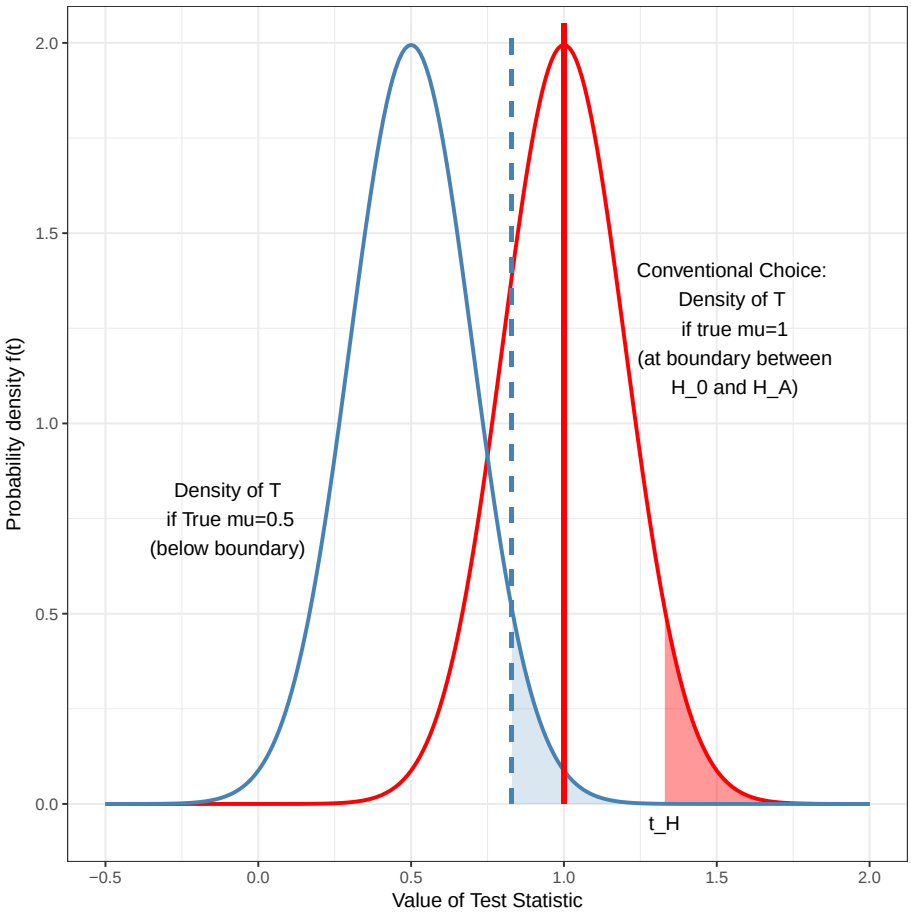


Figure 8.3: Examining a one-sided hypothesis

reject it.

Suppose, however, that we had chosen some other value on which to center the test statistic, say, $\mu = 0.5$, in which case the statistic is distributed as $T \sim \mathcal{N}(0.5, \frac{\sigma^2}{n})$. This value of the true μ is *also* consistent with the null hypothesis. The test statistic has the density shown by the curve to the left, whose rejection region and size (keeping the size of the test the same) is indicated in the blue-shaded region.

But note how this non-conventional choice would put us in a logical dilemma. All values of the test statistic to the right of the dashed vertical line would cause us to reject the null hypothesis; but many of these values in the rejection region would not have caused rejection in the conventional set-up. Indeed, values between the dashed vertical blue line and the thick red line are perfectly consistent with the null hypothesis $H_0 : \mu \leq 1$. They certainly cannot count as evidence *against* the null. Yet had we picked $\mu = 0.5$ as the point on which to center our test statistic, values of T in this range would have caused us to reject. And values of the test statistic between the thick red line at $\mu = 1$ and the beginning of the conventional rejection region at t_H, which would not have been taken as decisive evidence against the null in the conventional approach, would also cause rejection.

As a consequence of having mulled over this problem, we might now admit to some uneasiness about declaring the red-shaded area to be "the" size of the test in the conventional set-up for the testing procedure. If this shaded region has area 0.05, then any true $\mu < \mu_0$ would cause the test statistic $T = \hat{\mu}$ to fall in the rejection range less than 0.05 of the time. One example is already sketched out by the blue density function, for which this chance would be extremely low. In a sense, then, 0.05 is the *maximum* size of the test, that is, the greatest probability of mistakenly rejecting the null given that (a) the null is true, and (b) we have designed our testing procedure in the conventional way (as represented by the red density function). The value 0.05 thus represents a ceiling on the probability of making this mistake.

8.6 What's a "p-value"?

In the discussion so far, we have designed testing procedures with the size of the test being one of the things we specify. As mentioned, you would typically choose a value such as 0.05, 0.01, or 0.001 for the test size. However, someone reading your analysis might have a different size criterion in mind than what you chose, and might want to know (for instance) whether a particular test statistic would reject the null hypothesis at a smaller size than you've chosen. To give satisfaction to such readers, you might want to report not only whether a test yielded a rejection at the size you chose, but also the p-value of the test so that the reader can apply the size criterion she had in mind.

What, precisely, is a p-value? Let t be the value you obtain for your test statistic T, and suppose that as in a two-tailed test, the larger the (absolute) value of t, the greater is the evidence against the null. The p-value is calculated on the assumption that the null hypothesis is true. It indicates the probability of obtaining a test statistic value greater than or equal to t given that the null is true. That is, $p = \Pr(T \geq t)$ conditional on the null being true. Since you know the distribution of T when the null is true, and since you also know the value t of your test statistic, the p-value is easy to calculate.

Suppose that you choose the test size to be 0.05 and find that your test statistic is large enough to reject the null at that size. A p-value of (for example) $p = 0.0075$ indicates that with the value of the test statistic you have in hand, the null would also be rejected at a size of 0.01. However, it would not be rejected at the stricter size of 0.005. This can be helpful information for your reader. These days, many economists report p-values whenever they write up results of hypothesis tests.

8.7 Specification Errors and Hypothesis Tests

With either heteroskedasticity or cross-observation correlation in the disturbances, we can no longer make use of the key results involving the χ^2 distribution on which we have relied so far. This form of specification error destroys our ability to conduct hypothesis tests in the usual manner. There is, in fact, a way to proceed, using what is termed the "Eicker–White" correction that supplies estimates of the variance of $\hat{\mu}$ which are robust to heteroskedasticity, and which provides a means of conducting hypothesis tests in large samples. We're not yet in a position to understand this correction, which relies on laws of large numbers and

central limit theorems, but will come back to the issue later after those tools have been introduced.

Chapter 9

Multivariate Regression

In this chapter we present basic results on multivariate linear regression. If you have understood the material on the sample mean and variance, what's here should be readily comprehended. In its mathematics, the multivariate regression model is very similar to the approaches we've just explored in the univariate case. There are, however, more fundamental differences: new conceptual issues arise when we incorporate explanatory variables in an econometric model. We need to be especially careful about how we interpret the estimated coefficients attached to these variables. The section that follows introduces some of the issues.

9.1 Interpreting Regressions in Causal Terms

When you set out a regression model, the economic theory you bring to bear on it may be summarized in a representation of the behavior Y_i of agent i, expressed in a function of the general form

$$Y_i = g(X_i, \mathbf{Z}_i, \epsilon_i).$$

Let us assume that good data are available on Y_i and two of the three arguments of the function: X_i is the explanatory factor on which the theory is focused and $\mathbf{Z}_i$ is a vector of additional influences recognized in the theory and also available in the data. The third argument of the function, ϵ_i, represents all other factors influencing the agent's behavior that are not measured in our dataset.

If the theoretical model is skillfully manipulated, it may deliver a testable proposition in the form of a partial derivative. For instance, the proposition may be that

$$\frac{\partial Y_i}{\partial X_i} = \frac{\partial g(X_i, \mathbf{Z}_i, \epsilon_i)}{\partial X_i} > 0.$$

That is, with other things being held constant (*ceteris paribus*), a small increase in X_i is posited to *cause* a small increase in Y_i, the measure of the agent's behavior.

A fundamental concern in moving from theory to its implementation in an econometric model is whether we have a means of estimating this causal partial derivative. Unlike scientists who conduct experiments in laboratories, economists seldom find themselves in a position to fix the X_i value facing an agent and then change that value, holding all

other features of the experiment constant, to determine how the agent responds. Instead we must do what we can to draw inferences about the partial derivative from the "naturally-occurring" variation evident in a set of data $\{(Y_i, X_i, \mathbf{Z}_i),\ i = 1, \ldots, n\}$. We hope that the associations in the dataset between Y_i and X_i, with $\mathbf{Z}_i$ held constant by statistical methods, will give an adequate estimate of the partial derivative. But we cannot expect to hold ϵ_i constant in this way—it is, after all, an unmeasured, unobserved random variable—and we may worry that in the data-generating process, ϵ_i may be linked to X_i in some way that causes the two to move together. If they are in fact linked, our estimates will not represent the partial derivative $\partial Y_i / \partial X_i$ that we seek, but rather the compound quantity

$$\frac{d\,Y_i}{d\,X_i} = \frac{\partial g(X_i, \mathbf{Z}_i, \epsilon_i)}{\partial X_i} + \frac{\partial g(X_i, \mathbf{Z}_i, \epsilon_i)}{\partial \epsilon_i} \cdot \frac{d\,\epsilon_i}{d\,X_i},$$

from which the causal effect of interest, $\partial g / \partial X_i$, cannot be extracted. Hence, if we want to interpret our estimates in causal terms, we must be confident that X_i and ϵ_i are not systematically related. As we will see in a moment, this fundamental issue is addressed in a key assumption about the conditional mean of the unobserved ϵ given all right-hand side observed explanatory variables.

These issues were described in an amusing fashion in a 1973 interview published in George J. W. Goodman's book *Supermoney* (page 286), in which Daniel Yankelovich said the following in reference to regrettable common practices in applied statistics:

> The first step is to measure whatever can be easily measured. This is okay as far as it goes. The second step is to disregard that which cannot be measured, or give it an arbitrary quantitative value. This is artificial and misleading. The third step is to presume that what can't be measured easily isn't very important. This is blindness. The fourth step is to say that what can't easily be measured doesn't really exist. This is suicide.

Please keep these warnings about the ϵ disturbances in mind in specifying and interpreting your own regression models!

9.2 Basic Assumptions

In this chapter we examine linear representations of the g function. Let the random variable Y_i, which we term the *dependent variable*, be a linear function of a $k \times 1$ vector of *explanatory variables* $\mathbf{X}_i$ and a disturbance term ϵ_i, as follows:

$$Y_i = \mathbf{X}_i' \boldsymbol{\beta} + \epsilon_i. \tag{9.1}$$

Here $\boldsymbol{\beta}$ is a set of k unknown parameters to be estimated. To begin, we need to clarify what "linear" means in the present context. It means *linear in parameters*. The various explanatory variables can enter in a number of forms that imply a nonlinear relationship between a variable and its association with the dependent variable Y_i. For instance, if Y_i refers to the wage of person i, then we would expect to see grades of schooling, a key explanatory factor, will enter in linear form and also as grades squared. We'd normally expect the β coefficient

on grades to be positive but the coefficient on grades squared to be negative, so that the payoff in terms of wages of an extra year of schooling rises but at a decreasing rate.

In this linear relationship, the variables Y_i and $\mathbf{X}_i$ are observed, while the constants $\boldsymbol{\beta}$ and the random variable ϵ_i are not observed. Consider a dataset of n observations, each of which obeys equation (9.1). Stacking these observations, we obtain

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{9.2}$$

In this notation, the i-th *row* of the $\mathbf{X}$ matrix is $\mathbf{X}_i'$, the transpose of the $\mathbf{X}_i$ *column* vector. That is,

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix} = \begin{bmatrix} X_{1,1} & X_{1,2} & \cdots & X_{1,k} \\ X_{2,1} & X_{2,2} & \cdots & X_{2,k} \\ \vdots & \vdots & \vdots & \vdots \\ X_{n,1} & X_{n,2} & \cdots & X_{n,k} \end{bmatrix}$$

At times we'll find the notation of (9.1) to be more convenient, whereas at other times the compact vector–matrix notation of (9.2) will simplify things.

A fundamental assumption in the regression model is that

$$\mathrm{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{0}. \tag{9.3}$$

Note that this expression refers to the full $\boldsymbol{\epsilon}$ vector and the full $\mathbf{X}$ matrix. It is a very strong assumption.[1] Later, when we discuss asymptotic analysis, we will be able to weaken it to $\mathrm{E}(\epsilon_i|\mathbf{X}_i) = 0\ \forall i$. For now, however, we will maintain the strong version of the assumption. Taking expectations of equation (9.2) conditional on $\mathbf{X}$, we have

$$\mathrm{E}(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} + \mathrm{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$$

because $\mathrm{E}(\boldsymbol{\epsilon}|\mathbf{X}) = \mathbf{0}$ implies both $\mathrm{E}\,\boldsymbol{\epsilon} = \mathbf{0}$ and $\mathrm{E}\,\mathbf{X}'\boldsymbol{\epsilon} = \mathbf{0}$.

A second assumption in the regression model, important but not nearly so critical as the first, is that the variance matrix of the disturbances conditional upon $\mathbf{X}$ is

$$\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}. \tag{9.4}$$

Under this assumption, the variance of each ϵ_i is σ^2 and $\mathrm{E}\,\epsilon_i\epsilon_j = 0\ \forall i \neq j$.

A third assumption, which we will invoke only when it is absolutely necessary, posits that the distribution of the disturbances $\boldsymbol{\epsilon}$ is multivariate normal conditional upon $\mathbf{X}$,

$$\boldsymbol{\epsilon} \mid \mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{I}\right). \tag{9.5}$$

Because the multivariate normal distribution is completely specified by its mean vector and variance matrix, and neither of these is a function of $\mathbf{X}$, this assumption implies that the disturbances are also multivariate normal (with mean $\mathbf{0}$ and variance matrix $\sigma^2\mathbf{I}$) *unconditional* on the $\mathbf{X}$ covariates. In essence, the disturbance vector $\boldsymbol{\epsilon}$ is assumed fully independent of the $\mathbf{X}$ covariates.

[1] The assumption fails to hold in the simple time-series model $Y_t = \beta Y_{t-1} + \epsilon_t$. It would also fail to hold in a two-equation time-series model whose main equation of interest is $Y_t = \beta_0 + \beta_1 X_t + \epsilon_t$, with a second equation $X_{t+1} = \alpha_0 + \alpha_1 Y_t + u_{t+1}$. In this case, ϵ_t is associated with X_{t+1} through the second equation—but this kind of linkage is disallowed by the assumption that $\mathrm{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$.

9.3 The OLS Estimator

From the above, we have obtained $\mathrm{E}(Y_i|\mathbf{X}) = \mathbf{X}_i'\boldsymbol{\beta}$. How might we estimate these $\boldsymbol{\beta}$ parameters? There are two approaches, yielding the same result—the ordinary least squares estimator—both of which rely on the *analogy principle* of estimation by which sample moments are substituted for population moments.

Both approaches begin with the conditional expectation concept. The first approach proceeds via the method of moments. To develop this approach, it will simplify matters if we think of the sequence $\{(Y_i, \mathbf{X}_i, \epsilon_i)\}$ as an iid sequence.[2] Post-multiply the column vector $\mathbf{X}_i$ by the scalar random variable $Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i$, giving

$$\mathbf{X}_i Y_i = \mathbf{X}_i \mathbf{X}_i'\boldsymbol{\beta} + \mathbf{X}_i \epsilon_i.$$

Now take expectations of the $k \times 1$ vector $\mathbf{X}_i Y_i$, first by conditioning on $\mathbf{X}$ and then unconditioning. This yields a set of k equations expressed in terms of β and the unconditional expectations,

$$\mathrm{E}(\mathbf{X}_i Y_i) = \mathrm{E}(\mathbf{X}_i \mathbf{X}_i')\boldsymbol{\beta}.$$

The $\mathbf{X}_i \epsilon_i$ term disappeared in the first step of iterated expectations, since $\mathrm{E}(\mathbf{X}_i \epsilon_i|\mathbf{X}) = \mathbf{X}_i \mathrm{E}(\epsilon_i|\mathbf{X}) = \mathbf{0}$. Assuming that the $k \times k$ matrix $\mathrm{E}(\mathbf{X}_i \mathbf{X}_i')$ is invertible, we obtain an exact solution for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} = \left[\mathrm{E}(\mathbf{X}_i \mathbf{X}_i')\right]^{-1} \mathrm{E}(\mathbf{X}_i Y_i).$$

Of course, this equation for $\boldsymbol{\beta}$ does not yield an *estimator* as such, because the (unconditional) expectations are not known quantities. However, their presence in the equation suggests an approach that employs sample averages as *estimates* of the unknown expectations.

For the $k \times 1$ vector $\mathrm{E}\,\mathbf{X}_i Y_i$, we use the sample average $(1/n)\sum_{i=1}^n \mathbf{X}_i Y_i$, and for the $k \times k$ matrix $\mathrm{E}(\mathbf{X}_i \mathbf{X}_i')$, we use the average outer product $(1/n)\sum_{i=n}^n \mathbf{X}_i \mathbf{X}_i'$. Then, inserting averages in place of expectations, we obtain an actual estimator

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n \mathbf{X}_i Y_i\right).$$

Using the vector–matrix notation and cancelling the n's, we obtain the possibly more familiar expression $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. To fully understand this representation of $\hat{\boldsymbol{\beta}}$, you should verify for yourself that

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_n\end{bmatrix} \begin{bmatrix}\mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n'\end{bmatrix} = \sum_{i=1}^n \mathbf{X}_i \mathbf{X}_i'$$

[2] By "iid," we mean that the sequence $\{(Y_i, \mathbf{X}_i, \epsilon_i)\}$ is "independent and identically distributed" over i. Among other things, this implies that the expected values of $\mathbf{X}_i$ and functions of $\mathbf{X}_i$ such as $\mathrm{E}\,\mathbf{X}_i \mathbf{X}_i'$ do not vary with i. But note that the iid assumption is made here only for convenience. As we will see later, it is relatively easy to generalize to accommodate non-identically distributed sequences.

and that

$$\mathbf{X}'\mathbf{Y} = \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2 & \cdots & \mathbf{X}_n\end{bmatrix} \begin{bmatrix}Y_1 \\ Y_2 \\ \vdots \\ Y_n\end{bmatrix} = \sum_{i=1}^{n} \mathbf{X}_i Y_i.$$

The other approach to $\hat{\beta}$ is to consider the sample analog to $\mathrm{E}(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2$. Recall that the conditional mean is the best predictor of Y_i in the sense that it minimizes $\mathrm{E}(Y_i - h(\mathbf{X}))^2$ among all functions $h(\mathbf{X})$, and in the case at hand, the conditional mean is linear in the parameters $\boldsymbol{\beta}$. Examine the the sample average

$$S(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\boldsymbol{\beta})^2,$$

or, in vector–matrix form,

$$S(\boldsymbol{\beta}) = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

The idea is to find the value $\hat{\boldsymbol{\beta}}$ that minimizes this sum of squares, with $\hat{\boldsymbol{\beta}}$ filling in for the true value of $\boldsymbol{\beta}$ that minimizes the expectation. (The $1/n$ factor can be discarded, as its presence does not affect the value of $\hat{\boldsymbol{\beta}}$.) We proceed by first writing out the sum of squares in the form

$$S(\boldsymbol{\beta}) = \mathbf{Y}'\mathbf{Y} - 2 \cdot \mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta},$$

and then find $\hat{\boldsymbol{\beta}}$ by setting k partial derivatives to zero,

$$\frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}} = -2 \cdot \mathbf{X}'\mathbf{Y} + 2 \cdot \mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{0}_{k\times 1}, \tag{9.6}$$

which yields the solution $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. We have reached the same result via two different routes.

Let $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, a vector of *residuals* that are the difference between $\mathbf{Y}$ and its fitted value $\mathbf{X}\hat{\boldsymbol{\beta}}$. Rearrange the first-order condition (9.6) as $\mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) = \mathbf{0}$ to obtain

$$\mathbf{X}'\mathbf{e} = \mathbf{0}_{k\times 1}. \tag{9.7}$$

That is, the residual vector $\mathbf{e}$ is *orthogonal* to each of the column vectors of $\mathbf{X}$. (When viewed column-wise, the $\mathbf{X}$ matrix is a collection of k vectors of length n, each such vector containing observations on a given explanatory variable. We can also write the orthogonality condition in the arithmetically equivalent form

$$\sum_{i=1}^{n} \mathbf{X}_i e_i = \mathbf{0}.$$

With $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$, we see that the residual vector is also orthogonal to the vector of fitted values, that is

$$\hat{\mathbf{Y}}'\mathbf{e} = \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{e} = 0.$$

Indeed, and by the same reasoning, the residuals $\mathbf{e}$ are orthogonal to *any linear combination* of the individual columns of $\mathbf{X}$, that is, any linear combination of the explanatory variables $\mathbf{X}\alpha$ for $\alpha \neq \mathbf{0}$.

Orthogonality is a very important computational property of least squares. To gain a better sense of its role, we explore in the next section several computational measures of the "fit" between the actual $\mathbf{Y}$ and its fitted or predicted value $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$.

9.4 Measures of Goodness of Fit

Assume that the coefficients β in the linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ have been estimated by ordinary least squares, yielding estimated values $\hat{\beta}$. If our aim is to assess the goodness of fit of this model, how should we proceed? There are three main computational methods, each of which yields a descriptive measure of fit known as an R^2. The difference between these approaches has to do with the inclusion of a constant term in the model. (If the first column of $\mathbf{X}$ is ι, a column vector of ones, then $\hat{\beta}_1$ is the estimated constant term.) The thing to remember is to avoid the second R_2^2 measure—which is, in fact, the measure most often calculated in standard statistical packages—if you have specified a model without a constant term.

Decomposing the total sum of squares for Y

Let us first consider the approach that does not require a constant term. Write the fitted regression in the form

$$\mathbf{Y} = \mathbf{X}\hat{\beta} + \mathbf{e}$$

where $\mathbf{e}$ is an n-vector of regression residuals with the property $\mathbf{X}'\mathbf{e} = \mathbf{0}$. Then $\mathbf{Y}'\mathbf{Y}$, the total sum of squares for $\mathbf{Y}$, is

$$\mathbf{Y}'\mathbf{Y} = (\hat{\beta}'\mathbf{X}' + \mathbf{e}')(\mathbf{X}\hat{\beta} + \mathbf{e}) = \hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta} + \mathbf{e}'\mathbf{e}.$$

The proportion of $\mathbf{Y}'\mathbf{Y}$ that is "explained" (in a purely computational sense) by the regression is therefore

$$R_1^2 = \frac{\hat{\beta}'\mathbf{X}'\mathbf{X}\hat{\beta}}{\mathbf{Y}'\mathbf{Y}}.$$

This definition of R^2 makes sense whether or not a constant term is specified. If we substitute $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ for $\hat{\beta}$, the R_1^2 measure can also be written in the form

$$R_1^2 = \frac{\mathbf{Y}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}}{\mathbf{Y}'\mathbf{Y}}.$$

This is sometimes termed the "uncentered" version of R^2.

Decomposing the sum of squares of Y about $\bar{\mathbf{Y}}$

Now we re-cast the problem in terms of explaining the squared deviations of $\mathbf{Y}$ from its sample mean $\bar{\mathbf{Y}}$, that is, $(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})$ where $\bar{\mathbf{Y}}$ denotes the vector with all elements set

equal to the sample mean of $\mathbf{Y}$. To do this it is convenient to re-introduce the $n \times n$ matrix $\mathbf{M}$ that we have seen earlier,

$$\mathbf{M} = \mathbf{I} - \frac{\iota\iota'}{n},$$

where ι is a column vector of ones, $\mathbf{I}$ is an $n \times n$ identity matrix, and as you will recall, $\mathbf{M}$ is symmetric idempotent. $\mathbf{M}$ has the property that

$$\mathbf{MY} = \mathbf{Y} - \bar{\mathbf{Y}}.$$

We again write the fitted regression model as

$$\mathbf{Y} = \boldsymbol{\iota} \cdot \hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \mathbf{e}$$

with the constant term of the model $\hat{\beta}_1$ now distinguished from the other regression coefficients. The regression residual $\mathbf{e}$ has an arithmetic mean of zero, this last being an implication of the least squares first-order conditions $\mathbf{X}'\mathbf{e} = \mathbf{0}$ when a constant term $\boldsymbol{\iota}$ is in the model. Multiplying through by $\mathbf{M}$, we obtain

$$\mathbf{MY} = \mathbf{M}\boldsymbol{\iota}\hat{\beta}_1 + \mathbf{MX}_2\hat{\beta}_2 + \mathbf{Me}.$$

Since $\boldsymbol{\iota}\hat{\beta}_1$ is constant and $\mathbf{e}$ has an arithmetic mean of zero, this reduces to

$$\mathbf{MY} = \mathbf{MX}_2\hat{\beta}_2 + \mathbf{e}.$$

Forming the total sum of squares, we have

$$(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}) = \mathbf{Y}'\mathbf{M}'\mathbf{MY} = \hat{\beta}_2'\mathbf{X}_2'\mathbf{M}'\mathbf{MX}_2\hat{\beta}_2 + \mathbf{e}'\mathbf{e}$$

using $\mathbf{M}' = \mathbf{M}$, $\mathbf{Me} = \mathbf{e}$ and $\mathbf{X}_2'\mathbf{e} = \mathbf{0}$. Written differently, this is

$$(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}}) = \hat{\beta}_2'(\mathbf{X}_2 - \bar{\mathbf{X}}_2)'(\mathbf{X}_2 - \bar{\mathbf{X}}_2)\hat{\beta}_2 + \mathbf{e}'\mathbf{e}.$$

We now have a second definition of R^2,

$$R_2^2 = \frac{\hat{\beta}_2'(\mathbf{X}_2 - \bar{\mathbf{X}}_2)'(\mathbf{X}_2 - \bar{\mathbf{X}}_2)\hat{\beta}_2}{(\mathbf{Y} - \bar{\mathbf{Y}})'(\mathbf{Y} - \bar{\mathbf{Y}})} = \frac{\hat{\beta}_2'\mathbf{X}_2'\mathbf{MX}_2\hat{\beta}_2}{\mathbf{Y}'\mathbf{MY}},$$

which is occasionally described as the "centered" version of R^2 to distinguish it from the version we saw first. With a bit of additional manipulation, the centered R_2^2 measure can be linked to an $\mathcal{F}$-test of the hypothesis $H_0 : \beta_2 = \mathbf{0}$.

The sample correlation between Y and $\hat{\mathbf{Y}}$

A third descriptive measure of fit—and to my mind, at least, the best-justified—is the sample correlation between $\mathbf{Y}$ and its predicted value $\hat{\mathbf{Y}}$. As are the other two measures, this one is bounded between 0 and 1, but the sample correlation is more generally applicable in that it can be employed with nonlinear models of the form $Y_i = g(\mathbf{X}_i, \theta) + \boldsymbol{\epsilon}_i$ in which $\hat{Y}_i = g(\mathbf{X}_i, \hat{\theta})$. It is a little surprising, on reflection, that the sample correlation has not always been the dominant measure of goodness-of-fit.

In a sense, this measure has in fact been historically dominant: It can be shown that if the (linear) model contains a constant term, the conventional R_2^2 measure equals the square of the sample correlation between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$. So, at least for linear models with constant terms, there is no real difference between the sample correlation measure of fit and the square of this measure, which is the conventional R_2^2.

9.5 Statistical Properties of $\hat{\boldsymbol{\beta}}$

Having explored some of the computational features of the least-squares estimator, we now turn to more important matters. What are the *statistical* properties of such estimators?

Substituting $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for $\mathbf{Y}$ in the expression $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, we obtain

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}.$$

Hence, using the fundamental assumption $\mathrm{E}(\epsilon \mid \mathbf{X}) = \mathbf{0}$ and iterated expectations,

$$\mathrm{E}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathrm{E}\left(\mathrm{E}(\,(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}|\mathbf{X})\right) = \boldsymbol{\beta}. \tag{9.8}$$

We see that $\hat{\boldsymbol{\beta}}$ is *unbiased* for $\boldsymbol{\beta}$.

Note that the formula for the $\hat{\beta}$ coefficients involves a second set of regression coefficients, $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$, which would come from a hypothetical regression of ϵ on $\mathbf{X}$. The notion of regressing $\boldsymbol{\epsilon}$ on $\mathbf{X}$ can only be hypothetical because $\boldsymbol{\epsilon}$ is not observed. Nevertheless, the coefficients from this regression are always embedded in the $\hat{\beta}$ coefficients. In saying that $\hat{\beta} = \beta + \hat{\theta}$ is unbiased, what we mean is that the $\hat{\theta}$ coefficients from the hypothetical regression have an expected value of zero. In any given sample of data, however, the realized value of these coefficients will not be zero. We refer to the coefficients $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$ as *sampling error*. We will consider them further in the closing section of this chapter.

To derive the variance matrix of $\hat{\boldsymbol{\beta}}$, we consider the expected outer product

$$\begin{aligned}\mathrm{Var}\,\hat{\boldsymbol{\beta}} &= \mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' \\ &= \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}.\end{aligned}$$

Again using iterated expectations and invoking the assumption $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}) = \sigma^2\mathbf{I}$, we obtain

$$\mathrm{Var}\,\hat{\boldsymbol{\beta}} = \sigma^2\,\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}. \tag{9.9}$$

To make use of this expression, we will need estimators for its two unknowns: the scalar parameter σ^2 and the $k \times k$ matrix $\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}$. For the matrix, we'll use $(\mathbf{X}'\mathbf{X})^{-1}$. What can be used to estimate σ^2?

9.6 An Estimator of σ^2

Consider

$$\begin{aligned}s^2 &= \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}})^2 \\ &= \frac{1}{n-k}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \frac{1}{n-k}\mathbf{e}'\mathbf{e},\end{aligned}$$

where $\mathbf{e}$ is the residual vector $\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}$. Substituting $\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ for $\mathbf{Y}$ and $\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$ for $\hat{\boldsymbol{\beta}}$, we obtain

$$\begin{aligned}\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{\epsilon} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ &= \mathbf{M}_X\boldsymbol{\epsilon}\end{aligned}$$

with $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The $\mathbf{M}_X$ matrix is $n \times n$, symmetric, and idempotent. Its rank equals its trace, and

$$\begin{aligned}\operatorname{trace}\mathbf{M}_X &= \operatorname{trace}\mathbf{I} - \operatorname{trace}(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= n - \operatorname{trace}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= n - k.\end{aligned}$$

We can now determine the expected value of s^2 without further ado. First, by the properties of traces,

$$\operatorname{trace}(\boldsymbol{\epsilon}'\mathbf{M}_X\boldsymbol{\epsilon}) = \operatorname{trace}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{M}_X),$$

Now condition on $\mathbf{X}$ in the first step of an iterated expectations approach, which makes $\mathbf{M}_X$ akin to a matrix of constants while the conditioning is in effect, and undertake the following sequence of operations:

$$\mathrm{E}\left(\operatorname{trace}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{M}_X)\mid\mathbf{X}\right) = \operatorname{trace}\left(\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X})\,\mathbf{M}_X\right) = \operatorname{trace}(\sigma^2\mathbf{M}_X) = \sigma^2\operatorname{trace}\mathbf{M}_X = \sigma^2(n-k).$$

In other words, $\mathrm{E}(\epsilon'\mathbf{M}_X\epsilon \mid \mathbf{X}) = \sigma^2(n-k)$. Removing the conditioning on $\mathbf{X}$ (in the second step of iterated expectations) leaves this result unchanged. In short, $\mathrm{E}\,s^2 = \sigma^2$, that is, the estimator of the variance is unbiased.

Can we make further progress by invoking normality, that is, by assuming that $\epsilon|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$? Rewrite

$$s^2 = \frac{1}{n-k}\,\boldsymbol{\epsilon}'\mathbf{M}_X\boldsymbol{\epsilon} = \frac{\sigma^2}{n-k}\mathbf{u}'\mathbf{M}_X\mathbf{u}$$

with $\mathbf{u} \equiv \sigma^{-1}\boldsymbol{\epsilon}$. Conditional on $\mathbf{X}$, the Chapter 2 theorem about quadratic forms can be applied with $\mathbf{M}_X$ treated *as if* it were a matrix of constants. Hence, conditional on $\mathbf{X}$, the quadratic form $\mathbf{u}'\mathbf{M}_X\mathbf{u} \sim \chi^2_{n-k}$ and $\mathrm{E}\,\mathbf{u}'\mathbf{M}_X\mathbf{u} = n-k$. It follows that

$$\mathrm{E}\,s^2|\mathbf{X} = \frac{\sigma^2}{n-k}\cdot(n-k) = \sigma^2. \tag{9.10}$$

Note that the unconditional expectation of s^2 is also equal to σ^2. Also, we obtain a new result, that

$$\operatorname{Var}s^2|\mathbf{X} = \operatorname{Var}\left(\frac{\sigma^2}{n-k}\mathbf{u}'\mathbf{M}_X\mathbf{u}\right) = \frac{2\sigma^4}{n-k}, \tag{9.11}$$

this because $\operatorname{Var}\mathbf{u}'\mathbf{M}_X\mathbf{u} = 2(n-k)$ conditional on $\mathbf{X}$. The same result holds unconditionally.

9.7 Independence of $\hat{\beta}$ and s^2 under Normality

Recall that we can write

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \epsilon \\ (n-k) \cdot s^2 &= \epsilon' \mathbf{M}_X \epsilon\end{aligned}$$

with $\epsilon|\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Conditional on $\mathbf{X}$, then, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \cdot \epsilon$ can be viewed as if it were a matrix of constants pre-multiplying a mean zero, variance $\sigma^2\mathbf{I}$ normal vector. Likewise, conditional on $\mathbf{X}$, we see that $\epsilon'\mathbf{M}_X\epsilon$ is a quadratic form in that same normal vector, with $\mathbf{M}_X$ again being akin to a matrix of constants.

We can prove that $\hat{\boldsymbol{\beta}}$ is independent of s^2 conditional on $\mathbf{X}$ by showing that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_X = \mathbf{0}$. (Recall the theorem from Chapter 2.) Simply carry out the matrix multiplication,

$$\begin{aligned}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_X &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}') \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{0},\end{aligned}$$

and the result is proved. Note that this result is *conditional* on $\mathbf{X}$—we will use it in our next chapter on hypothesis testing.

Several of the previous results might lead you to think that $\hat{\beta}$ and s^2 are independent unconditionally—but please do not draw this conclusion. We *cannot* conclude from conditional independence given $\mathbf{X}$ that $\hat{\beta}$ and s^2 are independent unconditionally. What we have shown above is that the joint distribution of $\hat{\beta}$ and s^2 given $\mathbf{X}$ can be factored as $f(\hat{\beta}|\mathbf{X})g(s^2|\mathbf{X})$. But if we "integrate out" $\mathbf{X}$ to obtain the unconditional joint distribution, using

$$p(\hat{\beta}, s^2) = \int_{R_X} f(\hat{\beta}|\mathbf{X})g(s^2|\mathbf{X})h(\mathbf{x})d\mathbf{x},$$

then p, the unconditional joint distribution of $\hat{\beta}$ and s^2, does not necessarily factor neatly into the product of their unconditional marginal distributions.

9.8 The Frisch–Waugh–Lovell (FWL) theorem

The FWL theorem is an extremely useful computational result, which allows us to focus in on one (or a sub-set) of the OLS slope parameter estimates. A full treatment is deferred to Chapter 11, but we can supply the most often-used piece of the theorem here. We will make use of it extensively in our next chapter on hypothesis testing.

Consider a linear model written in the form $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\delta + \epsilon$, and suppose that the δ parameter is of main interest. The FWL theorem shows that

$$\hat{\delta} = \left(\mathbf{Z}'\mathbf{M}_X\mathbf{Z}\right)^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{Y}$$

and similarly, that

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{M}_Z\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{M}_Z\mathbf{Y}$$

in which the $n \times n$ matrices $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and $\mathbf{M}_Z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$. Note that when post-multiplied by any n-vector $\mathbf{w}$, the expression $\mathbf{M}_X\mathbf{w}$ creates OLS *residuals* from a regression of $\mathbf{w}$ on $\mathbf{X}$, and $\mathbf{M}_Z\mathbf{w}$ creates OLS residuals from a regression of $\mathbf{w}$ on $\mathbf{Z}$.

The proof of the FWL theorem provides a good work-out with the computational aspects of OLS. We begin by writing $\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}\hat{\delta} + \mathbf{e}$, in which $\mathbf{e}$ is a vector of OLS residuals that is orthogonal to each variable in $\mathbf{X}$ and each variable in $\mathbf{Z}$. That is, both $\mathbf{X}'\mathbf{e} = \mathbf{0}$ and $\mathbf{Z}'\mathbf{e} = \mathbf{0}$ from the OLS first-order conditions. Now multiply through by $\mathbf{Z}'\mathbf{M}_X$:

$$\mathbf{Z}'\mathbf{M}_X\mathbf{Y} = \mathbf{Z}'\mathbf{M}_X\mathbf{X}\hat{\boldsymbol{\beta}} + \mathbf{Z}'\mathbf{M}_X\mathbf{Z}\hat{\delta} + \mathbf{Z}'\mathbf{M}_X\mathbf{e}.$$

This can be simplified in three ways. First, note that we must have $\mathbf{M}_X\mathbf{X} = \mathbf{0}$ because (considering each column of $\mathbf{X}$ one at a time) the residuals from a regression of a variable in $\mathbf{X}$ on itself and the other $\mathbf{X}$ variables must be zero. So the first term on the right-hand side drops out. Look now at the last term, $\mathbf{Z}'\mathbf{M}_X\mathbf{e}$. Again from the OLS first-order conditions, we have $\mathbf{M}_X\mathbf{e} = \mathbf{e}$. (Write out $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and post-multiply by $\mathbf{e}$ to verify this.) Hence we have $\mathbf{Z}'\mathbf{M}_X\mathbf{e} = \mathbf{Z}'\mathbf{e}$ and finally—yet again by the OLS first-order conditions—we have $\mathbf{Z}'\mathbf{e} = \mathbf{0}$. We're thus left with

$$\mathbf{Z}'\mathbf{M}_X\mathbf{Y} = \mathbf{Z}'\mathbf{M}_X\mathbf{Z}\hat{\delta}$$

and the result of the theorem follows immediately.

9.9 Specification Errors

What are the consequences of omitting relevant explanatory variables from the specification of the regression model? This sort of specification error wreaks havoc with the least-squares estimators $\hat{\beta}$ and s^2. As we work through the issues, you will see connections to the problem of causal interpretation discussed in the opening section of this chapter. The likelihood of specification error is one of the reasons for being cautious about claiming that one's regression provides estimates of true causal effects.

To understand the consequences stemming from the omission of a relevant covariate, write the true model as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\gamma + \mathbf{u}$$

with $\gamma \neq \mathbf{0}$ (it is in this sense that the $\mathbf{Z}$ variable is relevant) and with $\mathbf{u}$ being a vector of disturbances having the property that $\mathrm{E}(\mathbf{u} \mid \mathbf{X}, \mathbf{Z}) = \mathbf{0}$ and $\mathrm{E}(\mathbf{u}\mathbf{u}' \mid \mathbf{X}, \mathbf{Z}) = \sigma_u^2\mathbf{I}$. Unfortunately you, the hapless researcher, have mistakenly understood the model to be

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \epsilon,$$

thus omitting the $\mathbf{Z}$ variables from your specification. The disturbance term of your mis-specified regression is actually $\epsilon \equiv \mathbf{Z}\gamma + \mathbf{u}$, a composite that includes the true disturbances and the omitted variables. What damage follows from this error?

In the mis-specified model, the least squares estimator of β is

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\gamma + \mathbf{u})$$

where we have used the true model for $\mathbf{Y}$. This simplifies to

$$\hat{\beta} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}.$$

Now apply iterated expectations to the expression, first conditionally on $\mathbf{X}$ and $\mathbf{Z}$, and then unconditionally. We obtain

$$\mathrm{E}\,\hat{\beta} = \boldsymbol{\beta} + \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma,$$

noting that the term involving $\mathbf{u}$ disappeared in the first step because of the fundamental assumption $\mathrm{E}(\mathbf{u} \mid \mathbf{X}, \mathbf{Z}) = \mathbf{0}$. Clearly $\hat{\boldsymbol{\beta}}$ is biased—its expectation no longer coincides with the true $\boldsymbol{\beta}$.

Let us study the bias term, $\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma$, in a bit more detail. Notice that $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$ is in the form of an OLS slopes estimator from a regression of the omitted $\mathbf{Z}$ on the included $\mathbf{X}$. (If more than one $\mathbf{Z}$ variable has been omitted, then $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$ can be viewed as a set of such least squares coefficients organized column-by-column.) In this situation we obviously cannot interpret the k-th slope coefficient $\hat{\beta}_k$ as an estimate of the partial derivative of Y_i with respect to $X_{i,k}$. It is instead an estimate of a compound quantity, reflecting not only the effects of $X_{i,k}$ but also the associations linking the $\mathbf{X}$ variables to the omitted $\mathbf{Z}$ variables that enter the composite disturbance $\epsilon = \mathbf{Z}\gamma + \mathbf{u}$. A causal interpretation of $\hat{\boldsymbol{\beta}}$ is thus rendered untenable.

The bias consists of two parts. Letting $\hat{\theta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}$ represent the coefficients from a regression of the omitted $\mathbf{Z}$ on $\mathbf{X}$, we may rewrite the expected value of $\hat{\boldsymbol{\beta}}$ as

$$\mathrm{E}\,\hat{\beta} = \beta + \mathrm{E}\,\hat{\theta} \cdot \gamma.$$

The first component of bias, $\mathrm{E}\,\hat{\theta}$, is often termed the "population regression" of $\mathbf{Z}$ on $\mathbf{X}$. If $\mathbf{Z}$ and $\mathbf{X}$ happen to be unrelated in the sense that the population regression coefficients are all zero, then $\hat{\beta}$ would be unbiased. But clearly this would be a very special case. If we have omitted $\mathbf{Z}$ because it is nowhere to be found in our dataset, we can at least speculate about what the population regression coefficients $\mathrm{E}\,\hat{\theta}$ are likely to be, and if we are really fortunate, we may be able to find estimates of this regression somewhere in the literature. The γ parameter is the second component of the bias. Again we may be able to speculate about the signs of this parameter and possibly even its magnitude. A good part of the art of writing about a mis-specified regression is in providing your reader with constructive and insightful speculation along these lines.

What happens to s^2 as a result of omitting $\mathbf{Z}$? We can anticipate that s^2 will be biased, but in this case there is no need for speculation about the direction of bias: we know that $\mathrm{E}\,s^2 \geq \sigma_u^2$, that is, the bias is in an upward direction. The proof is as follows.

The least-squares residuals $\mathbf{e}$ from the mis-specified regression are

$$\begin{aligned}\mathbf{e} &= \mathbf{Y} - \mathbf{X}\hat{\beta} \\ &= \mathbf{Z}\gamma - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\gamma + \mathbf{u} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \\ &= \mathbf{M}_X\mathbf{Z}\gamma + \mathbf{M}_X\mathbf{u}\end{aligned}$$

with $\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, the familiar idempotent matrix. Hence, the sum of squared residuals is

$$\mathbf{e}'\mathbf{e} = \gamma'\mathbf{Z}'\mathbf{M}_X\mathbf{Z}\gamma + \mathbf{u}'\mathbf{M}_X\mathbf{u} + \mathbf{u}'\mathbf{M}_X\mathbf{Z}\gamma + \gamma'\mathbf{Z}'\mathbf{M}_X\mathbf{u}.$$

By iterated expectations the expected values of the last two terms on the right are zero. By our earlier arguments, $\mathrm{E}\,\mathbf{u}'\mathbf{M}_X\mathbf{u} = \sigma_u^2 \cdot (n-k)$ with k being the number of included $\mathbf{X}$ variables. Furthermore, because $\mathbf{M}_X$ is positive semidefinite, the quadratic form $\gamma'\mathbf{Z}'\mathbf{M}_X\mathbf{Z}\gamma \geq 0$. Pulling all this together,

$$\mathrm{E}\,s^2 = \mathrm{E}\left(\frac{1}{n-k}\mathbf{e}'\mathbf{e}\right) = \mathrm{E}\left(\frac{1}{n-k}\gamma'\mathbf{Z}'\mathbf{M}_X\mathbf{Z}\gamma + \sigma_u^2\right) \geq \sigma_u^2.$$

Upward bias in s^2 relative to σ_u^2 is exactly what we should have expected, given that in the mis-specified regression the disturbance $\epsilon \equiv \mathbf{Z}\gamma + \mathbf{u}$ and thus s^2 reflects not only the variance of the elements of $\mathbf{u}$ but also the variation that comes from the omitted variables $\mathbf{Z}$.

It is important to see that the arguments we have made concerning the omission of a relevant covariate also apply to the model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon$$

from which *no* relevant covariate has been omitted but for which, nevertheless, we have $\mathrm{E}(\epsilon \mid \mathbf{X}) \neq \mathbf{0}$. When there exists a correlation of this sort linking one or more of the $\mathbf{X}$ variables to the disturbance term, then

$$\mathrm{E}\,\hat{\beta} = \beta + \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon,$$

which can be written as

$$\mathrm{E}\,\hat{\beta} = \beta + \mathrm{E}\,\hat{\theta},$$

where $\hat{\theta}$ is the set of regression coefficients from a hypothetical regression of ϵ on $\mathbf{X}$. Earlier we referred to $\hat{\theta}$ as "sampling error," which is an acceptable label if $\mathrm{E}\,\hat{\theta} = \mathbf{0}$, but with $\mathrm{E}(\epsilon \mid \mathbf{X}) \neq \mathbf{0}$, we have $\mathrm{E}\,\hat{\theta} \neq \mathbf{0}$. In this case we have what might be termed *systematic* sampling error and it compromises our ability to interpret $\hat{\beta}$ in causal terms.

The problem was recognized in the earliest days of econometrics when basic supply-and-demand models were being considered. Let the supply equation be rendered as $\mathbf{Q}_s = \beta_0 + \mathbf{P}\beta_1 + \epsilon_s$ and the demand equation as $\mathbf{Q}_d = \alpha_0 + \mathbf{P}\alpha_1 + \epsilon_d$, with $\mathbf{P}$ representing price. Price in its turn is determined by the intersection of the supply and demand curves, giving in equilibrium

$$\mathbf{P} = \frac{\alpha_0 - \beta_0}{\beta_1 - \alpha_1} + \frac{1}{\beta_1 - \alpha_1} \cdot (\epsilon_d - \epsilon_s).$$

We see that price is correlated with both ϵ_d and ϵ_s, and therefore bias will infect ordinary least squares estimates of the α parameters of the demand equation and the supply equation's β parameters. In this example both the supply and demand equations are correctly specified, but the economic model yielding these equations also implies that $\mathrm{E}(\epsilon_j \mid \mathbf{P}) \neq \mathbf{0}$ for $j = s, d$. The early econometricians realized that methods other than simple least squares would be needed in situations such as these, and this realization led to the development of the instrumental variables method.

Another prominent case—and this is a case that applied econometricians confront every working day—is that of measurement error. In this instance, the true model is

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\delta + \epsilon$$

but only a mis-measured version of $\mathbf{Z}$ is available, that is, $\tilde{\mathbf{Z}} = \mathbf{Z} + \mathbf{m}$, the vector $\mathbf{m}$ being a vector of measurement errors. Writing the model in terms of $\mathbf{X}$ and $\tilde{\mathbf{Z}}$, we have

$$\mathbf{Y} = \mathbf{X}\beta + \tilde{\mathbf{Z}}\delta + \mathbf{v}$$

where the composite disturbance term $\mathbf{v} = \epsilon - \mathbf{m}\delta$. Algebraically this is identical to the model written out in terms of $\mathbf{X}$ and the perfectly measured $\mathbf{Z}$, but statistically it is a very different creature. Since the measurement errors $\mathbf{m}$ are unobserved, they are included in the composite disturbance term $\mathbf{v}$ and (with $\mathbf{m}$ also being embedded in $\tilde{\mathbf{Z}} = \mathbf{Z} + \mathbf{m}$), a systematic association links $\tilde{\mathbf{Z}}$ to the composite disturbance.

There are many additional circumstances in which we would suspect that an otherwise correctly specified equation might suffer from the problem of $\mathrm{E}(\epsilon \mid \mathbf{X}) \neq \mathbf{0}$. In labor and health economics, for example, we are often interested in the effect of intervention or training programs on subsequent wages, employment, or health. But when such a program is offered, some individuals will choose to participate in it and others will not. It is likely that the individuals who are most strongly motivated to participate are the same people who are most likely to show better performance (in wages, health, and the like) after the intervention. Indeed, even in the absence of the program these motivated people might have earned higher wages and so forth. When people select themselves into programs on this basis, how can we separate out the causal effects of any given program from the confounding effects of unmeasured motivation? If the regression includes program participation as a covariate, the estimated coefficient on the variable will undoubtedly reflect the true causal effect of the program mixed together with the correlation between the program variable and the disturbance term that arises from the motivation–participation link. We would have little basis for claiming that the regression isolates *only* the causal effect of the program. Much the same argument would apply whenever we include on the right-hand side of the regression equation an explanatory variable that reflects individual choice. Economists are therefore reluctant to specify models with such choice variables on the right-hand side, recognizing that when theory leads them in this direction methods other than least squares will be required to estimate the parameters of the equation.

A quite different form of specification error stems from mis-specification of the covariance matrix of the disturbances. We have been assuming that $\mathrm{E}\,\epsilon\epsilon' \mid \mathbf{X} = \sigma^2\mathbf{I}$, but what would happen if instead $\mathrm{E}\,\epsilon\epsilon' \mid \mathbf{X} = \mathbf{V}$? Considering

$$\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon,$$

we see that so long as the fundamental assumption $\mathrm{E}(\epsilon \mid \mathbf{X}) = \mathbf{0}$ holds, $\hat{\beta}$ remains unbiased. However, the covariance matrix of $\hat{\beta}$ is no longer $\sigma^2\,\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}$. Rather,

$$\mathrm{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

by iterated expectations. The consequence is that all *hypothesis tests* based on the mistaken assumption of a scalar covariance matrix will be incorrect. In this case, we can still interpret $\hat{\beta}$ in causal terms, but we cannot test propositions about β, or at least we cannot do so in the usual way. Later we will return to this form of mis-specification and explore how the method of estimated generalized least squares restores our ability to conduct hypothesis tests.

9.10 Running OLS regressions in R

```
#-------------------------------------------------------------------------#
# Read in a panel-data data.frame on country savings percentages, real GDP,
# age composition, and world region. Then run some simple regressions and
# conduct a hypothesis test using the F distribution.
#-------------------------------------------------------------------------#

# NOTE: The full program is posted on Blackboard; this is only an excerpt.
#  The data are stored in DT, a data.table.

#-------------------------------------------------------------------------#
# OK, let's run some regressions!
#-------------------------------------------------------------------------#
# The lm() function (think "linear model") does the work, but it prints
# very little to the screen by default. Better to save the results to
# an object, and then explore that object.

# This simple regression is estimated on ALL countries.

# It includes a constant term by default, but the constant can be
# removed from the specification if you like. The "~" is part of R's
# formula syntax---see ?formula.
#-------------------------------------------------------------------------#

MyRegression <- lm(formula = SavingsPercent ~ GDP_PerWorker, data=DT)
typeof(MyRegression) # What kind of object is this? A list.
names(MyRegression) # What are the names of its components?
MyRegression$coefficients # the beta_hats

# But this is what we want to see. Much more informative output!
summary(MyRegression)

# The summary() function can save its results in a different list that can
   be
# quite helpful:
Summary <- summary(MyRegression)
names(Summary)

#-------------------------------------------------------------------------#
# Use functions to extract information from the MyRegression object
#-------------------------------------------------------------------------#
anova(MyRegression) # Note where s-squared appears
# Alternatively:
SumSquaredResiduals <- deviance(MyRegression)
# Constructing s_squared from the residuals: Note that
#  n-k is stored in df.residual
s_squared <- SumSquaredResiduals/MyRegression$df.residual

# The full variance matrix of the coefficients, which
#  is s_squared * (X'X)^{-1}:
Var_beta_hat <- vcov(MyRegression)
# Display the estimated coefficient standard errors; you can
#  compare these with the output from summary:
sqrt(diag(Var_beta_hat))
```

```
# To prepare for our next OLS regression, use a version of Region that
# is a factor rather than a character variable:
DT[, RegionFactor := as.factor(Region)]
DT[, levels(RegionFactor)]

#-------------------------------------------------------------------------#
# Add the region covariates to the linear model. Because RegionFactor is a
# factor, lm() will select its *first* level ("Africa") to serve as the
# omitted category. Here "+" means to add dummy variables for all
# factor levels *except* the omitted level.
#------------------------------------------------------------------------ -#
MyRegression2 <- lm(SavingsPercent ~ GDP_PerWorker + RegionFactor, data=DT)
```

Chapter 10

Tests of the Regression Model

This chapter will restrict itself—with one brief exception—to the case in which the regression disturbance $\boldsymbol{\epsilon}$ is normally distributed conditional on the explanatory variables. The distributions of test statistics will therefore involve the normal, chi-square, t and $\mathcal{F}$ distributions. If we suspect $\boldsymbol{\epsilon}$ of being non-normally distributed, then to conduct tests on β we will usually need to appeal to asymptotic arguments.

Before we explore the tests in detail, let us re-visit the general problem of constructing a test statistic for parameters θ given an unbiased, normally-distributed estimator

$$\hat{\theta} \sim \mathcal{N}(\theta, \mathbf{V})$$

where $\mathbf{V}$ is the covariance matrix of the estimator in question. We could imagine there being two types of tests for the null hypothesis $H_0 : \theta = \theta_0$ against the alternative $H_A : \theta \neq \theta_0$, based on the quadratic forms

$$T_1 = (\hat{\theta} - \theta_0)'(\hat{\theta} - \theta_0) \quad \text{and} \quad T_2 = (\hat{\theta} - \theta_0)'\mathbf{V}^{-1}(\hat{\theta} - \theta_0)$$

respectively. The T_2 form is generally preferred. This is because it makes sense to weight deviations from the null $\hat{\theta} - \theta_0$ inversely according to their variance, so that the "noisier" and less trustworthy elements of the estimator are down-weighted as evidence against the null.

This approach leads naturally to a chi-square test. Since $\mathbf{V}^{-1}$ is positive definite and (under the null) the random vector in the wings of the quadratic form $\hat{\theta} - \theta_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$, the T_2 test statistic has (under the null) a central chi-square distribution with degrees of freedom equal to the rank of $\mathbf{V}$. (See Chapter 2.) Under the alternative hypothesis that the true value of θ is not the assumed θ_0 but rather $\theta_0 + \delta$ for some $\delta \neq \mathbf{0}$, we would have $\hat{\theta} - \theta_0 \sim \mathcal{N}(\delta, \mathbf{V})$. For a given δ the test statistic T_2 would be distributed as non-central χ^2 with non-centrality parameter $\lambda = (1/2)\delta'\mathbf{V}^{-1}\delta$. We would use this information, together with the test size and rejection region, to calculate the power of the T_2 test statistic as a function of δ. We begin by exploring this approach.

10.1 χ^2 Tests

Consider the regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ in which $\epsilon \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and assume that σ^2 is *somehow already known*. Take the null hypothesis to be $H_0 : \boldsymbol{\beta} = \mathbf{0}$, that is, each of the

elements of the k-vector $\boldsymbol{\beta}$ is hypothesized to be zero. (We will develop a more compelling example in a moment.) Under the null hypothesis,

$$\hat{\boldsymbol{\beta}} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

and clearly any non-zero element of $\hat{\boldsymbol{\beta}}$ provides evidence against the null hypothesis. Conditional on $\mathbf{X}$, the test statistic

$$T = \frac{1}{\sigma^2}\,\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}} \sim \chi^2_k$$

under the null. Given the alternative hypothesis $H_A : \boldsymbol{\beta} \neq \mathbf{0}$, let us focus attention on a particular value $\boldsymbol{\beta}$ for the vector of slope parameters. For this value of the true parameters,

$$\hat{\boldsymbol{\beta}} \mid \mathbf{X} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Since $\boldsymbol{\beta} \neq \mathbf{0}$, conditional on $\mathbf{X}$ the test statistic

$$T = \frac{1}{\sigma^2}\hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\hat{\boldsymbol{\beta}}$$

is distributed as *non-central* χ^2_k with noncentrality parameter

$$\lambda = \frac{1}{2}\frac{1}{\sigma^2}\boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}$$

as explained in Chapter 2. Thus, we see that for a given $\mathbf{X}'\mathbf{X}$, the power of the test statistic is a function of $\boldsymbol{\beta}$. The greater is the departure of $\boldsymbol{\beta}$ from $\mathbf{0}$, the greater is the noncentrality parameter λ; and greater is λ, the further to the right is the distribution of the test statistic shifted (again see Chapter 2). For any given $\boldsymbol{\beta}$ and $\mathbf{X}$ we could calculate λ and use the result to find the power of the test (conditional on $\mathbf{X}$).

We now turn to a more realistic example in which the null hypothesis is that $\delta = \mathbf{0}$ for a model in which $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\delta + \epsilon$. By the FWL Theorem (introduced in the previous chapter), $\hat{\delta}$ is

$$\hat{\delta} = (\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_X\mathbf{Y} = \delta + (\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{M}_X\boldsymbol{\epsilon}$$

and its variance (conditional on $\mathbf{X}$, $\mathbf{Z}$) is

$$\operatorname{Var}\hat{\delta} = \sigma^2(\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1}.$$

Under the null and conditional on $\mathbf{X}$ and $\mathbf{Z}$, we have $\hat{\delta} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{Z}'\mathbf{M}_X\mathbf{Z})^{-1})$. This implies that the test statistic

$$T = \frac{1}{\sigma^2}\hat{\delta}'\mathbf{Z}'\mathbf{M}_X\mathbf{Z}\hat{\delta},$$

is distributed as central $\chi^2_{k_Z}$ under the null. Under the alternative hypothesis, for any true value of the parameter $\delta \neq \mathbf{0}$, the test statistic is distributed as non-central χ^2 with non-centrality parameter

$$\lambda = \frac{1}{2}\frac{1}{\sigma^2}\delta'\mathbf{Z}'\mathbf{M}_X\mathbf{Z}\delta$$

conditional on $\mathbf{X}$ and $\mathbf{Z}$. Again we have all the ingredients needed to derive the power of the test as a function of δ.

Please note that if $\mathbf{Z}$ is a *single* explanatory variable, the quantity $\mathbf{Z}'\mathbf{M}_X\mathbf{Z}$ can be viewed as the sum of squared residuals from a regression of $\mathbf{Z}$ on the other explanatory variables $\mathbf{X}$. If these $\mathbf{X}$ variables do not "predict" $\mathbf{Z}$ very well in a computational sense, so that the sum of squared residuals is large, then the non-centrality parameter will also be large and this will enhance the power of the test. There is, in effect, more "independent" variation in $\mathbf{Z}$ to use in testing propositions about the δ parameter. Or to put it differently, the less collinearity there is between $\mathbf{X}$ and $\mathbf{Z}$, the greater is the power of the test.

The preceding examples are admittedly artificial, in that the so-called test statistic requires knowledge of σ^2, which in reality is an unknown parameter. How then should we proceed to implement testing procedures in a real-world case? As you will remember from Chapter 8, we might try to devise a test statistic in which the unknown parameter conveniently cancels out—this route leads to the t text familiar from that chapter, and also leads to a related set of tests known as $\mathcal{F}$ tests, which we will explore at some length in this chapter.

10.2 The t-test

Recall that the t distribution is derived from the ratio of two independent random variables, a standard normal variable being in the numerator (under the null) and the square root of an independent chi-square variable, divided by its degrees of freedom, being in the denominator.

If $\boldsymbol{\epsilon} \mid \mathbf{X}$ is $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, the least-squares estimator is distributed as

$$\hat{\boldsymbol{\beta}} \mid \mathbf{X} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}).$$

Consider any null hypothesis that can be expressed as a *single linear constraint* on $\boldsymbol{\beta}$,

$$H_0 : \mathbf{R}\boldsymbol{\beta} = r,$$

in which $\mathbf{R}$ is a pre-specified $1 \times k$ row vector of constants and r is a pre-specified scalar. We will work with $\mathbf{R}\hat{\boldsymbol{\beta}} - r$ to derive the numerator of our test statistic. This statistic is a scalar, and if it departs from zero in either direction, it provides evidence against the null. Under the null,

$$\mathbf{R}\hat{\boldsymbol{\beta}} - r \mid \mathbf{X} \sim \mathcal{N}(0, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'),$$

and we can divide by the square root of the variance to obtain a standard normal random variable. The full test statistic is

$$T = \frac{\frac{\mathbf{R}\hat{\boldsymbol{\beta}} - r}{\sqrt{\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\mathbf{R}\hat{\boldsymbol{\beta}} - r}{s\sqrt{\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'}} \sim t_{n-k}$$

conditional on $\mathbf{X}$. We are assured of independence between numerator and denominator because $\hat{\boldsymbol{\beta}}$ is independent of s^2 (conditional on $\mathbf{X}$), and therefore so are functions of $\hat{\boldsymbol{\beta}}$ such as $\mathbf{R}\hat{\boldsymbol{\beta}} - r$.

To develop the *power function* of the t-test, let's narrow our focus to the test that is nearly always presented as part of standard regression output in statistical packages: the null hypothesis $H_0 : \beta = 0$ against the alternative $H_A : \beta \neq 0$ in a model specified in the form $Y = \beta \mathbf{X}_1 + \mathbf{X}_2\beta_2 + \epsilon$. The idea is that β is one of the k slope parameters of the larger model. In regression output, you are typically shown the results for all k hypothesis tests of this kind, with one result for each of the slope parameters.

To begin, note that the analysis of test power for the t-test makes use of the non-central $t_{n-k,\delta}$ distribution, which is produced by the ratio

$$\frac{\mathcal{N}(\delta, 1)}{\sqrt{\chi^2_{n-k}/(n-k)}}$$

provided the numerator and denominator are independent. As we've discussed earlier, the larger is δ, the more the distribution is shifted to the right (when $\delta > 0$) or to the left ($\delta < 0$) relative to the central t_{n-k} distribution.

Now, under the null and conditional on $\mathbf{X}_1, \mathbf{X}_2$, we have $\hat{\beta} \sim \mathcal{N}(0, \sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1})$ and therefore

$$\frac{\hat{\beta}}{\sqrt{\sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}}} \sim \mathcal{N}(0, 1).$$

Under the alternative hypothesis that the true $\beta = d \neq 0$, however,

$$\frac{\hat{\beta}}{\sqrt{\sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}}} \sim \mathcal{N}(\delta, 1)$$

with

$$\delta = \frac{d}{\sqrt{\sigma^2(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{-1}}} = d \cdot \frac{(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{1/2}}{\sigma}.$$

This δ is the non-centrality parameter of the non-central $t_{n-k,\delta}$ distribution.

We can rewrite the non-centrality parameter in a way that helps to clarify the roles of the size of the sample and other factors in test power:

$$\delta = d \cdot \sqrt{n} \cdot \frac{(\frac{1}{n}\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1)^{1/2}}{\sigma} \equiv d \cdot \sqrt{n} \cdot \frac{\bar{q}^{1/2}}{\sigma}.$$

In this way we see that the non-centrality parameter is larger when d, the difference between the true β and the hypothesized value $\beta = 0$, is larger. It is also larger in larger samples, in samples in which $\bar{q}$, which measures the association between $\mathbf{X}_1$ and the remaining covariates $\mathbf{X}_2$, indicates that there is more "independent" variation in $\mathbf{X}_1$ net of the other covariates, and it is larger when there is less underlying "noisiness" stemming from the disturbance term variance. All these factors increase the power of the t-test.[1]

[1] The sample size n also affects the degrees of freedom $n - k$ of the test statistic—see Chapter 2 for illustrations.

10.3 $\mathcal{F}$ Tests of Linear Hypotheses

Consider now a more general hypothesis about $\boldsymbol{\beta}$, positing that $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$, where the matrix $\mathbf{R}$ is $m \times k$ and the vector $\mathbf{r}$ is $m \times 1$. $\mathbf{R}$ should have a number of rows $m \leq k$; otherwise the constraint $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ contains redundant restrictions on β. Various versions of $\mathbf{R}$ and $\mathbf{r}$ yield different tests on $\boldsymbol{\beta}$, such as the test of equality between coefficients $\beta_j = \beta_l$, or the test that all $\boldsymbol{\beta}$'s other than the constant term are equal to zero, $\beta_2 = \beta_3 = \cdots = \beta_k = 0$. All these can be expressed as varieties of $\mathcal{F}$ tests, as we will now show.

For a particular $\mathbf{R}$ and $\mathbf{r}$, consider the transformation of the OLS estimator $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}$. Assuming normality and conditional on $\mathbf{X}$,

$$\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'),$$

under the null hypothesis that $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. Examine the quadratic form centered around the inverse of the $m \times m$ covariance matrix $\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$,

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}). \tag{10.1}$$

Under the null, this is distributed as central χ^2_m and, if σ^2 were known, tests of the null could be conducted using this distribution. The fact that σ^2 is unknown forces us to adopt a different approach, whereby this parameter is made to cancel out of the test statistic.

As we've just discussed, conditional on $\mathbf{X}$ (and assuming normality), the OLS estimator $\hat{\boldsymbol{\beta}}$ is independent of s^2, and since the expression in equation (10.1) is just a function of $\hat{\boldsymbol{\beta}}$, it too is independent of s^2. Furthermore, s^2/σ^2 is a χ^2_{n-k} variable divided by its degrees of freedom (again conditional on $\mathbf{X}$ and assuming normality). Hence, if we were to divide equation (10.1) by s^2/σ^2 and then divide the result by m, the numerator degrees of freedom, the unknown σ^2 would cancel out and we would have a test statistic that is distributed according to the $\mathcal{F}(m, n-k)$ distribution. (You will recall that the $\mathcal{F}$ distribution is produced by the ratio of two independent chi-square variables, each being divided by its degrees of freedom.) Just as with the t-tests, this result also holds unconditionally. With a given $\mathbf{R}$ and $\mathbf{r}$, one constructs the actual test statistic using

$$T = (\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[s^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})/m. \tag{10.2}$$

Looking back to the χ^2 statistic of equation (10.1), we see that the $\mathcal{F}$ statistic closely resembles it. In essence, s^2 has replaced the unknown σ^2 and m has been inserted in the numerator. All regression software packages supply s^2 when you run a regression, and if the software also allows you to manipulate the matrices of equation (10.1), then you can form the $\mathcal{F}$ test statistic yourself without difficulty.

10.4 Implementing $\mathcal{F}$ tests in R

```
#------------------------------------------------------------------------#
# Testing hypotheses: The F-test
#------------------------------------------------------------------------#
# Using MyRegression2 (see the previous chapter), test that the five
```

```
# region parameters are all equal to zero, using the R*beta=0 form for
# the null.
#-----------------------------------------------------------------------#

# We use the R matrix to implement the null hypothesis. The number of rows
# equals the number of distinct hypotheses to be tested, and there are k
# columns, one for each coefficient:

R <- matrix(data = 0, nrow=5, ncol=7)
for (i in 1:nrow(R)) {R[i,2+i] <- 1} # print R to see the result

W <- R %*% MyRegression2$coefficients # Note the dimensions of W

# Derive W's variance matrix (the t() function means "transpose"), which
# includes s_squared:
V <- R %*% vcov(MyRegression2) %*% t(R)
V_inv <- solve(V) # derive the inverse of V. See ?solve.

# The value of the F-statistic is:
Fstat <- (1.0/nrow(R)) * t(W) %*% V_inv %*% W

#-----------------------------
# The p-value of the F-test:
#-----------------------------
# This is the probability of having a test statistic value *as high or
    higher*
# than Fstat *if* the null is true. Here nrow(R) is the numerator degrees of
# freedom, and recall that df.residual is n-k, the denominator degrees of
# freedom:
pf(Fstat,
   nrow(R), MyRegression2$df.residual,
   lower.tail=FALSE)
```

The *power function* of the $\mathcal{F}$-test is derived with essentially the same approach as for the t-test. It relies on the non-central $\mathcal{F}_{m,n-k,\lambda}$ distribution, which is produced by the ratio

$$\frac{\chi^2_{m,\lambda}/m}{\chi^2_{n-k}/(n-k)}$$

provided the numerator and denominator are independent. The larger is λ, the more this non-central distribution is shifted to the right.

Recall that under the null hypothesis and conditional on $\mathbf{X}$, the quadratic form

$$(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}).$$

is distributed as central χ^2_m. But if $\mathbf{R}\boldsymbol{\beta} \neq \mathbf{r}$, that is, if the null hypothesis is false, then we can write $\mathbf{R}\boldsymbol{\beta} = \mathbf{r} + \mathbf{d}$, with $\mathbf{d} \neq \mathbf{0}$ indexing the extent to which the null is false in each of the m dimensions being simultaneously tested. In this case, $\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r} \sim \mathcal{N}(\mathbf{d}, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')$ conditional on $\mathbf{X}$. The non-centrality parameter of the $\chi^2_{m,\lambda}$ distribution is then

$$\lambda = (1/2) \cdot \mathbf{d}'(\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1}\mathbf{d}.$$

To draw out the implications of larger samples and greater disturbance term variances for the power function, this can be re-expressed in the form

$$\lambda = (1/2) \cdot n \cdot \mathbf{d}' \left(\mathbf{R}(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}' \right)^{-1} \mathbf{d} \cdot \frac{1}{\sigma^2}$$

much as we did with the non-centrality parameter of the t-test

10.5 $\mathcal{F}$ tests revisited

There is a longer route leading to the $\mathcal{F}$ test statistic, whose usefulness we will understand in a moment.[2] In the old days, before the arrival of statistical packages that allowed easy manipulation of matrices, it was customary to use an ingenious two-step method to perform $\mathcal{F}$ tests. The model would first be estimated with the restriction $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ imposed on the coefficients, and the constrained error sum of squares $\mathbf{e}_c'\mathbf{e}_c$ retained. Next it would be estimated without these restrictions and the unconstrained sum of squares $\mathbf{e}_u'\mathbf{e}_u$ retained. Given the normal disturbances assumption and conditional on $\mathbf{X}$, under the null $(\mathbf{e}_c'\mathbf{e}_c - \mathbf{e}_u'\mathbf{e}_u)/\sigma^2$ can be shown to be distributed as a central chi-square with degrees of freedom equal to m, the number of rows of $\mathbf{R}$. This difference in sums of squares provides us with the numerator of an $\mathcal{F}$ test statistic. As for the denominator, as we saw just above, $\mathbf{e}_u'\mathbf{e}_u$ is its key component.

Let $S(\tilde{\boldsymbol{\beta}})$ be the minimized sum of squared residuals $(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{Y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) = \mathbf{e}_c'\mathbf{e}_c$ using the restricted estimator $\tilde{\boldsymbol{\beta}}$ that satisfies $\mathbf{R}\tilde{\boldsymbol{\beta}} = \mathbf{r}$ and let $S(\hat{\boldsymbol{\beta}})$ be the minimized sum $\mathbf{e}_u'\mathbf{e}_u$ using $\hat{\boldsymbol{\beta}}$, the unrestricted ordinary least-squares estimator. Obviously, $S(\tilde{\boldsymbol{\beta}}) \geq S(\hat{\boldsymbol{\beta}})$. We can relate the two quantities via a Taylor series expansion

$$\begin{aligned} S(\tilde{\boldsymbol{\beta}}) &= S(\hat{\boldsymbol{\beta}}) + \frac{\partial S(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}}' (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) + \frac{1}{2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})' \frac{\partial^2 S(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \\ &= S(\hat{\boldsymbol{\beta}}) + \frac{1}{2}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})' \frac{\partial^2 S(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \end{aligned}$$

in which $\boldsymbol{\beta}^*$ lies between $\tilde{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}$. The vector of first derivatives equals zero by the ordinary least squares first-order condition. Since $\partial^2 S(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}' = 2\mathbf{X}'\mathbf{X}$, we can write the above as

$$\mathbf{e}_c'\mathbf{e}_c = \mathbf{e}_u'\mathbf{e}_u + (\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})'\mathbf{X}'\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) \tag{10.3}$$

where in place of $S(\tilde{\boldsymbol{\beta}})$ we write $\mathbf{e}_c'\mathbf{e}_c$ for the constrained sum of squares and let $S(\hat{\boldsymbol{\beta}}) = \mathbf{e}_u'\mathbf{e}_u$ be the unconstrained sum of squares. Note that the second expression on the right is a positive definite quadratic form, which confirms that the constrained sum of squares can be no smaller than $\mathbf{e}_u'\mathbf{e}_u$, the unconstrained sum of squares.

To work out an expression for $\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}$, we must explore the features of $\tilde{\boldsymbol{\beta}}$, the constrained estimator. The problem of estimating $\boldsymbol{\beta}$ subject to the constraint $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ can be set up as a

[2] As we will see later in the econometrics sequence, there is a close connection in theory between this two-step approach to motivating the $\mathcal{F}$ test and the construction of likelihood ratio tests for models estimated by the method of maximum likelihood.

Lagrangian,

$$\frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})+\lambda'(\mathbf{R}\boldsymbol{\beta}-\mathbf{r}),$$

where there are m multipliers in the λ vector corresponding to the m rows of $\mathbf{R}$. (A factor of $1/2$ has been introduced to simplify the derivation.) Differentiating with respect to $\boldsymbol{\beta}$ and λ gives

$$\begin{aligned}(\mathbf{X}'\mathbf{X})\tilde{\boldsymbol{\beta}}-\mathbf{X}'\mathbf{Y}+\mathbf{R}'\tilde{\lambda} &= \mathbf{0}_{k\times 1}\\ \mathbf{R}\tilde{\boldsymbol{\beta}}-\mathbf{r} &= \mathbf{0}_{m\times 1}.\end{aligned}$$

where $\tilde{\lambda}$ is the estimated Lagrange multiplier. We now multiply the first equation by the quantity $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$ and use $\mathbf{R}\tilde{\boldsymbol{\beta}}=\mathbf{r}$ to obtain the m multipliers

$$\tilde{\lambda}=\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r}), \tag{10.4}$$

in which $\hat{\boldsymbol{\beta}}$, that is, the ordinary, unconstrained estimator of β, appears. Under the null hypothesis, $\tilde{\lambda}$ is distributed as $\mathcal{N}(\mathbf{0},\sigma^2(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}')^{-1})$ conditional on $\mathbf{X}$, as can be shown by substituting for $\hat{\beta}$ in the expression above. One could test the constraint $\mathbf{R}\boldsymbol{\beta}=\mathbf{r}$ by testing whether the m multipliers $\lambda=\mathbf{0}$.

This approach is known as a Lagrange multiplier test. As we will see later in our course, the notion of testing constraints via Lagrange multiplier tests is a very important theme in econometrics. Were we to form the test statistic directly from the Lagrange multipliers, we would do so using

$$\tilde{\lambda}'\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\tilde{\lambda}/\sigma^2.$$

This quadratic form is distributed as χ^2_m under the null hypothesis. The test for $\mathbf{R}\boldsymbol{\beta}=\mathbf{r}$ is seldom formulated in this way, however, since few regression programs provide the required $\tilde{\lambda}$. Moreover, the variance parameter σ^2 is unknown. Therefore, we must proceed in a different fashion.

Substitution of the expression for $\tilde{\lambda}$ into the first-order conditions for the constrained estimator $\tilde{\boldsymbol{\beta}}$ shows how the constrained and unconstrained estimators are related,

$$\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}-(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r}).$$

Now return to equation (10.3) above, which expressed the constrained sum of squares $\mathbf{e}_c'\mathbf{e}_c$ as a quantity equal to the unconstrained sum of squares $\mathbf{e}_u'\mathbf{e}_u$ plus a quadratic form. Substituting the expression just derived into that quadratic form, we find

$$\mathbf{e}_c'\mathbf{e}_c-\mathbf{e}_u'\mathbf{e}_u=(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})'(\mathbf{X}'\mathbf{X})(\tilde{\boldsymbol{\beta}}-\hat{\boldsymbol{\beta}})=(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r}),$$

which is *identical* to (10.1) except for the fact that σ^2 is absent. Inserting σ^2, we obtain

$$\frac{\mathbf{e}_c'\mathbf{e}_c-\mathbf{e}_u'\mathbf{e}_u}{\sigma^2}=(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r})'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}}-\mathbf{r}).$$

We find that conditional on $\mathbf{X}$,

$$\frac{\mathbf{e}_c'\mathbf{e}_c-\mathbf{e}_u'\mathbf{e}_u}{\sigma^2}\sim\chi^2_m,$$

which is the key result we need to form the numerator of the $\mathcal{F}$ statistic. Combining this with the fact that (conditionally)

$$\frac{(n-k)\cdot s^2}{\sigma^2} = \frac{\mathbf{e}_u'\mathbf{e}_u}{\sigma^2} \sim \chi^2_{n-k},$$

and given the independence of numerator and denominator, which we have already established, it follows that the ratio

$$T = \frac{(\mathbf{e}_c'\mathbf{e}_c - \mathbf{e}_u'\mathbf{e}_u)/m}{\mathbf{e}_u'\mathbf{e}_u/(n-k)} \sim \mathcal{F}(m, n-k).$$

Note how easy it is to implement the $\mathcal{F}$ test: one runs the model with the constraint $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ imposed and obtains $\mathbf{e}_c'\mathbf{e}_c$; then one runs it without the constraint and obtains $\mathbf{e}_u'\mathbf{e}_u$. From here, a simple manipulation yields the $\mathcal{F}$ statistic needed to test the hypothesis $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$.

Robustness of $\mathcal{F}$-tests to non-normality

In the discussion above, we emphasized the benefits derived from the ratio form of the $\mathcal{F}$ statistic, which causes σ^2 to cancel out. There is a further benefit to this ratio form: it offers some protection against the possibility of non-normal regression disturbances.

Suppose that the true regression disturbance is $\boldsymbol{\epsilon}^* = z\boldsymbol{\epsilon}$, where z is a scalar random variable and $\boldsymbol{\epsilon}$ is a vector of normally-distributed random variables, independent of $\mathbf{X}$ and with variance matrix $\sigma^2\mathbf{I}$. The distribution of $\boldsymbol{\epsilon}^*$ will not be normal, in general, and in fact a variety of distributions for $\boldsymbol{\epsilon}^*$ can be generated by varying the distribution of the scalar random variable z. Nevertheless, the $\mathcal{F}$ test remains perfectly valid (Zaman (1996, Chapter 8)).

To see this, recall that the ratio forming the $\mathcal{F}$ test statistic is

$$T = \frac{\boldsymbol{\epsilon}^{*\prime}\mathbf{P}\boldsymbol{\epsilon}^*/m}{\boldsymbol{\epsilon}^{*\prime}\mathbf{M}_X\boldsymbol{\epsilon}^*/(n-k)}.$$

from which the scalar z^2 cancels out, leaving behind a ratio of quadratic forms in normal random variables that is distributed according to the $\mathcal{F}$ distribution.

10.6 Testing for Structural Change

The $\mathcal{F}$ test is often used to detect structural changes or differences in economic relationships. In this context, it is known as a Chow test after the Princeton econometrician Gregory Chow. The test is applied to cases in which

$$\begin{aligned} Y_{1t} &= \mathbf{X}_{1t}\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_{1t} \qquad t = 1, ..., \tau \\ Y_{2t} &= \mathbf{X}_{2t}\boldsymbol{\beta}_2 + \epsilon_{2t} \qquad t = \tau+1, ..., T \end{aligned}$$

Here τ can represent the point of structural change in the relationship between $\mathbf{Y}$ and $\mathbf{X}$, such that $\boldsymbol{\beta}_1$ is in effect for $t \leq \tau$, and $\boldsymbol{\beta}_2$ for $t > \tau$. In a time-series sample, τ indicates the point in time at which the relationship shifts. In a cross-section sample, it may indicate

a point of division in the sample when it is ordered according to an explanatory variable such as sex or urban/rural residence. For example, it is common practice to test for gender wage discrimination in the labor market by testing $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, where $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)'$ represents wages for men and women respectively.

How does one proceed to test the null hypothesis of no structural change across regimes, that is, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$? As we will demonstrate, an $\mathcal{F}$ test can be applied if $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ have the same variance. Unfortunately, if the variances differ then no satisfactory small-sample test is available, see Amemiya (1985, pp. 35–38) for discussion. We will highlight the point of difficulty below.

The equal-variances case is straightforward. If we stack the equations then the full unrestricted model can be written as

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{bmatrix}$$

or, in a more compact form, as

$$\mathbf{Y}_{T\times 1} = \mathbf{X}_{T\times 2k}\, \boldsymbol{\beta}_{2k\times 1} + \boldsymbol{\epsilon}_{T\times 1}$$

with the dimensions indicated in the subscripts. Under the assumption of equal variances, $\operatorname{Var} \boldsymbol{\epsilon} = \sigma^2 \mathbf{I}_T$.

Now, in principle, one could estimate the unrestricted model in the stacked form presented above. But since $\mathbf{X}$ is block-diagonal, it is simpler (yielding results that are numerically identical) to estimate $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ via two separate regressions of $\mathbf{Y}_1$ on $\mathbf{X}_1$ and $\mathbf{Y}_2$ on $\mathbf{X}_2$. The unrestricted error sum of squares $\mathbf{e}_u'\mathbf{e}_u$ should then be calculated over the full sample from $t = 1, \ldots, T$. As usual, $\mathbf{e}_u'\mathbf{e}_u/\sigma^2$ is distributed as a chi-square with $T - 2k$ degrees of freedom (but note the $2k$ rather than the usual k).

The null hypothesis of equality in coefficients, that is, $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$, can be expressed as a linear restriction $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ of the form

$$[\mathbf{I}_k, -\mathbf{I}_k] \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \mathbf{0}_k.$$

The restricted model is most easily estimated by changing the way in which the data are stacked,

$$\begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \beta + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{bmatrix},$$

where the $\mathbf{X}$ matrix is now $T \times k$ rather than $T \times 2k$. Estimating this model supplies us with the restricted sum of squares $\mathbf{e}_c'\mathbf{e}_c$.

As before, $(\mathbf{e}_c'\mathbf{e}_c - \mathbf{e}_u'\mathbf{e}_u)/\sigma^2$ is distributed as χ_k^2 and is independent of $\mathbf{e}_u'\mathbf{e}_u/\sigma^2$, which itself is distributed χ_{T-2k}^2. When each χ^2 variable is divided by its degrees of freedom and the ratio formed, the σ^2 cancels and the result is

$$T = \frac{(\mathbf{e}_c'\mathbf{e}_c - \mathbf{e}_u'\mathbf{e}_u)/k}{\mathbf{e}_u'\mathbf{e}_u/(T-2k)} \sim \mathcal{F}(k, T-2k).$$

10.7 Testing Whether Variances Are Equal

If we attempt to extend the analysis above to the case of unequal variances σ_1^2 and σ_2^2, the difficulty we encounter is that the error sums of squares cannot be normalized by a single factor σ^2 that will then cancel in the appropriate $\mathcal{F}$ ratio. It makes sense, therefore, to test for the equality of variances *before* we undertake a test for $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. If the data do not reject the hypothesis $\sigma_1^2 = \sigma_2^2$, we can proceed to the $\mathcal{F}$ test for $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2$. If the data reject equality of variances, however, we have little alternative but to consider asymptotic tests.

Suppose that there are T_1 observations in the first part of the sample governed by the relation

$$\mathbf{Y}_1 = \mathbf{X}_1\boldsymbol{\beta}_1 + \boldsymbol{\epsilon}_1, \quad \text{Var}\,\boldsymbol{\epsilon}_1 = \sigma_1^2\mathbf{I}_1,$$

and T_2 observations for the part of the sample in which

$$\mathbf{Y}_2 = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon}_2, \quad \text{Var}\,\boldsymbol{\epsilon}_2 = \sigma_2^2\mathbf{I}_2.$$

As we have throughout this chapter, we maintain the assumption that $\boldsymbol{\epsilon}_t$ is independent of $\epsilon_{t'}$ for any $t \neq t'$, and this implies that the vectors $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are independent.

Estimate $\boldsymbol{\beta}_1$ from the T_1 observations of the first regime and $\boldsymbol{\beta}_2$ from the T_2 observations of the second regime. Form the error sums of squares $\mathbf{e}_1'\mathbf{e}_1$ and $\mathbf{e}_2'\mathbf{e}_2$ for the regimes 1 and 2 respectively. Since $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are normal, it follows that

$$\frac{\mathbf{e}_1'\mathbf{e}_1}{\sigma_1^2} = \frac{\boldsymbol{\epsilon}_1'\mathbf{M}_1\boldsymbol{\epsilon}_1}{\sigma_1^2} \sim \chi^2_{T_1-k}$$

conditional on $\mathbf{X}$, and likewise

$$\frac{\mathbf{e}_2'\mathbf{e}_2}{\sigma_2^2} = \frac{\boldsymbol{\epsilon}_2'\mathbf{M}_2\boldsymbol{\epsilon}_2}{\sigma_2^2} \sim \chi^2_{T_2-k}.$$

The two chi-square variables are independent due to the assumed independence of $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$, and under the null hypothesis $\sigma_1^2 = \sigma_2^2$. Therefore, conditional on $\mathbf{X}$,

$$T = \frac{\mathbf{e}_1'\mathbf{e}_1/(T_1-k)}{\mathbf{e}_2'\mathbf{e}_2/(T_2-k)} \sim \mathcal{F}(T_1-k, T_2-k)$$

under the null. The result also holds unconditionally. See Amemiya (1985, p. 35) for further discussion of this test.

The following R program illustrates how to perform the test using the same set of data examined earlier in this and the previous chapter.

```
#-----------------------------------------------------------------------------#
# Let's try another test, this one focused on the values of the disturbance
# term variances in Asian and African countries.
#-----------------------------------------------------------------------------#
# The null hypothesis is that the variances are the same in these two
# regions. The alternative is that they differ. We form an F-statistic
# using the ratio of s_squared for Asia to s_squared for Africa.
#-----------------------------------------------------------------------------#
```

```
AfricaRegression <- lm(formula = SavingsPercent ~ GDP_PerWorker,
                       data=subset(DT, subset=Region=="Africa") )
summary(AfricaRegression)
# There are 1335 observations for Africa and 2 covariates, so the degrees of
# freedom is 1333:
AfricaRegression$df.residual
 # Constructing s_squared from the residuals:
Africa_s_squared <- deviance(AfricaRegression)/AfricaRegression$df.residual

AsiaRegression <- lm(formula = SavingsPercent ~ GDP_PerWorker,
                       data=subset(DT, subset=Region=="Asia") )
summary(AsiaRegression)
# There are 1172 observations for Asia and 2 covariates, so the degrees of
# freedom is 1170:
AsiaRegression$df.residual
Asia_s_squared <- deviance(AsiaRegression)/AsiaRegression$df.residual

Fstat = Asia_s_squared/Africa_s_squared

#-------------------------------------------------
# The p-value of the F-statistic for Test No. 2:
#-------------------------------------------------
# Our Fstat is 1.0628. What is the probability of having an Fstat this
#  high or higher if the null hypothesis of equal variances is true?
pf(Fstat,
   AsiaRegression$df.residual, AfricanRegression$df.residual,
   lower.tail=FALSE)
# We can't reject the null hypothesis of equal variances in Asia
# and Africa at any conventional significance level
```

Chapter 11

Least Squares and Projections

The lecture notes for Economics 590 present the basics of projection theory. In its current version, this chapter begins with material already found in the Eco 590 notes, but the next version of the chapter will eliminate the repetition.

11.1 Definitions and Basic Results

Let $\mathbf{Y}$ be an $n \times 1$ vector and let $\mathbf{X}$ be a matrix composed of k vectors of the same length. We will assume that $k < n$. Define $\mathcal{S}(\mathbf{X})$ to be the set of all linear combinations of the column vectors in $\mathbf{X}$, that is, the set of vectors $\mathbf{z}$ that can be generated via the relation $\mathbf{z} = \mathbf{X}\alpha$ for some $k \times 1$ vector α.

The *orthogonal projection* of $\mathbf{Y}$ onto $\mathcal{S}(\mathbf{X})$ is the $n \times 1$ vector $\hat{\mathbf{Y}}$ in $\mathcal{S}(\mathbf{X})$ that is closest to $\mathbf{Y}$. The "closeness" of two vectors $\mathbf{Y}$ and $\tilde{\mathbf{Y}}$ is defined in terms of the conventional (squared) distance metric

$$\|\mathbf{Y} - \tilde{\mathbf{Y}}\|^2 = (\mathbf{Y} - \tilde{\mathbf{Y}})'(\mathbf{Y} - \tilde{\mathbf{Y}}).$$

The projection $\hat{\mathbf{Y}}$ is the vector that minimizes this quantity among all vectors $\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$. It can be shown that $\hat{\mathbf{Y}}$ exists and is unique; see Luenberger (1969, pp. 49–52) and Ruud (2000, pp. 38–40).

Because $\hat{\mathbf{Y}}$ can be expressed as a linear combination of the $\mathbf{X}$ vectors, and among all such combinations, $\hat{\mathbf{Y}}$ is closest to $\mathbf{Y}$, the projection error vector $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ must be orthogonal to all vectors in $\mathcal{S}(\mathbf{X})$. That is, for any column vector $\mathbf{X}_i$ in $\mathbf{X}$, we have $\mathbf{X}_i'\mathbf{e} = 0$, and the same relation holds for any vector constructed from a linear combination of the vectors in $\mathbf{X}$. The proof of the orthogonality result, which I take from Ruud (2000), is instructive. (See Zaman (1996, Chapter 1) for further discussion.) Let $\tilde{\mathbf{Y}}$ be any vector in $\mathcal{S}(\mathbf{X})$. First, let's show that $\tilde{\mathbf{Y}}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \tilde{\mathbf{Y}}'\mathbf{e} = 0$ for all $\tilde{\mathbf{Y}}$ is a *sufficient* condition for $\hat{\mathbf{Y}}$ to be the distance-minimizing vector. That is, we will show that for $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, when $\tilde{\mathbf{Y}}'\mathbf{e} = 0$ for any $\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$, this implies that $\hat{\mathbf{Y}}$ must be the distance-minimizing vector. Given orthogonality, we have

$$\tilde{\mathbf{Y}}'(\mathbf{Y} - \hat{\mathbf{Y}}) = 0. \tag{11.1}$$

Also, because $\tilde{\mathbf{Y}} - \hat{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$ since the difference of two linear combinations of the $\mathbf{X}$ vectors is another linear combination of these vectors, we have

$$(\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = 0. \tag{11.2}$$

Now, $\mathbf{Y} - \tilde{\mathbf{Y}} = (\mathbf{Y} - \hat{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \tilde{\mathbf{Y}})$, and we have

$$\begin{aligned}(\mathbf{Y} - \tilde{\mathbf{Y}})'(\mathbf{Y} - \tilde{\mathbf{Y}}) &= (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) \\ &\quad + (\mathbf{Y} - \hat{\mathbf{Y}})'(\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \tilde{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) \\ &\quad + (\hat{\mathbf{Y}} - \tilde{\mathbf{Y}})'(\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}).\end{aligned} \tag{11.3}$$

By equation (11.2), we see that the two middle terms of equation (11.3) are zero. Hence,

$$(\mathbf{Y} - \tilde{\mathbf{Y}})'(\mathbf{Y} - \tilde{\mathbf{Y}}) = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \tilde{\mathbf{Y}})'(\hat{\mathbf{Y}} - \tilde{\mathbf{Y}}).$$

Clearly for any $\tilde{\mathbf{Y}} \neq \hat{\mathbf{Y}}$, the distance from $\mathbf{Y}$ to $\tilde{\mathbf{Y}}$ must exceed the distance from $\mathbf{Y}$ to $\hat{\mathbf{Y}}$.

Now let's complete the if-and-only-if proof by showing that $\tilde{\mathbf{Y}}'\mathbf{e} = 0$ is a *necessary* condition for $\hat{\mathbf{Y}}$ to be the distance-minimizing vector. That is, we show that when $\hat{\mathbf{Y}}$ is distance-minimizing, this implies $\tilde{\mathbf{Y}}'\mathbf{e} = 0$ for all $\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$. We'll develop a proof by contradiction. Suppose that there is a $\mathbf{Z}$ vector in $\mathcal{S}(\mathbf{X})$ such that $\mathbf{Z}'\mathbf{e} \neq 0$. Let

$$\tilde{\mathbf{Y}} = \hat{\mathbf{Y}} + \frac{\mathbf{Z}'\mathbf{e}}{\mathbf{Z}'\mathbf{Z}}\mathbf{Z}$$

and note that $\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$. Now

$$\begin{aligned}(\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})'(\mathbf{Y} - \tilde{\mathbf{Y}}) &= \left(\frac{\mathbf{Z}'\mathbf{e}}{\mathbf{Z}'\mathbf{Z}}\mathbf{Z}\right)'\left(\mathbf{Y} - \hat{\mathbf{Y}} - \frac{\mathbf{Z}'\mathbf{e}}{\mathbf{Z}'\mathbf{Z}}\mathbf{Z}\right) \\ &= \left(\frac{\mathbf{Z}'\mathbf{e}}{\mathbf{Z}'\mathbf{Z}}\mathbf{Z}\right)'\left(\mathbf{e} - \frac{\mathbf{Z}'\mathbf{e}}{\mathbf{Z}'\mathbf{Z}}\mathbf{Z}\right) \\ &= \frac{(\mathbf{Z}'\mathbf{e})^2}{\mathbf{Z}'\mathbf{Z}} - \frac{(\mathbf{Z}'\mathbf{e})^2}{\mathbf{Z}'\mathbf{Z}} = 0.\end{aligned} \tag{11.4}$$

Consider the decomposition

$$\mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{Y} - \tilde{\mathbf{Y}}) + (\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})$$

from which

$$(\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) = (\mathbf{Y} - \tilde{\mathbf{Y}})'(\mathbf{Y} - \tilde{\mathbf{Y}}) + (\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})'(\tilde{\mathbf{Y}} - \hat{\mathbf{Y}}) + 2 \cdot (\tilde{\mathbf{Y}} - \hat{\mathbf{Y}})'(\mathbf{Y} - \tilde{\mathbf{Y}}).$$

According to equation (11.4), the last term is zero. But with the last term equal to zero, this would imply that $\hat{\mathbf{Y}}$ is further from $\mathbf{Y}$ than is $\tilde{\mathbf{Y}}$. That cannot be the case, because we have assumed $\hat{\mathbf{Y}}$ to be distance-minimizing for this part of the proof. In view of the contradiction, we conclude that $\tilde{\mathbf{Y}}'\mathbf{e} = 0$ for all $\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$.

To this point, we have not required the collection of vectors in $\mathbf{X}$ to be linearly independent. Even if there were to exist linear dependencies among the columns of $\mathbf{X}$, the projection $\hat{\mathbf{Y}}$ would be well-defined and the projection error $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ would be orthogonal to all $\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$. Linear dependencies *do* matter, however, if we hope to express $\hat{\mathbf{Y}}$ as a *unique* linear combination of the columns of $\mathbf{X}$. Since $\hat{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$, it must be that $\hat{\mathbf{Y}} = \mathbf{X}\beta$ for *some* β—the question is whether there is a unique β vector.

If the k columns of $\mathbf{X}$ are linearly independent, we can derive an explicit form for the projection $\hat{\mathbf{Y}}$ in which the β coefficients are unique. We seek the k-vector $\hat{\beta}$ with the property

that $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta}$ is orthogonal to each column of $\mathbf{X}$. Solving the k orthogonality relations $\mathbf{X}'\mathbf{e} = \mathbf{X}'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{0}$ for $\hat{\beta}$ yields the familiar formula $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. From this we obtain the projection vector itself, $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. It is no more than the fitted value from a regression of $\mathbf{Y}$ on $\mathbf{X}$.

Let $\mathcal{S}^{\perp}(\mathbf{X})$ be the set of n-vectors that are orthogonal to all vectors in $\mathcal{S}(\mathbf{X})$. By the argument above, $\mathbf{e} \in \mathcal{S}^{\perp}(\mathbf{X})$. Another way to view $\mathbf{e}$ is as the projection of $\mathbf{Y}$ onto $\mathcal{S}^{\perp}(\mathbf{X})$; in this perspective it is $\mathbf{e}$ which is the fitted value. The original vector $\mathbf{Y}$ can therefore be decomposed into two projection vectors: $\hat{\mathbf{Y}}$ is the projection of $\mathbf{Y}$ onto $\mathcal{S}(\mathbf{X})$ and $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ is the projection of $\mathbf{Y}$ onto $\mathcal{S}^{\perp}(\mathbf{X})$.

Two matrices implement these projections. Let $\mathbf{P}_{\mathbf{X}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ and let $\mathbf{M}_{\mathbf{X}} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. When they operate on $\mathbf{Y}$, the matrices $\mathbf{P}_{\mathbf{X}}$ and $\mathbf{M}_{\mathbf{X}}$ project it onto $\mathcal{S}(\mathbf{X})$ and $\mathcal{S}^{\perp}(\mathbf{X})$ respectively. On occasion it will prove useful to write $\mathbf{Y}$ as the sum of the two projections, $\mathbf{Y} = \mathbf{P}_{\mathbf{X}}\mathbf{Y} + \mathbf{M}_{\mathbf{X}}\mathbf{Y}$.

Nonsingular Transformations of X

Suppose that $\mathbf{X}$, a $n \times k$ matrix, is post-multiplied by a nonsingular $k \times k$ matrix $\mathbf{A}$. If we define $\mathbf{Z} = \mathbf{X}\mathbf{A}$, then it is easy to show that $\mathbf{P}_{\mathbf{Z}} = \mathbf{P}_{\mathbf{X}}$. The transformation to $\mathbf{Z}$ leaves $\hat{\mathbf{Y}}$ unaffected and thus leaves the error vector $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$ unaffected as well. The result is due to the fact that $\mathcal{S}(\mathbf{X}) = \mathcal{S}(\mathbf{Z})$. In the regression context, if we happen to be concerned only with the fitted values $\hat{\mathbf{Y}}$ and residuals $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$, we can employ any nonsingular transform of the $\mathbf{X}$ matrix that we find to be convenient.

Another way to view $\mathbf{P}_{\mathbf{X}}$ is as the matrix that extracts all information relevant to $\hat{\mathbf{Y}}$ from the collection of vectors in $\mathbf{X}$. In other words, the only aspects of $\mathbf{X}$ that matter are those that are expressed through $\mathbf{P}_{\mathbf{X}}$. Evidently, these key features are left unaffected by nonsingular transformations.

Linearity

It is worth noting here one additional property of projections: they are linear. In other words, if $\mathbf{Y} = \mathbf{z} + \mathbf{w}$, then $\hat{\mathbf{Y}} = \hat{\mathbf{z}} + \hat{\mathbf{w}}$. As an example of this, consider the regression model $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$, where β is the true parameter vector and $\boldsymbol{\epsilon}$ is the vector of stochastic disturbances. Consider the left-hand side of the regression equation. As we have already seen, projecting $\mathbf{Y}$ on $\mathbf{X}$ yields $\hat{\mathbf{Y}} = \mathbf{X}\hat{\beta}$. Now, turning to the right-hand side, we see that the projection of $\mathbf{X}\beta$ onto $\mathbf{X}$ has no effect at all, yielding simply $\mathbf{X}\beta$. The projection of $\boldsymbol{\epsilon}$ on $\mathbf{X}$ yields $\hat{\boldsymbol{\epsilon}}$. Taking them together, we have

$$\hat{\mathbf{Y}} = \mathbf{X}\beta + \hat{\boldsymbol{\epsilon}}.$$

In the simple regression model, we assume that ϵ is statistically unrelated to $\mathbf{X}$, and in particular, $E(\epsilon \mid \mathbf{X}) = \mathbf{0}$. If this is true, then $\hat{\epsilon}$ can be no more than sampling error, a vector with expectation zero. In any given sample of data, however, ϵ need not be orthogonal to $\mathbf{X}$. Note that the sampling error $\hat{\epsilon} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$ by the projection formula. Using this, we obtain

$$\hat{\mathbf{Y}} = \mathbf{X}\left(\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\right) = \mathbf{X}\hat{\beta},$$

which shows that the sampling error is embedded in the estimate of $\hat{\beta}$ for that sample. We have arrived at a familiar expression by a new route.

Another example is provided by the projection of $\mathbf{z} = \mathbf{Y} - \mathbf{X}_1\alpha$ onto $\mathcal{S}(\mathbf{X}_1, \mathbf{X}_2)$. This yields

$$\hat{\mathbf{z}} = \hat{\mathbf{Y}} - \mathbf{X}_1\alpha$$

because $\mathbf{X}_1$ is already contained in $\mathcal{S}(\mathbf{X}_1, \mathbf{X}_2)$. Another way to write the result is

$$\hat{\mathbf{z}} = \hat{\mathbf{Y}} - \mathbf{X}_1\alpha = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix} \alpha \\ \mathbf{0} \end{pmatrix}$$

or

$$\hat{\mathbf{z}} = (\mathbf{X}_1, \mathbf{X}_2)\begin{pmatrix} \hat{\beta}_1 - \alpha \\ \hat{\beta}_2 \end{pmatrix}.$$

Thus, when we project $\mathbf{z} = \mathbf{Y} - \mathbf{X}_1\alpha$ onto $\mathcal{S}(\mathbf{X}_1, \mathbf{X}_2)$, the $\mathbf{X}_2$ projection coefficients are unaffected. The $\mathbf{X}_2$ coefficients are just what they would be if $\mathbf{Y}$ were projected onto $\mathbf{X}$.

For an example using this result, consider estimating a model of international migration from a given country over time, with the log of the total number of emigrants M_t serving as the dependent variable,

$$\ln \mathbf{M} = \alpha \ln \mathbf{P} + \boldsymbol{\epsilon},$$

where $\mathbf{P}$ is the country's total population size, which also enters in log form. Suppose that you have run this regression and have an estimate $\hat{\alpha}$. Now you are asked about the effect of population size on the *migration rate*, or M_t/P_t. Do you need to take the log of this new dependent variable and re-estimate the model to see what the coefficient of population will be? Not at all—rewriting the equation as

$$\ln \mathbf{M} - \ln \mathbf{P} = \alpha \ln \mathbf{P} - \ln \mathbf{P} + \boldsymbol{\epsilon},$$

we see that the estimated coefficient must be the original $\hat{\alpha} - 1$. But what if $\mathbf{P}$ had entered the right-hand side of the original regression rather than $\ln \mathbf{P}$? In that case, switching the dependent variable to the log of the migration rate *would* make it necessary to re-run the model.

Idempotent Matrices

As you already know, both $\mathbf{P}_\mathbf{X}$ and $\mathbf{M}_\mathbf{X}$ are symmetric and idempotent. This makes good sense if we think about the matter in terms of projections. Consider the sequence of projections $\mathbf{P}_\mathbf{X}(\mathbf{P}_\mathbf{X}\mathbf{Y})$, whereby we first project $\mathbf{Y}$ onto $\mathcal{S}(\mathbf{X})$, yielding $\hat{\mathbf{Y}}$, and then project $\hat{\mathbf{Y}}$ onto $\mathcal{S}(\mathbf{X})$. Since $\hat{\mathbf{Y}}$ is *already* in $\mathcal{S}(\mathbf{X})$, when we undertake the second projection we find that the closest vector in $\mathcal{S}(\mathbf{X})$ to $\hat{\mathbf{Y}}$ is just $\hat{\mathbf{Y}}$ itself. The "error" of the second projection is identically zero. Therefore $\mathbf{P}_\mathbf{X}\mathbf{P}_\mathbf{X}\mathbf{Y} = \mathbf{P}_\mathbf{X}\mathbf{Y}$; and because this is true for all $\mathbf{Y}$, $\mathbf{P}_\mathbf{X}$ must be idempotent. The same kind of argument applies to $\mathbf{M}_\mathbf{X}$.

The converse is also true: if a matrix is symmetric and idempotent, it is an orthogonal projection. The proof is sketched by Seber (1980, p. 14), who shows that if $\mathbf{P}$ is an $n \times n$ symmetric idempotent matrix of rank k, then it represents an orthogonal projection onto some subspace of dimension k.

Table 11.1: Spaces, subspaces and associated projection matrices

$\mathcal{S}(\mathbf{X}_1, \mathbf{X}_2), \mathbf{P}_\mathbf{X}$	$\mathcal{S}(\mathbf{X}_1)$, $\mathbf{P}_1$	$\mathcal{S}(\mathbf{X}_1, \mathbf{X}_2) \supset \mathcal{S}(\mathbf{X}_1), \mathbf{P}_\mathbf{X}\mathbf{P}_1 = \mathbf{P}_1\mathbf{P}_\mathbf{X} = \mathbf{P}_1$
$\mathcal{S}^\perp(\mathbf{X}_1, \mathbf{X}_2), \mathbf{M}_\mathbf{X}$	$\mathcal{S}^\perp(\mathbf{X}_1), \mathbf{M}_1$	$\mathcal{S}^\perp(\mathbf{X}_1, \mathbf{X}_2) \subset \mathcal{S}^\perp(\mathbf{X}_1), \mathbf{M}_\mathbf{X}\mathbf{M}_1 = \mathbf{M}_1\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{X}$

In summary, there is an intimate connection between idempotent matrices and projections.[1]

Annihilation

The projection matrix $\mathbf{P}_\mathbf{X}$ annihilates any vector in $\mathcal{S}^\perp(\mathbf{X})$, by which we mean that if $\mathbf{z} \in \mathcal{S}^\perp(\mathbf{X})$, then $\mathbf{P}_\mathbf{X}\mathbf{z} = \mathbf{0}_n$. The projection matrix $\mathbf{M}_\mathbf{X}$ does the same for any vector in $\mathcal{S}(\mathbf{X})$. Another way to express this is to write $\mathbf{P}_\mathbf{X}\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{X}\mathbf{P}_\mathbf{X} = \mathbf{0}_{n\times n}$. This last statement can be proven by straightforward algebra, but again it may help to consider things in terms of projections. Examine the sequence of projections $\mathbf{M}_\mathbf{X}\mathbf{P}_\mathbf{X}\mathbf{Y}$. The projection $\mathbf{P}_\mathbf{X}\mathbf{Y}$ will find the vector $\hat{\mathbf{Y}}$ that is closest to $\mathbf{Y}$ in $\mathcal{S}(\mathbf{X})$. Then $\mathbf{M}_\mathbf{X}$ will find the closest vector to $\hat{\mathbf{Y}}$ in $\mathcal{S}^\perp(\mathbf{X})$. But because $\hat{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$, it must be orthogonal to all vectors in $\mathcal{S}^\perp(\mathbf{X})$. Indeed, the closest vector to $\hat{\mathbf{Y}}$ that can be found in $\mathcal{S}^\perp(\mathbf{X})$ is $\hat{\hat{\mathbf{Y}}} = \mathbf{0}_n$, the zero vector. To sum up, then, $\mathbf{M}_\mathbf{X}\mathbf{P}_\mathbf{X}\mathbf{Y} = \mathbf{0}$, and applying the same reasoning establishes that $\mathbf{P}_\mathbf{X}\mathbf{M}_\mathbf{X}\mathbf{Y} = \mathbf{0}$.

If you feel uncomfortable with this geometric style of argument, then here's an easy algebraic proof that $\mathbf{P}_\mathbf{X}\mathbf{M}_\mathbf{X}\mathbf{Y} = \mathbf{0}$. Let $\bar{\mathbf{Y}} = \mathbf{M}_\mathbf{X}\mathbf{Y}$. To project $\bar{\mathbf{Y}}$ onto $\mathcal{S}(\mathbf{X})$, we examine all $\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$ to find the one that minimizes the distance

$$D = (\bar{\mathbf{Y}} - \tilde{\mathbf{Y}})'(\bar{\mathbf{Y}} - \tilde{\mathbf{Y}}) = \bar{\mathbf{Y}}'\bar{\mathbf{Y}} - 2\tilde{\mathbf{Y}}'\bar{\mathbf{Y}} + \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}.$$

But $\bar{\mathbf{Y}} \in \mathcal{S}^\perp(\mathbf{X})$ and it is therefore orthogonal to all vectors $\tilde{\mathbf{Y}} \in \mathcal{S}(\mathbf{X})$. As a result, the distance expression simplifies to $D = \bar{\mathbf{Y}}'\bar{\mathbf{Y}} + \tilde{\mathbf{Y}}'\tilde{\mathbf{Y}}$. The distance-minimizing $\tilde{\mathbf{Y}}$ is clearly a vector of zeroes.

Subspaces

We often want to consider projections onto subspaces of $\mathcal{S}(\mathbf{X})$. For example, we might want to partition $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2)$ and consider the subspace $\mathcal{S}(\mathbf{X}_1)$, which is smaller than and contained within the space $\mathcal{S}(\mathbf{X}_1, \mathbf{X}_2)$. Let W represent one such subspace of $\mathcal{S}(\mathbf{X})$. We might project $\mathbf{Y}$ onto W directly, a procedure that yields $\hat{\mathbf{Y}}_W$. Alternatively, we could proceed in two steps, by first projecting $\mathbf{Y}$ onto the larger space $\mathcal{S}(\mathbf{X})$, obtaining $\hat{\mathbf{Y}}$, and then project $\hat{\mathbf{Y}}$ onto the subspace W. The final result, $\hat{\mathbf{Y}}_W$, is the same for both procedures.

What we have just argued is that $\mathbf{P}_W\mathbf{Y} = \mathbf{P}_W\mathbf{P}_\mathbf{X}\mathbf{Y}$. This very important result follows algebraically from the orthogonality between the projection error $\mathbf{e} = \mathbf{Y} - \mathbf{P}_\mathbf{X}\mathbf{Y}$ and W. If we attempt to project $\mathbf{e}$ onto W, a subspace to which $\mathbf{e}$ is already orthogonal, we obtain the zero vector. Thus, $\mathbf{P}_W\mathbf{e} = \mathbf{P}_W(\mathbf{Y} - \mathbf{P}_\mathbf{X}\mathbf{Y}) = \mathbf{0}$ and it follows that $\mathbf{P}_W\mathbf{Y} = \mathbf{P}_W\mathbf{P}_\mathbf{X}\mathbf{Y}$.

Taking a different logical route, we can show that $\mathbf{P}_\mathbf{X}\mathbf{P}_W\mathbf{Y} = \mathbf{P}_W\mathbf{Y}$. In this case, we simply recognize that $\hat{\mathbf{Y}}_W = \mathbf{P}_W\mathbf{Y}$ is an element of $\mathcal{S}(\mathbf{X})$. Therefore, $\mathbf{P}_\mathbf{X}\hat{\mathbf{Y}}_W = \hat{\mathbf{Y}}_W = \mathbf{P}_W\mathbf{Y}$.

[1] This connection is fundamental to the general theory of projections; for a formal development see Pollock (1979), among others.

For an example of these results, let's consider the projection matrices $\mathbf{M}_{\mathbf{X}}$ and $\mathbf{M}_1$, which are associated with the spaces $\mathcal{S}^{\perp}(\mathbf{X}_1, \mathbf{X}_2)$ and $\mathcal{S}^{\perp}(\mathbf{X}_1)$, respectively. Now $\mathcal{S}^{\perp}(\mathbf{X}_1)$ is larger than $\mathcal{S}^{\perp}(\mathbf{X}_1, \mathbf{X}_2)$, and $\mathcal{S}^{\perp}(\mathbf{X}_1, \mathbf{X}_2)$ is contained within $\mathcal{S}^{\perp}(\mathbf{X}_1)$. It follows that $\mathbf{M}_{\mathbf{X}}\mathbf{M}_1 = \mathbf{M}_1\mathbf{M}_{\mathbf{X}} = \mathbf{M}_{\mathbf{X}}$. Table 11.1 summarizes these relationships.

An Economic Application

An interesting application of these ideas is found in some systems of regression equations, such as those describing a full set of demand functions. Consider the m-equation system

$$\begin{aligned} \mathbf{Y}_1 &= \alpha_1 + \mathbf{Y}\beta_1 + \mathbf{p}'\gamma_1 + \epsilon_1 \\ \mathbf{Y}_2 &= \alpha_2 + \mathbf{Y}\beta_2 + \mathbf{p}'\gamma_2 + \epsilon_2 \\ &\vdots \\ \mathbf{Y}_m &= \alpha_m + \mathbf{Y}\beta_m + \mathbf{p}'\gamma_3 + \epsilon_m \end{aligned}$$

In this system, the $\mathbf{Y}_i$ variable represents expenditures on the i-th commodity or service, and $\sum_i^m \mathbf{Y}_i = \mathbf{Y}$, where $\mathbf{Y}$ is total expenditure on all of the commodities. The notion is that total expenditure $\mathbf{Y}$, which appears on the right-hand side of each equation, plays a role analogous to total income. Also, the vector $\mathbf{p}$, an m-vector of prices, appears in the equation system. A key feature of this system is that the *same* explanatory variables $(\iota, \mathbf{Y}, \mathbf{p})$ appear on the right-hand side of each equation.

Suppose that ordinary least squares is applied equation-by-equation to this system. If we sum the individual parameter estimates, must an adding-up constraint apply? That is, must it be the case that $\sum_{i=1}^m \hat{\beta}_i = 1$ and that both $\hat{\alpha}_i$ and $\hat{\gamma}_i$ sum to zero? The answer, perhaps surprisingly, is yes.

To see this, let $\delta_i = (\alpha_i, \beta_i, \gamma_i)'$ be the coefficients for the i-th commodity, so that $\mathbf{Y}_i = \mathbf{X}\delta_i + \epsilon_i$, and define $\hat{\delta}_i$ as the least squares estimate of these coefficients. Here $\mathbf{X} = (\boldsymbol{\iota}, \mathbf{Y}, \mathbf{p})$ denotes the common set of right-hand side explanatory variables and $\hat{\delta}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_i$. Summing the δ_i vectors over i,

$$\sum_{i=1}^m \hat{\delta}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\sum_{i=1}^m \mathbf{Y}_i) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

Now, $\mathbf{Y}$ is itself one of the variables contained in $\mathbf{X}$. The expression $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ can thus be regarded as the set of least squares coefficients from a projection of $\mathbf{Y}$ onto a collection of vectors that includes itself and $(\boldsymbol{\iota}, \mathbf{p})$. It must be the case that in such a projection, $\mathbf{Y}$ perfectly explains itself. Thus its coefficient must equal unity and the coefficients associated with $\boldsymbol{\iota}$ and $\mathbf{p}$ must equal zero. In other words, $\sum_i^m \hat{\beta}_i = 1$, $\sum_m^i \hat{\alpha}_i = 0$, and $\sum_i^m \hat{\gamma}_i = \mathbf{0}$.

This argument also applies to equation systems in which the dependent variables $\mathbf{Y}_i$ represent shares (or percentages) that add to unity (or to 100) and a constant term is on the right-hand side. In this case, the coefficients associated with $\boldsymbol{\iota}$ will sum to unity (or to 100) and all the remaining coefficients will sum to zero. Remember that for this to occur, the equations in the system must have exactly the same explanatory variables, as in the example above.

11.2 The Frisch-Waugh-Lovell (FWL) Theorem

In the next sections we will be thinking of regression equations in purely computational terms. Partition $\mathbf{X}$ as $(\mathbf{X}_1, \mathbf{X}_2)$ and consider the regression equation

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \boldsymbol{\epsilon}. \tag{11.5}$$

The FWL Theorem asserts two things. First, if we apply least squares to equation (11.5), we obtain an estimate $\hat{\beta}_2$ that is *numerically identical* to the estimate derived from applying least squares to the transformed equation,

$$\mathbf{M}_1\mathbf{Y} = \mathbf{M}_1\mathbf{X}_2\beta_2 + \mathbf{M}_1\boldsymbol{\epsilon}. \tag{11.6}$$

In the latter equation, $\mathbf{M}_1 = \mathbf{I} - \mathbf{P}_1 = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1'$, that is, $\mathbf{M}_1\mathbf{z}$ generates residuals from a regression of a vector $\mathbf{z}$ on $\mathbf{X}_1$. Second, the FWL Theorem asserts that the *residuals* from the fitted version of equation (11.6) are numerically identical to the residuals from the fitted version of the full equation (11.5).

Conventional proofs of the FWL Theorem use partitioned inverses. (See, for example, Greene (2003, section 3.3).) The route we will follow, modeled on Davidson and MacKinnon (1993, pp. 20–22), employs projections.

To show that the least squares estimate $\hat{\beta}_2$ from the full equation (11.5) is the same as

$$\hat{\hat{\beta}}_2 = \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y},$$

which is the result of applying least squares to the transformed equation (11.6), we rewrite $\mathbf{Y}$ as follows:

$$\mathbf{Y} = \mathbf{P}_\mathbf{X}\mathbf{Y} + \mathbf{M}_\mathbf{X}\mathbf{Y} = \mathbf{X}_1\hat{\beta}_1 + \mathbf{X}_2\hat{\beta}_2 + \mathbf{M}_\mathbf{X}\mathbf{Y}.$$

This expresses $\mathbf{Y}$ as the sum of the fitted values from the full regression equation (11.5) and the residuals from that equation. Now multiply both sides by $\mathbf{X}_2'\mathbf{M}_1$, yielding

$$\mathbf{X}_2'\mathbf{M}_1\mathbf{Y} = \mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\hat{\beta}_2,$$

where we have used $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$, $\mathbf{M}_1\mathbf{M}_\mathbf{X} = \mathbf{M}_\mathbf{X}$ and $\mathbf{X}_2'\mathbf{M}_\mathbf{X} = \mathbf{0}$. The second equality is due to the fact we mentioned earlier, that the space of vectors orthogonal to $\mathbf{X}$, or $\mathcal{S}^\perp(\mathbf{X})$, is smaller than and contained within the space of vectors orthogonal to $\mathbf{X}_1$, or $\mathcal{S}^\perp(\mathbf{X}_1)$. The FWL result for $\hat{\hat{\beta}}_2$ follows immediately.

To prove that the residuals from equation (11.6) are the same as those from equation (11.5), we simply multiply $\mathbf{Y}$ by $\mathbf{M}_1$,

$$\mathbf{M}_1\mathbf{Y} = \mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 + \mathbf{M}_1\mathbf{M}_\mathbf{X}\mathbf{Y} = \mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 + \mathbf{M}_\mathbf{X}\mathbf{Y} = \mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 + \mathbf{e}.$$

Hence, the residuals from the transformed equation, $\mathbf{M}_1\mathbf{Y} - \mathbf{M}_1\mathbf{X}_2\hat{\beta}_2 = \mathbf{e}$, are the same as those from the untransformed equation.

Finally, it is worth noting that the estimator of the covariance matrix of $\hat{\hat{\beta}}_2$ based on the transformed equation, $\hat{\sigma}^2\left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}$, is identical to the estimate derived from the full equation. (But note that if s^2 is used to estimate σ^2, the sum of squared residuals from the transformed equation (11.6) should be divided by $n-k$ rather than by $n-k_2$.) Hence, if

statistical tests are to be constructed using the transformed equation, all of the necessary elements are in place.

Here's a little R program illustrating the theorem:

```
x1 = rnorm(100)
x2 = rnorm(100)
y1 = 1 + x1 - x2 + rnorm(100)

r1 = residuals(lm(y1 ~ x2))
r2 = residuals(lm(x1 ~ x2))
# ols
coef(lm(y1 ~ x1 + x2))
# fwl ols
coef(lm(r1 ~ -1 + r2))
```

11.3 The $\mathcal{F}$-test Revisited

Projection methods are a great aid in simplifying the algebra associated with $\mathcal{F}$ statistics for testing the relevance of a set of explanatory variables. Suppose that we partition $\mathbf{X}$ as $(\mathbf{X}_1, \mathbf{X}_2)$ and formulate a test for the relevance of $\mathbf{X}_2$. As you know, the $\mathcal{F}$ statistic can be written in terms of the unconstrained and constrained error sums of squares. To understand this from the projections point of view, let us first project $\mathbf{Y}$ on $\mathbf{X}$, yielding $\hat{\mathbf{Y}}$ and then project $\mathbf{Y}$ onto the subset $\mathbf{X}_1$, a procedure that yields $\hat{\mathbf{Y}}_1$. Note that the unconstrained error vector $\mathbf{e}_u = \mathbf{Y} - \hat{\mathbf{Y}}$ is orthogonal to $\mathbf{X}_1$ as well as to any linear combination of the vectors in $\mathbf{X}_1$, such as $\hat{\mathbf{Y}}_1$. It follows from these orthogonality conditions that the constrained error sum of squares can be written

$$\mathbf{e}_c'\mathbf{e}_c = (\mathbf{Y} - \hat{\mathbf{Y}}_1)'(\mathbf{Y} - \hat{\mathbf{Y}}_1) = (\mathbf{Y} - \hat{\mathbf{Y}})'(\mathbf{Y} - \hat{\mathbf{Y}}) + (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_1)'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_1),$$

or

$$\mathbf{e}_c'\mathbf{e}_c = \mathbf{e}_u'\mathbf{e}_u + (\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_1)'(\hat{\mathbf{Y}} - \hat{\mathbf{Y}}_1).$$

To make use of this relationship, we note that $\hat{\mathbf{Y}}$ is the vector of fitted values from a regression of $\mathbf{Y}$ on $\mathbf{X}$, and $\hat{\mathbf{Y}}_1$ is a vector of fitted values from the regression of $\mathbf{Y}$ on $\mathbf{X}_1$. Thus, one way of forming the numerator of the $\mathcal{F}$ test is to use the sum of squares of the differences in these fitted values.

Given the frequency with which such expressions appear in the literature, it may be worthwhile to state the result above using different notation. The constrained residuals $\mathbf{e}_c$ are given by $\mathbf{e}_c = \mathbf{M}_1\mathbf{Y}$ and the sum of squares of these residuals is $\mathbf{e}_c'\mathbf{e}_c = \mathbf{Y}'\mathbf{M}_1\mathbf{Y}$. From the FWL Theorem, the unconstrained residuals $\mathbf{e}_u = \mathbf{M}_X\mathbf{Y}$, are equal to the residuals from the fitted equation

$$\mathbf{M}_1\mathbf{Y} = \mathbf{M}_1\mathbf{X}_2\hat{\hat{\beta}}_2 + \text{residuals}.$$

Because of this, the unconstrained residuals can be written as $\mathbf{M}_{1,2}\mathbf{M}_1\mathbf{Y}$, where

$$\mathbf{M}_{1,2} = \mathbf{I} - \mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1.$$

On writing out the sum of squares of the unconstrained residuals, that is,

$$\mathbf{e}_u'\mathbf{e}_u = \mathbf{Y}'\mathbf{M}_1\mathbf{M}_{1,2}\mathbf{M}_1\mathbf{Y},$$

we find

$$\mathbf{Y}'\mathbf{M}_1\mathbf{M}_{1,2}\mathbf{M}_1\mathbf{Y} = \mathbf{Y}'\mathbf{M}_1\mathbf{Y} - \mathbf{Y}'\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y}.$$

Thus, the difference between the restricted and unrestricted sums of squared residuals is

$$\mathbf{e}_c'\mathbf{e}_c - \mathbf{e}_u'\mathbf{e}_u = \mathbf{Y}'\mathbf{M}_1\mathbf{X}_2(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{Y} = \mathbf{Y}'\mathbf{P}_{1,2}\mathbf{Y}$$

where $\mathbf{P}_{1,2}$ projects onto the space spanned by $\mathbf{M}_1\mathbf{X}_2$. This expression occurs often enough in the literature that you should be sure to understand its origins. As Davidson and MacKinnon (1993, p. 84) note, the $\mathcal{F}$ test statistic is

$$f = \frac{\mathbf{Y}'\mathbf{P}_{1,2}\mathbf{Y}/k_2}{\mathbf{Y}'\mathbf{M}_\mathbf{X}\mathbf{Y}/(n-k)}$$

and because the space spanned by $\mathbf{M}_1\mathbf{X}_2$ is a subspace of $\mathcal{S}(\mathbf{X})$—we know this from $\mathbf{M}_1\mathbf{X}_2 = \mathbf{X}_2 - \mathbf{P}_1\mathbf{X}_2$—it must be the case that $\mathbf{M}_\mathbf{X}\mathbf{P}_{1,2} = \mathbf{0}$, a result that establishes the statistical independence of numerator and denominator.

Now consider a more general $\mathcal{F}$ test that involves the restriction $\mathbf{R}\beta = \mathbf{r}$, where $\mathbf{R}$ is $q \times k$ and $\mathbf{r}$ is $q \times 1$. Let us reparameterize so that this can be written as a set of zero restrictions analogous to the above. (Our approach is taken from Davidson and MacKinnon (1993, pp. 16–19).) Since the rank of $\mathbf{R}$ is q, the restriction can be expressed in the form

$$\mathbf{R}_1\beta_1 + \mathbf{R}_2\beta_2 = \mathbf{r}$$

where $\mathbf{R}_1$ is a nonsingular $q \times q$ matrix and $\mathbf{R}_2$ is $q \times k - q$ matrix, with β_1 and β_2 partitioned accordingly. Note that to put things in the form just shown, it may be necessary to rearrange the columns of $\mathbf{X}$ and relabel the βs accordingly. Having done that, we may write the restriction as

$$\beta_1 = \mathbf{R}_1^{-1}\left(\mathbf{r} - \mathbf{R}_2\beta_2\right).$$

When we insert this expression into the original (suitably rearranged and relabelled) regression equation, we obtain

$$\mathbf{Y} = \mathbf{X}_1\left(\mathbf{R}_1^{-1}\left(\mathbf{r} - \mathbf{R}_2\beta_2\right)\right) + \mathbf{X}_2\beta_2 + \text{residuals}.$$

This restricted regression can then be written as

$$\begin{aligned}\mathbf{Y} - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r} &= \left(\mathbf{X}_2 - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{R}_2\right)\beta_2 + \text{residuals, or}\\ \mathbf{Y}^* &= \mathbf{Z}_2\theta_2 + \text{residuals}\end{aligned}$$

with $\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r}$, $\mathbf{Z}_2 = \mathbf{X}_2 - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{R}_2$ and $\theta_2 = \beta_2$. Estimating this equation by least squares produces $\hat{\theta}_2$, the equality-constrained estimate of β_2.

To obtain a counterpart unrestricted regression, we must introduce q regressors $\mathbf{Z}_1$ that, when taken together with $\mathbf{Z}_2$, will span the original space $\mathcal{S}(\mathbf{X})$. The choice $\mathbf{Z}_1 = \mathbf{X}_1$ will do the trick; in other words, $\mathbf{Z} \equiv (\mathbf{X}_1, \mathbf{Z}_2)$ spans $\mathcal{S}(\mathbf{X})$. (You can prove this by showing that $\mathbf{Z} = \mathbf{X}\mathbf{A}$ with $\mathbf{A}$ a nonsingular $k \times k$ matrix.) The unrestricted regression can now be written in terms of $\mathbf{Y}^*$, $\mathbf{X}_1$ and $\mathbf{Z}_2$ as

$$\mathbf{Y}^* = \mathbf{X}_1\theta_1 + \mathbf{Z}_2\theta_2 + \text{residuals}$$

and expressed in these terms, the restriction $\mathbf{R}\beta = \mathbf{r}$ amounts to the assertion $\theta_1 = \mathbf{0}$ and can be so tested.

In fact, from the new restricted and unrestricted regression equations, we can obtain all the ingredients we need to calculate the $\mathcal{F}$ statistic. Why? The residuals from the new unrestricted regression are numerically identical to the residuals from the original unrestricted regression. This equivalence arises from two facts. First, on the right-hand side of the equation, we see that $(\mathbf{X}_1, \mathbf{Z}_2)$ spans the same space $\mathcal{S}(\mathbf{X})$ as did the original variables $(\mathbf{X}_1, \mathbf{X}_2)$. Second, although the dependent variable of the new regression is $\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r}$, this transformed vector differs from the original $\mathbf{Y}$ only by $\mathbf{X}_1\mathbf{R}_1^{-1}\mathbf{r}$, a vector that already lies in $\mathcal{S}(\mathbf{X})$. Hence, when we project $\mathbf{Y}^*$ onto $\mathcal{S}(\mathbf{X})$, finding the closest vector to it in $\mathcal{S}(\mathbf{X})$, the projection error $\mathbf{e}^* = \mathbf{Y}^* - \hat{\mathbf{Y}}^*$ must be identical to the original projection error $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}}$. Note how the linearity property of projections was invoked in the course of the argument.

An example of this approach is given by the regression

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \text{residuals},$$

in which $\mathbf{X}_1$ and $\mathbf{X}_2$ are single variables, with the restriction being $\beta_1 + \beta_2 = 1$. The restricted regression is then

$$\mathbf{Y}^* = \mathbf{Y} - \mathbf{X}_1 = (\mathbf{X}_2 - \mathbf{X}_1)\beta_2 + \text{residuals} = \mathbf{Z}_2\theta_2 + \text{residuals}$$

with $\mathbf{Z}_2 = \mathbf{X}_2 - \mathbf{X}_1$. The counterpart unrestricted regression is

$$\mathbf{Y}^* = \mathbf{X}_1\theta_1 + \mathbf{Z}_2\theta_2 + \text{residuals}.$$

The $\mathcal{F}$ test can be carried out by running the latter regression with and without $\mathbf{X}_1$.

A more complicated and realistic example is provided by a model with five explanatory variables and two constraints, these being $\beta_1 + \beta_2 + \beta_3 = 1$ and $\beta_1 + \beta_4 = 2$. This can be written out as

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix},$$

or alternatively as

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

Inverting the leading matrix and multiplying through yields

$$\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \end{bmatrix} - \begin{bmatrix} 0 & 1 & 0 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}.$$

Substituting this into the equation for $\mathbf{Y}$, we obtain

$$\mathbf{Y} = \mathbf{X}_1(2 - \beta_4) + \mathbf{X}_2(-1 - \beta_3 + \beta_4) + \mathbf{X}_3\beta_3 + \mathbf{X}_4\beta_4 + \mathbf{X}_5\beta_5 + \epsilon,$$

or

$$\mathbf{Y} - 2\mathbf{X}_1 + \mathbf{X}_2 = (\mathbf{X}_3 - \mathbf{X}_2)\beta_3 + (\mathbf{X}_4 + \mathbf{X}_2 - \mathbf{X}_1)\beta_4 + \mathbf{X}_5\beta_5 + \epsilon.$$

The constrained model can be estimated in this form—you'd simply need to generate a new dependent variable and two new explanatory variables.

There is a practical point to be extracted from all this: No matter how complicated the restriction $\mathbf{R}\beta = \mathbf{r}$ might appear to be at first, there exists a simple way of computing the associated $\mathcal{F}$ statistic. Find the transformation of the data such that the restriction $\mathbf{R}\beta = \mathbf{r}$ can be written as a zero restriction; and then take the restricted and unrestricted error sums of squares from the transformed regression equation.

11.4 Projections with Linear Constraints

Returning to the general expression $\beta_1 = \mathbf{R}_1^{-1}\mathbf{r} - \mathbf{R}_1^{-1}\mathbf{R}_2\beta_2$, with β_1 being $m \times 1$ and β_2 being $k - m \times 1$, we can write

$$\begin{bmatrix}\beta_1 \\ \beta_2\end{bmatrix} = \begin{bmatrix}-\mathbf{R}_1^{-1}\mathbf{R}_2 \\ \mathbf{I}\end{bmatrix}\beta_2 + \begin{bmatrix}\mathbf{R}_1^{-1}\mathbf{r} \\ \mathbf{0}\end{bmatrix},$$

or $\beta = \mathbf{S}\gamma + \tau$, where γ is a subset of the βs and τ is a "translation vector" whose value is known, determined by the nature of the constraints. Following Ruud (2000, p. 79), we can implement the constrained estimator by inserting this expression into $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, yielding

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}(\mathbf{S}\gamma + \tau) + \epsilon \\ \mathbf{Y} - \mathbf{X}\tau &= \mathbf{X}\mathbf{S}\gamma + \epsilon,\end{aligned}$$

which implies that

$$\hat{\gamma} = \left(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\tau).$$

From this we obtain the constrained estimator

$$\tilde{\beta} = \mathbf{S}\hat{\gamma} + \tau.$$

We can think of the projection of $\mathbf{Y}$ onto $\mathcal{S}(\mathbf{X})$, subject to the constraints, as the vector

$$\begin{aligned}\hat{\mathbf{Y}}_c &= \mathbf{X}\tilde{\beta} = \mathbf{X}\mathbf{S}\left(\mathbf{S}'\mathbf{X}'\mathbf{X}\mathbf{S}\right)^{-1}\mathbf{S}'\mathbf{X}'(\mathbf{Y} - \mathbf{X}\tau) + \mathbf{X}\tau \\ &= \mathbf{P}_{\mathbf{X}\mathbf{S}}\mathbf{Y} + (\mathbf{I} - \mathbf{P}_{\mathbf{X}\mathbf{S}})\mathbf{X}\tau.\end{aligned}$$

That is, to implement this projection, we first project $\mathbf{Y}$ onto the space $\mathcal{S}(\mathbf{X}\mathbf{S})$, and then make a translation adjustment of $(\mathbf{I} - \mathbf{P}_{\mathbf{X}\mathbf{S}})\mathbf{X}\tau$.

Chapter 12

Influential Observations

This chapter follows the approach of Davidson and MacKinnon (1993, pp. 32–39) in formulating measures of the influence of a single observation on least squares estimates and residuals. Although indirect in approach, the proofs they have devised are less algebraic than the conventional proofs, yield side benefits in the form of additional results useful in other contexts, and also exploit the properties of projection and the FWL Theorem that were set out in our previous chapter.

One caution should be issued. The method discussed here is effective only in identifying *individual* observations that are influential. It will not, in general, identify *groups* of influential observations, such as multiple "outliers." For techniques that are effective more broadly, see the interesting review by Zaman (1996).

As will be shown below, the influence of the t-th observation depends on the t-th diagonal element of the projection matrix $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. We shall focus on the influence of the last, or T-th, observation in a sample of $t = 1, \ldots, T$ observations. The associated element of $\mathbf{P}$ is $\mathbf{P}_{TT} = \mathbf{D}_T'\mathbf{P}\mathbf{D}_T$ where $\mathbf{D}_T$ is a $T \times 1$ vector with a 1 in the T-th position and zeroes elsewhere.

A few facts about $\mathbf{P}_{TT}$ can be stated at the outset. Since $\mathbf{P}$ is positive semidefinite, $\mathbf{P}_{TT} \geq 0$. Furthermore, from the fact that $\mathbf{D}_T'(\mathbf{I} - \mathbf{P})\mathbf{D}_T \geq 0$ we see that $\mathbf{P}_{TT} \leq 1$. Recall that the trace of an idempotent matrix is equal to its rank. The trace of $\mathbf{P}$ is k, and the average $\mathbf{P}_{tt}$, or T^{-1} trace $\mathbf{P}$, is just k/T. This can be taken as a benchmark against which we can compare values of $\mathbf{P}_{TT}$.

We now show why $\mathbf{P}_{TT}$ can be interpreted as a measure of the influence or "leverage" exerted by the T-th observation. To begin, we will state the main result and then, following Davidson and MacKinnon (1993), show how that result can be derived. Let $\mathbf{Y}, \mathbf{X}$ denote the full set of T observations on $\mathbf{Y}$ and $\mathbf{X}$, and let $\mathbf{Y}_T$ and $\mathbf{X}_T$ denote the data for the T-th observation. If the T-th observation is omitted, the least squares estimator $\hat{\beta}^T$ is given by

$$\hat{\beta}^T = [\mathbf{X}'\mathbf{X} - \mathbf{X}_T\mathbf{X}_T']^{-1}[\mathbf{X}'\mathbf{Y} - \mathbf{X}_T\mathbf{Y}_T]$$

and the only difficulty is how to derive the inverse. Results presented by Greene (2003, equation A-66) can be applied, and from these we find

$$\hat{\beta}^T = \hat{\beta} - \frac{1}{1 - \mathbf{P}_{TT}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_T\mathbf{e}_T,$$

in which the role of the leverage factor $\mathbf{P}_{TT}$ is evident. In the above, $\hat{\beta}$ is the full-sample OLS estimator and $\mathbf{e}_T$ is the residual for the T-th observation when the full sample is used. The difference between $\hat{\beta}^T$ based on $T-1$ observations and the full sample $\hat{\beta}$ depends (in part) on the leverage $\mathbf{P}_{TT}$ exerted by the T-th observation. In what follows, we will take a different route that leads to the same result.

12.1 Leverage

Davidson and MacKinnon develop their proofs by inserting into the original regression of interest, or $\mathbf{Y} = \mathbf{X}\beta +$ residuals, a dummy variable that effectively nullifies the influence of the T-th observation in the sample. Consider a specification in which a dummy variable singles out that T-th observation,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}_T\alpha + \text{residuals}. \tag{12.1}$$

As above, the $T \times 1$ vector $\mathbf{D}_T$ has 1 in the T-th position and zeroes elsewhere.

By the FWL Theorem, the least-squares estimates of β are the same as the estimates from

$$\mathbf{M}_D\mathbf{Y} = \mathbf{M}_D\mathbf{X}\beta + \text{residuals}.$$

It is easy to see that projecting any vector $\mathbf{z}$ onto $\mathbf{D}_T$ gives a $\hat{\mathbf{z}}$ that has zeroes in positions $t = 1, \ldots, T-1$ and $\mathbf{z}_T$ in the T-th position; the projection error $\mathbf{M}_D\mathbf{z}$ is thus the original $\mathbf{z}$ vector except for a single zero in the T-th position. That is,

$$\mathbf{M}_D\mathbf{Y} = \begin{bmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \\ \vdots \\ \mathbf{Y}_{T-1} \\ 0 \end{bmatrix}, \mathbf{M}_D\mathbf{X} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_{T-1}' \\ \mathbf{0}' \end{bmatrix}$$

where $\mathbf{X}_t'$ is the t-th row of $\mathbf{X}$.

Since the T-th elements of the transformed data are zeros, we see that the estimate $\hat{\beta}$ derived from equation (12.1) must be numerically identical to the estimate that would have been obtained from a sample from which the T-th observation was omitted. Let $\hat{\beta}^T$ denote the least-squares estimate of β based on equation (12.1), identical to the estimate produced by simply dropping the T-th observation, and let $\mathbf{e}^T$ denote the vector of residuals from (12.1). A least-squares first-order condition for equation (12.1), $\mathbf{D}_T'\mathbf{e}^T = 0$, ensures that the residual $\mathbf{e}_T^T$ for observation T must be zero. Hence, the estimate of α, the scalar coefficient attached to $\mathbf{D}_T$, must be $\hat{\alpha} = \mathbf{Y}_T - \mathbf{X}_T'\hat{\beta}^T$. It follows that $\hat{\alpha}$ can be viewed as a kind of prediction error for the T-th observation based on data from observations $1, \ldots, T-1$.

Differences in residuals

One way to think about the influence of the T-th observation is to ask how the residual for that observation, $\mathbf{Y}_T - \hat{\mathbf{Y}}_T$, is affected by omitting $(\mathbf{Y}_T, \mathbf{X}_T)$ from the calculation of $\hat{\beta}$.

We think of influential observations as being ones that draw the regression line toward themselves, thereby reducing the residual difference.

The vector of residuals based on the full sample of data is denoted by $\mathbf{e}$—this is the residual from the full-sample regression of $\mathbf{Y}$ on $\mathbf{X}$ alone. As above, we denote by $\mathbf{e}^T$ the residual vector based on equation (12.1). The vector of differences in these residuals is

$$\mathbf{e}^T - \mathbf{e} = \mathbf{X}(\hat{\beta} - \hat{\beta}^T) - \mathbf{D}_T\hat{\alpha}. \tag{12.2}$$

Multiply both sides of equation (12.2) by $\mathbf{M}_\mathbf{X}$. Since $\mathbf{M}_\mathbf{X}\mathbf{X} = \mathbf{0}$ and neither residual vector is affected by multiplying by $\mathbf{M}_\mathbf{X}$, we obtain

$$\mathbf{e}^T - \mathbf{e} = -\mathbf{M}_\mathbf{X}\mathbf{D}_T\hat{\alpha}.$$

Another multiplication, this time by $\mathbf{D}_T'$, gives

$$\mathbf{D}_T'(\mathbf{e}^T - \mathbf{e}) = -\mathbf{e}_T = -\mathbf{D}_T'\mathbf{M}_\mathbf{X}\mathbf{D}_T\hat{\alpha}$$

since $\mathbf{e}_T^T = 0$. This result can be re-expressed as $\mathbf{e}_T = (1 - \mathbf{P}_{TT})\hat{\alpha}$. From this we have a second representation of $\hat{\alpha} = \mathbf{e}_T/(1 - \mathbf{P}_{TT})$.

With this result in hand, we can compare the residual for the T-th observation when that observation is omitted from the calculation of $\hat{\beta}$ to its value when the observation is included. As $\mathbf{Y}_T - \mathbf{X}_T'\hat{\beta}^T = \hat{\alpha}$, the change in residuals is

$$(\mathbf{Y}_T - \mathbf{X}_T'\hat{\beta}^T) - (\mathbf{Y}_T - \mathbf{X}_T'\hat{\beta}) = \hat{\alpha} - \mathbf{e}_T = \frac{\mathbf{P}_{TT}}{1 - \mathbf{P}_{TT}}\mathbf{e}_T,$$

an expression that depends on the full-sample residual $\mathbf{e}_T$ as well as the leverage measure $\mathbf{P}_{TT}$.

The effect on $\hat{\beta}$

Writing equation (2) as

$$\mathbf{e} - \mathbf{e}^T = \mathbf{X}\hat{\beta}^T + \mathbf{D}_T\hat{\alpha} - \mathbf{X}\hat{\beta},$$

and then multiplying by $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ yields

$$\mathbf{0} = \hat{\beta}^T - \hat{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_T\hat{\alpha}$$

or

$$\hat{\beta}^T - \hat{\beta} = -\frac{1}{1 - \mathbf{P}_{TT}}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_T\mathbf{e}_T.$$

This is the result that was presented at the outset.

12.2 An Example

The following example is taken from Davidson and MacKinnon. Applied to a sample of 10 observations, a least-squares regression of $\mathbf{Y}$ on a constant and $\mathbf{X}$ produces the following estimates $\hat{\mathbf{Y}} = 3.420 + .238\mathbf{X}$ and $s^2 = .908$.

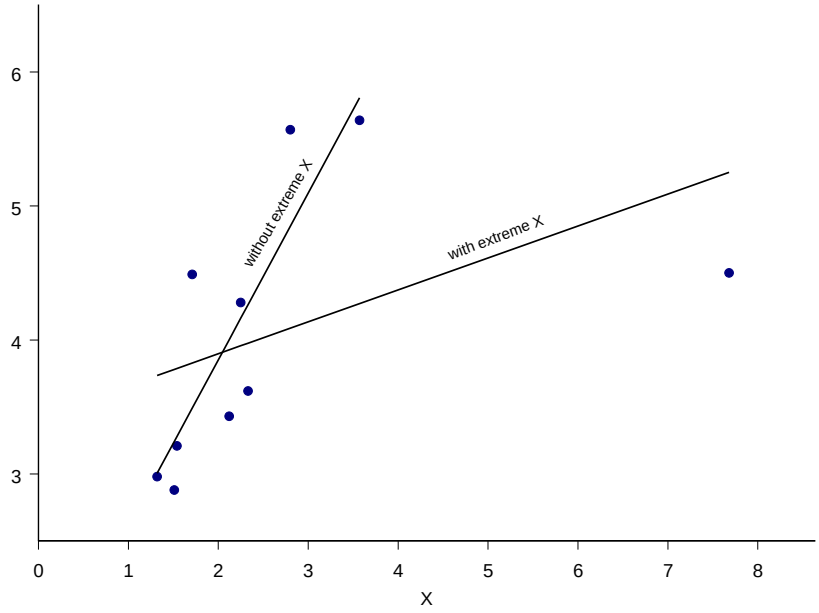


Figure 12.1: An example of influence

Obs	Y	X	e	P_tt	(P_tt/(1-P_tt))e
1	2.88	1.51	-.9003869	.143151	-.1504247
2	3.62	2.33	-.355854	.1039079	-.0412637
3	5.64	3.57	1.368562	.1246741	.1949265
4	3.43	2.12	-.4957952	.1099406	-.0612409
5	3.21	1.54	-.5775382	.140972	-.0947777
6	4.49	1.71	.6619379	.1296907	.0986398
7	4.50	7.68	-.7511563	.8830925	-5.674064
8	4.28	2.25	.3232163	.1058799	.0382746
9	2.98	1.32	-.7550959	.1582621	-.1419718
10	5.57	2.80	1.48211	.1004293	.1654648

The value of $\mathbf{X}_7$ (observation 7) was deliberately chosen to be extreme. Note that the full-sample regression residual $\mathbf{e}_7$ does not appear unusual in relation to the other residuals (see Figure (12.1)). The influence ($\mathbf{P}_{77}$) wielded by this observation is considerable, however, as is the net effect $(\mathbf{P}_{77}/(1-\mathbf{P}_{77}))\mathbf{e}_7$. What would have been a very large (negative) residual with observation 7 omitted, becomes a much smaller (negative) residual. The message is that residuals, by themselves, provide an inadequate basis for diagnosis. They are most useful when viewed in conjunction with other statistics.

12.3 Effects on s^2

Armed with the results derived above, we can also ask how the estimator of the variance s^2 is affected by the T-th observation. Belsley, Kuh, and Welsch (1980, p. 64) address this question. Let s^{2^T} denote the estimate from a sample from which the T-th observation is

omitted. After considerable algebraic manipulation, they are able to show that

$$(n-k-1)s^{2^T} = (n-k)s^2 + \frac{\mathbf{e}_T^2}{1-\mathbf{P}_{TT}}.$$

This is a useful result in that it provides a means of *testing* whether the T-th observation $(\mathbf{Y}_T, \mathbf{X}_T)$ is statistically different from the remainder of the observations.

12.4 Testing for Outliers

Recall that if the model is correctly specified, such that observation T obeys the same regression relationship $\mathbf{Y}_T = \mathbf{X}_T'\beta + \epsilon_T$ as the other observations, we have

$$\mathrm{E}\,\mathbf{e}\mathbf{e}' = \mathrm{E}\,\mathbf{M}\epsilon\epsilon'\mathbf{M} = \sigma^2\mathbf{M} = \sigma^2(\mathbf{I}-\mathbf{P}).$$

Therefore $\mathrm{E}\,\mathbf{e}_T^2 = \sigma^2(1-\mathbf{P}_{TT})$. That is, the variance of the least-squares residual will be smaller when that observation exerts greater leverage.

This result provides the foundation for a hypothesis test, the null being that the T-th observation is generated by the same model that generates the remaining data. Given $\hat{\beta}^T$ and s^{2^T}, we calculate the residual $\mathbf{Y}_T - \mathbf{X}_T'\hat{\beta}^T$. Recall that this is equal to $\hat{\alpha}$, the coefficient on $\mathbf{D}_T$ in equation (12.1). We then divide by an estimate of its standard deviation under the null. The test statistic is

$$T = \frac{\hat{\alpha}}{\left(s^{2^T}(1-\mathbf{P}_{TT})\right)^{1/2}}.$$

This should be approximately distributed as $\mathcal{N}(0,1)$ under the null hypothesis.

Chapter 13

Estimator Efficiency

When we search for an "efficient" estimator, we have in mind a set of alternative estimators of a parameter θ, all of which are unbiased, and we aim to determine which one has the smallest variance. For a scalar θ, suppose we have two unbiased estimators $\hat{\theta}$ and $\tilde{\theta}$, with variances $\hat{V}$ and $\tilde{V}$ respectively. If $\hat{V} < \tilde{V}$, then we prefer the $\hat{\theta}$ estimator to the $\tilde{\theta}$ estimator. After all, denoting by θ_0 the true parameter value, Markov's inequality tells us that

$$\Pr\left(|\hat{\theta} - \theta_0| > c\right) \leq \frac{\hat{V}}{c^2} \text{ and } \Pr\left(|\tilde{\theta} - \theta_0| > c\right) \leq \frac{\tilde{V}}{c^2}.$$

The estimator with smaller variance will be less likely to give an estimate further than c from the true value of the parameter.

For vector $\boldsymbol{\theta}$, let $\hat{\mathbf{V}}$ and $\tilde{\mathbf{V}}$ be the respective variance matrices of the two unbiased estimators. Consider the transformations $\boldsymbol{\alpha}'\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\alpha}'\tilde{\boldsymbol{\theta}}$ by which each estimator vector is converted to a scalar random variable using the same vector α of arbitrarily-chosen constants. It is easy to show that

$$\operatorname{Var} \boldsymbol{\alpha}'\hat{\boldsymbol{\theta}} < \operatorname{Var} \boldsymbol{\alpha}'\tilde{\boldsymbol{\theta}}$$

can be written in terms of the quadratic form

$$\boldsymbol{\alpha}'\left(\hat{\mathbf{V}} - \tilde{\mathbf{V}}\right)\boldsymbol{\alpha} < 0.$$

That is, when $\hat{\mathbf{V}} - \tilde{\mathbf{V}}$ is negative definite, $\operatorname{Var} \boldsymbol{\alpha}'\hat{\boldsymbol{\theta}} < \operatorname{Var} \boldsymbol{\alpha}'\tilde{\boldsymbol{\theta}}$ for any $\boldsymbol{\alpha} \neq 0$ that we might choose. When we say that the matrix $\hat{\mathbf{V}}$ is smaller than the matrix $\tilde{\mathbf{V}}$ in the matrix sense, we mean that the difference between the two is negative definite.

13.1 The Gauss–Markov Theorem

You may have heard of the famous *Gauss–Markov theorem*, which states that (under the assumptions we have been using) the ordinary least squares regression estimator $\hat{\boldsymbol{\beta}}$ is the best linear unbiased estimator of $\boldsymbol{\beta}$. To be specific, by "linear," we mean estimators of the form $\tilde{\boldsymbol{\beta}} = \mathbf{W}\mathbf{Y}$ where $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ and the $k \times n$ matrix $\mathbf{W}$ is a function of the $\mathbf{X}$ covariates. By "best," we mean the estimator with smallest variance among the linear estimators that

are unbiased (i.e., all those for which $\mathrm{E}\,\tilde{\boldsymbol{\beta}} = \boldsymbol{\beta}$). That is, $\hat{\boldsymbol{\beta}}$ is the most *efficient* of such linear estimators.

The proof is as follows. Without loss of generality, let us define a candidate linear estimator as $\bar{\boldsymbol{\beta}} = \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{A}\right)\mathbf{Y}$, with the $k \times n$ matrix $\mathbf{A}$ allowed to depend in an unspecified way upon the $\mathbf{X}$ variables. Then the conditional expectation

$$\begin{aligned}
\mathrm{E}\left[\bar{\boldsymbol{\beta}}|\mathbf{X}\right] &= \mathrm{E}\left[\left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{A}\right)(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon})\,|\mathbf{X}\right] \\
&= \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta} + \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}|\mathbf{X}\right] + \mathrm{E}\left[\mathbf{A}\boldsymbol{\epsilon}|\mathbf{X}\right] \\
&= \boldsymbol{\beta} + \mathbf{A}\mathbf{X}\boldsymbol{\beta},
\end{aligned}$$

and clearly we must have $\mathbf{A}\mathbf{X} = \mathbf{0}_{k\times k}$ for unbiasedness. With this restriction, the candidate estimator may be expressed as

$$\bar{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} + \mathbf{A}\boldsymbol{\epsilon},$$

and, taking expectations conditional upon $\mathbf{X}$, we find its variance to be

$$\begin{aligned}
\mathrm{E}(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})(\bar{\boldsymbol{\beta}} - \boldsymbol{\beta})' &= \mathrm{E}\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' + \mathbf{A}\right]\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\left[(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{A}'\right] \\
&= \sigma^2\left[(\mathbf{X}'\mathbf{X})^{-1} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}' + \mathbf{A}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{A}\mathbf{A}'\right] \\
&= \sigma^2\left[(\mathbf{X}'\mathbf{X})^{-1} + \mathbf{A}\mathbf{A}'\right].
\end{aligned}$$

Because $\mathbf{A}\mathbf{A}'$ is positive semidefinite, the variance matrix of $\bar{\boldsymbol{\beta}}$ (conditional upon $\mathbf{X}$) exceeds that of the ordinary least squares estimator $\hat{\boldsymbol{\beta}}$ by a positive semidefinite matrix. As Greene (2003) points out, since the result holds for any $\mathbf{X}$, it also holds unconditionally.

There is another way of expressing this result. Notice that the inefficient estimator $\bar{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \mathbf{A}\boldsymbol{\epsilon}$. It is the sum of the efficient estimator $\hat{\boldsymbol{\beta}}$ and another term, $\mathbf{A}\boldsymbol{\epsilon}$, that turns out to be uncorrelated with the efficient estimator. To see this, take expectations conditional on $\mathbf{X}$,

$$\mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\mathbf{A}\boldsymbol{\epsilon})' = \mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{A}' = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}' = \mathbf{0}.$$

We say that the inefficient estimator $\bar{\boldsymbol{\beta}}$ equals the efficient estimator $\hat{\boldsymbol{\beta}}$ plus uncorrelated "noise," with $\mathbf{A}\boldsymbol{\epsilon}$ representing the noise.

The Gauss–Markov result may seem to be thoroughly definitive. However, it is based on one unspoken assumption that needs further examination: It assumes *no prior knowledge* of any relationship among the $\boldsymbol{\beta}$ parameters. If we do have knowledge of this sort—say, that the true $\boldsymbol{\beta}$ parameters satisfy the relationship $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ with $\mathbf{R}$ a known $m \times k$ matrix of constants and $\mathbf{r}$ a known $m \times 1$ vector—then we can do even better than OLS in terms of efficiency.

Earlier we derived the form of the constrained estimator of $\boldsymbol{\beta}$ given $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$. It is

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\hat{\boldsymbol{\beta}} - \mathbf{r}).$$

Substituting $\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$ for the OLS estimator $\hat{\boldsymbol{\beta}}$, we find after some algebraic manipulation that

$$\mathrm{Var}\,\tilde{\boldsymbol{\beta}}|\mathbf{X} = \sigma^2\left((\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\right).$$

Note that the variance matrix for the constrained estimator equals the OLS variance matrix less a positive definite matrix. In other words, the constrained estimator $\tilde{\boldsymbol{\beta}}$ is more efficient than $\hat{\boldsymbol{\beta}}$ despite the fact that the Gauss–Markov result holds true. Of course, the greater efficiency comes about only if the constraint $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ is actually correct; otherwise the constrained estimator is biased and the question of its efficiency does not arise.

For those of you who are interested, here is how we derived the conditional variance of the constrained estimator. Following the substitution of $\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}$ for the OLS estimator $\hat{\boldsymbol{\beta}}$, and assuming that the constraint $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ is correct, we have

$$\begin{aligned}\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ &= (\mathbf{X}'\mathbf{X})^{-1/2}\left(\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{R}'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1/2}\right)(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\boldsymbol{\epsilon}.\end{aligned}$$

If we let $\mathbf{Z} \equiv (\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{R}'$, then the expression in parentheses is recognizable as $\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, a symmetric idempotent matrix that we will denote by $\tilde{\mathbf{M}}$. Hence,

$$\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1/2}\tilde{\mathbf{M}}(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\boldsymbol{\epsilon}.$$

To find the variance matrix, we then take the expected value of the outer product of this vector conditional on $\mathbf{X}$, which gives

$$\begin{aligned}\mathrm{E}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta})' \mid \mathbf{X} &= \mathrm{E}\left((\mathbf{X}'\mathbf{X})^{-1/2}\tilde{\mathbf{M}}(\mathbf{X}'\mathbf{X})^{-1/2}\mathbf{X}'\boldsymbol{\epsilon}\cdot\boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1/2}\tilde{\mathbf{M}}(\mathbf{X}'\mathbf{X})^{-1/2} \mid \mathbf{X}\right) \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1/2}\tilde{\mathbf{M}}(\mathbf{X}'\mathbf{X})^{-1/2},\end{aligned}$$

and putting $\tilde{\mathbf{M}}$ back into its original form brings us to the final result.

There is one further result that merits discussion. Consider the vector of *contrasts* $\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}$ between the OLS and constrained estimators. If the constraint is correct, then the expected value of the contrast vector is zero and its variance (conditional on $\mathbf{X}$) is

$$\mathrm{Var}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}})|\mathbf{X} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\left(\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}.$$

We see that $\mathrm{Var}(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = \mathrm{Var}\,\hat{\boldsymbol{\beta}} - \mathrm{Var}\,\tilde{\boldsymbol{\beta}}$, that is, the variance of the difference of the estimators equals the difference of their respective variances. This interesting and somewhat surprising result is one example of a larger class of results that we will examine at the end of the chapter.

13.2 Measuring Collinearity

A common frustration in empirical work is that the $\mathbf{X}$ variables entering one's regression can be highly inter-correlated, making it difficult to distinguish the net effect of any one of these variables. Although $\hat{\boldsymbol{\beta}}$ and s^2 remain unbiased, provided that $\mathbf{X}'\mathbf{X}$ is nonsingular, collinearity results in a lack of precision in the estimates of $\boldsymbol{\beta}$.

How does one measure this lack of precision in respect to any one $\hat{\boldsymbol{\beta}}$ coefficient? To study this question, consider the variance of an estimated coefficient, let us say that of $\hat{\boldsymbol{\beta}}_1$. By the FWL theorem, we know that, conditional on $\mathbf{X}$,

$$\operatorname{Var} \hat{\boldsymbol{\beta}}_1 = \frac{\sigma^2}{\mathbf{X}_1' \mathbf{M}_2 \mathbf{X}_1}.$$

In the denominator of this expression is the sum of squared residuals from a diagnostic regression of $\mathbf{X}_1$ on $\mathbf{X}_2$. When $\mathbf{X}_2$ is a good predictor of $\mathbf{X}_1$, yielding a small sum of squared residuals, this inflates the variance of $\hat{\boldsymbol{\beta}}_1$ relative to what it would be if $\mathbf{X}_2$ were not a good predictor of $\mathbf{X}_1$. In the best-case scenario when $\mathbf{X}_2$ has *no* explanatory power for $\mathbf{X}_1$, this variance would be

$$\operatorname{Var} \hat{\boldsymbol{\beta}}_1 = \frac{\sigma^2}{\mathbf{X}_1' \mathbf{X}_1}.$$

A very useful measure of the effects of collinearity is the so-called *variance inflation factor*,

$$\frac{\sigma^2 / \mathbf{X}_1' \mathbf{M}_2 \mathbf{X}_1}{\sigma^2 / \mathbf{X}_1' \mathbf{X}_1} = \frac{\mathbf{X}_1' \mathbf{X}_1}{\mathbf{X}_1' \mathbf{M}_2 \mathbf{X}_1}.$$

Note that if you divide numerator and denominator by $\mathbf{X}_1' \mathbf{X}_1$, the `vif` ratio can be expressed as

$$\frac{\mathbf{X}_1' \mathbf{X}_1}{\mathbf{X}_1' \mathbf{M}_2 \mathbf{X}_1} = \frac{1}{1 - R^2}$$

with the uncentered R^2 coming from a diagnostic regression of $\mathbf{X}_1$ on $\mathbf{X}_2$. The higher is this R^2, the higher is the variance of $\hat{\boldsymbol{\beta}}_1$.

Of course, this `vif` measure cannot be viewed uncritically and should not in general be taken as a trustworthy guide to your model's specification. Suppose, for example, that in estimating a model of wages you choose to include both labor market experience and the square of experience. There are excellent theoretical reasons, described at some length in the labor economics literature, for having both experience variables in the wage model. The `vif` measure will suggest that the inclusion of experience squared substantially inflates the variance of the coefficient on experience alone—but although true, this is the price you must pay in terms of variance in order to stay faithful to a theoretically well-justified specification.

Another problem in interpretation of the `vif` arises when you employ multiple "dummy variables" to handle nominal covariates such as race or region of residence. If a country has 10 regions, for instance, you would typically include dummy variables for 9 of them in your model, leaving the 10th to fall into the "omitted category". In such common specifications, only one of the included dummies can take the value of "1" for any given observation. In a sense, collinearity is built-in to the specification: knowing that one dummy equals 1 implies that the others must be 0. The `vif` will generally indicate a high degree of inflation in such cases, but that is the logical consequence of including dummy variables.

13.3 Adding Irrelevant Explanatory Variables

Is there any real harm in adding explanatory variables to a regression model if their coefficients, in truth, are zero? After all, we know that the OLS estimator is unbiased, so

that the expectation of these estimated coefficients will also be zero. In what sense is there a cost entailed in adding such variables?

The cost is evident in the variance matrix for the *other* estimated coefficients, which is larger (in the matrix sense of a positive semidefinite difference) than it needs to be. Write the model as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}$, and let $\boldsymbol{\delta} = 0$, i.e., the $\mathbf{Z}$ variables are the irrelevant ones. Let us compare two variance matrices, one for the least-squares estimator $\hat{\boldsymbol{\beta}}_0$ taken from a regression of $\mathbf{Y}$ on $\mathbf{X}$ alone, and the other, denoted simply $\hat{\boldsymbol{\beta}}$, from the regression of $\mathbf{Y}$ on both $\mathbf{X}$ and the irrelevant $\mathbf{Z}$.

By the Frisch–Waugh–Lovell (FWL) theorem, we have

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{Y}$$

for $\mathbf{M}_\mathbf{Z} = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, with $\mathbf{M}_\mathbf{Z}$ being a symmetric, idempotent matrix with the property that $\mathbf{M}_\mathbf{Z}\mathbf{Z} = \mathbf{0}$. Substituting for $\mathbf{Y}$, we obtain

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{Z}\boldsymbol{\epsilon},$$

and given the assumption $\mathrm{E}[\boldsymbol{\epsilon}|\mathbf{X}, \mathbf{Z}] = \mathbf{0}$, we have $\mathrm{E}\,\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}$. As for the variance matrix,

$$\mathrm{Var}\,\hat{\boldsymbol{\beta}} = \mathrm{E}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})' = \mathrm{E}(\mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{Z}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{M}_\mathbf{Z}\mathbf{X}(\mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X})^{-1}.$$

Using iterated expectations and assuming $\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}'|\mathbf{X}, \mathbf{Z}] = \sigma^2\mathbf{I}$, we obtain $\mathrm{Var}\,\hat{\boldsymbol{\beta}} = \sigma^2\,\mathrm{E}(\mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X})^{-1}$. This may be compared with the variance matrix for $\hat{\boldsymbol{\beta}}_0$ when $\mathbf{Z}$ is omitted from the model, $\mathrm{Var}\,\hat{\boldsymbol{\beta}}_0 = \sigma^2\,\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}$.

Let's now condition on both $\mathbf{X}$ and $\mathbf{Z}$. We'll show that the difference

$$(\mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}$$

is positive semidefinite. This is true if and only if the difference $\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X}$ is positive semidefinite. (See the Economics 590 notes on linear algebra for the proof.) Rewriting, we have

$$\mathbf{X}'\mathbf{X} - \mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X} = \mathbf{X}'(\mathbf{I} - \mathbf{M}_\mathbf{Z})\mathbf{X} \equiv \mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X},$$

with $\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ also being symmetric idempotent and thus positive semidefinite. It follows that $\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X}$ is positive semidefinite.

In summary, adding irrelevant explanatory variables $\mathbf{Z}$ to a regression model leaves unbiased the estimator $\hat{\boldsymbol{\beta}}$ associated with the $\mathbf{X}$ variables, but inflates the variance of $\hat{\boldsymbol{\beta}}$ so that this estimator is "noisier" than it needs to have been.

13.4 "Over-fitting" and Split-Sample Strategies

A 12 December 2014 post to stats.stackexchange.com explained one part of the issue very well:

> [Overfitting] happens every time you throw all the potential predictors into a regression model, should any of them in fact have no relationship with the response once the effects of others are partialled out.

> With this type of overfitting, the good news is that inclusion of these irrelevant terms does not introduce bias [into] your estimators, and in very large samples the coefficients of the irrelevant terms should be close to zero. But there is also bad news: because the limited information from your sample is now being used to estimate more parameters, it can only do so with less precision—so the standard errors on the genuinely relevant terms increase. That also means they're likely to be further from the true values than estimates from a correctly specified regression, which in turn means that if given new values of your explanatory variables, the predictions from the overfitted model will tend to be less accurate than for the correctly specified model.

Indeed, as we saw just above, adding irrelevant explanatory variables reduces the efficiency of the $\hat{\boldsymbol{\beta}}$ estimator, causing its variance matrix to be larger (in the matrix sense) that it needs to be. In addition to inflating the standard errors of the estimated coefficients, the addition of irrelevant variables also inflates the variance of *predictions* based on the estimated model.

To see this more formally, suppose that you have in hand the $\hat{\boldsymbol{\beta}}$ vector estimated from the original sample, and that you are given one new Y_f case and a new vector of its associated covariates $\mathbf{X}_f$ to use in assessing the forecast performance of your model. Conditional on $\mathbf{X}_f = \mathbf{x}_f$, the expected value of Y_f is $\mathbf{x}_f'\boldsymbol{\beta}$, which is an unbiased prediction of Y_f. However, the variance of this prediction (conditional on the original $\mathbf{X}$ and also on $\mathbf{X}_f = \mathbf{x}_f$ for the new observation) is

$$\text{Var}\left(\mathbf{x}_f'\hat{\boldsymbol{\beta}}\right) = \text{E}\,\mathbf{x}_f'(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{x}_f = \mathbf{x}_f'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_f$$

and much as with the OLS variance matrix, the prediction variance is also larger than it needs to be.

But how can you *know* whether you have "over-fit" a model? Because you don't know the true model—the model of the true data-generating process—you cannot be absolutely sure whether you have over-fit the model or not. You may have been so preoccupied with the consequences of omitting a relevant explanatory variable, that you have gone too far in the other direction and overloaded the specification with irrelevant variables. There are fundamental tradeoffs to think about here, and it is far from obvious how you should proceed.

An increasingly popular strategy in applied economics to help in deciding about such specification issues is the "split-sample" strategy, whereby the full sample is divided into one sub-sample for estimating the model and another for testing it. If you have enough data to employ this strategy, here's how it can help.[1]

One symptom of over-fitting, is that by adding a number of irrelevant explanatory variables to the specification, you increase the fit of the model to the original data (as evidenced in the R^2) while also increasing the prediction variance. Looking at the original data only (the "estimation sample"), you might observe a tight fit between the original Y_i

[1] In micro-econometric work, you would use a random-number generator to identify observations for the estimation and testing samples so that the samples show no systematic difference in composition. In time-series settings, by contrast, it is common to divide the samples according to time, into a "before" sample for model estimation and an "after" sample for assessing the forecasts. This before-and-after practice is vulnerable to the emergence of new time trends over the course of the "after" sample.

and $\mathbf{X}_i\hat{\boldsymbol{\beta}}$ because you have added enough irrelevant variables to substantially jack up the R^2 on the estimation sample. But in doing so, you may have increased the prediction variance to the point that the prediction $\mathbf{x}_f'\hat{\boldsymbol{\beta}}$ becomes quite noisy, seemingly too noisy to be useful as a guide to the value of Y_f. In a testing sample, the point-predicted values might well appear to be "all over the map" in relation to the Y_f values in the testing sample. Even so, there should be no *systematic* tendency evident in the testing sample for the predicted values to exceed or fall short of Y_f.

But another possibility is that in bringing many additional variables into the specification, you may have inadvertently introduced some variables that are correlated with the disturbance term. If this has happened, then the OLS estimator is no longer unbiased, and its predicted values will also be systematically off the mark. You might see symptoms of this problem if you have a good-sized set of additional $(Y_i, \mathbf{X}_i)$ observations in the testing sample and find that your predicted values seem to consistently undershoot or overshoot the testing sample Y's.

13.5 The Cramér–Rao Lower Bound on Variances

What is the theoretical minimum variance—that is, the greatest efficiency—that can be achieved by an estimator? We'll restrict ourselves to the case of a model with parameter vector θ, for which we have an estimator $\hat{\theta}$ that is unbiased, i.e., $\mathrm{E}\,\hat{\theta} = \theta_0$, in which we denote by θ_0 the true value of the θ parameter. (See Mittelhammer (1996, pp. 408–417) for discussion of cases in which $\hat{\theta}$ is biased.) In the simplest case, the $k \times 1$ estimator $\hat{\theta}$ depends on an $n \times 1$ data vector $Y = (Y_1, \ldots, Y_n)'$, and we let $L^*(y|\theta)$ represent the joint density function of Y given θ. With $\hat{\theta}$ being unbiased, the condition $\mathrm{E}\,\hat{\theta} = \theta_0$ can be expressed as

$$\int \hat{\boldsymbol{\theta}} L^*(y|\boldsymbol{\theta}_0)\, dy = \boldsymbol{\theta}_0, \tag{13.1}$$

in which the single integral sign is to be understood to represent multiple integrals over the components of Y, and dy is likewise a shorthand form of $dy_1\, dy_2 \cdots dy_n$. Note as well that

$$\int L^*(y|\boldsymbol{\theta})\, dy = 1 \tag{13.2}$$

as is the case with all joint densities.

Taking partial derivatives of equation (13.2) with respect to the k elements of $\boldsymbol{\theta}$—as discussed in the Economics 590 supplementary notes, we need to invoke regularity conditions that permit differentiation under the integral sign, and the range of Y must not depend on the $\boldsymbol{\theta}$ parameter—we obtain the $k \times 1$ vector of derivatives

$$\int \frac{\partial L^*(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} dy = \mathbf{0}_{k\times 1}.$$

Cleverly re-writing the derivative vector, we obtain

$$\int \left(\frac{1}{L^*(y|\boldsymbol{\theta})} \frac{\partial L^*(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \cdot L^*(y|\boldsymbol{\theta})\, dy = \mathbf{0}.$$

If we define $L(y|\boldsymbol{\theta})$ to be the log of the joint density L^*, this last expression can be rewritten again, as

$$\int \left(\frac{\partial L(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \cdot L^*(y|\boldsymbol{\theta})\, dy = \mathbf{0}. \tag{13.3}$$

If we now set $\boldsymbol{\theta} = \boldsymbol{\theta}_0$, which is the true value of the $\boldsymbol{\theta}$ parameter, we can express the result even more compactly as

$$\mathrm{E}\, \frac{\partial L(Y|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} = \mathbf{0}. \tag{13.4}$$

This is itself an important result in the context of maximum likelihood estimation. In that context, we refer to equation (13.4) as saying that when evaluated at the true $\boldsymbol{\theta}_0$, the *score vector* $\partial L(Y|\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$ has expectation zero. A class of statistical tests—termed Lagrange Multiplier tests—forms the sample counterpart to the score vector under the null hypothesis and assesses whether there is evidence against the proposition of a zero mean, which would indicate that the null is false.

Returning to equation (13.3), let us differentiate it again with respect to $\boldsymbol{\theta}$, obtaining a $k \times k$ matrix of expressions on the left-hand side that equal a $k \times k$ matrix of zeroes on the right-hand side,

$$\int \left[\frac{\partial^2 L(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} + \left(\frac{\partial L(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial L(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right)' \right] \cdot L^*(y|\boldsymbol{\theta})\, dy = \mathbf{0}_{k \times k}. \tag{13.5}$$

Remembering that $\mathrm{E}\, \partial L(Y|\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta} = \mathbf{0}$, when we set $\boldsymbol{\theta} = \boldsymbol{\theta}_0$ we note that the expectation of the outer product,

$$\mathrm{E} \left(\frac{\partial L(Y|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial L(Y|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right)',$$

is simply the variance matrix of $\partial L(Y|\boldsymbol{\theta}_0)/\partial \boldsymbol{\theta}$. This is the famous result for the variance of the score vector,

$$\begin{aligned} \mathrm{Var}\, \frac{\partial L(Y|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} &= \mathrm{E} \left(\frac{\partial L(Y|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right) \left(\frac{\partial L(Y|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} \right)' && (13.6) \\ &= -\mathrm{E}\, \frac{\partial^2 L(Y|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, && (13.7) \end{aligned}$$

which will also figure prominently in the analysis of maximum-likelihood methods of estimation. The variance matrix of the score is termed the *information matrix*.

With all this in hand, let us narrow our focus to the special case in which θ is a scalar, and determine the smallest variance that can be exhibited by any unbiased estimator $\hat{\theta}$. After we obtain this lower bound, we will generalize things and again allow $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ to be $k \times 1$ vectors.

Return to equation (13.1) defining unbiasedness and take its derivative with respect to θ_0, noting that although Y can appear in the estimator $\hat{\theta}$, the parameter θ itself cannot. If the regularity conditions hold, we obtain

$$\int \hat{\theta} \frac{dL^*(y|\theta_0)}{d\theta}\, dy = 1$$

and, upon re-writing this in a now-familiar fashion, have

$$\int \hat{\theta} \frac{dL(y|\theta_0)}{d\theta} L^*(y|\theta_0)\, dy = 1.$$

Because $\mathrm{E}\, dL(Y|\theta_0)/d\theta = 0$, the equation above can be summarized as saying

$$\mathrm{Cov}\left(\hat{\theta}, \frac{dL(Y|\theta_0)}{d\theta}\right) = 1.$$

Applying the Cauchy–Schwartz inequality,

$$1 = \mathrm{Cov}\left(\hat{\theta}, \frac{dL(Y|\theta_0)}{d\theta}\right) \leq \sqrt{\mathrm{Var}\,\hat{\theta}} \cdot \sqrt{\mathrm{Var}\,\frac{dL(Y|\theta_0)}{d\theta}}$$

or

$$1 \leq \mathrm{Var}\,\hat{\boldsymbol{\theta}} \cdot \mathrm{Var}\,\frac{dL(Y|\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}.$$

This brings us, finally, to the celebrated Cramér–Rao result:

$$\mathrm{Var}\,\hat{\theta} \geq \frac{1}{\mathrm{Var}\,\frac{dL(Y|\theta_0)}{d\theta}}. \tag{13.8}$$

In other words, the inverse of the variance of the score $dL(Y|\theta_0)/d\theta$ establishes a lower bound on the variance of *any* unbiased estimator of θ, at least under the regularity conditions that were mentioned above.

Let's summarize result (13.8) using the notation $V_{\hat{\theta}} \geq V_S^{-1}$ for the two variances, and go on to consider the case of a $k \times 1$ vector $\boldsymbol{\theta}$. The extension proceeds as follows. Consider the first element of the k equations $\int \hat{\boldsymbol{\theta}} L^*(y|\boldsymbol{\theta})\, dy = \boldsymbol{\theta}_0$, which is

$$\int \hat{\boldsymbol{\theta}}_1 L^*(y|\boldsymbol{\theta})\, dy = \boldsymbol{\theta}_1,$$

and differentiate this with respect to all k elements of $\boldsymbol{\theta}$. For $\boldsymbol{\theta}_1$ itself we obtain

$$\int \hat{\boldsymbol{\theta}}_1 \frac{\partial L(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} L^*(y|\boldsymbol{\theta})\, dy = 1,$$

but for all $\boldsymbol{\theta}_j$ with $j \geq 2$,

$$\int \hat{\boldsymbol{\theta}}_1 \frac{\partial L(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} L^*(y|\boldsymbol{\theta})\, dy = 0.$$

Because $\mathrm{E}\, \partial L(Y|\boldsymbol{\theta})/\partial \boldsymbol{\theta}_j = 0$ for all $j = 1, \ldots, k$, we can write the equation

$$\int \hat{\boldsymbol{\theta}}_1 \frac{\partial L(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} L^*(y|\boldsymbol{\theta})\, dy = 1$$

in terms of a covariance, as

$$\mathrm{Cov}\left(\hat{\boldsymbol{\theta}}_1, \frac{\partial L(Y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\right) = 1,$$

and, similarly, the equation

$$\int \hat{\boldsymbol{\theta}}_1 \frac{\partial L(y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j} L^*(y|\boldsymbol{\theta})\, dy = 0$$

can be written as

$$\text{Cov}\left(\hat{\boldsymbol{\theta}}_1, \frac{\partial L(Y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_j}\right) = 0.$$

We have been focusing attention only on the first element of the $\boldsymbol{\theta}$ vector, but this reasoning applies to the other elements of the vector as well. Hence, all these results can be summarized in a $k \times k$ matrix,

$$\text{Cov}\left(\hat{\boldsymbol{\theta}}, \frac{\partial L(Y|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right) = \mathbf{I}_{k\times k}.$$

Now consider the full variance matrix of the $2k \times 1$ vector $(\hat{\boldsymbol{\theta}}, \partial L/\partial \boldsymbol{\theta})'$, which can be written in the form

$$\text{Var}\begin{bmatrix} \hat{\boldsymbol{\theta}} \\ \frac{\partial L}{\partial \boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} \mathbf{V}_{\hat{\theta}} & \mathbf{I}_k \\ \mathbf{I}_k & \mathbf{V}_S \end{bmatrix}.$$

As we know, a variance matrix must be positive semi-definite. Let $\mathbf{a}$ be an arbitrary $k \times 1$ vector and consider the quadratic form

$$\begin{bmatrix} \mathbf{a}' & -\mathbf{a}'\mathbf{V}_S^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{V}_{\hat{\theta}} & \mathbf{I}_k \\ \mathbf{I}_k & \mathbf{V}_S \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ -\mathbf{V}_S^{-1}\mathbf{a} \end{bmatrix} = \begin{bmatrix} \mathbf{a}'\mathbf{V}_{\hat{\theta}} - \mathbf{a}'\mathbf{V}_S^{-1} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ -\mathbf{V}_S^{-1}\mathbf{a} \end{bmatrix} \geq 0,$$

or simply

$$\mathbf{a}' \left(\mathbf{V}_{\hat{\theta}} - \mathbf{V}_S^{-1}\right) \mathbf{a} \geq 0.$$

This proves that $\mathbf{V}_{\hat{\theta}}$, the variance matrix of $\hat{\boldsymbol{\theta}}$, must be greater than or equal to $\mathbf{V}_S^{-1}$ in the matrix sense of inequality. This is the generalization we have been seeking: The lower bound on the variance of an unbiased estimator is the inverse of the information matrix.

When we engage with maximum likelihood estimation methods, we will be using the Cramér–Rao results extensively. There we will be mainly concerned with data series that are either iid or inid (independent but not identically distributed). In these cases independence allows us to write $L(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ln f_i(y_i|\boldsymbol{\theta})$ and

$$\frac{\partial L}{\partial \boldsymbol{\theta}} = \sum_i \frac{\partial \ln f_i(y_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_i \frac{\partial \ln f_i}{\partial \boldsymbol{\theta}}.$$

By the same logic (and under the same regularity conditions) as used to develop the Cramér–Rao results, we have $\text{E}\,\partial \ln f_i/\partial \boldsymbol{\theta} = \mathbf{0}$, and note that the $k \times k$ score outer product matrix is

$$\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)' = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial \ln f_i}{\partial \boldsymbol{\theta}} \frac{\partial \ln f_j}{\partial \boldsymbol{\theta}}'.$$

Since the data series are independent and each term in the double sum is the (outer) product of two mean-zero random vectors, only n terms remain when expectations are taken, giving

$$\mathrm{E}\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)' = \sum_{i=1}^{n} \mathrm{E}\,\frac{\partial \ln f_i}{\partial \boldsymbol{\theta}}\frac{\partial \ln f_i}{\partial \boldsymbol{\theta}}'.$$

Again applying the logic developed above,

$$\mathrm{E}\,\frac{\partial \ln f_i}{\partial \boldsymbol{\theta}}\frac{\partial \ln f_i}{\partial \boldsymbol{\theta}}' = -\,\mathrm{E}\,\frac{\partial^2 \ln f_i}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} \equiv \mathcal{J}_i,$$

where the i subscript in $\mathcal{J}_i$ reminds us that the expected values can depend on i in the case of inid data series. Hence, the information matrix $\mathcal{R}$ is

$$\mathcal{R} = \mathrm{E}\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)\left(\frac{\partial L}{\partial \boldsymbol{\theta}}\right)' = \sum_{i=1}^{n} \mathcal{J}_i$$

for inid data series, and this simplifies to $\mathcal{R} = n \cdot \mathcal{J}$ for iid data.

An example may help to clarify these ideas. Suppose that we have an iid sample with n observations on the random variable Y_i, which is exponentially distributed with parameter r. Recall that $\mathrm{E}\,Y_i = 1/r$. Then

$$L^*(r) = r^n e^{-r\sum_{i=1}^{n} Y_i},$$

and

$$L(r) = n\ln r - r\sum_{i=1}^{n} Y_i.$$

The condition that $\mathrm{E}\,\partial L(r_0)/\partial r = 0$ is, in this case,

$$\mathrm{E}\left(\frac{n}{r_0} - \sum_{i=1}^{n} Y_i\right) = 0,$$

and the variance of $\partial L(r_0)/\partial r$ is $-\,\mathrm{E}\,\partial^2 L(r_0)/\partial r^2 = n/r_0^2$. Hence the smallest variance that any unbiased estimator of r can achieve is r_0^2/n.

13.6 Efficient versus Unbiased Estimators

Earlier we mentioned that an inefficient estimator of a parameter vector $\boldsymbol{\theta}$ can be written in terms of an efficient estimator plus some "noise" that is uncorrelated with the efficient estimator. We will prove the result in this section. But why is it important? The relationship between one estimator of $\boldsymbol{\theta}$ that is unbiased and efficient, and another that is simply unbiased, gives rise to an interesting family of test statistics termed *Hausman tests* after the famous paper by Hausman (1978). The idea behind these tests (which actually dates to work twenty years earlier by Durbin) is to assess the statistical significance of the difference between two estimators of $\boldsymbol{\theta}$ based on the same dataset. One of them is $\hat{\boldsymbol{\theta}}_e$, a normally-distributed estimator that is efficient under the null hypothesis but biased under the alternative hypothesis. The other is $\hat{\boldsymbol{\theta}}_u$, another normally-distributed estimator that

is unbiased under the null (although inefficient) but which maintains its unbiasedness under the alternative hypothesis. The Hausman test is based on the contrast between these estimators, i.e., on the vector of differences $(\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e)$. We saw an example of this earlier when we considered the constrained estimator $\tilde{\boldsymbol{\beta}}$ under the constraint $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ and the unconstrained estimator $\hat{\boldsymbol{\beta}}$, the first of these being unbiased and efficient but only when the assumption $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ is correct, and the second being unbiased whether or not that assumption holds.

The Hausman test is mainly applied with large-sample asymptotics in mind, using central limit theorems to prove that the two estimators (when suitably transformed) are normally distributed, and replacing the small-sample concept of unbiasedness by the assumption of consistency. Here we will explain the essence of the method using small-sample theory, simply assuming that the two estimators of $\boldsymbol{\theta}$ are both normally distributed.

Specifically, let us assume that under a null hypothesis that is to be tested (such as $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$), we have $\mathrm{E}\,\hat{\boldsymbol{\theta}}_e = \mathrm{E}\,\hat{\boldsymbol{\theta}}_u = \boldsymbol{\theta}$ and assume that both estimators are multivariate normal, with covariance matrices $\mathbf{V}_e$ and $\mathbf{V}_u$ respectively. We'll show that under the null, the variance matrix of the contrast vector $(\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e)$ is simply $\mathbf{V}_u - \mathbf{V}_e$. Thus, if the variance matrices were known and we were to proceed to construct a quadratic form

$$(\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e)' \left[\mathbf{V}_u - \mathbf{V}_e\right]^{-1} (\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e) \sim \chi^2_k,$$

we would have a basis for *testing* whether the differences in the two estimators are of statistical significance. We defer further discussion of the intricacies of the testing procedure to Chapter 30.

The key to the proof is to show that the two vectors $(\hat{\boldsymbol{\theta}}_e - \boldsymbol{\theta})$ and $(\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e)$ have zero covariance. To prove this, we begin with the identity $\hat{\boldsymbol{\theta}}_u = \hat{\boldsymbol{\theta}}_e + (\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e)$. The variance matrix is

$$\begin{aligned} \mathbf{V}_u &\equiv \mathrm{Var}\left((\hat{\boldsymbol{\theta}}_e - \boldsymbol{\theta}) + (\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e)\right) \\ &= \mathbf{V}_e + \mathbf{C} + \mathbf{C}' + \mathbf{V}_{ue} \end{aligned} \tag{13.9}$$

in which $\mathbf{V}_{ue} = \mathrm{Var}((\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e))$ is the $k \times k$ variance matrix of the contrast vector and the covariance matrix $\mathbf{C}$ is defined as

$$\mathbf{C} = \mathrm{E}(\hat{\boldsymbol{\theta}}_e - \boldsymbol{\theta})(\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e)'.$$

We aim to show that the $k \times k$ covariance matrix $\mathbf{C}$ is a zero matrix. Once this is proven, we see that

$$\mathbf{V}_u = \mathbf{V}_e + \mathbf{V}_{ue},$$

and the variance of the contrast vector is therefore $\mathbf{V}_{ue} = \mathbf{V}_u - \mathbf{V}_e$.

We already know that under the null, $\hat{\boldsymbol{\theta}}_e$ is efficient, that is, no other estimator can boast of a variance matrix $\mathbf{V}$ such that $\mathbf{V} < \mathbf{V}_e$ in the matrix sense. The proof that $\mathbf{C} = \mathbf{0}$ proceeds from this fact. Consider a new estimator $\bar{\boldsymbol{\theta}}$ formed as follows,

$$\bar{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_e + r\mathbf{A}(\hat{\boldsymbol{\theta}}_u - \hat{\boldsymbol{\theta}}_e),$$

where r is a scalar and $\mathbf{A}$ is a matrix whose elements are to be specified shortly. By construction, the new estimator $\bar{\boldsymbol{\theta}}$ is unbiased under the null. Subtracting $\boldsymbol{\theta}$ from both sides,

we find that the variance of $\bar{\boldsymbol{\theta}}$ is given by $\mathbf{V}(r) = \mathbf{V}_e + r\mathbf{C}\mathbf{A}' + r\mathbf{A}\mathbf{C}' + r^2\mathbf{A}\mathbf{V}_{ue}\mathbf{A}'$, which is a (matrix) function of the scalar r. Consider the derivative $\mathbf{V}'(r)$ evaluated at $r = 0$, that is, $\mathbf{V}'(0) = \mathbf{C}\mathbf{A}' + \mathbf{A}\mathbf{C}'$. The matrix $\mathbf{V}'(0)$ must be either a zero matrix or a positive definite matrix, as otherwise a small r could be found that would make the variance of $\bar{\boldsymbol{\theta}}$ smaller than the variance of the efficient estimator $\hat{\boldsymbol{\theta}}_e$, which would be impossible by definition. Now choose $\mathbf{A} = -\mathbf{C}$. For this choice of $\mathbf{A}$ we have $\mathbf{V}'(0) = -2\mathbf{C}\mathbf{C}'$, which is negative definite—a contradiction—unless $\mathbf{C} = \mathbf{0}$.

We thus conclude that $\mathbf{C} = \mathbf{0}$. Returning to equation (13.9) above, we see that this implies $\mathbf{V}_{ue} = \mathbf{V}_u - \mathbf{V}_e$. Writing this result as $\mathbf{V}_u = \mathbf{V}_e + \mathbf{V}_{ue}$, we can say that (under the null) the unbiased but inefficient estimator $\hat{\boldsymbol{\theta}}_u$ can be written as $\hat{\boldsymbol{\theta}}_u = \hat{\boldsymbol{\theta}}_e + \eta$ where η is an vector of "noise" (with variance $\mathbf{V}_{ue}$) that is uncorrelated with the efficient estimator $\hat{\boldsymbol{\theta}}_e$.

Part II

Asymptotic Analysis

Chapter 14

Laws of Large Numbers

An exploration of asymptotic theory in econometrics could easily occupy a course in itself. Fortunately, we require only a few of the basic results. This chapter will develop the background you need to understand laws of large numbers, applying these laws to the ordinary least squares estimators of the slope $\boldsymbol{\beta}$ and variance σ^2 parameters of the linear regression model. Accessible treatments of this material are available in Mittelhammer (1996, Chapter 5), Mittelhammer, Judge, and Miller (2000), Davidson and MacKinnon (1993, Chapter 4), and Ruud (2000). Higher-level treatments can be found in Bierens (2004) and Davidson (1994). For further background, I would recommend Greenberg and Webster (1983, Chapter 1), McFadden (1988), Serfling (1980), and Spanos (1986).

Although this material is a step up in level of difficulty by comparison with the techniques we've been using so far, it is actually of greater practical value to the applied econometrician. In day-to-day empirical research, you are likely to be involved in problems that require estimation methods and (especially) hypothesis tests whose justifications rest on the kinds of large-sample arguments that we will be developing. To give you a sense of the differences in approach, the top panel of Table 14.1 summarizes where we have been so far, and the bottom panel indicates where the asymptotic tools will be taking us.

Table 14.1: Overview for the Linear Model $Y_i = \mathbf{X}_i'\boldsymbol{\beta} + \epsilon_i$

Assumptions on $\boldsymbol{\epsilon}$ and $\mathbf{X}$	Distributions and Tests
Where we've been: We've assumed $\mathrm{E}(\epsilon \mid \mathbf{X}) = \mathbf{0}$ and $\mathrm{E}(\epsilon\epsilon'\|\mathbf{X}) = \sigma^2\mathbf{I}$. These strong assumptions yield strong results: $\hat{\boldsymbol{\beta}}$ is unbiased for $\boldsymbol{\beta}$ as is s^2 for σ^2, and $\mathrm{Var}\,\hat{\boldsymbol{\beta}} = \sigma^2\,\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}$. Furthermore, $\hat{\boldsymbol{\beta}}$ is best linear unbiased. Under the more general assumption $\mathrm{E}(\epsilon\epsilon'\|\mathbf{X}) = \mathbf{V}$, the OLS estimator $\hat{\boldsymbol{\beta}}$ remains unbiased, but its variance conditional on $\mathbf{X}$ is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$. Because $\mathbf{V}$ is unknown, however, we have not done much of anything with this result.	We've assumed $\epsilon \mid \mathbf{X} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ From normality we obtain further strong results: $\hat{\boldsymbol{\beta}} \mid \mathbf{X} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$, and $(n-k)s^2/\sigma^2 \mid \mathbf{X} \sim \chi^2_{n-k}$. Also, $\hat{\boldsymbol{\beta}}$ and s^2 are independent conditional on $\mathbf{X}$. These assumptions provide the justification for using t and $\mathcal{F}$ tests in hypothesis-testing. Note that as yet, we have no testing method available for $\boldsymbol{\beta}$ that does not rely on both normality and homoskedasticity assumptions.
Where we're going: We will assume $\mathrm{E}(\epsilon_i \mid \mathbf{X}_i) = 0$, and when a homoskedasticity assumption is needed, $\mathrm{E}(\epsilon_i^2 \mid \mathbf{X}_i) = \sigma^2$. These are much weaker and more acceptable assumptions about the relationship between the disturbances and the explanatory covariates. Under some mild additional conditions, *laws of large numbers* can be applied to show *consistency*, that is, to show that $\hat{\boldsymbol{\beta}}$ converges in probability to the true $\boldsymbol{\beta}$ as does s^2 to the true σ^2 value. Under these assumptions, however, we are *not* generally able to prove unbiasedness. If $\mathrm{E}(\epsilon_i^2 \mid \mathbf{X}_i) = \sigma_i^2$, we can also show that $\hat{\boldsymbol{\beta}}$ converges in probability to $\boldsymbol{\beta}$ (under additional conditions that are not unreasonable) .	No assumption on the distribution of ϵ will be needed, except in the case of the maximum-likelihood approach, which requires *some* assumption but does *not* require normality. Instead, we apply *central limit theorems* to show that $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ converges in distribution to normal, as does $\sqrt{n}(s^2 - \sigma^2)$ when the disturbance term is homoskedastic. These results provide us with the justification for using χ^2 tests for hypothesis-testing, which are easily modified to handle heteroskedastic disturbances.

14.1 Convergence Concepts

To begin our study of asymptotics, in this section we consider two kinds of convergence involving sequences of random variables: convergence in probability and (what amounts to a special case in our applications) convergence in mean of order r. A third and very different type of convergence—known as convergence in distribution—is addressed in the next chapter.

To understand convergence in probability, it is helpful to recall what we mean by the *limit* of a sequence of real numbers. We say that $\lim_{n\to\infty} x_n = c$ if for any positive number $\epsilon > 0$, no matter how small, there exists an integer N_ϵ with this property: for $n > N_\epsilon$, we have $c - \epsilon < x_n < c + \epsilon$. We can write this as $\mid x_n - c \mid < \epsilon$ for $n > N_\epsilon$. That is, we can construct a band of arbitrarily small width centered around c, and examine whether all x_n with $n > N_\epsilon$ fall in that band. If they do, and if this is true for all $\epsilon > 0$, then c is the limit of the sequence.

Convergence in Probability

How can we generalize the concept of limit to apply to $\{X_n\}$, a sequence of random variables? The trick is to use a function that effectively converts the distribution of each random variable in the sequence to a single real number. We say that $\{X_n\}$ converges in probability to a constant c if, for all values $\epsilon > 0$,

$$\lim_{n\to\infty} \Pr(\mid X_n - c \mid < \epsilon) = 1.$$

That is, the probability that X_n falls *inside* the band $[c - \epsilon, c + \epsilon]$ approaches 1 as n grows large, and this must be true no matter what $\epsilon > 0$ value we choose. Imagining very small values of ϵ, hardly bigger than zero, this can be visualized as the distribution of X_n collapsing on the point c as n grows large.

In formal proofs, this definition is formalized with a second tolerance value $\delta > 0$. Letting $p_n = \Pr(\mid X_n - c \mid < \epsilon)$, the limit of interest for the $\{p_n\}$ sequence is 1, and since we must have $p_n \leq 1$, to investigate the limit we only need to consider the half-band extending down from 1 to $1 - \delta$ for some arbitrarily small $\delta > 0$. If for every $\delta > 0$ and $\epsilon > 0$, an integer $N_{\epsilon,\delta}$ exists such that for all $n > N_{\epsilon,\delta}$ we have $p_n > 1 - \delta$, then c is the probability limit of the $\{X_n\}$ sequence. Note that *two tolerance values*, ϵ and δ, are involved in this definition.

Often the probability limit is written in an equivalent form,

$$\lim_{n\to\infty} \Pr(\mid X_n - c \mid > \epsilon) = 0$$

for any $\epsilon > 0$, no matter how small. That is, the probability that X_n falls *outside* the band $[c - \epsilon, c + \epsilon]$ approaches zero as n grows large, for any positive value of ϵ that we might choose.[1]

In the literature, convergence in probability is generally defined in the following compressed fashion. The sequence $\{X_n\}$ converges in probability to c if, for any $\epsilon > 0$ and $\delta > 0$,

[1]Note that instead of writing $\mid X_n - c \mid > \epsilon$, we could have written $\mid X_n - c \mid \geq \epsilon$ without affecting the definition in any significant way. This is because the condition must hold for all $\epsilon > 0$.

there exists an integer that is a function of both of them, $N_{\epsilon,\delta}$, such that for all $n > N_{\epsilon,\delta}$, $\Pr(\mid X_n - c \mid > \epsilon) < \delta$. Alternatively, we can write the definition in these terms: for all $n > N_{\epsilon,\delta}$, we have $\Pr(\mid X_n - c \mid < \epsilon) > 1 - \delta$.

Convergence in probability to c is commonly denoted by

$$X_n \xrightarrow{p} c$$

and

$$\operatorname{plim} X_n = c,$$

where the term "plim" is shorthand for the phrase "probability limit." In fact, "plim" is something of a misnomer, in that the concept is better described by "limp." Nevertheless, the usage persists. As will be seen below, we use plims in econometrics when we examine the consistency of an estimator. We also use them on our way to establishing the large-sample distributions of estimators and test statistics, as our next chapter will show.

Sometimes we speak of $\{X_n\}$ converging in probability to a random variable Y, denoted by $X_n \xrightarrow{p} Y$ in what might be viewed as slightly confusing notation. The phrase "convergence to a random variable" actually means that the sequence of random variables $Z_n \equiv X_n - Y$ converges in probability to the constant $c = 0$. That is,

$$\lim_{n\to\infty} \Pr(\mid X_n - Y \mid > \epsilon) = 0 \quad \forall\, \epsilon > 0.$$

To understand the idea, consider making a random draw from the distribution of Y and a sequence of such draws from the distributions of X_n for $n = 1, 2, \ldots$. As n becomes large, the chance that X_n differs from Y by more than ϵ becomes vanishingly small, no matter how small we choose ϵ itself to be. In a sense, the random variables X_n and Y become less and less distinguishable as n goes to infinity. Or, we could say that X_n becomes a better and better predictor of Y (and vice versa) as $n \to \infty$. Mittelhammer (1996, p. 242) elaborates on this point.

In the above, we might have considered two sequences $\{X_n\}$ and $\{Y_n\}$. If the difference between X_n and Y_n converges in probability to zero, i.e.,

$$\lim_{n\to\infty} \Pr(\mid X_n - Y_n \mid > \epsilon) = 0 \quad \forall\, \epsilon > 0$$

then $\{X_n\}$ and $\{Y_n\}$ are termed *asymptotically equivalent* and we denote this by $X_n \overset{a}{=} Y_n$. Note that asymptotic equivalence can obtain even if neither $\{X_n\}$ nor $\{Y_n\}$ has a probability limit as such, so long as their *difference* converges in probability to zero. However, if $X_n \overset{a}{=} Y_n$ and $Y_n \xrightarrow{p} Y$, then $X_n \xrightarrow{p} Y$ as well (Fuller 1976, pp. 186–187).

The probability limits of random vectors are examined in two ways. Some authors inspect the behavior of each element of a random vector, saying for example that the sequence of random vectors $\mathbf{X}_n$ converges in probability to the random vector $\mathbf{Y}$ if and only if this is true for all elements $X_{n,j}$ and Y_j. Others replace the absolute value operator in the definitions above with a scalar distance measure—the Euclidean norm—that is also non-negative,

$$\|\mathbf{X}_n - \mathbf{Y}\| = \sqrt{(\mathbf{X}_n - \mathbf{Y})'(\mathbf{X}_n - \mathbf{Y})}.$$

These approaches are equivalent (Fuller 1976, pp. 182–183). The latter approach can be extended to random matrices by "vectorizing" them, that is, by stacking their columns in long column vectors.

Probability limits of continuous functions The following is an exceedingly important result: If $g(x)$ is a continuous function and $\operatorname{plim} X_n = X$, then we have $\operatorname{plim} g(X_n) = g(X)$. Here X_n can be a random variable or random vector, and X can be a constant (or vector of constants) or a random variable (or random vector).

To get a sense of how the theorem is proven, consider the special case in which the sequence of random variables $\{X_n\}$ converges in probability to the scalar constant c. For any $\eta > 0$, we consider the band around $g(c)$ formed by $g(c) + \eta$ and $g(c) - \eta$. Because g is continuous at c, we know that for any η we pick, no matter how small, there must exist a $\delta_{\eta,c} > 0$ such that for all x in the range $(c - \delta_{\eta,c},\ c + \delta_{\eta,c})$, the value assumed by the function $g(x)$ must be in the range $(g(c) - \eta,\ g(c) + \eta)$. To express this definition compactly,

$$| x - c | < \delta_{\eta,c} \text{ implies } | g(x) - g(c) | < \eta.$$

We now consider a random variable X_n, and examine the probability that it falls into the range in question. It must be the case that

$$\Pr(| X_n - c | < \delta_{\eta,c}) \leq \Pr(| g(X_n) - g(c) | < \eta),$$

because if Event A implies Event B, then $\Pr B \geq \Pr A$. Taking limits,

$$\lim_{n\to\infty} \Pr(| X_n - c | < \delta_{\eta,c}) \leq \lim_{n\to\infty} \Pr(| g(X_n) - g(c) | < \eta).$$

Recognizing that $\lim_{n\to\infty} \Pr(|\ X_n - c\ | < \delta_{\eta,c}) = 1$ because X_n converges in probability to c, we conclude that $\lim_{n\to\infty} \Pr(|\ g(X_n) - g(c)\ | <\ \eta) = 1$ as well. In other words, $\operatorname{plim} g(X_n) = g(c)$.

Nothing essential about the proof would change if we were to consider a sequence of random vectors $\{\mathbf{X}_n\}$ and a vector $\mathbf{c}$ of constants. We would only need to remind ourselves that when the continuous function g has a vector argument, we say that g is continuous at the point $\mathbf{c}$ if $\forall \eta > 0$ we can find a scalar $\delta_{\eta,c}$ such that $\|\mathbf{x} - \mathbf{c}\| < \delta_{\eta,c}$ implies $|g(\mathbf{x}) - g(\mathbf{c})| < \eta$. We have simply replaced an absolute value measure of distance in the x dimension with the vector norm $\|\cdot\|$ measure of distance.

That takes care of the case in which X_n converges in probability to a constant, but what about $X_n \xrightarrow{p} X$ with X being a random variable or vector? For this case a more sophisticated proof is needed (Fuller 1976, pp. 188–189). Let W be a closed and bounded set such that $\Pr(X \in W) \geq 1 - \delta/2$ and therefore $\Pr(X \notin W) < \delta/2$. Because W is closed and bounded, the g function is not just continuous but *uniformly* continuous on W, and so for any $\epsilon > 0$, there is a $\delta_\epsilon > 0$ such that $| x_1 - x_2 | < \delta_\epsilon$ implies $| g(x_1) - g(x_2) | < \epsilon$ if both x_1 and x_2 are in W. (Uniform continuity means that the same δ_ϵ applies irrespective of the values of x_1 and x_2—only the distance between the two points matters.)

Since $\operatorname{plim} X_n = X$, there is an integer $N_{\delta,\delta_\epsilon}$ such that for $n > N_{\delta,\delta_\epsilon}$, we must have $\Pr(| X_n - X | > \delta_\epsilon) < \delta/2$. Now, for such n values, we can establish the following chain of relationships,

$$\begin{aligned}\Pr(\mid g(X_n) - g(X) \mid > \epsilon) &= \Pr(\mid g(X_n) - g(X) \mid > \epsilon \mid X \notin W) \cdot \Pr(X \notin W) \\ &\quad + \Pr(\mid g(X_n) - g(X) \mid > \epsilon \mid X \in W) \cdot \Pr(X \in W) \\ &\leq \Pr(X \notin W) + \Pr(\mid g(X_n) - g(X) \mid > \epsilon \mid X \in W) \cdot \Pr(X \in W).\end{aligned}$$

We now let Event A be $\mid X_n - X \mid \leq \delta_\epsilon$ given $X \in W$. Event A implies Event B, defined as $\mid g(X_n) - g(X) \mid \leq \epsilon$ given $X \in W$. We therefore know that $\Pr(B) \geq \Pr(A)$ and thus $\Pr(\bar{A}) \geq \Pr(\bar{B})$ where $\bar{A}$ is the complement of A (on W) and $\bar{B}$ is the complement of B. In particular, $\bar{B}$ is the event $\mid g(X_n) - g(X) \mid > \epsilon$ for $X \in W$, and its probability is less than $\Pr(\bar{A}) = \Pr(\mid X_n - X \mid > \delta_\epsilon \mid X \in W)$. Inserting this last expression where we left off,

$$\begin{aligned}\Pr(\mid g(X_n) - g(X) \mid > \epsilon) &\leq \Pr(X \notin W) + \Pr(\mid g(X_n) - g(X) \mid > \epsilon \mid X \in W) \cdot \Pr(X \in W) \\ &\leq \Pr(X \notin W) + \Pr(\mid X_n - X \mid > \delta_\epsilon \mid X \in W) \cdot \Pr(X \in W) \\ &= \Pr(X \notin W) + \Pr(\mid X_n - X \mid > \delta_\epsilon) \leq \delta/2 + \delta/2 = \delta.\end{aligned}$$

The proof makes use of techniques that you will see in the econometrics journals; note in particular how uniform continuity plays a role. Note, too, that the theorem goes through if X_n and X are vectors, provided you replace the absolute values $\mid X_n - X \mid$ with Euclidean norms. It also applies to vectors of continuous functions $(g^1(X_n), g^2(X_n), \ldots, g^k(X_n))'$.

Mittelhammer (1996, Theorem 5.6, page 245) lists some of the implications of this important theorem. Among them, we have the result that $\operatorname{plim}(X_n \cdot Y_n) = \operatorname{plim} X_n \cdot \operatorname{plim} Y_n$ when both the $\{X_n\}$ and $\{Y_n\}$ sequences have probability limits. The theorem extends to cover random vectors and matrices. See Bierens (2004) and Serfling (1980, p. 24) for further discussion and a guide to the proofs. These results are essential to proving the consistency of the OLS estimator.

Probability Limits and Expectations

There is a tendency to think of plims as being somehow equivalent to expected values (in simple cases) or (as we will shortly see) to limits of expected values. Often these concepts coincide, but sometimes they do not, and you should not give into the temptation to think of plims as if they were expectations.

Consider this example in which a sequence of random variables has a probability limit, but the limit *does not* correspond to its expected value or to the limit of a sequence of expected values. Let X_n take the value 0 with probability $1 - 1/n$ and take the value n with probability $1/n$. Its expectation is simply 1, for every n. Its probability limit, however, is 0:

$$\lim_{n \to \infty} \Pr(\mid X_n - 0 \mid > \epsilon) = \lim_{n \to \infty} 1/n = 0.$$

Note that the variance of X_n is $n - 1$, which clearly explodes as $n \to \infty$.

What, then, *is* the connection between probability limits and expectations or limits of expectations? There are in fact some important linkages between them. As Bierens (2004, Theorem 6.4, 142–143) explains, when the random variables in the sequence $\{X_n\}$ are bounded, then $X_n \xrightarrow{p} X$ implies that $\lim_{n \to \infty} \mathrm{E}\, X_n = \mathrm{E}\, X$. (Boundedness implies that the

means exist.) This result includes the case in which X is a constant. Note that boundedness is precisely the problem in the example above. For unbounded random variables, there is a technical condition termed *uniform integrability* that gives the same result provided that $\mathrm{E}\, X$ exists.[2]

Convergence in Mean of Order r

We say that X_n converges in mean of order r to X if

$$\lim_{n\to\infty} \mathrm{E}(\mid X_n - X \mid^r) = 0.$$

It can be shown (Ramanathan (1993, p. 148); Amemiya (1985, Chapter 3)) that if X_n converges in mean of order r to X, then X_n also converges to X in order $s < r$. In econometrics, we are most often concerned with convergence in order $r = 2$, that is, with *convergence in mean square*.

A very important result—one that we will make use of repeatedly—is that convergence in mean square implies convergence in probability. We can show this by invoking Markov's inequality, which, as you will recall, applies to a positively-valued random variable Y with finite mean $\mathrm{E}\, Y$,

$$\Pr(Y > c) \leq \frac{\mathrm{E}\, Y}{c}.$$

Sometimes we see the inequality expressed in terms of $Y = g(X)$, where $g(X)$ is a non-negatively-valued function of a random variable (or random vector). In this case the result is stated as $\Pr(g(X) > c) \leq \mathrm{E}\, g(X)/c$.

Let's apply Markov's inequality to the concept of convergence in mean square. Let there be a parameter θ and consider the expected squared deviation of an estimator of θ from its true value θ_0,

$$\mathrm{E}(\hat{\theta}_n - \theta_0)^2 \equiv MS_n,$$

where MS_n denotes "mean square" and n is the sample size on which the estimator $\hat{\theta}_n$ is based. We assume that the expectation exists.

Applying Markov's inequality,

$$\Pr((\hat{\theta}_n - \theta_0)^2 > \epsilon^2) \leq \frac{\mathrm{E}(\hat{\theta}_n - \theta_0)^2}{\epsilon^2} = \frac{MS_n}{\epsilon^2}.$$

This is equivalent to

$$\Pr(\mid \hat{\theta}_n - \theta_0 \mid > \epsilon) \leq \frac{MS_n}{\epsilon^2}.$$

Now let the sample size n go to infinity. If it is the case that $MS_n \to 0$, then

$$\lim_{n\to\infty} \Pr(\mid \hat{\theta}_n - \theta_0 \mid > \epsilon) \leq \lim_{n\to\infty} \frac{MS_n}{\epsilon^2} = 0,$$

[2] See Serfling (1980, pp. 13–15) and Bierens (2004, Theorem 6.5, 143). A sequence of random variables $\{X_n\}$ is termed uniformly integrable if $\lim_{c\to\infty} \sup_n \int_{|X_n|>c} \mid X_n \mid f_n(x)dx = 0$. Note that if a random variable Y exists such that $\mid X_n \mid \leq Y$ with probability one for all n and $\mathrm{E}\, Y$ is finite, then this implies that $\{X_n\}$ is uniformly integrable. This result is the reason we so often see proofs that introduce such "dominating" Y variables.

that is, $\operatorname{plim} \hat{\theta}_n = \theta_0$. When this occurs, we say that the estimator $\hat{\theta}_n$ is *consistent* for the θ parameter.[3]

Convergence in mean square is a very important concept in econometric work. To make use of it, however, we must always make assumptions sufficient to guarantee that the mean square actually exists. This usually requires assumptions about the existence of higher moments.

Given that convergence in mean square implies convergence in probability, what more can be said about the implications for convergence of moments? As Mittelhammer (1996, Theorem 5.12, 250–252) proves, when X_n converges to X in mean square, and $\mathrm{E}\, X^2$ exists, we have $\lim_{n\to\infty} \mathrm{E}\, X_n = \mathrm{E}\, X$. It is helpful to see why. Assuming that X_n converges to X in mean square, it follows that

$$| \mathrm{E}\, X_n - \mathrm{E}\, X | = | \mathrm{E}(X_n - X) | \leq \mathrm{E} | X_n - X | \leq \left[\mathrm{E}(X_n - X)^2\right]^{1/2}.$$

Here the first inequality is, of course, due to the fact that $X_n - X$ must be less than or equal to $| X_n - X |$. (It can also be understood in terms of Jensen's inequality, since the absolute value function is a convex function.) The second inequality stems from Jensen's inequality applied to another convex function, by which

$$(\mathrm{E} | X_n - X |)^2 \leq \mathrm{E}(| X_n - X |^2),$$

the inequality being preserved when we take positive square roots. The final piece of the argument uses the fact that $\mathrm{E}(X_n - X)^2 \to 0$ by the definition of convergence in mean square, implying that we must have $| \mathrm{E}\, X_n - \mathrm{E}\, X | \to 0$ as well. (See Chapter 2 on Holder's inequality for a similar argument.) Mittelhammer continues this style of proof in showing that when X_n converges to X in mean square, $\operatorname{Var} X_n \to \operatorname{Var} X$. He also proves sufficiency, so that these are "if and only if" results.[4]

14.1.1 The O_p, o_p Notation

The following notation is often helpful in dealing with the relative magnitudes of various asymptotic quantities. (Chapter 2 presents the non-stochastic versions.) Judge et al. (1985, pp. 148–149) give a compact but informative summary, and Davidson and MacKinnon (1993, pp. 108–113) and Mittelhammer (1996) provide more detailed discussion. Fuller (1976, Chapter 5) gives a well-organized treatment with proofs.

We say that the sequence of random variables $\{X_n\}$ is of smaller order of magnitude than n^k if

$$\operatorname{plim} \frac{X_n}{n^k} = 0.$$

[3] Recall that the mean square is simply the sum of the squared bias and the variance of the estimator. Hence, another way of stating the result is to say that an estimator is consistent if both its bias and its variance go to zero in the limit. If the estimator happens to be unbiased, it is consistent provided that its variance goes to zero in the limit.

[4] Similar results have been obtained for convergence in mean of order r generally; they are not limited to convergence in mean square. For example, Bierens (2004) shows that when X_n converges to X in mean (convergence in order $r = 1$) and $\mathrm{E}\, X$ exists, then $\lim_{n\to\infty} \mathrm{E}\, X_n = \mathrm{E}\, X$.

This is denoted by $\{X_n\} = o_p(n^k)$. The concept extends to normalizing functions $g(n)$, in that if

$$\text{plim} \left| \frac{X_n}{g(n)} \right| = 0$$

we write $\{X_n\} = o_p(g(n))$. It can also be generalized to the probability limits of ratios of random variables. When we write $\{X_n\} = o_p(1)$, we mean $X_n \xrightarrow{p} 0$.

The notation $O_p(n^k)$ expresses a related concept. We say that $\{X_n\}$ is $O_p(n^k)$ if for any $\epsilon > 0$, there exists a constant K_ϵ such that

$$\Pr\left(\left| \frac{X_n}{n^k} \right| > K_\epsilon \right) < \epsilon$$

for all n. In other words, the absolute value of the ratio of X_n to n^k is bounded for *every* member of the sequence, being greater than K_ϵ only in cases having arbitrarily small probability.

When the sequence $\{X_n\}$ is $O_p(1)$, it is said to be "stochastically bounded." For instance, if X is distributed $\mathcal{N}(0,1)$, it is stochastically bounded by the definition above even though X itself can take values from $-\infty$ to $+\infty$. However, $\{X_n\}$ need not converge in distribution to be of order $O_p(1)$. (We will discuss "convergence in distribution" in our next chapter.) An example is provided by the case of $X_n \equiv 0$ for n odd and $X_n = \mathcal{N}(0,1)$ for n even.

We will make use of several simple rules for manipulating sequences of various orders. Consider two such sequences $\{X_n\}$ and $\{Y_n\}$. Adding, subtracting, and multiplying their elements yield sequences of the following orders:

$$\begin{aligned}
O(n^p) \pm O(n^q) &= O(n^{\max(p,q)}) \\
o(n^p) \pm o(n^q) &= o(n^{\max(p,q)}) \\
O(n^p) \pm o(n^q) &= O(n^p) \text{ if } p \geq q \\
&= o(n^q) \text{ if } p < q \\
O(n^p)O(n^q) &= O(n^{p+q}) \\
o(n^p)o(n^q) &= o(n^{p+q}) \\
O(n^p)o(n^q) &= o(n^{p+q}).
\end{aligned}$$

Note that if $\{X_n\}$ is $O(n^p)$, then $\{X_n\}$ is also $o(n^{p+\delta}), \delta > 0$.

14.2 Laws of Large Numbers: Sample Means

Several laws of large numbers will be examined in this section. Most of them require that the sequence of random variables be uncorrelated, but we will close the section with some results for time-series of random variables that allow restricted forms of serial correlation. The concern throughout is with random variables that can be expressed in the form of sample means, $\hat{\mu}_n = (1/n)\sum_{i=1}^n Y_i$. Once you have mastered the simple cases that follow, you will be fully capable of applying them to establish the consistency of ordinary least squares estimators.

Figure 14.1 illustrates convergence in probability for the simplest model that we will consider. The figure shows how the sample mean converges in probability to a constant, with the constant being the true mean. When an estimator (such as a sample mean) converges in probability to the true value of the parameter that the estimator is supposed to estimate, then the estimator is declared to be *consistent*.

As we examine whether the sample mean $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ converges in probability, we are trying to understand which outcome obtains under a given set of assumptions about the data-generating process that produces the sequence $\{Y_1, Y_2, \ldots, Y_n\}$. Depending on assumptions, we can find that:

- $\hat{\mu}_n$ does not converge in probability at all.
- It converges to a constant that is the value of a parameter in the dgp—for the sample mean, this would be $\mu = \mathrm{E}\, Y_i$. We would then say that $\hat{\mu}_n$ is "consistent for μ" or simply "consistent".
- The sample mean converges to a constant that is not itself one of the parameters of the dgp, but which is informative about those parameters. The usual example here, which we'll shortly see, is when the sample mean converges to a constant that is the limit of the average of the $\mathrm{E}\, Y_i$ expectations.
- The sample mean converges to a constant that is not informative. We'll see examples of this case when we look at the consequences of specification errors.

So with these possible outcomes in mind, let's begin:

- Consider $\{Y_i\}$, an uncorrelated sequence, with $\mathrm{E}\, Y_i = \mu$ and $\mathrm{Var}\, Y_i = \sigma^2$. (Hence, the mean and variance both *exist*.) We have $\mathrm{E}\, \hat{\mu}_n = \mathrm{E}(1/n)\sum_{i=1}^{n} Y_i = \mu$, and $\mathrm{Var}\, \hat{\mu}_n = \sigma^2/n$. In more complicated settings, it will prove convenient to write the variance of the sample mean in the form

$$\mathrm{Var}\, \hat{\mu}_n = \frac{1}{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \sigma^2 \equiv \frac{1}{n} \bar{V}_n,$$

  in which $\bar{V}_n$ is the average variance among the n observations. In the case at hand, $\bar{V}_n$ is simply σ^2.

  Now, by Markov's inequality,

$$\Pr(|\hat{\mu}_n - \mu| > \epsilon) \leq \frac{\mathrm{E}(\hat{\mu}_n - \mu)^2}{\epsilon^2} = \frac{\mathrm{Var}\, \hat{\mu}_n}{\epsilon^2} = \frac{\sigma^2}{n}\frac{1}{\epsilon^2}.$$

  This implies

$$\lim_{n\to\infty} \Pr(|\hat{\mu}_n - \mu| > \epsilon) \leq \lim_{n\to\infty} \frac{\sigma^2}{n}\frac{1}{\epsilon^2} = 0.$$

  In other words, $\hat{\mu}_n \xrightarrow{p} \mu$. We say that $\hat{\mu}_n$ is *consistent for* μ, a parameter of the data-generating process.

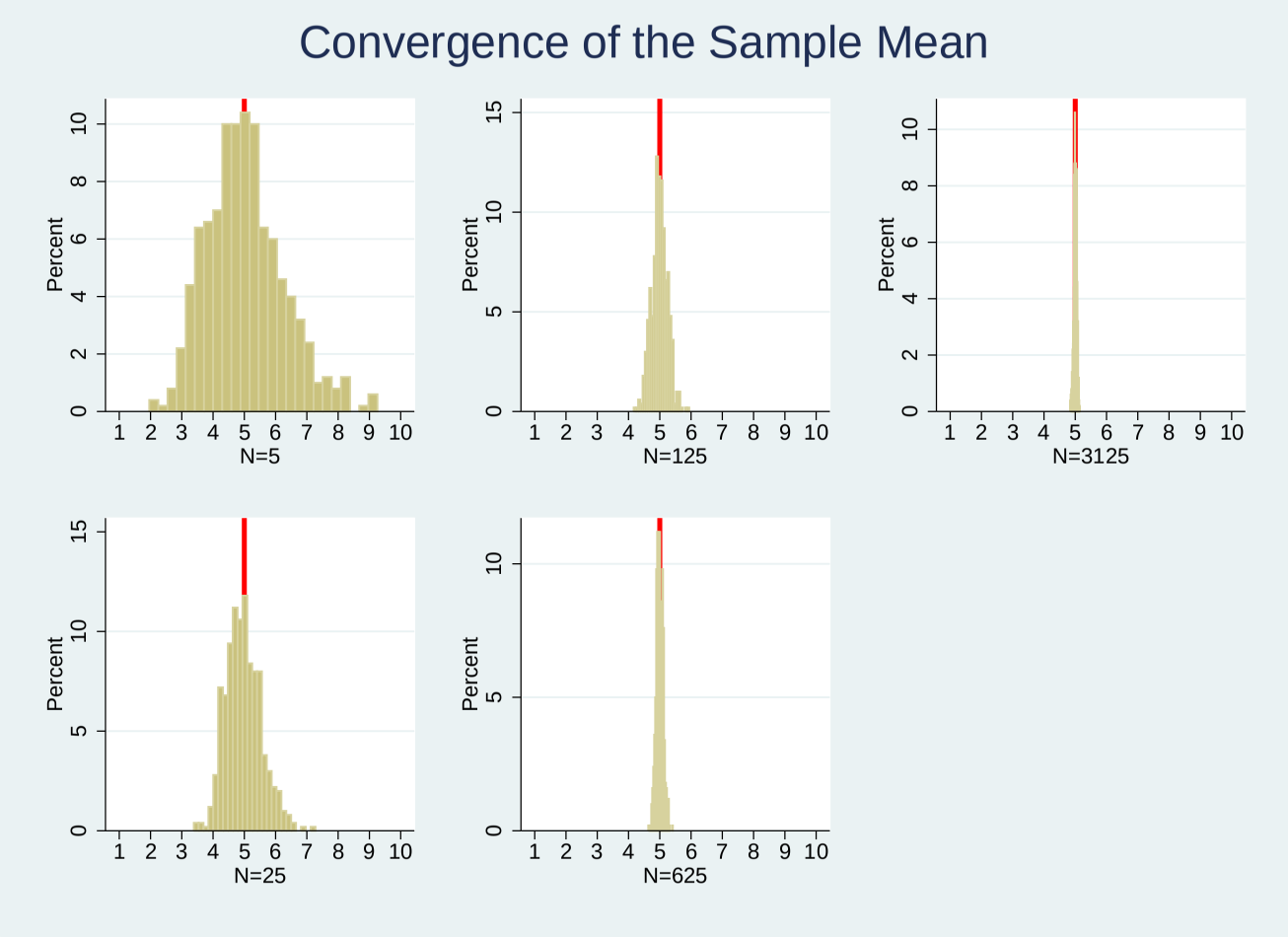


Figure 14.1: Convergence in probability of $\hat{\mu}$ with true $\mu = 5$ and variance $\sigma^2 = 8$, simulation results using sample sizes from 5 to 3125 observations. Disturbance term distributed as χ^2_4 with 4 subtracted from the disturbance to yield a zero mean.

- Continue to assume that $\{Y_i\}$ is uncorrelated with $\operatorname{Var} Y_i = \sigma^2$, but allow $\operatorname{E} Y_i = \mu_i$. We have

$$\operatorname{E} \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i \equiv \bar{\mu}_n,$$

and therefore

$$\hat{\mu}_n = (\hat{\mu}_n - \bar{\mu}_n) + \bar{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu_i) + \bar{\mu}_n.$$

The first term on the right-hand side has an expectation of zero, and

$$\operatorname{Var}(\hat{\mu}_n - \bar{\mu}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(Y_i - \mu_i) = \sigma^2/n.$$

Hence, for this term,

$$\Pr(|\hat{\mu}_n - \bar{\mu}_n| > \epsilon) \leq \frac{\operatorname{E}(\hat{\mu}_n - \bar{\mu}_n)^2}{\epsilon^2} = \frac{\sigma^2}{n} \frac{1}{\epsilon^2},$$

and by implication $\hat{\mu}_n - \bar{\mu}_n \xrightarrow{p} 0$.

At this point we can summarize what we've learned in the equation

$$\hat{\mu}_n = o_p(1) + \bar{\mu}_n,$$

or, expressed differently, as $\hat{\mu}_n \overset{a}{=} \bar{\mu}_n$. This is as far as we can go unless we introduce further assumptions. Note that both $\hat{\mu}_n$ and $\bar{\mu}_n$ can increase, decrease, oscillate, or otherwise vary with n; we know only that their *difference* is $o_p(1)$.

If the *additional assumption* is made that $\lim_{n\to\infty} \bar{\mu}_n = m$, then we at last obtain $\hat{\mu}_n \xrightarrow{p} m$. In this case, the sample mean converges to a constant that equals the limiting average of the individual means $\operatorname{E} Y_i$ (see Mittelhammer 1996, Theorem 5.23).

We cannot say that $\hat{\mu}_n$ is consistent, because m is not itself a parameter of the data-generating process. Rather, it is the limit of the average of such parameters.

- Continue to assume that $\{Y_i\}$ is uncorrelated, but now allow both $\operatorname{Var} Y_i = \sigma_i^2$ and $\operatorname{E} Y_i = \mu_i$. As you know, a sequence of random variables with variances differing over i is termed *heteroskedastic*. Again we have

$$\operatorname{E} \hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} \mu_i \equiv \bar{\mu}_n.$$

and again we write the sample mean as

$$\hat{\mu}_n = (\hat{\mu}_n - \bar{\mu}_n) + \bar{\mu}_n.$$

However, for the first term on the right, the variance is more complicated,

$$\operatorname{Var}(\hat{\mu}_n - \bar{\mu}_n) = \frac{1}{n^2} \sum_{i=1}^{n} \operatorname{Var}(Y_i - \mu_i) = \frac{1}{n^2} \sum_{i=1}^{n} \sigma_i^2 = \frac{1}{n} \bar{V}_n.$$

Does $\lim_{n\to\infty}(1/n)\bar{V}_n = 0$? We don't know, because to this point we have not imposed enough structure on the problem to determine the answer. If it does converge to zero, then using Markov's inequality as above, we would obtain $\hat{\mu}_n - \bar{\mu}_n \xrightarrow{p} 0$. But without further assumptions we cannot proceed.

One way to break the impasse is to assume that

$$\lim_{n\to\infty} \bar{V}_n = V,$$

that is, to assume that the average variance approaches a limiting value of V. This assumption would deliver the result we require, that $\lim_{n\to\infty}(1/n)\bar{V}_n = 0$ and therefore $\hat{\mu}_n - \bar{\mu}_n \xrightarrow{p} 0$.

Alternatively, consider the *bounded variances* case with $\sigma_i^2 < B\ \forall i$. Then $\sum_{i=1}^n \sigma_i^2 < \sum_{i=1}^n B = B \cdot n$, and

$$\frac{1}{n}\bar{V}_n = \frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 < \frac{1}{n^2} B \cdot n = \frac{B}{n},$$

which also implies $\hat{\mu}_n - \bar{\mu}_n \xrightarrow{p} 0$. Either one of these assumptions does the job; you would choose between them on the basis of the specifics of the economic model you are considering.

To resume the proof, with the aid of one assumption or the other, we have determined that $\hat{\mu}_n = o_p(1) + \bar{\mu}_n$. We now find ourselves back at a familiar road-block. Inserting the further assumption that $\lim_{n\to\infty} \bar{\mu}_n = m$ gives us passage to the final result we have been seeking, that $\hat{\mu}_n \xrightarrow{p} m$. Although $\hat{\mu}_n$ is not consistent, it at least converges to a constant (m) that is informative.

Please note: these results exploiting Markov's inequality require both means and variances to exist, which seems restrictive. What can be said for cases in which variances do not exist? In this connection, Mittelhammer (1996, Theorem 5.19) discusses *Khinchin's* weak law of large numbers. This law requires the sequence $\{Y_i\}$ to be not only uncorrelated, but actually iid. This strong assumption is counterbalanced by a weak assumption: the law requires only the existence of means.

Khinchin's law states that with $\{Y_i\}$ an iid sequence and $\mathrm{E}\, Y_i = \mu$, the sample mean $\hat{\mu}_n \xrightarrow{p} \mu$. The proof as outlined by Mittelhammer assumes the existence of the moment-generating function $\mathrm{MGF}_Y(t)$. From the iid assumption,

$$\begin{aligned}
\mathrm{MGF}_{\hat{\mu}_n}(t) &= \mathrm{E} \exp\left(t \cdot \frac{1}{n}\sum_{i=1}^n Y_i\right) \\
&= \prod_{i=1}^n \mathrm{E} \exp\left(\frac{t}{n} \cdot Y_i\right) \\
&= \left(\mathrm{MGF}_Y\left(\frac{t}{n}\right)\right)^n.
\end{aligned}$$

We proceed as follows,

$$\begin{aligned}\lim_{n\to\infty} \mathrm{MGF}_{\hat{\mu}_n}(t) &= \lim_{n\to\infty} \big(1 + \mathrm{MGF}_Y(t/n) - 1\big)^n \\ &= \lim_{n\to\infty} \left(1 + \frac{n(\mathrm{MGF}_Y(t/n) - 1)}{n}\right)^n \\ &= \exp\big(\lim_{n\to\infty} n(\mathrm{MGF}_Y(t/n) - 1)\big),\end{aligned}$$

using Lemma 5.4 from Mittelhammer (1996) in the last step (we gave the proof in Chapter 2). Write

$$\lim_{n\to\infty} n\big(\mathrm{MGF}_Y(t/n) - 1\big) = \lim_{n\to\infty} \frac{\mathrm{MGF}_Y(t/n) - 1}{n^{-1}}.$$

To find the limit, apply L'Hospital's Rule, taking derivatives with respect to n of both numerator and denominator. The resulting ratio is

$$\mathrm{MGF}'_Y(t/n) \cdot t.$$

We then have

$$\lim_{n\to\infty} \mathrm{MGF}'_Y(t/n) \cdot t = \mu t,$$

because $\mathrm{MGF}'_Y(0) = \mu$. This implies

$$\lim_{n\to\infty} \mathrm{MGF}_{\hat{\mu}_n}(t) = e^{\mu t},$$

which is the moment-generating function of a degenerate random variable with all of its probability mass concentrated at μ, or, in other words, the constant μ.

What about more challenging cases in which even means don't exist? In general, laws of large numbers cannot be derived for these cases. For instance, if the data series is iid with each Y_i being distributed according to the Cauchy distribution, it is known that the sample mean does not converge in probability. Distributions such as these, with unusually "heavy tails", are popular in finance and related fields in which the probabilities of extreme events are of great interest.

Unbiasedness versus consistency

Before going on, let us pause to reflect for a moment on the difference between *unbiasedness* and *consistency*. Each of these is a desirable property of an estimator, but when you are first learning about consistency, you may tend to confuse it with unbiasedness (in much the same way, and for much the same reason, that you may confuse probability limits with expectations). To see the difference between these concepts more clearly, let's return to the assumptions used in our first case above—in which $\mathrm{E}\, Y_i = \mu$, $\mathrm{Var}\, Y_i = \sigma^2$ and the $\{Y_i\}$ sequence is uncorrelated—but instead of focusing on the sample mean, let's examine a different estimator.

Let

$$\tilde{\mu}_n = \frac{1}{2} \cdot Y_1 + \frac{1}{2} \frac{1}{n-1} \sum_{i=2}^{n} Y_i.$$

It is easy to verify that $\tilde{\mu}_n$ is unbiased, with $\mathrm{E}\,\tilde{\mu}_n = \mu$. But is the estimator consistent? Examining its variance, we see that

$$\mathrm{Var}\,\tilde{\mu}_n = \frac{1}{4}\sigma^2 + \frac{1}{4}\frac{\sigma^2}{n-1},$$

and as $n \to \infty$, this variance converges to $\frac{1}{4}\sigma^2$ rather than to zero. The distribution of $\tilde{\mu}_n$ never completely collapses on μ; it always retains some variability. Hence $\tilde{\mu}_n$ is *unbiased but inconsistent*.

Another example is given by

$$\tilde{\mu}_n = \frac{1}{n-2}\sum_{i=1}^{n} Y_i.$$

Here $\tilde{\mu}_n$ is biased, with mean $\mathrm{E}\,\tilde{\mu}_n = \frac{n}{n-2}\mu$. To be sure, $n/(n-2) \to 1$ so that the bias gets smaller and smaller as n increases. Since

$$\tilde{\mu}_n = \frac{1}{n-2}\sum_i (Y_i - \mu) + \frac{n}{n-2}\mu,$$

we see that

$$\mathrm{Var}\,\tilde{\mu}_n = \mathrm{Var}\,\frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \mu) = \frac{n}{n-2}\cdot\frac{\sigma^2}{n-2},$$

The variance of the estimator goes to zero in the limit, leaving $\tilde{\mu}_n \stackrel{a}{=} \frac{n}{n-2}\mu$. And since $n/(n-2) \to 1$, the estimator is *biased but consistent* for μ.

Allowing for correlation

Now let us return to our examination of the sample mean and weaken the conditions on the Y_i further, allowing them to be inter-correlated. Is it still true that $\hat{\mu}_n$ is consistent for μ? The answer is "maybe"—it depends on the sum of the variances and on the sizes of the covariances between Y_i and Y_j. Serfling (1980) and Spanos (1986, p. 169) discuss these conditions briefly, as do most time-series textbooks in vastly greater detail (e.g., Hayashi 2000).

Earlier we saw that when variances σ_{ii} differ across observations, if we assume that the variances are bounded, then we can prove convergence in probability. In the present case, suppose that we assume that each variance $\sigma_{ii} < B$ and also assume that the covariances are bounded, that is, $|\,\sigma_{ij}\,| < C$ for $C > 0$. Is this enough to assure convergence in probability? No, or at least, not in general. In this case,

$$\begin{aligned}\mathrm{Var}(\hat{\mu}_n - \bar{\mu}_n) &= \frac{1}{n^2}\left(\sum_{i=1}^{n}\sigma_{ii} + \sum_{i=1}^{n}\sum_{j\neq i}\sigma_{ij}\right)\\ &\leq \frac{1}{n^2}\left(n\cdot B + n(n-1)\cdot C\right),\end{aligned}$$

so that the variance converges not to zero, but rather to $C > 0$. Evidently bounds on variances and covariances are not quite enough.

Here is one important time-series case for which, with one additional assumption, we can easily prove that the sample mean $\hat{\mu}_n$ converges in probability to μ. Let $Y_t = \mu + \epsilon_t$ and assume $\mathrm{E}\,\epsilon_t = 0$. Let's allow for arbitrary patterns of serial correlation in the $\{\epsilon_t\}$ sequence, but with one important restriction: When the time gap between ϵ_t and ϵ_s exceeds G, that is, $|t - s| > G$, then we impose the restriction that the covariance $\mathrm{E}(\epsilon_t \epsilon_s) = 0$. When $|t - s| \leq G$, however, we allow $\mathrm{E}(\epsilon_t \epsilon_s) \neq 0$, placing no conditions on its levels or time pattern.

In this set-up,

$$\mathrm{E}\,\hat{\mu}_T = \mu + \frac{1}{T}\sum_{t=1}^{T} \mathrm{E}\,\epsilon_t = \mu,$$

and

$$\mathrm{Var}\,\hat{\mu}_T = \frac{1}{T^2}\sum_{t=1}^{T}\sum_{s=1}^{T} \mathrm{E}(\epsilon_t \epsilon_s)$$

in general. The variance expression looks as if it might be troublesome, because the number of terms in each of the double sums increases with T, raising mathematically delicate questions about the convergence of infinite sums.

But if we think about our restriction that $\mathrm{E}(\epsilon_t \epsilon_s) = 0$ when $|t - s| > G$, we see that we can drastically simplify the task in front of us. For any value of t from the first sum, you should be able to recognize that

$$\sum_{s=1}^{T} \mathrm{E}(\epsilon_t \epsilon_s) = \sum_{s=\max(1,t-G)}^{s=\min(T,t+G)} \mathrm{E}(\epsilon_t \epsilon_s)$$

and you can easily confirm that the total number of terms in $\sum_{s=\max(1,t-G)}^{\min(T,t+G)} \mathrm{E}(\epsilon_t \epsilon_s)$ cannot exceed $2 \cdot G + 1$. (The number of terms is less than that at the beginning and end of the sequence: draw a picture and consider the possible range for the s subscript when $t = 1$ and $t = T$.) The key point is that the number of terms in this sum *does not go to infinity* as $T \to \infty$, but rather *remains capped* at $2 \cdot G + 1$.

We can therefore rewrite the variance of the sample mean to focus on the essentials, as

$$\mathrm{Var}\,\hat{\mu}_T = \frac{1}{T^2}\sum_{t=1}^{T} V_t$$

with $V_t \equiv \sum_{s=\max(1,t-G)}^{\min(T,t+G)} \mathrm{E}(\epsilon_t \epsilon_s)$. Having renamed things in this way, we can now easily see what conditions would guarantee that $\hat{\mu}_T \xrightarrow{p} \mu$. If $V_t < B$, that is V_t is bounded, this would do the job. Alternatively, if $\lim_{T\to\infty} \frac{1}{T}\sum_t^T V_t = v$, this would work as well.

The technique we've just used is fundamental to fixed-T, $N \to \infty$ panel-data econometrics, as we will see later in our course. Our simple, easily proven, and yet powerful result also provides us with an entryway into the somewhat arcane world of time-series econometrics. Indeed, one of the workhorse data-generating processes used in time series, the *moving-average of order G*, maps perfectly into the set-up we've just explored. In such a process, we write $Y_t = \mu + \epsilon_t$ as before, but then express ϵ_t as the sum of $G + 1$ current and previous period components,

$$\epsilon_t = u_t + u_{t-1} + u_{t-2} + \cdots u_{t-G}$$

with each u_t assumed to have zero mean and the $\{u_t\}$ series assumed to be serially uncorrelated. This yields a correlation pattern for the $\{\epsilon_t\}$ series with the property that when $|t-s| > G$, we have $\mathrm{E}(\epsilon_t \epsilon_s) = 0$ but when $|t-s| \leq G$ the covariance is non-zero. Another important linkage in the time-series literature is to series with the property of "M-dependence", in which random variables separated by more than M periods are assumed not just uncorrelated but fully independent. Most of the research attention in this part of the literature has directed to the development of central limit theorems.

On the whole, however, the time-series literature has been mainly focused on data-generating processes for which $\mathrm{E}(\epsilon_t \epsilon_s) \neq 0$ even as the gap $|t-s|$ grows toward infinity. These processes require more sophisticated mathematical tools that we have applied so far, exploring conditions under which the sum $\sum_{s=1}^{T} \mathrm{E}(\epsilon_t \epsilon_s)$ converges even as $T \to \infty$. Restrictive assumptions such as "stationarity" and "ergodicity" are generally needed to make the infinite sums behave appropriately.

Note that the result we've just given requires *no assumptions regarding "stationarity"* and indeed, requires no assumptions about the pattern of $\mathrm{E}(\epsilon_t \epsilon_s)$ when the gap between t and s is G or less. It is therefore worth asking whether, in any substantive applied problem, the assumption of a strictly bounded range on non-zero covariances is something that can be lived with. After all, to set the bounded range assumption aside and confront the intricacies of infinite series, you would have to be willing to impose in its place other strong assumptions about stationarity and ergodicity. For any given applied problem, the trade-offs between these sets of assumptions will need some careful consideration. In particular, with an applied problem in mind, you will need to ask yourself whether the assumption that the data-generating process is stationary can possibly be justified.

In these notes, we cannot begin to do justice to the rich and sophisticated time-series literature, but the following results will convey something of its flavor and provide some preliminary insight into the role of stationarity.

- Hayashi (2000, p. 401) presents a law of large numbers for cases in which the $\{Y_i\}$ sequence is assumed to be *covariance stationary* with common mean μ, by which we mean that the variance of Y_i is the constant $\sigma^2 \equiv \gamma_0$ and the covariance of Y_i and Y_j for $i \neq j$ is $\gamma_{|i-j|}$, a quantity that depends on the difference between the time points i and j, but not on the values of i and j as such. With covariance stationarity imposed, the key additional condition under which the sample mean $\hat{\mu}$ converges in probability to the true mean μ is that $| \gamma_j | \to 0$ as $j \to \infty$. In other words, the covariance between any two observations decreases as the gap in time between them increases and it approaches a limit of zero as the gap grows large.

  With $\hat{\mu}_n = \mu + (\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{\epsilon}$, the variance of $\hat{\mu}_n$ is

$$\mathrm{E}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1} = \frac{1}{n^2}\boldsymbol{\iota}' \, \mathrm{E}\, \boldsymbol{\epsilon}\boldsymbol{\epsilon}'\boldsymbol{\iota},$$

  and

$$\mathrm{E}\, \boldsymbol{\epsilon}\boldsymbol{\epsilon}' = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \cdots & \gamma_{n-1} \\ \gamma_1 & \gamma_0 & \gamma_1 & \cdots & \gamma_{n-2} \\ \gamma_2 & \gamma_1 & \gamma_0 & \cdots & \gamma_{n-3} \\ \vdots & & & \ddots & \vdots \\ \gamma_{n-1} & \cdots & & & \gamma_0 \end{bmatrix}.$$

Note that $\boldsymbol{\iota}' \operatorname{E} \boldsymbol{\epsilon}\boldsymbol{\epsilon}' \boldsymbol{\iota} = n\gamma_0 + 2(n-1)\gamma_1 + 2(n-2)\gamma_2 + \cdots + 2(n-(n-1))\gamma_{n-1}$, and this sum can be written as $n\gamma_0 + 2\sum_{j=1}^{n}(n-j)\gamma_j$, where we can allow the j subscript to go to n because the term associated with $j = n$ is zero. Hence, the variance of the sample mean is

$$\operatorname{Var} \hat{\mu}_n = \frac{1}{n}\left(\gamma_0 + 2\sum_{j=1}^{n}\left(1 - \frac{j}{n}\right)\gamma_j\right) = \frac{\gamma_0}{n} + \frac{2}{n}\sum_{j=1}^{n}\left(1 - \frac{j}{n}\right)\gamma_j.$$

Now, obviously $\gamma_0/n \to 0$ as $n \to \infty$. Also, because

$$\frac{1}{n}\sum_{j=1}^{n}\left(1 - \frac{j}{n}\right)\gamma_j \leq \frac{1}{n}\sum_{j=1}^{n} \mid \gamma_j \mid,$$

if we have $\mid \gamma_j \mid \to 0$ as $j \to \infty$, then this implies $\frac{1}{n}\sum \mid \gamma_j \mid \to 0$ as well (see the discussion in Chapter 2 on convergence of series for this result).

- Serfling presents another law of large numbers of this general type. If the Y_i have common mean μ, variances σ_i^2, and bounded covariances

$$\left|\operatorname{E}(Y_i - \mu)(Y_j - \mu)\right| \leq \rho_{|i-j|},$$

and meet the additional conditions that $\lim n^{-1}\sum_i^n \sigma_i = \alpha$ and $\sum_{i=0}^{\infty}\rho_i$ finite, then $\hat{\mu}_n \xrightarrow{p} \mu$. (Note that $\sum\rho$ is then $O(1)$.) This law does not require covariance stationarity as such.

If you are interested in time-series econometrics, you will need to study other laws of large numbers that deal with such inter-correlated $\{Y_n\}$ sequences.

The following interesting result is sometimes invoked to covered both uncorrelated and correlated data series. If

$$\lim_{n\to\infty} \operatorname{E} \frac{(\hat{\mu}_n - \bar{\mu}_n)^2}{1 + (\hat{\mu}_n - \bar{\mu}_n)^2} = 0,$$

this implies $\hat{\mu}_n - \bar{\mu}_n \xrightarrow{p} 0$. The proof is simple. Given any two non-negative numbers a and b, a little algebra shows that $a \geq b \Rightarrow a/(1+a) \geq b/(1+b)$. Hence,

$$(\hat{\mu}_n - \bar{\mu}_n)^2 \geq \epsilon^2 \Rightarrow \frac{(\hat{\mu}_n - \bar{\mu}_n)^2}{1 + (\hat{\mu}_n - \bar{\mu}_n)^2} \geq \frac{\epsilon^2}{1+\epsilon^2}.$$

It follows that

$$\Pr\Big((\hat{\mu}_n - \bar{\mu}_n)^2 \geq \epsilon^2\Big) \leq \Pr\left(\frac{(\hat{\mu}_n - \bar{\mu}_n)^2}{1 + (\hat{\mu}_n - \bar{\mu}_n)^2} \geq \frac{\epsilon^2}{1+\epsilon^2}\right).$$

By Markov's inequality,

$$\Pr\left(\frac{(\hat{\mu}_n - \bar{\mu}_n)^2}{1 + (\hat{\mu}_n - \bar{\mu}_n)^2} \geq \frac{\epsilon^2}{1+\epsilon^2}\right) \leq \frac{\operatorname{E}\left(\frac{(\hat{\mu}_n - \bar{\mu}_n)^2}{1+(\hat{\mu}_n - \bar{\mu}_n)^2}\right)}{\frac{\epsilon^2}{1+\epsilon^2}}$$

and by hypothesis the expectation appearing in the right-hand side numerator has a limit of zero. This implies

$$\lim_{n\to\infty} \Pr\left((\hat{\mu}_n - \bar{\mu}_n)^2 \geq \epsilon^2\right) = 0,$$

or, taking positive square roots, $\lim_{n\to\infty} \Pr\left(\mid \hat{\mu}_n - \bar{\mu}_n \mid \geq \epsilon\right) = 0$. Note that the mean-square assumption $\lim_{n\to\infty} \mathrm{E}(\hat{\mu}_n - \bar{\mu}_n)^2 = 0$ is not strictly required here; the result is more general than that.

14.3 Laws of Large Numbers: Sample Variances

The concepts and tools that we will use in this section are *exactly the same* as those we've just used in connection with sample means. Make sure that you see that—don't let the difference in notation obscure the fundamental similarities.

To understand the asymptotic behavior of the sample mean $\hat{\mu}_n$, we have had to make assumptions about both the mean and the variance of Y_i. To understand the behavior of s_n^2, we'll find that we need to make assumptions about even higher moments of Y_i.

For this analysis, let's adjust our notation to write things out in linear-model form, $Y_i = \mu + \epsilon_i$, and then examine s_n^2 expressed in terms of residuals $\mathbf{e} = \mathbf{Y} - \hat{\mu}_n \cdot \boldsymbol{\iota}$,

$$s_n^2 = \frac{1}{n-1}\mathbf{e}'\mathbf{e} \stackrel{a}{=} \frac{1}{n}\mathbf{e}'\mathbf{e}.$$

You will remember that $\mathbf{e} = \mathbf{Y} - \boldsymbol{\iota}\hat{\mu}_n = \left(\mathbf{I} - \boldsymbol{\iota}(\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\boldsymbol{\iota}'\right)\boldsymbol{\epsilon} = \mathbf{M}\boldsymbol{\epsilon}$, with $\mathrm{E}\,\boldsymbol{\epsilon} = \mathbf{0}$.

Assuming homoskedasticity

If we assume that the disturbances are homoskedastic, then $\mathrm{Var}\,\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$. We re-write s_n^2 in what by now is a familiar form,

$$\begin{aligned} s_n^2 &\stackrel{a}{=} \frac{1}{n}\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon} \\ &= \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} - \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\iota}(\frac{1}{n}\boldsymbol{\iota}'\boldsymbol{\iota})^{-1}\frac{1}{n}\boldsymbol{\iota}'\boldsymbol{\epsilon} \\ &= \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right). \end{aligned}$$

Using the assumptions we've made previously in our analysis of the sample mean, we have $(1/n)\sum_{i=1}^{n}\epsilon_i \stackrel{p}{\to} 0$. Why? The sequence $\{\epsilon_i\}$ is uncorrelated, with each member of the sequence having mean zero and variance σ^2. Our simplest law of large numbers applies. This takes care of the last two terms.

We're left to consider what happens with $s_n^2 \stackrel{a}{=} (1/n)\sum_i \epsilon_i^2$. We have assumed that $\mathrm{E}\,\epsilon_i^2 = \sigma^2$, so that $(1/n)\sum_i \epsilon_i^2$ is the average of n terms with σ^2 being their common mean. To proceed, we must now *assume* that the $\{\epsilon_i^2\}$ sequence is uncorrelated, which is an assumption that has *not* been required in our examination of the sample mean. With this assumption,

$$\mathrm{Var}(1/n)\sum_i \epsilon_i^2 = \frac{1}{n}\bar{V}_n = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\,\epsilon_i^2.$$

If we want to get into the details, we would examine $\operatorname{Var}\epsilon_i^2 = \operatorname{E}(\epsilon_i^2 - \sigma^2)^2$, or $\operatorname{E}\epsilon_i^4 - \sigma^4$ when we expand the square and take expectations. Up to now, we have not needed to consider $\operatorname{E}\epsilon_i^4$, which is a fourth moment.

Suppose that $\operatorname{E}\epsilon_i^4 = \mu_4$, a constant. In this case,

$$\frac{1}{n}\bar{V}_n = \frac{1}{n^2}\sum_{i=1}^{n}(\mu_4 - \sigma^4) = \frac{1}{n}(\mu_4 - \sigma^4),$$

and clearly $s_n^2 \xrightarrow{p} \sigma^2$ if we apply Markov's inequality. (See Figure 14.2 for an illustration.) Alternatively, suppose that $\operatorname{E}\epsilon_i^4 = \mu_{4,i}$, a quantity that varies with i. Then to determine the asymptotic behavior of s_n^2, we must ask whether

$$\lim_{n\to\infty}\frac{1}{n}\bar{V}_n = \lim_{n\to\infty}\frac{1}{n^2}\sum_{i=1}^{n}(\mu_{4,i} - \sigma^4) = 0.$$

If the limit is zero, then s_n^2 is consistent for σ^2. What would guarantee such a result? As we saw earlier, a bound $\mu_{4,i} < B$ would suffice, as would the assumption $\lim_{n\to\infty}\bar{V}_n = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\mu_{4,i} - \sigma^4) = V$.

Heteroskedasticity and the Sample Variance

Suppose now that $\operatorname{E}\epsilon_i^2 = \sigma_i^2$, that is, the disturbances are heteroskedastic. Continue to assume that the sequences $\{\epsilon_i\}$ and $\{\epsilon_i^2\}$ are uncorrelated over i. Then, with

$$s_n^2 \overset{a}{=} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\right),$$

consider the term $(1/n)\sum_{i=1}^{n}\epsilon_i$. This quantity has zero mean and a variance of $(1/n)\bar{V}_n = (1/n^2)\sum_{i=1}^{n}\sigma_i^2$. We know that either of two assumptions, that $\bar{V}_n \to V$ or $\sigma_i^2 < B\,\forall i$, will imply that $\sum_{i=1}\sigma_i^2$ is $o(n^2)$, which in turn implies $(1/n)\sum_i \epsilon_i \xrightarrow{p} 0$. That takes care of the last two terms.

Now consider $(1/n)\sum_i \epsilon_i^2$. Allow the fourth moment of ϵ_i to vary with i, that is, let $\operatorname{E}\epsilon_i^4 = \mu_{4,i}$. Then, writing $s_n^2 = (s_n^2 - \bar{V}_n) + \bar{V}_n$ with $\bar{V}_n = (1/n)\sum_{i=1}^{n}\sigma_i^2$, consider

$$s_n^2 - \bar{V}_n \overset{a}{=} \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2 - \sigma_i^2).$$

The i-th term in the sum has zero mean and variance $\mu_{4,i} - \sigma_i^4$. Hence,

$$\operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2 - \sigma_i^2)\right) = \frac{1}{n^2}\sum_{i=1}^{n}(\mu_{4,i} - \sigma_i^4).$$

If $\sum_{i=1}^{n}(\mu_{i,4} - \sigma_i^4)$ is $o(n^2)$, which we obtain by making assumptions about bounded variances or limits, as above, then it follows that $(1/n)\sum_{i=1}^{n}(\epsilon_i^2 - \sigma_i^2) \xrightarrow{p} 0$, which in turn implies that $s_n^2 - \bar{V}_n \xrightarrow{p} 0$. Making the *further assumption* that $\lim_{n\to\infty}\bar{V}_n = V$, we finally obtain $s_n^2 \xrightarrow{p} V$. That is, s_n^2 converges to a quantity that is the limit of the average variance.

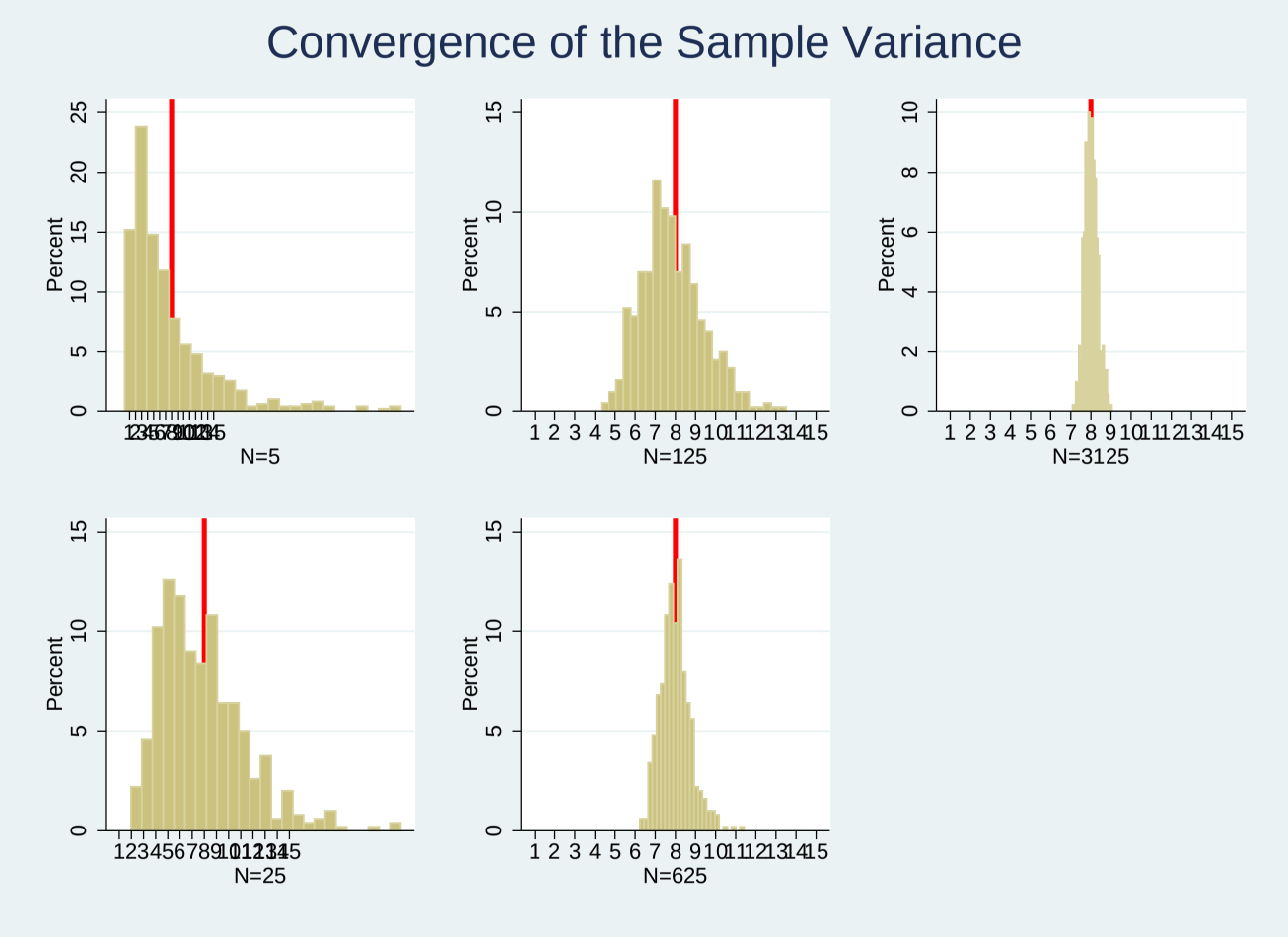


Figure 14.2: Convergence in probability of s^2 with true $\mu = 5$ and variance $\sigma^2 = 8$, , simulation results using sample sizes from 5 to 3125 observations. Disturbance term distributed as χ^2_4 with 4 subtracted from the disturbance to yield a zero mean.

14.4 Consistency of the OLS estimator

The basic ideas are all present in stripped-down form in the simple linear model $Y_i = \beta X_i + \epsilon_i$ with one right-hand side covariate. The OLS estimator of β is

$$\hat{\beta}_n = \beta + \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right)^{-1} \cdot \frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i.$$

As is usual, we can only determine the properties of the estimator from assumptions about the properties of its ingredients—the $\{X_n^2\}$ and $\{X_n \cdot \epsilon_n\}$ sequences.[5]

Since $\frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i$ is a sample average, it can be analyzed with the tools we've been developing. To begin, we assume that the $\{X_n \cdot \epsilon_n\}$ sequence is uncorrelated. To derive the expected value of each random variable, we introduce the *fundamental assumption justifying OLS regression*: $\mathrm{E}(\epsilon_i \mid X_i) = 0$. This assumption is far weaker than and thus preferable to the one we used to employ, $\mathrm{E}(\boldsymbol{\epsilon} \mid \mathbf{X}) = \mathbf{0}$, which disallowed non-zero conditional expectations for *any* pair of X_i and ϵ_j, ruling out many economic models of interest. From $\mathrm{E}(\epsilon_i \mid X_i) = 0$ we have $\mathrm{E}(X_i \cdot \epsilon_i) = 0$ by iterated expectations—so in our laws of large numbers tools, we needn't consider cases with i-varying means. The variance is $\mathrm{Var}(X_i \epsilon_i) = \mathrm{E}\, X_i^2 \epsilon_i^2 \equiv V_i$. Without assumptions regarding the V_i variances, we know that we do not have enough information to determine the probability limit of $\frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i$. If V_i is assumed constant, then our simplest law of large numbers applies, yielding $\frac{1}{n} \sum_{i=1}^{n} X_i \epsilon_i \xrightarrow{p} 0$. If V_i is not constant, then we would need to assume either that it is bounded, $V_i < B$, or that $\lim_n \bar{V}_n = \lim_n \frac{1}{n} \sum_i^n V_i = v$, both of which give the same result for the probability limit.

The other sample average in the OLS formula is $\frac{1}{n} \sum_{i=1}^{n} X_i^2$. We'll assume that $\{X_n^2\}$ is an uncorrelated sequence. In the absence of any additional structure, there is no option here but to proceed to consider all of the possible cases for laws of large numbers, involving the four combinations of $\mathrm{E}\, X_i^2$ and $\mathrm{Var}(X_i^2)$ being constant or i-varying. Since we know how to deal with all four cases, there is no need to rehearse them here: by applying the relevant assumptions, we reach the result $\frac{1}{n} \sum_{i=1}^{n} X_i^2 \xrightarrow{p} q$, in which q is interpreted either as $\mathrm{E}\, X_i^2$ when the expectation is constant over i, or $q \equiv \lim_n \frac{1}{n} \sum_i \mathrm{E}\, X_i^2$ for the non-constant case. Note that *some* assumption about the means and variances is required: Choose the weakest one that makes sense given your economic model.

At this point in the argument, we bring on board the theorem concerning probability limits of continuous functions of random variables, which yields:

$$\operatorname{plim} \left(\frac{1}{n} \sum_i X_i^2 \right)^{-1} = (q)^{-1}.$$

Obviously, the application of this result requires that $q \neq 0$, but pathological cases aside, that does not seem unreasonable. In summary, we obtain

$$\operatorname{plim} \hat{\beta}_n = \beta + q^{-1} \cdot 0 = \beta,$$

[5] Note that in the case at hand, neither the $\{\epsilon_n\}$ sequence nor the $\{X_n\}$ sequence directly influences the OLS estimator. But this is unusual. If a constant term were in the model—in which case there would be two $\boldsymbol{\beta}$ parameters rather than just one—then we would have $\frac{1}{n} \sum_{i=1}^{n} \epsilon_i$ as one of the averages to consider in the right-hand term, and $\frac{1}{n} \sum_{i=1}^{n} X_i$ would be one of the terms inside the inverse. Constant terms are usually included in multivariate OLS regressions, which we will investigate shortly.

in other words, we find that the OLS estimator is *consistent*.

With this as background, it is really easy to deal with the multivariate regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, yielding the OLS estimator

$$\hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}.$$

The fundamental assumption justifying OLS regression is $\mathrm{E}(\boldsymbol{\epsilon}_i|\mathbf{X}_i) = 0$, which involves only the disturbance ϵ_i and the $\mathbf{X}_i$ covariate vector for the i-th observation.

Recall that $\mathbf{X}'\mathbf{X} = \sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'$ with $\mathbf{X}_i$ the $k \times 1$ column vector of explanatory variables for observation i, and

$$\mathbf{X}_i\mathbf{X}_i' = \begin{bmatrix} X_{i,1} \\ X_{i,2} \\ \vdots \\ X_{i,k} \end{bmatrix}\begin{bmatrix} X_{i,1} & X_{i,2} & \dots & X_{i,k} \end{bmatrix} = \begin{bmatrix} X_{i,1}^2 & X_{i,1}X_{i,2} & \dots & X_{i,1}X_{i,k} \\ X_{i,2}X_{i,1} & X_{i,2}^2 & & \vdots \\ \vdots & & \ddots & \vdots \\ X_{i,1}X_{i,k} & \dots & \dots & X_{i,k}^2 \end{bmatrix}.$$

Hence, $\mathbf{X}'\mathbf{X}$ is a matrix composed of sums (over $i = 1, \dots, n$) of squares and cross-products of the explanatory variables.

Consider any element of $n^{-1}\mathbf{X}'\mathbf{X}$, say, the entry in the second row and first column. Under what conditions will

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,2}X_{i,1} \xrightarrow{p} q_{2,1}$$

where $q_{2,1}$ is a constant? To answer this question, we let the random variable $Z_i \equiv X_{i,2}X_{i,1}$ and consider the properties of the $\{Z_i\}$ sequence. That is, we ask whether $\{Z_i\}$ is an uncorrelated sequence, whether $\mathrm{E}\, Z_i$ is constant or varies with i, and whether $\mathrm{Var}\, Z_i$ is a constant or varies. We then apply the appropriate law of large numbers.

We proceed in this way to analyze each distinct element of $n^{-1}\mathbf{X}'\mathbf{X}$. Provided that every element obeys a law of large numbers (different laws can be applied to different elements if need be), we obtain

$$\frac{1}{n}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{Q}.$$

The theorem about probability limits of continuous functions now applies, with the inverse being a continuous function of the elements of the $k \times k$ matrix. *Assuming* that $\mathbf{Q}$ is invertible,

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \xrightarrow{p} \mathbf{Q}^{-1}.$$

To complete our analysis of $\hat{\boldsymbol{\beta}}_n$, we proceed to consider the $k \times 1$ column vector $n^{-1}\mathbf{X}'\boldsymbol{\epsilon}$, which can be written out as $n^{-1}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i$, or, in even more detail,

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} X_{i,1}\epsilon_i \\ X_{i,2}\epsilon_i \\ \vdots \\ X_{i,k}\epsilon_i \end{bmatrix}.$$

Consider one element of this vector, for example the j-th, letting

$$Z_i \equiv X_{i,j}\epsilon_i,$$

and write the j-th element of the vector $n^{-1}\mathbf{X}'\boldsymbol{\epsilon}$ as $n^{-1}\sum_{i=1}^{n} Z_i$. This quantity is recognizable as a sample mean, although with ϵ_i being unobservable, it is not the kind of sample mean that we could calculate. Nevertheless, we can analyze its large-sample behavior.

As before, we ask about the properties of the sequence $\{Z_i\}$, although this time we can bring a bit more structure to the problem than was the case with $n^{-1}\mathbf{X}'\mathbf{X}$. By the fundamental assumption that $\mathrm{E}(\epsilon_i|\mathbf{X}_i) = 0$, we have $\mathrm{E}\, Z_i = \mathrm{E}\, X_{i,j}\epsilon_i = 0$. We see that we are dealing with a sequence of mean zero random variables. To apply Markov's inequality, we will need to *assume* that the $\{Z_i\}$ sequence is uncorrelated over i, and this requires the sequence of products $\{X_{i,j}\epsilon_i\}$ to be uncorrelated. Then

$$\mathrm{Var}\,\frac{1}{n}\sum_{i=1}^{n} Z_i = \mathrm{Var}\,\frac{1}{n}\sum_{i=1}^{n} X_{i,j}\epsilon_i = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_{i,j}\epsilon_i).$$

Each member of the $\{Z_i\}$ sequence has mean zero and a variance that could either be constant or changing with i. We have laws of large numbers available to deal with both of these cases.

Under a range of assumptions, then, we can obtain the result

$$\frac{1}{n}\sum_{i=1}^{n} Z_i = \frac{1}{n}\sum_{i=1}^{n} X_{i,j}\epsilon_i \xrightarrow{p} 0,$$

and applying this kind of analysis to each element of the $k \times 1$ vector $n^{-1}\mathbf{X}'\boldsymbol{\epsilon}$, we obtain

$$\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon} \xrightarrow{p} \mathbf{0}.$$

Hence,

$$\begin{aligned}\hat{\boldsymbol{\beta}}_n &= \boldsymbol{\beta} + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\left(\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}\right)\\ &\stackrel{a}{=} \boldsymbol{\beta} + \mathbf{Q}^{-1}\cdot\mathbf{0} = \boldsymbol{\beta}.\end{aligned}$$

That is, $\hat{\boldsymbol{\beta}}_n$ is consistent for $\boldsymbol{\beta}$.

14.5 Consistency of s_n^2

This analysis is very similar to what we have already seen in the case of sample variances. We'll assume that $\mathrm{Var}\,\boldsymbol{\epsilon} = \sigma^2\mathbf{I}$, but the logic can be extended to cover the case of heteroskedasticity. We begin with

$$\begin{aligned}s_n^2 &= \frac{1}{n-k}\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}\\ &\stackrel{a}{=} \frac{1}{n}\boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}\\ &= \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} - \frac{1}{n}\boldsymbol{\epsilon}'\mathbf{X}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}.\end{aligned}$$

We have already analyzed the probability limits of $n^{-1}\mathbf{X}'\boldsymbol{\epsilon}$ and $(n^{-1}\mathbf{X}'\mathbf{X})^{-1}$, which we found to be $\mathbf{0}$ and $\mathbf{Q}^{-1}$ respectively. Hence,

$$\begin{aligned} s_n^2 &\stackrel{a}{=} \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} - \mathbf{0}\cdot\mathbf{Q}^{-1}\cdot\mathbf{0} \\ &\stackrel{a}{=} \frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2. \end{aligned}$$

From this point forward, the analysis is identical to that performed for the sample variance. Under a variety of assumptions, we obtain the result $s_n^2 \xrightarrow{p} \sigma^2$.

Hayashi (2000, pp. 115–117) offers another perspective on the proof that is illuminating. Writing the OLS residual as

$$e_i = Y_i - \mathbf{X}_i'\hat{\boldsymbol{\beta}}_n = \epsilon_i - \mathbf{X}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}),$$

we have

$$e_i^2 = \epsilon_i^2 - 2(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})'\mathbf{X}_i\epsilon_i + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})'\mathbf{X}_i\mathbf{X}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}),$$

and this yields

$$\frac{1}{n}\sum_i e_i^2 = \frac{1}{n}\sum_i \epsilon_i^2 - 2(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})'\frac{1}{n}\sum_i \mathbf{X}_i\epsilon + (\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta})'\frac{1}{n}\sum_i \mathbf{X}_i\mathbf{X}_i'(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}).$$

As we have seen, under the fundamental assumption $\mathrm{E}(\epsilon_i \mid \mathbf{X}_i) = 0$, the probability limit of $\frac{1}{n}\sum_i \mathbf{X}_i\epsilon_i$ equals zero and that the probability limit of $\frac{1}{n}\mathbf{X}_i\mathbf{X}_i' = \mathbf{Q}$. These results, coupled with the consistency of $\hat{\boldsymbol{\beta}}_n$, cause the last two terms on the right-hand side to vanish in the limit, leaving $\frac{1}{n}\sum_i e_i^2 \stackrel{a}{=} \frac{1}{n}\sum_i \epsilon_i^2$. Note, however, that this result also obtains if the probability limit of $\frac{1}{n}\sum_i \mathbf{X}_i\epsilon_i$ is any finite value, and if $\hat{\boldsymbol{\beta}}_n$ is replaced by any consistent estimator of the $\boldsymbol{\beta}$ parameter. For instance, if we allow the probability limit of $\frac{1}{n}\sum_i \mathbf{X}_i\epsilon_i$ to be non-zero and replace $\hat{\boldsymbol{\beta}}_n$ by the instrumental variables estimator, the result continues to hold.

Proof that $\operatorname{plim} e_i = \boldsymbol{\epsilon}_i$

Consider a single regression residual e_i and the counterpart disturbance $\boldsymbol{\epsilon}_i$. Hayashi's analysis indicates that $\operatorname{plim} e_i = \boldsymbol{\epsilon}_i$, with the subscript i being held fixed as $n \to \infty$. We can also see this from

$$\begin{aligned} \mathbf{e} &= \mathbf{M}\boldsymbol{\epsilon} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\boldsymbol{\epsilon}, \\ e_i - \epsilon_i &= -\mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon} \\ &= -\mathbf{X}_i'(\frac{1}{n}\mathbf{X}'\mathbf{X})^{-1}(\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon}), \end{aligned}$$

noting that as $n \to \infty$, the vector $\mathbf{X}_i$ itself remains unchanged. By the arguments already employed above we see that $\operatorname{plim} e_i - \epsilon_i = 0$.

Why is this result of interest? The reason is that many important tests address the properties of $\boldsymbol{\epsilon}$, with the null hypothesis being that $\boldsymbol{\epsilon}$ is distributed with mean $\mathbf{0}$ and variance $\sigma^2\mathbf{I}$. It may seem natural to use the regression residuals $\mathbf{e} = \mathbf{M}\boldsymbol{\epsilon}$ in forming a test,

but the difficulty is that even under the null hypothesis, $\mathrm{E}\,\mathbf{e}\mathbf{e}' = \sigma^2\mathbf{M} \neq \sigma^2\mathbf{I}$. That is, the regression residuals $\mathbf{e}$ are inter-correlated even when the $\boldsymbol{\epsilon}$ are not. Even so, the fact that $e_i \stackrel{a}{=} \epsilon_i$ is an important first step in showing that tests based on regression residuals are justified in large samples.

14.6 Specification Errors and Inconsistency of $\hat{\boldsymbol{\beta}}_n$

As we've seen, the key to the consistency proof is that $\operatorname{plim} \frac{1}{n}\sum_i \mathbf{X}_i\epsilon_i = \mathbf{0}$, which follows from the fundamental assumption that $\mathrm{E}(\epsilon_i \mid \mathbf{X}_i) = 0$ and conditions on variances. When the fundamental assumption is violated, such that $\mathrm{E}(\epsilon_i \mid \mathbf{X}_i) \neq 0$, then $\hat{\boldsymbol{\beta}}_n$ generally fails to converge to $\boldsymbol{\beta}$. Assuming that $\operatorname{plim} \frac{1}{n}\sum_i \mathbf{X}_i\epsilon_i = \mathbf{C}$, a $k \times 1$ vector, the probability limit is

$$\operatorname{plim} \hat{\boldsymbol{\beta}}_n = \boldsymbol{\beta} + \mathbf{Q}^{-1} \cdot \mathbf{C},$$

This is a case in which the probability limit of the OLS estimator exists, but is uninformative about the $\boldsymbol{\beta}$ parameter of the data-generating process. Note that we can interpret $\mathbf{Q}^{-1} \cdot \mathbf{C}$ as the probability limit of the slope coefficients from a hypothetical OLS regression of the disturbances $\boldsymbol{\epsilon}$ on $\mathbf{X}$. This way of thinking about the inconsistency can be helpful in establishing its likely direction.

Inconsistency arises whenever one or more right-hand side covariates $\mathbf{X}_i$ is correlated with the disturbance term ϵ_i, and the circumstances in which this can happen are those we discussed in Chapter 9. Omission of a relevant explanatory variable, measurement error in the covariates, time-series or panel-data models with both lagged dependent variables and serially correlated disturbances, specifications with choice variables used as covariates—all these are apt to generate inconsistency.

There are tools other than OLS that we can use to estimate $\boldsymbol{\beta}$ consistently, as we'll see later in the course. For some data-generating processes, as in panel data with so-called "fixed effects", it is possible to eliminate the source of inconsistency with simple arithmetic transformations of the $\mathbf{Y}$ and $\mathbf{X}$ data. The instrumental variables method offers an alternative to OLS that is expressly designed to address the inconsistency problem, but which requires additional data in the form of "instruments" and additional assumptions about the relationship of these instruments to the disturbance term—these variables and assumptions are often the subject of hot dispute in the applied literature. Some problems require data transformations followed by the use of instrumental variables. We'll have much to say about these issues in the coming months.

14.7 A Note on Strong Convergence

In our discussion to this point we have been concerned with convergence in probability, or what some authors term *weak convergence*. There is a related but more powerful concept, variously termed *strong convergence* and *almost-sure convergence*, that you will encounter in advanced treatments of econometrics. Although we will not make further use of the strong convergence concept, a brief discussion is warranted here.

According to the weak convergence concept we have been using, we say that the sequence $X_n \xrightarrow{p} c$, a constant, if for all $\epsilon > 0$, we have $\lim_{n\to\infty} \Pr(\mid X_n - c \mid > \epsilon) = 0$, or alternatively, $\lim_{n\to\infty} \Pr(\mid X_n - c \mid \leq \epsilon) = 1$. The probability that enters these expressions depends only on the marginal distribution of the single random variable X_n. By contrast, when invoking the strong concept of convergence, we say that $X_n \xrightarrow{a.s.} c$ if for all $\epsilon > 0$, we have $\lim_{n\to\infty} \Pr(\mid X_m - c \mid > \epsilon \quad \forall m \geq n) = 0$. Here the probability expression has to do with the *joint* distribution of the random variables $X_n, X_{n+1}, X_{n+2}, \ldots$. As you can well imagine, some sophisticated mathematical tools are needed to handle the joint distributions of such infinite collections of random variables.

Strong convergence implies weak convergence, but the reverse is not true. Bierens (2004, p. 144) supplies a simple example in which an $\{X_n\}$ sequence converges weakly in probability to the constant 0, but does not converge in the strong sense. Let $X_n = (1/n)U_n$ with the random variable $U_n > 0$ and let the $\{U_n\}$ sequence be iid with cdf $F(u) = e^{-\frac{1}{u}}$. Because X_n is positively-valued, $\Pr(\mid X_n \mid \leq \epsilon) = \Pr(X_n \leq \epsilon) = \Pr(U_n \leq n \cdot \epsilon)$ and, using the assumed cdf, this last expression is

$$\Pr(U_n \leq n \cdot \epsilon) = e^{-\frac{1}{n\cdot\epsilon}},$$

which approaches 1 as n goes to infinity. That is, $X_n \xrightarrow{p} 0$ as we would expect from the set-up of this particular example. Oddly, though, despite what our intuition about things would suggest, X_n does not converge almost surely to zero. We have $\Pr(X_m \leq \epsilon \quad \forall m \geq n) = \Pr(U_m \leq m \cdot \epsilon \quad \forall m \geq n)$, and by independence,

$$\Pr(U_m \leq m \cdot \epsilon \quad \forall m \geq n) = \Pi_{m\geq n}^{\infty} e^{-\frac{1}{\epsilon m}} = e^{-\frac{1}{\epsilon}\cdot\Sigma_{m\geq n}^{\infty}\frac{1}{m}}.$$

You may remember from a math course that the sum $\sum_{m\geq n}^{\infty} \frac{1}{m}$ diverges, going to positive infinity. Therefore, the limit of $\Pr(U_m \leq m \cdot \epsilon \quad \forall m \geq n)$ is zero!

14.8 A Note on Uniform Laws of Large Numbers

This concept will surface in our later discussion on maximum likelihood methods. (See Ruud (2000, Chapter 15) for a preview of the issues.) You may have studied the concept of uniform convergence with respect to non-random functions (we reviewed this concept in Chapter 2). Consider a sequence of such functions $\{f_n(\theta)\}$, $\theta \in \Theta$, in which the subscript n allows the form of the function to change as the sequence proceeds. We say that $\{f_n(\theta)\}$ converges uniformly to the limit function $f(\theta)$ if

$$\lim_{n\to\infty} \sup_{\theta\in\Theta} |f_n(\theta) - f(\theta)| = 0.$$

To handle random sequences $\{f_n(\theta)\}$, we switch from simple limits to probability limits, so that if

$$\operatorname{plim} \sup_{\theta\in\Theta} |f_n(\theta) - f(\theta)| = 0.$$

we say that the sequence $\{f_n(\theta)\}$ converges uniformly in probability to the limit function $f(\theta)$.

Now consider a function $g(Y_i, \theta)$ that is continuous in $\theta \in \Theta$, with Θ being closed and bounded. Let $\{Y_i\}$ be a sequence of iid random variables. If the expected value

$$\mathrm{E} \sup_{\theta \in \Theta} |g(Y, \theta)|$$

exists, then the function $\mathrm{E}\ g(Y, \theta)$ is continuous in $\theta \in \Theta$ and the sample average

$$n^{-1} \sum_{i}^{n} g(Y_i, \theta) \xrightarrow{p} \mathrm{E}\ g(Y, \theta)$$

uniformly. Here $n^{-1} \sum_{i=1}^{n} g(Y_i, \theta)$ plays the role of $f_n(\theta)$ above and $\mathrm{E}\, g(Y, \theta)$ plays the role of the limit function $f(\theta)$. See Bierens (2004, Appendix 6.A) for the proof, which is rather involved. We will use this uniform law, abbreviated as ULLN, in developing the properties of maximum likelihood and related estimators.

Chapter 15

Central Limit Theorems

15.1 Convergence in Distribution

The concept of *convergence in distribution* is very different from that of convergence in probability. Let $F_n(x)$ denote the cdf of X_n and let $G(y)$ denote the cdf of the random variable Y. We say that the sequence of random variables $\{X_n\}$ converges in distribution to Y if

$$\lim_{n\to\infty} F_n(x) = G(x)$$

for all points x at which G is continuous. The convergence concept used here is that of the "point-wise" convergence of functions. We represent this kind of convergence by $X_n \xrightarrow{d} Y$, a notation that takes some getting used to. When we write $X_n \xrightarrow{d} Y$, we mean only that the sequence of cdfs $\{F_n\}$ associated with $\{X_n\}$ assumes (in the limit) the same functional form as G, the cdf of Y. Convergence in distribution does not imply that Y and the random variables in $\{X_n\}$ are linked in any other way. Indeed, Y could be fully independent of all of the $\{X_n\}$ and yet we could have $X_n \xrightarrow{d} Y$.

The caveat about "continuity points" is required to cover cases in which the limiting distribution G has a discontinuous cdf. Davidson and MacKinnon (1993, pp. 107–108) give an instructive example showing why points of discontinuity in G are ignored. Suppose that $X_n \sim \mathcal{N}(0, n^{-1})$. Obviously $X_n \xrightarrow{p} 0$, a constant with cdf $G(x) = 0$ for $x < 0$ and $G(x) = 1$ for $x \geq 0$. It is only sensible to say that X_n converges in distribution to the constant zero, a degenerate random variable with G its cdf. Yet the limit of $F_n(x)$ does not equal $G(x)$ for all x values. Note that $\sqrt{n}X_n \sim \mathcal{N}(0,1)$ and that

$$F_n(x) = \Pr(X_n \leq x) = \Pr(\sqrt{n}X_n \leq \sqrt{n}x) = \Phi(\sqrt{n}x),$$

with $\Phi()$ denoting the cdf of a standard normal random variable. Now for strictly negative values of $x < 0$, we have $\lim_{n\to\infty} F_n(x) = \lim_{n\to\infty} \Phi(\sqrt{n}x) = 0$, and for strictly positive values $x > 0$, $\lim_{n\to\infty} F_n(x) = 1$. Hence, for all $x \neq 0$, the limit of $F_n(x)$ equals $G(x)$. But at $x = 0$, $\lim_{n\to\infty} F_n(0) = \Phi(0) = 1/2$ whereas $G(0) = 1$. Although we cannot say that $\lim_{n\to\infty} F(x) = G(x)$ for all x values, in defining convergence in distribution we agree to ignore points such as $x = 0$ where G is discontinuous.

It can be shown that if $\{X_n\}$ converges to Y in probability, then $\{X_n\}$ also converges to Y in distribution (keeping in mind the caveat about discontinuities). This is intuitively obvious, although a formal proof is more difficult than you might expect.[1] Here is an example that should convey the basic ideas. Let $X_n = \frac{1}{n} + Z$ with Z being standard normal. The cdf of a standard normal distribution is again denoted by Φ. Since X_n converges in probability to Z, we would expect it to converge in distribution as well. The result is easy to derive for this example. From $F_n(x) = \Pr(X_n \leq x) = \Pr(X_n - \frac{1}{n} \leq x - \frac{1}{n}) = \Phi(x - \frac{1}{n})$, we see that as $n \to \infty$ the right-most expression indeed converges to $\Phi(x)$ as we had anticipated it would.

Although convergence in distribution does not in general imply convergence in probability, there is one special case in which it does. If the sequence $\{X_n\}$ converges in distribution to the constant c, that is, to a degenerate random variable with all of its probability mass concentrated at c, then $\operatorname{plim} X_n = c$. The proof that $X_n \xrightarrow{d} c$ implies $X_n \xrightarrow{p} c$ is straightforward. As noted earlier, a degenerate random variable with all probability mass at c has a cdf $G(x)$ such that $G(x) = 0$ for $x < c$ and $G(x) = 1$ for $x \geq c$. Now, for any $\epsilon > 0$,

$$\Pr(\mid X_n - c \mid > \epsilon) = \Pr(X_n < c - \epsilon) + \Pr(X_n > c + \epsilon),$$

and we have $\Pr(X_n < c - \epsilon) \leq \Pr(X_n \leq c - \epsilon) = F_n(c - \epsilon)$. Also, $\Pr(X_n > c + \epsilon) = 1 - F_n(c + \epsilon)$. Therefore,

$$\Pr(\mid X_n - c \mid > \epsilon) \leq F_n(c - \epsilon) + 1 - F_n(c + \epsilon),$$

By the definition of convergence in distribution to c, we have $\lim_{n\to\infty} F_n(c - \epsilon) = 0$ and $\lim_{n\to\infty} F_n(c + \epsilon) = 1$. It follows that $\lim_{n\to\infty} \Pr(\mid X_n - c \mid > \epsilon) = 0$, or to put the result more succinctly, $X_n \xrightarrow{p} c$. For alternative presentations of this proof, see Ruud (2000, p. 274) or Mittelhammer (1996, Theorem 5.8, page 246).

The extension of these ideas to sequences of random vectors $\{\mathbf{X}_n\}$ is straightforward. Note, however, that when we work out the probability limits of random vectors, we can proceed on an element-by-element basis, or alternatively we can convert such vectors to random scalars using $\|\mathbf{X}_n\|$ and carry out the analysis on these scalars. Approaches such as these will not do where convergence in distribution is concerned: we must deal simultaneously with all of the arguments of $F_n(\mathbf{X}_n)$ and $G(\mathbf{Y})$.

There is an important *continuity rule* for convergence in distribution that is similar to the one we saw earlier for convergence in probability. If the function $f(x)$ is continuous, then $X_n \xrightarrow{d} X$ implies $f(X_n) \xrightarrow{d} f(X)$. See Mittelhammer (1996, Theorem 5.3, page 240) for this result.

The following theorems—sometimes termed the *Slutzky theorems*—have to do with combinations of convergence in probability and in distribution. We list them without proof, but see Mittelhammer (1996, Theorem 5.10) and Ruud (2000, pp. 275–277) for discussion.

Sum Rule If $\operatorname{plim} X_n = c$ and $\{Y_n\}$ converges in distribution to Y, then the sequence $\{X_n + Y_n\}$ converges in distribution to $c + Y$.

[1] See Serfling (1980) or Greenberg and Webster (1983, p. 13) for the proof.

Product Rule 1 If $\operatorname{plim} X_n = c$ and $\{Y_n\}$ converges in distribution to Y, then the sequence $\{X_n Y_n\}$ converges in distribution to cY.

Product Rule 2 If $\operatorname{plim} X_n = 0$ and $\{Y_n\}$ converges in distribution to Y, then the probability limit of $X_n Y_n = 0$. (Note the differences between Product Rules 1 and 2.)

Quotient Rule If $\operatorname{plim} X_n = c \neq 0$ and $\{Y_n\}$ converges in distribution to Y, then $\{Y_n / X_n\}$ converges in distribution to Y/c.

Difference Rule If $\operatorname{plim}(X_n - Y_n) = 0$ and $\{Y_n\}$ converges in distribution to Y, then $\{X_n\}$ also converges in distribution to Y. Stated differently, if $X_n \stackrel{a}{=} Y_n$ and $Y_n \xrightarrow{d} Y$, then $X_n \xrightarrow{d} Y$.

As Mittelhammer (1996, p. 247) notes, several of the rules above are special cases of a more general proposition. Let $X_n \xrightarrow{d} X$, $Y_n \xrightarrow{p} y$, a constant, and let $\{a_n\}$ be a sequence of constants that converges to a. If $g(X_n, Y_n, a_n)$ is a continuous function, then it can be shown that $g(X_n, Y_n, a_n) \xrightarrow{d} g(X, y, a)$.

15.2 The Cramér–Wold Device

This "device," which is a mathematician's term for a useful tool, may be stated as follows. A sequence of $k \times 1$ random vectors $\{\mathbf{X}_n\}$ converges in distribution to the random vector $\mathbf{X}$ if and only if $\mathbf{a}'\mathbf{X}_n \xrightarrow{d} \mathbf{a}'\mathbf{X}$ for any $\mathbf{a} \neq \mathbf{0}$, with $\mathbf{a}$ being a $k \times 1$ vector of constants. In consequence, if we are interested in the limiting distribution of a vector of random variables, and are able to work out the properties of the associated sequence of scalar random variables $\{Y_n = \mathbf{a}'\mathbf{X}_n\}$ and find its limiting distribution, we can then determine the limiting distribution of the vector. This very nice result is presented in Mittelhammer (1996, Theorem 5.36). He proves it in the following way.

Sufficiency Supposing that

$$\mathrm{MGF}_{\mathbf{a}'\mathbf{X}_n}(t) \to \mathrm{MGF}_{\mathbf{a}'\mathbf{X}}(t),$$

we show that this implies

$$\mathrm{MGF}_{\mathbf{X}_n}(\boldsymbol{\tau}) \to \mathrm{MGF}_{\mathbf{X}}(\boldsymbol{\tau})$$

for t a scalar and $\boldsymbol{\tau}$ a $k \times 1$ vector. Note that

$$\begin{aligned} \mathrm{MGF}_{\mathbf{a}'\mathbf{X}_n}(t) &\equiv \mathrm{E}\exp(t \cdot \mathbf{a}'\mathbf{X}_n) \\ &= \mathrm{E}\exp(\boldsymbol{\tau}'\mathbf{X}_n) \\ &= \mathrm{MGF}_{\mathbf{X}_n}(\boldsymbol{\tau}) \end{aligned}$$

for $\boldsymbol{\tau} \equiv t \cdot \mathbf{a}$. Hence, for a given t and arbitrary $\mathbf{a}$, convergence of $\mathrm{MGF}_{\mathbf{a}'\mathbf{X}_n}(t) \to \mathrm{MGF}_{\mathbf{a}'\mathbf{X}}(t)$ implies $\mathrm{MGF}_{\mathbf{X}_n}(\boldsymbol{\tau}) \to \mathrm{MGF}_{\mathbf{X}}(\boldsymbol{\tau})$ assuming of course that the moment-generating functions exist. As you know, random variables with the same moment-generating functions have essentially the same cumulative distributions.

Necessity This follows from our earlier theorem on the convergence in distribution of continuous functions of random variables. That is, for g a continuous function,

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X} \text{ implies } g(\mathbf{X}_n) \xrightarrow{d} g(\mathbf{X}).$$

In the case at hand, $g(\mathbf{X}_n) = \mathbf{a}'\mathbf{X}_n$.

15.3 The Lindeberg–Levy CLT

We will prove this central limit theory assuming that the random variables in question possess moment-generating functions. Because moment-generating functions are not guaranteed to exist, formal proofs instead make use of characteristic functions. The method of proof is similar, however, and the logic is a bit easier to grasp for moment-generating functions. The proof we present here is modelled on that of Taylor (1974, pp. 184–186, 331–332)

Let the sequence of random variables $\{Y_i\}$ be iid with $\mathrm{E}\, Y_i = \mu$ and $\mathrm{Var}\, Y_i = \sigma^2$. Denote by $\hat{\mu}_n$ the average of the sequence through n terms. In what follows, we will examine the limiting distribution not of $\hat{\mu}_n$ as such, but rather of

$$\frac{\sqrt{n}\,(\hat{\mu}_n - \mu)}{\sigma} = \sqrt{n} \cdot \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right).$$

Because $\mathrm{Var}\, \hat{\mu}_n = \sigma^2/n$, this expression has mean zero and unit variance. Let $Z_i \equiv (Y_i - \mu)/\sigma$ and let $M_Z(t)$ be its moment-generating function. Because we have linearly transformed (via the factor $\sqrt{n}/n = 1/\sqrt{n}$) the sum of n independent and identically distributed Z_i, we can write the moment-generating function of the transformed variable as

$$\left(M_Z\left(\frac{t}{\sqrt{n}} \right) \right)^n.$$

To prepare for what comes next, recall from the properties of moment-generating functions that $M_Z(0) = 1$, $M'_Z(0) = \mathrm{E}\, Z = 0$, and $M''_Z(0) = \mathrm{E}\, Z^2 = 1$. We will also need a result on limits that was discussed in Chapter 2, where we showed that when $\lim_{n\to\infty} a_n = a$, then

$$\lim_{n\to\infty} \left(1 + \frac{a_n}{n} \right)^n = e^a.$$

We now put these results to use in proving the Lindeberg–Levy theorem.

Begin by Taylor-expanding the moment-generating function $M_Z(t/\sqrt{n})$ to second order,

$$M_Z(t/\sqrt{n}) = M_Z(0) + M'_Z(0) \cdot \frac{t}{\sqrt{n}} + \frac{1}{2} \cdot M''_Z(\tilde{t}_n) \cdot \frac{t^2}{n}$$

with $\tilde{t}_n$ lying between $t/\sqrt{n}$ and zero. Because $\mathrm{E}\, Z = 0$, this simplifies to

$$M_Z(t/\sqrt{n}) = 1 + \frac{1}{2} \cdot M''_Z(\tilde{t}_n) \cdot \frac{t^2}{n}.$$

Since $\lim_{n\to\infty} M_Z''(\tilde{t}_n) = M_Z''(0) = \mathrm{E}\, Z^2 = 1$, the limit of $M_Z''(\tilde{t}_n) \cdot (t^2/2) = t^2/2$. Hence,

$$\lim_{n\to\infty} M_{Y_n}(t) = \lim_{n\to\infty} \left(1 + \frac{M_Z''(\tilde{t}_n) \cdot (t^2/2)}{n}\right)^n = e^{\frac{t^2}{2}},$$

which is the moment-generating function of a standard normal random variable. We have proved that the transformed sample mean

$$\frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0,1),$$

which is a truly remarkable result given how little has been assumed.[2] We can also write the result in a slightly easier-to-remember form, which is $\sqrt{n}(\hat{\mu}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$. Note that this way of expressing the result relies on one of the Slutzky theorems.

15.4 Lindeberg's CLT

This more advanced central limit theorem is discussed in Mittelhammer (1996, Theorem 5.31). Let $\{Y_i\}$ be a sequence of independent but *not* identically distributed random variables, with $\mathrm{E}\, Y_i = \mu_i$ and $\mathrm{Var}\, Y_i = \sigma_i^2$. Denote the average mean and average variance by $\bar{\mu}_n = (1/n)\sum_i^n \mu_i$ and $\bar{V}_n = (1/n)\sum_i \sigma_i^2$. Also, let the sum of the variances be $V_n = \sum_i \sigma_i^2$. Consider

$$\hat{\mu}_n - \bar{\mu}_n = \frac{1}{n}\sum_{i=1}^n (Y_i - \mu_i),$$

which has zero mean and variance

$$\mathrm{Var}\,(\hat{\mu}_n - \bar{\mu}_n) = \frac{1}{n^2}\sum_{i=1}^n \sigma_i^2 = \frac{1}{n}\bar{V}_n.$$

Dividing $\hat{\mu}_n - \bar{\mu}_n$ by the square root of its variance gives us a transformed random variable with a mean of zero and unit variance,

$$\frac{\sqrt{n}(\hat{\mu}_n - \bar{\mu}_n)}{\sqrt{\bar{V}_n}}.$$

The Lindeberg theorem shows that

$$\frac{\sqrt{n}(\hat{\mu}_n - \bar{\mu}_n)}{\sqrt{\bar{V}_n}} \xrightarrow{d} \mathcal{N}(0,1) \tag{15.1}$$

provided that a condition termed the *Lindeberg condition* is satisfied. The condition is that for all $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{V_n}\sum_{i=1}^n \int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} (y_i - \mu_i)^2 f_i(y_i)\, dy_i = 0. \tag{15.2}$$

[2] Woodroofe (1975, p. 254) presents a similar proof using a Taylor expansion to third order, and Mittelhammer (1996, Theorem 5.30) develops an unusual proof in which L'Hospital's Rule is applied to the moment-generating function.

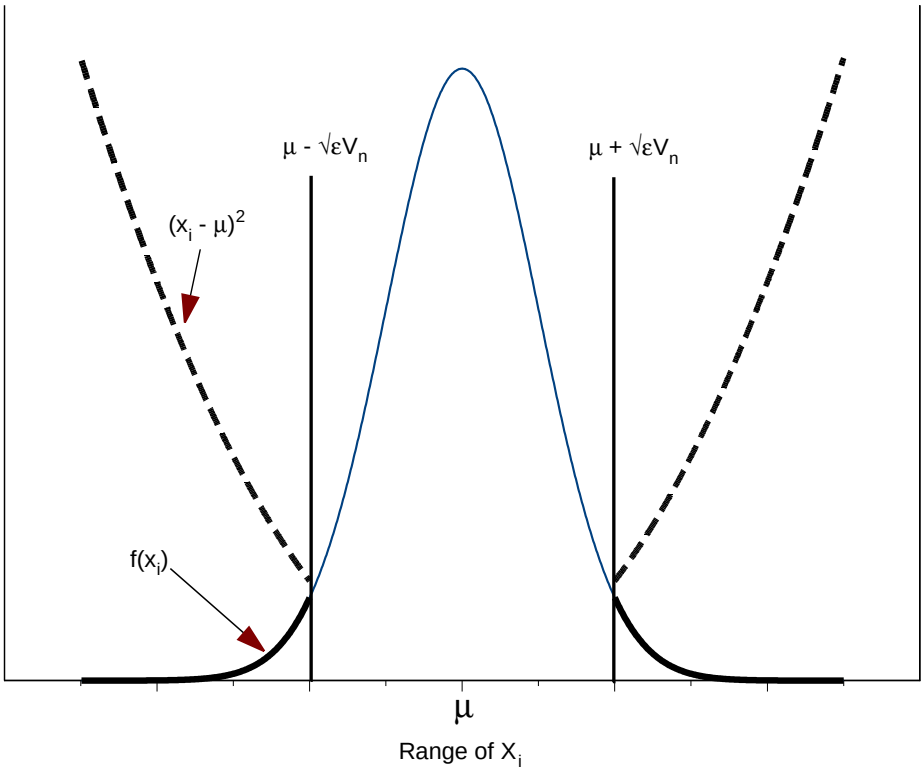


Figure 15.1: Essence of the Lindeberg condition

In this form the meaning of the condition is not easily comprehended, and several authors have found more intuitive conditions that, when met, imply that the Lindeberg condition is satisfied. (We note in passing that some authors write the range of integration as $|y_i - \mu_i| \geq \epsilon\sqrt{V_n}$, which is equivalent to the expression used above if ϵ is redefined.) Figure 15.1 depicts the essentials of the Lindeberg condition for the iid case with $\mathrm{E}\, Y_i = \mu$. Here the range of integration can be represented in terms of $\mu \pm \sqrt{\epsilon \cdot V_n}$, shown in the two vertical bars. The term $(y_i - \mu)^2$ is indicated in dashed lines, and the relevant portion of the density function is depicted in the thick solid lines. In the iid case, the density function would remain fixed as $n \to \infty$, but the vertical bars indicating the range of integration would move outwards as the sum $V_n = \sum_{i=1}^{n} \sigma_i^2$ grows, and in the integral the higher values of $(y_i - \mu)^2$ would be weighted by lower values of the density (at least in the case shown).

Bounded random variables

As the figure suggests, if the random variables Y_i in the sequence are all bounded, and if V_n grows with n, then eventually $\epsilon \cdot V_n$ will become so large that no y_i values exist such that $(y_i - \mu_i)^2 \geq \epsilon \cdot V_n$. When $\epsilon \cdot V_n$ gets this large, the integral in (15.2) falls to zero for all i and the Lindeberg condition is satisfied. The following argument, taken from Mittelhammer (1996, Theorem 5.32), formalizes this logic.

The theorem states that if $\Pr(|Y_i| \leq m) = 1$ for all i (in other words, the probability of $|Y_i|$ exceeding m is zero), and if $V_n = \sum_{i=1}^{n} \sigma_i^2 \to \infty$ as $n \to \infty$, then the Lindeberg condition

is satisfied. To see this, note first that $|y_i| < m$ implies $|y_i - \mu_i| < 2m$. (Draw a picture if you need to understand why.) Hence,

$$\int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} (y_i - \mu_i)^2 f(y_i) dy_i \leq \int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} 4m^2 f(y_i) dy_i$$

because $|y_i - \mu_i| < 2m$ implies $(y_i - \mu_i)^2 < 4m^2$. Going further,

$$\begin{aligned}
\int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} 4m^2 f(y_i) dy_i &= 4m^2 \Pr((y_i - \mu_i)^2 \geq \epsilon \cdot V_n) \\
&\leq 4m^2 \frac{\mathrm{E}((Y_i - \mu_i)^2)}{\epsilon \cdot V_n} \\
&= \frac{4m^2 \sigma_i^2}{\epsilon \cdot V_n}
\end{aligned}$$

by Markov's inequality. Summing over i and dividing by V_n,

$$\begin{aligned}
\frac{1}{V_n} \sum_{i=1}^{n} \int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} (y_i - \mu_i)^2 f(y_i) dy_i &\leq \frac{1}{V_n} \cdot \sum_{i=1}^{n} \frac{4m^2 \sigma_i^2}{\epsilon \cdot V_n} \\
&= \frac{1}{V_n} \frac{4m^2}{\epsilon \cdot V_n} \sum_{i=1}^{n} \sigma_i^2 \\
&= \frac{4m^2}{\epsilon \cdot V_n}.
\end{aligned}$$

By assumption, $V_n \to \infty$ as $n \to \infty$, and thus the Lindeberg condition is satisfied.

Note how important it is that the variances $\sigma_i^2 > 0$. If, as the sequence proceeds, it should happen that the random variables Y_i become constants (such that $Y_i = \mu_i$), then after a point V_n will cease to grow and the Lindeberg condition may not be met.

We often see the Lindeberg CLT expressed in a slightly different form that requires one additional assumption. The theorem is stated in terms involving

$$\frac{1}{\sqrt{\bar{V}_n}} \cdot \sqrt{n}(\hat{\mu}_n - \bar{\mu}_n) \xrightarrow{d} \mathcal{N}(0, 1)$$

provided that the Lindeberg condition is met. Suppose that we now add the assumption that $\lim_{n\to\infty} \bar{V}_n = V$, that is, as $n \to \infty$ the average variance approaches a constant. Then by virtue of the Slutzky theorems, we can write the Lindeberg result as

$$\sqrt{n}(\hat{\mu}_n - \bar{\mu}_n) \xrightarrow{d} \mathcal{N}(0, V).$$

We'll often use this version of the result.

The Liapounov CLT

This CLT also provides a limiting condition that ensures that the Lindeberg condition (15.2) is met. With the sequence $\{Y_i\}$ having the properties defined in the Lindeberg CLT, the Lindeberg condition is satisfied if for *some* $\delta > 0$,

$$\lim_{n\to\infty} \frac{\sum_{i=1}^{n} \mathrm{E}\,|Y_i - \mu_i|^{2+\delta}}{V_n^{1+\delta/2}} = 0. \tag{15.3}$$

The proof proceeds through a series of inequalities associated with the Lindeberg condition.

$$\int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} (y_i - \mu_i)^2 f(y_i) dy_i =$$
$$\int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} |y_i - \mu_i|^{-\delta} |y_i - \mu_i|^{\delta} (y_i - \mu_i)^2 f(y_i) dy_i =$$
$$\int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} |y_i - \mu_i|^{-\delta} |y_i - \mu_i|^{2+\delta} f(y_i) dy_i.$$

Because $(y_i - \mu_i)^2 \geq \epsilon \cdot V_n$ over the range of integration, it follows that $|y_i - \mu_i| \geq (\epsilon \cdot V_n)^{1/2}$ and hence that $|y_i - \mu_i|^{-\delta} \leq (\epsilon \cdot V_n)^{-\delta/2}$, where you should note the reversal of the inequality. Hence, the last integral shown above is

$$\leq \int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} (\epsilon \cdot V_n)^{-\delta/2} |y_i - \mu_i|^{2+\delta} f(y_i) dy_i,$$

and this, in turn, is

$$\leq (\epsilon \cdot V_n)^{-\delta/2} \int_{-\infty}^{\infty} |y_i - \mu_i|^{2+\delta} f(y_i) dy_i = (\epsilon \cdot V_n)^{-\delta/2} \cdot \mathrm{E} |y_i - \mu_i|^{2+\delta}.$$

Insert this into the expression for the Lindeberg condition (15.2), and we obtain

$$\frac{1}{V_n} \sum_{i=1}^{n} \int_{(y_i-\mu_i)^2 \geq \epsilon \cdot V_n} (y_i - \mu_i)^2 f(y_i) dy_i \leq$$
$$\frac{1}{V_n} \cdot \sum_{i=1}^{n} (\epsilon \cdot V_n)^{-\delta/2} \cdot \mathrm{E} |y_i - \mu_i|^{2+\delta}$$

The limit of the right-hand side can be written as

$$\lim_{n \to \infty} \epsilon^{-\delta/2} \left(\frac{1}{V_n^{1+\delta/2}} \sum_{i=1}^{n} \mathrm{E} |y_i - \mu_i|^{2+\delta} \right).$$

By assumption, the limit of the expression in parentheses is zero, and thus the Lindeberg condition is met.

A Liapounov example

Suppose that $Y_i \sim \text{uniform}(-c_i, c_i)$, with $c_i > 0$ and bounded by $c_i \in [\tau, m]$. Therefore Y_i has zero mean and $\mathrm{Var}\, Y_i = \frac{1}{3} c_i^2$ as can be readily verified. Since $\mu_i = 0$ and $c_i \leq m$, it follows that $\mathrm{E} |Y_i - \mu_i|^{2+\delta} \leq m^{2+\delta}$. Also, $V_n = \sum_i^n \mathrm{Var}\, Y_i = \frac{1}{3} \sum_{i=1}^n c_i^2$.

The theorem focuses on

$$\lim_{n \to \infty} \left(\frac{1}{V_n^{1+\delta/2}} \sum_{i=1}^{n} \mathrm{E} |y_i - \mu_i|^{2+\delta} \right).$$

Note that because $c_i \geq \tau$, it follows that $V_n = \frac{1}{3} \sum_{i=1}^n c_i^2 \geq \frac{n\tau^2}{3}$.

Choose the value of $\delta = 1$. Replace $\sum_i^n \mathrm{E}\,|Y_i - \mu_i|^{2+\delta}$ with the larger number $\sum_{i=1}^n m^{2+\delta} = n\,m^3$. Replace $V_n^{1+\delta/2}$ with the smaller number $(n\tau^2/3)^{3/2}$. Thus, remembering that $\delta = 1$, we have

$$\lim_{n\to\infty} \frac{1}{V_n^{3/2}} \sum_{i=1}^n \mathrm{E}\,|y_i - \mu_i|^{3/2} \le \lim_{n\to\infty} \frac{nm^3}{(n\tau^2/3)^{3/2}}$$
$$= \lim_{n\to\infty} \frac{1}{\sqrt{n}} \cdot \frac{m^3}{\tau^3 3^{3/2}} = 0.$$

Hence the Lindeberg condition is met. Of course, this is an artificial example, and because Y_i is bounded, a simpler proof was available to us.

See Mittelhammer (1996, pp. 274–278), Serfling (1980, p. 29), Greenberg and Webster (1983, pp. 19–21), and Spanos (1986, p. 174) for further discussion and applications. Mittelhammer (1996) is especially good on how alternative forms of the Lindeberg condition yield a family of useful central limit theorems.

15.5 The Limiting Distribution of $\sqrt{n}(\hat{\beta}_n - \beta)$

The essentials of the argument are perhaps easiest to grasp for the simple model $Y_i = \beta X_i + \epsilon_i$, in which both X_i and β are scalars. (Extensions to multiple covariates require the Cramér-Wold device, which we will use after the essentials of the proofs have been laid out here.) We will maintain the fundamental assumption that justifies ordinary regression analysis, that $\mathrm{E}(\epsilon_i|X_i) = 0$. Also, the sequence of pairs of random variables $\{(X_i, \epsilon_i)\}$ will be assumed to be *independent* over i, this being a stronger condition than is needed for the analysis of estimator consistency. We'll present first the proof for the iid case, and then address the inid case, which is of far greater practical value.

The iid case

For the simple model $Y_i = \beta X_i + \epsilon_i$ with only one explanatory variable, assume the sequence $\{(X_i, \epsilon_i)\}$ is iid. We can write

$$\hat{\beta}_n = \beta + \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n X_i\epsilon_i\right), \tag{15.4}$$

and therefore

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^{-1} \left(\sqrt{n}\frac{1}{n}\sum_{i=1}^n X_i\epsilon_i\right). \tag{15.5}$$

By Khinchin's law of large numbers for iid sequences,

$$\frac{1}{n}\sum_{i=1}^n X_i^2 \xrightarrow{p} q,$$

where $q \equiv \mathrm{E}\, X_i^2$, and using the results about probability limits of continuous functions of random variables (the inverse being one such function), we obtain

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{-1} \xrightarrow{p} q^{-1}.$$

That takes care of the first factor on the right-hand side of equation (15.5).

Consider the second factor,

$$\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i.$$

What sort of central limit theorem might apply to this? Because $\mathrm{E}\, X_i\epsilon_i = 0$ and $\mathrm{Var}\, X_i\epsilon_i = \sigma^2 \mathrm{E}\, X_i^2 = \sigma^2 q$, we see that we are dealing with $\sqrt{n}$ times the average of n indepdendent, mean zero, variance $\sigma^2 q$ random variables. Evidently the Lindeberg-Levy CLT applies, and it yields

$$\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i \xrightarrow{d} \mathcal{N}(0, \sigma^2 q).$$

That takes care of the second factor.

Combining both results, we find that

$$\sqrt{n}(\hat{\beta}_n - \beta) \stackrel{a}{=} q^{-1} \cdot \mathcal{N}(0, \sigma^2 q),$$

which in turn implies

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 q^{-1}). \tag{15.6}$$

Before we move on to the inid case, let's re-examine this proof from a slightly different angle. Begin with

$$\hat{\beta}_n - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i\right)$$

and note that

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \epsilon_i\right) = \frac{1}{n}\bar{V}_n,$$

with $(1/n)\bar{V}_n = (1/n) \cdot (1/n)\sum_i q\sigma^2 = q\sigma^2/n$. If we normalize $\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i$ by the square root of its variance, we can write

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{-1} \left(\frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i}{\sqrt{q\sigma^2}}\right) \cdot \sqrt{q\sigma^2}.$$

The Lindeberg–Levy theorem applies directly to the middle factor in parentheses, and we see that

$$\sqrt{n}(\hat{\beta}_n - \beta) \stackrel{a}{=} q^{-1} \cdot \mathcal{N}(0,1) \cdot \sqrt{q\sigma^2},$$

from which we obtain $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 q^{-1})$. This is the approach we will now apply to the inid case.

The inid case

In your research, you will generally be confronted with data sequences whose members are independent but *not* identically distributed, and so it is especially important to get a sense of how the Lindeberg central limit theorem is applied. We'll consider two specifications: in the first the variance of the disturbance term is assumed to be constant (*homoskedasticity*) and in the second we generalize things to allow for non-constant variances (*heteroskedasticity*). In both of these cases, however, we will need to allow $\mathrm{E}\, X_i^2$ to vary freely (in an unspecified way) with i. The assumption of homoskedasticity buys you a little bit of simplification, but not much.

Homoskedastic disturbances

We maintain the assumptions that $\mathrm{E}(\epsilon_i|X_i) = 0$ and $\mathrm{E}(\epsilon_i^2|X_i) = \sigma^2$. Even with these assumptions, the analysis is more complicated than in the iid case, because we must now allow for the mean and higher moments of X_i to vary with i, and although $\mathrm{E}(\epsilon_i^2|X_i) = \sigma^2$ is assumed constant over i, other moments of ϵ_i may not be. The powerful Lindeberg central limit theorem is needed to handle this case, and you'll recall that bounds on the random variables imply that the Lindeberg condition is met. Hence, to deal with the case at hand we need to add a new assumption, that both X_i and ϵ_i are *bounded*.[3]

Begin as before with

$$\hat{\beta}_n - \beta = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{-1} \cdot \left(\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i\right). \tag{15.7}$$

As we now understand, the key factor in this expression is $(1/n)\sum_{i=1}^{n} X_i\epsilon_i$. By iterated expectations, each of the $X_i\epsilon_i$ terms in the sum has mean zero. Let $q_i = \mathrm{E}\, X_i^2$, in which you should take note of the i subscript. Applying iterated expectations again, we see that the variance of each term is $\mathrm{E}\,\epsilon_i^2 X_i^2 = \sigma^2 q_i$. Hence, the expected value of $(1/n)\sum_{i=1}^{n} X_i\epsilon_i$ is zero and its variance is

$$\frac{1}{n}\bar{V}_n = \frac{1}{n}\cdot\frac{1}{n}\sum_{i=1}^{n}\sigma^2 q_i = \frac{1}{n}\sigma^2\bar{q}_n$$

with $\bar{q}_n$ being the average of the q_i terms. Let's rewrite our equation to accommodate the transformation that normalizes the key factor to have mean zero and unit variance,

$$\frac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i)}{\sqrt{\bar{V}_n}}.$$

The rewritten equation is

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{-1} \cdot \left(\frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i}{\sqrt{\bar{V}_n}}\right) \cdot \sqrt{\bar{V}_n}. \tag{15.8}$$

[3] You will not often see this assumption invoked in the econometric journals, where most authors prefer to assume that some more exotic condition is satisfied (such as used in the Liapounov CLT), which in turn guarantees that the Lindeberg condition is met.

Provided that the Lindeberg condition is met, the middle factor on the right-hand size converges in distribution to $\mathcal{N}(0,1)$. That constitutes the full contribution of the Lindeberg CLT to our proof.

Having extracted what we can from Lindeberg, we still have work to do. If the equation is to deliver the results about $\sqrt{n}(\hat{\beta}_n - \beta)$ that we seek, the first factor on the right must have a probability limit, and the third factor, $\sqrt{\bar{V}_n}$, must have an ordinary limit. If these conditions are met, then the right-hand side will converge to the product of a constant, a $\mathcal{N}(0,1)$ random variable, and another constant.

Let's deal first with $\sqrt{\bar{V}_n}$. Recalling for the case at hand that

$$\bar{V}_n = \sigma^2 \frac{1}{n} \sum_{i=1}^{n} q_i = \sigma^2 \bar{q}_n,$$

we now *assume* that $\bar{V}_n \to \sigma^2 q$ with q being the limit of $\bar{q}_n$, the average of the q_i terms. We did not need this assumption to apply the Lindeberg CLT, but we need it now.

We finally turn our attention to $\frac{1}{n}\sum_{i=1}^{n} X_i^2$, which is an average of n random variables with means $\mathrm{E}\, X_i^2 = q_i$ and variances $\mathrm{E}(X_i^2 - q_i)^2$. We want a law of large numbers to apply here—and note that Khinchin's law of large numbers is no longer applicable because we have an inid data series. Nevertheless, the variances must be bounded because (by assumption) the random variables X_i are bounded. Applying a law of large numbers for heteroskedastic sequences with i-specific means, we can say

$$\frac{1}{n} \sum_{i=1}^{n} X_i^2 \stackrel{a}{=} \frac{1}{n} \sum_{i=1}^{n} q_i = \bar{q}_n.$$

Since $\frac{1}{n}\sum_{i=1}^{n} X_i^2 \stackrel{a}{=} \bar{q}_n$, and since we have already assumed that the limit of $\bar{q}_n$ exists and equals q, we have

$$\operatorname{plim} \left(\frac{1}{n} \sum_{i=1}^{n} X_i^2 \right)^{-1} = q^{-1}.$$

Hence,

$$\sqrt{n}(\hat{\beta}_n - \beta) \stackrel{a}{=} q^{-1} \cdot \mathcal{N}(0,1) \cdot \sqrt{q\sigma^2}. \tag{15.9}$$

We are at last at the finish line. We have the limiting distribution result that we have been seeking,

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, \sigma^2 q^{-1}). \tag{15.10}$$

This concluding result is dressed out in the same notation as the iid case, although the meaning of q is different and certainly the logic leading to the final result is much more complicated.

Heteroskedastic disturbances

We maintain the assumption that $\mathrm{E}(\epsilon_i | X_i) = 0$ but now allow the disturbance variances $\mathrm{E}(\epsilon_i^2 | X_i) = \sigma_i^2$ to differ across the observations. This alters $\bar{V}_n$ to

$$\bar{V}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(X_i^2 \epsilon_i^2),$$

but so long as the Lindeberg condition is met, we can still apply the Lindeberg CLT to obtain

$$\frac{\sqrt{n}(\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i)}{\sqrt{\bar{V}_n}} \xrightarrow{d} \mathcal{N}(0,1)$$

Normalizing as before, we have

$$\sqrt{n}(\hat{\beta}_n - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right)^{-1} \cdot \left(\frac{\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} X_i\epsilon_i}{\sqrt{\bar{V}_n}}\right) \cdot \sqrt{\bar{V}_n}. \tag{15.11}$$

If this expression is to deliver the result we seek, it must be the case that the first factor on the right has a probability limit, and the third factor, $\sqrt{\bar{V}_n}$, must have an ordinary limit. If we continue to assume as before that $\frac{1}{n}\sum_i q_i$ converges to a limit of q, then the analysis we carried out above for the first factor will give the result that $\frac{1}{n}\sum_i X_i^2 \xrightarrow{p} q$. We also need a new assumption, that

$$\bar{V}_n = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(X_i^2\epsilon_i^2) \to V.$$

Then

$$\sqrt{n}(\hat{\beta}_n - \beta) \stackrel{a}{=} q^{-1} \cdot \mathcal{N}(0,1) \cdot \sqrt{V}. \tag{15.12}$$

and this implies

$$\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(0, q^{-1}Vq^{-1}). \tag{15.13}$$

This is not too different from the limiting result we obtained in the homoskedastic case.

15.6 Generalizing to Multiple Covariates

Earlier we mentioned the Cramér-Wold device—how can we put this tool to work to extend our univariate results to the multivariate case in which $\hat{\boldsymbol{\beta}}_n$ is a $k \times 1$ vector? We have

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \sqrt{n}\frac{1}{n}\mathbf{X}'\boldsymbol{\epsilon},$$

and, assuming that various laws of large numbers apply to the averages of all squares and cross-products of the explanatory variables, this gives us

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \stackrel{a}{=} \mathbf{Q}^{-1} \cdot \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\epsilon_i,$$

with $\mathbf{X}_i$ being the $k \times 1$ column vector of explanatory variables for the i-th observation. For the iid case, $\mathbf{Q} = \mathrm{E}\,\mathbf{X}_i\mathbf{X}_i'$, and for the inid case, we define

$$\mathbf{Q} = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\,\mathbf{X}_i\mathbf{X}_i'$$

implicitly assuming that this limit exists.

To examine the behavior of the second factor, we follow the Cramer-Wold approach and define the scalar random variable $Z_i = \mathbf{a}'\mathbf{X}_i\epsilon_i$ with $\mathbf{a} \neq \mathbf{0}$ being a vector of constants. Note that $\mathrm{E}\, Z_i = 0$ by iterated expectations and $\mathrm{Var}\, Z_i = \mathbf{a}'\,\mathrm{E}(\mathbf{X}_i\epsilon_i^2\mathbf{X}_i')\mathbf{a} \equiv \mathbf{a}'\mathbf{W}_i\mathbf{a}$ in which $\mathbf{W}_i$ is a $k \times k$ matrix. To proceed, we will want to consider separately the iid and inid cases.

If the sequence $\{(\mathbf{X}_i, \epsilon_i)\}$ is iid, this implies that $\{Z_i\}$ is iid as well, and we can omit the i subscript from $\mathbf{W}_i$. For this case, we can also assume that $\mathrm{E}(\epsilon_i^2 \mid \mathbf{X}_i) = \sigma^2$ and therefore $\mathrm{Var}\, Z_i = \mathbf{a}'\sigma^2\mathbf{Q}\mathbf{a}$. Evidently we have in $\{Z_i\}$ a sequence of independent, mean zero, scalar random variables with constant variances. Using the Lindeberg–Levy CLT,

$$\sqrt{n}\frac{1}{n}\sum_i Z_i \xrightarrow{d} \mathcal{N}(0, \mathbf{a}'\sigma^2\mathbf{Q}\mathbf{a}).$$

By the Cramér–Wold device, this implies

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i\epsilon_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}).$$

Why, exactly, does this result follow? Here we rely on a study of the relevant moment-generating functions. A scalar random variable distributed as $\mathcal{N}(0, w)$ has the moment-generating function

$$M(t) = \exp\left(\frac{1}{2}t'wt\right),$$

and if we substitute $\mathbf{a}'\sigma^2\mathbf{Q}\mathbf{a}$ for w, this is

$$M(t) = \exp\left(\frac{1}{2}t'\mathbf{a}'\sigma^2\mathbf{Q}\mathbf{a}t\right) = \exp\left(\frac{1}{2}\boldsymbol{\tau}'\sigma^2\mathbf{Q}\boldsymbol{\tau}\right) = M(\boldsymbol{\tau})$$

with $\boldsymbol{\tau} = \mathbf{a} \cdot t$ being a $k \times 1$ vector. The right-hand expression is recognizable as the moment-generating function of a *multivariate* $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q})$ random vector. Our main result then follows, that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \overset{a}{=} \mathbf{Q}^{-1} \cdot \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}).$$

The analysis of the inid case proceeds in a similar fashion. For this case, recall that $\mathrm{Var}\, Z_i = \mathbf{a}'\mathbf{W}_i\mathbf{a}$ with $\mathbf{W}_i = \mathrm{E}\,\mathbf{X}_i\epsilon_i^2\mathbf{X}_i'$. The average variance is

$$\bar{V}_n = \frac{1}{n}\sum_i \mathbf{a}'\mathbf{W}_i\mathbf{a} = \mathbf{a}' \cdot \frac{1}{n}\sum_i \mathbf{W}_i \cdot \mathbf{a}$$

and we assume that

$$\lim_{n\to\infty} \frac{1}{n}\sum_i \mathbf{W}_i = \mathbf{W}.$$

The Lindeberg CLT then implies,

$$\sqrt{n}\frac{1}{n}\sum_i Z_i \xrightarrow{d} \mathcal{N}(0, \mathbf{a}'\mathbf{W}\mathbf{a}),$$

and when we examine the moment-generating function as we did above, it is evident that

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{W}).$$

Our main result for the inid case is therefore

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}) \stackrel{a}{=} \mathbf{Q}^{-1} \cdot \mathcal{N}(\mathbf{0}, \mathbf{W}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1}\mathbf{W}\mathbf{Q}^{-1}).$$

15.7 The Limiting Distribution of $\sqrt{n}(s_n^2 - \sigma^2)$

Under the assumption that $\mathrm{E}(\epsilon_i^2|\mathbf{X}_i) = \sigma^2$, write

$$\sqrt{n}(s_n^2 - \sigma^2) \stackrel{a}{=} \sqrt{n} \cdot \frac{1}{n}\left(\boldsymbol{\epsilon}'\boldsymbol{\epsilon} - \boldsymbol{\epsilon}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\epsilon}\right) - \sqrt{n}\sigma^2$$

Rearrange and re-express the right-hand side as

$$\sqrt{n}\frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} - \sqrt{n}\sigma^2 - \left(\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i\right).$$

From the assumptions made above, we know that in the iid case,

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i &\xrightarrow{p} \mathbf{0}, \\
\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\mathbf{X}_i'\right)^{-1} &\xrightarrow{p} \mathbf{Q}^{-1}, \text{ and} \\
\sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\epsilon_i &\xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q})
\end{aligned}$$

with a similar result holding for the inid case. Hence, taken together the three terms in parentheses have a probability limit of zero.

We're left with

$$\begin{aligned}
\sqrt{n}(s_n^2 - \sigma^2) &\stackrel{a}{=} \sqrt{n}\frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} - \sqrt{n}\sigma^2 \\
&= \sqrt{n} \cdot \frac{1}{n}\sum_{i=1}^{n}(\epsilon_i^2 - \sigma^2).
\end{aligned}$$

Each term in the sum has zero mean and a variance equalling $\mathrm{E}\,\epsilon_i^4 - \sigma^4$. If $\{\epsilon_i\}$ is iid, this variance is constant and the Lindeberg–Levy CLT applies, giving

$$\sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, \mu_4 - \sigma^4), \tag{15.14}$$

where $\mu_4 = \mathrm{E}\,\epsilon_i^4$. Alternatively, if $\{\epsilon_i\}$ is not identically distributed but is bounded, we can appeal to the Lindeberg CLT for the same general result,

$$\sqrt{n}(s_n^2 - \sigma^2) \xrightarrow{d} \mathcal{N}(0, \delta), \tag{15.15}$$

in which $\delta \equiv \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}(\mu_{4,i} - \sigma^4)$.

Approximations

We have now seen quite a number of theoretical results about the limiting distributions of suitably transformed estimators. But as a practical matter, what are we to make of a result such as $\sqrt{n}(\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1})$? The two factors defining the variance of the limiting distribution are both unknown. To lay the foundation for hypothesis tests, we need to understand how such factors are approximated.

To review the theory set out above, the matrix $\mathbf{Q}$ stands for two related but distinct quantities depending on whether the $\{(\mathbf{X}_i, \epsilon_i)\}$ series is assumed to be iid or inid, and for the latter type of series, whether we assume that ϵ_i is conditionally homoskedastic. For iid data,

$$\mathbf{Q} = \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i',$$

in which $\mathrm{E}\, \mathbf{X}_i \mathbf{X}_i' = \mathbf{Q}$ and by the iid assumption $\mathbf{Q}$ has no i subscript. For inid homoskedastic data, however, we would have to write $\mathrm{E}\, \mathbf{X}_i \mathbf{X}_i' = \mathbf{Q}_i$ and then consider the probability limit of

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \frac{1}{n} \sum_{i=1}^{n} \left(\mathbf{X}_i \mathbf{X}_i' - \mathbf{Q}_i \right) + \frac{1}{n} \sum_{i=1}^{n} \mathbf{Q}_i.$$

Under additional assumptions, the probability limit of the first term on the right-hand side is $\mathbf{0}_{k \times k}$ by application of one or more laws of large numbers, leaving

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' \overset{a}{=} \frac{1}{n} \sum_{i=1}^{n} \mathbf{Q}_i.$$

If we assume that the right-hand side average has a limit, and agree to call that limit $\mathbf{Q}$, then we can say

$$\operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i' = \mathbf{Q},$$

a result which looks just like the iid result.

In both cases, a consistent estimator of $\mathbf{Q}$ suggests itself,

$$\hat{\mathbf{Q}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i',$$

and we have already seen that $\hat{\sigma}_n^2$ (or s_n^2 if you prefer) is consistent for σ^2.

For the inid case with heteroskedasticity, we have another unknown quantity to address, the matrix

$$\mathbf{W} = \operatorname{plim} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i^2 \mathbf{X}_i \mathbf{X}_i'.$$

What can we use to approximate $\mathbf{W}$?

White (1980) answered this question in a celebrated paper. His result is easiest to understand for the case of a single explanatory variable. White was able to show, first, that

$$\frac{1}{n} \sum_i X_i^2 \epsilon_i^2 \overset{a}{=} \frac{1}{n} \sum_i \mathrm{E}(X_i^2 \epsilon_i^2) = \bar{V}_n$$

which is not too surprising, and second, in his breakthrough result, that

$$\frac{1}{n}\sum_i X_i^2 e_i^2 \stackrel{a}{=} \bar{V}_n.$$

This is the key finding, because it involves on the left-hand side only the observed X_i variables and the ordinary least squares residuals e_i, each of them squared, which serve as proxies for the squares of the unobserved disturbances. We will provide further insight into White's approach in Chapter 24, but for the moment the important point is that we have a valid approximation to the variance of the limiting distribution,

$$\left(\frac{1}{n}\sum_i X_i^2\right)^{-1} \cdot \frac{1}{n}\sum_i X_i^2 e_i^2 \cdot \left(\frac{1}{n}\sum_i X_i^2\right)^{-1}. \tag{15.16}$$

The computational simplicity of this result has meant that the variance matrix of an ordinary least squares model is now commonly estimated allowing the disturbances to be heteroskedastic, but without imposing any particular assumptions on the form that this heteroskedasticity takes. In good-quality statistical software, *robust standard errors* make use of the variance estimate shown in (15.16). The use of robust standard errors has become common practice to such an extent that Cameron and Trivedi (2005) have written their textbook assuming that OLS regressions will generally be run with these standard errors.

15.8 Hypothesis Testing

In the asymptotic context, the chi-square distribution provides the foundation for most hypothesis testing. The three major procedures used in econometrics—the Wald, Lagrange Multiplier, and Likelihood Ratio tests—all involve test statistics that are distributed as central chi-squared under the null hypothesis, and as non-central chi-squared under the alternative hypothesis.

Before we plunge into the details of the chi-squared approach, let's reconsider the simple OLS model with one explanatory covariate. In the i.i.d. case, we have seen that $\sqrt{n}(\hat{\beta}_n - \beta) \stackrel{d}{\to} \mathcal{N}(0, \sigma^2 q^{-1})$ with $q = \mathrm{E}\, X_i^2$. Given a null hypothesis on the value of β, how can we turn this result—with its unknown parameters σ^2 and q whose values are not addressed in the null—into a test statistic based on the standard normal distribution?

Under the null hypothesis (which specifies the correct value for β),

$$\frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\sqrt{\sigma^2 q^{-1}}} \stackrel{d}{\to} \mathcal{N}(0,1).$$

As it turns out, we can get rid of the two unknowns by inserting the consistent estimator s_n^2 for σ^2 and the consistent estimator $\hat{q}_n = \frac{1}{n}\sum_i X_i^2$ for q. The resulting test statistic,

$$T = \frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\sqrt{s_n^2 \hat{q}_n^{-1}}}$$

converges to standard normal just like its theoretical counterpart using σ^2 and q. To see this, simply rewrite the test statistic T as

$$T = \left(\frac{\sqrt{\sigma^2 q^{-1}}}{\sqrt{s_n^2 \hat{q}_n^{-1}}} \right) \cdot \frac{\sqrt{n}(\hat{\beta}_n - \beta)}{\sqrt{\sigma^2 q^{-1}}}.$$

The key is to recognize that the factor in parentheses has a probability limit of 1. This argument is a nice illustration of the value of the Slutzky theorems on combinations of convergence in probability and in distribution. We'll shortly see another illustration in the context of chi-squared tests.

The following results are needed to understand the role of the chi-square distribution. The first is a result you have seen before in Chapter 2, but it will be expressed here in slightly different notation.

- Let $\mathbf{Y}$, a $k \times 1$ vector, be distributed as $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, with $\mathbf{\Sigma}$ being nonsingular (of rank k). Then the quadratic form
$$\mathbf{Y}'\mathbf{\Sigma}^{-1}\mathbf{Y} \sim \chi^2_k.$$
The proof is modified only slightly from what appeared in the earlier chapter. Write $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2}$ and re-express the quadratic form as
$$\mathbf{Y}'\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2}\mathbf{Y},$$
noting the symmetry of $\mathbf{\Sigma}^{-1/2}$. Consider the random vector $\mathbf{Z} \equiv \mathbf{\Sigma}^{-1/2}\mathbf{Y}$. Since $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, we have $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}\mathbf{\Sigma}^{-1/2})$, and because $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}$, we see that $\mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. Hence, the quadratic form
$$\mathbf{Y}'\mathbf{\Sigma}^{-1}\mathbf{Y} = \mathbf{Z}'\mathbf{Z},$$
and the right-hand side is the sum of squares of k independent standard normal random variables.

- If the k-vector $\mathbf{Z} \sim \mathcal{N}(\boldsymbol{\delta}, \mathbf{I})$, then $\mathbf{Z}'\mathbf{Z}$ has the non-central chi-square distribution with k degrees of freedom and *non-centrality parameter*
$$\lambda = \frac{1}{2}\boldsymbol{\delta}'\boldsymbol{\delta}.$$
Consider the case in which $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ and proceed to derive the distribution of the quadratic form $\mathbf{Y}'\mathbf{\Sigma}^{-1}\mathbf{Y}$. Again factor $\mathbf{\Sigma}^{-1} = \mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2}$ and let $\mathbf{Z} = \mathbf{\Sigma}^{-1/2}\mathbf{Y}$, giving $\mathbf{Y}'\mathbf{\Sigma}^{-1}\mathbf{Y} = \mathbf{Z}'\mathbf{Z}$ with $\mathbf{Z} \sim \mathcal{N}(\mathbf{\Sigma}^{-1/2}\boldsymbol{\mu}, \mathbf{I})$. Hence, $\mathbf{Y}'\mathbf{\Sigma}^{-1}\mathbf{Y}$ is distributed as non-central $\chi^2_{k,\lambda}$ with k degrees of freedom and non-centrality parameter
$$\lambda = \frac{1}{2}\boldsymbol{\mu}'\mathbf{\Sigma}^{-1/2}\mathbf{\Sigma}^{-1/2}\boldsymbol{\mu} = \frac{1}{2}\boldsymbol{\mu}'\mathbf{\Sigma}^{-1}\boldsymbol{\mu}.$$

With these results in hand, consider a test of the null hypothesis $H_0 : \beta = \beta_0$ for the linear model $\mathbf{Y} = \mathbf{X}\beta + \boldsymbol{\epsilon}$. Using the multivariate versions of the central limit theorems

described earlier, and assuming that the data series is i.i.d. with homoskedastic disturbances, we have the result that

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{Q}^{-1}),$$

where $\mathbf{Q}^{-1} = \text{plim}\ (n^{-1}\mathbf{X}'\mathbf{X})^{-1}$. Hence, as $n \to \infty$,

$$\sqrt{n}(\hat{\beta}_n - \beta_0)' \cdot \left(\sigma^2 \mathbf{Q}^{-1}\right)^{-1} \cdot \sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \chi^2_k$$

if the null hypothesis is true. This is still not quite in the form of a test statistic, as it depends on the unknown parameter σ^2 and unknown matrix $\mathbf{Q}$, no value for either having been specified in the null hypothesis. But as we've just seen, we can substitute consistent estimators $\hat{\sigma}^2_n$ for σ^2 and $\hat{\mathbf{Q}}_n$ for $\mathbf{Q}$ without altering the asymptotic properties,

$$\sqrt{n}(\hat{\beta}_n - \beta_0)' \cdot \left(\frac{1}{\hat{\sigma}^2_n}\frac{1}{n}\mathbf{X}'\mathbf{X}\right) \cdot \sqrt{n}(\hat{\beta}_n - \beta_0) \stackrel{a}{=} \sqrt{n}(\hat{\beta}_n - \beta_0)' \cdot \left(\frac{1}{\sigma^2}\mathbf{Q}\right) \cdot \sqrt{n}(\hat{\beta}_n - \beta_0)$$

The left-hand side is now a test statistic: it depends only on the null hypothesis and estimated or calculable quantities. To use this result, we would decide what size of test we want and pick the critical value t_h accordingly, such that the size equals $\Pr(T \geq t_h)$ under the null.

A better example is given by the null hypothesis that the true $\boldsymbol{\beta}_0$ satisfies

$$H_0 : \mathbf{R}\boldsymbol{\beta}_0 = \mathbf{r}$$

with $\mathbf{R}$ being an $m \times k$ matrix of constants and $\mathbf{r}$ an $m \times 1$ vector. Many common hypotheses about $\boldsymbol{\beta}$ can be put in this form. To develop a test statistic for this kind of null hypothesis, consider $\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{r}$ and substitute $\mathbf{R}\boldsymbol{\beta}_0$ for $\mathbf{r}$ as implied by that hypothesis, to obtain

$$\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{r}) = \sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{R}\boldsymbol{\beta}_0) = \mathbf{R}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$$

under the null. Because $\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$,

$$\mathbf{R}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}').$$

Hence, using $\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{r}$ in place of $\mathbf{R}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$ to form the test statistic, we obtain

$$\sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{r})' \left(\sigma^2\mathbf{R}\mathbf{Q}^{-1}\mathbf{R}'\right)^{-1} \sqrt{n}(\mathbf{R}\hat{\boldsymbol{\beta}}_n - \mathbf{r}) \sim \chi^2_m$$

under the null. We convert the left-hand side to a test statistic by substituting $\hat{\sigma}^2_n$ for σ^2 and substituting $(n^{-1}\mathbf{X}'\mathbf{X})^{-1}$ for $\mathbf{Q}^{-1}$.

Using "Pitman drift" to analyze power

In the asymptotic context, the alternative hypothesis cannot usefully be posed in terms of a specific fixed value of β. To see why, let's return to the simple null hypothesis $H_0 : \beta = \beta_0$ we first examined above, and consider the alternative hypothesis $H_A : \beta = \beta_0 + \delta$, in which δ indexes the extent to which the true value of β departs from what was assumed in the

null hypothesis. Hence, the hypothesized value $\beta_0 = \beta - \delta$. Then since the least-squares estimator can be written in terms of the true β as

$$\hat{\beta} = \beta + \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{1}{n}\mathbf{X}'\epsilon$$

we have for the vector $\sqrt{n}(\hat{\beta} - \beta_0)$ that will appear in the wings of the test statistic quadratic form,

$$\sqrt{n}(\hat{\beta} - \beta_0) = \sqrt{n}(\hat{\beta} - \beta) + \sqrt{n}\delta$$

As we've already seen, the quantity on the right-hand side $\sqrt{n}(\hat{\beta} - \beta)$ in which the true β appears, converges in distribution to $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$. But under the alternative hypothesis, the left-hand side expression actually used in the test statistic, which is $\sqrt{n}(\hat{\beta} - \beta_0)$, does *not* converge in distribution: Its mean (so to speak) is not zero but rather $\sqrt{n}\delta$, which changes as n changes. If we allow ourselves to think of $\sqrt{n}(\hat{\beta} - \beta_0)$ as being approximately normal with mean $\sqrt{n}\delta$ for some large n, then the quadratic form of our Wald test will have (approximately) a non-centrality parameter that depends positively on n, and as $n \to \infty$ the non-centrality parameter will also go to ∞ resulting in a certain rejection of the null hypothesis. Asymptotically, any departure of the true β from the null, no matter how trivial it may be numerically, is therefore sure to be rejected. If we wanted to compare two tests of this kind—they are termed *consistent tests*—on the basis of their power, we would be unable to discriminate between them.

Somehow, then, we need to take a different perspective on the power of a test that has the potential to be useful in comparisons of tests. This is how the concept of *Pitman drift* comes into play. Let $\phi(\beta_n)$ be the power of a test for alternative $\beta_n \neq \beta_0$. We will be able to analyze the following limit:

$$\lim_{\beta_n \to \beta_0} \phi(\beta_n).$$

We hope that this limit of the power function might be useful in comparing tests of the null hypothesis against alternatives that are "close" to the null.

To get a sense of how the non-central χ^2 distribution is involved in determining the "limiting" power of a hypothesis test, consider the null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0$ and the alternative hypothesis $H_A : \boldsymbol{\beta}_n \neq \boldsymbol{\beta}_0$,

$$H_A : \boldsymbol{\beta}_n = \boldsymbol{\beta}_0 + \frac{1}{\sqrt{n}}\boldsymbol{\delta},$$

with $\boldsymbol{\beta}_n$ now being the *true value* of the parameter that we are considering to understand test power, whereas $\boldsymbol{\beta}_0$ is the value specified in the null hypothesis. We have

$$\begin{aligned}\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) &= \sqrt{n}\left(\hat{\boldsymbol{\beta}}_n - (\boldsymbol{\beta}_n - \frac{1}{\sqrt{n}}\boldsymbol{\delta})\right) \\ &= \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) + \boldsymbol{\delta}.\end{aligned}$$

Because $\boldsymbol{\beta}_n$ is the true value of the parameter under the alternative we're considering,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_n) + \boldsymbol{\delta} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}) + \boldsymbol{\delta} \xrightarrow{d} \mathcal{N}(\boldsymbol{\delta}, \sigma^2\mathbf{Q}^{-1}).$$

Therefore

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)' \left(\sigma^2 \mathbf{Q}^{-1}\right)^{-1} \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \xrightarrow{d} \chi^2_{k,\lambda}$$

with non-centrality parameter

$$\lambda = \frac{1}{2}\delta' \left(\frac{1}{\sigma^2}\mathbf{Q}\right)\delta.$$

You will recall that $\frac{1}{n}\mathbf{X}\mathbf{X} \overset{a}{=} \mathbf{Q}$, and for a given sample size n, this expression, along with a consistent estimator of σ^2, would be used to approximate the power of the test in regions of β values near the value β_0 given in the null hypothesis. You approximate power by finding out the probability that a $\chi^2_{k,\lambda}$ variable exceeds the critical value t_h for the test statistic as discussed above, which you calculate under the null hypothesis for whatever test size you desire. See Cameron and Trivedi (2005, pp. 247–250) for an interesting discussion of how to think about Pitman drift.

Tests on σ^2

What about hypothesis tests on σ^2, the disturbance term variance, such as $H_0 : \sigma^2 = \sigma_0^2$? Under the null, $\sqrt{n}(s_n^2 - \sigma_0^2) \xrightarrow{d} \mathcal{N}(0, \mu_4 - \sigma_0^4)$ with $\mu_4 = \mathrm{E}\,\epsilon_i^4$ in the iid case. The only problem facing us in the construction of a test statistic is that μ_4 is unknown and to this point, we have not considered any estimator of μ_4 and proven that the estimator converges in probability to the unknown value of the parameter.

A plausible candidate for such an estimator is the average of the regression residuals raised to the fourth power, that is

$$\hat{\mu}_4 = \frac{1}{n}\sum_{i=1}^{n} e_i^4.$$

Because regression residuals are not the same thing as disturbances, proving that $\hat{\mu}_4 \xrightarrow{p} \mu_4$ is no easy task—we will not undertake the proof here—but consistency can be shown to occur under reasonably general conditions. If we insert $\hat{\mu}_4$ in place of μ_4, we can proceed to create the test statistic.

It is not difficult to extend our scope to hypotheses on the variances from two or more data-generating processes, if we assume that these processes operate independently. For example, if the linear model $Y_{i,W} = \mathbf{X}'_{i,W}\beta_W + \epsilon_{i,W}$ is a model of wage rates for women, and $Y_{i,M} = \mathbf{X}'_{i,M}\beta_M + \epsilon_{i,M}$ is the counterpart model for men, and if the two processes are independent, we can easily test the null hypothesis $H_0 : \sigma_W^2 = \sigma_M^2$. Write the null as

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \sigma_W^2 \\ \sigma_M^2 \end{bmatrix} = 0$$

which is of the familiar $\mathbf{R}\Sigma = \mathbf{r}$ form with Σ representing the 2×1 vector of the true variances. Assume two samples each with n observations. Let $\hat{\mathbf{S}}_n = [s_W^2, s_M^2]'$ and consider the variance of the limiting distribution of

$$\sqrt{n}\mathbf{R}\hat{\mathbf{S}}_n = \mathbf{R}\sqrt{n}(\hat{\mathbf{S}}_n - \Sigma) = \mathbf{R}\sqrt{n}\begin{bmatrix} s_W^2 - \sigma_W^2 \\ s_M^2 - \sigma_M^2 \end{bmatrix},$$

under the null. The variance is

$$\mathbf{R}\begin{bmatrix}\mu_{4,W}-\sigma_W^4 & 0\\ 0 & \mu_{4,M}-\sigma_M^4\end{bmatrix}\mathbf{R}'$$

or simply $\mu_{4,W}-\sigma_W^4+\mu_{4,M}-\sigma_M^4$. Note that the explanatory covariates need not be the same in the two models, and even if they are the same, their β coefficients can differ.

15.9 Limiting Distributions of Differentiable Functions

Suppose we are interested in approximating the behavior of a differentiable function of $\hat{\beta}_n$ as n increases. We've just seen how to handle the case of linear functions. What about nonlinear functions?

The approach we'll describe is often termed the *delta method*, essentially because it uses derivatives to approximate the function in question in a neighborhood of β_0, the probability limit of the $\hat{\beta}_n$ estimator. In fact, the method can be explored with reference to a sequence of random vectors that, when appropriately transformed, converge in distribution to multivariate normal—there's no need to link it to $\hat{\beta}_n$ as such. Let's discuss the method in these general terms.

Let $g(\mathbf{X})$ be a function of the $k\times 1$ vector $\mathbf{X}$ and let the sequence of random vectors $\{\mathbf{X}_n\}$ have a probability limit $\boldsymbol{\mu}$ and obey a central limit theorem, such that

$$\sqrt{n}(\mathbf{X}_n-\boldsymbol{\mu})\xrightarrow{d}\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}).$$

Let g have first-order partial derivatives in a neighborhood of $\boldsymbol{\mu}$ that are continuous at $\boldsymbol{\mu}$, and further suppose that the $k\times 1$ vector

$$\mathbf{G}=(\partial g(\boldsymbol{\mu})/\partial x_1,\ldots,\partial g(\boldsymbol{\mu})/\partial x_k)'\neq\mathbf{0}.$$

We'll show that

$$\sqrt{n}(g(\mathbf{X}_n)-g(\boldsymbol{\mu}))\xrightarrow{d}\mathcal{N}(0,\mathbf{G}(\boldsymbol{\mu})'\cdot\boldsymbol{\Sigma}\cdot\mathbf{G}(\boldsymbol{\mu})).$$

To do this, we appeal to the mean value theorem and expand $g(\mathbf{X}_n)$ around $\boldsymbol{\mu}$, yielding

$$g(\mathbf{X}_n)=g(\boldsymbol{\mu})+\mathbf{G}(\tilde{\boldsymbol{\mu}}_n)'\cdot(\mathbf{X}_n-\boldsymbol{\mu}) \tag{15.17}$$

or

$$\sqrt{n}(g(\mathbf{X}_n)-g(\boldsymbol{\mu}))=\mathbf{G}(\tilde{\boldsymbol{\mu}}_n)'\cdot\sqrt{n}(\mathbf{X}_n-\boldsymbol{\mu})$$

in which $\tilde{\boldsymbol{\mu}}$ lies between μ and $\mathbf{X}_n$. Because the probability limit of $\mathbf{X}_n$ is $\boldsymbol{\mu}$,

$$\sqrt{n}(g(\mathbf{X}_n)-g(\boldsymbol{\mu}))\overset{a}{=}\mathbf{G}(\boldsymbol{\mu})'\cdot\sqrt{n}(\mathbf{X}_n-\boldsymbol{\mu}).$$

Given that $\sqrt{n}(\mathbf{X}_n-\boldsymbol{\mu})\xrightarrow{d}\mathcal{N}(\mathbf{0},\boldsymbol{\Sigma})$, then,

$$\sqrt{n}(g(\mathbf{X}_n)-g(\boldsymbol{\mu}))\xrightarrow{d}\mathcal{N}(0,\mathbf{G}(\boldsymbol{\mu})'\cdot\boldsymbol{\Sigma}\cdot\mathbf{G}(\boldsymbol{\mu})).$$

For m-vectors of functions $\mathbf{g}(\mathbf{X})$, with $\mathbf{g}(\mathbf{X})=(g_1(\mathbf{X}),\ldots,g_m(\mathbf{X}))'$, there is a straightforward generalization of this result; see Mittelhammer (1996, Theorem 5.40)).

Applying this general theory to the special case of $\hat{\beta}$, consider the null hypothesis $H_0 : g(\beta_0) = 0$ and the associated $g(\hat{\beta}_n)$. Using a Taylor expansion

$$g(\hat{\beta}_n) = g(\beta_0) + \mathbf{G}(\tilde{\beta}_n)'(\hat{\beta}_n - \beta_0),$$

multiplying through by $\sqrt{n}$ and using the fact that under the null $g(\beta_0) = 0$, we have

$$\sqrt{n}g(\hat{\beta}_n) = \mathbf{G}(\tilde{\beta}_n)'\sqrt{n}(\hat{\beta}_n - \beta_0).$$

Because $\tilde{\beta}_n \to \beta_0$,

$$\sqrt{n}g(\hat{\beta}_n) \xrightarrow{d} \mathcal{N}(0, \mathbf{G}(\beta_0)'\sigma^2\mathbf{Q}^{-1}\mathbf{G}(\beta_0)),$$

and we would proceed to make a χ^2_1 test statistic as usual, putting the inverse of the variance in the middle of the quadratic form and putting $\sqrt{n}g(\hat{\beta}_n)$ in the wings. When we turn this into a proper test statistic, of course, we evaluate the derivative vector $\mathbf{G}(\beta_0)$ at $\hat{\beta}_n$ rather than at the unknown β_0.

15.10 Asymptotic Theory and Regression Output

We know that in the homoskedastic case, the limiting distribution of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$ with

$$\mathbf{Q} = \text{plim}\,\frac{1}{n}\sum_i \mathbf{X}_i\mathbf{X}_i',$$

and further, that we can approximate the variance using s_n^2 for σ^2 and

$$\hat{\mathbf{Q}} = \frac{1}{n}\sum_i \mathbf{X}_i\mathbf{X}_i'$$

for $\mathbf{Q}$. In R, however, the `lm` command by which we obtain OLS regression estimates delivers not an estimate of the variance of the limiting distribution as such, but rather an approximation to the variance of

$$\hat{\beta}_n - \beta_0 = \frac{1}{\sqrt{n}}\sqrt{n}(\hat{\beta}_n - \beta_0)$$

with this approximate variance being (in theory)

$$\frac{1}{n} \cdot \sigma^2\mathbf{Q}^{-1},$$

a quantity that is estimated by

$$s_n^2 \left(\sum_i \mathbf{X}_i\mathbf{X}_i'\right)^{-1}.$$

You need to keep this in mind when you construct test statistics that are based on asymptotic theory but make use of R's output. This is not an R-specific issue: All statistical software of which I'm aware follows the same practice.

Similarly, in the heteroskedastic case, we know from the multivariate version of the Lindeberg central limit theorem that

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1}\mathbf{W}\mathbf{Q}^{-1})$$

with

$$\mathbf{W} = \operatorname{plim} \frac{1}{n} \sum_i \mathbf{X}_i \epsilon_i^2 \mathbf{X}_i'$$

and we estimate the variance matrix using

$$\left(\frac{1}{n} \sum_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \cdot \frac{1}{n} \sum_i \mathbf{X}_i e_i^2 \mathbf{X}_i' \cdot \left(\frac{1}{n} \sum_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1}$$

with e_i^2 being the square of the i-th OLS residual. Note that in our approximation, two of the three n's cancel. When the variance matrix is constructed, it will be calculated as $1/n$ times the matrix expression above, so that the third n also cancels out. We therefore have

$$\left(\sum_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1} \cdot \sum_i \mathbf{X}_i e_i^2 \mathbf{X}_i' \cdot \left(\sum_i \mathbf{X}_i \mathbf{X}_i' \right)^{-1}.$$

You should be aware that in these cases, software programs often have an option for inserting a "finite-sample correction factor" which causes the matrix to differ slightly from the expression just shown. In samples of the size that we usually work with in economics—several hundred observations or more—this correction factor makes a negligible difference and can be safely ignored. In small samples, however, the correction may be worth considering.

Chapter 16

Nonlinear Regression

Students: Prepare for this chapter by reading Cameron and Trivedi (2005, Section 4.4), which applies the Lindeberg central limit theorem to the ordinary least squares model (assuming heteroskedastic disturbances.) With that material providing background, supplement this chapter by reading Cameron and Trivedi (2005, Section 5.8).

In this chapter, we will explore the properties of nonlinear least-squares (NLS) estimators of the model $Y_i = g(\mathbf{X}_i, \beta) + \epsilon_i$, in which $g(\mathbf{X}_i, \beta)$ is a nonlinear function of explanatory covariates $\mathbf{X}_i$ and β parameters, there being k such parameters. Our discussion draws from Davidson and MacKinnon (1993, pp. 162–167), Amemiya (1985), and Mittelhammer, Judge, and Miller (2000), all of which are highly recommended. Greene (2008) presents a number of economic applications of the NLS method.

The members of the sequence $\{(\mathbf{X}_i, \epsilon_i)\}$ are assumed to be independent but not identically distributed (inid) and, as a consequence, we will need to apply the Lindeberg central limit theorem when we derive the limiting distribution of the NLS estimator. The fundamental assumption about the disturbance term ϵ_i is that it has a conditional mean of zero, i.e., $\mathrm{E}(\epsilon_i|\mathbf{X}_i) = 0$. It is not difficult to allow the disturbances to exhibit heteroskedasticity of unknown form. Much as in ordinary least squares models with White-corrected (robust) standard errors, this generalization yields what some term a "sandwich" estimator for the variance of the limiting distribution, as Cameron and Trivedi (2005) describe in detail. However, to keep our exposition focused on the essentials, we will make the simplifying assumption that the disturbances are homoskedastic, with $\mathrm{E}(\epsilon_i^2|\mathbf{X}_i) = \sigma^2$. To further economize on notation, we will generally abbreviate $g(\mathbf{X}_i, \beta)$ as $g_i(\beta)$, leaving implicit the role of the $\mathbf{X}_i$ covariates.

Note that although it is considerably more general than the linear regression model, the model we are examining remains restricted, in that we rule out cases such as $Y_i^\theta = g_i(\beta) + \epsilon_i$ in which the dependent variable is transformed in a nonlinear fashion. To analyze such cases, we would have to leave the framework of regression altogether and proceed on the basis of the maximum likelihood method or the generalized method of moments.

A good example of nonlinear regression is provided by a time-series model with first-order serial correlation in the disturbances. Here $Y_t = \mathbf{X}_t'\beta + \epsilon_t$ with $\epsilon_t = \rho\epsilon_{t-1} + u_t$ where u_t is independent over t and also independent of all ϵ_{t-s} disturbances for $s \geq 1$. We can

transform the model for Y_t by multiplying Y_{t-1} by ρ and subtracting the result from Y_t, yielding

$$Y_t = \rho Y_{t-1} + (\mathbf{X}_t - \rho \mathbf{X}_{t-1})' \beta + u_t$$

for observations Y_2 to Y_T. This is a nonlinear equation whose parameters can be estimated by the method of nonlinear least squares. (Why did we need to assume that u_t is independent of ϵ_{t-s}? The reason is that Y_{t-1} appears on the right-hand side of the equation, and since it obviously depends on ϵ_{t-1}, we need the assumption to eliminate the possibility of correlation between u_t and a right-hand side explanatory variable.) Later we will examine alternative estimation methods (estimated generalized least squares and maximum likelihood) that can also be applied to this model.

16.1 Properties of $\hat{\beta}$

In what follows, we let β_0 denote the true value of the β vector. The NLS estimator $\hat{\beta}_n$ is the value of β that minimizes the sum of squares

$$S_n(\beta) = \frac{1}{n}(\mathbf{Y} - \mathbf{g}(\beta))'(\mathbf{Y} - \mathbf{g}(\beta)),$$

where $\mathbf{Y}$ and $\mathbf{g}$ are n-vectors, and where we restrict the admissible β values to the set Ω, about which we will have more to say below. The solution $\hat{\beta}_n$ must usually be obtained through numerical optimization methods and it is not generally available in a closed form. Therefore we cannot analyze the properties of $\hat{\beta}_n$ using the techniques that we applied to the ordinary least squares model—more advanced methods are called for.

To understand the NLS estimator, write the sum of squares minimization problem in summation form as

$$\min_{\beta \in \Omega} S_n(\beta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - g_i(\beta))^2.$$

If $g()$ is differentiable in β, the first-order conditions are

$$-\frac{2}{n} \sum_{i=1}^{n} (Y_i - g_i(\hat{\beta}_n)) \cdot \frac{\partial g_i(\hat{\beta}_n)}{\partial \beta} = \mathbf{0}.$$

This orthogonality condition is similar to the one for ordinary least squares, with $Y_i - g_i(\hat{\beta}_n) \equiv e_i$ being the residual and with the $k \times 1$ vector of derivatives $\partial g_i(\hat{\beta}_n)/\partial \beta$ replacing the vector $\mathbf{X}_i$ of the linear model. The orthogonality condition can be viewed as a set of k nonlinear equations in k unknowns.

If we ignore the $-2/n$ factor, the first-order conditions can be written in matrix form as

$$\begin{bmatrix} \frac{\partial g_1(\hat{\beta}_n)}{\partial \beta} & \frac{\partial g_2(\hat{\beta}_n)}{\partial \beta} & \cdots & \frac{\partial g_n(\hat{\beta}_n)}{\partial \beta} \end{bmatrix} \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \mathbf{0}.$$

It is more conventional to express these conditions as

$$\mathbf{G}(\hat{\beta}_n)'\mathbf{e} = \mathbf{0}.$$

in which $\mathbf{G}$ is the $n \times k$ transpose of the matrix of derivatives and $\mathbf{e}$ is the vector of NLS residuals. When they are written out in this way, we see that the NLS first-order condition resembles the condition $\mathbf{X}'\mathbf{e} = \mathbf{0}$ for OLS estimators, in that at the minimizing $\hat{\beta}_n$, each column of $\mathbf{G}$ is orthogonal to the residual vector $\mathbf{e} = \mathbf{Y} - \mathbf{g}(\hat{\beta}_n)$.

It can be shown that the NLS estimator $\hat{\beta}_n$ is consistent and that $\sqrt{n}(\hat{\beta}_n - \beta_0)$ converges in distribution to multivariate normal. The proof of consistency uses concepts that we will discuss in more detail in our next chapter on the maximum likelihood method. In essence, the sufficient conditions for consistency of the NLS estimator are as follows.

Let $Q_n(\beta) \equiv -S_n(\beta) = -\frac{1}{n}(\mathbf{Y} - \mathbf{g}(\beta))'(\mathbf{Y} - \mathbf{g}(\beta))$, and let $\hat{\beta}_n$ be the value of β that maximizes $Q_n(\beta)$ and therefore minimizes the sum of squares. If the following conditions hold, that

- The parameter space Ω containing all possible values for β is compact, that is, closed and bounded;
- $Q_n(\beta)$ converges uniformly in probability to a limit function $Q(\beta)$;
- The limit function $Q(\beta)$ is continuous; and
- The limit function $Q(\beta)$ is uniquely maximized at β_0, the true value of β,

then the estimator $\hat{\beta}_n \xrightarrow{p} \beta_0$. Mittelhammer, Judge, and Miller (2000, Chapters 7 and 8) and Davidson and MacKinnon (1993) provide illuminating discussions of the issues and Newey and McFadden (1994) give a rigorous proof. Of the sufficient conditions mentioned above, the most difficult to establish is the one involving uniform convergence. Newey and McFadden (1994) also provide an alternative set of conditions in which it is assumed only that $Q_n(\beta) \xrightarrow{p} Q(\beta)$ pointwise, but the limit function $Q(\beta)$ is required to be concave in β and the parameter space Ω must be a convex set.

Given consistency, it is not too difficult to derive the limiting distribution of the NLS estimator. The essence of the approach is to carry out a Taylor-series expansion of the first-order conditions and manipulate it to discover the limiting distribution. Let's simplify our notation temporarily by defining the $k \times 1$ vector

$$h_i(\hat{\beta}_n) = (Y_i - g_i(\hat{\beta}_n)) \cdot \frac{\partial g_i(\hat{\beta}_n)}{\partial \beta}$$

and Taylor-expand it around the true β_0, giving

$$h_i(\hat{\beta}_n) = h_i(\beta_0) + \frac{\partial h_i(\tilde{\beta}_n)}{\partial \beta} \cdot (\hat{\beta}_n - \beta_0)$$

with $\tilde{\beta}_n$ lying between $\hat{\beta}_n$ and β_0. Note that $\frac{\partial h_i(\tilde{\beta}_n)}{\partial \beta}$ is a $k \times k$ matrix. Now,

$$h_i(\beta_0) = (Y_i - g_i(\beta_0)) \cdot \frac{\partial g_i(\beta_0)}{\partial \beta} = \epsilon_i \cdot \frac{\partial g_i(\beta_0)}{\partial \beta}.$$

Because of the fundamental assumption $\mathrm{E}\,\epsilon_i|\mathbf{X}_i = 0$, the expected value of $h_i(\beta_0)$ must be zero and its variance is

$$\operatorname{Var} h_i(\beta_0) = \sigma^2 \,\mathrm{E}\, \frac{\partial g_i(\beta_0)}{\partial \beta} \frac{\partial g_i(\beta_0)}{\partial \beta}'.$$

We'll make use of these two results in a moment.[1]

Turning our attention to $\partial h_i(\tilde{\beta}_n)/\partial\beta$, we have

$$\frac{\partial h_i(\tilde{\beta}_n)}{\partial\beta} = \frac{\partial}{\partial\beta}\left((Y_i - g_i(\tilde{\beta}_n)) \cdot \frac{\partial g_i(\tilde{\beta}_n)}{\partial\beta}\right).$$

This derivative can be written as

$$\frac{\partial h_i(\tilde{\beta}_n)}{\partial\beta} = (Y_i - g_i(\tilde{\beta}_n))\frac{\partial^2 g_i(\tilde{\beta}_n)}{\partial\beta\partial\beta'} - \frac{\partial g_i(\tilde{\beta}_n)}{\partial\beta}\frac{\partial g_i(\tilde{\beta}_n)}{\partial\beta}'$$

where to find the second term (like the first, it is a $k \times k$ matrix) we have made use of the general result that if $\mathbf{c}$ is a column vector and $a(\theta)$ is a (scalar) function of the θ vector, then the derivative of the column vector $a(\theta)\mathbf{c}$ with respect to θ is the outer product $\mathbf{c} \cdot \partial a/\partial\theta'$.

Note that because $\hat{\beta}_n$ is consistent, for any given i

$$(Y_i - g_i(\tilde{\beta}_n))\frac{\partial^2 g_i(\tilde{\beta}_n)}{\partial\beta\partial\beta'} \xrightarrow{p} \epsilon_i \frac{\partial^2 g_i(\beta_0)}{\partial\beta\partial\beta'},$$

the right-hand side being a matrix with a mean of zero because $\mathrm{E}\,\epsilon_i|\mathbf{X}_i = 0$. Likewise, for the second term of this derivative,

$$-\frac{\partial g_i(\tilde{\beta}_n)}{\partial\beta}\frac{\partial g_i(\tilde{\beta}_n)}{\partial\beta}' \xrightarrow{p} -\frac{\partial g_i(\beta_0)}{\partial\beta}\frac{\partial g_i(\beta_0)}{\partial\beta}'.$$

Let's now insert all these results into the first-order conditions, after multiplying those conditions by $\sqrt{n}$:

$$\mathbf{0} = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n} h_i(\beta_0) + \frac{1}{n}\sum_{i=1}^{n}\frac{\partial h_i(\tilde{\beta}_n)}{\partial\beta} \cdot \sqrt{n}(\hat{\beta}_n - \beta_0).$$

Assuming that the data series $\{(\mathbf{X}_i, \epsilon_i)\}$ is inid, and that the Lindeberg condition is met, we can apply the Lindeberg central limit theorem to the first of the terms on the right-hand side, finding

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} h_i(\beta_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i \cdot \frac{\partial g_i(\beta_0)}{\partial\beta} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{V})$$

with

$$\mathbf{V} = \lim\frac{1}{n}\sum_i \mathrm{E}\left(\frac{\partial g_i(\beta_0)}{\partial\beta}\frac{\partial g_i(\beta_0)}{\partial\beta}'\right) = \operatorname{plim}\frac{1}{n}\sum_i \frac{\partial g_i(\beta_0)}{\partial\beta}\frac{\partial g_i(\beta_0)}{\partial\beta}'.$$

Furthermore—here we pass lightly over some technical matters that are discussed in the Newey–McFadden reference—we obtain

$$\frac{1}{n}\sum_i \frac{\partial h_i(\tilde{\beta}_n)}{\partial\beta} \xrightarrow{p} -\operatorname{plim}\frac{1}{n}\sum_i \frac{\partial g_i(\beta_0)}{\partial\beta}\frac{\partial g_i(\beta_0)}{\partial\beta}' = -\mathbf{V}.$$

[1] Had we not imposed homoskedasticity on the disturbances, the variance term would have been

$$\mathrm{Var}\, h_i(\beta_0) = \mathrm{E}\left(\epsilon_i^2\frac{\partial g_i(\beta_0)}{\partial\beta}\frac{\partial g_i(\beta_0)}{\partial\beta}'\right).$$

Drawing all of this together, we see that

$$\mathbf{V}\sqrt{n}(\hat{\beta}_n - \beta_0) \stackrel{a}{=} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{V}),$$

and at last obtain our main result,

$$\sqrt{n}(\hat{\beta}_n - \beta_0) \stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{V}^{-1}). \tag{16.1}$$

To make use of this result, we would approximate $\mathbf{V}$ with

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_i \frac{\partial g_i(\hat{\beta}_n)}{\partial \beta}\frac{\partial g_i(\hat{\beta}_n)}{\partial \beta}'$$

and, as discussed next, would approximate the unknown variance σ^2 with a consistent estimator of it.

16.2 Properties of $\hat{\sigma}_n^2$

Consider the estimator of the variance $\hat{\sigma}_n^2 = \mathbf{e}'\mathbf{e}/n$ based on NLS residuals. It can be shown (actually, a great deal of manipulation is needed) that $\hat{\sigma}_n^2 \stackrel{p}{\to} \sigma^2$. Also, $\sqrt{n}(\hat{\sigma}_n^2 - \sigma^2) \stackrel{d}{\to} \mathcal{N}(0, \mu^4 - \sigma^4)$. See Davidson and MacKinnon (2004) for further discussion. To test hypotheses about the variance, we would require a consistent estimator of μ^4.

16.3 The Gauss–Newton Regression

In what follows, we discuss a useful adjunct to nonlinear regression, which is the artificial linear regression known as *Gauss–Newton regression* or GNR. We explore two of its uses in this chapter, and will examine additional uses in a later chapter.

Suppose that one is given an estimate $\hat{\beta}_n$ that is thought to minimize $S_n(\beta)$, the regression sum of squares. Consider the Gauss–Newton regression

$$\mathbf{Y} - \mathbf{g}(\hat{\beta}_n) = \mathbf{G}(\hat{\beta}_n)b + \text{ residuals},$$

where $\mathbf{G}$ is the $n \times k$ derivative matrix to which we referred above and b is a vector of artificial regression coefficients. This linear regression supplies us with two potentially useful outputs.

First, note that if $\hat{\beta}_n$ is actually the β value that minimizes the sum of squares, then the NLS first-order conditions guarantee that $\hat{b} = \mathbf{0}$. Hence, the Gauss–Newton regression may help us to assess the accuracy of the NLS minimization routine that was used to locate $\hat{\beta}_n$. The GNR may be of particular value in difficult problems, where, for example, one coefficient can be found only through grid search or other non-standard methods.

Provided that $\hat{\beta}_n$ is indeed the minimizing value, a second use of the Gauss–Newton regression is in calculating its covariance matrix. Note that the estimated covariance matrix of $\hat{b}$ is

$$\text{Cov}(\hat{b}) = s_n^2(\mathbf{G}(\hat{\beta}_n)'\mathbf{G}(\hat{\beta}_n))^{-1} = s_n^2\left(\sum_i \frac{\partial g_i(\hat{\beta}_n)}{\partial \beta}\frac{\partial g_i(\hat{\beta}_n)}{\partial \beta}'\right)^{-1}$$

and, since $\hat{b} = 0$,

$$s_n^2 = \frac{(\mathbf{Y} - \mathbf{g}(\hat{\beta}_n))'(\mathbf{Y} - \mathbf{g}(\hat{\beta}_n))}{n - k}.$$

(We can of course substitute n for $n - k$.) These are precisely the results we require to approximate the variance of $\hat{\beta}_n$.

16.4 Partially linear (semiparametric) models: Revisiting FWL

There is a nice approach by Robinson (1988), summarized by Cameron and Trivedi (2005, pp. 324–325), which not only can be viewed as an extension of the Frisch–Waugh–Lovell linear algebraic theorem, but which also provides the jumping-off place for a class of semiparametric models (a large topic not otherwise discussed in these notes). Robinson examined the partially linear model

$$Y_i = \mathbf{X}_i'\beta + \phi(Z_i) + \epsilon_i$$

on the assumption that $\mathrm{E}(\epsilon_i|\mathbf{X}_i, Z_i) = 0$. In the semiparametric literature, ϕ is a function of unknown form and owing to the formidable computational difficulties of dealing with multivariate explanatory vectors in functions of unknown form, Z_i is usually assumed to be a scalar random variable. In the parametric literature, by contrast, the form of ϕ is taken to be known, expressed for instance as $\phi(\mathbf{Z}_i, \theta)$, and $\mathbf{Z}_i$ is allowed to be a vector of explanatory variables.

Note that conditioning on Z_i, we have

$$\mathrm{E}(Y_i|Z_i) = \mathrm{E}(\mathbf{X}_i|Z_i)'\beta + \phi(Z_i)$$

and therefore

$$Y_i - \mathrm{E}(Y_i|Z_i) = (\mathbf{X}_i - \mathrm{E}(\mathbf{X}_i|Z_i))'\beta + \epsilon_i$$

with $\phi(Z_i)$ having been subtracted away. If consistent estimators $\hat{\mathrm{E}}(Y_i|Z_i)$ and $\hat{\mathrm{E}}(\mathbf{X}_i|Z_i)$ are available for the respective conditional expectations, then β can be consistently estimated via the regression

$$Y_i - \hat{\mathrm{E}}(Y_i|Z_i) = (\mathbf{X}_i - \hat{\mathrm{E}}(\mathbf{X}_i|Z_i))'\beta + \epsilon_i$$

This is a result not unlike the FWL theorem for fully linear regressions: rather than being written in terms of deviations from projections of $\mathbf{Y}$ and $\mathbf{X}$ on $\mathbf{Z}$, it is expressed in terms of deviations from conditional means.

Robinson (1988) establishes conditions under which the estimator of β is consistent and asymptotically normal. With a consistent $\hat{\beta}$ in hand, the relationship

$$\phi(Z_i) = \hat{\mathrm{E}}(Y_i|Z_i) - \hat{\mathrm{E}}(\mathbf{X}_i|Z_i)'\hat{\beta}$$

can be used to derive a nonparametric estimate of the unknown ϕ function. Cameron and Trivedi (2005, Chapter 9) provide an introduction to nonparametric methods in econometrics; much progress has been made in this difficult area since their textbook was published.

Chapter 17

Optimization and Related Numerical Methods

17.1 Maximization and minimization of differentiable functions

A good treatment is available in Cameron and Trivedi (2005, Chapter 10). Consider the problem of maximizing a function $L(\theta)$, assuming it to be twice differentiable in the θ parameter. We will describe the most famous iterative approach to this problem, which is known as the *Newton–Raphson* method. Functions that are not differentiable are (obviously) more difficult to maximize and require tools other than those we will discuss here.

The idea of the method is as follows. Suppose that we have an initial estimate θ_1 and an initial value of the function $L(\theta_1)$, and that starting from here, we want to find the θ_2 value that maximizes a quadratic approximation to $L(\theta)$ in the neighborhood of θ_1. Write

$$L(\theta) \approx L(\theta_1) + \frac{\partial L(\theta_1)}{\partial \theta}' (\theta - \theta_1) + \frac{1}{2}(\theta - \theta_1)' \frac{\partial^2 L(\theta_1)}{\partial \theta \partial \theta'} (\theta - \theta_1).$$

The first-order conditions that the maximizing θ_2 must satisfy are

$$\frac{\partial L(\theta_1)}{\partial \theta} + \frac{\partial^2 L(\theta_1)}{\partial \theta \partial \theta'} (\theta_2 - \theta_1) = \mathbf{0}_k,$$

and the second-order condition is that the $k \times k$ matrix

$$\frac{\partial^2 L(\theta_1)}{\partial \theta \partial \theta'}$$

must be negative definite. Assuming the second-order condition to be met—this is a major concern in applied work, as we will discuss shortly—and denoting the "gradient" by $\mathbf{g}_1 = \partial L(\theta_1)/\partial \theta$ and the "Hessian" by $\mathbf{H}_1 = \partial^2 L(\theta_1)/\partial \theta \partial \theta'$, the solution to the first-order conditions can be expressed as

$$\theta_2 = \theta_1 - \mathbf{H}_1^{-1} \mathbf{g}_1. \tag{17.1}$$

Obviously we cannot calculate this solution if at θ_1 the Hessian matrix of second derivatives is singular. Assuming $\mathbf{H}_1$ to be invertible, we can substitute the solution into

$$L(\theta_2) = L(\theta_1) + \mathbf{g}_1'(\theta_2 - \theta_1) + \frac{1}{2}(\theta_2 - \theta_1)' \mathbf{H}_1 (\theta_2 - \theta_1)$$

which yields

$$L(\theta_2) - L(\theta_1) = -\frac{1}{2}\mathbf{g}_1'\mathbf{H}_1^{-1}\mathbf{g}_1.$$

This shows that if $\mathbf{H}_1$ is *negative* definite, the value of $L(\theta_2)$ will exceed that of $L(\theta_1)$, so that θ_2 improves upon θ_1. However, if $\mathbf{H}_1$ is invertible but *positive* definite, $L(\theta_2)$ will be smaller than $L(\theta_1)$, and the Newton–Raphson algorithm will lead us astray.

This sensitivity to second derivatives has prompted investigation into a range of alternative methods that are similar in spirit to the Newton–Raphson, but which introduce various "fixes" to ensure that in practice, the algorithm picks a sequence of θ_n so as to produce successively higher values of the $L(\theta)$ function.

Application to OLS

Let $L(\theta) = -\frac{1}{2}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i'\theta)^2$, which is minus the sum of squares. We already know that the least squares estimator $\hat{\theta}$ solves the problem of maximizing $L(\theta)$. If we apply the Newton–Raphson iterative approach, does it lead us to the least-squares solution?

To find out, suppose that we have an initial estimate θ_1. The first-order conditions are then

$$\frac{\partial L(\theta_1)}{\partial \theta} = \sum_i \mathbf{X}_i(Y_i - \mathbf{X}_i'\theta_1) = \mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\theta_1 = \mathbf{g}_1,$$

and the second-order conditions are

$$\partial^2 L(\theta_1)/\partial\theta\partial\theta' = -\sum_i \mathbf{X}_i\mathbf{X}_i' = -\mathbf{X}'\mathbf{X} = \mathbf{H}_1.$$

We see that in this case, there can be no concern about whether $\mathbf{H}_1$ is negative definite. Hence the first-order conditions can be solved; they yield

$$\theta_2 = \theta_1 - \mathbf{H}_1^{-1}\mathbf{g}_2 = \theta_1 + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y} - \mathbf{X}'\mathbf{X}\theta_1) = \hat{\theta}.$$

Starting from any arbitrary θ_1, we reach the OLS estimator $\hat{\theta}$ in a single bound! In essence, this is because the sum of squares function $S(\theta)$ is globally convex (hence $L(\theta)$ is globally concave) and the Newton–Raphson method works exceedingly well for such functions.

Application to NLS

Where nonlinear models are concerned, matters are a bit more complicated and certainly there can be no guarantee of the negative definiteness of $\mathbf{H}$ when you are far away from the maximizing value of the parameter. There are some computational adjustments that are typically made to ensure that a modified Newton–Raphson approach proceeds steadily toward a maximum. These adjustments are not obviously well-justified in theoretical terms, but they do seem to work in practice.

To see the issues, consider the nonlinear model $Y_i = \phi(\mathbf{X}_i, \theta) + \epsilon_i$, which we represent in the simplified form $Y_i = \phi_i(\theta) + \epsilon_i$. The maximization problem is to

$$\max_\theta L(\theta) = -\frac{1}{2}\sum_i (Y_i - \phi_i(\theta))^2,$$

which, for initial estimate θ_1, leads to the $k \times 1$ gradient

$$\mathbf{g}_1 = \sum_i (Y_i - \phi_i(\theta_1)) \frac{\partial \phi_i(\theta_1)}{\partial \theta},$$

and $k \times k$ Hessian

$$\mathbf{H}_1 = \sum_i (Y_i - \phi_i(\theta_1)) \frac{\partial^2 \phi_i(\theta_1)}{\partial \theta \partial \theta'} - \sum_i \frac{\partial \phi_i(\theta_1)}{\partial \theta} \frac{\partial \phi_i(\theta_1)}{\partial \theta}'.$$

The second term of the Hessian is minus an outer product, and therefore it must be negative definite. The first term, however, could well cause us trouble. Now, at the *true* value of the θ parameter, the expectation of the first term is zero provided that $\mathrm{E}\, \epsilon_i | \mathbf{X}_i = 0$ as we have been assuming. Hence, we might proceed to optimize by ignoring the first term—admittedly, the grounds for doing so are a bit thin—using only the component of the Hessian which is guaranteed to be negative definite in our optimization algorithm.

Application to maximum likelihood

The notation that we have been using applies without much modification if we take $L(\theta)$ to be the log-likelihood function. Here as in other applications, the problem lies in the Hessian matrix of second derivatives. Recall that at the true θ value, which we'll denote here by θ^*, we have this equality for the normalized log-likelihood function,

$$-\mathrm{E}\, \frac{\partial^2 L_n(\theta^*)}{\partial \theta \partial \theta'} = \mathrm{E} \left(\frac{\partial L_n(\theta^*)}{\partial \theta} \right) \left(\frac{\partial L_n(\theta^*)}{\partial \theta} \right)'.$$

This provides a bit of theoretical justification for using

$$-\tilde{\mathbf{H}}_1 = \left(\frac{\partial L(\theta_1)}{\partial \theta} \right) \left(\frac{\partial L(\theta_1)}{\partial \theta} \right)'$$

in which we have dispensed with n, the normalizing factor needed in the theory. The updating rule is then

$$\theta_2 = \theta_1 + \left(\frac{\partial L(\theta_1)}{\partial \theta} \frac{\partial L(\theta_1)}{\partial \theta}' \right)^{-1} \mathbf{g}_1$$

where you will note the change of sign by comparison with the usual rule.

In addition, where the data series are inid, the expected outer product

$$\mathrm{E} \left(\frac{\partial L_n(\theta^*)}{\partial \theta} \right) \left(\frac{\partial L_n(\theta^*)}{\partial \theta} \right)' = \frac{1}{n} \sum_i \mathrm{E}\, \frac{\partial \ln f_i(\theta^*)}{\partial \theta} \frac{\partial \ln f_i(\theta^*)}{\partial \theta}'$$

and in the numerical implementation, this becomes the sum of the outer product of each observation's contribution to the score, that is,

$$\sum_i \frac{\partial \ln f_i(\theta_1)}{\partial \theta} \frac{\partial \ln f_i(\theta_1)}{\partial \theta}'.$$

This method—termed "BHHH" after the economists Berndt, Hall, Hall, and Hausman—is popular because it requires the coding only of first partial derivatives.

17.2 Finding the roots of nonlinear equations

On these issues, Miranda and Fackler (2002) provide a clear and easily readable account. As is the case with maximization of differentiable functions, Newton's method provides the core of the technique, but often needs to be supplemented with other methods when a matrix of derivatives is not guaranteed to be invertible.

Consider the case of k equations in k unknowns, the unknowns being denoted by θ as above. Stack the k equations in a k-vector $\mathbf{h}(\theta)$ and pose the problem in terms of finding the θ^* values that solve the equation system

$$\mathbf{0}_k = \mathbf{h}(\theta^*).$$

Begin with a trial value θ_1 and its associated vector of functions $\mathbf{h}(\theta_1)$. Starting at θ_1, we want to extrapolate $\mathbf{h}$ by linearizing it, using both the level and the slope of the function at this starting point, to find the value θ_2 at which the extrapolated (linearized) function $\mathbf{h}^e(\theta_2) = \mathbf{0}$. The linearized version of the function is

$$\mathbf{h}^e(\theta_2) = \mathbf{h}(\theta_1) + \mathbf{H}(\theta_1)(\theta_2 - \theta_1).$$

in which $\mathbf{H}$ is a $k \times k$ matrix of derivatives of the k component functions in the $\mathbf{h}()$ vector with respect to the k unknown θs.

Equating the right-hand side to zero, we find the candidate solution

$$\theta_2 = \theta_1 - \mathbf{H}(\theta_1)^{-1}\mathbf{h}(\theta_1).$$

With this candidate θ_2 in hand, we now check whether $\mathbf{h}(\theta_2) \approx \mathbf{0}$, which would usually be done by setting a tolerance level $\epsilon > 0$ and checking whether $\|\mathbf{h}(\theta_2)\| < \epsilon$. If $\mathbf{h}(\theta_2)$ is close enough to $\mathbf{0}$ in this sense, we would declare that the root has been found. Otherwise, we begin again to linearize the function, this time using θ_2 as our starting point. The algorithm proceeds until we find a candidate solution that meets the threshold test.

For this algorithm to succeed, it is important that $\mathbf{H}(\theta_j)$ be invertible at all evaluation points; otherwise Newton's method breaks down. This is an important practical weakness of the method, and various "fixes" have been applied to get around it. These fixes are closely analogous to those used in the Newton–Raphson method described above. Indeed, if we think back, we see that Newton–Raphson is simply a root-finding method by which solutions to first-order conditions are located, as can be seen via

$$\mathbf{0}_k = \frac{\partial L(\theta^*)}{\partial \theta} \equiv \mathbf{h}(\theta^*).$$

Hence the maximization–minimization techniques we described earlier are essentially identical to those employed in a root-finding problem.

Part III

The Maximum Likelihood Method: Theory and Applications

Chapter 18

Maximum Likelihood Estimation

Students: You should read this chapter carefully, except for sections 9–11 which can be skimmed through. If you have already have access to this textbook (which will be required for Economics 522) you can supplement the chapter's material by reading Cameron and Trivedi (2005, Sections 5.1–5.6, 5.9–5.10).

The presentation that follows is modeled on Newey and McFadden (1994), which is something of a modern classic. They provide a thorough, rigorous, and extremely well-organized tour of the theory. For most beginning students of econometrics, however, parts of the Newey–McFadden piece are likely to prove hard going, and in what follows I will adopt the less formal presentation style of McFadden (1988), Mittelhammer, Judge, and Miller (2000), and Hayashi (2000), all of whom do a wonderful job of explicating the Newey and McFadden (1994) proofs in more accessible language. All these efforts owe a great deal to Amemiya (1985), an early textbook that has had an enduring influence on the field. For further reading, I would also recommend Ruud (2000), Bierens (2004), and Davidson and MacKinnon (1993, Chapter 8), and helpful background discussions can be found in Greenberg and Webster (1983, Chapter 1), Serfling (1980), and Spanos (1986).

In the maximum likelihood approach, we begin by *fully specifying the distribution* that governs the data-generating process. This distribution depends on the θ parameter, a $k \times 1$ vector, with θ_0 being its true value. The maximum-likelihood estimator $\hat{\theta}_n$ is defined to be the value of θ that maximizes the log of the sample likelihood function for a sample having n observations. What this means will be explained shortly, but in general the ML estimator is found by application of numerical maximization methods. There are special cases in which it can be expressed in a closed form (that is, as an explicit formula) but these are exceptions to the general rule.

To understand the properties of the ML estimator, we will need notation to distinguish between the random variables for the ith observation—for which we'll use upper-case Y_i and $\mathbf{X}_i$ taking the explanatory variables to be a k-vector—and their realized values in the sample that is on hand—for this we'll use the lower-case y_i and $\mathbf{x}_i$. Ignoring explanatory covariates for the moment, we define the sample likelihood function as $L^*(y_1, y_2, \ldots, y_n \mid \theta)$, which can be viewed either as the sample joint density function $f(y_1, y_2, \ldots, y_n \mid \theta)$ when

we direct attention to its dependence on $\mathbf{y}$, or as $L^*(\theta)$, a function of θ, in which the role of $\mathbf{y}$ is suppressed to simplify the notation. We denote the log of the likelihood function by $L(y_1, y_2, \ldots, y_n \mid \theta)$. To work out the asymptotic properties of the maximum-likelihood estimator, however, we'll need to use the *normalized* log-likelihood function in which $\mathbf{Y}$ represents the random variables,

$$L_n(\mathbf{Y} \mid \theta) = \frac{1}{n} L(\mathbf{Y} \mid \theta).$$

I will say more below about how to think about the $\mathbf{X}$ explanatory random variables in this set-up.

For any given sample data $\mathbf{y}$, the maximum-likelihood estimator is the value $\hat{\theta}_n$ that maximizes $L_n(\mathbf{y}|\theta)$, the sample likelihood function. Since in general we do not know the functional form of this estimator, we need some new mathematical tools to help us understand its properties. We begin the sections addressing the consistency and limiting distribution of the maximum likelihood estimator with a statement of some general conditions that do not require the $\{Y_i\}$ sequence to be an iid sequence. We will then narrow our focus to the iid case to show why the results hold, and go on to discuss how to handle inid (independent but not identically distributed) data series.

Although our notation might be read to suggest that Y_i must be a scalar random variable for observation i, this is not at all necessary. We could replace Y_i with a vector $\mathbf{Y}_i$ without altering the essence of the argument in any substantial way. Even for relatively simple models it can be convenient to make use of a $\mathbf{Y}_i$ vector. For example, in the multinomial case in which there are three outcomes, call them Outcome 1, Outcome 2, and Outcome 3, which occur with probabilities p_1, p_2, and $1 - p_1 - p_2$ respectively, it is convenient to let $Y_{i1} = 1$ when Outcome 1 occurs and let it be zero otherwise, and also to let $Y_{i2} = 1$ when Outcome 2 occurs and zero otherwise. A compact expression for the probability of any one of the three outcomes for the i-th observation is

$$f(\mathbf{Y}_i \mid p_1, p_2) = p_1^{Y_{i1}} \cdot p_2^{Y_{i2}} \cdot (1 - p_1 - p_2)^{1 - Y_{i1} - Y_{i2}},$$

in which $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$. There are many other cases in which we would make use of a $\mathbf{Y}_i$ vector—for instance, in modeling a series of longitudinal data on individuals over time periods $t = 1$ to T, with $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{iT})$.

When the data-generating process includes explanatory covariates $\mathbf{X}_i$, the iid assumption does not generally apply. The required modifications to accommodate inid processes are not great, but a little discussion is in order. Assume that the data-generating process yields a sequence $\{(Y_i, \mathbf{X}_i)\}$ that is independent over i but is not necessarily identically distributed. Let the joint density function for the observed data be

$$L^* \left((y_1, \mathbf{x}_1), (y_2, \mathbf{x}_2), \ldots, (y_n, \mathbf{x}_n) \mid \gamma \right)$$

in which we explicitly take note of the dependent variable y_i and covariates $\mathbf{x}_i$ for the i-th observation, as well as γ, a vector of parameters. Under the independence assumption, this can be factored into the product of i-specific densities. Let $g_i(y_i, \mathbf{x}_i \mid \gamma)$ be the joint density for the i-th observation. The normalized log-likelihood function is then

$$L_n(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \ln g_i(y_i, \mathbf{x}_i \mid \gamma).$$

We now take the key step, which is to factor g_i into the product of a conditional and marginal density,

$$g_i(y_i, \mathbf{x}_i \mid \gamma) = f_i(y_i \mid \mathbf{x}_i, \theta) \cdot h_i(\mathbf{x}_i \mid \alpha).$$

and assume—this is the important point—that there is no link whatsoever between the θ parameters that enter the conditional density $f_i(y_i \mid \mathbf{x}_i, \theta)$ and the α parameters that appear in the marginal density for $\mathbf{x}_i$. Given this factorization, we may write

$$L_n(\gamma) = L_n^c(\theta) + L_n^m(\alpha)$$

with $L_n^c(\theta) = \frac{1}{n}\sum_i \ln f_i(y_i \mid \mathbf{x}_i, \theta)$ and similarly for the $L_n^m(\alpha)$ component. Clearly the value $\hat{\theta}_n$ that maximizes the $L_n^c(\theta)$ component of the log-likelihood can be obtained without reference to the $L_n^m(\alpha)$ component. Where θ is concerned, therefore, we can proceed to treat $L_n^c(\theta)$ just *as if* it were the full log-likelihood function.

But what about cases in which the θ and α parameters are linked in some way? In these cases the likelihood function is not separable in the parameters, and we would have to proceed using the full likelihood function. Two-equation sample selection models are one prominent example in microeconometrics, in which probability expressions are developed for the joint distribution of the two random dependent variables of the two equations, conditional on the exogenous explanatory covariates that enter these equations. We will have more to say about covariates later in the chapter.

18.1 The Fundamental Rationale

Although the idea of finding $\hat{\theta}$ by maximizing a log-likelihood function makes sense intuitively, it is worth seeing the formal rationale here at the outset. The case for the ML technique rests on what is termed the *strict expected log-likelihood inequality*. Consider Θ, the set of all allowable values for the parameter. The inequality is that for all θ_0 and $\theta \neq \theta_0$ values in Θ,

$$\mathrm{E}\, L_n(\theta_0) > \mathrm{E}\, L_n(\theta).$$

The condition says that the expected value of the normalized log-likelihood function is highest when that function is evaluated at the true parameter value—i.e., θ_0 maximizes $\mathrm{E}\, L_n(\mathbf{Y}|\theta)$ among all $\theta \in \Theta$. For now we will set aside the question of whether and under what conditions these expectations *exist* and simply assume that they do.

To see the essence of the argument, consider the counterpart expression for the density (or probability mass, if we replace integrals with sums in what follows) of a single Y_i variable, $f_Y(Y_i \mid \theta)$:

$$\mathrm{E}\, \frac{f_Y(Y_i \mid \theta)}{f_Y(Y_i \mid \theta_0)} = \int \left(\frac{f_Y(y \mid \theta)}{f_Y(y \mid \theta_0)} \right) f_Y(y \mid \theta_0)\, dy$$

The right-hand side can be simplified, yielding

$$\mathrm{E} \left(\frac{f_Y(Y_i \mid \theta)}{f_Y(Y_i \mid \theta_0)} \right) = 1.$$

because $f_Y(y \mid \theta)$, although it is not actually the density function of the data-generating process (its argument is θ rather than θ_0), nevertheless has the functional form of a density and therefore integrates to 1. (We'll look more closely at this point in a moment.)

Now consider Jensen's inequality applied to the strictly concave function $\ln()$, from which we have $\mathrm{E}\ln(X) \leq \ln(\mathrm{E}\,X)$. The inequality is strict provided that X is not a constant. Taking X to be $f_Y(Y_i \mid \theta)/f_Y(Y_i \mid \theta_0)$ for the case at hand, we obtain

$$\mathrm{E}\ln\left(\frac{f_Y(Y_i \mid \theta)}{f_Y(Y_i \mid \theta_0)}\right) \leq \ln\left(\mathrm{E}\,\frac{f_Y(Y_i \mid \theta)}{f_Y(Y_i \mid \theta_0)}\right) = \ln(1) = 0,$$

from which it follows that $\mathrm{E}\ln f_Y(Y_i \mid \theta) \leq \mathrm{E}\ln f_Y(Y_i \mid \theta_0)$.

To achieve *strict* inequality in the relationship, we reconsider

$$\mathrm{E}\ln\left(\frac{f_Y(Y_i \mid \theta)}{f_Y(Y_i \mid \theta_0)}\right) = \int \ln\left(\frac{f_Y(y \mid \theta)}{f_Y(y \mid \theta_0)}\right) f_Y(y \mid \theta_0)\, dy$$

and assess the probability of cases with $f_Y(Y_i \mid \theta) \neq f_Y(Y_i \mid \theta_0)$. Clearly, if the maximum-likelihood approach is to work at all, such cases must occur with positive probability for any $\theta \neq \theta_0$, because otherwise the ratio would be constant (with probability 1) for all $\theta \neq \theta_0$ and we could never hope to locate θ_0 by the maximum likelihood method. Assuming that this minimal "identification" assumption is met, we have Jensen's inequality holding in its strict form.

Let's look at a special case that may clarify some of these ideas. Suppose the dgp depends on three parameters a, b, and c, and let the first two of these establish the allowable range for the random variable, such that $Y_i \in [a, b]$, with the density being 0 outside of this range. When integrated over $[a, b]$ the density integrates to 1 for any allowable value of the c parameter. However, note that

$$\mathrm{E}\,\frac{f_Y(Y_i \mid a, b, c)}{f_Y(Y_i \mid a_0, b_0, c_0)} = \int_{a_0}^{b_0} \left(\frac{f_Y(y \mid a, b, c)}{f_Y(y \mid a_0, b_0, c_0)}\right) f_Y(y \mid a_0, b_0, c_0)\, dy = \int_{a_0}^{b_0} f_Y(y \mid a, b, c) dy.$$

The last integral does not necessarily equal 1, depending on where a and b are in relation to the true parameter values a_0 and b_0. We can say, of course, that $0 \leq \int_{a_0}^{b_0} f_Y(y \mid a, b, c) dy \leq 1$. If the integral is 0, we obviously can't go on to take logs, but the general conclusion still goes through (think of $\ln 0 = -\infty$). If the integral is a quantity less than 1, we can proceed to take logs and the remainder of the argument applies without any significant modification.

In iid or inid data, what we have done for a single Y_i and its density (or probability mass) is readily applied to the normalized log-likelihood function $L_n(\mathbf{Y} \mid \theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(Y_i \mid \theta)$. Even with non-independent data, the argument goes through, as can be seen by revisiting the proof using $L^*(\mathbf{Y} \mid \theta)$ rather than $f_Y(Y_i \mid \theta)$ and using multiple rather than single integrals in taking expectations.

These results thus provide us with the essential motivation for the ML approach in general. Knowing that the global maximum of $\mathrm{E}\,L_n(\mathbf{Y} \mid \theta)$ is located at $\theta = \theta_0$, we have some reason to hope that a technique that employs sample averages in place of expectations will have good large-sample properties. The first order of business is to verify our intuition, by investigating more formally the consistency of the ML estimator.

18.2 Consistency

The key to the consistency proof lies in the following important theorem due to Amemiya (1985) and discussed at length in Newey and McFadden (1994), Mittelhammer, Judge, and Miller (2000), Ruud (2000), and Bierens (2004).

Consistency of Maxima Let $\{Q_n(\theta)\}$ be a sequence of functions (depending on random variables) that converges in probability uniformly to a limit function $Q(\theta)$ on a closed and bounded ("compact") parameter space Θ. If the limit function $Q(\theta)$ is continuous and uniquely maximized at θ_0, then the sequence $\{\hat{\theta}_n\}$, with

$$\hat{\theta}_n \equiv \arg\max_{\theta\in\Theta} Q_n(\theta),$$

converges in probability to θ_0.

This remarkable result is discussed further in the references cited above.[1]

No iid or inid assumption enters this theorem, although as we will see, the assumption that the data series is iid makes it easy to establish uniform convergence. Note, too, that the theorem makes no specific reference to maximum likelihood—it can be applied to a variety of estimation problems in which an estimator is derived from the maximization of some reasonably well-behaved Q_n function.

Amemiya's proof of the theorem is simple but ingenious. Let N be an open neighborhood in Θ that contains the point θ_0. The set $\Theta - N$, consisting of all points in Θ that are not in N, is compact. Let the quantity $\epsilon > 0$ be defined as

$$\epsilon = Q(\theta_0) - \max_{\theta\in\Theta-N} Q(\theta)$$

We know that $\epsilon > 0$ because of the assumption that Q is uniquely maximized at θ_0, and since Q is continuous, the compactness of $\Theta - N$ guarantees that Q attains a maximum over this smaller set as well.

Figure 18.1 illustrates the two maxima for the case of $\{Y_i\}$ distributed as iid exponential. Here the true parameter value is $r_0 = 0.10$ and we focus on the neighborhood $N = (0.08, 0.11)$ around r_0. Interpreting $Q(r) = \ln r - r/r_0$ as the expected value of the normalized log-likelihood, we see that its overall maximum (denoted by M in the figure) occurs at r_0 and the maximum achieved by $Q(r)$ outside the neighborhood, which is of course smaller, is indicated by m in the figure. We have $\epsilon \equiv M - m$.

Now consider the event A_n,

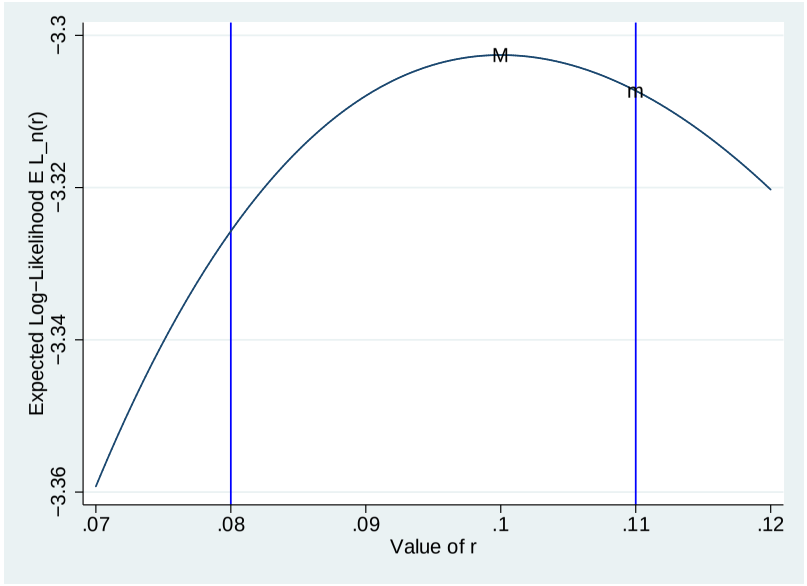
$$A_n \equiv |Q_n(\theta) - Q(\theta)| < \frac{\epsilon}{2} \quad \forall\theta.$$

Clearly A_n is defined with uniform convergence in mind: A_n occurs (for a given n) when $Q_n(\theta)$ is trapped in a sleeve of total width ϵ around $Q(\theta)$ for all values of $\theta \in \Theta$. By the assumption of uniform convergence in probability, as $n \to \infty$ we have $\Pr(A_n) \to 1$. For our exponential example, the sleeve is indicated in dashed lines in the second figure. With this in mind, let's see how the proof unfolds.

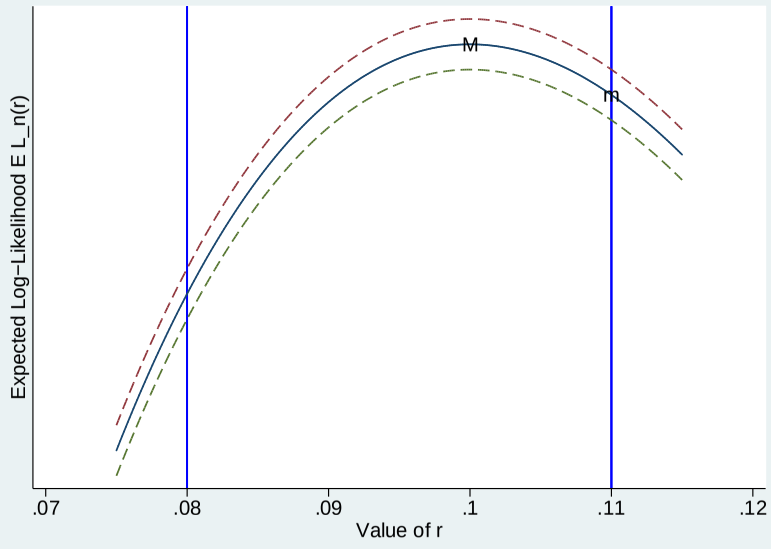
[1] Recall that $\{Q_n(\theta)\}$ is said to converge uniformly in probability to $Q(\theta)$ if for any $\epsilon > 0$ and $\delta > 0$, there is an integer $N_{\epsilon,\delta}$ such that $n > N_{\epsilon,\delta}$ implies $\Pr(\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| < \epsilon) > 1 - \delta$.

Figure 18.1: An illustration of Amemiya's consistency theorem for the iid exponential case with $r_0 = 0.10$.

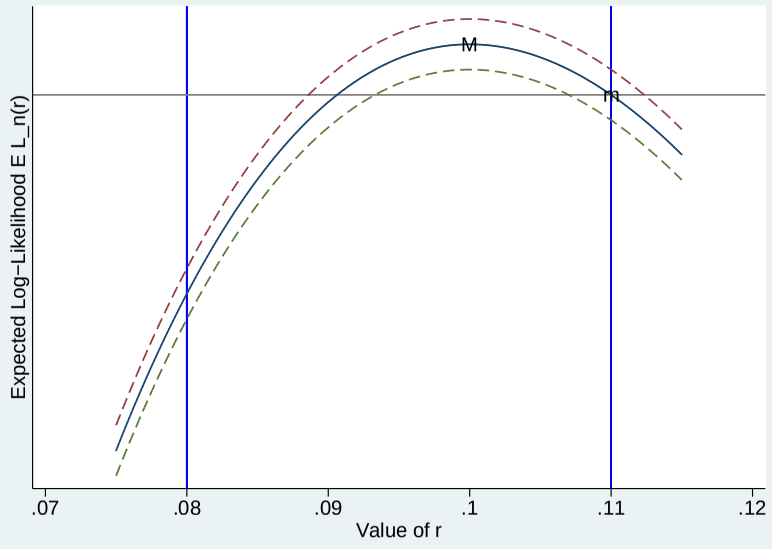
(a) Global maximum $M = Q(r_0)$ and maximum m outside $(0.08, 0.11)$ neighborhood.



(b) Sleeve of height $M - m$.



(c) $Q(\hat{r}) > m$.



As we've said, when A_n occurs this implies

$$Q(\theta) - \frac{\epsilon}{2} < Q_n(\theta) < Q(\theta) + \frac{\epsilon}{2}$$

for all $\theta \in \Theta$. For $\hat{\theta}_n$ in particular, the occurrence of A_n implies that $Q_n(\hat{\theta}_n) < Q(\hat{\theta}_n) + \frac{\epsilon}{2}$, or

$$Q(\hat{\theta}_n) > Q_n(\hat{\theta}_n) - \frac{\epsilon}{2}. \tag{18.1}$$

Also, when A_n occurs,

$$Q_n(\theta_0) > Q(\theta_0) - \frac{\epsilon}{2} \tag{18.2}$$

and we know that in general $Q_n(\hat{\theta}_n) \geq Q_n(\theta_0)$ because $\hat{\theta}_n$ maximizes Q_n. From equation (18.1),

$$Q(\hat{\theta}_n) > Q_n(\hat{\theta}_n) - \frac{\epsilon}{2} > Q_n(\theta_0) - \frac{\epsilon}{2}.$$

Adding this inequality to the inequality of equation (18.2), we obtain

$$Q_n(\theta_0) + Q(\hat{\theta}_n) > Q(\theta_0) - \frac{\epsilon}{2} + Q_n(\theta_0) - \frac{\epsilon}{2},$$

or simply

$$Q(\hat{\theta}_n) > Q(\theta_0) - \epsilon. \tag{18.3}$$

In short, the event $A_n \Rightarrow Q(\hat{\theta}_n) > Q(\theta_0) - \epsilon$.

This result implies that $\hat{\theta}_n \in N$, the neighborhood in which the true θ_0 is found. Why does this last implication follow? If we return to the definition of ϵ and rearrange it, we have

$$\max_{\theta \in \Theta - N} Q(\theta) = Q(\theta_0) - \epsilon.$$

Hence, the occurrence of

$$A_n \Rightarrow Q(\hat{\theta}_n) > \max_{\theta \in \Theta - N} Q(\theta).$$

Clearly we must have $\hat{\theta}_n \in N$ if this inequality is to hold. (In our example, as shown in the third figure, $Q(\hat{\theta}_n) > m$. We do not know what value of $\hat{\theta}$ obtains, but we do know that whatever its value, $Q(\hat{\theta}_n)$ exceeds the level indicated by the horizontal line.) Since the event $A_n \Rightarrow \hat{\theta}_n \in N$, it must be the case that

$$\Pr(\hat{\theta}_n \in N) \geq \Pr(A_n)$$

Given uniform convergence, $\lim_{n\to\infty} \Pr(A_n) = 1$, and thus $\lim_{n\to\infty} \Pr(\hat{\theta}_n \in N) = 1$ as well, no matter how small we make the N neighborhood of the true θ_0. Indeed, we can make N arbitrarily small and still have the result $\lim_{n\to\infty} \Pr(\hat{\theta}_n \in N) = 1$. To express this in other words, we find that $\hat{\theta}_n \xrightarrow{p} \theta_0$.

To put this remarkable result to use in the context of maximum likelihood, we will work with the normalized log-likelihood function $L_n(\mathbf{Y} \mid \theta)$, which will play the role of $Q_n(\theta)$ in Amemiya's theorem, and $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ will play the role of the $Q(\theta)$ limit function. The essential conditions for consistency of the maximum likelihood estimator, which apply whether the $\{Y_i\}$ series is iid, inid, or dependent over i, may be stated as follows (Mittelhammer, Judge, and Miller 2000).

C1. The parameter $\theta \in \Theta$, a set that is closed and bounded (compact).

C2. The expectation $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ exists.

C3. The function $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ is continuous in θ.

C4. The true parameter value θ_0 maximizes $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$, that is,

$$\theta_0 = \arg\max_{\theta \in \Theta} \mathrm{E}\, L_n(\mathbf{Y} \mid \theta),$$

and θ_0 is unique.

C5. The normalized log-likelihood function converges uniformly in probability to the limit function $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$. In other words, for any $\epsilon > 0$ and $\delta > 0$, there is an integer $N_{\epsilon,\delta}$ such that $n > N_{\epsilon,\delta}$ implies $\Pr(\sup_{\theta \in \Theta} |L_n(\theta) - \mathrm{E}\, L_n(\theta)| < \epsilon) > 1 - \delta$.

When these consistency conditions are met, the maximum-likelihood estimator $\hat{\theta}_n \xrightarrow{p} \theta_0$, the true parameter value. In this chapter we will focus on the iid and inid cases, but you should know that more general cases can also be explored, such as in time-series analysis.

Discussion

There is an alternative to the first consistency condition **C1** but we take it as given and move on to examine the circumstances under which conditions **C2**–**C5** hold.[2] Condition **C2** is met if what Ruud (2000) terms the *Dominance I* condition holds.

Dominance I The expectation $\mathrm{E} \sup_{\theta \in \Theta} |L_n(\mathbf{Y} \mid \theta)|$ exists. As explained by Ruud (2000, p. 290), this is described in terms of "dominance" because it directs attention to a function $H(y) \equiv \sup_{\theta \in \Theta} |L_n(y \mid \theta)|$ that does not depend on θ and which is always at least as large as $\mid L_n(y \mid \theta) \mid$ for any admissible θ. If $\mathrm{E}\ H(\mathbf{Y})$ exists, then this implies the existence of $\mathrm{E} \mid L_n(\mathbf{Y} \mid \theta) \mid$ and thereby the existence of $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ for all $\theta \in \Theta$. The dominance condition is related to conditions that permit differentiation under an integral sign (see below) but its main role here is to establish the existence of the limit function in **C5**, the uniform convergence condition.

If the continuity condition **C3** is met, it will ensure that the maximum of the expected log-likelihood function $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ exists, since a continuous function must attain a maximum over a closed and bounded (compact) set (here, the set is Θ, whose compactness is assumed in condition **C1**). As we will see shortly, in the iid case we will be able to prove continuity of $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ by assuming that $L_n(y, \theta)$ is continuous and invoking the uniform law of large numbers (ULLN) to show that the continuity property carries over to $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$. For non-iid data series, however, it is more difficult to establish that condition **C3** holds.[3]

We have already examined condition **C4** in connection with the fundamental rationale for maximum likelihood. Condition **C4** holds under a global identifiability assumption.

[2] See Newey and McFadden (1994) on how concavity of the limit function and convexity of Θ can substitute for compactness of Θ and uniform convergence.

[3] Note that it is possible for $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ to be continuous in θ even when $L_n(\mathbf{Y} \mid \theta)$ is not, provided that the act of taking expectations somehow "smooths away" any discontinuities.

Global Identification The true parameter value θ_0 is globally identified; that is, for all $\theta \in \Theta$ such that $\theta \neq \theta_0$,

$$\Pr\bigl(L^*(\mathbf{Y} \mid \theta_0) \neq L^*(\mathbf{Y} \mid \theta)\bigr) > 0.$$

We addressed the identification idea earlier. To understand identification, suppose that for some $\theta \neq \theta_0$, it happens that the joint density $L^*(y \mid \theta_0) = L^*(y \mid \theta)$ for all measureable values of y, that is, all y values that occur with positive probability. Then the data would not allow us to determine whether θ or θ_0 is the true value of the parameter—not even an infinite amount of data could settle the question. Evidently, if the parameter θ_0 is to be identified, the joint densities $L^*(y \mid \theta_0)$ and $L^*(y \mid \theta)$ must differ for at least some measureable y values.

As we've already discussed, the identifiability assumption generates what is termed the *strict expected log-likelihood inequality*, which is consistency condition **C4** under another name. To re-state it, the inequality is

$$\mathrm{E}\, L_n(\theta_0) > \mathrm{E}\, L_n(\theta),$$

for all $\theta \in \Theta$ such that $\theta \neq \theta_0$.

Condition **C5**, having to do with uniform convergence in probability, is difficult to verify in general (see Newey and McFadden (1994) on how the concept of *stochastic equicontinuity* can help), but if the data series is iid, we can establish **C5** using the uniform law of large numbers, which you may recall from the very end of our earlier chapter on laws of large numbers. In the iid case $L_n(\mathbf{Y} \mid \theta) = n^{-1} \sum_{i=1}^{n} \ln f(Y_i \mid \theta)$, we see that the normalized log-likelihood function is an average of iid random variables. Hence the strict expected log-likelihood inequality can be expressed in the form $\mathrm{E} \ln f(Y_i \mid \theta) < \mathrm{E} \ln f(Y_i \mid \theta_0)$ for any i and for any $\theta \neq \theta_0$, so that the limit function simplifies to

$$\mathrm{E}\, L_n(\mathbf{Y} \mid \theta) = \mathrm{E} \ln f(Y_i \mid \theta).$$

The uniform law of large numbers applies to an iid sequence of random variables such as $\{\ln f(Y_i \mid \theta)\}$. It can be restated for the present context as follows.

ULLN *Assume* that $\ln f(Y \mid \theta)$ is continuous in $\theta \in \Theta$, a closed and bounded subset. If the expected value

$$\mathrm{E} \sup_{\theta \in \Theta} |\ln f(Y \mid \theta)|$$

exists, then the function $\mathrm{E}\ \ln f(Y \mid \theta)$ exists and is *continuous* in $\theta \in \Theta$. The sample average converges to its expected value uniformly in θ, that is

$$\operatorname{plim} \left(\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i}^{n} \ln f(Y_i \mid \theta) - \mathrm{E} \ln f(Y_i \mid \theta) \right| \right) = 0,$$

as discussed by Ruud (2000).

To put the result differently, in the iid case $L_n(\mathbf{Y} \mid \theta)$ converges uniformly in probability to the limit function $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ and $\mathrm{E}\, L_n(\mathbf{Y} \mid \theta)$ is continuous in θ. Given the iid assumption,

we obtain these two results at a modest price: we must assume only that $\ln f(y \mid \theta)$ is continuous in θ and that $\mathrm{E} \sup_{\theta \in \Theta} |\ln f(Y_i \mid \theta)|$ exists. Since to implement the ML estimator you would need to specify the functional form of $f(y \mid \theta)$ in any case, verifying that f is continuous in θ would generally be an easy task.[4]

18.3 Asymptotic Normality

With consistency established, it is not too difficult to see why the ML estimator, when suitably transformed, should converge to a normal distribution. The conditions yielding a limiting normal distribution are as follows.

N1. The estimator $\hat{\theta}_n \xrightarrow{p} \theta_0$, the true value of θ.

N2. θ_0 is in the *interior* of Θ.

N3. $L_n(\mathbf{y} \mid \theta)$ is twice continuously differentiable in θ, at least in a neighborhood of θ_0.

N4. The normalized score vector, when evaluated at the true θ_0 and multiplied by $\sqrt{n}$, converges in distribution to multivariate normal with mean $\mathbf{0}$ and variance matrix Σ, i.e.,

$$\sqrt{n}\frac{\partial L_n(\mathbf{Y} \mid \theta_0)}{\partial \theta} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Sigma).$$

N5. The $k \times k$ matrix of second derivatives of the normalized log-likelihood converges uniformly in probability to a matrix function $\mathbf{H}(\theta)$,

$$\operatorname{plim}\left(\sup_{\theta \in \Theta}\left|\frac{\partial^2 L_n(\mathbf{Y} \mid \theta)}{\partial \theta \partial \theta'} - \mathbf{H}(\theta)\right|\right) = \mathbf{0}.$$

N6. The matrix $\mathbf{H}(\theta)$ is continuous and nonsingular at $\theta = \theta_0$.

Under these conditions, $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{H}_0^{-1}\Sigma\,\mathbf{H}_0^{-1})$ with $\mathbf{H}_0^{-1}$ being the inverse of $\mathbf{H}(\theta_0)$. As we will see shortly, with the aid of a few additional assumptions, the variance of the limiting distribution can be simplified.

Discussion

Condition **N2** implies that the maximum likelihood estimator $\hat{\theta}_n$ can be found by locating the value of θ that solves the first-order condition, that the (normalized) score equals zero,

$$\frac{\partial L_n(\hat{\theta}_n)}{\partial \theta} = \mathbf{0}.$$

[4]For those of you interested in time-series analysis in which the iid assumption does not hold, note that the ULLN also applies if the data series is strictly stationary and ergodic—see Hayashi (2000) for this important case.

Our first step is to expand the score about θ_0,

$$\mathbf{0} = \frac{\partial L_n(\hat{\theta}_n)}{\partial \theta} = \frac{\partial L_n(\theta_0)}{\partial \theta} + \frac{\partial^2 L_n(\bar{\theta}_n)}{\partial \theta \partial \theta'}(\hat{\theta}_n - \theta_0),$$

where $\bar{\theta}_n$ lies between $\hat{\theta}_n$ and θ_0. We now make use of the uniform convergence condition N5, together with the consistency of $\hat{\theta}_n$ that ensures the convergence of $\bar{\theta}_n \to \theta_0$. By these conditions,

$$\frac{\partial^2 L_n(\bar{\theta}_n)}{\partial \theta \partial \theta'} \xrightarrow{p} \mathbf{H}(\theta_0).$$

See Newey and McFadden (1994), Ruud (2000), Mittelhammer, Judge, and Miller (2000), and Davidson and MacKinnon (1993, pp. 261–262) to learn exactly why this is so. (The proof involves advanced mathematical techniques that we have not explored in this course.) Taking the result as given and inserting it into the above, we have

$$\mathbf{0} \overset{a}{=} \frac{\partial L_n(\theta_0)}{\partial \theta} + \mathbf{H}(\theta_0) \cdot (\hat{\theta}_n - \theta_0).$$

Multiplying by $\sqrt{n}$ and rearranging, using condition **N6** in the process, we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{a}{=} -\mathbf{H}_0^{-1} \cdot \sqrt{n}\frac{\partial L_n(\theta_0)}{\partial \theta}.$$

We now appeal to condition **N4**, by which

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \overset{a}{=} -\mathbf{H}_0^{-1} \cdot \mathcal{N}(\mathbf{0}, \Sigma)$$

and thus

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{H}_0^{-1}\Sigma\,\mathbf{H}_0^{-1}\right),$$

which is what we sought to prove.

We can make further progress if we introduce some simplifying assumptions.

IID The $\{Y_i\}$ data series is iid.

Support The support of Y does not depend on θ. This allows us to differentiate an integral involving $L_n(y \mid \theta)$, and functions (such as joint densities) that are related to it, without attending to the limits of integration.

Interchanging integrals and derivatives Differentiation and integration are interchangeable in the sense that for the joint densities L^*,

$$\frac{\partial}{\partial \theta}\int L^*(\mathbf{y} \mid \theta)dy = \int \frac{\partial L^*(\mathbf{y} \mid \theta)}{\partial \theta}dy,$$

and likewise for the second derivatives. The same sort of condition applies to first and second derivatives of the normalized log-likelihoods $L_n(y \mid \theta)$. As Chapter 2 points out (also see Ruud 2000, p. 299), there are "dominance conditions" that show precisely what is needed to interchange these operations.

Expectation of the score The $k \times 1$ vector of expected values $\mathrm{E}(\partial L_n(\mathbf{Y} \mid \theta)/\partial\theta)$ exists. (This was likely to be a required condition for the convergence in distribution condition **N4** in any case.)

Dominance II The expectation

$$\mathrm{E} \sup_{\theta \in \Theta} \left| \frac{\partial^2 L_n(\mathbf{Y}|\theta)}{\partial\theta\partial\theta'} \right|$$

exists—compare this with the *Dominance I* condition.

As noted above, the iid assumption yields $L_n(\mathbf{Y} \mid \theta) = n^{-1} \sum_{i=1}^{n} \ln f(Y_i \mid \theta)$, with $\{\ln f(Y_i \mid \theta)\}$ being an iid sequence of random variables. The next steps in the argument will remind you of what we did to find the Cramér-Rao lower bound (you may want to re-read Chapter 13), which also required us to interchange differentiation and integration and required that the support of the random variable not be a function of θ. In that discussion, we showed that the expected value of the $k \times 1$ score vector, when it is evaluated at the true θ_0, is

$$\mathrm{E}\, \frac{\partial L_n(\mathbf{Y} \mid \theta_0)}{\partial\theta} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\, \frac{\partial \ln f(Y_i \mid \theta_0)}{\partial\theta} = \mathbf{0}.$$

Furthermore, because the expected value of $\partial \ln f(Y_i \mid \theta_0)/\partial\theta$ is a vector of zeroes, its $k \times k$ variance matrix is

$$\mathrm{Var}\, \frac{\partial \ln f(Y_i \mid \theta_0)}{\partial\theta} = \mathrm{E} \left(\frac{\partial \ln f(Y_i \mid \theta_0)}{\partial\theta} \right) \left(\frac{\partial \ln f(Y_i \mid \theta_0)}{\partial\theta} \right)' = -\mathrm{E}\, \frac{\partial^2 \ln f(Y_i \mid \theta_0)}{\partial\theta\partial\theta'},$$

as we also showed in Chapter 13. Let

$$\mathcal{J} \equiv \mathrm{Var}\, \frac{\partial \ln f(Y_i \mid \theta_0)}{\partial\theta}$$

and note that $\mathcal{J}$ can be expressed using either of the forms above, that is, as the expected value of an outer product or as the negative of the expected value of the second partial derivatives.

Consider normality condition **N4**. By writing out the derivative, we find that

$$\sqrt{n} \frac{\partial L_n(\mathbf{Y} \mid \theta_0)}{\partial\theta} = \sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \frac{\partial \ln f(Y_i \mid \theta_0)}{\partial\theta}.$$

Because the Y_i are iid and each term in the sum has mean $\mathbf{0}$ and variance $\mathcal{J}$, the multivariate version of the Lindeberg–Levy central limit theorem can be applied. It yields

$$\sqrt{n} \frac{\partial L_n(\theta_0)}{\partial\theta} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}).$$

This is an important result in itself; as we will see, it provides the foundation for a class of statistical testing procedures known as Lagrange Multiplier tests.

The Dominance II condition implies (we ignore some technical details here) that the matrix $\partial^2 L_n(\hat{\theta}_n)/\partial\theta\partial\theta'$ is uniformly integrable, which in turn implies that its probability limit coincides with (the limit of) its expectation. Now, in the iid case

$$\frac{\partial^2 L_n(\hat{\theta}_n)}{\partial\theta\partial\theta'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln f(Y_i \mid \hat{\theta}_n)}{\partial\theta\partial\theta'},$$

and if we could evaluate each term in the sum at θ_0 rather than $\hat{\theta}_n$, each would have expectation equal to $-\mathcal{J}$. We have already established that $\hat{\theta}_n \to \theta_0$ under conditions **C1**–**C5**, and (again passing lightly over the technical detail) we obtain the result that,

$$\operatorname{plim}\frac{\partial^2 L_n(\hat{\theta}_n)}{\partial\theta\partial\theta'} = \mathrm{E}\,\frac{\partial^2 L_n(\theta_0)}{\partial\theta\partial\theta'} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\,\frac{\partial^2 \ln f(Y_i \mid \theta_0)}{\partial\theta\partial\theta'} = -\mathcal{J}.$$

In terms of our earlier notation, the variance matrix of the score $\Sigma = \mathcal{J}$ and $-\mathbf{H}_0 = \mathcal{J}$. Making these substitutions, we obtain a much-simplified expression for the limiting distribution,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}^{-1}) \tag{18.4}$$

in the iid case. An often-seen alternative expression is

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \stackrel{a}{=} \mathcal{J}^{-1}\sqrt{n}\frac{\partial L_n(\theta_0)}{\partial\theta}. \tag{18.5}$$

Introducing explanatory covariates

What is different about this analysis when the data-generating process includes covariates? Earlier in discussing the inid case with covariates, we distinguished the conditional $L_n^c(\theta)$ from the marginal $L_n^m(\alpha)$ component of the full log-likelihood function and explained the conditions under which we could estimate θ using the conditional component alone. Consider for the inid case the limiting behavior of

$$\sqrt{n}\frac{\partial L_n^c(\theta_0)}{\partial\theta} = \sqrt{n}\frac{1}{n}\sum_{i}\frac{\partial \ln f_i(\theta_0)}{\partial\theta},$$

in which $f_i(\theta_0)$ is shorthand for the conditional density $f_i(Y_i|\mathbf{X}_i, \theta_0)$. By iterated expectations (conditioning first on $\mathbf{X}_i$), each of the terms $\partial \ln f_i(\theta_0)/\partial\theta$ in the sum will have an expectation of zero under the Cramér–Rao regularity conditions we spelled out earlier. Their variances, however, will generally differ from one observation to the next. Let

$$\mathcal{J}_i = \mathrm{E}\,\frac{\partial \ln f_i(\theta_0)}{\partial\theta}\frac{\partial \ln f_i(\theta_0)}{\partial\theta}'$$

represent the variance matrix of $\partial \ln f_i(\theta_0)/\partial\theta$. When the variances differ over i, we require the Lindeberg central limit theorem. Assuming that the Lindeberg condition is met, which allows that theorem to be applied, and further assuming that

$$\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathcal{J}_i = \mathcal{J},$$

we can proceed much as before to obtain the result

$$\sqrt{n}\frac{\partial L_n^c(\theta_0)}{\partial \theta} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}),$$

and from this,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}^{-1}).$$

The notation in these expressions is the same as in the iid case, but you should understand that the meaning of the $\mathcal{J}$ matrix is slightly different.

18.4 Consistent Estimation of the ML Variance

For the iid case, Newey and McFadden (1994) show that $\hat{\mathcal{J}}_n \xrightarrow{p} \mathcal{J}$, where either

$$\hat{\mathcal{J}}_n = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial \ln f(y_i \mid \hat{\theta}_n)}{\partial \theta}\right)\left(\frac{\partial \ln f(y_i \mid \hat{\theta}_n)}{\partial \theta}\right)',$$

or

$$\hat{\mathcal{J}}_n = -\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 \ln f(y_i \mid \hat{\theta}_n)}{\partial \theta \partial \theta'}.$$

The formal proof of consistency once again uses the concept of uniform convergence.

Note what is actually involved in estimating $\mathcal{J}$. Given the maximum likelihood estimate $\hat{\theta}_n$, you find the derivative of each observation's contribution to the sample log-likelihood. Each derivative is then squared—for a θ vector, calculate the outer product—and the average is taken over the sample. Alternatively, you can find the matrix of second derivatives and (after multiplying by -1) take the average over them instead. In either case, if you are programming the likelihood function yourself for some new problem, $\mathcal{J}$ can be estimated with a bit of extra code.

18.5 Examples

Here we describe the concepts we have been developing by way of two examples, the first involving the exponential distribution and the second the linear regression model assuming normally-distributed disturbance terms.

The exponential model

Consider the case in which $\{Y_i\}$ is iid exponential with parameter r and let r_0 denote the true value of the parameter. Then $\ln f(Y_i|r) = \ln r - rY_i$. The normalized log-likelihood function is $L_n(r) = \ln r - r\bar{Y}$ with $\bar{Y}$ being the sample mean of the Y_i variables. We find the maximum likelihood estimate by differentiating L_n and setting the derivative to zero,

$$\frac{d}{dr}L_n(r) = \frac{1}{\hat{r}_n} - \bar{Y} = 0,$$

which gives $\hat{r}_n = 1/\bar{Y}$. Recalling that $\mathrm{E}\, Y_i = 1/r_0$, it is easy to show that $\mathrm{E} \ln f(Y_i|r)$ is maximized at $r = r_0$, and of course the same is true for $\mathrm{E}\, L_n(r)$.

To find $\mathcal{J}$ in the iid case, we consider the first and second derivatives of the log density, which are

$$\frac{d}{dr} \ln f_i(r) = \frac{1}{r} - Y_i,$$

which has expectation zero when $r = r_0$ (as does the derivative of the normalized log-likelihood), and

$$\frac{d^2}{dr^2} \ln f_i(r) = -\frac{1}{r^2}.$$

In this case

$$\mathcal{J} = -\,\mathrm{E}\, \frac{d^2}{dr^2} \ln f_i(r_0) = \frac{1}{r_0^2}.$$

Hence, $\sqrt{n}(\hat{r}_n - r_0) \xrightarrow{d} \mathcal{N}(0, r_0^2)$ and we would estimate the variance by $\hat{r}_n^2$. Also,

$$\sqrt{n} \frac{d}{dr} L_n(r_0) = \sqrt{n} \left(\frac{1}{r_0} - \bar{Y} \right) \xrightarrow{d} \mathcal{N}(0, 1/r_0^2).$$

To adapt the argument to the inid case in which Y_i is exponential with parameter r_i, we introduce the $\mathbf{X}_i$ covariates via the specification $\ln r_i = \mathbf{X}_i'\theta$. The conditional mean of Y_i is

$$\mathrm{E}\, Y_i|\mathbf{X}_i = \frac{1}{e^{\mathbf{X}_i'\theta_0}}.$$

The log of the density is

$$\ln f_i(\theta) = \mathbf{X}_i'\theta - e^{\mathbf{X}_i'\theta} Y_i$$

and its derivative with respect to θ is the $k \times 1$ vector

$$\frac{\partial}{\partial \theta} \ln f_i(\theta) = \mathbf{X}_i - \mathbf{X}_i e^{\mathbf{X}_i'\theta} Y_i.$$

Conditional on $\mathbf{X}_i$, this derivative has expectation zero when it is evaluated at the true θ_0. Hence, its unconditional expectation is also zero. The matrix of second partial derivatives is

$$\frac{\partial^2}{\partial \theta \partial \theta'} \ln f_i(\theta) = -\mathbf{X}_i \mathbf{X}_i' \cdot e^{\mathbf{X}_i'\theta} Y_i.$$

To find its expectation at θ_0, we again first condition on $\mathbf{X}_i$, which gives us $-\mathbf{X}_i\mathbf{X}_i'$ for the conditional expectation, and then we uncondition. After multiplying by -1 we obtain

$$\mathcal{J}_i = \mathrm{E}\, \mathbf{X}_i \mathbf{X}_i'.$$

Assuming that the Lindeberg condition is met and that

$$\mathcal{J} = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathcal{J}_i = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}\, \mathbf{X}_i \mathbf{X}_i'$$

we have the result that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}^{-1})$$

and we would estimate $\mathcal{J}$ by

$$\hat{\mathcal{J}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i \mathbf{X}_i'.$$

The normal linear model

Assume that in the linear model $Y_i = \mathbf{X}_i'\beta + \epsilon_i$ the disturbance terms are distributed as $\mathcal{N}(0, \sigma^2)$ conditional on $\mathbf{X}_i$, and assume that the disturbances are independent across observations. The log of the conditional density for the i-th observation is then

$$\ln f_i = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}(Y_i - \mathbf{X}_i'\beta)^2,$$

and we can use this to obtain

$$\frac{\partial \ln f_i}{\partial \beta} = \frac{1}{\sigma^2}\mathbf{X}_i \cdot (Y_i - \mathbf{X}_i'\beta)$$

and

$$\frac{\partial \ln f_i}{\partial \sigma^2} = -\frac{1}{2}\frac{1}{\sigma^2} + \frac{1}{2}\frac{1}{\sigma^4}(Y_i - \mathbf{X}_i'\beta)^2.$$

When evaluated at the true values of β and σ^2, these derivatives have conditional expectations of zero (and thus unconditional expectations of zero as well). The second derivatives of interest (it is easy to see that the cross-partial has an expectation of zero) are these:

$$\frac{\partial^2 \ln f_i}{\partial \beta \partial \beta'} = -\frac{1}{\sigma^2}\mathbf{X}_i\mathbf{X}_i',$$

and

$$\frac{\partial^2 \ln f_i}{\partial \sigma^2 \partial \sigma^2} = \frac{1}{2}\frac{1}{\sigma^4} - \frac{1}{\sigma^6}(Y_i - \mathbf{X}_i'\beta)^2.$$

Multiplying each by -1 and taking expected values (again setting the parameters to their true values) yields

$$\mathrm{E}\left(-\frac{\partial^2 \ln f_i}{\partial \beta \partial \beta'}\right) = \frac{1}{\sigma_0^2}\mathrm{E}\,\mathbf{X}_i\mathbf{X}_i',$$

and

$$\mathrm{E}\left(-\frac{\partial^2 \ln f_i}{\partial \sigma^2 \partial \sigma^2}\right) = \frac{1}{2\sigma_0^4}.$$

Because the cross-partial $\partial^2 \ln f_i / \partial\beta\partial\sigma^2$ has an expected value of zero, we obtain

$$\mathcal{J}_i = \begin{bmatrix} \frac{1}{\sigma_0^2}\mathrm{E}\,\mathbf{X}_i\mathbf{X}_i' & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_0^4} \end{bmatrix},$$

and thus, assuming that the limit exists,

$$\mathcal{J} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathcal{J}_i = \lim_{n\to\infty}\begin{bmatrix} \frac{1}{\sigma_0^2}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\,\mathbf{X}_i\mathbf{X}_i' & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_0^4} \end{bmatrix}.$$

We would approximate this by substituting $\hat{\sigma}_n^2$ for σ_0^2 and by using the sum

$$\frac{1}{n}\sum_i \mathbf{X}_i\mathbf{X}_i'$$

in place of the limiting average of the expected outer products.

18.6 The Wald and Lagrange Multiplier Tests

The Wald, Lagrange Multiplier, and Likelihood Ratio tests dominate hypothesis testing in econometrics. The three tests are asymptotically equivalent, and the choice among them usually hinges on ease of computation, an area in which the Wald or Lagrange tests may have an advantage in complicated problems. In exploring the Wald and Lagrange Multiplier tests, we follow the approach of McFadden (1988), which is developed more formally in Newey and McFadden (1994). We begin by assuming that the null hypothesis is of the form $H_0 : \theta = \theta_0$. Later we will see how to handle linear and non-linear constraints that may involve only a subset of the parameters.

The Wald Test

The Wald test is straightforward. We have already seen that

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}^{-1}),$$

and it follows that

$$\sqrt{n}(\hat{\theta}_n - \theta_0)' \mathcal{J} \sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \chi^2_k,$$

where k is the dimension of θ. The asymptotic distribution is not affected if $\mathcal{J}$ is replaced by a consistent estimator, giving

$$W = \sqrt{n}(\hat{\theta}_n - \theta_0)' \hat{\mathcal{J}}_n \sqrt{n}(\hat{\theta}_n - \theta_0)$$

as the Wald statistic.[5] In short, to calculate the Wald statistic given a null hypothesis $H_0 : \theta = \theta_0$, we require the unconstrained ML estimator $\hat{\theta}_n$ and the unconstrained estimator $\hat{\mathcal{J}}_n$.

It is almost as easy to derive the test statistic for a set of q linear constraints on the parameters, $H_0 : \mathbf{R}\theta_0 = \mathbf{r}$, with q being the number of rows (hence, the rank) of $\mathbf{R}$. Under the null, $\mathbf{r} = \mathbf{R}\theta_0$ and thus $\mathbf{R}\hat{\theta}_n - \mathbf{r} = \mathbf{R}(\hat{\theta}_n - \theta_0)$. Multiplying by $\sqrt{n}$, we see that $\sqrt{n}(\mathbf{R}\hat{\theta} - \mathbf{r}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{R}\mathcal{J}^{-1}\mathbf{R}')$ under the null, yielding the result

$$\sqrt{n}(\mathbf{R}\hat{\theta}_n - \mathbf{r})'[\mathbf{R}\hat{\mathcal{J}}^{-1}\mathbf{R}']^{-1}\sqrt{n}(\mathbf{R}\hat{\theta}_n - \mathbf{r}) \xrightarrow{d} \chi^2_q.$$

We will discuss nonlinear constraints shortly, from which the result just shown can be obtained as a special case.

The Lagrange Multiplier or Score Test

We showed above that the normalized score

$$\sqrt{n}\,\frac{\partial L_n(\theta_0)}{\partial \theta} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}).$$

[5] Note that $\mathcal{R} = n\mathcal{J}$ is the information matrix in the iid case. The test statistic is often presented in this way, by combining the two $\sqrt{n}$ factors and drawing them to the middle of the quadratic form.

If the null hypothesis $H_0 : \theta = \theta_0$ is true, then

$$\sqrt{n}\left(\frac{\partial L_n(\theta_0)}{\partial \theta}\right)' \mathcal{J}^{-1} \sqrt{n}\left(\frac{\partial L_n(\theta_0)}{\partial \theta}\right) \xrightarrow{d} \chi^2_k,$$

which is the same asymptotic distribution as for the Wald test.

To form the test statistic, we need a consistent estimator for $\mathcal{J}$, and under the null this is provided by

$$\tilde{\mathcal{J}} = \frac{1}{n}\sum_i \left(\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta}\right)\left(\frac{\partial \ln f(y_i \mid \theta_0)}{\partial \theta}\right)'.$$

The alternative expression for $\tilde{\mathcal{J}}$, which involves second partial derivatives, can also be used. Note the use of the constrained θ_0 in $\tilde{\mathcal{J}}$.

18.7 Tests of Nonlinear Hypotheses

Now consider the problem of testing nonlinear hypotheses. Let $r(\theta)$ be a single nonlinear function of θ. The null hypothesis is expressed as $H_0 : r(\theta_0) = 0$. We will assume that r possesses continuous first and second derivatives. Let $\mathbf{R}(\theta)$ be the $k \times 1$ first derivative vector.

An expansion of $r(\hat{\theta})$ around the true value θ_0 yields

$$r(\hat{\theta}) = r(\theta_0) + \mathbf{R}(\theta^*)'(\hat{\theta} - \theta_0)$$

where θ^* lies between θ_0 and $\hat{\theta}$. Insert this Taylor expansion into the expression $\sqrt{n}(r(\hat{\theta}) - r(\theta_0))$. Then

$$\sqrt{n}\left(r(\hat{\theta}) - r(\theta_0)\right) = \mathbf{R}(\theta^*)'\sqrt{n}(\hat{\theta} - \theta_0).$$

We know that $\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}^{-1})$. The right-hand side will therefore converge in distribution to $\mathcal{N}(0, \mathbf{R}'\mathcal{J}^{-1}\mathbf{R})$ with $\mathbf{R} = \mathbf{R}(\theta_0)$. The Wald test of the null hypothesis $H_0 : r(\theta_0) = 0$ is implemented via the test statistic

$$W = \sqrt{n}r(\hat{\theta})'[\mathbf{R}(\hat{\theta})'\hat{\mathcal{J}}^{-1}\mathbf{R}(\hat{\theta})]^{-1}\sqrt{n}r(\hat{\theta}) \xrightarrow{d} \chi^2_1.$$

The same kind of argument would apply to a vector of nonlinear constraints.

Where nonlinear constraints are concerned, the Lagrange Multiplier approach is likely to be impractical—it would require embedding the constraint(s) in the specification of the null model. In general this would be much more difficult to implement than the Wald test. However, it is worth seeing how it might be done, and there are two side benefits. The proof supplies an expression for the asymptotic distribution of the constrained estimator. Also, it gives us an essential ingredient that will be needed to establish the properties of Likelihood Ratio tests. Our presentation follows Hayashi (2000, pp. 487–493), and similar versions can be found in Newey and McFadden (1994), Davidson and MacKinnon (1993, pp. 276–278), and Mittelhammer (1996, pp. 616–619).

The constrained maximization problem can be written in terms of the Lagrangian $L_n(\theta) - \mathbf{r}(\theta)'\lambda$ where θ is of dimension k and λ is a q-vector of Lagrange multipliers that

are associated with q nonlinear constraints. The first-order conditions for the constrained $\tilde{\theta}$ and $\tilde{\lambda}$ are

$$\frac{\partial L_n(\tilde{\theta})}{\partial \theta} - \mathbf{R}(\tilde{\theta})'\tilde{\lambda} = \mathbf{0} \tag{18.6}$$

$$-\mathbf{r}(\tilde{\theta}) = \mathbf{0} \tag{18.7}$$

where $\mathbf{R}(\tilde{\theta})$ is a $q \times k$ matrix with typical element $\partial r_i/\partial \theta_j$ for the i-th constraint and j-th parameter. You may want to work out the first-order conditions for the special case of q linear constraints, the null being expressed as $\mathbf{R}\theta_0 = \mathbf{r}$, which yields a very similar expression.

Expand $\partial L_n(\tilde{\theta})/\partial \theta$ about the true θ_0 and do likewise for $r(\tilde{\theta})$. We obtain

$$\frac{\partial L_n(\theta_0)}{\partial \theta} + \frac{\partial^2 L_n(\bar{\theta})}{\partial \theta \partial \theta'}(\tilde{\theta} - \theta_0) - \mathbf{R}(\tilde{\theta})'\tilde{\lambda} = \mathbf{0} \tag{18.8}$$

$$-\mathbf{R}(\dot{\theta})(\tilde{\theta} - \theta_0) = \mathbf{0} \tag{18.9}$$

where $\bar{\theta}$ and $\dot{\theta}$ both lie between $\tilde{\theta}$ and θ_0, and we have used $\mathbf{r}(\theta_0) = \mathbf{0}$ in the second line.

Multiply the equations by $-\sqrt{n}$ and rearrange to put things in matrix form. This yields

$$\begin{bmatrix} -\frac{\partial^2 L_n(\bar{\theta})}{\partial \theta \partial \theta'} & \mathbf{R}(\tilde{\theta})' \\ \mathbf{R}(\dot{\theta}) & \mathbf{0} \end{bmatrix} \begin{bmatrix} \sqrt{n}(\tilde{\theta} - \theta_0) \\ \sqrt{n}\tilde{\lambda} \end{bmatrix} = \begin{bmatrix} \sqrt{n}\frac{\partial L_n(\theta_0)}{\partial \theta} \\ \mathbf{0} \end{bmatrix}.$$

Asymptotically, this equation system is

$$\begin{bmatrix} \mathcal{J} & \mathbf{R}' \\ \mathbf{R} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \sqrt{n}(\tilde{\theta} - \theta_0) \\ \sqrt{n}\tilde{\lambda} \end{bmatrix} \stackrel{a}{=} \begin{bmatrix} Z \\ \mathbf{0} \end{bmatrix}.$$

where $Z \sim \mathcal{N}(\mathbf{0}, \mathcal{J})$ and $\mathbf{R} = \mathbf{R}(\theta_0)$. Inverting the leading matrix and solving yields

$$\sqrt{n}(\tilde{\theta} - \theta_0) \stackrel{a}{=} \left(\mathcal{J}^{-1} - \mathcal{J}^{-1}\mathbf{R}'(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1}\right) \cdot Z. \tag{18.10}$$

To obtain this result, we use a partitioned inverse formula.[6] Also,

$$\sqrt{n}\,\tilde{\lambda} = \left(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}'\right)^{-1}\mathbf{R}\mathcal{J}^{-1} \cdot Z.$$

Let

$$\mathbf{M} = \mathbf{I} - \mathcal{J}^{-1/2}\mathbf{R}'(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1/2}.$$

[6] The formula applies to the inverse of

$$\begin{bmatrix} \mathbf{B} & \mathbf{A}' \\ \mathbf{A} & \mathbf{0} \end{bmatrix}$$

with $\mathbf{B}$ symmetric positive definite. The inverse is

$$\begin{bmatrix} \mathbf{B}^{-1} - \mathbf{B}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}')^{-1}\mathbf{A}\mathbf{B}^{-1} & \mathbf{B}^{-1}\mathbf{A}'(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}')^{-1} \\ (\mathbf{A}\mathbf{B}^{-1}\mathbf{A}')^{-1}\mathbf{A}\mathbf{B}^{-1} & -(\mathbf{A}\mathbf{B}^{-1}\mathbf{A}')^{-1} \end{bmatrix}$$

We apply the formula letting $\mathcal{J} = \mathbf{B}$ and $\mathbf{R} = \mathbf{A}$.

Then the asymptotic distribution of the constrained estimator is

$$\sqrt{n}(\tilde{\theta} - \theta_0) \stackrel{a}{=} \mathcal{J}^{-1/2}\mathbf{M}\mathcal{J}^{-1/2} \cdot \mathcal{N}(\mathbf{0}, \mathcal{J}) \tag{18.11}$$

$$\stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \mathcal{J}^{-1/2}\mathbf{M}\mathcal{J}^{-1/2}). \tag{18.12}$$

We also have the asymptotic distribution of the multipliers,

$$\sqrt{n}\,\tilde{\lambda} \stackrel{a}{=} (\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1} \cdot \mathcal{N}(\mathbf{0}, \mathcal{J}) \tag{18.13}$$

$$\stackrel{d}{\to} \mathcal{N}(\mathbf{0}, (\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}). \tag{18.14}$$

With these important results in hand, we can now formulate the LM test. The LM test statistic, as expressed in terms of the q multipliers $\tilde{\lambda}$ themselves, is

$$LM_1 = \sqrt{n}\,\tilde{\lambda}' \left(\tilde{\mathbf{R}}\tilde{\mathcal{J}}^{-1}\tilde{\mathbf{R}}'\right) \sqrt{n}\,\tilde{\lambda}. \tag{18.15}$$

This statistic is distributed in the limit as χ^2_q under the null. Since $\partial L_n(\tilde{\theta})/\partial\theta = \mathbf{R}(\tilde{\theta})'\tilde{\lambda}$ by the first-order conditions for the constrained model, this test statistic is numerically identical to

$$LM_2 = \sqrt{n}\frac{\partial L_n(\tilde{\theta})}{\partial\theta}' \tilde{\mathcal{J}}^{-1}\sqrt{n}\frac{\partial L_n(\tilde{\theta})}{\partial\theta}. \tag{18.16}$$

In the latter expression, the role of the constraints is somewhat hidden, but their presence is implicit in $\tilde{\theta}$. In any case, although the vectors in the wings of the quadratic form LM_2 are $k \times 1$, the statistic is distributed as χ^2_q not χ^2_k.

What are the implications of this result for tests of linear hypotheses, that is, for $H_0 : \mathbf{R}\theta_0 - \mathbf{r} = \mathbf{0}$? We would specify the Lagrangian as

$$L_n(\theta) - (\mathbf{R}\theta - \mathbf{r})'\lambda$$

and proceed very much as above to find $\tilde{\theta}$, the constrained estimator. The implied test statistic takes exactly the same form as that of equation (18.16) above. For instance, a commonly-specified linear hypothesis posits that some of the θ parameters are zero, and in this case $\mathbf{R} = [\mathbf{0}_1, \mathbf{I}_2]$, and $\mathbf{r} = \mathbf{0}_2$ when the null hypothesis is $H_0 : \theta_2 = \mathbf{0}_2$ (the subscripts indicate that $\mathbf{r}$ and the components of $\mathbf{R}$ have dimensions that accord with θ_1 or θ_2 as appropriate). Obviously, since θ_1 is not constrained, $\partial L_n/\partial\theta_1 = \mathbf{0}_1$ whereas $\partial L_n/\partial\theta_2 \neq \mathbf{0}_2$. With zeroes appearing in the score vector, the only component of $\tilde{\mathcal{J}}^{-1}$ that is relevant to the test statistic (18.16) is the block of terms corresponding to θ_2 in the lower right of the matrix. As we will see in the next chapter, to obtain useful analytic results we will generally need to isolate this block of terms using partitioned inversion.

If the null hypothesis is true, then the constrained estimator $\tilde{\theta}$ is more *efficient* than the unconstrained $\hat{\theta}$ estimator, as we can see by comparing the limiting variance matrices of $\sqrt{n}(\hat{\theta} - \theta_0)$ and $\sqrt{n}(\tilde{\theta} - \theta_0)$. The two variances are, respectively,

$$\hat{\mathbf{V}} = \mathcal{J}^{-1} \tag{18.17}$$

$$\tilde{\mathbf{V}} = \mathcal{J}^{-1/2}\mathbf{M}\mathcal{J}^{-1/2} = \mathcal{J}^{-1} - \mathcal{J}^{-1}\mathbf{R}'(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1} \tag{18.18}$$

and the difference $\hat{\mathbf{V}} - \tilde{\mathbf{V}}$ is a positive definite matrix.

We'll need one last result from this section to set up our discussion of likelihood ratio tests. From equation (18.5) we have

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \stackrel{a}{=} \mathcal{J}^{-1} \cdot \sqrt{n}\frac{\partial L_n(\theta_0)}{\partial \theta},$$

and from equation (18.10),

$$\sqrt{n}(\tilde{\theta} - \theta_0) \stackrel{a}{=} \left(\mathcal{J}^{-1} - \mathcal{J}^{-1}\mathbf{R}'(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1}\right) \cdot \sqrt{n}\frac{\partial L_n(\theta_0)}{\partial \theta}.$$

Subtracting, we obtain

$$\sqrt{n}(\hat{\theta} - \tilde{\theta}) \stackrel{a}{=} \mathcal{J}^{-1}\mathbf{R}'(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1} \cdot \sqrt{n}\frac{\partial L_n(\theta_0)}{\partial \theta}.$$

Note that because $\sqrt{n}\partial L_n(\theta_0)/\partial\theta \stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \mathcal{J})$,

$$\sqrt{n}(\hat{\theta} - \tilde{\theta}) \stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{W}),$$

with $\mathbf{W} = \mathcal{J}^{-1}\mathbf{R}'(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1}$.

18.8 The Distribution of the Likelihood Ratio

Suppose that we have a null hypothesis regarding the full parameter vector $H_0 : \theta = \theta_0$ and want to test this against the alternative $H_A : \theta \neq \theta_0$. The first step is to expand $L_n(\theta_0)$ to second order around the unconstrained ML estimator $\hat{\theta}$,

$$L_n(\theta_0) = L_n(\hat{\theta}) + \frac{\partial L_n(\hat{\theta})}{\partial \theta}'(\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})'\frac{\partial^2 L_n(\bar{\theta})}{\partial\theta\partial\theta'}(\theta_0 - \hat{\theta}).$$

Noting that $\partial L_n(\hat{\theta})/\partial\theta = \mathbf{0}$, we now subtract $L_n(\hat{\theta})$ from both sides and then multiply by $-2n$, so that

$$\begin{aligned} 2n\left(L_n(\hat{\theta}) - L_n(\theta_0)\right) &= \sqrt{n}(\hat{\theta} - \theta_0)'\left[-\frac{\partial^2 L_n(\bar{\theta})}{\partial\theta\partial\theta'}\right]\sqrt{n}(\hat{\theta} - \theta_0) \\ &\stackrel{a}{=} \sqrt{n}(\hat{\theta} - \theta_0)'\mathcal{J}\sqrt{n}(\hat{\theta} - \theta_0). \end{aligned}$$

Since $\mathcal{J} = \mathcal{J}^{1/2} \cdot \mathcal{J}^{1/2}$ and $\mathcal{J}^{1/2}$ is symmetric, the quadratic form can be represented in the form of an inner product,

$$\mathcal{N}(\mathbf{0}, \mathcal{J}^{-1})'\mathcal{J}^{1/2} \cdot \mathcal{J}^{1/2}\mathcal{N}(\mathbf{0}, \mathcal{J}^{-1}).$$

Given that $\mathcal{J}^{1/2}\mathcal{N}(\mathbf{0}, \mathcal{J}^{-1}) \stackrel{a}{=} \mathcal{N}(\mathbf{0}, \mathbf{I})$, the result is distributed in the limit as χ^2_k. Hence, the test statistic

$$LR = 2n\left(L_n(\hat{\theta}) - L_n(\theta_0)\right) \stackrel{d}{\to} \chi^2_k.$$

Most computer programs report the maximized value of the log-likelihood $L(\hat{\theta})$ rather than the normalized log-likelihood $L_n(\hat{\theta})$, so that the usual form of the LR statistic is

$$LR = 2\left(L(\hat{\theta}) - L(\theta_0)\right) \xrightarrow{d} \chi^2_k.$$

This is a very important result; it appears in a number of forms throughout econometrics.

With a little extra work, we can generalize the result to accommodate null hypotheses expressed in terms of linear or nonlinear constraints, yielding $\tilde{\theta}$ as the constrained estimator. The key is to use the limiting variance of the difference $\sqrt{n}(\hat{\theta} - \tilde{\theta})$ which we derived at the end of the preceding section.

The analysis of this case again begins with a second-order expansion of $L_n(\tilde{\theta})$ around the ML estimator $\hat{\theta}$, and, proceeding as above, we obtain

$$\begin{aligned} 2n\left(L_n(\hat{\theta}) - L_n(\tilde{\theta})\right) &\stackrel{a}{=} \sqrt{n}(\hat{\theta} - \tilde{\theta})'\mathcal{J}\sqrt{n}(\hat{\theta} - \tilde{\theta}) \\ &\stackrel{a}{=} \mathcal{N}(\mathbf{0}, \mathbf{W})'\mathcal{J}\mathcal{N}(\mathbf{0}, \mathbf{W}) \end{aligned}$$

with $\mathbf{W} = \mathcal{J}^{-1}\mathbf{R}'(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1}$ as we showed above. We factor $\mathcal{J} = \mathcal{J}^{1/2} \cdot \mathcal{J}^{1/2}$ and examine the variance of $\mathcal{J}^{1/2} \cdot \mathcal{N}(\mathbf{0}, \mathbf{W})$. This variance matrix is

$$\tilde{\mathbf{M}} = \mathcal{J}^{-1/2}\mathbf{R}'(\mathbf{R}\mathcal{J}^{-1}\mathbf{R}')^{-1}\mathbf{R}\mathcal{J}^{-1/2},$$

which is a symmetric idempotent matrix of rank q, with q being equal to the number of constraints. Hence,

$$\begin{aligned} \mathcal{N}(\mathbf{0}, \mathbf{W})'\mathcal{J}^{1/2} \cdot \mathcal{J}^{1/2}\mathcal{N}(\mathbf{0}, \mathbf{W}) &\stackrel{a}{=} \mathcal{N}(\mathbf{0}, \tilde{\mathbf{M}})'\mathcal{N}(\mathbf{0}, \tilde{\mathbf{M}}) \\ &\stackrel{a}{=} \mathcal{N}(\mathbf{0}, \mathbf{I})'\,\tilde{\mathbf{M}}\,\mathcal{N}(\mathbf{0}, \mathbf{I}) \xrightarrow{d} \chi^2_q \end{aligned}$$

using the familiar result from Chapter 2 on quadratic forms in standard normal random vectors and idempotent matrices.

To sum up, the general Likelihood Ratio test statistic is

$$LR = 2n\left(L_n(\hat{\theta}) - L_n(\tilde{\theta})\right) = 2\left(L(\hat{\theta}) - L(\tilde{\theta})\right) \xrightarrow{d} \chi^2_q,$$

so that the LR statistic has the same limiting distribution (under the null) as the Wald and Lagrange Multiplier statistics. For formal proofs of what we have presented rather informally, see Newey and McFadden (1994), who develop the proofs for generalized method-of-moments estimators, with maximum likelihood estimators being a special case.

18.9 Implementing LM Tests by Artificial Regressions

In this section, we will show that LM tests can be represented, in general, by the nR^2 of an auxiliary regression. As above, let $\tilde{\theta}$ denote the constrained estimator.

Let $\mathbf{W}$ be the $n \times k$ matrix,

$$\mathbf{W} = \begin{bmatrix} \dfrac{\partial \ln f(y_1 \mid \tilde{\theta})}{\partial \theta_1} & \cdots & \dfrac{\partial \ln f(y_1 \mid \tilde{\theta})}{\partial \theta_k} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial \ln f(y_n \mid \tilde{\theta})}{\partial \theta_1} & \cdots & \dfrac{\partial \ln f(y_n \mid \tilde{\theta})}{\partial \theta_k} \end{bmatrix}.$$

Then the (normalized) score vector can be written as

$$n^{-1}\mathbf{W}'\iota \;=\; \frac{\partial L_n(\tilde{\theta})}{\partial \theta} = \left[\begin{array}{c} \frac{\partial L_n}{\partial \theta_1} \\ \frac{\partial L_n}{\partial \theta_2} \\ \vdots \\ \frac{\partial L_n}{\partial \theta_k} \end{array}\right],$$

where ι is an n-vector of ones. The $k \times k$ matrix $n^{-1}\mathbf{W}'\mathbf{W}$ is a consistent estimator of $\mathcal{J}$, since

$$n^{-1}\mathbf{W}'\mathbf{W} = n^{-1}\sum_{i=1}^{n}\left(\frac{\partial \ln f(y_i \mid \tilde{\theta})}{\partial \theta}\right)\left(\frac{\partial \ln f(y_i \mid \tilde{\theta})}{\partial \theta}\right)'.$$

Making the appropriate substitutions, we see that the LM statistic,

$$LM \;=\; \sqrt{n}\frac{\partial L_n}{\partial \theta}' \;\; \tilde{\mathcal{J}}^{-1} \;\; \sqrt{n}\frac{\partial L_n}{\partial \theta},$$

can be estimated by

$$\iota'\mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\iota$$

since the various n's cancel. Now consider an auxiliary regression of the form

$$\iota = \mathbf{W}\rho + \text{residuals},$$

to which we apply ordinary least squares, yielding

$$\hat{\rho} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\iota.$$

The uncentered R^2 from this regression is

$$R^2 = \frac{\iota'\mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\iota}{\iota'\iota} = \frac{\iota'\mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\iota}{n}.$$

From this we find

$$nR^2 \;=\; \iota'\mathbf{W}\left(\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\iota \;=\; LM.$$

This is an intriguing result. However, as Davidson and MacKinnon (1993) caution, this artificial regression does not seem to perform especially well in small or even moderately-sized samples. The difficulty resides in its outer-product estimate of $\mathcal{J}$, which tends to be quite noisy.

18.10 One-Step Efficient Estimation

We close our review of maximum likelihood theory with an important result having to do with the relationship between consistent estimators of θ and the ML estimator. What we will show, following Davidson and MacKinnon (1993, pp. 472–474), is that if one begins with a estimate of θ derived from a consistent estimator, and adjusts it by means of an

artificial regression much like that described above, the procedure delivers an estimator that has all of the desirable asymptotic properties of the ML estimator.

Let $\tilde{\theta}$ be the consistent estimator—note that we have changed notation—and let $\tilde{\mathbf{W}}$ be the $n \times k$ matrix of contributions to the score, but evaluated at $\tilde{\theta}$ rather than at the ML estimate $\hat{\theta}$. In other words, a typical element of $\tilde{\mathbf{W}}$ is $\partial \ln f(y_i \mid \tilde{\theta})/\partial\theta_j$ for row i and column j. We will show that the new estimator $\dot{\theta}$, defined as $\dot{\theta} = \tilde{\theta} + \tilde{c}$ where $\tilde{c}$ is taken from the artificial regression

$$\iota = \tilde{\mathbf{W}}c + \text{residuals},$$

shares the asymptotic properties of the MLE. Note that $\tilde{c} = (\tilde{\mathbf{W}}'\tilde{\mathbf{W}})^{-1}\tilde{\mathbf{W}}'\iota$.

To prove this, expand $\partial L_n(\tilde{\theta})/\partial\theta$ around θ_0 and multiply by $\sqrt{n}$, so that

$$\sqrt{n}\frac{\partial L(\tilde{\theta})}{\partial\theta} = \sqrt{n}\frac{\partial L(\theta_0)}{\partial\theta} + \frac{\partial^2 L_n(\bar{\theta})}{\partial\theta\partial\theta'}\sqrt{n}(\tilde{\theta} - \theta_0).$$

Similarly expand the score $\partial L_n(\hat{\theta})/\partial\theta$ around θ_0,

$$\mathbf{0} = \sqrt{n}\frac{\partial L_n(\hat{\theta})}{\partial\theta} = \sqrt{n}\frac{\partial L_n(\theta_0)}{\partial\theta} + \frac{\partial^2 L_n(\ddot{\theta})}{\partial\theta\partial\theta'}\sqrt{n}(\hat{\theta} - \theta_0).$$

Upon subtracting and rearranging, we see that we have

$$\sqrt{n}\frac{\partial L_n(\tilde{\theta})}{\partial\theta} \stackrel{a}{=} \mathcal{J}\sqrt{n}(\hat{\theta} - \tilde{\theta})$$

or

$$\sqrt{n} \cdot n^{-1}\tilde{\mathbf{W}}'\iota \stackrel{a}{=} (n^{-1}\tilde{\mathbf{W}}'\tilde{\mathbf{W}})\sqrt{n}(\hat{\theta} - \tilde{\theta})$$

or

$$\sqrt{n}\,\tilde{c} \stackrel{a}{=} \sqrt{n}(\hat{\theta} - \tilde{\theta}).$$

Then, recalling the definition of the adjusted estimator $\dot{\theta} = \tilde{\theta} + \tilde{c}$, we have

$$\sqrt{n}(\dot{\theta} - \theta_0) \stackrel{a}{=} \sqrt{n}(\tilde{\theta} + \hat{\theta} - \tilde{\theta} - \theta_0) = \sqrt{n}(\hat{\theta} - \theta_0).$$

18.11 A Note on Optimization

The Newton–Raphson approach we described in Chapter 17 applies with little modification to maximization of log-likelihood functions. As you will recall, following the notation of that chapter and denoting the "gradient" of the likelihood function at a trial value θ_1 by $\mathbf{g}_1 = \partial L(\theta_1)/\partial\theta$ and the "Hessian" by $\mathbf{H}_1 = \partial^2 L(\theta_1)/\partial\theta\partial\theta'$, the updating algorithm can be summarized as

$$\theta_2 = \theta_1 - \mathbf{H}_1^{-1}\mathbf{g}_1. \tag{18.19}$$

Obviously we cannot calculate this solution if at θ_1 the Hessian matrix of second derivatives is singular, which it make well be if the θ_1 vector is far from the optimizing value; indeed, as we discussed, the matrix $\mathbf{H}_1$ must be both invertible and negative definite.

However, we know that at the *true* value of θ, which in the current chapter we have denoted by θ_0, we have

$$-\operatorname{E}\frac{\partial^2 L(\theta_0)}{\partial\theta\partial\theta'} = \operatorname{E}\left(\frac{\partial L(\theta_0)}{\partial\theta}\right)\left(\frac{\partial L(\theta_0)}{\partial\theta}\right)'.$$

Assuming that the data series is inid, we can write the expected outer product as

$$\operatorname{E}\left(\frac{\partial L(\theta_0)}{\partial\theta}\right)\left(\frac{\partial L(\theta_0)}{\partial\theta}\right)' = \sum_{i=1}^{n}\operatorname{E}\left(\frac{\partial \ln f_i(\theta_0)}{\partial\theta}\right)\left(\frac{\partial \ln f_i(\theta_0)}{\partial\theta}\right)'.$$

The $k \times k$ matrix on the right is the sum of expected outer products. If we approximate it by dropping the expectation operator and evaluating the derivatives not at the true θ_0 but at any arbitrary θ_1,

$$\sum_{i=1}^{n}\left(\frac{\partial \ln f_i(\theta_1)}{\partial\theta}\right)\left(\frac{\partial \ln f_i(\theta_1)}{\partial\theta}\right)',$$

we have an matrix that, like its theoretical counterpart, is positive definite.

Armed with this admittedly loose justification, we proceed to *approximate* the $-\mathbf{H}_1$ matrix by using an alternative matrix expression that is guaranteed to be positive definite,

$$-\mathbf{H}_1 \approx \sum_{i=1}^{n}\left(\frac{\partial \ln f_i(\theta_1)}{\partial\theta}\right)\left(\frac{\partial \ln f_i(\theta_1)}{\partial\theta}\right)',$$

and of course we may write the gradient in terms of a sum as well,

$$\mathbf{g}_1 = \sum_{i=1}^{n}\frac{\partial \ln f_i(\theta_1)}{\partial\theta}.$$

The updating algorithm thus becomes

$$\theta_2 = \theta_1 + \left(\sum_{i=1}^{n}\left(\frac{\partial \ln f_i(\theta_1)}{\partial\theta}\right)\left(\frac{\partial \ln f_i(\theta_1)}{\partial\theta}\right)'\right)^{-1}\cdot\sum_{i=1}^{n}\frac{\partial \ln f_i(\theta_1)}{\partial\theta}.$$

These are the basic ingredients in applying a modified version of the Newton–Raphson method to maximum likelihood problems. It is often termed the "Berndt–Hall–Hall–Hausman" method in honor of the four economists who introduced it into econometrics.

Chapter 19

LR, LM, and Wald Tests for the Normal Linear Model

Students: You can think of this chapter as being mainly a set of exercises showing how to apply the concepts of the previous chapter to the linear model with normally distributed disturbances.

This chapter develops the Likelihood Ratio, Wald and Lagrange Multiplier tests in the context of the normal linear model $Y_i = \mathbf{X}_i'\beta + \epsilon_i$, with the disturbance vector ϵ distributed as $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ and independent of the $\mathbf{X}$ covariates. Of course we don't really need these large-sample tests in this context, as we already have $\mathcal{F}$ and t tests available for hypotheses about β and χ^2 tests available for hypotheses about σ^2. These tests require no asymptotic approximations. The motivation for what follows is simply to explore the LR, Wald and LM tests in an otherwise familiar context, thereby preparing ourselves to apply them where they are really needed: in the general linear model with $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{V})$, and in nonlinear models of all sorts.

Regression models generally include covariates, and we will therefore examine the *conditional* density functions in what follows. The key results we need from our previous chapter are that

$$\sqrt{n}\frac{\partial L_n(\theta_0)}{\partial \theta} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}),$$

where $L_n(\theta_0)$ should be understood to be the normalized log-likelihood for the conditional component of the full likelihood function; and, from this,

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}^{-1}).$$

The first limiting distribution result leads to the Lagrange multiplier test statistic

$$LM = \sqrt{n}\frac{\partial L_n(\tilde{\theta})'}{\partial \theta}\tilde{\mathcal{J}}^{-1}\sqrt{n}\frac{\partial L_n(\tilde{\theta})}{\partial \theta}$$

where $\tilde{\mathcal{J}}$ is the $\mathcal{J}$ matrix evaluated at the restricted estimator $\tilde{\theta}$ that satisfies the null hypothesis. The second of the expressions above motivates the Wald test statistic

$$W = \sqrt{n}(\hat{\theta}_n - \theta_0)'\hat{\mathcal{J}}\sqrt{n}(\hat{\theta}_n - \theta_0) = (\hat{\theta}_n - \theta_0)'n\hat{\mathcal{J}}(\hat{\theta}_n - \theta_0)$$

where θ_0 is specified by the null hypothesis and $\hat{\mathcal{J}}$ is the $\mathcal{J}$ matrix evaluated at the unrestricted maximum likelihood estimator $\hat{\theta}$. (Although the link to the asymptotic theory is clearer in the first expression for the test statistic, the second expression, in which the information matrix appears in the center of the quadratic form, is more often seen in the literature.) Let us now put these concepts to work on the simple linear model.

19.1 The Normal Log-Likelihood

We explored the likelihood function for this model in our previous chapter. In the literature, the results are usually obtained by applying calculus to vector–matrix expressions. Let's retrace our steps and derive the results all over again using the more common approach. The full-sample normalized log-likelihood function is given by

$$L_n = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{1}{n}\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta).$$

Differentiating this, we obtain the $k+1$ elements of the score vector,

$$\begin{aligned}\frac{\partial L_n}{\partial \beta} &= \frac{1}{\sigma^2}\frac{1}{n}\mathbf{X}'(\mathbf{Y}-\mathbf{X}\beta)\\ \frac{\partial L_n}{\partial \sigma^2} &= -\frac{1}{2\sigma^2} + \frac{1}{n}\frac{(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)}{2\sigma^4}.\end{aligned}$$

Using the first-order conditions, we can find the unconstrained maximum-likelihood estimators by proceeding in a recursive fashion. Set both score elements to zero, solve the first equation for $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and substitute the result into the second equation to find $\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/n$, with $\mathbf{e}$ being the residual $\mathbf{Y} - \mathbf{X}\hat{\beta}$. This recursive feature simplifies a number of derivations.

Another point to note from the first-order conditions is that if we have a null hypothesis on the variance, $H_0 : \sigma^2 = \sigma_0^2$, and impose this null to derive the constrained maximum-likelihood estimators $\tilde{\beta}$ and $\tilde{\sigma}^2 = \sigma_0^2$, we find that $\tilde{\beta} = \hat{\beta}$, that is, the constrained estimator of β is identical to the unconstrained estimator. In this case, imposing a constraint on the variance makes no difference to the estimator of the slope coefficients. For null hypotheses about β, however, such as $\mathbf{R}\beta_0 = \mathbf{r}$, the imposition of the constraint generates different implications. The constrained maximum-likelihood estimator $\tilde{\beta}$ will obviously differ from the unconstrained estimator (in fact, the form of the constrained estimator $\tilde{\beta}$ is identical to that derived in Chapter 10 without the normality assumption), producing residuals $\tilde{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\tilde{\beta}$, and the constrained estimator of the variance is $\tilde{\sigma}^2 = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}/n$, which also differs from the unconstrained version.

The second partial derivatives are expressed in vector–matrix notation as

$$\begin{aligned}\frac{\partial^2 L_n}{\partial\beta\,\partial\beta'} &= -\frac{1}{\sigma^2}\frac{1}{n}\mathbf{X}'\mathbf{X}\\ \frac{\partial^2 L_n}{\partial(\sigma^2)^2} &= \frac{1}{2\sigma^4} - \frac{1}{n}\frac{(\mathbf{Y}-\mathbf{X}\beta)'(\mathbf{Y}-\mathbf{X}\beta)}{\sigma^6}\\ \frac{\partial^2 L_n}{\partial\beta\,\partial\sigma^2} &= -\frac{1}{n}\frac{\mathbf{X}'(\mathbf{Y}-\mathbf{X}\beta)}{\sigma^4}.\end{aligned}$$

Multiplying each by -1 and taking expectations, we obtain

$$\mathcal{J} = \lim_{n\to\infty} \begin{bmatrix} \frac{1}{\sigma^2}\frac{1}{n}\mathrm{E}\,\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma^4} \end{bmatrix}$$

just as before.

19.2 Likelihood-Ratio Tests

The normalized log-likelihood function, when evaluated at the unconstrained estimates $(\hat{\beta}, \hat{\sigma}^2)$, is

$$L_n(\hat{\beta}, \hat{\sigma}^2) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) - \frac{1}{2}\frac{n}{\mathbf{e}'\mathbf{e}}\frac{1}{n}\mathbf{e}'\mathbf{e} = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\left(\frac{\mathbf{e}'\mathbf{e}}{n}\right) - \frac{1}{2}.$$

The nature of the constrained log-likelihood function depends, of course, on the nature of the constraint. Consider a null hypothesis $H_0 : \sigma^2 = \sigma_0^2$. Because imposing this constraint does not affect the likelihood-maximizing estimator of β, it yields the constrained likelihood

$$L_n(\tilde{\beta}, \sigma_0^2) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\left(\sigma_0^2\right) - \frac{1}{2}\frac{1}{\sigma_0^2}\frac{1}{n}\mathbf{e}'\mathbf{e},$$

in which the unconstrained sum of squared residuals $\mathbf{e}'\mathbf{e}$ appears owing to the fact that $\tilde{\beta} = \hat{\beta}$ in this case. For this hypothesis, the likelihood-ratio test statistic is $2n\left(L_n(\hat{\beta}, \hat{\sigma}^2) - L_n(\tilde{\beta}, \sigma_0^2)\right) \xrightarrow{d} \chi_1^2$ under the null.

By contrast, if we estimate the model under the constraint $\mathbf{R}\beta_0 = \mathbf{r}$, the maximized log-likelihood value is

$$L_n(\tilde{\beta}, \tilde{\sigma}^2) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\left(\tilde{\sigma}^2\right) - \frac{1}{2}\frac{1}{\tilde{\sigma}^2}\frac{1}{n}\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\left(\tilde{\sigma}^2\right) - \frac{1}{2}.$$

Here $2n\left(L_n(\hat{\beta}, \hat{\sigma}^2) - L_n(\tilde{\beta}, \tilde{\sigma}^2)\right) \xrightarrow{d} \chi_q^2$ with q being the number of constraints (i.e., the number of rows of the $\mathbf{R}$ matrix.)

19.3 Tests of σ^2

We now show how to construct Wald and Lagrange Multiplier tests, illustrating the principles of test construction for the null hypothesis $H_0 : \sigma^2 = \sigma_0^2$. To understand these tests, we need to think about two matrices: the $\mathbf{R}$ matrix by which the null hypothesis is stated, $\mathbf{R}\theta_0 = \mathbf{r}$ with $\theta_0 = (\beta_0, \sigma_0^2)'$, and the $\mathcal{J}$ matrix, which in the case of normal distributions is block-diagonal. As we'll see, because $\mathcal{J}$ is block-diagonal, the component relevant to hypotheses regarding σ^2 is simply $\mathcal{J}_{\sigma^2} = 1/(2\sigma^4)$.

Wald test

Under the null hypothesis $\mathbf{R}\theta_0 = \mathbf{r}$, the difference $\sqrt{n}(\mathbf{R}\hat{\theta} - \mathbf{r}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{R}\mathcal{J}^{-1}\mathbf{R}')$. In this instance, the $1 \times k+1$ matrix $\mathbf{R} = \begin{bmatrix}\mathbf{0} & 1\end{bmatrix}$ and $\mathbf{r} = \sigma_0^2$. Hence,

$$\mathbf{R}\hat{\mathcal{J}}^{-1}\mathbf{R}' = \begin{bmatrix}\mathbf{0} & 1\end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_0^2}\frac{1}{n}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_0^4} \end{bmatrix}^{-1} \begin{bmatrix}\mathbf{0} \\ 1\end{bmatrix} = 2\sigma_0^4.$$

The unconstrained estimate of this scalar, inverted, goes in the middle of the quadratic form defining the Wald test statistic. Thus, for a Wald test with null hypothesis $\sigma^2 = \sigma_0^2$, we have

$$W = \sqrt{n}(\hat{\sigma}^2 - \sigma_0^2)\frac{1}{2\hat{\sigma}^4}\sqrt{n}(\hat{\sigma}^2 - \sigma_0^2) = \frac{n}{2\hat{\sigma}^4}(\hat{\sigma}^2 - \sigma_0^2)^2.$$

Lagrange Multiplier test

To construct a Lagrange Multiplier test of the same hypothesis, we recall that under the null, $\sqrt{n}\partial L_n(\tilde{\theta})/\partial\theta \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J})$, where in the case at hand $\theta = (\beta, \sigma^2)'$. The null hypothesis does not address β and so when we differentiate the constrained Lagrangian to solve the first-order conditions for $\tilde{\beta}$ and $\tilde{\sigma}^2$, we can set $\partial L_n(\tilde{\beta}, \tilde{\sigma}^2))/\partial\beta = \mathbf{0}$. Hence,

$$\sqrt{n}\begin{bmatrix}\partial L_n(\tilde{\beta}, \tilde{\sigma}^2)/\partial\beta & \partial L_n(\tilde{\beta}, \tilde{\sigma}^2)/\partial\sigma^2\end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_0^2}\frac{1}{n}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_0^4} \end{bmatrix}^{-1} \sqrt{n}\begin{bmatrix}\partial L_n(\tilde{\beta}, \tilde{\sigma}^2)/\partial\beta \\ \partial L_n(\tilde{\beta}, \tilde{\sigma}^2)/\partial\sigma^2\end{bmatrix}$$

reduces to

$$\sqrt{n}\begin{bmatrix}\mathbf{0} & \partial L_n(\tilde{\beta}, \tilde{\sigma}^2)/\partial\sigma^2\end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_0^2}\frac{1}{n}\mathbf{X}'\mathbf{X} & \mathbf{0} \\ \mathbf{0}' & \frac{1}{2\sigma_0^4} \end{bmatrix}^{-1} \sqrt{n}\begin{bmatrix}\mathbf{0} \\ \partial L_n(\tilde{\beta}, \tilde{\sigma}^2)/\partial\sigma^2\end{bmatrix},$$

which equals

$$\sqrt{n}\frac{\partial L_n(\tilde{\beta}, \tilde{\sigma}^2)}{\partial\sigma^2} \cdot 2\sigma_0^4 \cdot \sqrt{n}\frac{\partial L_n(\tilde{\beta}, \tilde{\sigma}^2)}{\partial\sigma^2}.$$

The relevant element of the score is

$$\frac{\partial L_n(\tilde{\beta}, \tilde{\sigma}^2)}{\partial\sigma^2} = -\frac{1}{2\sigma_0^2} + \frac{1}{n}\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{2\sigma_0^4} = -\frac{1}{2\sigma_0^2} + \frac{\hat{\sigma}^2}{2\sigma_0^4}.$$

The reason that the unconstrained variance estimate $\hat{\sigma}^2$ appears here is that the constrained residuals $\tilde{\mathbf{e}}$ are identical to the usual unconstrained least-squares residuals $\mathbf{e}$—as mentioned earlier, this is because imposing the null hypothesis $\sigma^2 = \sigma_0^2$ does not affect the maximum likelihood estimate of β. After a little algebraic manipulation, we obtain a Lagrange Multipler test statistic that is nearly identical to the Wald statistic,

$$LM = \frac{n}{2\sigma_0^4}(\hat{\sigma}^2 - \sigma_0^2)^2.$$

19.4 Tests on the Full β Vector

Essentially the same principles of test construction apply to hypothesis tests of the β slope parameters. To begin, consider the null hypothesis $H_0 : \beta = \beta_0$. Since this hypothesis refers only to β, the $\mathbf{R}$ matrix of the constraint $\mathbf{R}\theta_0 = \mathbf{r}$ will have zeroes in its last column, which corresponds to the σ^2 parameter. Since $\mathcal{J}$ is block-diagonal, the implied $(\mathbf{R}\hat{\mathcal{J}}^{-1}\mathbf{R}')^{-1}$ term for the Wald test, which appears in the center of the quadratic form (19.1), is estimated by $\hat{\mathcal{J}}_\beta = n^{-1}\mathbf{X}'\mathbf{X}/\hat{\sigma}^2$, with $\hat{\sigma}^2 = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta})/n$.

Similarly, for the Lagrange Multiplier test the element of the score that pertains to σ^2 is zero, i.e., $\partial L_n(\tilde{\beta}, \tilde{\sigma}^2)/\partial\sigma^2 = 0$, and only the block of terms in $\tilde{\mathcal{J}}$ that corresponds to β will be relevant. Note that when forming the LM test statistic, all terms in $\tilde{\mathcal{J}}$ and the score vector are evaluated using the constrained estimators $\tilde{\beta}$ and $\tilde{\sigma}^2$.

The Wald test

As just mentioned, the Wald test statistic is

$$W = \sqrt{n}(\hat{\beta} - \beta_0)' \left(\frac{1}{n}\frac{\mathbf{X}'\mathbf{X}}{\hat{\sigma}^2}\right) \sqrt{n}(\hat{\beta} - \beta_0)$$

or simply

$$W = (\hat{\beta} - \beta_0)' \left(\frac{\mathbf{X}'\mathbf{X}}{\hat{\sigma}^2}\right) (\hat{\beta} - \beta_0) \xrightarrow{d} \chi^2_k.$$

This expression has appeared more than once in previous chapters.

The Lagrange Multiplier test

For this test we require the score vector $\partial L_n(\tilde{\beta}, \tilde{\sigma}^2)/\partial\beta$. Note that the constrained $\tilde{\beta} = \beta_0$ since the null hypothesis specifies values for each element of the β vector. We have

$$\frac{\partial L_n(\tilde{\beta}, \tilde{\sigma}^2)}{\partial\beta} = \frac{\mathbf{X}'(\mathbf{Y} - \mathbf{X}\beta_0)}{n\tilde{\sigma}^2} = \frac{\mathbf{X}'\tilde{\mathbf{e}}}{n\tilde{\sigma}^2}$$

where $\tilde{\mathbf{e}} = \mathbf{Y} - \mathbf{X}\beta_0$ and $\tilde{\sigma}^2 = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}/n$. Note that both the residual $\tilde{\mathbf{e}}$ and the variance estimator $\tilde{\sigma}^2$ are calculated at the null hypothesis β_0. As a result, the residuals $\tilde{\mathbf{e}}$ *do not* have the usual properties associated with least-squares residuals. In particular, $\mathbf{X}'\tilde{\mathbf{e}} \neq \mathbf{0}$, and the LM test statistic assesses the degree to which $\mathbf{X}$ and $\tilde{\mathbf{e}}$ depart from orthogonality.

The LM statistic is

$$\begin{aligned} \sqrt{n}\frac{\partial L_n}{\partial\beta}' \tilde{\mathcal{J}}_\beta^{-1} \sqrt{n}\frac{\partial L_n}{\partial\beta} &= \frac{\tilde{\mathbf{e}}'\mathbf{X}}{\tilde{\sigma}^2} \left(\frac{\mathbf{X}'\mathbf{X}}{\tilde{\sigma}^2}\right)^{-1} \frac{\mathbf{X}'\tilde{\mathbf{e}}}{\tilde{\sigma}^2} \\ &= \frac{\tilde{\mathbf{e}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{e}}}{\tilde{\sigma}^2} \end{aligned}$$

and this is also distributed as χ^2_k in the limit. Since the denominator of this expression is $\tilde{\sigma}^2 = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}/n$, on re-writing we find

$$LM = n\frac{\tilde{\mathbf{e}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\tilde{\mathbf{e}}}{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}} = n \cdot R^2$$

where the R^2 is taken from a regression of the residuals $\tilde{\mathbf{e}}$ on the $\mathbf{X}$ variables. (Note that the un-centered definition of R^2 is used here.) The R^2 summarizes the extent to which $\tilde{\mathbf{e}}$ and $\mathbf{X}$ depart from orthogonality.

19.5 Testing a Subset of the β Parameters

Now partition the β vector as $\beta = (\beta_1, \beta_2)$ and let the null hypothesis be $\beta_2 = \mathbf{0}$, where β_2 has dimension $k_2 \times 1$. The null hypothesis does not address the values of β_1, which are left unrestricted. This is probably the most common form in which null hypotheses are expressed in econometric work.

Expressed in terms of the constraint $\mathbf{R}\theta_0 = \mathbf{r}$, we have $\mathbf{R} = \begin{bmatrix}\mathbf{0}_1 & \mathbf{I}_2 & \mathbf{0}\end{bmatrix}$ and the Wald and Lagrange Multiplier test statistics now depend only on a sub-block of the terms in $\tilde{\mathcal{J}}$ that correspond to β_2. To obtain analytic results, we will need to use partitioned inversion to isolate this sub-block.

The Wald test

For a Wald test of the hypothesis $\beta_2 = \mathbf{0}_2$, we use

$$\hat{\mathcal{J}}_{\beta_2} = \frac{\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2}{n\hat{\sigma}^2}$$

in the center of the quadratic form (this expression should be familiar from our earlier treatment of the FWL theorem), yielding the test statistic

$$W = \hat{\beta}_2' \frac{\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2}{\hat{\sigma}^2}\hat{\beta}_2,$$

where $\hat{\sigma}^2 = \hat{\mathbf{e}}'\hat{\mathbf{e}}/n$ and $\hat{\mathbf{e}} = \mathbf{Y} - \mathbf{X}_1\hat{\beta}_1 - \mathbf{X}_2\hat{\beta}_2$ just as in an unrestricted OLS estimator.

The Lagrange Multiplier test

Since β_1 is unrestricted, the score in the β_1 direction, that is, $\partial L_n/\partial\beta_1$, will be zero. The only relevant component of the full score vector is

$$\frac{\partial L_n(\tilde{\beta}_1, \tilde{\beta}_2, \tilde{\sigma}^2)}{\partial \beta_2} = \frac{\mathbf{X}_2'(\mathbf{Y} - \mathbf{X}_1\tilde{\beta}_1)}{\tilde{\sigma}^2} = \frac{\mathbf{X}_2'\tilde{\mathbf{e}}}{n\tilde{\sigma}^2}.$$

since the restricted $\tilde{\beta}_2 = \mathbf{0}$. The residual $\tilde{\mathbf{e}}$ and likewise $\tilde{\sigma}^2$ are defined using the restricted estimator $\tilde{\beta}_1$, which in this case is simply the OLS estimator from a regression of $\mathbf{Y}$ on $\mathbf{X}_1$ alone.

Because $\partial L_n/\partial\beta_1 = \mathbf{0}_1$, we require only the portion of $\mathcal{J}^{-1}$ that has to do with β_2. With $\tilde{\mathcal{J}}_\beta = n^{-1}\mathbf{X}'\mathbf{X}/\tilde{\sigma}^2$, partitioned inversion on $\mathbf{X}'\mathbf{X}$ yields

$$\tilde{\mathcal{J}}_{\beta_2}^{-1} = \left(\frac{\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2}{n\tilde{\sigma}^2}\right)^{-1}.$$

Drawing these elements together, we obtain the LM test statistic

$$LM = \frac{\tilde{\mathbf{e}}'\mathbf{X}_2 \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1} \mathbf{X}_2'\tilde{\mathbf{e}}}{\tilde{\sigma}^2} \xrightarrow{d} \chi^2_{k_2}.$$

Although it may not be immediately obvious, this expression can also be reduced to an nR^2 form. To see this, recall that β_1 is estimated by least squares, so that the residuals $\tilde{\mathbf{e}}$ have the property that $\mathbf{X}_1'\tilde{\mathbf{e}} = \mathbf{0}$. The numerator of the LM test is therefore numerically identical to the explained sum of squares from a regression of $\tilde{\mathbf{e}}$ on $(\mathbf{X}_1, \mathbf{X}_2)$, and the denominator is again $\tilde{\mathbf{e}}'\tilde{\mathbf{e}}/n$. Thus, although the regression from which it is taken includes $\mathbf{X}_1$ as well as $\mathbf{X}_2$, the R^2 properly measures the degree to which $\tilde{\mathbf{e}}$ and $\mathbf{X}_2$ depart from being orthogonal.

19.6 Tests of General Linear Hypotheses $\mathbf{R}\theta = \mathbf{r}$

Let a set of q linear restrictions on $\theta = (\beta, \sigma^2)'$ be represented in the form $\mathbf{R}\theta = \mathbf{r}$. Given the unrestricted maximum-likelihood estimator $\hat{\theta}$, we have

$$\sqrt{n}(\mathbf{R}\hat{\theta} - \mathbf{r}) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{R}\mathcal{J}^{-1}\mathbf{R}').$$

This immediately yields the Wald statistic,

$$W = \sqrt{n}(\mathbf{R}\hat{\theta} - \mathbf{r})' \left(\mathbf{R}\hat{\mathcal{J}}^{-1}\mathbf{R}'\right)^{-1} \sqrt{n}(\mathbf{R}\hat{\theta} - \mathbf{r}) \tag{19.1}$$

which is calculated using the unconstrained, maximum-likelihood estimates $\hat{\beta}$ and $\hat{\sigma}^2$. This statistic could be used to test hypotheses that address both the β and σ^2 parameters. The LM test statistic is, in general,

$$\sqrt{n}\frac{\partial L_n(\tilde{\theta})}{\partial \theta}' \tilde{\mathcal{J}}^{-1} \sqrt{n}\frac{\partial L_n(\tilde{\theta})}{\partial \theta}$$

with $\tilde{\theta}$ being the constrained estimates of β and σ^2. In cases in which the constraint $\mathbf{R}\theta = \mathbf{r}$ is easy to impose, the Lagrange Multiplier statistic is readily calculated. If the constraint is difficult to impose, however, as it will be if both σ^2 and β are addressed in the null hypothesis, the Wald test will generally be preferred.

Chapter 20

Binary Dependent Variables

Students: if you have access to this textbook, supplement this chapter by reading Cameron and Trivedi (2005, Chapter 14, Sections 14.1 to 14.4 and 14.8).

Economists are often interested in models of choice among discrete alternatives. In the simplest case, the economic agent chooses one of two options. Consider an example in which a consumer's utility depends on a durable good D and all other goods G, with $D = 0$ and $D = 1$ the only affordable choices given prices p_D, p_G, and income I. It is helpful to model the choice problem using *conditional indirect utility functions* $V_d(p_D, p_G, I|\theta)$, which give the maximum utility obtained if the consumer were to choose $D = d$, with this utility level determined by prices, income, and parameters θ of the utility function. (Functional forms for conditional indirect utility functions can be found in a number of microeconomic textbooks.) A consumer would choose $D = 0$ if $V_1(\cdot) \leq V_0(\cdot)$ and would choose $D = 1$ if $V_1(\cdot) > V_0(\cdot)$.

To construct an econometric counterpart to this discrete choice problem, we would augment the theoretical model with continuously distributed disturbance terms ϵ_d that represent the unobserved elements of conditional indirect utility, making the further assumption that (ϵ_0, ϵ_1) are distributed independently of prices and income. We would set up the model as

$$D = \begin{cases} 0 & \text{if } V_1(p_D, p_G, I|\theta) + \epsilon_1 \leq V_0(p_D, p_G, I|\theta) + \epsilon_0 \\ 1 & \text{if } V_1(p_D, p_G, I|\theta) + \epsilon_1 > V_0(p_D, p_G, I|\theta) + \epsilon_0. \end{cases}$$

From here, we proceed to derive the *conditional probabilities* that $D = 0$ and $D = 1$ given the values of the covariates. The conditional probability of $D = 1$ can be written as

$$\Pr\Big(\epsilon_1 - \epsilon_0 > -\big(V_1(p_D, p_G, I|\theta) - V_0(p_D, p_G, I|\theta)\big) \;\Big|\; p_D, p_g, I\Big).$$

Given data on prices and incomes, and given a functional form for $V_1(\cdot)$ and $V_0(\cdot)$, all we would need to calculate this probability for a given θ is the distribution of the random variable $\epsilon \equiv \epsilon_1 - \epsilon_0$ representing the difference across alternatives in the unobserved components of utility.

The literature on binary discrete choice has been dominated by two distributions. The assumption that $\epsilon \sim \mathcal{N}(0,1)$ leads to the probit model and the assumption that ϵ is distributed according to the logistic distribution leads to the logit model. In what follows, we explore the probit and logit specifications in some detail. As you read through the material, keep in mind the consumer choice example that we have introduced here. In particular, remember that the expression $-\big(V_1(p_D, p_G, I|\theta) - V_0(p_D, p_G, I|\theta)\big)$ that enters the conditional probability is likely to be nonlinear in the covariates and the θ parameters. In the conventional set-up of the probit and logit models, however, a linear approximation to this expression is substituted for the nonlinear form that would be suggested by economic theory. Although linear approximations are both conventional and convenient, they are by no means a necessary feature of discrete choice econometrics. Once you see how the conventional probit and logit models are constructed, you should be able to see how easy it would be to generalize them to accommodate any nonlinearities implied by your own theory.

20.1 Overview

The probit and logit models can both be fit within the following general framework for a binary dependent variable Y_i, which takes two values conventionally defined to be zero and one. Economists find it useful to think of such binary outcomes as labels or indicators that (as in our consumer choice problem) refer back to a latent variable model

$$Y_i^* = \mathbf{X}_i'\theta + \epsilon_i$$

where $\epsilon_i \sim (0, \sigma_\epsilon^2)$ and the disturbance ϵ_i is assumed independent of $\mathbf{X}_i$, the $k \times 1$ vector of explanatory variables. In this formulation, we observe the indicator $Y_i = 1$ when $Y_i^* > 0$ and have $Y_i = 0$ otherwise. That is, $Y_i = 1$ if $\epsilon_i > -\mathbf{X}_i'\theta$ and $Y_i = 0$ if $\epsilon_i \leq -\mathbf{X}_i'\theta$. (Note the similarity to the consumer choice problem.) Since the *sign* of Y_i^* is observed but *not its magnitude*, we really cannot hope to identify σ_ϵ^2 itself; only the ratio θ/σ_ϵ can be identified. This means that the θ coefficients are themselves identified only up to a normalizing scale factor.

Let $\Pr(Y_i = 1 \mid \mathbf{X}_i, \theta) \equiv p(\mathbf{X}_i, \theta) = 1 - F(-\mathbf{X}_i'\theta)$ with F being the cumulative distribution function of ϵ. Note that conditional on $\mathbf{X}_i$,

$$\begin{aligned} \mathrm{E}\, Y_i \mid \mathbf{X}_i &= p(\mathbf{X}_i, \theta_0) \equiv p_i \\ \mathrm{Var}\, Y_i \mid \mathbf{X}_i &= p_i(1 - p_i). \end{aligned}$$

A compact representation of the probability function for observation i is

$$f_i(Y_i \mid \mathbf{X}_i, \theta) = p(\mathbf{X}_i, \theta)^{Y_i} \cdot (1 - p(\mathbf{X}_i, \theta))^{1-Y_i}.$$

Using this representation, we obtain the normalized log-likelihood function

$$L_n = \frac{1}{n} \sum_{i=1}^{n} \Big(Y_i \ln p(\mathbf{X}_i, \theta) + (1 - Y_i) \ln(1 - p(\mathbf{X}_i, \theta)) \Big)$$

and the score is

$$\begin{aligned}\frac{\partial L_n}{\partial \theta} &= \frac{1}{n}\sum_i \left(Y_i \frac{1}{p_i}\,\frac{\partial p_i}{\partial \theta} - (1-Y_i)\frac{1}{1-p_i}\,\frac{\partial p_i}{\partial \theta}\right) \\ &= \frac{1}{n}\sum_i \frac{1}{p_i(1-p_i)}\,\frac{\partial p_i}{\partial \theta}\,(Y_i - p_i),\end{aligned}$$

with $\partial p_i/\partial\theta$ being a $k \times 1$ vector.

The task of an ML estimation algorithm is to find the $\hat{\theta}$ that satisfies the first-order conditions for the sample log-likelihood with data $\mathbf{y}, \mathbf{x}$,

$$\frac{\partial L_n(\hat{\theta})}{\partial \theta} = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\hat{p}_i(1-\hat{p}_i)}\frac{\partial \hat{p}_i}{\partial \theta}\cdot(y_i - \hat{p}_i) = \mathbf{0}$$

with $\hat{p}_i = p_i(\mathbf{x}_i'\hat{\theta})$. Another way to view the first-order conditions is in terms of *orthogonality* between the $n \times 1$ vector of "residuals" $\mathbf{e} = \mathbf{y} - \hat{\mathbf{p}}$ and each $n \times 1$ vector of functions $(\partial \hat{p}_i/\partial\theta_k)/\hat{p}_i(1-\hat{p}_i)$ associated with θ_k. As usual, this orthogonality can be expressed in a matrix–vector form.

Given the log-probability for one observation,

$$\ln f(Y_i \mid \mathbf{X}_i, \theta) = Y_i \ln p(\mathbf{X}_i, \theta) + (1-Y_i)\ln(1 - p(\mathbf{X}_i, \theta)),$$

and its derivative

$$\frac{\partial \ln f_i}{\partial \theta} = \frac{1}{p_i(1-p_i)}\frac{\partial p_i}{\partial \theta}\cdot(Y_i - p_i),$$

it is easy to see that if the derivative is evaluated at the true θ_0, its expectation conditional on $\mathbf{X}_i$ will be zero. When we condition on $\mathbf{X}_i$, the only remaining random variable is Y_i, and since $\mathrm{E}\, Y_i \mid \mathbf{X}_i \equiv p_i(\theta_0)$, the conditional expectation of $\partial \ln f_i(\theta_0)/\partial\theta$ is zero. The unconditional expectation is therefore also zero.

To calculate $\mathcal{J}_i = \mathrm{Var}\,\partial \ln f_i(\theta_0)/\partial\theta$, we use the outer-product form

$$\mathrm{E}\,\frac{\partial \ln f_i(\theta_0)}{\partial \theta}\frac{\partial \ln f_i(\theta_0)}{\partial \theta}' = \mathrm{E}\,\frac{1}{p_i(1-p_i)}\frac{\partial p_i}{\partial \theta}\cdot(Y_i - p_i)^2\frac{\partial p_i}{\partial \theta}'\frac{1}{p_i(1-p_i)}$$

with all terms on the right-hand side evaluated at θ_0. Conditioning on $\mathbf{X}_i$ and then unconditioning, this becomes

$$\mathcal{J}_i = \mathrm{E}\,\frac{1}{p_i(\theta_0)(1-p_i(\theta_0))}\frac{\partial p_i(\theta_0)}{\partial \theta}\frac{\partial p_i(\theta_0)}{\partial \theta}'$$

As usual, we assume $\mathcal{J} = \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathcal{J}_i$, that is, we assume the limit exists. Given $\hat{\theta}$, we would estimate $\mathcal{J}$ as

$$\hat{\mathcal{J}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i(\hat{\theta})(1-p_i(\hat{\theta}))}\frac{\partial p_i(\hat{\theta})}{\partial \theta}\frac{\partial p_i(\hat{\theta})}{\partial \theta}'.$$

Although these expressions may look complicated, they are not hard to compute for the models most often used in econometrics, the *logit* and *probit* models.

20.2 Logit Model

The logit model can be expressed in terms of a latent variable equation with a disturbance term that follows the logistic distribution. Let the latent variable equation be

$$Y_i^* = \mathbf{X}_i'\theta + \epsilon_i$$

and assume that the disturbance term ϵ_i is drawn from the logistic distribution with cumulative distribution function

$$F(\epsilon) = \frac{e^{\epsilon}}{1+e^{\epsilon}} = \frac{1}{1+e^{-\epsilon}}$$

and density

$$f(\epsilon) = \frac{e^{-\epsilon}}{(1+e^{-\epsilon})^2}.$$

When so distributed, ϵ has mean zero and variance $\pi^2/3$, see Johnson and Kotz (1970b, Chapter 22).

Returning to the latent variable equation $Y_i^* = \mathbf{X}_i'\theta + \epsilon_i$, we find that

$$p_i(\theta) = \Pr\left(Y_i^* > 0 \mid \mathbf{X}_i\right) = \Pr\left(\epsilon_i > -\mathbf{X}_i'\theta \mid \mathbf{X}_i\right) = 1 - F(-\mathbf{X}_i'\theta) = \frac{e^{\mathbf{X}_i'\theta}}{1+e^{\mathbf{X}_i'\theta}}.$$

This simple functional form for $p_i(\theta)$ implies

$$\frac{\partial p_i(\theta)}{\partial \theta} = p_i(\theta)(1-p_i(\theta)) \cdot \mathbf{X}_i.$$

Substituting this result into our general expression for the ML first-order conditions of a binary choice model, we obtain

$$\frac{\partial L_n(\hat{\theta})}{\partial \theta} = \frac{1}{n}\sum_i \mathbf{x}_i \cdot (y_i - p_i(\hat{\theta}) = \mathbf{0}.$$

In other words, with data $\mathbf{y}, \mathbf{x}$, for the logit model the columns of the $\mathbf{x}$ matrix (of dimension $n \times k$) are orthogonal to the residual vector $\mathbf{e} = \mathbf{y} - \hat{\mathbf{p}}$, just as in a linear least squares regression model. If the specification includes a constant term, then the orthogonality condition $\mathbf{x}'\mathbf{e} = \mathbf{0}$ will imply that the arithmetic mean of the residuals is zero. To put this in another way, the average predicted probability in the sample will equal the sample mean of the y_i variables.

The $\mathcal{J}_i$ matrix for the logit case is

$$\mathcal{J}_i = \mathrm{E}\, p_i(\theta_0)(1-p_i(\theta_0)) \cdot \mathbf{X}_i\mathbf{X}_i',$$

and the estimated version of $\mathcal{J}$ is

$$\hat{\mathcal{J}} = \frac{1}{n}\sum_i p_i(\hat{\theta})(1-p_i(\hat{\theta})) \cdot \mathbf{x}_i\mathbf{x}_i'.$$

20.3 Probit Model

For this model it is conventional to begin with a latent variable specification in which the error variance is normalized to unity,

$$Y_i^* = \mathbf{X}_i'\theta + \epsilon_i$$

with $\epsilon \sim \mathcal{N}(0,1)$. Then conditional on $\mathbf{X}_i$,

$$\begin{aligned} p_i(\theta) = \Pr(Y_i = 1|\mathbf{X}_i) &= \Pr(Y_i^* > 0|\mathbf{X}_i) \\ &= \Pr(\epsilon_i > -\mathbf{X}_i'\theta \mid \mathbf{X}_i) \\ &= 1 - \Phi(-\mathbf{X}_i'\theta) \\ &= \Phi(\mathbf{X}_i'\theta), \end{aligned}$$

where Φ is the standard normal c.d.f. and we have invoked in the last line the symmetry property of the normal distribution. For $Y_i = 0$ the conditional probability is $\Phi(-\mathbf{X}_i'\theta)$.

When we examine the derivative for the probit case, we find

$$\frac{\partial p_i}{\partial \theta} = \frac{\partial}{\partial \theta}\Phi(\mathbf{X}_i'\theta) = \phi(\mathbf{X}_i'\theta)\mathbf{X}_i,$$

in which $\phi(\mathbf{X}_i'\theta)$ is the standard normal density function. Therefore, the maximum-likelihood estimate $\hat{\theta}$ for the sample of $\mathbf{y}, \mathbf{x}$ data satisfies the first-order condition

$$\frac{\partial L_n(\hat{\theta})}{\partial \theta} = \frac{1}{n}\sum_i \frac{\phi(\mathbf{x}_i'\hat{\theta})}{\Phi(\mathbf{x}_i'\hat{\theta})\ \Phi(-\mathbf{x}_i'\hat{\theta})}\mathbf{x}_i \cdot (y_i - \hat{p}_i) = \mathbf{0}.$$

(Note that in contrast to the logit case, in a probit model the columns of the $\mathbf{x}$ matrix and the $\mathbf{e} = \mathbf{y} - \hat{\mathbf{p}}$ residual are not orthogonal.) The estimated $\hat{\mathcal{J}}$ matrix for the probit model is

$$\hat{\mathcal{J}} = \frac{1}{n}\sum_{i=1}^{n} \frac{\phi(\mathbf{x}_i'\hat{\theta})^2}{\Phi(\mathbf{x}_i'\hat{\theta})\ \Phi(-\mathbf{x}_i'\hat{\theta})}\mathbf{x}_i\mathbf{x}_i'.$$

20.3.1 Computational considerations

The main computational issue in probit models is how to calculate the standard normal cdf $\Phi()$, a function which is not available in closed form. As you can imagine, quite a lot of effort over the years has gone to finding good numerical approximations to various areas under the standard normal density function. In my own programming in modern Fortran, I've made use of a function originally published in 1973 in the journal *Applied Statistics*, which the Australian statistician Alan Miller recoded into Fortran 90: see http://jblevins.org/mirror/amiller/as66.f90. It is possible that improvements have been made since 1973 that render this particular algorithm obsolete, but I have never noticed a problem with it. These computational and algorithmic concerns are more pressing in the bivariate and multivariate generalizations of probit models, which we will discuss later in this chapter.

For probits, a nice computational trick to know about exploits the symmetry of the normal distribution. Recall that for the $Y_i = 1$ case, the conditional probability $\Pr(\epsilon_i > -\mathbf{X}_i'\theta) = \Pr(\epsilon_i \leq \mathbf{X}_i'\theta)$ by symmetry and for the $Y_i = 0$ case, we use $\Pr(\epsilon_i \leq -\mathbf{X}_i'\theta)$. If Y_i were redefined so that $\epsilon_i \leq -\mathbf{X}_i'\theta$ maps to $Y_i = -1$ rather than to 0, we could represent the probabilities of both outcomes in the compact notation

$$\Pr(Y_i = y_i \mid \mathbf{X}_i) = \Phi(y_i \cdot \mathbf{X}_i'\theta)$$

This considerably simplifies the programming. Some authors prefer to leave Y_i coded in its original $1, 0$ form, but for compactness of notation write the two probabilities as

$$\Pr(Y_i = y_i \mid \mathbf{X}_i) = \Phi((2y_i - 1) \cdot \mathbf{X}_i'\theta)$$

which is obviously the same thing.

20.4 Comparing Logit and Probit Results

When the logit and probit specifications are represented in their latent variable form $Y_i^* = \mathbf{X}_i'\theta + \epsilon_i$, it is evident that one difference between them has to do with the variance assumed for ϵ_i. In the logit specification the variance is $\pi^2/3$, whereas in the probit specification σ_ϵ^2 is normalized to unity. In comparing estimated coefficients $\hat{\theta}$ across the two models, therefore, one should multiply the logit estimates by $\sqrt{3}/\pi$ to put the models on a roughly equal footing. This is often sufficient for approximate comparisons, but other standardization methods are occasionally used, see Johnson and Kotz (1970b, Chapter 22) and Greene (2003, Chapter 21).

20.5 Consequences of Omitting Variables

The omission of a relevant explanatory variable is a common form of specification error. In general, the maximum likelihood method does not produce consistent coefficient estimates under mis-specification of the likelihood function.[1] To see how things go wrong, the following probit example may be helpful.

Beginning with the latent variable specification (and suppressing for now the i subscript that indexes observations), let the latent equation of the true model be

$$Y^* = \beta_0 + X\beta_1 + Z\delta + \epsilon \tag{20.1}$$

in which both X and Z are single variables and the disturbance ϵ is standard normal and assumed independent of both X and Z. Suppose, however, that we mistakenly omit the Z variable—with $\delta \neq 0$ this is a specification error. The composite disturbance $u \equiv Z\delta + \epsilon$ of the mis-specified model will not be independent of X, in general, and it is also unlikely to

[1] To be sure, interesting special cases exist in which mis-specification is not so disastrous. In these cases, the first-order conditions of the misspecified ML problem, although derived from the wrong likelihood function, nevertheless function as generalized "moments" that produce consistent estimates of the parameters. These cases will be addressed later in the context of the generalized method-of-moments estimator.

be standard normal. Two sorts of consequences follow from this specification error. First, just as in linear models, the estimated $\hat{\beta}_1$ coefficient on the included X variable will be inconsistent, having a probability limit that differs from β_1 owing to the association between X and Z, and the estimator $\hat{\beta}_0$ of the constant term will also be inconsistent. Second, unless the composite disturbance u happens to follow the standard normal distribution, the use of the standard normal cdf $\Phi(\beta_0 + \beta_1 X)$ to represent $\Pr(u > -(\beta_0 + X'\beta_1) \mid X)$ will be incorrect.

To gain insight into the first source of inconsistency, imagine that the X and Z covariates are distributed as bivariate normal with means of zero (Yatchew and Griliches 1984). We can then express Z in terms of X, as

$$Z = \frac{\sigma_{ZX}}{\sigma_{XX}} X + w,$$

with w being independent of X and normally distributed with a mean of zero and variance σ_w^2. Substituting this expression for Z, we obtain

$$Y^* = \beta_0 + \left(\beta_1 + \delta \frac{\sigma_{ZX}}{\sigma_{XX}}\right) X + w\delta + \epsilon.$$

The term in parentheses combines β_1, the true effect of X, with Z's coefficient δ as well as the association between the included X and the omitted Z variable. The new composite disturbance $w\delta + \epsilon$ is normally distributed. However, its variance is $\sigma_w^2\delta^2 + 1$, whereas a probit model assumes a disturbance term variance of one. To put the model into a probit form, we must therefore divide through by the standard deviation of the composite disturbance. This yields

$$\begin{aligned} Y_i^{**} &= \frac{\beta_0}{\sqrt{\sigma_w^2\delta^2 + 1}} + \left(\frac{\beta_1 + \delta\frac{\sigma_{ZX}}{\sigma_{XX}}}{\sqrt{\sigma_w^2\delta^2 + 1}}\right) X + v \\ &= \gamma_0 + \gamma_1 X + v. \end{aligned}$$

Here at last is a properly-specified probit model. Its estimated $\hat{\gamma}_0$ and $\hat{\gamma}_1$ coefficients will converge to the corresponding γ_0 and γ_1 parameters and the limiting variances of the coefficients will be correctly estimated. The problem, of course, is that the γ_0 and γ_1 parameters are messy composites that are only distantly related to the β_0 and β_1 parameters of the true model.

20.6 Lagrange Multiplier Tests

Recall that the LM statistic can be written as

$$LM = \sqrt{n}\frac{\partial L_n(\tilde{\theta})}{\partial \theta}' \tilde{\mathcal{J}}^{-1} \sqrt{n}\frac{\partial L_n(\tilde{\theta})}{\partial \theta} = \frac{\partial L(\tilde{\theta})}{\partial \theta}' \tilde{\mathcal{R}}^{-1} \frac{\partial L(\tilde{\theta})}{\partial \theta}$$

where $\tilde{\theta}$ is the constrained estimator of θ and $\tilde{\mathcal{R}}$ is the information matrix evaluated at the constrained estimates. We now show how to rewrite this test statistic in a form easily calculable by an artificial regression.

Let $\tilde{\mathbf{S}}$ be the $N \times K$ matrix

$$\tilde{\mathbf{S}} = \begin{bmatrix} \frac{\partial \tilde{p}_1 / \partial \theta_1}{\sqrt{\tilde{p}_1(1-\tilde{p}_1)}} & \cdots & \frac{\partial \tilde{p}_1 / \partial \theta_K}{\sqrt{\tilde{p}_1(1-\tilde{p}_1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \tilde{p}_N / \partial \theta_1}{\sqrt{\tilde{p}_N(1-\tilde{p}_N)}} & \cdots & \frac{\partial \tilde{p}_N / \partial \theta_K}{\sqrt{\tilde{p}_N(1-\tilde{p}_N)}} \end{bmatrix}$$

where $\tilde{p}_i = p_i(\mathbf{X}_i'\tilde{\theta})$. We estimate the information matrix $\tilde{\mathcal{R}}$ using $\tilde{\mathbf{S}}'\tilde{\mathbf{S}}$. Redefining the vector of residuals $\tilde{\mathbf{e}}$ so that it has i-th element

$$\tilde{\mathbf{e}}_i = \frac{y_i - \tilde{p}_i}{\sqrt{\tilde{p}_i(1-\tilde{p}_i)}},$$

it follows that $\tilde{\mathbf{S}}'\tilde{\mathbf{e}} = \partial L(\tilde{\theta})/\partial\theta$, the score evaluated at the constrained estimates. Therefore,

$$\text{LM } = \tilde{\mathbf{e}}'\tilde{\mathbf{S}} \left(\tilde{\mathbf{S}}'\tilde{\mathbf{S}}\right)^{-1} \tilde{\mathbf{S}}'\tilde{\mathbf{e}},$$

a quantity recognizable as the explained sum of squares from an artificial regression of $\tilde{\mathbf{e}}$ on $\tilde{\mathbf{S}}$. This can be taken to a χ^2_q table, with q being the number of constraints.

Testing for omitted variables

Consider the probit case with $Y_i^* = \mathbf{X}_i'\theta + \mathbf{Z}_i'\delta + \epsilon_i$, where the null hypothesis is $\delta = 0$. In this case, the constrained estimates $\tilde{\theta}$ are simply the usual ML estimates of θ obtained by omitting $\mathbf{Z}_i$ from the model. Recalling that $p_i = \Phi(\mathbf{X}_i'\theta + \mathbf{Z}_i'\delta)$, we obtain

$$\begin{aligned} \frac{\partial \tilde{p}_i}{\partial \theta_j} &= \phi(\mathbf{X}_i'\tilde{\theta}) X_{i,j} \\ \frac{\partial \tilde{p}_i}{\partial \delta_j} &= \phi(\mathbf{X}_i'\tilde{\theta}) Z_{i,j}. \end{aligned}$$

The i-th row of the matrix $\tilde{\mathbf{S}}$ is then

$$\tilde{\mathbf{S}}_i = \frac{\phi(\mathbf{X}_i'\tilde{\theta})}{\sqrt{\Phi(\mathbf{X}_i'\tilde{\theta})\Phi(-\mathbf{X}_i'\tilde{\theta})}} \left[X_{i,1}, \ldots, X_{i,K_1}, Z_{i,1}, \ldots, Z_{i,K_2}\right].$$

The normalized i-th residual $\tilde{\mathbf{e}}_i$ is

$$\tilde{\mathbf{e}}_i = \frac{y_i - \Phi(\mathbf{X}_i'\tilde{\theta})}{\sqrt{\Phi(\mathbf{X}_i'\tilde{\theta})\Phi(-\mathbf{X}_i'\tilde{\theta})}}.$$

Of course, one would not normally use a Lagrange Multiplier test in this case, because with modern statistical software it is far easier test $\delta = \mathbf{0}$ with a Wald test or run the model with and without $\mathbf{Z}$ and use a likelihood ratio test. However, the analysis provides a helpful benchmark for the next test that we will consider.

Testing for heteroskedasticity

What we will show is that, first, it is relatively easy to form a Lagrange Multiplier test for heteroskedasticity in the probit case; and second, because the test is so similar in form to the omitted variables LM test, we should be *very cautious* about interpreting rejection of the null hypothesis as evidence in favor of heteroskedasticity.

Given the latent specification $Y_i^* = \mathbf{X}_i'\theta + \epsilon_i$, let the standard deviation of ϵ_i be represented as

$$\sigma_{\epsilon_i} = e^{\mathbf{Z}_i'\delta}$$

where we have (implicitly) assumed that with $\delta = \mathbf{0}$ the variance of ϵ_i is unity. Recall that the scale of ϵ is not identifiable; nevertheless, it is possible to incorporate into the probit model heteroskedasticity of the form just shown. In principle, at least, the δ coefficients are identified. Even if your statistical software provides an option to estimate a heteroskedastic probit model, you should certainly conduct a Lagrange Multiplier test to see if the data are suggestive of heteroskedasticity.

Now, given that

$$p_i = \Phi\left(\frac{\mathbf{X}_i'\theta}{e^{\mathbf{Z}_i'\delta}}\right),$$

in the heteroskedastic case, under the null ($\delta = \mathbf{0}$) the constrained estimates $\tilde{\theta}$ are the usual ML estimates assuming homoskedasticity. We have

$$\begin{aligned}\frac{\partial \tilde{p}_i}{\partial \theta_j} &= \phi(\mathbf{X}_i'\tilde{\theta})X_{i,j} \\ \frac{\partial \tilde{p}_i}{\partial \delta_j} &= \phi(\mathbf{X}_i'\tilde{\theta})(-\mathbf{X}_i'\tilde{\theta})Z_{i,j}.\end{aligned}$$

Thus, the i-th row of $\tilde{\mathbf{S}}$ is

$$\tilde{\mathbf{S}}_i = \frac{\phi(\mathbf{X}_i'\tilde{\theta})}{\sqrt{\Phi(\mathbf{X}_i'\tilde{\theta})\Phi(-\mathbf{X}_i'\tilde{\theta})}}\left[X_{i,1}, \ldots, X_{i,K_1}, (-\mathbf{X}_i'\tilde{\theta})Z_{i,1}, \ldots, (-\mathbf{X}_i'\tilde{\theta})Z_{i,K_2}\right].$$

Apart from the scalar $(-\mathbf{X}_i'\tilde{\theta})$, this is identical to the expression used in the omitted variables test. Therefore, rejection of the null hypothesis $\delta = \mathbf{0}$ may mean either that heteroskedasticity is present, or that the $\mathbf{Z}_i$ variables properly belong with $\mathbf{X}_i$ in the specification of the mean of the latent variable Y_i^*.

20.7 Understanding the Results

In highly nonlinear models such as the probit and logit models, it is very difficult to get a sense of whether a given coefficient estimate $\hat{\theta}_k$ implies that X_{ik} will have a large or small effect on the outcome probability. In my view, by far the best way to understand the substantive implications of your estimates is to summarize things in terms of predicted probabilities.

Predicted probabilities actually help us in two ways: First, they can show how well the estimated model fits a portion of the data; and second, as just mentioned, they can reveal the economic importance of the estimated coefficients. These are different reasons to use predicted probabilities, and they call for slightly different calculations. To illustrate, we consider the probit case in which the latent variable $Y_i^* = \mathbf{X}_i'\theta + Z_i\theta_Z + \epsilon_i$ and the dummy variable Z_i is the covariate in which we are interested. This covariate takes two values in our example, corresponding to male, for which $Z_i = 0$, and female ($Z_i = 1$).

We ask first whether the predicted probabilities of the model, which for individual i are given by $\Phi(\mathbf{X}_i'\hat{\theta} + Z_i\hat{\theta}_Z)$, reproduce the empirical probabilities for men and women. Suppose that π_0 is the proportion of men in the sample for whom $Y_i = 1$ and let π_1 be the corresponding proportion for women, with the subscript indicating the value taken on by the Z_i variable. To calculate the predicted versions, we average predicted probabilities over the *sub-sample* of men and (separately) the sub-sample of women. That is, we calculate

$$\hat{p}_0 = \frac{1}{n_0} \sum_{Z_i=0} \Phi(\mathbf{X}_i'\hat{\theta})$$

for men, and for women,

$$\hat{p}_1 = \frac{1}{n_1} \sum_{Z_i=1} \Phi(\mathbf{X}_i'\hat{\theta} + \hat{\theta}_Z).$$

We expect that if the model fits the data reasonably well, then the value of $\hat{p}_0$ should be close to the empirical probability π_0 for men, and similarly for women. Notice that the values of the other covariates $\mathbf{X}_i$ entering these calculations are the values that prevail in the male and female sub-samples. That is, $\hat{p}_0$ is an estimate of the *conditional probability* of $Y_i = 1$ among men, using the values for other covariates that are seen in the male sub-sample..

A different objective would be to use predicted probabilities to learn about the substantive importance of the $\hat{\theta}_Z$ coefficient. For this purpose we would also calculate two predicted probabilities. Setting $Z_i = 0$ for all $i = 1, \ldots, n$ yields one set of $\hat{p}_{i,0}$ predicted probabilities, and setting all $Z_i = 1$ yields the other set, $\hat{p}_{i,1}$. Note that each predicted probability depends on the values taken by the other covariates, which again we would leave unchanged. The overall predicted probabilities are then calculated via

$$\hat{p}_0 = \frac{1}{n} \sum_i \Phi(\mathbf{X}_i'\hat{\theta}) \text{ and}$$

$$\hat{p}_1 = \frac{1}{n} \sum_i \Phi(\mathbf{X}_i'\hat{\theta} + \hat{\theta}_Z).$$

Note that in contrast to the first way we used predicted probabilities, the summations in these two expressions range over the *full* sample. Forming the predicted probabilities in this way, by flipping Z_i from 0 to 1 for all observations while leaving intact the values of all other covariates, and then averaging over the full sample, yields a difference $\hat{p}_1 - \hat{p}_0$ that is the discrete analog to a partial derivative. When we apply this method, we do not necessarily expect $\hat{p}_1$ to closely resemble the empirical probability π_1, nor should $\hat{p}_0$ necessarily be close to π_0, although often the values do turn out to be similar. The idea is not to reproduce the empirical probabilities, but rather to see clearly the size of the Z_i effect.

The same general methods can be applied when the covariate of interest X_{ik} is a continuous variable. In this instance we could calculate predicted derivatives of the outcome probability with respect to X_{ik}, or—this is the approach I myself would follow—we could calculate average outcome probabilities for selected values of X_{ik} that span a range of interest.

Whether X_{ik} is discrete or continuous, it is likely that its implications for outcome probabilities will be best depicted in a figure, such as a bar chart or line graph. In addition to the point estimates of the effect, some readers will want to see in the figure some indication of the standard errors associated with the average predicted values. How might these standard errors be calculated?

Here we would use what is termed the "delta method" to approximate the variance of a predicted value. Consider a given $\hat{p}_{i,0}$ derived by setting $X_{ik} = 0$ as above. Taylor-expand this function around the true parameter value θ_0,

$$\hat{p}_{i,0} = p_{i,0} + \frac{\partial \tilde{p}_{i,0}}{\partial \theta}' \left(\hat{\theta} - \theta_0\right),$$

in which $p_{i,0}$ is evaluated at θ_0 and the derivative vector $\partial \tilde{p}_{i,0}/\partial\theta$ is evaluated at a $\tilde{\theta}$ that lies between $\hat{\theta}$ and θ_0. Then

$$\sqrt{n}(\hat{p}_{i,0} - p_{i,0}) \xrightarrow{d} \mathcal{N}\left(0, \partial p_{i,0}/\partial\theta' \mathcal{J}^{-1} \partial p_{i,0}/\partial\theta\right).$$

Although the limiting variance depends on the true θ_0, we would estimate it using $\hat{\theta}$ as usual.

Now, in maximum-likelihood problems most statistical package supply us with $\left(n\hat{\mathcal{J}}\right)^{-1}$, which is the inverse of the information matrix, rather than the $\hat{\mathcal{J}}^{-1}$ matrix that figures in the asymptotic theory. The reason is that the inverse of the information matrix provides an approximation to the variance of $\hat{\theta}$ itself, whereas the asymptotic theory delivers results about the limiting variance of $\sqrt{n}(\hat{\theta} - \theta_0)$, a transformation of this estimator. Hence, we take $\partial \hat{p}_{i,0}/\partial\theta' \left(n\hat{\mathcal{J}}\right)^{-1} \partial \hat{p}_{i,0}/\partial\theta$ to be the approximate variance of $\hat{p}_{i,0}$.

After estimating a probit or logit model, programs such as R and STATA temporarily store the inverse of the information matrix in a matrix. Furthermore, for the probit case note that $\partial p_i(\hat{\theta})/\partial\theta = \phi(\mathbf{X}_i'\hat{\theta})\mathbf{X}_i$. The approximate variance of $p_i(\hat{\theta})$ in this case is therefore

$$\phi(\mathbf{X}_i'\hat{\theta})^2 \cdot \mathbf{X}_i' \left(n\hat{\mathcal{J}}\right)^{-1} \mathbf{X}_i.$$

If we have saved the inverse of the information matrix, the quadratic form is easy enough to calculate and for the $\phi(.)$ component of the expression we can use the built-in standard normal density function.

Where does all this leave us in terms of calculating error bounds on an average predicted probability? We have learned how to approximate the variance of $\hat{p}_{i,0}$ in the probit case (the logit case is of course very similar), but you should take note of one particular feature of the results: the variance $\phi(\mathbf{X}_i'\hat{\theta})^2 \cdot \mathbf{X}_i' \left(n\hat{\mathcal{J}}\right)^{-1} \mathbf{X}_i$ differs across i, that is, the $\hat{p}_{i,0}$ are heteroskedastic. Furthermore, each $\hat{p}_{i,0}$ depends on the same random vector $\hat{\theta}$, and this implies that the $\hat{p}_{i,0}$ will be intercorrelated to some degree. It is no trivial task to calculate

rigorously the standard error bounds on an average of n heteroskedastic and intercorrelated random variables. (Consider extending the analysis above to consider a vector of predicted probabilities.) When we discuss heteroskedasticity in the linear model—an average is the simplest such model—we will see how we might proceed to deal with the heteroskedasticity part of the problem using the method of weighted least squares.

20.8 Bivariate and Multivariate Probit Models

It is conceptually straightforward to generalize the single-equation probit model to the case of two "seemingly unrelated" latent variable equations connected by a non-zero correlation term. Let $(Y^*_{1,i}, Y^*_{2,i})$ represent the latent variables, which are jointly normally distributed conditional on covariates $(\mathbf{X}_{1,i} = \mathbf{x}_{1,i}, \mathbf{X}_{2,i} = \mathbf{x}_{2,i})$ such that

$$\begin{pmatrix} Y^*_{1,i} \\ Y^*_{2,i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{x}'_{1,i}\beta_1 \\ \mathbf{x}'_{2,i}\beta_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right)$$

We can rewrite this in the form of a two-equation latent-variables system

$$\begin{aligned} Y^*_{1,i} &= \mathbf{x}'_{1,i}\beta_1 + \epsilon_{1,i} \\ Y^*_{2,i} &= \mathbf{x}'_{2,i}\beta_2 + \epsilon_{2,i} \end{aligned}$$

with the distribution of ϵ_i given $\mathbf{X}_i$ being normal with means of zero, unit variances, and covariance ρ. The latent variables map to the observed $(Y_{1,i}, Y_{2,i})$ much as in the standard probit case, with $Y^*_{1,i} > 0$ implying $Y_{1,i} = 1$, and $Y^*_{1,i} \leq 0$ implying $Y_{1,i} = 0$, and similarly for $Y^*_{2,i}$. Apart from the cross-equation linkage due to the covariance ρ, then, the two equations are identical to their counterparts in standard probit models. Indeed, β_1 and β_2 can be consistently estimated via two separate probits.

In a full maximum-likelihood approach, β_1, β_2 and ρ are all estimated jointly (perhaps with starting values for β_1, β_2 supplied by the two standard probit models). This approach requires efficient and accurate approximations to the cumulative distribution functions of bivariate (and more generally, multivariate) normal distributions. Specialists have made a considerable investment in methods for calculating areas under multivariate normal densities, and this remains an active area of research. One family of such methods has been developed by Alan Genz (http://www.math.wsu.edu/faculty/genz/homepage) and implemented in both an R package (*mvtnorm*) and in downloadable Fortran 77 code.[2] As will be discussed later, in economics the Geweke–Hajivassiliou–Keane simulation approach is a popular alternative. Both approaches are able to handle high-dimensional distributions, although in economics it is unusual to encounter multivariate normal distributions with more than 20 dimensions or so.

In a pair of NBER working papers, Mullahy (2011) and Mullahy (2015) proves the following very useful (but apparently little-known) result about the partial derivative of a bivariate cumulative distribution function $F(v_1, v_2) = \Pr(V_1 \leq v_1, V_2 \leq v_2)$ with respect to

[2] For modern Fortran versions, see Alan Miller's site http://jblevins.org/mirror/amiller/ where some of Genz' routines are available under the "Miscellaneous" category.

its v_i arguments:

$$\frac{\partial F(v_1, v_2)}{\partial v_1} = f(v_1) \cdot G(v_2 \mid v_1)$$

in which $f()$ is the univariate density of V_1 and G is the cdf of V_2 conditional on $V_1 = v_1$. (The derivative of F with respect to other parameters of the joint distribution—such as ρ in the joint normal case—is not addressed in this result.) Upon differentiating again, we see that the result is consistent with the second-partials translation from joint cdf's to joint densities familiar from introductory mathematical statistics:

$$\frac{\partial^2 F(v_1, v_2)}{\partial v_1 \partial v_2} = f(v_1) \cdot \frac{\partial G(v_2 \mid v_1)}{\partial v_2} = f(v_1) \cdot g(v_2 \mid v_1) = h(v_1, v_2)$$

Mullahy also derives similar expressions for higher-dimensional cases.

To make use of Mullahy's result for the bivariate case at hand, consider the joint probability of $Y_1 = 0, Y_2 = 0$, or equivalently, $\epsilon_1 \leq -\mathbf{x}_1'\beta_1, \epsilon_2 \leq -\mathbf{x}_2'\beta_2$. (We omit the observation index i.) In terms of the joint normal cdf conditional on $\mathbf{X}_1, \mathbf{X}_2$, this is

$$\Pr(Y_1 = 0, Y_2 = 0 \mid \mathbf{X}_1, \mathbf{X}_2) = F(-\mathbf{x}_1'\beta_1, -\mathbf{x}_2'\beta_2)$$

in which the role of ρ in F is left implicit. We can differentiate this expression either with respect to the β parameters (to get the elements of the score vector for the βs), or with respect to the $\mathbf{X}$ covariates to investigate the marginal effects of these covariates on the joint probability.

Let's first consider the score components in a simple case in which the only explanatory "variables" in the model are the two constant terms β_1 and β_2. Then

$$\Pr(Y_1 = 0, Y_2 = 0) = F(-\beta_1, -\beta_2)$$

The partial with respect to β_1 is then

$$\partial F / \partial \beta_1 = -\phi(-\beta_1) \cdot \Pr(\epsilon_2 \leq -\beta_2 \mid \epsilon_1 = -\beta_1).$$

with $\phi()$ being the standard normal density. For bivariate normal random variables, the distribution of ϵ_2 conditional on the value of ϵ_1 is normal with mean $\rho\epsilon_1$ and variance $1 - \rho^2$. The event $\epsilon_2 \leq -\beta_2$ is the same as $\epsilon_2 - \rho\epsilon_1 \leq -\beta_2 - \rho\epsilon_1$, and with $\epsilon_1 = -\beta_1$, this is the same as

$$\frac{\epsilon_2 + \rho\beta_1}{\sqrt{1 - \rho^2}} \leq \frac{-\beta_2 + \rho\beta_1}{\sqrt{1 - \rho^2}}$$

Conditional on ϵ_1, the random variable on the left is distributed as standard normal and therefore

$$\Pr\left(\frac{\epsilon_2 - \rho\epsilon_1}{\sqrt{1 - \rho^2}} \leq \frac{-\beta_2 + \rho\beta_1}{\sqrt{1 - \rho^2}}\right) = \Phi\left(\frac{-\beta_2 + \rho\beta_1}{\sqrt{1 - \rho^2}}\right)$$

Therefore,

$$\partial F / \partial \beta_1 = -\phi(-\beta_1) \cdot \Phi\left(\frac{-\beta_2 + \rho\beta_1}{\sqrt{1 - \rho^2}}\right).$$

The derivative of the joint cdf with respect to β_1 is thus expressed in terms of a univariate standard normal density and a marginal (but implicitly conditional) cdf. Similarly, reversing the roles of β_1 and β_2, we have

$$\partial F/\partial \beta_2 = -\phi(-\beta_2) \cdot \Phi\left(\frac{-\beta_1 + \rho\beta_2}{\sqrt{1-\rho^2}}\right).$$

In a more realistic case with actual $\mathbf{X}_1, \mathbf{X}_2$ covariates, the derivative vectors are

$$\partial F/\partial \beta_1 = -\mathbf{x}_1 \cdot \phi(-\mathbf{x}_1'\beta_1) \cdot \Phi\left(\frac{-\mathbf{x}_2'\beta_2 + \rho\mathbf{x}_1'\beta_1}{\sqrt{1-\rho^2}}\right),$$

which is a $k_1 \times 1$ vector (k_1 being the number of β_1 parameters), and

$$\partial F/\partial \beta_2 = -\mathbf{x}_2 \cdot \phi(-\mathbf{x}_2'\beta_2) \cdot \Phi\left(\frac{-\mathbf{x}_1'\beta_1 + \rho\mathbf{x}_2'\beta_2}{\sqrt{1-\rho^2}}\right),$$

a $k_2 \times 1$ vector. Using these analytic results, we could fill in all but one of the elements of the score contribution $\partial \ln F/\partial \beta$ for a single observation, the entry for ρ being the sole missing element.

As with standard probit models, there is a computational trick (using the symmetry of normal distributions) that lets us handle all four of the bivariate possibilities in one compact notation. Let $s_{1,i} = 2y_{1,i} - 1$ and $s_{2,i} = 2y_{2,i} - 1$ so that s_1 and s_2 are in a $1, -1$ form. Then

$$\Pr\left(Y_{1,i} = y_{1,i}, Y_{2,i} = y_{2,i} \mid \mathbf{X}_{1,i}, \mathbf{X}_{2,i}\right) = F(s_{1,i}\mathbf{x}_{1,i}'\beta_1, s_{2,i}\mathbf{x}_{2,i}'\beta_2, s_{1,i}s_{2,i}\rho)$$

Note that when only one of the pair of outcome variables $Y_{1,i}, Y_{2,i}$ equals 1, the sign of ρ must be switched. Mullahy shows how to apply this notation to higher-dimensional probit models.

Chapter 21

The Conditional Logit Model

Students: You can skim through section 21.3. Supplement this chapter with Cameron and Trivedi (2005, Chapter 15), which effectively distinguishes the conditional logit from the multinomial logit model.

This chapter explores the conditional logit model, an important econometric representation used in discrete choice problems. Although we discuss here only the simple version of that model, you should be aware that a more general "nested" version exists which is useful in many applications. Cameron and Trivedi (2005) give the best textbook description I have seen of the standard and nested models.

21.1 Overview

Consider a discrete decision problem of the following sort. Our task is to choose the alternative j that yields maximum utility among $J+1$ alternatives, that is, $U_j = \text{argmax}\ (U_0, ..., U_J)$, where the alternative-specific utility for the c-th choice is given by

$$U_c = V_c + \epsilon_c,$$

in which the functional form of V_c may be taken from a conditional indirect utility function as in our previous chapter. The decision-maker is assumed to know the values of V_c and ϵ_c for all alternatives, and chooses from these the alternative yielding maximum utility. Note that V_c is a function of the characteristics of the c-th choice. It may also depend on the characteristics of the decision-maker, provided that these "interact" in some way with the characteristics of the alternative. We denote all relevant characteristics by the vector $\mathbf{X}_c$. In most econometric software, the linear specification $V_c = \mathbf{X}_c'\beta$ is assumed. If your theory suggests a nonlinear relationship—and it probably will—then it will be necessary to write your own computer code in order to maximize the correct likelihood function.

Under an assumption about the disturbance terms $(\epsilon_0, ...\epsilon_J)$ that will spelled out in a moment, the j-th alternative will be the utility-maximizing choice with probability

$$P_j = \frac{e^{V_j}}{\sum_{c=0}^{J} e^{V_c}}.$$

The form of this probability makes it clear why the model bears the logit name—it is a generalization of the functional form seen in the binary logit model. Although we cannot know the exact value of the utility obtained when the utility-maximizing choice is made, we can calculate its expected value, which is

$$\mathrm{E}\max(U_0, U_1, \ldots, U_J) = \ln\left(\sum_{j=0}^{J} e^{V_j}\right).$$

This result is very useful in applied welfare problems. Suppose, for example, that the choices being modelled involve selecting one consumer durable out of a set of $J+1$ such durables. We may be interested in the effect of a change in the price of the j-th durable good on consumer well-being.

To obtain these results, we require distributional assumptions on the disturbance terms. Let each ϵ_c be Gumbel-distributed with parameters (η, μ), also known as a Type I extreme value, with cdf

$$F(\epsilon \mid \eta, \mu) \;=\; \exp(-e^{-\mu(\epsilon-\eta)}),$$

in which η is a location parameter and μ is a scale parameter (like a standard deviation). Letting $\gamma \cong .577$ denote Euler's constant, a $\mathcal{G}(\eta, \mu)$ variable can be characterized as follows,

$$\begin{aligned} \mathrm{E}\ \epsilon &= \eta + \frac{\gamma}{\mu} \\ \text{mode}\ \epsilon &= \eta \\ \mathrm{Var}\ \epsilon &= \frac{\pi^2}{6\mu^2}. \end{aligned}$$

The following are additional useful properties of the Gumbel distribution:

- $\epsilon \sim \mathcal{G}(\eta, \mu)$ implies $\alpha\epsilon + V \sim \mathcal{G}(\alpha\eta + V, \frac{\mu}{\alpha})$ for $\alpha > 0$.
- If ϵ_1 and ϵ_2 are independent $\mathcal{G}(\eta_1, \mu)$ and $\mathcal{G}(\eta_2, \mu)$, then $\epsilon_1 - \epsilon_2$ is logistically distributed
$$F(\epsilon_1 - \epsilon_2) \;=\; \frac{1}{1 + \exp(\mu(\eta_2 - \eta_1 - (\epsilon_1 - \epsilon_2)))}.$$
- If ϵ_1, ϵ_2 are independent $\mathcal{G}(\eta_1, \mu)$, $\mathcal{G}(\eta_2, \mu)$, this implies
$$\max(\epsilon_1, \epsilon_2) \sim \mathcal{G}(\frac{1}{\mu}\ \log(e^{\mu\eta_1} + e^{\mu\eta_2}), \mu).$$
Hence, for $\epsilon_0, ..., \epsilon_J$ independent with $\mathcal{G}(\eta_j, \mu)$,
$$\max(\epsilon_0, ...\epsilon_J) \sim \mathcal{G}(\frac{1}{\mu}\ \log \sum_{j=0}^{J} e^{\mu\eta_j}, \mu).$$

We will use these properties in order to reduce a $J+1$ dimensional choice problem to a simpler problem involving only two alternatives.

21.2 $\mathcal{G}(0,1)$ Disturbances

Now assume that in the choice problem with $J+1$ alternatives,

$$\begin{aligned} U_0 &= V_0 + \epsilon_0 \\ &\vdots \\ U_J &= V_J + \epsilon_J, \end{aligned}$$

the ϵ's are mutually independent $\mathcal{G}(0,1)$ disturbances. Let us derive the probability P_0 that alternative "0" is utility-maximizing.

Note that each $U_j \sim \mathcal{G}(V_j, 1)$ and let $U_* = \max(U_1, ..., U_J)$. Then

$$U_* \sim \mathcal{G}(\log \sum_{j=1}^{J} e^{V_j}, 1).$$

We can rewrite this as $U_* = V_* + \epsilon_*$ with $V_* \equiv \log \sum_{j=1}^{J} e^{V_j}$ and $\epsilon_* \sim \mathcal{G}(0,1)$. Hence,

$$\begin{aligned} P_0 &= \Pr(V_0 + \epsilon_0 \geq V_* + \epsilon_*) \\ &= \Pr(\epsilon_0 - \epsilon_* \geq -(V_0 - V_*)) \\ &= \frac{1}{1 + e^{V_* - V_0}} \\ &= \frac{e^{V_0}}{e^{V_0} + e^{V_*}}, \end{aligned}$$

where in the third row we have made use of the logistic property mentioned above. With $V_* = \log \sum_{j=1}^{J} e^{V_j}$, we have

$$\begin{aligned} P_0 &= \frac{e^{V_0}}{e^{V_0} + \sum_{j=1}^{J} e^{V_j}} \\ &= \frac{e^{V_0}}{\sum_{j=0}^{J} e^{V_j}}, \end{aligned}$$

as we set out to prove.

21.3 The Score and Information Matrix

In what follows, we will derive the log-likelihood contribution and the contribution to the score and information matrix for one observation, generally omitting the i subscript for notational simplicity.

Let each choice be associated with covariates $\mathbf{X}_c$. Under the conventional linear specification positing $V_c = \mathbf{X}_c'\beta$, we would have this choice probability for alternative c,

$$P_c = \frac{e^{\mathbf{X}_c'\beta}}{\sum_j e^{\mathbf{X}_j'\beta}}.$$

One observation's contribution to the log-likelihood can be expressed as

$$\ln f_i = \sum_c d_c \, \log P_c = \sum_c d_c \left(\mathbf{X}_c'\beta - \log(\sum_j e^{\mathbf{X}_j'\beta}) \right)$$

where $d_c = 1$ if option c is actually chosen and $d_c = 0$ otherwise. The score contribution for this observation is

$$\begin{aligned} \frac{\partial \ln f_i}{\partial \beta_k} &= \sum_c d_c \left(X_{ck} - \frac{\sum_j X_{jk} e^{\mathbf{X}_j'\beta}}{\sum_l e^{\mathbf{X}_l'\beta}} \right) \\ &= \sum_c d_c (X_{ck} - \sum_j X_{jk} P_j) \\ &\equiv \sum_c d_c (X_{ck} - \bar{X}_k) \end{aligned}$$

where

$$\bar{X}_k = \sum_j X_{jk} P_j$$

is the weighted average of X_{jk} with weights equal to the probability that choice j is made.

Another way to write the score contribution (for one observation) is

$$\begin{aligned} \frac{\partial \ln f_i}{\partial \beta_k} &= \sum_c (d_c - P_c)(X_{ck} - \bar{X}_k) + P_c(X_{ck} - \bar{X}_k) \\ &= \sum_c (d_c - P_c)(X_{ck} - \bar{X}_k) + \sum_c P_c(X_{ck} - \bar{X}_k) \\ &= \sum_c (d_c - P_c)(X_{ck} - \bar{X}_k) + \sum_c P_c X_{ck} - \bar{X}_k \sum P_c \\ &= \sum_c (d_c - P_c)(X_{ck} - \bar{X}_k) + \bar{X}_k - \bar{X}_k \cdot 1 \\ &= \sum_c (d_c - P_c)(X_{ck} - \bar{X}_k). \end{aligned}$$

Note that $d_c - P_c$ can be viewed as a residual, so clearly here we have the makings of an orthogonality condition. If we arrange the $\mathbf{X}$ data for this one observation as

$$\mathbf{X} = \begin{bmatrix} x_{11} - \bar{x}_1 & \cdots & x_{1K} - \bar{x}_K \\ \vdots & & \vdots \\ x_{C1} - \bar{x}_1 & \cdots & x_{CK} - \bar{x}_K \end{bmatrix}$$

and let $\mathbf{e}$ represent a vector of residuals for this observation,

$$\mathbf{e} = \begin{bmatrix} d_1 - P_1 \\ \vdots \\ d_C - P_C \end{bmatrix}$$

then the observation's contribution to the score is $\mathbf{X}'\mathbf{e}$. Stacking the data observation by observation will then give the full score.

To obtain the information matrix, return to the first representation of the score contribution. For one observation i,

$$\begin{aligned}\frac{\partial \ln f_i}{\partial \beta_k} &= \sum_c d_c (X_{ck} - \sum_j X_{jk} P_j) \\ &= \sum_c d_c \left(X_{ck} - \sum_j X_{jk} \frac{e^{\mathbf{X}_j'\beta}}{\sum_l e^{\mathbf{X}_l'\beta}} \right).\end{aligned}$$

As we prepare to take second derivatives, we see that the only relevant portion of the above is

$$-\sum_c d_c \sum_j X_{jk} \frac{e^{\mathbf{X}_j'\beta}}{\sum_l e^{\mathbf{X}_l'\beta}} = -\sum_j X_{jk} \frac{e^{\mathbf{X}_j'\beta}}{\sum_l e^{\mathbf{X}_l'\beta}}.$$

Now we proceed to take the second derivative with respect to β_s. We obtain

$$\begin{aligned}-\frac{\partial^2 \ln f_i}{\partial \beta_k \partial \beta_s} &= \sum_j X_{jk} \left[\frac{(\sum_l e^{\mathbf{X}_l'\beta}) X_{js} e^{\mathbf{X}_j'\beta} - e^{\mathbf{X}_j'\beta} (\sum_l X_{ls} e^{\mathbf{X}_l'\beta})}{(\sum_l e^{\mathbf{X}_l'\beta})^2} \right] \\ &= \sum_j X_{jk} \ (X_{js} P_j - P_j \bar{X}_s) \\ &= \sum_j X_{jk} \ (X_{js} - \bar{X}_s) \ P_j.\end{aligned}$$

This is the same as

$$\sum_j (X_{jk} - \bar{X}_j) \ (X_{js} - \bar{X}_s) \ P_j.$$

The (k, s)-th element of the information matrix is formed by summing terms like the above across individuals.

21.4 The Independence of Irrelevant Alternatives

Consider a case with three possible choices,

$$\begin{aligned}U_0 &= \mathbf{X}_0'\beta + \epsilon_0 \\ U_1 &= \mathbf{X}_1'\beta + \epsilon_1 \\ U_2 &= \mathbf{X}_2'\beta + \epsilon_2.\end{aligned}$$

The ratio of the choice probabilities P_0 to P_1 is

$$\frac{P_0}{P_1} = \frac{e^{\mathbf{X}_0'\beta}}{e^{\mathbf{X}_1'\beta}} = e^{(\mathbf{X}_0 - \mathbf{X}_1)'\beta}.$$

Oddly enough, this is the same ratio as in the binary choice problem

$$\begin{aligned}U_0 &= \mathbf{X}_0'\beta + \epsilon_0 \\ U_1 &= \mathbf{X}_1'\beta + \epsilon_1,\end{aligned}$$

which also yields

$$\frac{P_0}{P_1} = e^{(\mathbf{x}_0 - \mathbf{x}_1)'\beta}.$$

So it seems that, at least as far as these ratios are concerned, the existence of Choice 2 is irrelevant. This disconcerting outcome, termed "independence of irrelevant alternatives" or IIA for short, is a serious weakness of the simple conditional logit specification.

To appreciate the difficulties, consider the famous transportation mode choice problem in which Choice 0 represents a car, Choice 1 a red bus, and Choice 2 a blue bus. If the consumer views the red bus and blue bus as being near-perfect substitutes, then when Choice 2 is available that should reduce P_1 by approximately half, making the ratio P_0/P_1 much different from the situation when only Choices 0 and 1 are available. The conditional logit model does not permit this.

21.5 Discussion and Example

An important feature of the conditional logit model is that it requires data on the characteristics of *all* alternatives considered by agent i, whether or not they were actually selected. Data such as these are not always available. For example, you may well be able to determine the price of a consumer durable that was purchased by consumer i but would not typically know the prices of other durables that were considered but not purchased. Sometimes, however, you have access to market-level or other aggregate data that can be used to calculate the prices of the nonpurchased items.

Also, in the otherwise excellent discussion of the conditional logit model given by Cameron and Trivedi (2005), the authors may leave you with the impression that *all* explanatory covariates must vary by choice alternative. This is certainly the case when the systematic component of utility is expressed linearly as $V_c = \mathbf{X}_c'\theta$, because if we attempted to augment this specification using agent-specific but alternative-invariant data $\mathbf{Z}_i$, via $V_{c,i} = \mathbf{Z}_i'\beta + \mathbf{X}_c'\theta$, the $\mathbf{Z}_i'\beta$ term would drop out of the choice probability. At first glance this would seem to rule out the use of individual income in models of the demand for durables, since income varies across people but not across durables.

However, when you step away from linear specifications and allow economic theory to suggest the form of the model, income can re-enter the picture. Suppose for illustration that the consumer faces the problem of choosing $c = 2, 3, 4$ units of a durable good, which has a price p per unit, and that all income remaining after the purchase of the durable goes to a composite consumption good. Further suppose that utility is Cobb–Douglas in the composite good and the durable, such that

$$\tilde{U}_{c,i} = (\Omega_i - p \cdot c)^{1-\alpha} \cdot c^{\alpha} \cdot e^{\epsilon_c}$$

with Ω_i being individual income. Taking logs,

$$U_{c,i} = (1-\alpha)\ln(\Omega_i - p \cdot c) + \alpha \ln c + \epsilon_c$$

in which the systematic component is $V_{c,i} = (1-\alpha)\ln(\Omega_i - p \cdot c) + \alpha \ln c$. In this model, the individual-specific income variable is bound up with the choice alternative.

Chapter 22

Poisson Models

Students: Supplement this chapter with Cameron and Trivedi (2005, Chapter 20). You can skim sections 20.4.2, 20.4.3, 20.5 and 20.6 of that chapter.

Poisson models are commonly used in studies of the number of events occuring over a given period of observation, typically when only a few such events are likely. The model is often motivated by an assumption that the waiting times between events are independent and exponentially distributed. In a study of workplace injuries in a given plant, for instance, we could assume that the time between injuries is distributed in this way, leading to a Poisson representation of the number of injuries over an observation period. In empirical industrial organization, the Poisson model is often used to study the number of patent applications submitted by a firm over a given period of time.

22.1 Likelihood and Score Vector

In the Poisson model, the probability associated with $Y_i = y$ events over an observation period of duration d, given covariates $\mathbf{X}_i$ and parameters β, is

$$\Pr(y \mid \mathbf{X}_i, \beta) = \frac{e^{-\lambda_i}\lambda_i^y}{y!} = P_i(y)$$

with $\lambda_i = \exp(\mathbf{X}_i'\beta + \ln d)$. The expected value of Y_i conditional on $\mathbf{X}_i$ is λ_i, and this is also the conditional variance. The fact that the Poisson mean must equal the variance is restrictive, and has motivated a search for alternative models as Cameron and Trivedi explain.

We have for the i-th observation

$$\frac{\partial \ln P_i}{\partial \beta} = (y_i - \lambda_i) \cdot \mathbf{X}_i,$$

a $k \times 1$ vector. When this derivative is evaluated at the true β_0, it has expectation zero as we have come to expect. The ML first-order condition, obtained by setting the overall score to

zero, is

$$\frac{\partial L_n}{\partial \beta} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{X}_i(y_i - \hat{\lambda}_i) = \mathbf{0}.$$

Just as in the linear regression and logit models, we have orthogonality between the columns of $\mathbf{X}$ and the residual vector $\mathbf{y} - \hat{\lambda}$.

22.2 Predicting the Probability of an Event

Let $P_i(1, \hat{\beta})$ represent the predicted probability that one event occurs over duration d, and consider its variance. A Taylor expansion yields

$$\frac{\partial P_i(1, \beta)}{\partial \beta} = P_i(1, \beta)\left(1 - \lambda_i(\beta)\right)\mathbf{X}_i$$

and this implies

$$\sqrt{n}\left(P_i(1, \hat{\beta}) - P_i(1, \beta_0)\right) = P_i(1, \beta_0)\left(1 - \lambda_i(\beta_0)\right) \cdot \mathbf{X}_i' \sqrt{n}\left(\hat{\beta} - \beta_0\right).$$

Let $\tilde{\mathcal{J}}$ be the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta_0)$, and let $(n\hat{\mathcal{J}})^{-1}$ be the inverse of the information matrix that STATA provides in the `e(V)` matrix following estimation. Then the finite sample variance is approximately

$$\text{Var } P_i(1, \hat{\beta}) = P_i(1, \hat{\beta})^2 \cdot \left(1 - \lambda_i(\hat{\beta})\right)^2 \cdot \mathbf{X}_i'(n\hat{\mathcal{J}})^{-1}\mathbf{X}_i.$$

In STATA, an option to the Poisson regression procedure calculates the square root of $\mathbf{X}_i'(n\hat{\mathcal{J}})^{-1}\mathbf{X}_i$, so the full expression above is easy to program.

Chapter 23

Hazard-Rate and Multiple-State Models

Students: Supplement this chapter with Cameron and Trivedi (2005, Chapters 17 and 19).

Let us consider the distribution of a non-negative continuous random variable T that represents the waiting time between two events. The classic demographic example is the time between birth and death, with T being the length of life. In economic contexts, waiting times of interest include the length of spells of unemployment or the interval between the time when a firm submits one patent application and the time it submits the next application. Cameron and Trivedi (2005) give an excellent textbook presentation of the issues involved in analyzing waiting times. What follows should complement their work.

The distribution of T can be described in varying degrees of detail. For some purposes, we require only a measure of central tendency, that is, a mean or median, to summarize the distribution in question. For other purposes, however, more detail is needed than can be captured in a single summary measure, and we may want to inspect the probability density function of T. Like the density, the hazard function gives a detailed description of one aspect of T's distribution.

We first ask what information the hazard function provides that is not already readily apparent in the density function. Next, we will ask why this additional information is necessary or useful in economic and demographic modelling.

We begin with the relationships between the hazard function and the density. Let the density function for T be given by $f(t)$ and the cumulative distribution function, representing $\Pr(T \leq t)$, by $F(t)$. Then the hazard or risk function $r(t)$ is defined as

$$r(t) = \frac{f(t)}{1 - F(t)}. \tag{23.1}$$

That is, the level of the hazard function when evaluated at t is simply the ratio of the density $f(t)$ to the area under the density function to the right of t.

In Figure 23.1 we show the density and the hazard function of one waiting-time distribution, the Weibull, that is often used in economic and demographic studies. Point a is the

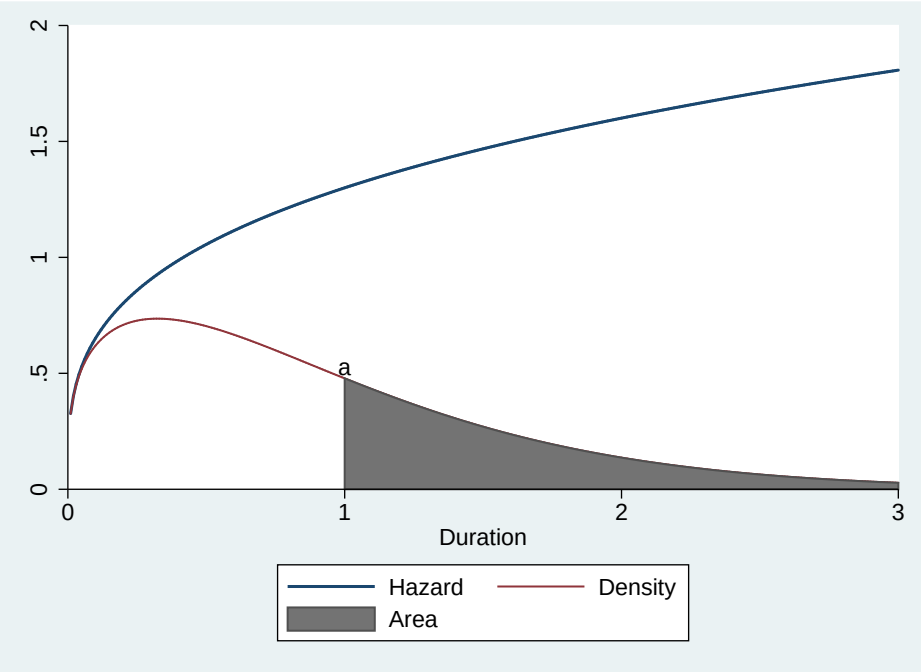


Figure 23.1: Density and hazard rate for the Weibull model, with $\alpha = 1.3$.

level of the density at $t = 1$, and the shaded area is $1 - F(1)$. The hazard function evaluated at the point $t = 1$ is the ratio of point a to this area.

A more informative approach to the hazard function is through *conditional* probabilities. A glance at equation (23.1) above makes it clear that the hazard function, when evaluated at t, is related to the chance of "failure" in the small interval $(t, t + \delta)$, given that failure has not occurred as of time t. Put more formally,

$$r(t) = \lim_{\delta \to 0} \frac{\Pr\left(t < T \leq t + \delta\right) / \delta}{\Pr\left(t < T\right)}.$$

Let $S(t)$ denote the survival probability $1 - F(t)$. Then we have $S'(t) = -f(t)$ and

$$r(t) = -\frac{S'(t)}{S(t)}.$$

This is a simple differential equation which has the following solution,

$$S(t) = \exp\left(-\int_0^t r(v)dv\right). \tag{23.2}$$

Equation (23.2) implies

$$f(t) = r(t)e^{-\int_0^t r(v)dv} = r(t)S(t). \tag{23.3}$$

In short, equation (23.1) shows us how to how to calculate the hazard function r given the density function f (or cdf F); equation (23.3) shows us how to calculate the density function

given the hazard. Hazards and densities describe different aspects of the distribution under study, but they are interchangeable in the sense that one can be derived from the other.

An often-used relationship links the survival function $S(t)$ to the mean of T,

$$\mathrm{E}\,T = \int_0^\infty tf(t)dt = \int_0^\infty S(t)dt. \tag{23.4}$$

We're familiar with this expression in demography via the intuitive connection between the expectation of life at birth and the concept of "total years lived" in a population. The formal proof of the relationship can be developed either through integration by parts or by changing the order of integration.

The integration-by-parts method is as follows. You may recall that if we have functions $A(t)$ and $B(t)$, and let $a(t)$ and $b(t)$ be defined, respectively, as $a(t) = A'(t)$, $b(t) = B'(t)$, we obtain

$$\int_0^x \frac{d}{dt}\Big(A(t)B(t)\Big)dt = \int_0^x a(t)B(t)dt + \int_0^x A(t)b(t)dt.$$

From this,

$$\int_0^x A(t)b(t)dt = A(t)B(t)|_0^x - \int_0^x a(t)B(t)dt.$$

Make the analogies $t = A(t)$ and $1 = a(t)$; also, $f(t) = b(t)$ and $F(t) = B(t)$. Then

$$\begin{aligned}\int_0^x tf(t)dt &= tF(t)|_0^x - \int_0^x F(t)dt = xF(x) - \int_0^x F(t)dt \\ &= \int_0^x S(t)dt - xS(x).\end{aligned}$$

Taking the limit as $x \to \infty$, and assuming that the mean exists, which implies $\lim_{x\to\infty} xS(x) = 0$, we obtain the result that $\mathrm{E}\,T = \int_0^\infty S(t)dt$.

A more elegant approach involves changing the order of integration. We rewrite the mean as

$$\mathrm{E}\,T = \int_0^\infty tf(t)dt = \int_0^\infty \left(\int_0^t dv\right) f(t)dt$$

and then change the order of integration, such that

$$\int_0^\infty \int_0^t f(t)dvdt = \int_0^\infty \left(\int_v^\infty f(t)dt\right) dv.$$

The last expression is recognizable as $\int_0^\infty S(v)dv$.

23.1 Why Model the Hazard Rate?

Why should any substantive investigation, whether of mortality, birth intervals, lengths of employment or the like, be couched in terms of the hazard function rather than in more familiar terms? There are two reasons, one stemming from practical considerations and the other from substantive and theoretical concerns, for regarding the hazard function as a fundamental concept of interest.

Right-censored data

It is very often the case that in a sample of individuals at risk of experiencing an event, some do and some do not experience that event within the period in which they are under observation. For instance, if our interest concerns female age at first marriage and we have data drawn from a retrospective survey of women, some of the women in our sample will have had a marriage as of the survey date, whereas others will be recorded as never-married at survey. Let T represent age at first marriage. The ever-married women in the sample provide us with the relatively precise information $T = t_i$, where t_i is the i-th woman's age at marriage. Those who are never-married at survey are effectively right-censored, that is, they contribute the information $T > \tau_i$, where τ_i is the i-th woman's age at survey.

It should be clear that both kinds of information must be combined in some fashion if we are to obtain an accurate picture of the underlying distribution of age at marriage. If, for instance, we analyzed only the closed or uncensored intervals, our sample would be composed of intervals that are shorter than is typical in the population. (In the case at hand, the average age at marriage in the sample would be too low, in relation to the true average.) A sample of closed intervals is inherently biased, unless the period of observation is long enough for everyone in the sample to have made the transition in question.

When statistical inference is based on the principle of maximum likelihood, as it ususally is in the case of hazard models, the solution to the right-censoring problem is easy and straightforward. The density $f(t)$ is taken to be the contribution made to the sample likelihood by those who make a transition (e.g., marry) at age t, whereas $1 - F(\tau)$ is the contribution of observations that are right-censored (e.g., not yet married) at age τ. It can be shown that the combination of these two types of data resolves the censoring problem and permits the underlying distribution to be estimated without large-sample bias.

Now, although we made use of the density f and the survivor function $1 - F$ in the discussion above, we nowhere made reference to the concept of the hazard rate. Indeed, so long as we have closed-form or easily calculable expressions for f and F (as we do for the exponential, log-normal, gamma, Weibull, log-logistic, and other waiting-time distributions) there is not the slightest need to resort to hazard functions as such to solve a right-censoring problem. Nevertheless, since to estimate a hazard model one must always calculate

$$1 - F(t) = S(t) = e^{-\int_0^t r(v)dv},$$

the hazard approach always supplies us with the ingredients needed to solve a right-censoring problem. From this perspective, the hazard model approach is convenient in the presence of right-censoring, even if it is not strictly necessary.

Time-varying covariates

In my view, the fundamental rationale for the use of hazard rate models has less to do with right-censoring than with the theoretical and substantive importance of time- or age-varying covariates. It is very difficult to understand the way such covariates affect waiting times without recourse to the concept of a hazard function. It is equally difficult to estimate the effects of time-varying variables outside of the hazards-model context. Perhaps an example will make this clear.

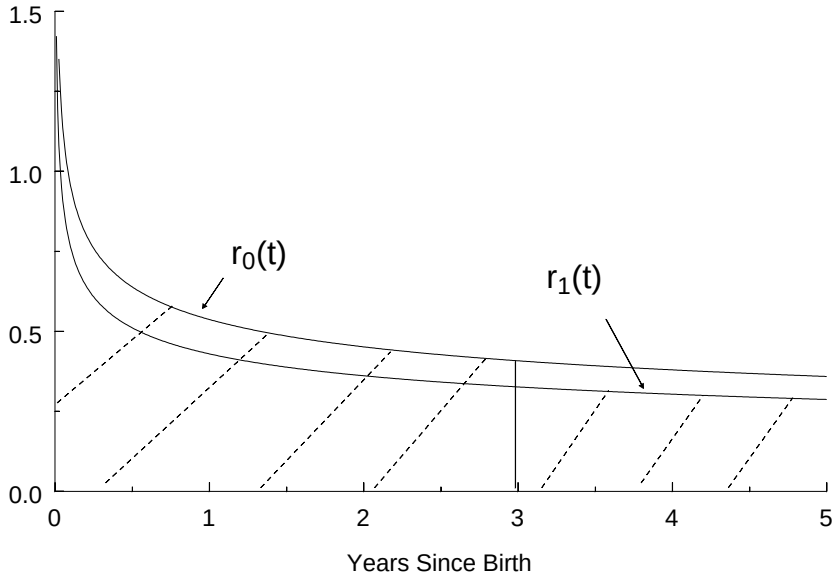


Figure 23.2: Urban (r_1) and rural (r_0) mortality hazard functions

Let T represent the length of life, and focus on infant and child survival in a developing country. Suppose that mortality risks differ between urban and rural environments. A child is born into a particular family at time 0, at which time the family lives in a rural area. Some 3 years after the child's birth ($t = 3$), the family moves to an urban area where mortality risks are lower. We are interested in the effect of this move on the probability that the child lives to be age 5.

Imagine a sample of such families and children. How can the impact of the change in location be estimated? It would be inappropriate to assess the impact of rural-urban moves by regressing a dependent variable "did the child survive to age 5" on an explanatory variable "did a rural to urban move take place?" There can be no effect if the child has already died by the time of the move. Neither is it appropriate to substitute the explanatory variable "did a move take place while the child was alive". Why? The longer a child was alive, the greater the time span in which a move could take place with the child alive—the explanatory variable is logically confounded with the dependent variable. Clearly a defensible estimation strategy must keep track of the timing of events and remove those individuals who are no longer at risk of a transition. This is precisely what a hazards model accomplishes.

Let the location of the family at point t be indexed by the dummy variable $x(t)$, where $x(t) = 1$ indicates urban residence and $x(t) = 0$ rural residence. The hazard function for point t (recall that t is the duration since the child's birth, or age provided that the child is still alive) is $r_0(t)$ if $x(t) = 0$ or $r_1(t)$ if $x(t) = 1$. We can assume that $r_0(t) > r_1(t)$.

Now the question of the effect of a move can be addressed with precision. Consider Figure 23.2. The probability of survival to age 5, given a change in the family's residence

three years after the child's birth, is

$$\exp\Big(-\left(\int_0^3 r_0(v)dv + \int_3^5 r_1(v)dv\right)\Big),$$

the magnitude of which is indicated by the shaded area under the respective hazard functions. Had the family not moved, leaving the child exposed to the risk $r_0(t)$ throughout childhood, the probability of survival to age 5 would have been

$$\exp\left(-\int_0^5 r_0(v)dv\right).$$

The difference between these two expressions gives the net effect of the move. Note that when we set out the problem in terms of hazard functions, we impose the logical condition that a change in environment at duration t can matter only for those who survive to experience the change.

Many researchers turn to hazard models out of the feeling that time-varying, systematic influences on risks must be important, but find that they lack the detailed data on time-varying covariates to estimate such influences directly. For a demographic example, consider an analysis of birth intervals in a society characterized by periods of postpartum sexual abstinence. One might be able to obtain data on intervals between births, but very often one would not have information on the point within each birth interval at which sexual relations were resumed. In this example, the time-varying covariate $x(t)$ would indicate whether sexual relations have resumed as of duration t in the interval. Even if $x(t)$ cannot itself be measured, there are suggestions of its influence in the upward slope of a birth interval hazard function. Just as a time trend in a regression model hints at systematic but unmeasured changes in fundamental variables, so the slope of the hazard function may suggest systematic variation in time-varying covariates.

The association between age and mortality risks provides another example. For as long as demography has been a science, demographers have explored the relationship between age a and the force of mortality $\mu(a)$, which is the name demographers give to the hazard function. But it is the rare study that can probe into this relationship, to identify the systematic changes in physiological variables that give rise to the overall age pattern of mortality. The age dependence displayed by $\mu(a)$ is nonetheless interesting for what it implies about these more fundamental age-varying variables.

Hazard functions with $r(t)$ non-constant with respect to t are said to exhibit *duration dependence*. When time-varying covariates enter the specification of the hazard, giving $r(t, x(t))$, duration dependence is said to exist if $\partial r/\partial t \neq 0$, where the derivative is taken with respect to t alone. The only constant hazard model is the exponential waiting-time model, which has enjoyed a prominent position in both demographic and economic analyses.

23.2 Estimation of Hazard-rate Models

In getting a sense of what sort of hazard model to estimate, a useful first step is to estimate the empirical survivor function using the life-table model (discussed below) with subintervals specified to be as short as possible, and then use the results to graph the empirical

hazard function. STATA does this for you in the `sts graph, hazard` command. This will give you an idea of the functional form the hazard is likely to take. You can then choose among a variety of parametric functional forms, or if all these appear inadequate to the task, you can generalize the life-table model to include covariates. We begin by considering the simplest functional form for the hazard—the exponential model—and show how it can be estimated in models without explanatory covariates. We then describe the life-table model.

The exponential model

The exponential model is a benchmark duration distribution with $r(t) = r$, giving $S(t) = e^{-rt}$ and $f(t) = re^{-rt}$. The percentiles of the distribution are easily derived from the survivor function. If we let s represent the proportion still surviving and ask to what duration t_s this corresponds, we can see that $t_s = -\ln s/r$. For instance, the median duration is $t_{.5} = \ln 2/r$. The mean of an exponential variable is easily derived from its survivor function,

$$ET = \int_0^\infty e^{-rt}dt = \frac{1}{r}.$$

Now, suppose we have a sample of N independent observations generated by an exponential distribution, NC of which make transitions at times $(t_1, t_2, \ldots, t_{NC})$ and C of which are right-censored at times $(\tau_1, \tau_2, \ldots, \tau_C)$. Each non-censored observation contributes $f(t_i) = r\exp(-rt_i)$ to the sample likelihood, whereas each right-censored observation contributes $S(\tau_i) = \exp(-r\tau_i)$ to the likelihood. The full log-likelihood, combining both types of observations, is simply

$$L(r) = NC \cdot \ln r - r\sum_{i=1}^{NC} t_i - r\sum_{i=1}^{C} \tau_i.$$

To solve for the maximum-likelihood estimate $\hat{r}$, we differentiate $L(r)$ and find the value of r such that $L'(r)$ equals zero,

$$\hat{r} = \frac{NC}{\sum t_i + \sum \tau_i}.$$

To demographers this expression is very familiar: It is the ratio of the total number of transitions (NC) to the total amount of "exposure" to the risk of transition. Note that if the observation period is very long, so that all observations make the transition, NC becomes equivalent to the sample size and $\sum \tau_i = 0$. In this case $\hat{r}$ is no more than the inverse of the sample mean, just as we would expect.

Another way to understand the general formula for $\hat{r}$ is to consider the probability expressions to which the numerator and denominator would converge in large samples. This is most easily done for the special case where all right-censoring takes place at a single duration τ. Then, given NC non-censored cases and C censored cases, the estimator becomes

$$\hat{r} = \frac{NC}{\sum t_i + C \cdot \tau}.$$

Dividing through by the full sample size $N = NC + C$, we have

$$\hat{r} = \frac{NC/N}{\frac{NC}{N}\frac{1}{NC}\sum t_i + \frac{C}{N}\tau}.$$

The numerator NC/N is the fraction of the sample observations that make a transition before the censoring point τ. This ratio will converge to $F(\tau) = 1 - S(\tau)$. The denominator converges to

$$(1 - S(\tau))\,\mathrm{E}(T|T < \tau) + S(\tau)\tau,$$

and after some manipulation, the denominator can be shown to reduce to $r^{-1}(1 - S(\tau))$. Thus, taking both numerator and denominator into account, we find that $\hat{r} \xrightarrow{p} r$.

The Life Table Model

The life table is, in essence, a piece-wise exponential duration model. We divide the range of T into sub-intervals, which need not be equal in length, and assume that within each sub-interval the hazard rate is constant. The hazard function $r(t)$ is a step function, equalling r_j for $t_{j-1} \leq t < t_j$ with $t_0 \equiv 0$. Within each subinterval the density resembles that of an exponential model, but the hazard rate is allowed to vary across sub-intervals so as to represent in a flexible way the varying risks of transition with duration.

To see how such a model might be estimated, consider a mortality problem with sample of 6 people. Person 1 dies at age 4; person 2 dies at age 55; person 3 dies at age 73; person 4 dies at age 2; person 5 is right-censored at age 61; and person 6 is right-censored at age 85. We assume that from birth (age 0) to exact age 5 the hazard rate is r_1; from age 5 to age 65 it is r_2; and for older ages it is r_3.

The table shows the likelihood contribution made by each observation.

Person	Probability Expression
Dies at age 4	$r_1 \exp(-r_1 4)$
Dies at age 55	$\exp(-r_1 5) r_2 \exp(-r_2(55-5))$
Dies at age 73	$\exp(-r_1 5) \exp(-r_2(65-5)) r_3 \exp(-r_3(73-65))$
Dies at age 2	$r_1 \exp(-r_1 2)$
Censored at age 61	$\exp(-r_1 5) \exp(-r_2(61-5))$
Censored at age 85	$\exp(-r_1 5) \exp(-r_2(65-5)) \exp(-r_3(85-65))$

If we take logs of each person's contribution to the sample likelihood, sum across people, and group the results according to the parameters r_1, r_2 and r_3, we obtain

$$L(r) = 2\ln r_1 - 26r_1 + \ln r_2 - 226r_2 + \ln r_3 - 28r_3.$$

Taking partial derivatives of L with respect to each of these parameters and solving, we find that $\hat{r}_1 = 2/26$, $\hat{r}_2 = 1/226$, and $\hat{r}_3 = 1/28$. In each instance, the estimate of the hazard rate is given by the ratio of deaths in the sub-interval to which the hazard corresponds, to the total exposure in the sub-interval. When viewed on a sub-interval by sub-interval basis, the results remind us of the simple exponential case.

There's really no more to the underlying concepts of the life table than this. Of course, if the data handed to us do not contain information on exact ages at death, but rather give the results in terms of broad age groups, complications can arise. The conventional treatment of the life table gives a great deal of emphasis to such problems of aggregation. For instance, we may have information on total deaths occurring in infancy, but not on the ages at death.

This raises two issues: (i) we cannot precisely calculate exposure to the risk of death in the first year of life, even if the underlying hazard rate could be assumed to be constant over the year; and (ii) the hazard itself is non-constant, being much higher in the neonatal period than in the post-neo-natal. These aggregation questions have been considered at mind-numbing length by actuaries and demographers, who have developed various rules of thumb that allow the available data to be translated into estimates of the underlying life table concepts.

23.3 Competing-Risk Models

We focus intially on competing-risk models as examples of the more general class of multiple-state increment-decrement models. A competing-risk model is characterized by a single origin state and multiple destination states $j = 1, \ldots, J$. These destination states are defined so as to be mutually exclusive, and taken together, they must exhaust the set of possible destinations. We can think of these J states as representing mutually exclusive "causes of death."

Associated with each such destination is a *transition intensity* $r_j(t)$, which approximates the concept "among those surviving in the origin state to exact duration t, what is the probability of exit to destination state j in the next small instant of time?" Like hazard rates, transition intensities are more formally defined as instantaneous probability densities.

Since the hazard rate $r(t)$ represents the risk of an exit of any kind, and the transition intensities are associated with mutually exclusive events, we must have

$$r(t) = \sum_{j=1}^{J} r_j(t),$$

and the survivor function $S(t)$ is itself

$$\begin{aligned} S(t) &= e^{-\int_0^t r(v)dv} \\ &= e^{-\int_0^t r_1(v)dv} e^{-\int_0^t r_2(v)dv} \cdots e^{-\int_0^t r_J(v)dv} \\ &= S_1(t)S_2(t)\cdots S_J(t) \end{aligned}$$

where the $S_j(t)$ are pseudo-survivor functions representing the hypothetical probability of survival in the presence of only the j-th cause of death.

The unconditional density of death at duration t due to cause k is

$$f_k(t) = r_k(t)S(t) = r_k(t)e^{-\int_0^t \sum_{j=1}^{J} r_j(v)dv}$$

and this is analogous to the unconditional density of death at duration t due to any cause, $f(t) = r(t)S(t)$. Notice, however, that the levels and age patterns of *all* causes of death figure into the expression for deaths due to a specific cause.

In cause-specific life tables for mortality, we find expressions for the probability that a newborn will eventually die of cause k,

$$\int_0^\omega f_k(t)dt = \int_0^\omega r_k(t)S(t)dt,$$

the probability of death due to cause k among survivors to exact age x,

$$\frac{\int_x^\omega f_k(t)dt}{S(x)},$$

the expected age at death, given that one dies of cause k,

$$\frac{\int_0^\omega t f_k(t)dt}{\int_0^\omega f_k(t)dt},$$

and so on, where ω represents the oldest age to which anyone could live.

23.4 Estimation of Competing-Risk Models

All these expressions can be estimated if one has enough information on exposure times, ages at death and causes of death to estimate the $r_j(t)$ transition intensities. Often it proves necessary to aggregate all causes but the one of greatest interest into a composite category, labelled "all other causes," to carry out the estimation.

Suppose that one has a sample in which only two causes of death are possible. The transition intensities for Cause 1 and Cause 2 are, respectively,

$$\begin{aligned} r_1(t \mid \theta) &= \theta \\ r_2(t \mid \beta) &= \beta_1 \beta_2 t^{\beta_2 - 1} \end{aligned}$$

so that the risk of Cause 1 is independent of age whereas the risk of Cause 2 varies in a Weibull-like manner. As the researcher, your aim is to estimate the parameters θ, β_1, β_2.

One key estimation issue can be seen in the hypothetical case of three observations. Person 1 dies of Cause 1 at t_1. Person 2 dies of Cause 2 at t_2. Person 3 survives to the end of the observation period, at which point she is age t_3. The sample likelihood is

$$\mathcal{L} = r_1(t_1)S(t_1) \cdot r_2(t_2)S(t_2) \cdot S(t_3).$$

This can be re-expressed using pseudo-survivor functions as

$$\begin{aligned} \mathcal{L} &= r_1(t_1)S_1(t_1)S_2(t_1) \cdot r_2(t_2)S_1(t_2)S_2(t_2) \\ &\quad \cdot S_1(t_3)S_2(t_3). \end{aligned}$$

From this it is clear that we can arrange the likelihood factors into two distinct groups according to the parameters that are involved, such that $\mathcal{L} = \mathcal{L}_\theta \cdot \mathcal{L}_\beta$. The factors involving θ, the parameter governing the risk of death due to Cause 1, are

$$\mathcal{L}_\theta = r_1(t_1)S_1(t_1)S_1(t_2)S_1(t_3).$$

If we happened to be interested only in θ, then we could ignore the remaining factors of the likelihood function, which involve only the β parameters. (The likelihood function is *strongly separable* in θ and β.) Indeed, an inspection of the expression just above shows that we can estimate θ by coding the data as if deaths to Cause 2 were, instead, right-censored with the censoring point equal to the age at death. This separability property is an enormous convenience.

23.5 General Multiple-State Models

It is a short step from competing-risk models to more general models in which individuals can circulate among states. Consider the case of child health and survival and define the states as follows:

- *State 1*: The child in question is dead;
- *State 2*: The child is alive, but in ill health;
- *State 3*: The child is alive and healthy.

Associated with these three states are three transition intensities: $r_{32}(t)$ is the intensity associated with a move from good health to ill health at duration t since birth; $r_{23}(t)$ is the intensity associated with recovery from illness; and $r_{21}(t)$ is associated with the transition from illness to death. (We have assumed that healthy children must pass through a stage of illness before they die, but this assumption could be relaxed to handle deaths by accident and so forth. We have further assumed, for simplicity, that all children are born in a state of good health, although this, too, could be generalized.) A child could make the passage from a state of good health to ill health and back to good health many times, although there is unfortunately no return from death, the ultimate "absorbing state" in the terminology of Markov chains.

With this apparatus, consider the probability expression that would be attached to the following history. A child has a first episode of illness at age t_1, recovers her good health at age t_2, falls ill again at t_3 and then dies at t_4. The probability expression associated with this sequence is

$$r_{32}(t_1)e^{-\int_0^{t_1} r_{32}(v)dv} \cdot r_{23}(t_2)e^{-\int_{t_1}^{t_2}(r_{21}(v)+r_{23}(v))dv}$$
$$\cdot r_{32}(t_3)e^{-\int_{t_2}^{t_3} r_{32}(v)dv} \cdot r_{21}(t_4)e^{-\int_{t_3}^{t_4}(r_{21}(v)+r_{23}(v))dv}.$$

Note that there are two possible destination states for a child in ill health, but only one destination state for a child in good health.

When the model is expressed in these simple terms, the risks of contracting illness and effecting a recovery from it, and likewise the risks of dying, are affected by the age of the child, t, but not by the history of previous illnesses that the child has experienced. A more realistic model would certainly take into account such "history dependence," using the number and duration of previous illnesses as explanatory factors that enter the specification of the transition intensities $r_{ij}(t)$.

Part IV

Generalizations of the Linear Model

Chapter 24

Generalized Least Squares

Students: Supplement this chapter with Cameron and Trivedi (2005, Sections 4.4, 4.5).

We return in this chapter to the linear model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, again making the strong assumption $\mathrm{E}(\epsilon \mid \mathbf{X}) = \mathbf{0}$ about the conditional means of the disturbance terms, but weakening what is assumed about their variances. Here we allow $\mathrm{E}(\epsilon\epsilon' \mid \mathbf{X}) = \mathbf{V}$, which some writers describe (a bit confusingly) as a *non-scalar covariance matrix* to distinguish it from $\sigma^2\mathbf{I}$, which they describe (no less confusingly) as a scalar covariance matrix. We first explore the implications of this new assumption for the OLS estimator and then investigate the properties of the generalized least squares (GLS), estimated generalized least squares (EGLS), and maximum likelihood approaches to estimation.

24.1 What Goes Wrong with OLS?

Write the OLS estimator as $\hat{\beta} = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$. By using iterated expectations and the fundamental assumption $\mathrm{E}(\epsilon \mid \mathbf{X}) = \mathbf{0}$, we obtain $\mathrm{E}\,\hat{\beta} = \beta$, that is, the OLS estimator is unbiased. However, the new assumption about the variance of ϵ alters the form of the variance matrix of the OLS estimator, as can be seen from

$$\mathrm{E}(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = \mathrm{E}\left((\mathbf{X}'\mathbf{X})^{-1} \cdot \sum_{i=1}^{n}\sum_{j=1}^{n} \epsilon_i\epsilon_j\mathbf{X}_i\mathbf{X}_j' \cdot (\mathbf{X}'\mathbf{X})^{-1} \right).$$

An application of iterated expectations gives us the result that

$$\mathrm{Var}\,\hat{\beta} = \mathrm{E}\left((\mathbf{X}'\mathbf{X})^{-1} \cdot \sum_{i}\sum_{j} v_{ij}\mathbf{X}_i\mathbf{X}_j' \cdot (\mathbf{X}'\mathbf{X})^{-1} \right)$$

or, in a more compact notation, $\mathrm{Var}\,\hat{\beta} = \mathrm{E}((\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1})$. Clearly, any hypothesis test based on the mistaken assumption that $\mathrm{Var}\,\hat{\beta} = \sigma^2\,\mathrm{E}(\mathbf{X}'\mathbf{X})^{-1}$ will be incorrectly formulated.

To investigate the consistency of the OLS estimator, we can weaken our core assumptions to $\mathrm{E}(\epsilon_i|\mathbf{X}_i) = 0$ and $\mathrm{E}(\epsilon_i\epsilon_j \mid \mathbf{X}_i, \mathbf{X}_j) = v_{ij}$. Consider

$$\operatorname{plim}\hat{\beta} = \beta + \mathbf{W}_{xx}^{-1} \operatorname{plim}\frac{1}{n}\mathbf{X}'\epsilon,$$

assuming $(n^{-1}\mathbf{X}'\mathbf{X})^{-1} \xrightarrow{p} \mathbf{W}_{xx}^{-1}$ as usual. Now, $\mathrm{E}\, n^{-1}\mathbf{X}'\epsilon = \mathbf{0}$ and

$$\operatorname{Var}\frac{1}{n}\mathbf{X}'\epsilon = \frac{1}{n^2}\mathrm{E}\sum_i\sum_j \epsilon_i\epsilon_j\mathbf{X}_i\mathbf{X}_j' = \frac{1}{n^2}\mathrm{E}\sum_i\sum_j v_{ij}\mathbf{X}_i\mathbf{X}_j'.$$

Note that there are n^2 terms in the double summation, which can can be written as $\mathrm{E}(\mathbf{X}'\mathbf{V}\mathbf{X})$. If $\mathrm{E}(\mathbf{X}'\mathbf{V}\mathbf{X})$ is of order $o(n^2)$, then we have $\operatorname{plim}\hat{\beta} = \beta$ by the convergence in mean square criterion. However, as Greene (2003) warns, the $o(n^2)$ condition may not be satisfied.

24.2 The GLS Method

In many treatments of the GLS method, the covariance matrix of the disturbances $\mathrm{E}(\epsilon\epsilon'|\mathbf{X}) = \sigma^2\mathbf{V}$, with σ^2 being an unknown scale factor that will be estimated, and $\mathbf{V}$ being a known $n \times n$ positive definite matrix. Let's follow this conventional approach. Since $\mathbf{V}$ is positive definite, we can find a matrix $\mathbf{V}^{-1/2}$ such that $\mathbf{V}^{-1/2}\mathbf{V}^{-1/2} = \mathbf{V}^{-1}$. Pre-multiplying $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ by $\mathbf{V}^{-1/2}$, we obtain

$$\mathbf{V}^{-1/2}\mathbf{Y} = \mathbf{V}^{-1/2}\mathbf{X}\beta + \mathbf{V}^{-1/2}\epsilon.$$

Note that

$$\operatorname{Var}\mathbf{V}^{-1/2}\epsilon = \sigma^2\mathbf{V}^{-1/2}\mathbf{V}\mathbf{V}^{-1/2} = \sigma^2\mathbf{I},$$

that is, the transformed disturbance term has a "scalar" covariance matrix. The GLS estimator is defined in terms of the transformed data,

$$\tilde{\beta} = \left(\mathbf{X}'\mathbf{V}^{-1/2}\mathbf{V}^{-1/2}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1/2}\mathbf{V}^{-1/2}\mathbf{Y} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y},$$

from which σ^2 has cancelled out, and we can also write it as

$$\tilde{\beta} = \beta + \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\epsilon.$$

From this and the $\mathrm{E}(\epsilon \mid \mathbf{X}) = \mathbf{0}$ assumption it is evident that $\mathrm{E}\,\tilde{\beta} = \beta$ and $\operatorname{Var}\tilde{\beta} = \sigma^2\,\mathrm{E}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. Provided that $\operatorname{plim} n^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \equiv \mathbf{Q}^{-1}$ exists, the GLS estimator $\tilde{\beta}$ is consistent, and with suitable additional assumptions, we have asymptotic normality in the sense that $\sqrt{n}(\tilde{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$.

Define the GLS residual $\tilde{\mathbf{e}} = \mathbf{V}^{-1/2}(\mathbf{Y} - \mathbf{X}\tilde{\beta})$. The estimator of σ^2 is $\tilde{\sigma}^2 = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}/(n-k)$, for which it can be shown that $\mathrm{E}\,\tilde{\sigma}^2 = \sigma^2$. Furthermore, $\operatorname{plim}\tilde{\sigma}^2 = \sigma^2$. The proofs are identical to those we used for the simple OLS model—all we need to do is to apply the same methods to the transformed data.

As you would expect, it can be shown that the difference between $\text{Var}\,\hat{\beta}$ of the OLS estimator and $\text{Var}\,\tilde{\beta}$ of the GLS estimator is a positive semidefinite matrix; in other words, $\tilde{\beta}$ is more efficient than $\hat{\beta}$. Indeed, it is more efficient than any other linear unbiased estimator. The proof of this is a variation on the Gauss–Markov proof we saw a few chapters ago. Let $\tilde{\beta} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} \cdot \mathbf{Y}$ and express the alternative estimator as

$$\bar{\beta} = \left((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1} + \mathbf{B}\right)\mathbf{Y}$$

for some matrix $\mathbf{B}$ that is an unspecified function of the $\mathbf{X}$ covariates. Following the logic of the earlier proof, we see that we must have $\mathbf{BX} = \mathbf{0}$ for unbiasedness, and then, conditional on the $\mathbf{X}$ variables,

$$\text{Var}\,\bar{\beta} = \sigma^2\left((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} + \mathbf{BVB}'\right),$$

from which it is obvious that $\mathbf{B}$ must be set to zero to produce the smallest possible variance matrix.

A few remarks are in order. First, we have presented $\mathbf{V}^{-1}$ as if it could be easily obtained from $\mathbf{V}$, but for samples of the size often encountered in economics—in the thousands of observations—there are formidable numerical difficulties to be faced in calculating $\mathbf{V}^{-1}$. The easily-solved cases are those in which $\mathbf{V}$ is so highly structured that its inverse can be obtained analytically. As we'll see, a time-series model in which the disturbances are first-order autoregressive is one of the cases with an analytic solution. Another easy case is that of heteroskedastic disturbances with zero covariances between ϵ_i and ϵ_j for $i \neq j$. But in general, the problem of obtaining the inverse can be numerically very difficult.

Second, it may be helpful to work through two alternative derivations of the GLS estimator, one based on minimization of a (weighted) sum of squares and the other based on the method of moments, much as we did in deriving the ordinary least squares estimator. If we pose the problem in terms of minimizing a sum of squares,

$$\tilde{\beta} = \arg\min\,(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\beta)$$

and factor $\mathbf{V}^{-1} = \mathbf{V}^{-1/2}\mathbf{V}^{-1/2}$, the problem can be re-expressed as

$$\arg\min\,(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta)'(\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta)$$

with $\tilde{\mathbf{Y}} = \mathbf{V}^{-1/2}\mathbf{Y}$ and $\tilde{\mathbf{X}} = \mathbf{V}^{-1/2}\mathbf{X}$. The solution to this problem is the GLS estimator $\tilde{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$.

A method-of-moments approach begins with the model written in the form

$$\mathbf{Y} - \mathbf{X}\beta_0 = \epsilon$$

using β_0 to denote the true value of β. Suppose that you have an $n \times k$ matrix $\mathbf{W}$ that satisfies the condition $\text{E}(\epsilon|\mathbf{W}) = \mathbf{0}_{n\times 1}$. Then

$$\text{E}\,\mathbf{W}'(\mathbf{Y} - \mathbf{X}\beta_0) = \text{E}\,\mathbf{W}'\epsilon = \mathbf{0}_{k\times 1}$$

from which the true

$$\beta_0 = \left(\text{E}(\mathbf{W}'\mathbf{X})\right)^{-1}\text{E}\,\mathbf{W}'\mathbf{Y},$$

suggesting the method-of-moments estimator

$$\bar{\beta} = \left((\mathbf{W}'\mathbf{X})\right)^{-1}\mathbf{W}'\mathbf{Y}.$$

Note that if we choose $\mathbf{W} = \mathbf{V}^{-1}\mathbf{X}$, the method-of-moments estimator is identical to the GLS estimator. So the GLS estimator can be viewed as a special case of the method-of-moments estimator.

What are the properties of the estimator for general $\mathbf{W}$? We can rewrite the estimator as

$$\bar{\beta} = \beta + \left((\mathbf{W}'\mathbf{X})\right)^{-1}\mathbf{W}'\epsilon$$

and examine its covariance matrix

$$\mathrm{E}(\bar{\beta} - \beta)(\bar{\beta} - \beta)' = \mathrm{E}\left((\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\epsilon\epsilon'\mathbf{W}(\mathbf{X}'\mathbf{W})^{-1}\right)$$

Conditional on $\mathbf{X}$ and $\mathbf{W}$, the covariance matrix is $\sigma^2(\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{V}\mathbf{W}(\mathbf{X}'\mathbf{W})^{-1}$, whereas the covariance matrix of the GLS estimator (conditional on $\mathbf{X}$) is $\sigma^2(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$. We now show that the method of moments estimator is in general less efficient than the GLS estimator. Examining whether the difference in the covariance matrices is positive semidefinite, we know that

$$(\mathbf{W}'\mathbf{X})^{-1}\mathbf{W}'\mathbf{V}\mathbf{W}(\mathbf{X}'\mathbf{W})^{-1} - (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

is positive semidefinite if and only if

$$\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} - \mathbf{X}'\mathbf{W}(\mathbf{W}'\mathbf{V}\mathbf{W})^{-1}\mathbf{W}'\mathbf{X}$$

is positive semidefinite. The difference can be factored as follows:

$$\mathbf{X}'\mathbf{V}^{-1/2}\left(\mathbf{I} - \mathbf{V}^{1/2}\mathbf{W}(\mathbf{W}'\mathbf{V}\mathbf{W})^{-1}\mathbf{W}'\mathbf{V}^{1/2}\right)\mathbf{V}^{-1/2}\mathbf{X}$$

The matrix in the middle is symmetric idempotent, hence the difference is positive semidefinite. To sum up, we see that the method of moments estimator is identical to GLS for one particular choice of $\mathbf{W}$, but in general differs from, and is less efficient than, the GLS estimator.

24.3 The Algebra of Projections

Much of the logic of projections employed in the case of simple least squares carries over to the GLS case if a different metric is used to define orthogonality and distance (Zaman 1996, pp. 12–13). Consider the n-vectors $\mathbf{y}$ and $\mathbf{z}$ and a positive definite $n \times n$ matrix $\mathbf{H}$. These vectors are said to be orthogonal with respect to $\mathbf{H}$ if $\mathbf{y}'\mathbf{H}\mathbf{z} = 0$. The (squared) length of $\mathbf{y}$ is defined as

$$\|\mathbf{y}\|^2 = \mathbf{y}'\mathbf{H}\mathbf{y}$$

and the distance between $\mathbf{y}$ and $\mathbf{z}$ is denoted by $\|\mathbf{y} - \mathbf{z}\|$ using the new $\mathbf{H}$ metric. The projection of $\mathbf{y}$ onto $\mathcal{S}(\mathbf{X})$, which we denote by $\hat{\mathbf{y}}_{\mathbf{X}}^{H}$, is the closest vector to $\mathbf{y}$ in $\mathcal{S}(\mathbf{X})$ according to the new definition of distance. The projection produces an orthogonal error, so

that to find the coefficients $\tilde{\beta}$ of the projection $\hat{\mathbf{y}}_{\mathbf{X}}^{H} = \mathbf{X}\tilde{\beta}$ we must solve the orthogonality conditions

$$\mathbf{X}'\mathbf{H}(\mathbf{y} - \mathbf{X}\tilde{\beta}) = \mathbf{0}.$$

Doing so yields the new projection formula $\tilde{\beta} = (\mathbf{X}'\mathbf{H}\mathbf{X})^{-1}\mathbf{X}'\mathbf{H}\mathbf{y}$.

If we take $\mathbf{H} = \mathbf{V}^{-1}$, we obtain the GLS estimator. The associated projection matrices are

$$\mathbf{P}_{\mathbf{X}}^{\mathbf{V}^{-1}} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$$

and

$$\mathbf{M}_{\mathbf{X}}^{\mathbf{V}^{-1}} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}$$

Note that the $\mathbf{M}_{\mathbf{X}}^{\mathbf{V}^{-1}}$ matrix projects vectors into the space $\mathcal{S}^{\perp}(\mathbf{X})$, which is the space of vectors $\mathbf{z}$ orthogonal to $\mathbf{X}$ according to the new definition of orthogonality, which is $\mathbf{z}'\mathbf{V}^{-1}\mathbf{x} = 0$.

24.4 Forecasting

In an influential paper, Goldberger (1962) showed how to construct minimum variance linear unbiased forecasts when the disturbances are correlated over time, as in autoregressive models (to be discussed below) and random-effects error component models for panel data (for more on the latter, see Chapter 31).

Consider a time-series model $Y_t = \mathbf{X}_t'\beta + \epsilon_t$ for which there are T observations and k covariates. Assume that $\mathbf{X}_{T+1}$ is somehow known and denote by F the forecast of Y_{T+1} based on all the information that is available: $\mathbf{X}_{T+1}, \mathbf{X}$, and $\mathbf{Y}$. We restrict ourselves to *linear* forecasting functions of the form $F = \mathbf{c}'\mathbf{Y}$ in which $\mathbf{c}$ is an $T \times 1$ vector whose elements, which we will shortly determine, are in some way functions of $\mathbf{X}$ and $\mathbf{X}_{T+1}$. For simplicity, let $\Omega = \mathrm{E}(\epsilon\epsilon' \mid \mathbf{X})$. Once we arrive at the result, we can revert to using $\sigma^2\mathbf{V}$ instead of Ω.

For unbiasedness, we must have $\mathrm{E}(F - Y_{T+1}) = 0$ and rewriting the forecast error as

$$F - Y_{T+1} = \mathbf{c}'\mathbf{Y} - Y_{T+1} = \mathbf{c}'\mathbf{X}\beta + \mathbf{c}'\epsilon - \mathbf{X}_{T+1}'\beta - \epsilon_{T+1} = (\mathbf{c}'\mathbf{X} - \mathbf{X}_{T+1}')\beta + \mathbf{c}'\epsilon - \epsilon_{T+1},$$

we see that $\mathbf{c}'\mathbf{X} = \mathbf{X}_{T+1}'$ is required if the forecast is to be unbiased. Imposing these k conditions on $\mathbf{c}$ yields the forecast error of an unbiased forecast,

$$F - Y_{T+1} = \mathbf{c}'\epsilon - \epsilon_{T+1}.$$

The variance of the forecast error, conditional on $\mathbf{X}$ and $\mathbf{X}_{T+1}$, is then

$$\mathrm{Var}(F - Y_{T+1}) = \mathbf{c}'\Omega\mathbf{c} + \sigma^2_{T+1,T+1} - 2\,\mathrm{E}\,\mathbf{c}'\epsilon \cdot \epsilon_{T+1}.$$

Letting

$$\mathbf{w} = \mathrm{E}\,\epsilon_{T+1} \cdot \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_T \end{bmatrix},$$

we have

$$\mathrm{Var}(F - Y_{T+1}) = \mathbf{c}'\Omega\mathbf{c} + \sigma^2_{T+1,T+1} - 2\mathbf{c}'\mathbf{w}.$$

Having structured the problem in this way, we now set ourselves the task of deriving the optimal, variance-minimizing $\mathbf{c}$, and for this purpose the middle term of the forecast error variance can be disregarded. Setting up a Langrangian with k multipliers,

$$\mathcal{L} = \mathbf{c}'\Omega\mathbf{c} - 2\mathbf{c}'\mathbf{w} - 2\lambda'(\mathbf{X}'\mathbf{c} - \mathbf{X}_{T+1}),$$

differentiating and solving, we find that

$$\mathbf{c} = \Omega^{-1}\left(\mathbf{I} - \mathbf{X}(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}'\Omega^{-1}\right)\mathbf{w} + \Omega^{-1}\mathbf{X}(\mathbf{X}'\Omega^{-1}\mathbf{X})^{-1}\mathbf{X}_{T+1},$$

and from this we obtain the minimum-variance linear forecast,

$$F = \mathbf{X}'_{T+1}\tilde{\beta} + \mathbf{w}'\Omega^{-1}\mathbf{e},$$

in which $\tilde{\beta}$ is the GLS estimator and $\mathbf{e} = \mathbf{Y} - \mathbf{X}\tilde{\beta}$ is the (untransformed) vector of residuals. We can also substitute the optimal $\mathbf{c}$ into the expression for the forecast error variance to obtain the minimized level of variance.

How does Goldberger's approach compare with the use of conditional expectations? Recall that the function $\mathrm{E}(Y_{T+1} \mid \mathbf{X}_{T+1}, \mathbf{X}, \mathbf{Y})$ minimizes the expected squared forecast error. This function is not necessarily linear. However, when it is applied to a linear model for Y_{T+1}, it gives

$$\mathrm{E}(Y_{T+1} \mid \mathbf{X}_{T+1}, \mathbf{X}, \mathbf{Y}) = \mathbf{X}'_{T+1}\beta + \mathrm{E}(\epsilon_{T+1} \mid \mathbf{X}_{T+1}, \mathbf{X}, \mathbf{Y}),$$

and the second term can be simplified (assuming that the covariates $\mathbf{X}_{T+1}$ and $\mathbf{X}$ are independent of ϵ and ϵ_{T+1}), so that

$$\mathrm{E}(Y_{T+1} \mid \mathbf{X}_{T+1}, \mathbf{X}, \mathbf{Y}) = \mathbf{X}'_{T+1}\beta + \mathrm{E}(\epsilon_{T+1} \mid \epsilon).$$

This looks very much like the Goldberger forecasting function. Note, though, that the Goldberger approach *assumes* that the forecaster already knows $\mathbf{X}_{T+1}$. In most situations, of course, $\mathbf{X}_{T+1}$ would not be known in advance. The conditional expectations approach for this more realistic case yields

$$\mathrm{E}(Y_{T+1} \mid \mathbf{X}, \mathbf{Y}) = \mathrm{E}(\mathbf{X}_{T+1} \mid \mathbf{X})'\beta + \mathrm{E}(\epsilon_{T+1} \mid \epsilon)$$

assuming independence of the covariates and disturbances. This is a harder forecasting problem: to proceed we need to estimate β, devise a method of forecasting $\mathbf{X}_{T+1}$ and also determine how to estimate the conditional expectation of ϵ_{T+1} using the data available.

24.5 Estimated GLS (EGLS)

Taking $\mathrm{E}(\epsilon\epsilon' \mid \mathbf{X}) = \sigma^2\mathbf{V}$, we now explore what happens when the $\mathbf{V}$ matrix is unknown and must be estimated by $\hat{\mathbf{V}}$. The following conditions are needed to establish the consistency and asymptotic normality of the EGLS estimators of β and the consistency of the estimator of σ^2:

i) $\operatorname{plim} n^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{Q}$, as above;

ii) $\text{plim}\Big((n^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}) - (n^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\Big) = \mathbf{0}$;

iii) $\text{plim}\Big((n^{-1/2}\mathbf{X}'\hat{\mathbf{V}}^{-1}\epsilon) - (n^{-1/2}\mathbf{X}'\mathbf{V}^{-1}\epsilon)\Big) = \mathbf{0}$; and

iv) $\text{plim}\Big(n^{-1}\epsilon'\hat{\mathbf{V}}^{-1}\epsilon - n^{-1}\epsilon'\mathbf{V}^{-1}\epsilon\Big) = 0$, this last to establish $\tilde{\sigma}^2 \xrightarrow{p} \sigma^2$.

Under these conditions, the GLS and EGLS estimators are asymptotically equivalent, and both dominate OLS in terms of efficiency. The proofs make for a good exercise in manipulation of asymptotic expressions. Note, however, that in finite samples nothing guarantees that EGLS will out-perform OLS.

To implement the EGLS estimator, one must have in mind a specification of $\mathbf{V}$, an $n \times n$ matrix, that involves a much smaller set of parameters than the dimension of $\mathbf{V}$ would suggest. We will discuss simple specifications of $\mathbf{V}$ immediately below and explore more detailed specifications in Chapter 25 on spatial econometrics. In general, however, the EGLS approach requires the following steps:

1. Having verified that the OLS estimator is consistent in the case being explored—this is certainly not a given, as we mentioned earlier—then run OLS and extract the OLS residuals.

2. Specify a model in which $\text{E}(\epsilon_i\epsilon_j|\mathbf{X}_i, \mathbf{X}_j) = \sigma^2\phi(\mathbf{X}_i, \mathbf{X}_j, \theta)$. The form of $\phi()$ and the roles played in it by the covariates will of course differ from one problem to the next. If you cannot specify a model, you will have to abandon EGLS and fall back on the robust methods of estimating the OLS covariance matrix.

3. Estimate θ using an auxiliary regression such as $e_ie_j = \sigma^2\phi(\mathbf{X}_i, \mathbf{X}_j, \theta) + w_{ij}$, in which you will note that the OLS residuals e_i and e_j—which we view as *proxies* for the unknown disturbances ϵ_i and ϵ_j—have been substituted for those disturbances. If we knew the disturbances, then the auxiliary regression would be $\epsilon_i\epsilon_j = \sigma^2\phi(\mathbf{X}_i, \mathbf{X}_j, \theta) + u_{ij}$, but this is not implementable.

4. Verify that $\hat{\theta}$ is consistent. This is not at all an easy task, as we'll show below in the simple case of heteroskedasticity.

5. Form $\hat{v}_{ij} = \phi(\mathbf{X}_i, \mathbf{X}_j, \hat{\theta})$ and the full matrix $\hat{\mathbf{V}}$.

6. Compute the EGLS estimator $\tilde{\beta} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$.

24.6 Heteroskedasticity

Heteroskedasticity arises when $\mathbf{V}$ is a diagonal matrix with different elements on its diagonal. We allow these elements v_{ii} to be functions of the covariate vector $\mathbf{X}_i$. In economic analyses, we would expect disturbances to exhibit heteroskedasticity when: there is *measurement error in the dependent variable* (with the variance of the measurement error being determined in part by right-hand side covariates); when the data being analyzed are *averages of more disaggregated data* (as would be the case, for instance, if we had data on average

income in US counties, with the average for county i being based on the incomes of n_i individual residents, and with the disturbance term thus having a variance of $v_{ii} = \sigma^2/n_i$); and when the *dependent variable has a floor or ceiling* (as in the case of grades of school attained in a sample of children aged 6 to 18, with all 6 year-olds having zero grades attained and the variance of attainment growing as a child gets older). There are, of course, many other circumstances in which heteroskedasticity may arise.

Properties of the OLS estimator

Let us first explore the implications of applying OLS to the untransformed model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$. Earlier we saw that under the appropriate assumptions, the OLS estimator $\hat{\beta}$ is unbiased and consistent for the β parameters. But when the disturbances are heteroskedastic, what quantity does the variance estimator s^2 actually estimate?

To address this question, consider the asymptotically equivalent estimator $\hat{\sigma}^2 = (n-k)s^2/n = \mathbf{e}'\mathbf{e}/n$, or

$$\hat{\sigma}^2 = \frac{\mathbf{e}'\mathbf{e}}{n} = \frac{\epsilon'\epsilon}{n} - \frac{\epsilon'\mathbf{X}}{n}\left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1}\frac{\mathbf{X}'\epsilon}{n}$$

As noted above, if the $k \times k$ matrix

$$\mathrm{E}(\mathbf{X}'\mathbf{V}\mathbf{X}) = \sigma^2 \mathrm{E}\sum_i v_{ii}\mathbf{X}_i\mathbf{X}_i'$$

is of order $o(n^2)$, then $n^{-1}\mathbf{X}'\epsilon \xrightarrow{p} \mathbf{0}$, and with $(n^{-1}\mathbf{X}'\mathbf{X})^{-1} \xrightarrow{p} \mathbf{W}_{xx}^{-1}$ we have

$$\hat{\sigma}^2 \stackrel{a}{=} \frac{1}{n}\sum_i^n \epsilon_i^2.$$

To determine if the expression on the right has a probability limit, we need to consider the behavior of a normalized sum of non-identically distributed random variables. (For the i-th element in the sum, we have $\mathrm{E}\,\epsilon_i^2 = \sigma^2 v_{ii}$.) Assuming that the higher moments of ϵ_i^2 are appropriately behaved (see your chapter on laws of large numbers for non-iid random variables) and that $\lim_{n\to\infty} n^{-1}\sum_i^n v_{ii} = \alpha$, we obtain $\operatorname{plim}\hat{\sigma}^2 = \operatorname{plim} s^2 = \sigma^2\alpha$. That is, the estimator s^2 converges to a quantity that is the limiting average of the disturbance term variances.

White's OLS Covariance Estimator

White (1980) has devised a means of correcting the estimated covariances of the OLS estimates of β for heteroskedasticity in a way that does not require us to specify a model of how that heteroskedasticity is generated. Without such a model, of course, one could not hope to proceed to a GLS or EGLS estimator. White's approach is therefore extremely useful in cases where not enough is known about the sources of heteroskedasticity to formulate a model of it. To keep the exposition simple, let's subsume σ^2 in the $\mathbf{V}$ matrix.

Assume that the OLS estimator $\hat{\beta}$ is consistent and recall

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \cdot \sqrt{n}\frac{1}{n}\sum_i \mathbf{X}_i\epsilon_i.$$

The first factor on the right-hand side will converge in probability to a $k \times k$ matrix, while the second factor will under suitable assumptions converge in distribution to multivariate normal. The assumptions required are those needed in the Lindeberg central limit theorem and its variants. This is because although each term in the sum has a mean of zero, the variance of these terms differs, this variance being $v_{ii} \, \mathrm{E} \, \mathbf{X}_i \mathbf{X}_i'$, in which we are subsuming σ^2 in v_{ii}. As you will recall from our discussion of the Lindeberg theorem, a further assumption is needed to make productive use of its conclusions, which is that $\lim_{n \to \infty} \frac{1}{n} \sum_i^n v_{ii} \, \mathrm{E} \, \mathbf{X}_i \mathbf{X}_i'$ exists.

Define the $k \times k$ matrices

$$\mathbf{M}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E} \, \mathbf{X}_i \mathbf{X}_i'$$

and

$$\bar{\mathbf{V}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E} \, \epsilon_i^2 \mathbf{X}_i \mathbf{X}_i'$$

and let $\mathbf{X}_i$ be independent over i (if not necessarily identically distributed). Assume that $\hat{\beta}$ is consistent. With these assumptions, the Lindeberg central limit theorem can be applied to show that

$$\bar{\mathbf{V}}_n^{-1/2} \mathbf{M}_n \sqrt{n} (\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}),$$

and thus, with the further assumption that the limits of $\mathbf{M}_n$ and $\mathbf{V}_n$ exist,

$$\sqrt{n} (\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \lim_{n \to \infty} \mathbf{M}_n^{-1} \bar{\mathbf{V}}_n \mathbf{M}_n^{-1} \right).$$

To obtain an estimate of the limiting value of $\bar{\mathbf{V}}_n$, we use $\hat{\mathbf{V}}_n = n^{-1} \sum_i e_i^2 \mathbf{X}_i \mathbf{X}_i'$ where e_i^2 is the square of the i-th OLS residual. Likewise, for $\mathbf{M}_n$ we employ $\hat{\mathbf{M}}_n = n^{-1} \sum_i \mathbf{X}_i \mathbf{X}_i'$. Then, as White shows, plim $(\hat{\mathbf{V}}_n - \bar{\mathbf{V}}_n) = \mathbf{0}$ and plim $(\hat{\mathbf{M}}_n^{-1} \hat{\mathbf{V}}_n \hat{\mathbf{M}}_n^{-1} - \mathbf{M}_n^{-1} \bar{\mathbf{V}}_n \mathbf{M}_n^{-1}) = \mathbf{0}$. It follows that

$$\sqrt{n} (\hat{\beta}_n - \beta) \xrightarrow{d} \mathcal{N} \left(\mathbf{0}, \mathrm{plim} (n^{-1} \mathbf{X}'\mathbf{X})^{-1} \mathrm{plim} (n^{-1} \sum_i e_i^2 \mathbf{X}_i \mathbf{X}_i') \mathrm{plim} (n^{-1} \mathbf{X}'\mathbf{X})^{-1} \right)$$

or, to approximate,

$$\hat{\beta}_n - \beta \approx \mathcal{N} \left(\mathbf{0}, (\mathbf{X}'\mathbf{X})^{-1} (\sum_i e_i^2 \mathbf{X}_i \mathbf{X}_i') (\mathbf{X}'\mathbf{X})^{-1} \right).$$

Many statistical packages, including R and STATA, allow White corrections to be made to the OLS variance matrix.

The flavor of the White approach can be appreciated via a simple example. Suppose that we have a model with only one covariate, such that $Y_i = X_i \beta + \epsilon_i$, and let $\mathrm{E}(\epsilon_i^2 \mid X_i) = \sigma_{ii}$. In this case $\bar{\mathbf{V}}_n$ is simply $n^{-1} \, \mathrm{E} \sum_{i=1}^n \epsilon_i^2 X_i^2$. Consider one OLS residual, written as

$$e_i = Y_i - X_i \hat{\beta} = \epsilon_i - X_i (\hat{\beta} - \beta).$$

Squaring the residual gives

$$e_i^2 = \epsilon_i^2 - 2 X_i \epsilon_i \cdot (\hat{\beta} - \beta) + X_i^2 \cdot (\hat{\beta} - \beta)^2.$$

Multiply the result by X_i^2 and then average to obtain

$$\hat{\mathbf{V}}_n = \frac{1}{n}\sum_i e_i^2 X_i^2 = \frac{1}{n}\sum_i \epsilon_i^2 X_i^2 - 2\frac{1}{n}\sum_i X_i^3 \epsilon_i \cdot (\hat{\beta} - \beta) + \frac{1}{n}\sum_i X_i^4 \cdot (\hat{\beta} - \beta)^2.$$

If the probability limits of $\frac{1}{n}\sum_i X_i^3 \epsilon_i$ and $\frac{1}{n}\sum_i X_i^4$ both exist, the consistency of the OLS estimator $\hat{\beta}$ will ensure that the second and third terms on the right-hand side converge to zero. Thus

$$\frac{1}{n}\sum_i e_i^2 X_i^2 \stackrel{a}{=} \frac{1}{n}\sum_i \epsilon_i^2 X_i^2,$$

and the right-hand side is asymptotically equivalent to $\bar{\mathbf{V}}_n$.

The EGLS estimator

The EGLS estimator for the heteroskedastic case is generally quite straightforward to implement. We first apply OLS to the model and extract the residuals $\mathbf{e}$. We then form $\hat{v}_{ii}$, the details of which will depend on the specifics of the model of heteroskedasticity that is being considered. Having done this, we proceed to arrange the elements $1/\sqrt{\hat{v}_{ii}}$ along the diagonal of the diagonal matrix $\hat{\mathbf{V}}^{-1/2}$ and premultiply the data by $\hat{\mathbf{V}}^{-1/2}$, yielding

$$\hat{\mathbf{V}}^{-1/2}\mathbf{Y} = \hat{\mathbf{V}}^{-1/2}\mathbf{X}\beta + \hat{\mathbf{V}}^{-1/2}\epsilon.$$

The final step is to estimate β and σ^2 by applying OLS to these transformed data. The approach is often described as "weighted regression," because to implement it all we really have to do is to divide each Y_i by $\sqrt{\hat{v}_{ii}}$ and do likewise for each covariate including the constant term (which, as a consequence, is no longer constant).

What are the statistical properties of the $\hat{v}_{ii}$ estimates based on OLS residuals? To establish these properties, we need first to posit the model by which heteroskedasticity is generated. The easiest case to consider is a linear model, and so, drawing from Amemiya (1977) and Pagan and Hall (1983), let us consider the specification $v_{ii} = \mathrm{E}\,\epsilon_i^2 = \sigma_i^2$, writing the variance as

$$\sigma_i^2 = \sigma^2 + \mathbf{Z}_i'\gamma = \mathbf{d}_i'\delta,$$

where $\mathbf{d}_i = (1, \mathbf{Z}_i)'$ and $\delta = (\sigma^2, \gamma)'$. The $\mathbf{Z}_i$ covariates can overlap the $\mathbf{X}_i$ covariates of the main structural model.

Is it the case that a regression equation with e_i^2 substituted for ϵ_i^2,

$$e_i^2 = \mathbf{d}_i'\delta + \text{residuals},$$

gives consistent estimates of δ? Are the standard errors of $\hat{\delta}$ meaningful? To answer these questions, we write the residuals of the equation as follows,

$$e_i^2 = \mathbf{d}_i'\delta + e_i^2 - \sigma_i^2 = \mathbf{d}_i'\delta + (\epsilon_i^2 - \sigma_i^2) + (e_i^2 - \epsilon_i^2).$$

The least-squares estimator $\hat{\delta}$ is

$$\hat{\delta} = \delta + \left(n^{-1}\sum_i \mathbf{d}_i\mathbf{d}_i'\right)^{-1}\left(n^{-1}\sum_i \mathbf{d}_i\left(\epsilon_i^2 - \sigma_i^2 + e_i^2 - \epsilon_i^2\right)\right).$$

Assume plim $n^{-1}\sum_i \mathbf{d}_i\mathbf{d}_i' = \mathbf{D}$. Also assume that Var $\epsilon_i^2 = \mu_{4,i} - \sigma_i^4$ is bounded and likewise that $\sigma_i^2 < \bar{\sigma}^2$ for all i. Finally, assume that $\mathbf{Z}_i$ and ϵ_i^2 are uncorrelated over i.

We can now show that $\hat{\delta} \xrightarrow{p} \delta$. We must consider two terms,

$$\mathbf{D}^{-1}\left(n^{-1}\sum_i \mathbf{d}_i(\epsilon_i^2 - \sigma_i^2)\right),$$

and

$$\mathbf{D}^{-1}\left(n^{-1}\sum_i \mathbf{d}_i(e_i^2 - \epsilon_i^2)\right).$$

For the first of these, note that each term in the sum has zero mean; the variances of the terms are bounded; and the terms are uncorrelated over i. Therefore, by a standard law of large numbers this term has a probability limit of zero.

As for the second term, note that the least squares residual for a single observation is $e_i = \epsilon_i - \mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$. Squaring this, we find that

$$e_i^2 - \epsilon_i^2 = -2\epsilon_i\mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon + \epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i\mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon$$

where $\mathbf{X}_i$ is the vector of covariates for the i-th observation. Using this, examine the k-th element of

$$n^{-1}\sum \mathbf{d}_i(e_i^2 - \epsilon_i^2),$$

which is

$$n^{-1}\sum_i d_{k,i}\left(-2\epsilon_i\mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\right) + n^{-1}\sum_i d_{k,i}\left(\epsilon'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_i\mathbf{X}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\epsilon\right).$$

In the first of these expressions, insert n and assume that $n^{-1}\mathbf{X}'\mathbf{X} \xrightarrow{p} \mathbf{Q}$ and $n^{-1}\mathbf{X}'\epsilon \xrightarrow{p} \mathbf{0}$. Given that $\mathrm{E}\, d_{k,i}\epsilon_i\mathbf{X}_i' = \mathbf{0}'$, and noting the bounded variance assumption, we find that $n^{-1}\sum_i d_{k,i}\epsilon_i\mathbf{X}_i' \xrightarrow{p} \mathbf{0}'$. In dealing with the second expression, again inserting n where necessary, we see that if plim $n^{-1}\sum_i d_{k,i}\mathbf{X}_i\mathbf{X}_i' = \mathbf{M}_k$, then the entire expression has a plim of zero.

We conclude that $\hat{\delta} \xrightarrow{p} \delta$, in other words, estimates of δ derived from a regression of squared residuals e_i^2 on $(1, \mathbf{Z}_i)$ are consistent. What about limiting distributions—are the estimated standard errors of $\hat{\delta}$ also meaningful? To understand the issues, consider

$$\sqrt{n}(\hat{\delta} - \delta) \stackrel{a}{=} \mathbf{D}^{-1}\left(n^{-1/2}\sum_i \mathbf{d}_i(\epsilon_i^2 - \sigma_i^2)\right) + \mathbf{D}^{-1}\left(n^{-1/2}\sum_i \mathbf{d}_i(e_i^2 - \epsilon_i^2)\right)$$

and examine the first term on the right-hand side. Under the *null hypothesis of homoskedasticity*, we have $\mathrm{E}\,\epsilon_i^2 = \sigma_i^2 = \sigma^2$. Further assume that $\mathbf{d}_i$ and ϵ_i are independent over i. From this it follows that

$$n^{-1/2}\sum_i \mathbf{d}_i(\epsilon_i^2 - \sigma^2) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\mu_4 - \sigma^4)\mathbf{D}\right).$$

By the same kind of analysis as employed above, the second term is asymptotically negligible. Therefore, *under the null*, $\sqrt{n}(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, (\mu_4 - \sigma^4)\mathbf{D}^{-1})$ and $\mu_4 - \sigma^4$ will be

estimated consistently by the s^2 of the regression. In short, under the null hypothesis of homoskedasticity, the standard errors produced by a regression of e_i^2 on $(1, \mathbf{Z}_i)$ are asymptotically valid. Thus, one can test for heteroskedasticity by testing whether the coefficients on $\mathbf{Z}_i$ are significantly different from zero using an $\mathcal{F}$ or χ^2 test.

However, under the *alternative hypothesis of heteroskedasticity*, the expression

$$n^{-1/2} \sum_i \mathbf{d}_i (\epsilon_i^2 - \sigma_i^2)$$

will have a *different*, although still normal, limiting distribution. It remains the case that $\mathrm{E}\, d_{k,i}(\epsilon_i^2 - \sigma_i^2) = 0$, but the variance of this term differs across observations and we therefore require a central limit theorem that will fit this situation. The problem is that the heteroskedasticity regression is itself heteroskedastic! One approach is to correct its standard errors using the White robust covariance matrix formula.

24.7 Autocorrelation

In the time-series context, disturbance terms ϵ_t are correlated over t when they exhibit some persistence over time, the value of the disturbance for time t depending in some fashion on values of the disturbance in period $t-1$ or earlier. Here we consider only the case of a first-order autocorrelated ($AR(1)$) model, in which $\mathbf{Y}_t = \mathbf{X}_t'\beta + \epsilon_t$, where $\epsilon_t = \rho\epsilon_{t-1} + u_t$ for $t = 1, \ldots, T$ and the vector $\mathbf{u}$ has mean zero and covariance matrix $\sigma_u^2\mathbf{I}$. We assume $\mid \rho \mid < 1$, a necessary condition for what is termed *covariance stationarity*. Under these assumptions, it is easily shown that ϵ_t has variance $\sigma_\epsilon^2 = \sigma_u^2/(1-\rho^2)$. The covariance matrix of ϵ is

$$\mathrm{E}\, \epsilon\epsilon' = \sigma_\epsilon^2 \mathbf{V} = \frac{\sigma_u^2}{1-\rho^2} \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \rho & \cdots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \ddots & \\ \vdots & \vdots & & \ddots & \rho \\ \rho^{T-1} & \cdots & & \rho & 1 \end{bmatrix}.$$

It can be shown that

$$\mathbf{V}^{-1} = \frac{1}{1-\rho^2} \begin{bmatrix} 1 & -\rho & 0 & \cdots & & \\ -\rho & 1+\rho^2 & -\rho & 0 & \cdots & \\ 0 & -\rho & 1+\rho^2 & -\rho & \ddots & \\ \vdots & 0 & \ddots & \ddots & \ddots & 0 \\ & & \ddots & -\rho & 1+\rho^2 & -\rho \\ & & & 0 & -\rho & 1 \end{bmatrix}.$$

The orthogonal matrix $\mathbf{P}$ such that $\mathbf{P}'\mathbf{P} = \mathbf{V}^{-1}$ is

$$\mathbf{P} = \frac{1}{(1-\rho^2)^{1/2}} \begin{bmatrix} (1-\rho)^{1/2} & 0 & \cdots & & \\ -\rho & 1 & 0 & \cdots & \\ 0 & -\rho & \ddots & \ddots & \\ \vdots & \ddots & \ddots & \ddots & 0 \\ & & 0 & -\rho & 1 \end{bmatrix}.$$

An alternative transformation matrix that is often used is $\mathbf{P}^* = (1-\rho^2)^{1/2}\mathbf{P}$. Note that whereas $\mathrm{E}\,\mathbf{P}\epsilon\epsilon'\mathbf{P}' = \sigma_\epsilon^2\mathbf{I}$, when we use the alternative $\mathbf{P}^*$ transformation, we have $\mathrm{E}\,\mathbf{P}^*\epsilon\epsilon'\mathbf{P}^* = \sigma_u^2\mathbf{I}$. Writing the transformed data as $\mathbf{P}^*\mathbf{Y} = \mathbf{P}^*\mathbf{X}\beta + \mathbf{P}^*\epsilon$, we obtain a GLS model in which the covariance matrix of $\mathbf{P}^*\epsilon$ is $\sigma_u^2\mathbf{I}$.

Estimating ρ from OLS residuals

Imagine that the true ϵ_t disturbances were available. Then we could estimate ρ from the auxiliary regression $\epsilon_t = \rho\epsilon_{t-1} + u_t$. Proving that this regression yields a consistent estimator of ρ is not easy: see Amemiya (1985, Chapter 5) for the proof or Hayashi (2000) for a law of large numbers that applies to correlated but stationary time series. Yet another proof is required—again see Amemiya (1985, pp. 188–189)—to demonstrate that the OLS residual e_t can be substituted for the corresponding ϵ_t. (The arguments are not unlike those we needed in the heteroskedasticity discussion above.) The EGLS estimator is then formed by inserting $\hat{\rho}$ from the auxiliary regression into the transformation matrix $\mathbf{P}$ or $\mathbf{P}^*$.

Forecasting

Consider forecasting the value of Y_{T+1} given knowledge of $\mathbf{X}_{T+1}$ by using the conditional expectations approach. This yields

$$\mathrm{E}(Y_{T+1} \mid \mathbf{X}_{T+1}, \mathbf{X}, \mathbf{Y}) = \mathbf{X}'_{T+1}\beta + \mathrm{E}(\epsilon_{T+1} \mid \epsilon)$$

given the assumption that the disturbances are independent of the covariates. Now, because $\epsilon_{T+1} = \rho\epsilon_T + u_{T+1}$,

$$\mathrm{E}(\epsilon_{T+1} \mid \epsilon) = \rho\epsilon_T + \mathrm{E}(u_{T+1} \mid \epsilon) = \rho\epsilon_T.$$

The forecasting function is therefore $\mathbf{X}'_{T+1}\beta + \rho\epsilon_T$, which we would implement by using the EGLS estimator for β, some estimator for ρ, and by inserting the residual e_T in place of the unobserved disturbance ϵ_T. If the model is estimated by maximum likelihood or nonlinear least squares as we discuss below, the estimates of β and ρ and the residual could be taken from those results instead.

How would the Goldberger forecasting approach differ? As it turns out, not at all. Recall that the forecast function is $\mathbf{X}'_{T+1}\tilde{\beta} + \mathbf{w}'(\sigma_\epsilon^2\mathbf{V})^{-1}\mathbf{e}$, and for the case at hand,

$$\mathbf{w}' = \sigma_\epsilon^2 \cdot \begin{bmatrix}\rho^T & \rho^{T-1} & \cdots & \rho^2 & \rho\end{bmatrix}$$

so that σ_ϵ^2 cancels. Using the expression for $\mathbf{V}^{-1}$ that was given above, you can verify that the Goldberger approach leads to exactly the same answer as the conditional expectations approach. Remember, however, that the Goldberger method does not apply unless $\mathbf{X}_{T+1}$ is known.

24.8 ML Estimation: Heteroskedasticity

In this section, we explore maximum-likelihood estimation of a model with heteroskedasticity, assuming that the disturbance terms are normally distributed. Recall that in general,

when $\mathbf{Y}$ is multivariate normal with mean $\mathbf{X}\beta$ and variance Ω, the log-likelihood function is

$$L = -\frac{n}{2}\ln 2\pi - \frac{1}{2}\ln \mid \Omega \mid -\frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)'\Omega^{-1}(\mathbf{Y}-\mathbf{X}\beta)$$

with $\mid \Omega \mid$ being the absolute value of the determinant of Ω. In the case of heteroskedasticity the Ω matrix is diagonal and its determinant is the product of the elements on the diagonal.

To keep things simple, we will suppress the role of covariates, letting the mean of the dependent variable $\mathbf{Y}$ be $\boldsymbol{\iota} \cdot \mu$. We therefore have $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\iota}\mu, \sigma^2\mathbf{V})$, with

$$\mathbf{V} = \begin{pmatrix} f_1(\alpha) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & f_n(\alpha) \end{pmatrix}$$

where the function $f_i > 0$ has the properties $f_i(0) = 1, \partial f_i(0)/\partial\alpha \neq 0$. A useful special case is that of $f_i(\alpha) = e^{\mathbf{Z}_i'\alpha}$ in which α is a vector of dimension J attached to covariates $\mathbf{Z}_i$. The log of the determinant is $\ln \mid \sigma^2\mathbf{V} \mid = n\ln\sigma^2 + \sum_i \ln f_i(\alpha)$.

Hence, the log-likelihood function for this case is

$$L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2}\sum_i \ln f_i - \frac{1}{2}\sum_i \frac{(y_i-\mu)^2}{\sigma^2 f_i}.$$

From this, we obtain the scores

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_i \frac{(y_i-\mu)}{f_i}$$

and

$$\frac{\partial L}{\partial \sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_i \frac{(y_i-\mu)^2}{f_i}.$$

Also,

$$\frac{\partial L}{\partial \alpha} = -\frac{1}{2}\sum_i \frac{1}{f_i}\frac{\partial f_i}{\partial \alpha} + \frac{1}{2\sigma^2}\sum_i \frac{(y_i-\mu)^2}{f_i^2}\frac{\partial f_i}{\partial \alpha}$$

or

$$\frac{\partial L}{\partial \alpha} = \frac{1}{2}\sum_i \frac{1}{f_i}\left(\frac{(y_i-\mu)^2}{\sigma^2 f_i} - 1\right)\frac{\partial f_i}{\partial \alpha},$$

where $\partial f_i/\partial\alpha$ is a $J \times 1$ vector. Note that the term in large parentheses has an expectation of zero when evaluated at the true parameter values.

The diagonal terms of the second derivative matrix are as follows,

$$\frac{\partial^2 L}{\partial\mu\partial\mu} = -\frac{1}{\sigma^2}\sum_i \frac{1}{f_i}$$

$$\frac{\partial^2 L}{\partial\sigma^2\partial\sigma^2} = \frac{n}{2}\frac{1}{\sigma^4} - \frac{1}{\sigma^6}\sum_i \frac{(y_i-\mu)^2}{f_i}$$

and

$$\begin{aligned}\frac{\partial^2 L}{\partial\alpha\partial\alpha} &= \frac{1}{2}\left\{\sum_i \frac{1}{f_i}\left(\frac{(y-\mu)^2}{\sigma^2 f_i}-1\right)\frac{\partial^2 f_i}{\partial\alpha\partial\alpha}\right.\\ &\quad -\sum_i \frac{1}{f_i^2}\left(\frac{(y_i-\mu)^2}{\sigma^2 f_i}-1\right)\left(\frac{\partial f_i}{\partial\alpha}\right)\left(\frac{\partial f_i}{\partial\alpha}\right)'\\ &\quad \left.-\sum_i \frac{1}{f_i^2}\left(\frac{(y_i-\mu)^2}{\sigma^2 f_i}\right)\left(\frac{\partial f_i}{\partial\alpha}\right)\left(\frac{\partial f_i}{\partial\alpha}\right)'\right\}.\end{aligned}$$

Taking minus the expectations of these expressions yields

$$-\,\mathrm{E}\,\frac{\partial^2 L}{\partial\mu\partial\mu} = \frac{1}{\sigma^2}\sum_i \frac{1}{f_i}$$

$$-\,\mathrm{E}\,\frac{\partial^2 L}{\partial\sigma^2\partial\sigma^2} = \frac{n}{2\sigma^4}$$

and

$$-\,\mathrm{E}\,\frac{\partial^2 L}{\partial\alpha\partial\alpha} = \frac{1}{2}\sum_i \frac{1}{f_i^2}\left(\frac{\partial f_i}{\partial\alpha}\right)\left(\frac{\partial f_i}{\partial\alpha}\right)'.$$

The off-diagonal terms are derived as follows,

$$\frac{\partial^2 L}{\partial\mu\partial\sigma^2} = -\frac{1}{\sigma^4}\sum_i \frac{(y_i-\mu)}{f_i}$$

$$\frac{\partial^2 L}{\partial\mu\partial\alpha} = -\frac{1}{\sigma^2}\sum_i \frac{(y_i-\mu)}{f_i^2}\frac{\partial f_i}{\partial\alpha}$$

$$\frac{\partial^2 L}{\partial\sigma^2\partial\alpha} = -\frac{1}{2\sigma^4}\sum_i \frac{(y_i-\mu)^2}{f_i^2}\frac{\partial f_i}{\partial\alpha}.$$

The first two of the off-diagonal derivatives have expectation zero. The last has (minus) expectation

$$-\,\mathrm{E}\,\frac{\partial^2 L}{\partial\sigma^2\partial\alpha} = \frac{1}{2\sigma^2}\sum_i \frac{1}{f_i}\frac{\partial f_i}{\partial\alpha}.$$

The expressions above, when arranged in the information matrix $\mathcal{R}_n$, can be summarized as follows. Let ι be an $n \times 1$ vector of ones and form the $n \times J$ matrix $\mathbf{G}$ with typical element $G_{i,j} = \partial f_i/\partial\alpha_j$. We then have

$$\mathcal{R}_n = \begin{pmatrix} \frac{1}{\sigma^2}\iota'\mathbf{V}^{-1}\iota & 0 & 0 \\ 0 & \frac{n}{2\sigma^4} & \frac{1}{2\sigma^2}\iota'\mathbf{V}^{-1}\mathbf{G} \\ 0 & \frac{1}{2\sigma^2}\mathbf{G}'\mathbf{V}^{-1}\iota & \frac{1}{2}\mathbf{G}'\mathbf{V}^{-2}\mathbf{G} \end{pmatrix},$$

an expression that is also shown in Greene (2003).

The Lagrange Multiplier Test

Under the null hypothesis of homoskedasticity, $\mathbf{V} = \mathbf{I}$. The Lagrange Multiplier test will involve the $J \times 1$ score element $\partial L/\partial\alpha$ and the appropriate sub–matrix of the inverse of the information matrix $\mathcal{R}_n$.

Using the partitioned inversion formula, we find that under the null, the $J \times J$ element of the inverse that corresponds to α is $2(\mathbf{G}'\mathbf{M}_\iota\mathbf{G})^{-1}$ where $\mathbf{M}_\iota = \mathbf{I} - \boldsymbol{\iota}\boldsymbol{\iota}'/n$, an idempotent matrix that generates deviations from means. Under the null, the score element associated with α can be written as

$$\frac{\partial L}{\partial \alpha} = \frac{1}{2}\mathbf{G}'\mathbf{l}$$

where the $n \times 1$ vector $\mathbf{l}$ has typical element $e_i^2/\hat{\sigma}^2 - 1$ with $\hat{\sigma}^2$ being the maximum likelihood estimate of σ^2 under the null, or simply the sum of squared residuals divided by n. Note that $\mathbf{l}$ has an arithmetic mean of zero.

The LM statistic is then

$$LM = \frac{1}{2}\mathbf{l}'\mathbf{G}(\mathbf{G}'\mathbf{M}_\iota\mathbf{G})^{-1}\mathbf{G}'\mathbf{l}.$$

Since $\mathbf{M}_\iota\mathbf{l} = \mathbf{l}$, this is identical to

$$LM = \frac{1}{2}\mathbf{l}'\mathbf{M}_\iota\mathbf{G}(\mathbf{G}'\mathbf{M}_\iota\mathbf{G})^{-1}\mathbf{G}'\mathbf{M}_\iota\mathbf{l}.$$

The last expression is one-half the explained sum of squares from an artificial regression of $\mathbf{l}$ on $(\boldsymbol{\iota}, \mathbf{G})$. (This is because $\mathbf{l}$ has zero arithmetic mean, hence $\boldsymbol{\iota}'\mathbf{l} = 0$.) Remember that $l_i = e_i^2/\hat{\sigma}^2 - 1$. If we now transform the dependent variable l_i to l_i^* by omitting the 1, the explained sum of squares (about the mean) is left unchanged.

Thus, the LM test can be implemented by an artificial regression in which the dependent variable for observation i is $e_i^2/\hat{\sigma}^2$ and the explanatory variables include a constant and the J derivatives $\partial f_i(0)/\partial\alpha_j$. The specification most likely to be encountered in practice is $f_i(\alpha) = \exp(\mathbf{z}_i'\alpha)$ and for this case the derivatives are simply the $z_{i,j}$. The test statistic is one-half of the explained sum of squares (about the mean) from this artificial regression.

Davidson and MacKinnon (1993, pp. 562–563) provide an enlightening discussion of the properties of the test statistic, showing in particular why the statistic is distributed as χ^2 in this case.

24.9 ML Estimation: Autocorrelation

In this section, we will derive the likelihood function for the model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ with $\epsilon_t = \rho\epsilon_{t-1} + u_t$ where $\mathbf{u}$ is $\mathcal{N}(0, \sigma_u^2\mathbf{I})$. We will consider a sample of T observations. In deriving the likelihood function, it is helpful to factor it as

$$L^*(Y_1, \ldots, Y_T \mid \theta) = L_2^*(Y_2, \ldots, Y_T \mid Y_1, \theta)L_1^*(Y_1 \mid \theta)$$

and then deal with each factor separately.

For the $T-1$ observations Y_2 to Y_T, we employ the transformation

$$Y_t - \rho Y_{t-1} - (\mathbf{X}_t - \rho\mathbf{X}_{t-1})'\beta = u_t.$$

In changing variables we always need to consider the Jacobian of the transformation, and luckily in this case the Jacobian is unity. Hence, the (conditional) log-likelihood is

$$L_2 = -\frac{(T-1)}{2}\ln 2\pi - \frac{(T-1)}{2}\ln\sigma_u^2 - \frac{1}{2\sigma_u^2}\mathbf{u}'\mathbf{u}$$

or

$$L_2 = -\frac{(T-1)}{2}\ln 2\pi - \frac{(T-1)}{2}\ln\sigma_u^2 - \frac{1}{2\sigma_u^2}\sum_{t=2}^{T}\left(Y_t - \rho Y_{t-1} - (\mathbf{X}_t - \rho\mathbf{X}_{t-1})'\beta\right)^2.$$

Now, for the first observation Y_1, recall that $\epsilon_1 \sim \mathcal{N}\left(0, \frac{\sigma_u^2}{1-\rho^2}\right)$. Therefore,

$$L_1 = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\frac{\sigma_u^2}{1-\rho^2} - \frac{1}{2}\frac{1-\rho^2}{\sigma_u^2}(Y_1 - \mathbf{X}_1'\beta)^2.$$

The full log–likelihood is the sum of L_1 and L_2. See Greene (2003) for the second derivatives and the information matrix $\mathcal{R}_T$.

Although the asymptotic properties of the ML estimator are not affected if we omit the first observation, including it makes good sense on both numerical and practical grounds. Inspection of L_1 above shows that the full log–likelihood will include the term $\frac{1}{2}\ln(1-\rho^2)$. The presence of this term ensures that $|\hat{\rho}| < 1$, as is required for stationarity. Furthermore, when one is estimating a model using a relatively short time-series of observations in which important variables are trended, the first observation will often exert considerable leverage, see Davidson and MacKinnon (1993, pp. 348–350). Accumulated experience suggests that it is wiser to include the first observation in such circumstances.

Testing for autocorrelation

If you have already estimated the model by maximum likelihood, testing for $\rho = 0$ using a Wald or likelihood ratio test is straightforward, and with modern software either method can be used. It may be of interest, however, to see how to carry out a Lagrange Multiplier test for an $AR(1)$ specification with $\rho = 0$ being the null hypothesis. One reason to work through the derivation is that the LM test is fairly easily generalized to $AR(p)$ error processes, which take the form $\epsilon_t = \sum_{i=1}^{p}\rho_i\epsilon_{t-i} + u_t$, whereas the Wald and likelihood ratio tests become much more difficult in high-order autoregressive specifications. With some additional effort, the LM test can also be shown to apply to moving average error processes (Breusch and Pagan 1980).

Although we have advocated the inclusion of the first observation in estimating AR models and the Wald and likelihood ratio tests would use the ML estimate $\hat{\rho}$ derived from that full likelihood, the formulation of the LM test is much simplified if we omit the first observation. We can then re-cast the linear model as a nonlinear regression model having parameters β, ρ, σ_u^2. From here we proceed to develop the LM test for the null hypothesis $\rho = 0$, following the procedures laid out in Chapter 19 for nonlinear models.[1]

[1]Note that some technical issues arise in the time-series context in defining the $\mathcal{J}$ matrix when the right-hand side explanatory variables of the model include the lagged value of the dependent variable.

Using observations $t = 2, ..., T$, we transform the linear model $Y_t = \mathbf{X}_t'\beta + \epsilon_t$ by ρ-differencing, giving

$$\begin{aligned} Y_t &= \rho Y_{t-1} + (\mathbf{X}_t' - \rho \mathbf{X}_{t-1}')\beta + u_t \\ &\equiv g(\mathbf{Z}_t, \beta, \rho) + u_t \end{aligned}$$

with $\mathbf{Z}_t = (Y_{t-1}, \mathbf{X}_t, \mathbf{X}_{t-1})$. To test the null hypothesis $\rho = 0$, we need the constrained ML estimates $\tilde{\beta}$ and $\tilde{\sigma}_u^2$ under this hypothesis—but these are simply the ordinary least squares estimates $\hat{\beta}$ and $\hat{\sigma}_u^2$. As Chapter 19 discussed, we also need the residuals evaluated at the null hypothesis, which are the ordinary least squares residuals $e_t = Y_t - \mathbf{X}_t'\hat{\beta}$. We need as well the derivatives of the g_t function evaluated at $\rho = 0$, and these are

$$\begin{aligned} \frac{\partial g_t}{\partial \rho} &= Y_{t-1} - \mathbf{X}_{t-1}'\hat{\beta} = e_{t-1} \\ \frac{\partial g_t}{\partial \beta} &= \mathbf{X}_t. \end{aligned}$$

Because the constrained estimator of β is the OLS estimator, e_{t-1} is simply a lagged ordinary least squares residual taken from a model in which Y_t is regressed on $\mathbf{X}_t$.

The LM test statistic is thus executed by regressing the OLS residual e_t on its own lagged value e_{t-1} and covariates $\mathbf{X}_t$. The uncentered R^2 from this regression, when multiplied by $T - 1$, is then compared with the critical values of the χ_1^2 distribution. A nice feature of the test is that it remains valid even when $\mathbf{X}_t$ already contains lagged values of Y_t, so that there is no need to refer to specialized tests such as Durbin's "h" for models of this kind.

An alternative procedure—which like the LM test can be extended to cover higher-order AR specifications—is to estimate the ρ-differenced model $Y_t = \rho Y_{t-1} + (\mathbf{X}_t' - \rho \mathbf{X}_{t-1}')\beta + u_t$ *directly* by nonlinear least squares, and use a Wald test focused on the estimated $\hat{\rho}$ coefficient to test the null. This approach is more general than the LM approach in that it does not require a normality assumption. When the LM tests were being worked out in the early 1980s, the ability to execute them using ordinary rather than nonlinear least squares was regarded as one of their most appealing features. With modern software such as R and STATA, however, it is not very difficult to specify and estimate nonlinear models, and today there is little reason to avoid them.

24.10 Lagged Dependent Variables: Introduction

Suppose that the structural model is specified as

$$Y_t = \alpha Y_{t-1} + \mathbf{X}_t'\beta + \epsilon_t$$

with $\epsilon_t = \rho \epsilon_{t-1} + u_t$ as above. Obviously, ϵ_t is correlated with the right-hand side variable Y_{t-1} and ordinary least squares will fail to produce consistent estimates of α and β. At first glance, this problem would appear to call for the use of instrumental variables as we will discuss in Chapter 27. But if we ρ-difference the model, yielding

$$Y_t = (\alpha + \rho)Y_{t-1} - \rho\alpha Y_{t-2} + (\mathbf{X}_t - \rho \mathbf{X}_{t-1})'\beta + u_t.$$

and apply nonlinear least squares, we can in fact estimate α, β, and ρ consistently, provided that the structural model includes $\mathbf{X}_t$ covariates.

To understand why $\mathbf{X}_t$ is needed for this approach to work, consider what happens with ρ-differencing in a model that has no such covariates. In this case,

$$Y_t = (\alpha + \rho)Y_{t-1} - \rho\alpha Y_{t-2} + u_t,$$

and relabeling parameters gives

$$Y_t = aY_{t-1} + bY_{t-2} + u_t.$$

Because u_t is uncorrelated with Y_{t-1} and Y_{t-2}, the a and b parameters can be estimated consistently by ordinary least squares. Imagine having a sample so large that the OLS estimates $\hat{a}$ and $\hat{b}$ are essentially equal to a and b, and consider solving for α and ρ given a and b, from the relations

$$\begin{aligned} a &= \alpha + \rho \\ b &= -\alpha\rho. \end{aligned}$$

Although we have here two equations in two unknowns, a little investigation will show you that there are, unfortunately, two solutions. The parameters α and ρ are therefore not identified. Not even a time-series approaching infinite length would allow you to determine which of the two solutions to the equations corresponds to the true values of the α and ρ parameters.

Now add a single $\mathbf{X}_t$ covariate to the structural equation and repeat the exercise, with

$$\begin{aligned} Y_t &= (\alpha + \rho)Y_{t-1} - \rho\alpha Y_{t-2} + (X_t - \rho X_{t-1})'\beta + u_t \\ &= aY_{t-1} + bY_{t-2} + cX_t + dX_{t-1} + u_t. \end{aligned}$$

We can identify β directly from the c coefficient and can determine the value of ρ from the ratio of d to c. We could then proceed to use this information to identify α. Adding X_t to the specification gives us enough information to identify all of the parameters.

Chapter 25

Spatial Econometrics

Written with Donghwan Kim. Still Under Construction!

This chapter was written to provide background for a set of spatial econometric routines that are being coded in Fortran. Only the spatial error model is discussed here, although much of the material applies to the spatial lag model as well. The literature on these models has been dominated by—one could almost say "obsessed with"—the computational limits that until recently have frustrated efforts to estimate spatial regressions with large samples. These limits stem from the nature of the weight matrix used to specify spatial linkages among the n observations of the dataset. The past ten years have seen the introduction of several new techniques that have pushed back the computational barriers, and it is now possible to estimate spatial models in samples with many thousands of observations, at least for certain types of weight matrices.

Where the computational details deserve mention, we have included notes on how to take maximum advantage of Fortran's column-by-column storage of arrays, which is a fundamental consideration in memory usage and speed of execution. Further revisions in the paper are likely as more material is gathered on sparse matrix routines and new algorithms that have appeared in the recent literature.

25.1 The Spatial Error Model

Viewed in the abstract, the spatial error model is merely a special case of the generalized least squares model, although there is much less mathematical structure in the covariance matrix of its disturbances than we are accustomed to seeing. The lack of structure gives rise to a host of computational complications, which must be confronted even in moderately-sized samples of one or two thousand observations.

The spatial error model is specified as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \tag{25.1}$$

with the pattern of spatial correlation in the disturbance term $\mathbf{u}$ being generated by

$$\mathbf{u} = \rho\mathbf{W}\mathbf{u} + \boldsymbol{\epsilon}. \tag{25.2}$$

In this equation, $\mathbf{W}$ is a spatial weight matrix of dimension $n \times n$ and $\boldsymbol{\epsilon}$ is a vector of n zero mean, homoskedastic, uncorrelated disturbances. The diagonal elements of $\mathbf{W}$ are zeroes and the off-diagonal elements, when multiplied by ρ, establish the connections between the spatial units i and j insofar as their disturbance terms are concerned. We can see this more clearly by isolating u_i, the disturbance for the i-th area,

$$u_i = \rho \sum_{j \neq i} w_{ij} u_j + \epsilon_i,$$

which is directly related via ρw_{ik} to u_k, the disturbance term of the k-th area.

Broadly speaking, there are three general types of weight matrices that can be considered: $\mathbf{W}$ can be a symmetric matrix, or the asymmetric result of row-standardizing a symmetric matrix, or it can be fundamentally asymmetric. The row-standardized specification is used so often in spatial econometrics that it has almost become the default specification. In this approach, each weight w_{ij} is scaled by $\sum_j w_{ij}$, so that for every area i the scaled weights sum to one. Standardizing the weights in this way often makes good sense. The popularity of row-standardization has been further enhanced by the numerical properties that it imparts to the scaled matrix, as will be seen. However, the approach does not apply to all cases of interest. We obviously cannot row-standardize spatially isolated observations for which $\sum_j w_{ij} = 0$, and the conversion of weights to averages is not always the right thing to do. Unfortunately, the literature I have seen rarely develops results for fundamentally asymmetric weight matrices, mainly, it would appear, because numerical issues arise with such weights that need to be studied on a case-by-case basis. In addition to these features of the $\mathbf{W}$ weight matrix, the nature of its elements also matters, in the sense that computational short-cuts are available when these elements take a zero–one, "binary" form.

From equation (25.2) we have $(\mathbf{I} - \rho\mathbf{W})\,\mathbf{u} = \boldsymbol{\epsilon}$ and assuming that the inverse exists, $(\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}$ can be substituted for $\mathbf{u}$ to obtain

$$\mathbf{Y} = \mathbf{X}\beta + (\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}. \tag{25.3}$$

The assumption that $\mathbf{I} - \rho\mathbf{W}$ is invertible is a necessary assumption in spatial error models. For some ρ values, at least, invertibility may preclude certain specifications of the weight matrix.[1]

As Anselin (1988, p. 108) observes, the substitution of $(\mathbf{I} - \rho\mathbf{W})^{-1}\boldsymbol{\epsilon}$ for $\mathbf{u}$ easily yields an expression for the covariance matrix of $\mathbf{u}$,

$$\Omega \equiv \mathrm{E}\,\mathbf{u}\mathbf{u}' = \sigma^2 \cdot (\mathbf{I} - \rho\mathbf{W})^{-1}\big((\mathbf{I} - \rho\mathbf{W})^{-1}\big)' = \sigma^2 \cdot \big[(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})\big]^{-1}. \tag{25.4}$$

An inspection of equation (25.4) shows that the relationship between the weight matrix $\mathbf{W}$ and the covariance matrix Ω is complicated. It is not intuitively obvious how a given specification of $\mathbf{W}$ is expressed in the pattern of variances and covariances that make up Ω, nor is it obvious how the nature of the weight matrix $\mathbf{W}$ can be divined from knowledge of the Ω covariance matrix. Equation (25.4) suggests that the diagonal elements of Ω will in general be unequal. That is, $\mathbf{u}$ is likely to be heteroskedastic as well as spatially correlated.

[1] I have not seen much discussion of this point in the spatial econometrics literature, but it may be addressed elsewhere.

If ρ and thus Ω were actually known, the GLS estimator of β would be

$$\begin{aligned}\hat{\beta} &= \left(\mathbf{X}'\Omega^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\Omega^{-1}\mathbf{Y} \\ &= \left(\mathbf{X}'(\mathbf{I}-\rho\mathbf{W})'(\mathbf{I}-\rho\mathbf{W})\mathbf{X}\right)^{-1}\mathbf{X}'(\mathbf{I}-\rho\mathbf{W})'(\mathbf{I}-\rho\mathbf{W})\mathbf{Y}.\end{aligned}$$

It is helpful to view the GLS estimator as an ordinary least squares regression applied to "spatially filtered" $\mathbf{X}$ and $\mathbf{Y}$ data. To see this, rewrite the model in the form

$$(\mathbf{I}-\rho\mathbf{W})\mathbf{Y} = (\mathbf{I}-\rho\mathbf{W})\mathbf{X}\beta + \boldsymbol{\epsilon}. \tag{25.5}$$

or, in an even more compact notation, as

$$\tilde{\mathbf{Y}} = \tilde{\mathbf{X}}\beta + \boldsymbol{\epsilon}. \tag{25.6}$$

Here $\tilde{\mathbf{Y}} = (\mathbf{I}-\rho\mathbf{W})\mathbf{Y}$ and $\tilde{\mathbf{X}} = (\mathbf{I}-\rho\mathbf{W})\mathbf{X}$ are the spatially filtered transformations. Note that the GLS estimator $\hat{\beta} = (\tilde{\mathbf{X}}'\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$. In short, if a consistent estimator of ρ is available, we can proceed using the method of feasible GLS.

25.2 The Kelejian–Prucha Method

It is surprising that to date, most of the spatial econometric literature has been given over to the maximum likelihood method of estimation rather than to feasible GLS. The problem was that for many years, no easily implemented, consistent estimator of ρ was available other than the maximum-likelihood estimator. Recently, however, a simple method-of-moments approach to estimation has been developed by Kelejian and Prucha (1999) and this may ultimately prove to be the method of choice in very large samples.

Three moment conditions involving ϵ and the weight matrix $\mathbf{W}$ form the basis of the method, with $(\mathbf{I}-\rho\mathbf{W})\,\mathbf{u}$ substituted for ϵ. The essence of the approach can be grasped by considering the first such moment,

$$\mathrm{E}\,\frac{1}{n}\boldsymbol{\epsilon}'\boldsymbol{\epsilon} = \sigma^2.$$

Written in terms of $\mathbf{u}$, this can be expressed as

$$\sigma^2 = \mathrm{E}\,\frac{1}{n}\mathbf{u}'\left(\mathbf{I}-\rho\mathbf{W}'-\rho\mathbf{W}+\rho^2\mathbf{W}'\mathbf{W}\right)\mathbf{u},$$

and further reduced to

$$\sigma^2 = \mathrm{E}\,\frac{1}{n}\mathbf{u}'\mathbf{u} - 2\rho\,\mathrm{E}\,\frac{1}{n}\mathbf{u}'\mathbf{W}\mathbf{u} + \rho^2\,\mathrm{E}\,\frac{1}{n}\mathbf{u}'\mathbf{W}'\mathbf{W}\mathbf{u}.$$

Similar logic applies to the other moment conditions

$$\mathrm{E}\,\frac{1}{n}\boldsymbol{\epsilon}'\mathbf{W}'\mathbf{W}\boldsymbol{\epsilon} = \sigma^2\frac{1}{n}\,\mathrm{trace}\,\mathbf{W}'\mathbf{W}$$

and

$$\mathrm{E}\,\frac{1}{n}\boldsymbol{\epsilon}'\mathbf{W}\boldsymbol{\epsilon} = 0,$$

the latter moment stemming from the fact that the diagonal elements of $\mathbf{W}$ are all zero. When all substitutions for $\boldsymbol{\epsilon}$ are made, we arrive at a three-equation system in two unknown parameters,

$$\mathrm{E}\,\frac{1}{n}\begin{bmatrix}\mathbf{u}'\mathbf{u}\\ \mathbf{u}'\mathbf{W}'\mathbf{W}\mathbf{u}\\ \mathbf{u}'\mathbf{W}\mathbf{u}\end{bmatrix} = \mathrm{E}\,\frac{1}{n}\begin{bmatrix}2\mathbf{u}'\mathbf{W}\mathbf{u} & -\mathbf{u}'\mathbf{W}\mathbf{W}\mathbf{u} & n\\ 2\mathbf{u}'\mathbf{W}'\mathbf{W}'\mathbf{W}\mathbf{u} & -\mathbf{u}'\mathbf{W}'\mathbf{W}'\mathbf{W}\mathbf{W}\mathbf{u} & \mathrm{trace}\,\mathbf{W}'\mathbf{W}\\ \mathbf{u}'\mathbf{W}\mathbf{W}\mathbf{u}+\mathbf{u}'\mathbf{W}'\mathbf{W}\mathbf{u} & -\mathbf{u}'\mathbf{W}'\mathbf{W}\mathbf{W}\mathbf{u} & 0\end{bmatrix}\begin{bmatrix}\rho\\ \rho^2\\ \sigma^2\end{bmatrix}$$

or, in a more compact notation, $\mathbf{g} = \mathbf{G}\gamma$ with the dimension of $\mathbf{g}$ being 3×1 and the $\mathbf{G}$ matrix having dimension 3×3.

As written above, the equalities hold exactly. The problem, of course, is that $\mathbf{u}$ is unobservable. Kelejian and Prucha (1999) prove that it is permissible to insert the ordinary least-squares residual $\hat{\mathbf{u}}$ in its place. Making this substitution and dropping the expectations operator, we can summarize the empirical moments in the form

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\gamma + \text{residuals}.$$

The parameters ρ and σ^2 in γ can be estimated by minimizing the sum of squares function

$$S = \left(\hat{\mathbf{g}} - \hat{\mathbf{G}}\gamma\right)'\left(\hat{\mathbf{g}} - \hat{\mathbf{G}}\gamma\right)$$

using Gauss-Newton regression or a similar nonlinear least squares algorithm. However, the estimate of ρ must fall within the admissible range for this parameter as determined by the eigenvalues of the weight matrix, as we will discuss below.

With an acceptable $\hat{\rho}$ in hand, the feasible GLS estimator of β is implemented in the manner suggested above, that is, by "spatially filtering" $\mathbf{Y}$ and $\mathbf{X}$ using $\mathbf{I} - \hat{\rho}\mathbf{W}$ as in equation (25.5) and applying least squares to the transformed data. Note that a second estimator of σ^2 can be derived from the feasible GLS residuals, so that the method-of-moments estimator of this parameter can be discarded.

Kelejian and Prucha (1999) conduct an extensive Monte Carlo study of the properties of the moment estimator of ρ, finding that $\hat{\rho}$ performs about as well as the maximum-likelihood estimator (to be described below). The great advantage of the method-of-moments approach to estimating ρ is that it can be applied in very large samples. In addition to the trace of $\mathbf{W}'\mathbf{W}$, which equals $\sum_i \sum_j w_{ij}^2$, only three n-vectors are needed to construct the empirical moments: the least-squares residual vector $\hat{\mathbf{u}}$ and the two vectors derived from it, $\hat{\mathbf{u}}_1 = \mathbf{W}\hat{\mathbf{u}}$ and $\hat{\mathbf{u}}_2 = \mathbf{W} \cdot \hat{\mathbf{u}}_1$. These vectors can be computed at low cost.

25.3 Maximum Likelihood Approaches

To assemble the likelihood function, we first need to find the Jacobian of the transformation from the unobserved disturbances $\boldsymbol{\epsilon}$ to the observed $\mathbf{Y}$ variables, and equation (25.5) provides this. As Anselin (1988, p. 182) and others have shown, with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ the log-likelihood function for the spatial error model is then

$$L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 + \ln \mid \mathbf{I} - \rho\mathbf{W} \mid -\frac{1}{2\sigma^2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon}. \tag{25.7}$$

In this equation, $\boldsymbol{\epsilon} = \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta$ and the term $\ln \mid \mathbf{I} - \rho\mathbf{W} \mid$ is the log of the absolute value of the determinant, in other words, the log of the Jacobian.

The first-order conditions of the model are as follows (see Burridge (1980) or the appendix of this paper for more detail):

$$\frac{\partial L}{\partial \beta} = \frac{1}{\sigma^2} \cdot \mathbf{X}'(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{0}_{k\times 1}, \tag{25.8}$$

$$\frac{\partial L}{\partial \sigma^2} = -\frac{1}{2\sigma^2}\left(n + \frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)'(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\beta)\right) = 0, \tag{25.9}$$

and

$$\frac{\partial L}{\partial \rho} = \frac{\partial \ln \mid \mathbf{I} - \rho\mathbf{W} \mid}{\partial \rho} \quad + \\ \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)'\Big((\mathbf{I} - \rho\mathbf{W})'\mathbf{W} + \mathbf{W}'(\mathbf{I} - \rho\mathbf{W})\Big)(\mathbf{Y} - \mathbf{X}\beta) = 0. \tag{25.10}$$

As we will discuss below, the last of these equations provides the basis for a Lagrange Multiplier (or "score") test of the null hypothesis that $\rho = 0$.

The asymptotic variance matrix of the estimators has $k + 2$ rows and columns, and is block-diagonal with two blocks: the first corresponds to the k parameters of β and the second to the pair of parameters (σ^2, ρ). (See Anselin and Bera (1998, p. 258) or the appendix to this paper.) For the β vector the limiting variance matrix is estimated as $\mathcal{I}_\beta = \sigma^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}$ just as in a generalized least-squares model. The 2×2 block for σ^2 and ρ is

$$\mathcal{I}_{\sigma^2,\rho} = \begin{bmatrix} \frac{n}{2\sigma^4} & \frac{1}{\sigma^2}\operatorname{trace}\bar{\mathbf{W}} \\ \frac{1}{\sigma^2}\operatorname{trace}\bar{\mathbf{W}} & \operatorname{trace}(\bar{\mathbf{W}}\bar{\mathbf{W}}) + \operatorname{trace}(\bar{\mathbf{W}}'\bar{\mathbf{W}}) \end{bmatrix}^{-1} \tag{25.11}$$

in which $\bar{\mathbf{W}} \equiv \mathbf{W}\left(\mathbf{I} - \rho\mathbf{W}\right)^{-1}$.

Unfortunately, for large n it may be difficult to compute this last block of terms. Even if $\mathbf{W}$ itself contains relatively few non-zero elements (allowing it to be stored efficiently), the matrix $\left(\mathbf{I} - \rho\mathbf{W}\right)^{-1}$ will not be similarly sparse (Anselin and Bera 1998, p. 261). To skirt the problem, Anselin and others advocate the use of likelihood-ratio tests for ρ rather than Wald tests, which would require the 2×2 block of the information matrix. Wald tests can still be used for hypotheses about β, however, since these tests involve only $\sigma^2\left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}$, a $k \times k$ matrix.

Although the maximum-likelihood first-order equations can be solved iteratively, another approach is simpler computationally and seems to work well in practice. Imagine that ρ is known. The log-determinant term is then irrelevant to the maximizing values of β and σ^2 and, given ρ, we have the conditional GLS-like estimator

$$\tilde{\beta} = \left(\tilde{\mathbf{X}}'\tilde{\mathbf{X}}\right)^{-1}\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}.$$

Using the "spatially filtered" residual

$$\tilde{\mathbf{e}} \equiv \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\tilde{\beta} = (\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\tilde{\beta}),$$

we obtain the conditional estimator for the variance,

$$\tilde{\sigma}^2 = \frac{1}{n}\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \frac{1}{n}(\mathbf{Y} - \mathbf{X}\tilde{\beta})'(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\tilde{\beta}).$$

With the $\tilde{\beta}$ and $\tilde{\sigma}^2$ estimators inserted in equation (25.7), the likelihood function reduces to

$$L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\tilde{\sigma}^2 + \ln \mid \mathbf{I} - \rho\mathbf{W} \mid -\frac{n}{2}. \tag{25.12}$$

As Anselin (1988, pp. 110, 182) notes, this is a concentrated likelihood function that depends directly only on the ρ parameter.

The function can be computed for each distinct value of ρ in a grid of possible values. The largest likelihood locates an initial estimate of ρ, and this estimate can be further refined using another grid with more closely-spaced points. Pace (2000a) and Pace (2000b) has explored this approach in large-n models.[2] All that is needed to implement the method is a tractable expression for the log-determinant.

Ord's Simplification

Ord (1975) showed that

$$\mid \mathbf{I} - \rho\mathbf{W} \mid = \prod_{i=1}^{n}(1 - \rho\lambda_i), \tag{25.13}$$

with λ_i being the i-th eigenvalue of $\mathbf{W}$. He derived the result in the following way. The characteristic equation defining eigenvalues is a polynomial, and by manipulating the equation one can show that for a scalar a,

$$\mid a\mathbf{I} - \mathbf{W} \mid = \prod_{i=1}^{n}(a - \lambda_i) = \left(\prod_{i=1}^{n}(1 - \frac{1}{a}\lambda_i)\right) \cdot a^n$$

Now, recall that for square matrices $\mathbf{B}$ and $\mathbf{C}$, the determinant

$$\mid \mathbf{BC} \mid = \mid \mathbf{B} \mid \cdot \mid \mathbf{C} \mid .$$

To apply the result to the case at hand, let $\mathbf{Z} = \frac{1}{a}\mathbf{I}$ and note that $\mid \mathbf{Z} \mid = a^{-n}$. Also,

$$\mathbf{Z}(a\mathbf{I} - \mathbf{W}) = \mathbf{I} - \frac{1}{a}\mathbf{W}$$

and thus

$$\mid \mathbf{I} - \frac{1}{a}\mathbf{W} \mid = a^{-n} \cdot \left(\prod_{i=1}^{n}(1 - \frac{1}{a}\lambda_i)\right) \cdot a^n = \prod_{i=1}^{n}(1 - \frac{1}{a}\lambda_i).$$

Letting $\rho = a^{-1}$ completes the proof.

[2] Anselin (1988, pp. 109–110) prefers an alternative back-and-forth method that proceeds as follows. Beginning with $\rho = 0$ and using the simple maximum-likelihood estimates of β and σ^2, insert these estimates in the concentrated likelihood function (25.12) and find the $\tilde{\rho}$ value that maximizes this function. Then, with $\tilde{\rho}$ in hand, re-transform $\mathbf{Y}$ and $\mathbf{X}$ using $(\mathbf{I} - \tilde{\rho}\mathbf{W})$, find new $\tilde{\beta}$ and $\tilde{\sigma}^2$ values, and repeat the procedure until convergence is achieved.

Ord's result is useful in three ways. First, it provides us with a closed-form expression for the log-determinant, $\sum_{i=1}^{n} \ln(1 - \rho\lambda_i)$, and because the task of finding the $\{\lambda_i\}$ eigenvalues can be completed before model estimation begins, this expression can be programmed as a simple function of the ρ parameter alone. The Ord method also provides us with an explicit representation of the derivative of the log-determinant with respect to ρ,

$$\frac{\partial \ln \mid \mathbf{I} - \rho\mathbf{W} \mid}{\partial \rho} = -\sum_i \frac{\lambda_i}{1 - \rho\lambda_i}.$$

The ML first-order condition for ρ is thus

$$\frac{\partial L}{\partial \rho} = -\sum_i \frac{\lambda_i}{1 - \rho\lambda_i} \quad + \\ \frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)'\Big((\mathbf{I} - \rho\mathbf{W})'\mathbf{W} + \mathbf{W}'(\mathbf{I} - \rho\mathbf{W})\Big)(\mathbf{Y} - \mathbf{X}\beta) = 0. \quad (25.14)$$

Third and most important, Ord's result establishes a range in which ρ must lie. To allow the log of the determinant to be calculated, each $1 - \rho\lambda_i$ term must exceed zero, and therefore for all i we must have $1 > \rho\lambda_i$. Because $\mathbf{W}$ has zeroes on its diagonal, yielding trace $\mathbf{W} = 0$, and we also have trace $\mathbf{W} = \sum \lambda_i$, some of the eigenvalues of $\mathbf{W}$ must be positive and others negative. Hence, considering the largest eigenvalue (necessarily positive) and the smallest (negative), the following bounds apply to the ρ parameter: $\rho > 1/\lambda_{min}$ and $\rho < 1/\lambda_{max}$. In practice negative values for ρ are seldom seen, and the lower bound on ρ is often set to zero to speed grid searches on this parameter.

Of course, if we are to make use of Ord's formula, we must be sure that the eigenvalues $\{\lambda_i\}$ are real numbers. This is guaranteed to be the case when the weight matrix $\mathbf{W}$ is symmetric. When $\mathbf{W}$ is asymmetric, however, there is a possibility of complex eigenvalues and this generally makes it necessary to set Ord's approach aside and pursue a different method for finding $\mid \mathbf{I} - \rho\mathbf{W} \mid$. The one exception to the rule, which we now discuss, is when $\mathbf{W}$ is a row-standardized version of an otherwise symmetric weight matrix $\mathbf{W}_1$.

Row-Standardization

Addressing this important special case, Ord (1975) showed that if we begin with a symmetric weight matrix $\mathbf{W}_1$, but wish to use a row-standardized version of this matrix in the spatial regression, we can do so even though the row-standardized version is not itself symmetric.[3] The proof is as follows.

Let $\mathbf{s} = \mathbf{W}_1\boldsymbol{\iota}$, an $n \times 1$ vector of the row sums of $\mathbf{W}_1$, with $\boldsymbol{\iota}$ a column vector of ones. (Note that because $\mathbf{W}_1$ is symmetric, its row and column sums are the same and the latter are faster to calculate in a column-oriented programming language such as Fortran.) Then,

[3] To see this, consider the case of a 3-area model in which area 1 is neighbored by the two other areas (call them areas 2 and 3) but the only neighbor of area 2 is area 1 and area 3 likewise has only area 1 as a neighbor. The unstandardized weight matrix is symmetric, but the row-standardized version is not.

defining the diagonal matrix

$$\mathbf{D} = \begin{bmatrix} s_1^{-1} & 0 & \cdots & 0 \\ 0 & s_2^{-1} & 0 & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & s_n^{-1} \end{bmatrix},$$

$\mathbf{W} = \mathbf{D}\mathbf{W}_1$ is the row-standardized version of $\mathbf{W}_1$. Because the original weight matrix $\mathbf{W}_1$ is symmetric, it can be shown that the eigenvalues of the asymmetric $\mathbf{W}$ are all real and in fact, they equal the eigenvalues of the symmetric matrix $\mathbf{W}^* = \mathbf{D}^{1/2}\mathbf{W}_1\mathbf{D}^{1/2}$, which can be easily found using standard numerical routines for symmetric matrices.

The proof relies on the well-known theorem—see for instance Schott (1997, p. 88)—that the eigenvalues of the square matrix $\mathbf{A}$ and of $\mathbf{B}\mathbf{A}\mathbf{B}^{-1}$ are the same (for conformable and invertible $\mathbf{B}$). To implement this theorem, we let $\mathbf{A} \equiv \mathbf{W} = \mathbf{D}\mathbf{W}_1$, which is the weight matrix we want to use in our regression, and take $\mathbf{B} = \mathbf{D}^{-1/2}$. Then, with $\mathbf{B}\mathbf{A}\mathbf{B}^{-1} \equiv \mathbf{W}^*$, we have

$$\mathbf{W}^* = \mathbf{D}^{-1/2} \cdot (\mathbf{D}\mathbf{W}_1) \cdot \mathbf{D}^{1/2} = \mathbf{D}^{1/2}\mathbf{W}_1\mathbf{D}^{1/2},$$

which is a symmetric matrix. In short, there is no price to be paid for using a row-standardized version $\mathbf{W}$ of the original weight matrix $\mathbf{W}_1$ so long as we remember to use $\mathbf{W}^*$ and not $\mathbf{W}$ when we calculate eigenvalues with algorithms that assume symmetry. In the concluding section of this chapter, we discuss how to compute $\mathbf{W}^*$ efficiently.

Note that for a row-standardized weight matrix, 1 must be an eigenvalue because $\mathbf{W}\boldsymbol{\iota} = \boldsymbol{\iota}$. Indeed, by the famous Perron–Frobenius theorem, which applies to square matrices with non-negative entries of the kind encountered in Markov chains (including models of population growth), we know that 1 is the maximum eigenvalue. All other eigenvalues lie below 1 in absolute value, so that -1 is the minimum possible eigenvalue.[4] The admissible range for the ρ parameter is therefore $-1 < \rho < 1$ in the row-standardized case.

Because the $(-1, 1)$ bounds on ρ are known *a priori*, there is no need to calculate minimum and maximum eigenvalues, as we would have to do to fix bounds on ρ in the general case. This is one of the features of row-standardization that accounts for its popularity, and for the reluctance evident in the literature to explore alternative specifications. In the general case, if n is not too large, finding the largest and smallest eigenvalues should present little difficulty (with $\rho > 0$, the maximum eigenvalue is of greater interest). For large n, however, this can be a demanding computational task.

Alternative Log-Determinant Methods

According to Smirnov and Anselin (2001), for $n > 1000$ the direct computation of eigenvalues required to implement Ord's method is numerically too unstable to be trusted.[5] In a problem of this size, one can turn to Cholesky or LU decompositions of the $\mathbf{I} - \rho\mathbf{W}$ matrix to find the log of the determinant. Neither approach requires eigenvalues to be

[4] Smirnov and Anselin (2001, pp. 305–306) seem to dispute this last point, but I do not understand why. The result would appear to be well-established.

[5] Anderson et al. (1999) and Barker et al. (2001) present a discussion of error bounds in computatations using LAPACK. I need to explore this further.

calculated, although the bounds $1/\lambda_{min} < \rho < 1/\lambda_{max}$ must still somehow be respected.[6] A disadvantage of these decompositions is that they apply to the matrix $\mathbf{I} - \rho\mathbf{W}$ as a whole, and therefore needed to be executed for each value of ρ that is considered during estimation.

Cholesky and LU decompositions

The Cholesky decomposition is evidently the method of choice when $\mathbf{W}$ is either symmetric or a row-standardized version of a symmetric matrix. Suppose that $\mathbf{W}$ is symmetric. The Cholesky decomposition finds the lower triangular matrix $\mathbf{L}$ such that $\mathbf{I} - \rho\mathbf{W} = \mathbf{L}\mathbf{L}'$. Hence,

$$\mid \mathbf{I} - \rho\mathbf{W} \mid = \mid \mathbf{L} \mid\mid \mathbf{L}' \mid = \mid \mathbf{L} \mid^2$$

since the determinant of a matrix is the same as the determinant of its transpose.[7] Moreover, because $\mathbf{L}$ is lower triangular, $\mid \mathbf{L} \mid$ is simply the product of its diagonal elements. The log of the Jacobian term simplifies to

$$\log \mid \mathbf{I} - \rho\mathbf{W} \mid = 2\sum_{i=1}^{n} \ln \mid L_{ii} \mid,$$

which is easily programmed.

To handle the row-standardized case, we again write $\mathbf{W} = \mathbf{D}\mathbf{W}_1$ with $\mathbf{W}_1$ symmetric. Note that

$$\mathbf{I} - \rho\mathbf{W} = \mathbf{I} - \rho\mathbf{D}\mathbf{W}_1 = \mathbf{D}^{1/2}\left(\mathbf{I} - \rho\mathbf{D}^{1/2}\mathbf{W}_1\mathbf{D}^{1/2}\right)\mathbf{D}^{-1/2},$$

and, from this,

$$\begin{aligned} \mid \mathbf{I} - \rho\mathbf{W} \mid &= \mid \mathbf{D}^{1/2} \mid\mid \mathbf{I} - \rho\mathbf{D}^{1/2}\mathbf{W}_1\mathbf{D}^{1/2} \mid\mid \mathbf{D}^{-1/2} \mid \\ &= \mid \mathbf{I} - \rho\mathbf{D}^{1/2}\mathbf{W}_1\mathbf{D}^{1/2} \mid \\ &= \mid \mathbf{I} - \rho\mathbf{W}^* \mid \end{aligned}$$

with $\mathbf{I} - \rho\mathbf{W}^*$ being symmetric. In this case, we would decompose $\mathbf{I} - \rho\mathbf{W}^*$ rather than $\mathbf{I} - \rho\mathbf{W}$.

The LU decomposition of $\mathbf{I} - \rho\mathbf{W}$ is worth considering when the weight matrix is fundamentally asymmetric, and it handles the symmetric case as well, although not as efficiently as the Cholesky decomposition. In this case the determinant is given by the product of the diagonal elements of $\mathbf{U}$, the upper triangular matrix returned by the LU routine, and the log of the Jacobian is the sum of the logs of the absolute values of these elements.[8]

The Cholesky and LU decompositions are better-behaved than eigenvalue computations in large problems, but they, too, require the storage of the $n \times n$ matrix $\mathbf{I} - \rho\mathbf{W}$ (or its equivalent) and this may be infeasible or impractical when the sample size is very large.

[6] MM: Explore what happends to the LU decomposition if ρ strays outside the bounds.

[7] For the Cholesky decomposition to be applied, we must have $\mathbf{I} - \rho\mathbf{W}$ positive semi-definite; see Schott (1997, p. 139).

[8] Although the diagonal elements U_{ii} of $\mathbf{U}$ can be positive or negative, and as a consequence so can the determinant, the Jacobian equals the absolute value of the determinant and we have $\mid \prod_{i=1}^{n} U_{ii} \mid = \prod_{i=1}^{n} \mid U_{ii} \mid$. Hence, $\ln \mid I - \rho\mathbf{W} \mid = \sum_i \ln \mid U_{ii} \mid$.

Fortunately, there exist implementations of the Cholesky and LU decompositions that make use of sparse matrix methods, which require only the storage of the non-zero entries of this matrix (O'Leary 2005; O'Leary 2006).[9]

The Barry–Pace method

The difficulties entailed in setting up such sparse-matrix routines have motivated a search for alternative approximation methods that have a very different character. Barry and Pace (1999) have discovered an algorithm by which the log of the determinant is estimated with good precision through Monte Carlo simulation.[10] I believe the essence of the algorithm is as follows.

Given Ord's expression for the log-determinant, $\sum_{i=1}^{n} \ln(1 - \rho\lambda_i)$, consider Taylor-expanding each of the $\ln(1 - \rho\lambda_i)$ components. Recall that for the function $\ln(1 + x)$, a fortuitous cancellation of terms yields a very simple expression for a Taylor series expansion around $x = 0$,

$$\ln(1 + x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots,$$

provided that $\mid x \mid < 1$. Substituting $-\rho\lambda_i$ for x (note here the connection to the $\rho < 1/\lambda_{max}$ condition), we obtain

$$\ln(1 - \rho\lambda_i) = -\sum_{k=1}^{\infty} \frac{\rho^k \lambda_i^k}{k}.$$

For this infinite sum to converge, we must have $\mid \rho\lambda_i \mid < 1$, and this condition is met if both ρ and all λ_i are less than one in absolute value. The Barry–Pace method therefore applies when the weight matrix is a row-standardized version of a symmetric matrix; there may be other special cases that also meet the $\mid \rho\lambda_i \mid < 1$ condition.

---

[9] Pace (2000a) expresses concern about the performance of the standard sparse matrix algorithms used in these decompositions,

> The main problem with the LU or Cholesky decompositions used to find the determinant in n by n problems [is] the sensitivity of these procedures to the pattern of non-zeroes.

In his reference to "sensitivity," it seems that Pace is saying that efficiency requires an acceptable ordering of the non-zero elements of a sparse matrix, and evidently finding the right ordering is a non-trivial task. LeSage (1999, pp. 57–63) does not share his concern, at least with respect to speed and memory use. This issue needs further investigation.

[10] A variant of the Monte Carlo method evidently allows asymmetric spatial weight matrices to be specified (Pace 2000b). [MM: check code to see exactly what kind of asymmetry is allowed. The usual row-standardized stuff only? Also, at the end of the article Barry and Pace suggest scaling the weight matrix by **W**'s largest eigenvalue (its "spectral radius") in cases in which it can't be assumed that the eigenvalues are bounded by one. But finding that largest eigenvalue would be a problem, presumably, when n is very large.]

Insert the sum into Ord's expression for the log-determinant,

$$\begin{aligned} -\sum_{i=1}^{n}\sum_{k=1}^{\infty}\frac{\rho^k\lambda_i^k}{k} &= -\sum_{k=1}^{\infty}\sum_{i=1}^{n}\frac{\rho^k\lambda_i^k}{k} \\ &= -\sum_{k=1}^{\infty}\frac{\rho^k}{k}\sum_{i=1}^{n}\lambda_i^k \\ &= -\sum_{k=1}^{\infty}\frac{\rho^k}{k}\operatorname{trace}(\mathbf{W}^k). \end{aligned}$$

The key to the Barry–Pace approach is to approximate the first m terms of this infinite sum with a sample average whose expected value equals these terms. The remaining $k = m+1, \ldots, \infty$ terms form one part of the approximation error of the method; randomness in the sample average is the other source of error. Both sources of error can be reduced by increasing m and by generating more replications from which the sample average is computed.

We note in passing that the $k = 1$ term in the sum is zero because $\operatorname{trace}\mathbf{W} = 0$ and (as will be discussed in the next section) we can readily calculate the second term as $\operatorname{trace}(\mathbf{W}^2) = \sum_i \sum_j w_{ij}w_{ji}$. For these two terms no approximation is really necessary. Given that $\mathbf{W}$ is $n \times n$, however, it is difficult to compute exactly the traces of the higher powers, and here is where Monte Carlo averaging helps out.

To approximate the first m terms of the Taylor expansion, Barry and Pace consider the random variable

$$T = \sum_{k=1}^{m}\frac{\mathbf{u}'\mathbf{W}^k\mathbf{u}}{\mathbf{u}'\mathbf{u}},$$

in which $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, an $n \times 1$ vector of independent standard normal random variables. They note that for a given k,

$$\mathrm{E}\,\frac{\mathbf{u}'\mathbf{W}^k\mathbf{u}}{\mathbf{u}'\mathbf{u}} = \frac{1}{n}\operatorname{trace}(\mathbf{W}^k),$$

a result that follows from the properties of the Dirichlet distribution and an orthogonal decomposition of $\mathbf{W}$ (see the excellent appendix to their paper). Therefore

$$\mathrm{E}\big(-n \cdot T\big) = -\sum_{k=1}^{m}\operatorname{trace}(\mathbf{W}^k).$$

This is very close to the expression we want. Re-formulating in terms of

$$V = -n\sum_{k=1}^{m}\frac{\rho^k}{k}\cdot\frac{\mathbf{u}'\mathbf{W}^k\mathbf{u}}{\mathbf{u}'\mathbf{u}},$$

we obtain a quantity whose expected value is exactly the first m terms of the Taylor expansion. Moreover, because the same standard normal vector $\mathbf{u}$ is employed throughout the calculation of V, there is no need to compute the $\mathbf{W}^k$ matrix directly. We can store the vector $\mathbf{c}_1 = \mathbf{W}\mathbf{u}$ and then pre-multiply it by $\mathbf{W}$ to find $\mathbf{c}_2 = \mathbf{W}\mathbf{c}_1 = \mathbf{W}^2\mathbf{u}$. The same procedure can

be executed again to find $\mathbf{c}_3 = \mathbf{W}^3\mathbf{u} = \mathbf{W}\mathbf{c}_2$, and so on to the m-th power. To carry out this recursion, we require only $\mathbf{W}$ and storage space for two n-vectors $\mathbf{c}_{new}$ and $\mathbf{c}_{old}$.[11]

Let `reps` indicate the number of times this procedure is repeated, with the r-th repetition using an independent draw of the $\mathbf{u}$ standard normal vector. Proceeding as above, we generate `reps` independent random variables $\{V_r\}$ and finish up by averaging them to approximate the expected value. Barry and Pace (1999) establish error bounds on the approximation which are remarkably easy to program.

A very attractive feature of the method is that the log-determinant can be approximated for many values of ρ almost simultaneously, at very little computational cost. Store in an m-vector $\mathbf{V}$ the random elements

$$\mathbf{V} = \begin{bmatrix} \mathbf{u}'\mathbf{W}\mathbf{u} \\ \mathbf{u}'\mathbf{W}^2\mathbf{u} \\ \vdots \\ \mathbf{u}'\mathbf{W}^m\mathbf{u} \end{bmatrix}$$

and in a $G \times m$ matrix $\mathbf{R}$ store the various powers of ρ (appropriately normalized) for a grid of values $\rho_1, \rho_2, \ldots, \rho_G$,

$$\begin{bmatrix} \rho_1 & \rho_1^2/2 & \ldots & \rho_1^m/m \\ \rho_2 & \rho_2^2/2 & \ldots & \rho_2^m/m \\ \vdots & \vdots & & \vdots \\ \rho_G & \rho_G^2/2 & \ldots & \rho_G^m/m \end{bmatrix}.$$

Then using

$$-\frac{n}{\mathbf{u}'\mathbf{u}} \cdot \mathbf{R}\mathbf{V}$$

we obtain estimates of the log-determinant for all G values of the ρ parameter. These can be averaged over the `reps` repetitions of the Monte Carlo algorithm.

Pace and LeSage (2004) have developed another approximation method using Chebyshev polynomials for the special case in which $\rho \in [0, 1)$. This approach does not require Monte Carlo simulation. Also deserving of mention are two additional algorithms, that of Pace and LeSage (2000) for row-standardized, nearest-neighbor weight matrices and the method of Smirnov and Anselin (2001)

25.4 Lagrange Multiplier Tests of $\rho = 0$

The heart of the test statistic is given by the first-order condition for ρ evaluated at $\rho = 0$. The derivative of the log-determinant term with respect to ρ reduces to $-\sum_i \lambda_i$ with $\rho = 0$ and because $-\sum_i \lambda_i = -\operatorname{trace}\mathbf{W}$ this equals zero.[12] The other term in the first-order condition simplifies to

$$\frac{1}{2\hat{\sigma}^2}\mathbf{e}'\left(\mathbf{W} + \mathbf{W}'\right)\mathbf{e} = \frac{1}{\hat{\sigma}^2}\mathbf{e}'\mathbf{W}\mathbf{e},$$

[11] Having cycled through the recursions for $k = 1, \ldots, m$ in this way, LeSage (1999) then overwrites the $k = 1$ term with zero and the $k = 2$ term with the trace of $\mathbf{W}\mathbf{W}$ to obtain greater precision.

[12] MM: Check into what happens if $\mathbf{W}$ is asymmetric with possibly complex eigenvalues.

in which $\mathbf{e}$ is the residual vector (equivalent to the residual in an ordinary least-squares model) and $\hat{\sigma}^2 = n^{-1}\mathbf{e}'\mathbf{e}$ is the estimate of the variance (also equivalent to the least-square estimate except that the denominator is n rather than $n - k$). The equality holds because $\mathbf{e}'\mathbf{W}\mathbf{e}$ is a scalar and equals its transpose $\mathbf{e}'\mathbf{W}'\mathbf{e}$.

When it is evaluated at $\rho = 0$, the 2×2 block of the information matrix becomes diagonal, and the entry for ρ simplifies to $\text{trace}(\mathbf{W}'\mathbf{W}) + \text{trace}(\mathbf{W}\mathbf{W})$. Assembling all these components, Anselin and Bera (1998, p. 270) arrive at an expression for the Lagrange Multiplier test statistic,

$$T = \left[\frac{\mathbf{e}'\mathbf{W}\mathbf{e}}{\frac{\mathbf{e}'\mathbf{e}}{n}}\right]^2 \cdot \frac{1}{\text{trace}(\mathbf{W}'\mathbf{W}) + \text{trace}(\mathbf{W}\mathbf{W})}.$$

Under the null, this test statistic converges in distribution to central chi-square with 1 degree of freedom.

As Anselin and Bera (1998, p. 280) note, the denominator of the factor on the right can be computed without carrying out the matrix multiplications, making use of the fact that $\text{trace}(\mathbf{W}'\mathbf{W}) = \sum_i \sum_j w_{ij}^2$ and $\text{trace}(\mathbf{W}\mathbf{W}) = \sum_i \sum_j w_{ij} w_{ji}$ as mentioned earlier. These sums depend only on the non-zero weights, which typically make up a low proportion of all the weight matrix entries.

In some cases further simplifications are available. When $\mathbf{W}$ is symmetric, the two traces are equal. Also, if all of the w_{ij} spatial weights are either 0 or 1, then $\text{trace}(\mathbf{W}'\mathbf{W}) = \sum_i \sum_j w_{ij}$, which is simply the total number of non-zero entries in the weight matrix. Hence, for a symmetric binary weight matrix, $\text{trace}(\mathbf{W}'\mathbf{W}) + \text{trace}(\mathbf{W}\mathbf{W})$ is given by twice the number of non-zero weights.

Alternatively, for row-standardized transformations of binary weight matrices, each non-zero $w_{ij} = s_i^{-1}$ with s_i being the row sum, that is, the number of ones (e.g., neighbors) in the i-th row of the unstandardized binary matrix. In this case, there are only s_i non-zero terms in $\sum_j w_{ij}^2$ and therefore

$$\begin{aligned}\text{trace}(\mathbf{W}'\mathbf{W}) &= \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}^2 = \sum_{i=1}^{n}\sum_{j=1}^{s_i} s_i^{-2} \\ &= \sum_i s_i s_i^{-2} = \sum_i s_i^{-1},\end{aligned}$$

which is easy to compute. The other trace does not simplify quite so neatly (we cannot appeal to symmetry in the case of row-standardized weights) but see Anselin and Bera (1998, p. 280) for further discussion.

25.5 Final Notes on Computation

Because $\mathbf{W}$ is an $n \times n$ matrix and $\mathbf{X}$ is $n \times k$, the number of operations implied by expressions such as $\tilde{\mathbf{I}} = \mathbf{I} - \rho\mathbf{W}$ and $\tilde{\mathbf{X}} = \mathbf{X} - \rho\mathbf{W}\mathbf{X}$ increases dramatically as the sample size n grows. These computational costs can be kept in check by storing only the non-zero entries of $\mathbf{W}$ and using these entries instead of the full $\mathbf{W}$ matrix where that is possible. There are

sophisticated sparse matrix libraries available for such problems, but the simpler methods discussed below are also worth considering.

Suppose that the weight matrix is created and stored in the form of an $n_{max} \times n$ `weight_index` matrix, such as

$$\begin{bmatrix} 5 & 3 & \cdots & 17 \\ 7 & 5 & \cdots & 35 \\ 0 & 12 & & 67 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & & 72 \\ 0 & 0 & \cdots & 78 \end{bmatrix},$$

whose i-th column contains the indices of all neighbors j whose spatial weights $w_{ij} > 0$, and with zeroes used to fill in the remainder of the column. (The number of rows n_{max} is the maximum number of neighbors for any area i in the dataset.) An isolated area without any neighbors would be represented in `weight_index` by a column of zeroes. The idea is to store in the i-th *column* of `weight_index` all relevant information about the i-th *row* of the **W** matrix.

In a Fortran program, the intrinsic function `COUNT` can be used to count the number of neighbors of area i, that is, those with positive indices in column i of the `weight_index` matrix. The `PACK` function will arrange these indices in a short vector called, say, `iwt`, which can be used as a vector of subscripts to simplify subsequent calculations.[13]

To see how this approach works, consider the case of binary weights. We can find the i-th entry of the $n \times 1$ vector **WY** by using the Fortran command `SUM (Y(iwt))`, which is exactly what would be found in the i-th row of the matrix–vector product **WY** in the binary weights case. The (i, j) entry of **WX** is likewise found as `SUM (X(iwt,j))`. If the weights are symmetric, the $n \times n$ weight matrix **W** can be found by initializing $\mathbf{W} = \mathbf{0}$ and then, for each column i, assigning ones to the correct rows by setting `W(iwt,i)`$= 1$. If the weights are not symmetric, **W** could either be filled in row-by-row with the code `W(i,iwt)`, although that is inefficient, or calculated column-by-column as in the symmetric case and then transposed.

A minor modification of the approach is needed for row-standardized binary weights, where for each i the number of neighbors is required. For the i-th entry of the $n \times 1$ vector **WY** we would simply divide `SUM (Y(iwt))` by the integer number of neighbors converted to a real value, that is, by `REAL(SIZE(iwt))` in Fortran, and would proceed similarly for the (i, j) entry of **WX**. For non-binary weights we would need to read in a companion matrix, say `weight_values`, organized in precisely the same way as `weight_index` but with the values of the weights as the entries.

The efficiency of these operations will depend to some degree on the numbering scheme that is used to label the spatial units. When neighboring areas are similarly numbered—as when **W** has most of its non-zero entries clustered about the diagonal—the vector of subscripts `iwt` will usually point to entries that are either adjacent or quickly fetched from cache memory. But even in the best of circumstances, a neighbor will occasionally be encountered whose index is very different from the other neighbors, and in such cases memory access will momentarily slow down. In the ideal numbering scheme, such access

[13] Not until Fortran 90 was it possible to make use of such vector subscripts.

penalties would not be faced very often. If they are common and unavoidable, however, more sophisticated sparse matrix techniques may be necessary.

Note that once $\mathbf{W}$, $\mathbf{WY}$, and $\mathbf{WX}$ are calculated, their values are fixed for the duration of the estimation. While estimation is underway, these matrices appear only in the contexts $\tilde{\mathbf{I}} = \mathbf{I} - \rho\mathbf{W}$, $\tilde{\mathbf{Y}} = \mathbf{Y} - \rho\mathbf{WY}$, and $\tilde{\mathbf{X}} = \mathbf{X} - \rho\mathbf{WX}$ and here we can expect the multiplication by the scalar ρ to be executed with high efficiency, especially by modern compilers with access to the `BLAS` library of optimized linear algebra routines.

In general it is important to arrange all calculations so as to exploit Fortran's column-by-column storage of arrays. For instance, the $k \times k$ matrix $\tilde{\mathbf{X}}'\tilde{\mathbf{X}}$, which figures into the conditional estimate of β given ρ, can be written as

$$\tilde{\mathbf{X}}'\tilde{\mathbf{X}} = \begin{bmatrix} \tilde{\mathbf{x}}_1' \\ \tilde{\mathbf{x}}_2' \\ \vdots \\ \tilde{\mathbf{x}}_k' \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{x}}_1 & \tilde{\mathbf{x}}_2 & \dots & \tilde{\mathbf{x}}_k \end{bmatrix},$$

in which $\tilde{\mathbf{x}}_i$ is the $n \times 1$ column vector that resides in the i-th column of $\tilde{\mathbf{X}}$. Despite appearances, there is no need to store $\tilde{\mathbf{X}}'$ to execute this operation—it can be carried out more efficiently via a series of dot-products $\tilde{\mathbf{x}}_i'\tilde{\mathbf{x}}_j$ using $\tilde{\mathbf{X}}$ alone, thereby taking advantage of Fortran's column storage method. Likewise, for the $k \times 1$ vector $\tilde{\mathbf{X}}'\tilde{\mathbf{Y}}$,

$$\begin{bmatrix} \tilde{\mathbf{x}}_1' \\ \tilde{\mathbf{x}}_2' \\ \vdots \\ \tilde{\mathbf{x}}_k' \end{bmatrix} \tilde{\mathbf{Y}} = \begin{bmatrix} \tilde{\mathbf{x}}_1'\tilde{\mathbf{Y}} \\ \tilde{\mathbf{x}}_2'\tilde{\mathbf{Y}} \\ \vdots \\ \tilde{\mathbf{x}}_k'\tilde{\mathbf{Y}} \end{bmatrix}.$$

As we discussed earlier, when the eigenvalues of row-standardized weight matrices are desired, these are most efficiently calculated by finding the eigenvalues of a symmetric matrix

$$\mathbf{W}^* = \mathbf{D}^{1/2}\mathbf{W}_1\mathbf{D}^{1/2}$$

whose eigenvalues are the same as those of the row-standardized matrix. Although $\mathbf{W}^*$ would seem to require the multiplication of three $n \times n$ matrices, it can be calculated in Fortran using much less costly column-by-column operations. Let $\mathbf{d}$, an $n \times 1$ vector, represent the diagonal of $\mathbf{D}^{1/2}$. Then the matrix product $\mathbf{D}^{1/2}\mathbf{W}_1$ is

$$\begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & d_n \end{bmatrix} \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & & w_{nn} \end{bmatrix} = \begin{bmatrix} w_{11}d_1 & w_{12}d_1 & \cdots & w_{1n}d_1 \\ w_{21}d_2 & w_{22}d_2 & \cdots & w_{2n}d_2 \\ \vdots & \vdots & & \vdots \\ w_{n1}d_n & w_{n2}d_n & & w_{nn}d_n \end{bmatrix}$$

and this can be calculated with a sequence of element-by-element multiplications of column $\mathbf{w}_i$ by $\mathbf{d}$. Using $\odot$ to denote element-by-element multiplication, for the i-th column we have

$$\mathbf{w}_i \odot \mathbf{d} = \begin{bmatrix} w_{1i} \cdot d_1 \\ w_{2i} \cdot d_2 \\ \vdots \\ w_{ni} \cdot d_n \end{bmatrix}$$

and

$$\mathbf{D}^{1/2}\mathbf{W}_1 = \begin{bmatrix}\mathbf{w}_1 \odot \mathbf{d} & \mathbf{w}_2 \odot \mathbf{d} & \cdots & \mathbf{w}_n \odot \mathbf{d}\end{bmatrix}.$$

Similarly, the matrix product $\mathbf{W}_1\mathbf{D}^{1/2}$ is

$$\begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{22} & \cdots & w_{2n} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & & w_{nn} \end{bmatrix} \begin{bmatrix} d_1 & 0 & \cdots & 0 \\ 0 & d_2 & & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & d_n \end{bmatrix} = \begin{bmatrix} w_{11}d_1 & w_{12}d_2 & \cdots & w_{1n}d_n \\ w_{21}d_1 & w_{22}d_2 & \cdots & w_{2n}d_n \\ \vdots & \vdots & & \vdots \\ w_{n1}d_1 & w_{n2}d_2 & & w_{nn}d_n \end{bmatrix},$$

and this simplifies to

$$\mathbf{W}_1\mathbf{D}^{1/2} = \begin{bmatrix}\mathbf{w}_1 \cdot d_1 & \mathbf{w}_2 \cdot d_2 & \cdots & \mathbf{w}_n \cdot d_n\end{bmatrix}$$

in which each column $\mathbf{w}_i$ is multiplied by the scalar d_i.

Appendix: The SEM First-Order Conditions and Information Matrix

To begin, we should remind ourselves of the notation used in matrix calculus. If L is a scalar function of a $k \times 1$ vector $\boldsymbol{\beta}$, then the derivative of L, $\partial L/\partial \boldsymbol{\beta}$, is defined as a $k \times 1$ column vector, the elements of which are the partial derivatives $\partial L/\partial \beta_i$, $i = 1, 2, \ldots, k$. The transpose, written $\partial L/\partial \boldsymbol{\beta}'$, is a $1 \times k$ row vector. If $\boldsymbol{\epsilon}$ is an $n \times 1$ vector function of a $k \times 1$ vector $\boldsymbol{\beta}$, then the derivative of $\boldsymbol{\epsilon}$, $\partial \boldsymbol{\epsilon}/\partial \boldsymbol{\beta}'$, is defined as the $n \times k$ matrix, the element of which are the partial derivatives $\partial \epsilon_i/\partial \beta_j$, $i = 1, \cdots, n$, $j = 1, \cdots, k$. And $\partial \boldsymbol{\epsilon}'/\partial \boldsymbol{\beta} = (\partial \boldsymbol{\epsilon}/\partial \boldsymbol{\beta}')'$. For example, let $\mathbf{X}$ be an $n \times k$ matrix, $\boldsymbol{\beta}$ be a $k \times 1$ vector, $\boldsymbol{\epsilon}$ be an $n \times 1$ vector, and $\mathbf{M}$ be an $n \times n$ matrix. The derivative, $\partial \mathbf{X}\boldsymbol{\beta}/\partial \boldsymbol{\beta}' = \mathbf{X}$. The derivative, $\partial \boldsymbol{\epsilon}' M \boldsymbol{\epsilon}/\partial \boldsymbol{\epsilon} = (\mathbf{M} + \mathbf{M}')\boldsymbol{\epsilon}$, and if $\mathbf{M}$ is symmetric, this simplifies to $\partial \boldsymbol{\epsilon}'\mathbf{M}\boldsymbol{\epsilon}/\partial \boldsymbol{\epsilon} = 2 \cdot \mathbf{M}\boldsymbol{\epsilon}$.

The first-order conditions

The log-likelihood function for the spatial error model is

$$L = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln \sigma^2 + \ln \mid \mathbf{I} - \rho\mathbf{W} \mid - \frac{1}{2\sigma^2}\boldsymbol{\epsilon}'\boldsymbol{\epsilon}.$$

with $\boldsymbol{\epsilon} = (\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. The derivative of L with respect to $\boldsymbol{\beta}$ is

$$\frac{\partial L}{\partial \beta} = -\frac{1}{2\sigma^2}\frac{\partial \boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{\partial \boldsymbol{\beta}} = \frac{1}{\sigma^2} \cdot \mathbf{X}'(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\beta)$$

in which we have used

$$\frac{\partial \boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{\partial \boldsymbol{\beta}} = 2 \cdot \left(\frac{\partial(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right)'(I - \rho\mathbf{W})'(I - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).$$

The partial with respect to σ^2 is trivial and therefore not repeated here.

As Abadir and Magnus (2005) show, for any symmetric and nonsingular matrix $\mathbf{X}$,

$$\frac{d \mid \mathbf{X} \mid}{d\mathbf{X}} = \mid \mathbf{X} \mid \cdot \operatorname{trace} \mathbf{X}^{-1} \; ; \; \frac{d\mathbf{X}^{-1}}{d\mathbf{X}} = -\mathbf{X}^{-1}\mathbf{X}^{-1}$$

Hence,

$$\begin{aligned}\frac{\partial L}{\partial \rho} &= \frac{\partial \ln \mid \mathbf{I} - \rho\mathbf{W} \mid}{\partial \rho} - \frac{1}{2\sigma^2}\frac{\partial \boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{\partial \rho} \\ &= -\operatorname{trace} \bar{\mathbf{W}} + \frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(I - \rho\mathbf{W})'\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\end{aligned}$$

where $\bar{\mathbf{W}} = \mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}$. The first term follows from

$$\frac{\partial \ln \mid \mathbf{I} - \rho\mathbf{W} \mid}{\partial \rho} = \operatorname{trace}(\mathbf{I} - \rho\mathbf{W})^{-1}\frac{\partial(\mathbf{I} - \rho\mathbf{W})}{\partial \rho} = -\operatorname{trace} \bar{\mathbf{W}}$$

The second term follows from

$$\begin{aligned}\frac{\partial \boldsymbol{\epsilon}'\boldsymbol{\epsilon}}{\partial \rho} &= -[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{W}'(\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \rho\mathbf{W})'\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})] \\ &= -2 \cdot [(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{I} - \rho\mathbf{W})'\mathbf{W}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})]\end{aligned}$$

The Information Matrix

The second derivative with respect to β is

$$\frac{\partial^2 L}{\partial\beta\partial\beta'} = -\frac{1}{\sigma^2} \cdot \mathbf{X}'(\mathbf{I} - \rho\mathbf{W})'(\mathbf{I} - \rho\mathbf{W})\mathbf{X} = -\frac{1}{\sigma^2}\bar{\mathbf{X}}'\bar{\mathbf{X}}$$

Hence, for the $\boldsymbol{\beta}$ vector the information matrix block is

$$-\operatorname{E}\left(\frac{\partial^2 L}{\partial\beta\partial\beta'}\right) = \frac{1}{\sigma^2}\operatorname{E}(\bar{\mathbf{X}}'\bar{\mathbf{X}}).$$

For the σ^2 parameter,

$$\frac{\partial^2 L}{(\sigma^2)^2} = \frac{n}{2\sigma^2} - \frac{1}{2\sigma^6}\boldsymbol{\epsilon}'\boldsymbol{\epsilon}$$

$$-\operatorname{E}\left(\frac{\partial^2 L}{(\sigma^2)^2}\right) = \frac{1}{2\sigma^6}\operatorname{E}(\boldsymbol{\epsilon}'\boldsymbol{\epsilon}) = \frac{n}{2\sigma^4}$$

For the ρ parameter,

$$\frac{\partial^2 L}{\partial\rho^2} = -\operatorname{trace} \bar{\mathbf{W}}\bar{\mathbf{W}} - \frac{1}{\sigma^2}(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{W}'\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)$$

The first term comes from

$$\frac{\partial \operatorname{trace} \mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}}{\partial\rho} = \operatorname{trace} \frac{\partial \mathbf{W}(\mathbf{I} - \rho\mathbf{W})^{-1}}{\partial\rho}$$

and

$$\frac{\partial(\mathbf{I}-\rho\mathbf{W})^{-1}}{\partial\rho} = -(\mathbf{I}-\rho\mathbf{W})^{-1}\frac{\partial(\mathbf{I}-\rho\mathbf{W})}{\partial\rho}(\mathbf{I}-\rho\mathbf{W})^{-1}.$$

The diagonal element for ρ is therefore

$$-\operatorname{E}\left(\frac{\partial^2 L}{\partial\rho^2}\right) = \operatorname{trace}\bar{\mathbf{W}}\bar{\mathbf{W}} + \frac{1}{\sigma^2}\operatorname{E}[(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{W}'\mathbf{W}(\mathbf{Y}-\mathbf{X}\beta)]$$

Notice that the expectation is applied to a scalar. Therefore

$$\begin{aligned}
\operatorname{E}[(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{W}'\mathbf{W}(\mathbf{Y}-\mathbf{X}\beta)] &= \operatorname{E}[\boldsymbol{\epsilon}'(\mathbf{I}-\rho\mathbf{W})^{-1'}\mathbf{W}'\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}] \\
&= \operatorname{E}[\operatorname{trace}\boldsymbol{\epsilon}\boldsymbol{\epsilon}'(\mathbf{I}-\rho\mathbf{W})^{-1'}\mathbf{W}'\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}] \\
&= \operatorname{trace}\operatorname{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')(\mathbf{I}-\rho\mathbf{W})^{-1'}\mathbf{W}'\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1} \\
&= \sigma^2\operatorname{trace}(\mathbf{I}-\rho\mathbf{W})^{-1'}\mathbf{W}'\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1} \\
&= \sigma^2\operatorname{trace}\bar{\mathbf{W}}'\bar{\mathbf{W}}
\end{aligned}$$

The first equality follows from $\mathbf{Y}-\mathbf{X}\beta = (\mathbf{I}-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}$. The second equality follows from one of properties of trace, that is, $\operatorname{trace}(AB) = \operatorname{trace}(BA)$ for any matrix A and B such that AB and BA exists. The fourth equality follows from the assumption $\operatorname{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma^2\mathbf{I}$. Therefore,

$$-\operatorname{E}\left(\frac{\partial^2 L}{\partial\rho^2}\right) = \operatorname{trace}\bar{\mathbf{W}}\bar{\mathbf{W}} + \operatorname{trace}\bar{\mathbf{W}}'\bar{\mathbf{W}}$$

As for the cross-product terms, we have

$$\frac{\partial^2 L}{\partial\beta\partial\sigma^2} = \frac{1}{\sigma^4}\mathbf{X}'(\mathbf{I}-\rho\mathbf{W})'(\mathbf{I}-\rho\mathbf{W})(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \frac{1}{\sigma^4}\bar{\mathbf{X}}'\boldsymbol{\epsilon}$$

and this yields

$$-\operatorname{E}\left(\frac{\partial^2 L}{\partial\beta\sigma^2}\right) = -\frac{1}{\sigma^4}\bar{\mathbf{X}}'\operatorname{E}(\boldsymbol{\epsilon}) = \mathbf{0}_{k\times 1}.$$

The next cross-product term is

$$\begin{aligned}
\frac{\partial^2 L}{\partial\beta\partial\rho} &= -\frac{1}{\sigma^2}\cdot[\mathbf{X}'\mathbf{W}'(\mathbf{I}-\rho\mathbf{W})(\mathbf{Y}-\mathbf{X}\beta) + \mathbf{X}'(\mathbf{I}-\rho\mathbf{W})'\mathbf{W}(\mathbf{Y}-\mathbf{X}\beta)] \\
&= -\frac{1}{\sigma^2}\cdot[\mathbf{X}'\mathbf{W}'\boldsymbol{\epsilon} + \bar{\mathbf{X}}'\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}]
\end{aligned}$$

and this too yields

$$-\operatorname{E}\left(\frac{\partial^2 L}{\partial\beta\partial\rho}\right) = \mathbf{0}_{k\times 1}.$$

The last cross-product term is

$$\frac{\partial^2 L}{\partial\sigma^2\partial\rho} = -\frac{1}{\sigma^4}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{I}-\rho\mathbf{W})'\mathbf{W}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = -\frac{1}{\sigma^4}\boldsymbol{\epsilon}'\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}$$

and

$$
\begin{aligned}
-\operatorname{E}\left(\frac{\partial^2 L}{\partial\sigma^2\partial\rho}\right) &= \frac{1}{\sigma^4}\operatorname{E}\operatorname{trace}(\boldsymbol{\epsilon}'\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}\boldsymbol{\epsilon}) = \frac{1}{\sigma^4}\operatorname{E}\operatorname{trace}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}) \\
&= \frac{1}{\sigma^4}\operatorname{trace}(\operatorname{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}) = \frac{1}{\sigma^2}\operatorname{trace}(\mathbf{W}(\mathbf{I}-\rho\mathbf{W})^{-1}) \\
&= \frac{1}{\sigma^2}\operatorname{trace}\bar{\mathbf{W}}
\end{aligned}
$$

Chapter 26

Quantile Regression

Conventional regression models estimate the conditional mean of the dependent variable and the variance of the disturbance term. In some circumstances, however, we want to know more about the distribution than means and variances can tell us. Studies of inequality often focus on the top-most or bottom-most deciles of the distribution of income and wealth; studies of school achievement often single out the worst-performing 10 or 25 percent of students for special attention. In cases such as these, it is very helpful to have access to a multivariate method that estimates the conditional percentiles of a distribution. Quantile regression is the leading example of such a method.

26.1 Background

To motivate our discussion of the quantile regression model, we begin with some background on medians and percentiles. (Many authors use the term quantile as a general label that applies to medians, percentiles, deciles, and so on.) It is easy to show that the median minimizes the quantity $\mathrm{E}\,|Y-\tau|$ over τ if we assume that Y is a continuous random variable (Hao and Naiman 2007). Let

$$\begin{aligned}\phi(\tau) \equiv \mathrm{E}\,|Y-\tau| &= \int_{-\infty}^{\tau} |y-\tau| f(y)dy + \int_{\tau}^{\infty} |y-\tau| f(y)dy \\ &= \int_{-\infty}^{\tau} (\tau - y) f(y)dy + \int_{\tau}^{\infty} (y-\tau) f(y)dy,\end{aligned}$$

which gives

$$\begin{aligned}\phi'(\tau) &= \int_{-\infty}^{\tau} f(y)dy - \int_{\tau}^{\infty} f(y)dy \\ &= F(\tau) - (1 - F(\tau)) = 2F(\tau) - 1.\end{aligned}$$

The ϕ function is convex (check the second derivative) and its minimum is the value of τ^* that satisfies $F(\tau^*) = 1/2$, which is obviously the median. Lee (1995) provides a more general proof that encompasses discrete random variables.

Quantiles other than the median can also be found by minimizing the expectation of absolute values. Consider the p-th percentile, with p expressed as a proportion (thus

lying between 0 and 1). We define the function $g(\tau) = (1-p)|Y-\tau|$ if $Y < \tau$ and $g(\tau) = p|Y-\tau|$ when $Y \geq \tau$ and let $\phi(\tau) = \mathrm{E}\, g(\tau)$. Following the same steps as above, we arrive at

$$\phi'(\tau) = (1-p)F(\tau) - p(1-F(\tau)) = F(\tau) - p,$$

and from this we can see that the minimizing τ^* is the p-th percentile.

There is a side result that may be of interest: When the mean and standard deviation both exist, the median must lie no further than one standard deviation away from the mean. Letting m denote the median of Y and μ the mean,

$$\begin{aligned} |\mu - m| &= |\,\mathrm{E}(Y-m)| \leq \mathrm{E}\,|Y-m| \leq \mathrm{E}\,|Y-\mu| \\ &= \mathrm{E}\,\sqrt{(Y-\mu)^2} \leq \sqrt{\mathrm{E}(Y-\mu)^2} = \sigma. \end{aligned}$$

The first of the inequalities results from Jensen's inequality (the absolute value function is a V-shaped convex function); the second arises because as we just showed, the median m minimizes $\mathrm{E}\,|(Y-\tau)|$ among all τ; and the third inequality again follows from Jensen's inequality because the square root function is concave.

There is another way by which we can link the mean μ to the median and the percentiles of the distribution. Letting $Y_i = \mu + \epsilon_i$, the p-th percentile is the value τ that satisfies the condition

$$p = \Pr(Y_i \leq \tau)$$

or what is equivalent,

$$p = \Pr(Y_i - \mu \leq \tau - \mu) = \Pr(\epsilon_i \leq \tau - \mu).$$

Denoting by F_ϵ the cdf of ϵ, we thus have

$$p = F_\epsilon(\tau - \mu)$$

and by inverting the cdf (this is possible if the density $f(\epsilon) > 0$ in the neighborhood of $\tau - \mu$) we obtain

$$\tau - \mu = F_\epsilon^{-1}(p).$$

We see that the p-th percentile can be expressed in terms of the mean plus an adjustment factor, i.e., $\tau = \mu + F_\epsilon^{-1}(p)$. The previous result on medians implies that $F_\epsilon^{-1}(0.5)$ can be no larger than σ_ϵ in absolute value.

These results have to do with expected values, but there are counterpart results for samples. In particular, the sample median $\hat{m}$ minimizes the average absolute deviation $\frac{1}{n}\sum_{i=1}^{n}|Y_i - \tau|$. This function is not differentiable in τ at points where $\tau = Y_i$, but it is continuous everywhere and convex in τ (being a sum of convex functions) and its minimum can be located without too much difficulty (Blyth 1990).

The proof begins by arranging the Y_i in order from smallest to largest. In a dataset having an odd number of observations, the observation corresponding to the sample median is the one that is numbered $(n+1)/2$ in this ordered series of data points. If the number of observations is even, the median is not unique—instead it is defined as the range of values from $Y_{n/2}$ to $Y_{1+n/2}$—although the median is uniquely defined in the special case that $Y_{n/2} = Y_{1+n/2}$.

Consider a point τ such that $Y_k < \tau < Y_{k+1}$, noting the strict inequalities. Let

$$\begin{aligned}\phi(\tau) &= \sum_{i=1}^{n} |Y_i - \tau| = \sum_{i=1}^{k} (\tau - Y_i) + \sum_{i=k+1}^{n} (Y_i - \tau) \\ &= k \cdot \tau - (n-k) \cdot \tau - \sum_{i=1}^{k} Y_i + \sum_{i=k+1}^{n} Y_i.\end{aligned}$$

For the range of τ values under consideration, the function is differentiable and the value of the derivative, $\phi'(\tau) = k - (n-k)$, is negative when $k < n-k$ and positive when $k > n-k$.

To see how to use this result, consider an ordered sample with three observations. When τ lies strictly to the left of observation 2 (which we know is the sample median), then $k = 1$ and $n - k = 2$, giving a negative derivative; when τ lies to the right of observation 2, then $k = 2$ and $n - k = 1$, giving a positive derivative. The switch of sign therefore locates the median. A similar argument applies to samples with an even number of observations. To understand the logic, you might work through the $n = 4$ case—in the interior of the range defining the median the derivative is zero, whereas it is negative to the left of this range and positive to the right of the range. Blyth (1990) provides an illustration.

Similarly, the sample p-th percentile can be shown to be the value of τ that minimizes $\frac{1}{n}\sum_{i=1}^{n} g_i(\tau)$ with $g_i(\tau)$ defined to equal $(1-p)|Y_i - \tau|$ if $Y_i < \tau$ and $g_i(\tau) = p|Y_i - \tau|$ when $Y_i \geq \tau$. You will occasionally see this sum written out in an alternative form, as

$$\frac{1}{n}\sum_{i=1}^{n} \left(p - \frac{1}{2} + \frac{1}{2}\text{sign}(Y_i - \tau) \right) \cdot (Y_i - \tau),$$

where the sign function takes the value -1 for strictly negative arguments, $+1$ for strictly positive arguments, and 0 when the argument is zero.

By the late 1970s these results had been known in one form or another for some time, but they were regarded as have no discernible practical value. It was suddenly realized that they could supply the basis for a multivariate estimation method focused on conditional medians and percentiles, in much the same way that multivariate regression focuses on conditional means. Cameron and Trivedi (2005, pp. 85–90) give a very good account of the multivariate approach that has been developed since then, and Ruud (2000, pp. 251–255, 270–273) provides a nicely complementary perspective with emphasis on symmetric distributions.

26.2 Multivariate Extensions

We have grown accustomed to writing linear models in the form $Y_i = \mathbf{X}_i'\beta + \epsilon_i$ and, given the assumption $\mathrm{E}\,\epsilon_i|\mathbf{X}_i$, are used to thinking of $\mathbf{X}_i'\beta$ as a conditional mean. Is there a similar way to write down expressions for conditional medians and percentiles? How is the conditional mean $\mathbf{X}_i'\beta$ related to these conditional quantiles? And how would we link the concept of heteroskedasticity in the familiar linear model to the coefficients of a quantile regression?

To understand the link between conditional means and percentiles, let's begin with the linear model $Y_i = \mathbf{X}_i'\beta + \epsilon_i$ and focus on the p-th conditional percentile of Y_i given $\mathbf{X}_i$, which

by definition is the value τ_i that satisfies

$$p = \Pr(Y_i \leq \tau_i \,|\mathbf{X}_i) = \Pr(Y_i - \mathbf{X}_i'\beta \leq \tau_i - \mathbf{X}_i'\beta \mid \mathbf{X}_i)$$

Assume that we can write the ϵ_i disturbance term as $\epsilon_i = h(\mathbf{X}_i, \theta) \cdot u_i$ in which the function $h(\mathbf{X}_i, \theta) > 0$ and the u_i disturbance has mean zero and variance σ_u^2. Further assume that u_i is fully independent of the $\mathbf{X}_i$ covariates. Note that ϵ_i is heteroskedastic with a standard deviation equal to $\sigma_u \cdot \mathrm{E}\, h(\mathbf{X}_i, \theta)$.

Since $h > 0$, we can re-express the above as

$$\begin{aligned} p &= \Pr\left(\frac{Y_i - \mathbf{X}_i'\beta}{h(\mathbf{X}_i, \theta)} \leq \frac{\tau_i - \mathbf{X}_i'\beta}{h(\mathbf{X}_i, \theta)} \mid \mathbf{X}_i\right) \\ &= \Pr\left(u_i \leq \frac{\tau_i - \mathbf{X}_i'\beta}{h(\mathbf{X}_i, \theta)} \mid \mathbf{X}_i\right) \\ &= F_u\left(\frac{\tau_i - \mathbf{X}_i'\beta}{h(\mathbf{X}_i, \theta)}\right), \end{aligned} \tag{26.1}$$

in which F_u is the cdf of u and we have made use of the independence assumption. Inverting the cdf gives

$$F_u^{-1}(p) = \frac{\tau_i - \mathbf{X}_i'\beta}{h(\mathbf{X}_i, \theta)}$$

and rearranging terms gives us the p-th conditional percentile τ_i, expressed as

$$\tau_i = \mathbf{X}_i'\beta + h(\mathbf{X}_i, \theta) \cdot F_u^{-1}(p).$$

In this way we see that τ_i is the sum of the conditional mean $\mathbf{X}_i'\beta$ and a non-linear function of the $\mathbf{X}_i$ covariates. That provides a little insight into the relationship between conditional means and percentiles, but admittedly not much.

Now suppose that we adopt a particular functional form for h, setting $h(\mathbf{X}_i, \theta) = \mathbf{X}_i'\theta > 0$. In this special case the standard deviation of ϵ_i conditional on $\mathbf{X}_i$ is simply $\sigma_\epsilon = \sigma_u \cdot \mathbf{X}_i'\theta$. We can rewrite τ_i as

$$\tau_i = \mathbf{X}_i'\beta + \mathbf{X}_i'\theta \cdot F_u^{-1}(p).$$

In other words, in this special case the conditional p-th percentile is equal to the conditional mean $\mathbf{X}_i'\beta$ plus an adjustment factor,

$$\mathbf{X}_i'\theta \cdot F_u^{-1}(p) = \sigma_\epsilon \cdot \frac{F_u^{-1}(p)}{\sigma_u},$$

that is proportional to the standard deviation of the ϵ_i disturbance term. Also, letting the p-th conditional percentile $\tau_i \equiv \mathbf{X}_i'\beta_p$,

$$\mathbf{X}_i'\beta_p = \mathbf{X}_i'\left(\beta + \theta \cdot F_u^{-1}(p)\right). \tag{26.2}$$

Like the conditional mean, the conditional percentile is linear in $\mathbf{X}_i$, at least in this special case.

How would we proceed to estimate β_p, the quantile coefficients? The quantile regression estimator for the p-th percentile is the solution to the minimization problem

$$\min_{\beta_p} S = \frac{1}{n}\sum_{i=1}^{n} g_i(\mathbf{X}_i'\beta_p),$$

with $g_i(\mathbf{X}_i'\beta_p)$ defined to equal $(1-p)|Y_i - \mathbf{X}_i'\beta_p|$ if $Y_i < \mathbf{X}_i'\beta_p$ and $g_i(\mathbf{X}_i'\beta_p) = p|Y_i - \mathbf{X}_i'\beta_p|$ when $Y_i \geq \mathbf{X}_i'\beta_p$. Note that the conditional percentile $\mathbf{X}_i'\beta_p$ is *assumed* to be linear in the $\mathbf{X}_i$ covariates. (Of course, in conventional regression we assume something very much like this: the conditional mean $\mathbf{X}_i'\beta$ is assumed linear in the covariates.) If there is some reason to think that the conditional percentile is not linear in $\mathbf{X}_i$—a graph of the percentiles of Y_i against a given $X_{i,k}$ may suggest this—then nonlinear functions of the covariates can be added to the specification.

Because S is not differentiable everywhere, the minimization problem cannot be solved through calculus. However, a solution can be found by applying methods from linear programming. The quantity $\mathbf{X}_i'\hat{\beta}_p$ is an estimate of the conditional percentile. The limiting distribution of $\hat{\beta}_p$ can be determined through methods that are too advanced for us to discuss here. STATA calculates standard errors and test statistics using the asymptotic approximation to the variance matrix, and also allows you to request bootstrapped standard errors as an alternative. See Buchinsky (1998) for an overview of the literature on these topics accompanied by a well-worked empirical example.

Finally, even if you are not interested in conditional medians as such, you may well find yourself drawn toward quantile regression for a different reason. When compared with ordinary least squares, quantile regression estimates are much less sensitive to outlying values in the data. You may want to estimate the quantile model simply to check whether the results of OLS are sensitive to outliers. Quantile regression can therefore be viewed as a useful variant of robust regression techniques. To be sure, OLS and quantile regression are designed to estimate different parameters, but since conditional means and conditional quantiles are both measures of the "central tendencies" of a distribution, the methods are at least roughly comparable.

Part V

Violations of Exogeneity

Chapter 27

Instrumental Variables

Students: You can skim section 4. If you have this textbook, you may supplement this chapter with readings from Cameron and Trivedi (2005, Sections 4.7–4.10, 6.1–6.5, 8.4.4).

This discussion draws extensively from Bowden and Turkington (1984), an old-fashioned but still excellent compendium of material on instrumental-variables methods. The chapter by Davidson and MacKinnon (1993, Chapter 7) is also very good.

27.1 The Linear Model

We begin with the linear model $Y_i = \mathbf{X}_i'\beta + \epsilon_i$, with $\mathbf{X}_i$ a $k \times 1$ vector of explanatory variables. The feature that sets apart this model from those of earlier chapters is that we now allow the $\mathbf{X}_i$ vector to be potentially correlated with ϵ_i, in the sense that $\mathrm{E}(\epsilon_i|\mathbf{X}_i) \neq 0$. When might a situation like this arise?

There are many possible causes of correlation between $\mathbf{X}_i$ and ϵ_i. Among the most common are these:

- Omission of a relevant explanatory variable typically induces a correlation between the remaining (included) explanatory variables and the (composite) disturbance.
- When an explanatory variable is the result of choices made by an economic agent, it may depend on unobserved factors that are themselves associated with ϵ. This possibility accounts for the reluctance of economists to specify models with right-hand-side choice variables even when such models are well-justified in theoretical terms.
- Measurement error may afflict one or more of the explanatory variables. As we will see in Chapter 35, this also brings about a correlation between the (measured) explanatory variables and the composite disturbance.
- In panel-data and time-series models (Chapter 31), we are often interested in specifications in which a lagged value of the dependent variable enters the model as one of

the explanatory variables. If the disturbance term is itself serially correlated, it will be associated with the lagged dependent variable.[1]

- If an explanatory variable is determined in a simultaneous-equations system, as in simple supply-and-demand models, it may well be correlated with the disturbance term.[2]

As this partial list suggests, the possibility of correlation between explanatory variables and the disturbance term is something that needs to be considered in almost all multivariate modelling.

How, then, should we proceed? In the structural model, there are k explanatory variables in total. We begin by organizing these variables into two groups: the $\mathbf{X}_{1,i}$ variables are assumed not to be correlated with the disturbance, whereas the $\mathbf{X}_{2,i}$ variables may be correlated with it. We write the structural model in a way that underscores this distinction,

$$Y_i = \mathbf{X}'_{1,i}\beta_1 + \mathbf{X}'_{2,i}\beta_2 + \epsilon_i.$$

Note that in classifying the $\mathbf{X}$ *variables into "safe" and problematic groups, we are making a fundamental and possibly wrong assumption*—when criticisms are aimed at a instrumental-variables model, this is one of the places where they are likely to be directed. In your own work, please give careful thought to this matter of safe and problematic right-hand side variables. We will return to this issue when we discuss the Sargan test, or as it is better known these days, the "test of over-identifying restrictions".

We now introduce another large and potentially controversial assumption, that a set of $m \geq k$ *valid instrumental variables* is available. We collect these instruments in an $m \times 1$ vector $\mathbf{Z}_i$. What characteristics of these instruments make them valid? They must satisfy several criteria, but the two most important are as follows. First, it must be the case that $\mathrm{E}(\epsilon_i|\mathbf{Z}_i) = 0$. That is, the instruments must be uncorrelated with the disturbances. Second, the instruments must be correlated with the problematic explanatory variables. At first glance, these two conditions would appear to be contradictory: how can we find $\mathbf{Z}_i$ variables that, on the one hand, are uncorrelated with the disturbances, and yet, on the other hand, are correlated with the $\mathbf{X}$s, at least some of which are themselves correlated with the disturbance term? This is indeed the crux of the problem with the method. You will generally find it easy to locate candidate instruments that are correlated with your problematic explanatory variables; but you will often find it very difficult to convince yourself (and others) that these candidate instruments are uncorrelated with the disturbance.

The $\mathbf{X}_i$ and $\mathbf{Z}_i$ vectors can and usually do have some variables in common; that is, if a given variable is known not to be correlated with ϵ, it enters both $\mathbf{X}_i$ and $\mathbf{Z}_i$ and is said to

[1] Consider $Y_t = \beta_1 + \beta_2 Y_{t-1} + \epsilon_t$ with $\epsilon_t = \rho\epsilon_{t-1} + u_t$. Here Y_{t-1} is correlated with ϵ_t because both variables depend on ϵ_{t-1}.

[2] Writing the supply equation as $Q_S = \alpha_1 + \alpha_2 P + \epsilon_S$ and the demand equation as $Q_D = \beta_1 + \beta_2 P + \epsilon_D$ with P representing price, setting $Q_S = Q_D$ in equilibrium, and solving for the equilibrium price, we have $P = (\alpha_1 - \beta_1)/(\alpha_2 - \beta_2) + (\epsilon_D - \epsilon_S)/(\alpha_2 - \beta_2)$. In this model, the price P depends on both the supply and the demand disturbances. Hence, neither the supply nor the demand equation can be estimated consistently by ordinary least squares.

"serve as its own instrument". Arrange the $\mathbf{X}_i$ and $\mathbf{Z}_i$ vectors as

$$\mathbf{X}_i = \begin{bmatrix} \mathbf{X}_1 \\ \cdots \\ \mathbf{X}_2 \end{bmatrix}_{k \times 1} \quad \mathbf{Z}_i = \begin{bmatrix} \mathbf{X}_1 \\ \cdots \\ \mathbf{A} \end{bmatrix}_{m \times 1},$$

in which $\mathbf{X}_1$, a subset of k_1 of the $\mathbf{X}$ covariates, takes a place in the $\mathbf{Z}_i$ vector of instruments, which also includes *additional instruments*, denoted by $\mathbf{A}$, of which there are $m - k_1$ in total. Note that the additional instruments $\mathbf{A}$ are *excluded from the structural model*. By assumption, their only role is to help us estimate that model consistently.

The case in which the number of instruments m equals the number of β parameters, such that $m = k$, is termed *just-identified*. If there is one problematic $\mathbf{X}_{2,i}$ variable, we have available one additional instrument (one variable in $\mathbf{A}$); if we have two problematic explanatory variables, we have two additional instruments; and so on. When $m > k$ we have what is termed an *over-identified* model, with more additional instruments $\mathbf{A}$ than there are problematic $\mathbf{X}_2$ variables. If $m < k$, however, we lack enough instruments to implement the method.

Before we plunge into the details, it is important to see the big picture and to understand why non-economists, as well as an increasing number of economists, are skeptical of the instrumental variables method. Two fundamental questions arise when this method is being considered, and to neither of these questions is there a fully satisfactory answer. The first question is: How can we tell whether any given explanatory variable is statistically exogenous, that is, uncorrelated with ϵ? We've listed above some of the situations that would tend to arouse suspicion, but we would like to have harder evidence on hand in the form of a statistical test before applying the IV method. Why? If all the $\mathbf{X}$s are exogenous, we have no reason to employ instrumental variables. Although the IV estimator would be consistent in this case (as we'll see), its variance would likely exceed by a considerable margin the variance of the OLS estimator. There is a tool available to diagnose statistical endogeneity—the Hausman test, to be discussed in Chapter 30—and it takes into account the typically greater variance of the instrumental variables estimator.[3] Unfortunately, we must have at least k valid instrumental variables in order to implement the Hausman test.

The second of the fundamental questions has to do with establishing the validity of candidate instrumental variables. In particular, how do we test the "null hypothesis" that there is no correlation between the instrument vector $\mathbf{Z}_i$ and the disturbance ϵ_i? We'll see later that a diagnostic test is available for this purpose—it is termed the *test of over-identifying restrictions*, or sometimes the *Sargan test* after the econometrician who devised it—but for various reasons this test falls well short of the ideal.[4] It can only be applied to over-identified models with at least k valid instruments and used in such models to assess the validity of the remaining $m - k$ candidate instruments. Furthermore, when the test rejects the null, it difficult to know what interpretation and course of action should be adopted. Under

[3] Nevertheless, caution is always in order. Monte Carlo studies have commonly found that consistent IV estimators can be disturbingly "noisy" by comparison to inconsistent OLS estimators, and can produce parameter estimates that can often be further from the truth, even in relatively large samples and especially when the instruments are weak predictors of the endogenous $\mathbf{X}$ variables.

[4] When ϵ_i is heteroskedastic, serially correlated, or both, there is a generalization of the Sargan test available that is termed the *J test*.

one interpretation, we would conclude that some of the candidate instruments are invalid and set these variables aside. But under an alternative interpretation, which is equally well justified, we would conclude that the variables in question should have been included in the structural specification of the model. Because the test does not discriminate between these two quite different possibilities, it often leaves us in limbo.

27.2 Validity Conditions

Assuming that we have at least enough candidate instruments ($m \geq k$) to proceed, what are the conditions under which these are valid instruments in the sense that they help us estimate β and σ^2 consistently? As noted above, one essential condition is that $\mathrm{E}(\epsilon_i|\mathbf{Z}_i) = 0$, that is, the *instruments must be uncorrelated with the disturbance*. To simplify matters, we add to this a *homoskedasticity* condition, that $\mathrm{E}(\epsilon_i^2|\mathbf{Z}_i) = \sigma^2$ (as we'll see, it is easy to generalize to allow heteroskedastic disturbances). We also assume that the sequence $\{(\mathbf{X}_i, \mathbf{Z}_i, \epsilon_i)\}$ is an *inid sequence*.

Note for future reference that with these assumptions, each term of the average

$$\frac{1}{n}\sum_i \mathbf{Z}_i\epsilon_i$$

has a mean vector of $\mathbf{0}$ and variance $\mathbf{V}_i = \mathrm{E}\,\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'$, which in turn equals $\sigma^2\,\mathrm{E}\,\mathbf{Z}_i\mathbf{Z}_i'$ under homoskedasticity. A standard law of large numbers for inid sequences ensures that

$$\frac{1}{n}\sum_i \mathbf{Z}_i\epsilon_i \xrightarrow{p} \mathbf{0}_{m\times 1}.$$

We will put this result to use in a moment.

We must add two further assumptions to secure consistent estimates of the β parameter. They are:

- $\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{X}_i' = \frac{1}{n}\mathbf{Z}'\mathbf{X} \xrightarrow{p} \mathbf{W}_{zx}$, a non-zero $m \times k$ matrix; and
- $\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\mathbf{Z}_i' = \frac{1}{n}\mathbf{Z}'\mathbf{Z} \xrightarrow{p} \mathbf{W}_{zz}$, a nonsingular $m \times m$ matrix.

Note that the first of these assumptions addresses, albeit in an uninformative way, the issue of correlation between the instruments and the $\mathbf{X}$ variables. We will draw out the meaning of this condition in more detail later in the chapter. For now, it is enough to say that if $\mathbf{Z}_i$ meets all of these validity conditions, the instrumental variables estimator produces consistent estimates of the β parameters. As we will see shortly, two more assumptions are needed to ensure that σ^2 is estimated consistently.

To yield estimators of β that have a limiting normal distribution, we must add the condition

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^n \mathbf{Z}_i\epsilon_i = \sqrt{n}\frac{1}{n}\mathbf{Z}'\boldsymbol{\epsilon} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}),$$

which requires the Lindeberg central limit theorem to apply to the normalized average of cross-products of the instrument vector $\mathbf{Z}_i$ and the scalar disturbance ϵ_i. In particular,

the Lindeberg condition must be met and since the variance of each term in the sum is $\mathbf{V}_i = \mathrm{E}(\epsilon_i^2 \mathbf{Z}_i \mathbf{Z}_i')$, the average of the variances must have a limit

$$\lim_{n\to\infty} \frac{1}{n} \sum_i \mathbf{V}_i \equiv \mathbf{V}.$$

For cross-sectional econometric applications, the additional conditions ensuring a limiting normal distribution would appear to be mild, at least in general.

The just-identified case

To begin our formal analysis of the IV estimator, suppose that we have exactly k valid instruments, so that the dimension of $\mathbf{Z}_i$ is $k \times 1$. In this instance we have the same number of additional instruments $\mathbf{A}_i$ as we do problematic $\mathbf{X}_{2,i}$ variables. Multiplying the vector $\mathbf{Z}_i$ by the scalar $Y_i = \mathbf{X}_i'\beta + \epsilon_i$ and then taking averages, we obtain

$$\frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i Y_i = \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \mathbf{X}_i' \beta + \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \epsilon_i.$$

Because the probability limit of the last term on the right is zero, as we discussed above, the true β satisfies

$$\beta = \left(\operatorname{plim} \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i \mathbf{X}_i' \right)^{-1} \operatorname{plim} \frac{1}{n} \sum_{i=1}^n \mathbf{Z}_i Y_i,$$

assuming that the probability limits exist. Note that because $m = k$, the expression in parentheses, $n^{-1} \sum_i \mathbf{Z}_i \mathbf{X}_i'$, is a square $k \times k$ matrix and its probability limit must be invertible if we are to solve for β. Invertibility is therefore an additional assumption that must be met for the just-identified case.

This method-of-moments approach—which we used in an earlier chapter to provide motivation for OLS regression—suggests an estimator in which simple averages replace the probability limits. After canceling the n values, we obtain

$$\hat{\beta}_{IV} = \left(\sum_{i=1}^n \mathbf{Z}_i \mathbf{X}_i' \right)^{-1} \sum_{i=1}^n \mathbf{Z}_i Y_i = (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\mathbf{Y} = \beta + (\mathbf{Z}'\mathbf{X})^{-1} \mathbf{Z}'\boldsymbol{\epsilon}.$$

If we assume that the disturbances are *homoskedastic*, that is $\mathrm{E}\left(\epsilon_i^2 \mid \mathbf{Z}_i\right) = \sigma^2$, then it is easy to show that the limiting distribution of this estimator is

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2 \mathbf{W}_{zx}^{-1} \mathbf{W}_{zz} \mathbf{W}_{xz}^{-1}\right)$$

assuming that the Lindeberg central limit theorem for inid data can be applied. As usual, we would use $\frac{1}{n} \sum_i \mathbf{Z}_i \mathbf{X}_i'$ as the estimator for $\mathbf{W}_{zx}$ and $\frac{1}{n} \sum_i \mathbf{Z}_i \mathbf{Z}_i'$ for $\mathbf{W}_{zz}$. If the disturbances are heteroskedastic, then the variance of the limiting distribution is $\mathbf{W}_{zx}^{-1} \mathbf{V} \mathbf{W}_{xz}^{-1}$ and we would estimate $\mathbf{V}$ by $\frac{1}{n} \sum_i e_i^2 \mathbf{Z}_i \mathbf{Z}_i'$ with e_i^2 being the square of the instrumental variables residual, that is, $e_i = Y_i - \mathbf{X}_i' \hat{\beta}_{IV}$. Hence, one approach to heteroskedasticity is to keep the

form of the just-identified IV estimator unchanged, but adjust the estimate of its variance matrix to reflect heteroskedasticity.

Exactly where in this analysis has the requirement that the instruments be correlated with the explanatory variables been addressed? Suppose that we are interested in a very simple model, $Y_i = \beta X_i + \epsilon_i$, in which the single explanatory variable X_i is measured in terms of deviations from its mean. Let the single instrument Z_i also be measured in deviations from its mean. The IV estimator for this case can be written

$$\hat{\beta}_{IV} = \beta + (\sum_i Z_i X_i)^{-1} \sum_i Z_i \epsilon_i \xrightarrow{p} \beta + \frac{\text{Cov}(Z, \epsilon)}{\text{Cov}(Z, X)}.$$

Note, though, that if the covariance between the problematic X_i and the instrument Z_i is zero, the IV method falls apart! In the limit, the estimator cannot even be defined. (This is the equivalent of the matrix $\mathbf{W}_{zx}$ not being invertible.) So long as the covariance is non-zero—even if it is negative—the method works. We will look closer at the covariance issue later in this chapter.

An example may help to clarify the mechanics of the instrumental variables method and show where your research judgements will come into play. Consider a model in which the wage rate for an adult, Y_i, is determined by a set of exogenous variables $\mathbf{X}_i$ and the grades of schooling S_i that this person has attained,

$$Y_i = \mathbf{X}_i' \beta + S_i \beta_S + \epsilon_i.$$

As we contemplate the disturbance term of the wage model, we may come to think of it as having two main parts, $\epsilon_i = a_i + v_i$, with a_i representing the kind of individual ability that pays off in terms of wages and v_i representing all of the other unobservables that influence wages (for instance, conditions at the firm in which person i is employed).

Considering further this interpretation, we realize that there may be a problem in estimating the wage model: it seems likely that in addition to earning higher wages once they enter the labor market, people with greater ability a_i might have been more highly motivated to acquire schooling. In other words, ability a_i is likely to affect wages directly on the job but also indirectly by raising the level of S_i that person i brings to the job. Hence, the OLS coefficient on S_i is apt to overstate the direct impact of schooling on wages; the coefficient will also reflect the effects of unobserved ability.

What instrument might be used to secure consistent estimates of the β parameter given the presumed correlation between S_i and the ϵ_i disturbance? A common strategy in the literature is to use parental variables as instruments for their children's schooling. Let the variable P_i represent the years of schooling attained by person i's parents, and assume that P_i is available in our dataset. We would expect P_i to be strongly correlated with S_i and can easily check this by examining the empirical correlation. If P_i is to be a valid instrument, we must make a further assumption, that parental schooling P_i is uncorrelated with the child's ability a_i and in this way uncorrelated with the full ϵ_i disturbance. Of course, since ϵ_i is not observed, nothing in our data set will directly verify this assumption.

If P_i and ϵ_i are indeed uncorrelated, then P_i can serve as a valid instrument. The full $k \times 1$ vector of instruments is

$$\mathbf{Z}_i = \begin{bmatrix} \mathbf{X}_i \\ P_i \end{bmatrix},$$

and the instrumental variables estimator is

$$\hat{\beta}_{IV} = \left(\sum_i \begin{bmatrix} \mathbf{X}_i \\ P_i \end{bmatrix} \begin{bmatrix} \mathbf{X}_i' & S_i \end{bmatrix} \right)^{-1} \sum_i \begin{bmatrix} \mathbf{X}_i \\ P_i \end{bmatrix} Y_i = \begin{bmatrix} \sum_i \mathbf{X}_i \mathbf{X}_i' & \sum_i \mathbf{X}_i S_i \\ \sum_i P_i \mathbf{X}_i' & \sum_i P_i S_i \end{bmatrix}^{-1} \begin{bmatrix} \sum_i \mathbf{X}_i Y_i \\ \sum_i P_i Y_i \end{bmatrix}.$$

If our key assumption is correct and P_i is a valid instrument, the IV estimator is consistent.

The heart of the argument supporting P_i as an instrument is the assumption that while ability a_i is expressed directly in person i's wages and indirectly through schooling, a_i is not correlated with the schooling attained by i's parents. How reasonable is this assertion? A counter-argument is easy to mount. Better-educated parents are apt to bring up their children in fundamentally different ways than do parents with less education. Educated parents may tend repeatedly to emphasize to their children how success in school is linked to success in the workplace; possibly they would reinforce various attitudes, habits of discipline, and modes of behavior in their children that would tend ultimately to be rewarded by higher wages. Through routes such as these, the child-rearing strategies of educated parents may well be associated with the kinds of unmeasured abilities a_i of their children that pay off in workplace wages and promotions when these children grow to adulthood. Although economists were late in coming to this realization, a massive research effort led by James Heckman is focused on the apparently high lifetime returns to these early "soft-skills" investments—see the web site https://heckmanequation.org/ for a comprehensive overview.

Evidently, then, the case for P_i being a valid instrument depends on a naïve and increasingly disputable model of child-rearing. Many arguments supporting the use of instruments turn out, on further reflection, to be less convincing than they initially appear. As in this case, to evaluate the validity of instruments we often need to step outside statistics proper and appeal to non-statistical theories and evidence from other fields.

The over-identified case

When we have more instruments $\mathbf{Z}_i$ available than $\mathbf{X}_i$ variables, that is, $m > k$, the multiply-through approach does not work, in that it yields m equations which would need to be solved for the $k < m$ unknown β parameters. For this *over-identified* case, the instrumental variables estimator of β is therefore defined as the vector $\hat{\beta}_{IV}$ that minimizes the following quadratic form,

$$S = (\mathbf{Y} - \mathbf{X}\beta)' \mathbf{P}_\mathbf{Z} (\mathbf{Y} - \mathbf{X}\beta) = (\mathbf{Y} - \mathbf{X}\beta)' \mathbf{Z} (\mathbf{Z}'\mathbf{Z})^{-1} \mathbf{Z}' (\mathbf{Y} - \mathbf{X}\beta),$$

where $\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, a projection matrix. Below we will ask why $\mathbf{P}_\mathbf{Z}$ appears in the quadratic form rather than some other matrix. For now, however, let us simply explore the properties of the resulting estimator.

To begin, instead of differentiating S with respect to β, let us exploit the fact that $\mathbf{P}_\mathbf{Z}$ is a projection matrix. Because $\mathbf{P}_\mathbf{Z}$ is symmetric idempotent, the quadratic form simplifies to

$$S = (\hat{\mathbf{Y}} - \hat{\mathbf{X}}\beta)' (\hat{\mathbf{Y}} - \hat{\mathbf{X}}\beta)$$

in which $\hat{\mathbf{Y}}$ is the result of projecting the $\mathbf{Y}$ vector upon the full set of instruments $\mathbf{Z}$, and $\hat{\mathbf{X}}$ is the result of projecting each column of $\mathbf{X}$ (i.e., each explanatory variable) onto $\mathbf{Z}$. The $\mathbf{X}$

projection onto $\mathbf{Z}$ actually needs to be done only for the $\mathbf{X}_2$ variables that are potentially correlated with ϵ. The other $\mathbf{X}_1$ variables, which are assumed to be uncorrelated with ϵ, are already in $\mathbf{Z}$ where they serve as their own instruments, and projecting such variables onto $\mathbf{Z}$ leaves them unchanged.

We already know what $\hat{\beta}$ parameter minimizes this sum of squares—it is

$$\hat{\beta}_{IV} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\hat{\mathbf{Y}},$$

the ordinary least squares estimator from a regression of $\hat{\mathbf{Y}}$ on $\hat{\mathbf{X}}$. An alternative formula for the IV estimator is

$$\hat{\beta}_{IV} = \left(\hat{\mathbf{X}}'\hat{\mathbf{X}}\right)^{-1}\hat{\mathbf{X}}'\mathbf{Y} = (\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{Y}.$$

This is known as the two-stage least squares estimator (2SLS); it is presented in every econometrics textbook. Clearly it makes no numerical difference whether $\mathbf{Y}$ or its projection $\hat{\mathbf{Y}}$ is used in the formula.

[Digression] A short digression is in order on an alternative route to the 2SLS estimator, given its importance in historical and present-day econometric thinking. Beginning with the structural model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

take expectations of both sides with respect to the (valid) instruments $\mathbf{Z}$,

$$\mathrm{E}(\mathbf{Y}|\mathbf{Z}) = \mathrm{E}(\mathbf{X}|\mathbf{Z})\beta.$$

In principle, these conditional expectations could either be linear or nonlinear in $\mathbf{Z}$. If they are linear—a non-trivial assumption—we could empirically approximate the conditional expectations using projections of $\mathbf{Y}$ and $\mathbf{X}$ on $\mathbf{Z}$, yielding

$$\mathbf{P}_{\mathbf{Z}}\mathbf{Y} = \mathbf{P}_{\mathbf{Z}}\mathbf{X}\,\hat{\beta}_{IV}.$$

Note that $\mathbf{P}_{\mathbf{Z}}\mathbf{X}$ is a collection of 2SLS first-stage coefficients $\hat{\pi}_j$, each derived from an OLS regression of a (potentially) endogenous explanatory variable $\mathbf{X}_j$ on the instruments $\mathbf{Z}$. That is, we would obtain these $\hat{\pi}_j$ coefficients from first-stage regressions of the form

$$\mathbf{X}_j = \mathbf{Z}\pi_j + \mathbf{u}_j$$

for each such explanatory variable, with (as explained more fully below) $\mathbf{P}_{\mathbf{Z}}\mathbf{X}_j = \mathbf{X}_j$ for the assumed-to-be-exogenous explanatory variables.

Now—in a step that is decidedly optional here, but helpful in other contexts such as nonparametric modelling of instrumental variables estimators—further project these projections onto $\mathbf{X}$, so that

$$\mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Z}}\mathbf{Y} = \mathbf{P}_{\mathbf{X}}\mathbf{P}_{\mathbf{Z}}\mathbf{X}\,\hat{\beta}_{IV}.$$

Write out the $\mathbf{P}_{\mathbf{X}}$ and $\mathbf{P}_{\mathbf{Z}}$ matrices in full, and then multiply the equation through by $\mathbf{X}'$. (Looking back a bit, you can see that this multiply-through operation could have been executed without the additional projection onto $\mathbf{X}$, which is why the projection onto $\mathbf{X}$ is optional for present purposes.) The equation simplifies to

$$\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{Y} = (\mathbf{X}'\mathbf{P}_{\mathbf{Z}}\mathbf{X})\,\hat{\beta}_{IV},$$

from which we obtain the 2SLS estimator $\hat{\beta}_{IV} = (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{Y}$. **[End of Digression]**

To get a sense of the first-order conditions that the instrumental variables estimator must satisfy, let us now return to the quadratic form as it was originally expressed and differentiate it with respect to the β vector. We find that the estimator satisfies

$$-2\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{Y} + 2(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})\hat{\beta}_{IV} = \mathbf{0}.$$

These first-order conditions can be re-expressed in terms of k orthogonality conditions

$$\mathbf{X}'\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}) = \mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{e} = \mathbf{0}$$

where $\mathbf{e}$ is the IV residual. They can also be written as $\hat{\mathbf{X}}'\mathbf{e} = \mathbf{0}$ with $\hat{\mathbf{X}}$ being the projection of $\mathbf{X}$ onto the instruments $\mathbf{Z}$.

Note that in contrast to ordinary least squares, these first-order conditions *do not* imply that all of the columns of $\mathbf{X}$ will be orthogonal to the IV residual vector. For the $\mathbf{X}_1$ variables that are known to be exogenous and therefore serve as their own instruments, we have $\mathbf{P}_\mathbf{Z}\mathbf{X}_1 = \hat{\mathbf{X}}_1 = \mathbf{X}_1$ and thus $\mathbf{X}_1'\mathbf{e} = \mathbf{0}$ for these variables. But the other $\mathbf{X}_2$ variables—the problematic ones—are *not* orthogonal to the IV residual vector.

The first-order conditions can be solved directly for the IV estimator, yielding

$$\begin{aligned}
\hat{\beta}_{IV} &= (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{Y} && (27.1)\\
&= \beta + (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\epsilon && (27.2)\\
&= \beta + \left(\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\epsilon. && (27.3)
\end{aligned}$$

This is the same result that we have already obtained.

Note one important feature of the just-identified case we dealt with earlier: the minimized value of its IV quadratic form is zero. That is, when $m = k$,

$$(\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV})'\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}) = 0,$$

a result that follows from the fact that $\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}) = \mathbf{0}$ in the just-identified case.[5] In the over-identified case, by contrast, the minimized value of the quadratic form will exceed zero. (Here, $\mathbf{X}'\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}) = \mathbf{0}$ but $\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}) \neq \mathbf{0}$.) The difference in these minimized values provides the basis for a family of specification tests, as we will see shortly.

As for the large-sample properties of the over-identified estimator, under the assumptions set out above it is straightforward to show that $\hat{\beta}_{IV} \xrightarrow{p} \beta$ and, assuming homoskedastic disturbances $\mathrm{E}\left(\epsilon_i^2 \mid \mathbf{Z}_i\right) = \sigma^2$,

$$\sqrt{n}(\hat{\beta}_{IV} - \beta) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2(\mathbf{W}_{xz}\mathbf{W}_{zz}^{-1}\mathbf{W}_{zx})^{-1}\right) \tag{27.4}$$

with $\mathbf{W}_{xz}$ being the transpose of $\mathbf{W}_{zx}$. Another expression for the asymptotic covariance matrix is $\sigma^2(\operatorname{plim} n^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}$. The variance of the limiting distribution under heteroskedasticity is given by a longer and more complicated expression involving $\mathbf{V} = \frac{1}{n}\sum_i \mathrm{E}\left(\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'\right)$, but the details are easy enough to work out. We'll take a closer look shortly, when we examine the instrumental variables method from the GMM perspective.

[5] Students: Prove this for yourselves. Note that the result implies $\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{e} = \mathbf{0}$ for the just-identified as well as for the over-identified case.

Estimating σ^2 consistently

For the homoskedastic case, we consider the estimator $\hat{\sigma}^2 = \mathbf{e}'\mathbf{e}/n$ where $\mathbf{e}$ is the IV residual $\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}$. It can be shown that $\hat{\sigma}^2 \xrightarrow{p} \sigma^2$. The proof is complicated by the fact that when $\mathbf{e}$ is expressed as a matrix times the disturbance vector ϵ, or

$$\mathbf{e} = (\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z})\epsilon \equiv \mathbf{M}_{\mathbf{X},\mathbf{Z}}\epsilon$$

the matrix $\mathbf{M}_{\mathbf{X},\mathbf{Z}}$ is not idempotent. To show consistency we must make two additional assumptions, namely that plim $n^{-1}\mathbf{X}'\mathbf{X}$ and plim $n^{-1}\mathbf{X}'\epsilon$ are finite matrices. Interestingly, neither of these assumptions is required to prove the consistency of $\hat{\beta}_{IV}$.

27.3 Rationale for the Quadratic Form

We return now to the formulation of the IV estimator $\hat{\beta}_{IV}$ as the vector of estimates that minimizes a particular quadratic form. The question of interest is why $\mathbf{P}_\mathbf{Z} = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$ should appear in the middle of that quadratic form rather than some other expression. We can gain insight into this issue if we view the IV estimator from the perspective of the Generalized Method of Moments (GMM) approach to estimation—this is a large family of estimators whose members include ordinary least squares, maximum likelihood, and instrumental variables. We give a brief introduction to the GMM approach here, and study it in more detail in Chapter 28.

In the GMM approach the researcher specifies a vector of $m \geq k$ functions of the observed $(Y_i, \mathbf{X}_i, \mathbf{Z}_i)$ variables and k unknown θ parameters, which we denote by $\mathbf{g}_i(\theta) = \mathbf{g}(Y_i, \mathbf{X}_i, \mathbf{Z}_i, \theta)$. This vector must have the property that $\mathrm{E}\,\mathbf{g}_i(\theta_0) = \mathbf{0}$, with θ_0 being the true value of θ. These m expectation conditions are termed *moment conditions*.

To estimate θ with $m > k$, we define a quadratic form $Q(\theta)$ in which, on the wings, we insert sample averages that are the sample analogs to expectations, and in the middle of the quadratic form we insert an $m \times m$ matrix $\mathbf{M}_n$ that is symmetric and positive definite, giving us

$$Q(\theta) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\theta)\right)' \mathbf{M}_n \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\theta)\right).$$

It turns out that under reasonable assumptions, the estimator $\hat{\theta}$ that minimizes the quadratic form is consistent for any $\mathbf{M}_n$ having the stated properties. But there is a best choice of $\mathbf{M}_n$—one yielding the smallest limiting variance of $\sqrt{n}(\hat{\theta} - \theta_0)$—which is the inverse of the limiting variance $\mathbf{V}$ of $\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\theta_0)$. If we define $\mathbf{M}_n$ in such a way that it converges to $\mathbf{V}^{-1}$, we obtain the *efficient* GMM estimator. Generally this requires an estimation procedure that entails two steps, with the output of the first step used to construct $\hat{\mathbf{V}}_n^{-1}$, the efficient form of the $\mathbf{M}_n$ matrix.

To apply these ideas to the over-identified IV case, let us first expand $\mathbf{P}_\mathbf{Z}$ and then write the IV quadratic form $S(\beta)$ in terms of averages, giving

$$\frac{1}{n}S(\beta) = \left(\frac{1}{n}\sum_i \mathbf{Z}_i(Y_i - \mathbf{X}_i'\beta)\right)' \left(\frac{1}{n}\sum_i \mathbf{Z}_i\mathbf{Z}_i'\right)^{-1} \left(\frac{1}{n}\sum_i \mathbf{Z}_i(Y_i - \mathbf{X}_i'\beta)\right).$$

The $m \times 1$ vector $\mathbf{Z}_i(Y_i - \mathbf{X}_i'\beta) = \mathbf{Z}_i\epsilon_i$ when β is set to its true value β_0 and at this β_0 the vector has an expected value of zero. As mentioned earlier,

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i(Y_i - \mathbf{X}_i'\beta_0) = \sqrt{n}\frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i\epsilon_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V})$$

in which

$$\mathbf{V} \equiv \lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(\epsilon_i^2\mathbf{Z}_i\mathbf{Z}_i'\right)$$

with the existence of this limit being required for the Lindeberg central limit theorem to hold. This yields the result

$$\sqrt{n}\frac{1}{n}\sum_{i}\mathbf{Z}_i\epsilon_i \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \mathbf{V}\right).$$

Applying the robust-standard-errors approach of White to the present case, we have a feasible estimator $\hat{\mathbf{V}}$ of the variance matrix,

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}\left(e_i^2\mathbf{Z}_i\mathbf{Z}_i'\right)$$

in which e_i^2 is the square of the instrumental variables residual $e_i = Y_i - \mathbf{X}_i'\hat{\beta}_{IV}$. Of course, this estimator cannot be implemented until we have a consistent estimator for β already in hand, from which we could compute the e_i residuals.

How do we proceed? Looking back at the sum-of-squares quadratic form,

$$\frac{1}{n}S(\beta) = \left(\frac{1}{n}\sum_{i}\mathbf{Z}_i(Y_i - \mathbf{X}_i'\beta)\right)'\left(\frac{1}{n}\sum_{i}\mathbf{Z}_i\mathbf{Z}_i'\right)^{-1}\left(\frac{1}{n}\sum_{i}\mathbf{Z}_i(Y_i - \mathbf{X}_i'\beta)\right).$$

we see that the matrix in the middle, $1/n\sum_i \mathbf{Z}_i\mathbf{Z}_i$, does not converge to $\mathbf{V}$ as would be required for an *efficient* GMM estimator, but nevertheless meets the symmetric, positive definite criterion needed for an *inefficient* GMM estimator. We can therefore use precisely this version of the quadratic form to deliver a first-step, inefficient estimator $\hat{\beta}_{IV}$ and, with its residuals $e_i = Y_i - \mathbf{X}_i'\hat{\beta}_{IV}$ in hand, we are then able to compute $\hat{\mathbf{V}}$ to implement the efficient GMM estimator in the second step. Note that the "heteroskedasticity of unknown form" approach can be modified to handle cases in which we have a particular model of heteroskedasticity that we want to impose, using the first-step residuals to estimate that model.

In short, *the standard 2SLS estimator can be viewed as an inefficient GMM estimator*. Interestingly, if the disturbances actually happen to be homoskedastic with $\mathrm{E}\left(\epsilon_i^2 \mid \mathbf{Z}_i\right) = \sigma^2$, then

$$\mathbf{V} \overset{a}{=} \sigma^2\frac{1}{n}\sum_{i}\mathbf{Z}_i\mathbf{Z}_i'$$

and the constant σ^2 simply scales the sum of squares in a way that does not affect the value of the minimizing $\hat{\beta}_{IV}$. In this special case of homoskedasticity, in other words, the standard 2SLS estimator *is* the efficient GMM estimator. You should bear this special-case result in mind when you have good reason to assume homoskedasticity.

27.4 Finding and Assessing Instruments

It is far from obvious how to locate valid instrumental variables—exactly what to do will vary with the specifics of the structural model. However, there are some important cases in which the form of the structural model suggests possible instruments. In other cases, economic theory may identify candidate instruments. We'll consider three examples.

Let a time-series model with a lagged dependent variable be written as $Y_t = \mathbf{X}_t'\beta + \delta Y_{t-1} + \epsilon_t$, and suppose that the disturbance term is serially correlated, such that $\epsilon_t = \rho\epsilon_{t-1} + u_t$, with u_t itself being serially uncorrelated and uncorrelated with all previous ϵ_{t-s} for $s \geq 1$. Obviously, the Y_{t-1} variable of the structural model is correlated with ϵ_t, as noted earlier in this chapter, and the structural model cannot be estimated consistently by ordinary least squares.

What instruments suggest themselves in this case? One possibility is to use lagged values $\mathbf{X}_{t-1}$ as instruments, on the assumption that the $\{\mathbf{X}_t\}$ sequence is generated in such a way that $\mathbf{X}_{t-1}$ is uncorrelated with ϵ_t disturbance. Indeed, additional lags in the $\mathbf{X}_t$ variables can also be considered—an insight that is exploited in a family of instrumental variables methods developed by Arellano and Bond; see the discussion in Greene (2008, p. 12.8.2) for an introduction.[6]

For an example in which economic theory suggests instruments, consider a textbook Cobb–Douglas production function for the i-th firm, $Q_i = E_i L_i^{\alpha} K_i^{1-\alpha}$, in which E_i is a firm-specific efficiency parameter (known to the firm but not observed by the researcher) which enters the production function along with labor and capital. Letting the wage be denoted by w_i and letting r_i denote the rental rate on capital, a cost-minimizing firm will choose levels of labor and capital such that for any given output level Q_i,

$$L_i = Q_i E_i^{-1} \phi^{1-\alpha}$$
$$K_i = Q_i E_i^{-1} \phi^{-\alpha}$$

with

$$\phi = \left(\frac{\alpha}{1-\alpha} \frac{r_i}{w_i} \right).$$

In an econometric model using data on n firms, we might observe output Q_i and labor and capital inputs L_i and K_i for all firms, and could write the production function in terms of the logs of these variables,

$$\ln Q_i = \alpha \ln L_i + (1-\alpha) \ln K_i + \ln E_i.$$

The log of the firm-specific efficiency parameter appears here in the role of the disturbance term of our structural model. Unfortunately, economic theory has just shown us that both $\ln L_i$ and $\ln K_i$ are correlated (negatively) with $\ln E_i$. Fortunately, the theory also suggests candidate instruments: if we observe the wage w_i and rental rate r_i facing the i-th firm, we may be able to employ these input prices as instruments. We can do so if the firm is

[6] Interestingly, as we saw in Chapter 24, it is possible to estimate the parameters of the structural model without using instruments at all. Multiply the equation for Y_{t-1} by ρ and subtract it from the equation for Y_t. The result is an equation that is nonlinear in the parameters, but which can be estimated consistently by nonlinear least squares.

a price-taker in its input markets (so that w_i and r_i are exogenous to the firm, unaffected by its decisions) and if there is sufficient variation in input prices for these variables to be informative.[7] If the firm is not a price-taker in the input markets, however, the input prices will not generally be valid instruments.

Economic theory is also helpful in the case of simultaneously determined variables. Returning to the supply-and-demand model mentioned above, write the supply equation as $Q_S = \alpha_1 + \alpha_2 P + \alpha_3 Z + \epsilon_S$, with Z being an exogenous variable whose value affects the quantity supplied for any given price. (If the supply equation reflects marginal costs, then Z can be viewed as a "cost-shifter" variable.) The demand equation of the system is $Q_D = \beta_1 + \beta_2 P + \epsilon_D$. Setting $Q_S = Q_D$ in equilibrium and solving for the equilibrium price, we have

$$P = \frac{\beta_1 - \alpha_1}{\alpha_2 - \beta_2} - \frac{\alpha_3}{\alpha_2 - \beta_2} Z + \frac{1}{\alpha_2 - \beta_2}(\epsilon_D - \epsilon_S).$$

The Z variable is obviously correlated with price P, and if it is uncorrelated with both ϵ_D and ϵ_S, it can play the role of an instrumental variable in the demand equation, allowing us to estimate consistently the β parameters and the disturbance term variance σ_D^2. Note, however, that Z does not help in estimating the parameters of the supply equation. We would need something akin to Z on the demand side in order to apply the instrumental variables method to the supply equation.

Adding instruments

Let $\mathbf{Z}_1$ denote one set of instruments and suppose that $\mathbf{Z}_1$ is augmented with additional instrumental variables, with $\mathbf{Z}_2$ being the expanded set. We will show that the covariance matrix of $\hat{\beta}_1$, the IV estimator using the smaller instrument set, is larger in the matrix sense than the covariance matrix of $\hat{\beta}_2$, the estimator that uses $\mathbf{Z}_2$ as the instrument set.

Davidson and MacKinnon (1993, p. 219) set up the problem in the following way. The difference in the respective covariance matrices of $\hat{\beta}_1$ and $\hat{\beta}_2$ is

$$\sigma^2 \left((\mathbf{X}'\mathbf{P}_1\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{P}_2\mathbf{X})^{-1} \right)$$

which is positive semidefinite if

$$(\mathbf{X}'\mathbf{P}_2\mathbf{X}) - (\mathbf{X}'\mathbf{P}_1\mathbf{X}) = \mathbf{X}'(\mathbf{P}_2 - \mathbf{P}_1)\mathbf{X}$$

is positive semidefinite. Both $\mathbf{P}_2$ and $\mathbf{P}_1$ are projection matrices; furthermore, $\mathbf{P}_1\mathbf{P}_2 = \mathbf{P}_2\mathbf{P}_1 = \mathbf{P}_1$ because $\mathbf{Z}_2$ contains $\mathbf{Z}_1$. Using these facts, we find that $\mathbf{P}_2 - \mathbf{P}_1$ is both symmetric and idempotent. Therefore, it must be a projection matrix of some sort. Projection matrices are positive semidefinite. From this it follows that $\hat{\beta}_1$ cannot be more efficient than $\hat{\beta}_2$.

Do we conclude that adding valid instruments is always desirable? In asymptotic terms, the answer must be yes, by the argument just presented. However, as has been established in a rather specialized and difficult literature on the finite-sample properties of the IV

[7] In this example, too, there is an alternative to estimating the production function: the Cobb–Douglas cost function could be estimated instead, provided that production costs are observed in the data, the firm is a price-taker in input markets and there is sufficient variation across firms in wages and rental rates.

estimator, the addition of instruments may cause the finite-sample properties of $\hat{\beta}_{IV}$ to deteriorate! An important theme in this literature is that the finite-sample and asymptotic properties of instrumental variables estimators can be strikingly different.

27.5 Sargan's Test of Over-identifying Restrictions

For the case in which more instruments are available than is strictly necessary to estimate the model, that is, $m > k$, it is possible to assess the validity of the additional $m - k$ instruments. By "validity" we mean the assumption that the instruments are asymptotically uncorrelated with the disturbance ϵ.

The test that we will discuss is very simple, but it has some curious features. First, it cannot be applied to the just-identified case $m = k$. Second, with respect to the $\mathbf{X}_1$ variables that are assumed to be exogenous and serve as their own instruments, the test is silent as to whether this exogeneity assumption is correct. It cannot help us in any way to evaluate this assumption. Third, to use the test we must maintain the hypothesis that we have at least k valid instruments; the test addresses only the validity of the remaining $m - k$ instruments.

The test statistic itself is quite straightforward. The minimized instrumental variables sum of squares function evaluated at $\hat{\beta}_{IV}$ is

$$S(\hat{\beta}_{IV}) = (\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV})'\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}) = \mathbf{e}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{e},$$

with $\mathbf{e}$ being the IV residual vector. As we mentioned earlier, when $m = k$ (the just-identified case) $S(\hat{\beta}_{IV}) = 0$ and obviously S cannot in this case serve as the basis for a test statistic. Provided that $m > k$, however, the minimized sum of squares $S(\hat{\beta}_{IV}) > 0$, and we will show that under the null hypothesis, the test statistic

$$T = S(\hat{\beta}_{IV})/\hat{\sigma}^2 \xrightarrow{d} \chi^2_{m-k},$$

where $m - k$ is the degree of over-identification and $\hat{\sigma}^2$ is the consistent estimator of σ^2 based on the IV residuals.

As Davidson and MacKinnon (1993, p. 236) point out, the test statistic T can also be expressed as n times the uncentered R^2 from a diagnostic regression of the IV residuals on the full set of instruments $\mathbf{Z}$. This way of formulating the test is instructive. Note that the IV residuals are *not* orthogonal to all columns of $\mathbf{Z}$, for, as you will recall, the first-order condition for the IV estimator is given by $\mathbf{X}'\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}) = \mathbf{0}$. However, as mentioned earlier, the IV residuals *are* orthogonal to each of the explanatory variables in $\mathbf{X}_1$. Therefore, the $\mathbf{X}_1$ variables will have no explanatory power in the diagnostic regression and will contribute nothing to its R^2. The test essentially ignores these exogenous $\mathbf{X}_1$ variables and focuses attention on the validity of $m - k$ of the remaining instruments.

We now return to the quadratic form defining $S(\hat{\beta}_{IV})$ and show how it leads to a χ^2_{m-k} test. The proof exploits some special properties of the IV residuals. With the residual vector $\mathbf{e} = (\mathbf{I}_n - \mathbf{X}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z})\epsilon$, we see that $\mathbf{P}_\mathbf{Z}\mathbf{e}$ is equal to $\left(\mathbf{P}_\mathbf{Z} - \hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\right)\epsilon$. The $n \times n$ matrix in parentheses is symmetric and idempotent, of rank $m - k$. Hence

$$S(\hat{\beta}_{IV}) = \epsilon'\left(\mathbf{P}_\mathbf{Z} - \hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\right)\epsilon.$$

Although idempotent, the $n \times n$ matrix is not in the right form for asymptotic analysis. However, by writing out $\mathbf{P_Z}$ in full in three of the places where it appears, we obtain an expression that, with the aid of clever factorization of $(\mathbf{Z}'\mathbf{Z})^{-1}$, judicial insertions of n and $\sqrt{n}$, and application of a central limit theorem, can be manipulated into an expression that (asymptotically) takes the form of two multivariate normal random vectors surrounding an $m \times m$ idempotent matrix (whose probability limit is fixed).

We begin with

$$S(\hat{\beta}_{IV}) = \epsilon' \left(\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}' \right) \epsilon. \tag{27.5}$$

If we now factor $(\mathbf{Z}'\mathbf{Z})^{-1} = (\mathbf{Z}'\mathbf{Z})^{-1/2} \cdot (\mathbf{Z}'\mathbf{Z})^{-1/2}$, we see that the quadratic form in (27.5) can be re-expressed as

$$S(\hat{\beta}_{IV}) = \epsilon'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2} \cdot \tilde{\mathbf{M}} \cdot (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\epsilon, \tag{27.6}$$

in which

$$\tilde{\mathbf{M}} = \mathbf{I}_m - (\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\mathbf{X}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1/2},$$

a matrix that is easily shown to be symmetric idempotent of rank $m - k$. Note that n can be inserted in the usual places without changing the value of $\tilde{\mathbf{M}}$, and therefore $\tilde{\mathbf{M}}$ converges in probability to a fixed $m \times m$ matrix.

The wings of this quadratic form converge in distribution to normal random variables with covariance matrix $\sigma^2\mathbf{I}_m$, that is,

$$(\mathbf{Z}'\mathbf{Z})^{-1/2}\mathbf{Z}'\epsilon = \left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1/2} \frac{1}{\sqrt{n}}\mathbf{Z}'\epsilon \stackrel{a}{=} \mathbf{W}_{zz}^{-1/2} \cdot \sqrt{n}\frac{1}{n}\mathbf{Z}'\epsilon \stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}_m).$$

The remaining task is to multiply through by a factor that converts each of the random vectors to in the wings to multivariate *standard* normal, as required by the quadratic form theorem. This is accomplished by dividing by $\frac{1}{n}\mathbf{e}'\mathbf{e} \stackrel{p}{\to} \sigma^2$, and in this way we arrive at the Sargan result: $T = S(\hat{\beta}_{IV})/\hat{\sigma}^2 \stackrel{d}{\to} \chi^2_{m-k}$.

Now that the essentials of the test have been laid out, we might ask why the test is described as a test of "over-identifying restrictions." What is meant by this phrase? In fact it is a warning signal: a rejection of the null hypothesis may not be interpretable in terms of instrument validity at all. To see the problem, imagine that the true structural model is not $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ but rather $\mathbf{Y} = \mathbf{X}\beta + \mathbf{A}^s\delta + u$, with $\mathbf{A}^s$ being a subset of the additional instruments. However, we (the researchers) proceed on the mistaken assumption that the true model is $\mathbf{Y} = \mathbf{X}\beta + \epsilon$, and thus unwittingly include $\mathbf{A}^s$ in ϵ giving $\epsilon = \mathbf{A}^s\delta + u$. In executing the test described above, we regress the IV residuals from our mis-specified model, $\mathbf{e} = \mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}$, on the full set of $\mathbf{Z}$ and find what we take to be evidence against the null hypothesis, that all our instruments are valid. In fact, what the test shows in this case is that some of the $\mathbf{Z}$ variables actually belonged in the model of $\mathbf{Y}$ itself; they should not have been excluded from the specification. In short, the result of the test is open to two conflicting interpretations and nothing in the structure of the test tells us which one of these is the correct interpretation.

27.6 Correlation of Instruments and Explanatory Variables

The requirement that the instruments $\mathbf{Z}_i$ be correlated with the explanatory variables needs more discussion that we have given it thus far. We saw in one simple just-identified example that a correlation of zero causes the IV estimator to become undefined in the limit. Let's explore this issue in a more general context.

Consider the two-stage least squares formula for the IV estimator,

$$\hat{\beta}_{IV} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{Y},$$

and, applying the FWL theorem, examine the IV coefficients on the problematic $\mathbf{X}_2$ variables,

$$\hat{\beta}_2 = (\hat{\mathbf{X}}_2'\hat{\mathbf{M}}_1\hat{\mathbf{X}}_2)^{-1}\hat{\mathbf{X}}_2'\hat{\mathbf{M}}_1\mathbf{Y},$$

with

$$\hat{\mathbf{M}}_1 = \mathbf{I} - \hat{\mathbf{X}}_1(\hat{\mathbf{X}}_1'\hat{\mathbf{X}}_1)^{-1}\hat{\mathbf{X}}_1' = \mathbf{I} - \mathbf{X}_1(\mathbf{X}_1'\mathbf{X}_1)^{-1}\mathbf{X}_1' = \mathbf{M}_1$$

because of the fact that the $\mathbf{X}_1$ variables serve as their own instruments. We have

$$\mathbf{M}_1\hat{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}_2.$$

To keep the notation simple, let's assume that there is only one problematic $\mathbf{X}_2$ variable. Then

$$\mathbf{M}_1\hat{\mathbf{X}}_2 = \mathbf{M}_1\mathbf{Z}\hat{\pi}_n,$$

with $\hat{\pi}_n$ being the coefficients from the first-stage regression of $\mathbf{X}_2$ on $\mathbf{Z}$. This simplifies further,

$$\mathbf{M}_1\mathbf{Z}\hat{\pi}_n = \mathbf{M}_1\begin{bmatrix}\mathbf{X}_1 & \mathbf{A}\end{bmatrix}\begin{bmatrix}\hat{\pi}_1 \\ \hat{\pi}_A\end{bmatrix} = \mathbf{M}_1\mathbf{X}_1\hat{\pi}_1 + \mathbf{M}_1\mathbf{A}\hat{\pi}_A = \mathbf{M}_1\mathbf{A}\hat{\pi}_A$$

because $\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$.

The variance of the limiting distribution of $\sqrt{n}(\hat{\beta}_2 - \beta_2)$ is

$$\sigma^2\left(\operatorname{plim}\frac{1}{n}\hat{\mathbf{X}}_2'\mathbf{M}_1\hat{\mathbf{X}}_2\right)^{-1} = \sigma^2\left(\operatorname{plim}\frac{1}{n}\hat{\pi}_A'\mathbf{A}'\mathbf{M}_1\mathbf{A}\hat{\pi}_A\right)^{-1},$$

and the matrix in the middle of the quadratic form

$$\begin{aligned}\frac{1}{n}\mathbf{A}'\mathbf{M}_1\mathbf{A} \xrightarrow{p} &\operatorname{plim}\frac{1}{n}\mathbf{A}'\mathbf{A} - \operatorname{plim}\frac{1}{n}\mathbf{A}'\mathbf{X}_1\left(\operatorname{plim}\frac{1}{n}\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\operatorname{plim}\frac{1}{n}\mathbf{X}_1'\mathbf{A} \\ \equiv &\mathbf{W}_{AA} - \mathbf{W}_{A1}\mathbf{W}_{11}^{-1}\mathbf{W}_{1A}.\end{aligned}$$

Therefore the quadratic form itself is asymptotically equal to,

$$\hat{\pi}_A'\left(\mathbf{W}_{AA} - \mathbf{W}_{A1}\mathbf{W}_{11}^{-1}\mathbf{W}_{1A}\right)\hat{\pi}_A.$$

Let π_A represent the probability limit of the coefficients on $\mathbf{A}$ in the first-stage regression of $\mathbf{X}_2$ on $\mathbf{X}_1$ and $\mathbf{A}$. Clearly if $\pi_A = \mathbf{0}$, the quadratic form will also equal zero in the limit and the instrumental variables estimator then becomes undefined. In general, the larger

are the π_A coefficients (it is irrelevant whether they take large positive or large negative values), the larger is the limiting value of the quadratic form, and the smaller is the limiting variance of the normalized estimator.

So the key findings from this analysis are, first, that the *additional instruments must be correlated (in the limit) with the problematic explanatory variables* if the instrumental variables method is to be valid, producing consistent estimates of the β parameters. Second, the size of the first-stage regression coefficients on the *additional instruments* affects the variance of the IV estimator. When those coefficients are small, you can expect instrumental variables to provide noisy and unreliable estimates of the true parameters.

Problems with weak "instruments"

Suppose that we are interested in a very simple model, $Y_i = \beta X_i + \epsilon_i$, in which the single explanatory variable X_i is measured in terms of deviations from its sample mean (Cameron and Trivedi 2005, p. 106). Let Z_i, a possible instrumental variable, also be measured in deviations from its mean. Then the OLS estimator is

$$\hat{\beta} = \beta + (\sum_i X_i^2)^{-1} \sum_i X_i \epsilon_i \xrightarrow{p} \beta + \frac{\text{Cov}(X, \epsilon)}{\text{Var}(X)} = \beta + \rho_{X,\epsilon} \cdot \frac{\sigma_\epsilon}{\sigma_X},$$

with $\rho_{X,\epsilon}$ being the correlation coefficient. Similarly, the formula for the IV estimator is

$$\hat{\beta}_{IV} = \beta + (\sum_i Z_i X_i)^{-1} \sum_i Z_i \epsilon_i \xrightarrow{p} \beta + \frac{\text{Cov}(Z, \epsilon)}{\text{Cov}(Z, X)} = \beta + \rho_{Z,\epsilon} \cdot \frac{\sigma_\epsilon \sigma_Z}{\text{Cov}(Z, X)}.$$

Dividing the second expression by the first, we find

$$\text{plim} \frac{\hat{\beta}_{IV} - \beta}{\hat{\beta} - \beta} = \frac{\rho_{Z,\epsilon}}{\rho_{X,\epsilon}} \cdot \frac{1}{\rho_{Z,X}}.$$

Now, if Z is in fact a valid instrument, its correlation with ϵ is zero. What if Z is not strictly valid, but is only very weakly correlated with ϵ? Suppose, for instance, that its correlation with ϵ is only one-tenth of the correlation of X with ϵ, so that $\rho_{Z,\epsilon}/\rho_{X,\epsilon} = 0.1$. If, in addition, the correlation between the "instrument" Z and X is low, being something less than 0.1, then the IV estimator will converge to a number that is further from the true β than the probability limit of the OLS estimator!

To put the result differently, imagine that you are considering a candidate instrument that is weak in the sense that it has little correlation with the X variable. (You know this because you've checked the empirical correlation between the two in your dataset.) Should you use this instrument? If you are very sure that the instrument is valid, then the answer is yes on asymptotic grounds. But if you have doubts as to its validity, you should be extra cautious—using a weak but slightly invalid "instrument" can lead you further astray than using ordinary least squares.

27.7 Imposing and Testing Linear Constraints

In deriving the constrained instrumental variables estimator, we follow much the same route as with similarly constrained ordinary least squares estimators. The problem of

estimating β subject to the constraint $\mathbf{R}\beta = \mathbf{r}$ is formulated using Lagrange multipliers, defining

$$L = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{P}_\mathbf{Z}(\mathbf{Y} - \mathbf{X}\beta) + \lambda'(\mathbf{R}\beta - \mathbf{r}).$$

Taking derivatives with respect to β and λ yields the estimate of the Lagrange multiplier,

$$\tilde{\lambda} = -\left(\mathbf{R}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\hat{\beta}_{IV} - \mathbf{r})$$

where $\hat{\beta}_{IV}$ is the unconstrained instrumental variables estimator. Using this, we obtain

$$\tilde{\beta}_{IV} = \hat{\beta}_{IV} - (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{R}'\left(\mathbf{R}(\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{R}'\right)^{-1}(\mathbf{R}\hat{\beta}_{IV} - \mathbf{r})$$

for the equality-constrained estimator.

Tests of the constraint can be carried out using a Wald-type test. In a later section of this chapter we will explore the Gauss-Newton regression, which provides an alternative approach. Yet another alternative to the Wald-type test is a modified $\mathcal{F}$ test, whose properties we shall develop here and compare to the Wald.

First consider the Wald test for $\beta_2 = \mathbf{0}$ in the model $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$. Recall that with the constraint matrix $\mathbf{R} = [0, \mathbf{I}_2]$ and null hypothesis $\mathbf{R}\beta_2 = \mathbf{0}$,

$$n^{1/2}\mathbf{R}\hat{\beta}_{IV} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2 \text{plim } \mathbf{R}(n^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}\mathbf{R}')\right),$$

and the relevant component of the covariance matrix is

$$(\hat{\mathbf{X}}_2'\mathbf{M}_1\hat{\mathbf{X}}_2)^{-1}$$

with

$$\mathbf{M}_1 = \mathbf{I} - \hat{\mathbf{X}}_1(\hat{\mathbf{X}}_1'\hat{\mathbf{X}}_1)^{-1}\hat{\mathbf{X}}_1' = \mathbf{I} - \mathbf{P}_\mathbf{Z}\mathbf{X}_1(\mathbf{X}_1'\mathbf{P}_\mathbf{Z}\mathbf{X}_1)^{-1}\mathbf{X}_1'\mathbf{P}_\mathbf{Z}.$$

Thus,

$$(\hat{\mathbf{X}}_2'\mathbf{M}_1\hat{\mathbf{X}}_2)^{-1} = (\mathbf{X}_2'\mathbf{P}_\mathbf{Z}\mathbf{M}_1\mathbf{P}_\mathbf{Z}\mathbf{X}_2)^{-1},$$

and the inverse of this matrix will appear in the middle of the quadratic form that defines the Wald statistic.

In the wings of that quadratic form $\hat{\beta}_2$ will appear. Applying the FWL theorem to the second stage of the 2SLS estimator, we find

$$\begin{aligned}\hat{\beta}_2 &= (\hat{\mathbf{X}}_2'\mathbf{M}_1\hat{\mathbf{X}}_2)^{-1}\hat{\mathbf{X}}_2'\mathbf{M}_1\mathbf{Y} \\ &= (\hat{\mathbf{X}}_2'\mathbf{M}_1\hat{\mathbf{X}}_2)^{-1}\mathbf{X}_2'\mathbf{P}_\mathbf{Z}\mathbf{M}_1\epsilon\end{aligned}$$

under the null since $\mathbf{P}_\mathbf{Z}\mathbf{M}_1\mathbf{X}_1 = \mathbf{0}$ as can be verified.

Assembling all these elements gives the Wald test statistic

$$W = \frac{\epsilon'\mathbf{M}_1\mathbf{P}_\mathbf{Z}\mathbf{X}_2(\mathbf{X}_2'\mathbf{P}_\mathbf{Z}\mathbf{M}_1\mathbf{P}_\mathbf{Z}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{P}_\mathbf{Z}\mathbf{M}_1\epsilon}{\hat{\sigma}^2} \xrightarrow{d} \chi^2_{k_2}$$

in which $\hat{\sigma}^2$ would be estimated from the residuals $\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}$.

As for the $\mathcal{F}$ test, the constrained model is given by the second stage regression

$$\mathbf{Y} = \hat{\mathbf{X}}_1\beta_1 + \text{residuals},$$

with sum of squared errors $\mathbf{e}_c'\mathbf{e}_c = \mathbf{Y}'\mathbf{M}_1\mathbf{Y}$, $\mathbf{M}_1$ being the matrix that appeared above. The unconstrained second stage regression is

$$\mathbf{Y} = \hat{\mathbf{X}}_1\beta_1 + \hat{\mathbf{X}}_2\beta_2 + \text{residuals}$$

and β_2 can be estimated from this by using the FWL theorem. Doing this yields the unconstrained error sum of squares,

$$\begin{aligned}\mathbf{e}_u'\mathbf{e}_u &= \mathbf{Y}'\mathbf{M}_1\mathbf{Y} - \mathbf{Y}'\mathbf{M}_1\mathbf{P}_\mathbf{Z}\mathbf{X}_2(\mathbf{X}_2'\mathbf{P}_\mathbf{Z}\mathbf{M}_1\mathbf{P}_\mathbf{Z}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{P}_\mathbf{Z}\mathbf{M}_1\mathbf{Y} \\ &= \mathbf{Y}'\mathbf{M}_1\mathbf{Y} - \epsilon'\mathbf{M}_1\mathbf{P}_\mathbf{Z}\mathbf{X}_2(\mathbf{X}_2'\mathbf{P}_\mathbf{Z}\mathbf{M}_1\mathbf{P}_\mathbf{Z}\mathbf{X}_2)^{-1}\mathbf{X}_2'\mathbf{P}_\mathbf{Z}\mathbf{M}_1\epsilon.\end{aligned}$$

Given this, we see that the numerator of the $\mathcal{F}$ test—the difference between the constrained and unconstrained sums of squares of the second stage regressions—is identical to the numerator of the Wald test statistic.[8]

Does this mean that one can test $\beta_2 = \mathbf{0}$ simply by applying a standard $\mathcal{F}$ test to the second stage regression? Unfortunately, it does not. Although the numerator of the $\mathcal{F}$ statistic will be correct asymptotically, the denominator—the probability limit of $\mathbf{e}_u'\mathbf{e}_u/(n-k)$—is not a consistent estimator of σ^2. The estimate of the variance based on $\mathbf{e}_u'\mathbf{e}_u$ will tend to be too large, see Davidson and MacKinnon (1993, p. 221). Thus, in using a standard regression package one would have to separately generate the residuals $\mathbf{Y} - \mathbf{X}\hat{\beta}_{IV}$ in order to estimate σ^2. Provided this denominator is correctly formed, one can then test $\beta_2 = \mathbf{0}$ using an $\mathcal{F}$ test based on two second stage regressions. Davidson and MacKinnon (1993) advocate this approach, basing their view on evidence suggesting that $\mathcal{F}$ statistics tend to exhibit better finite sample performance than the Wald statistic.

27.8 Nonlinear Models

Much of the discussion of the linear case carries over to the non-linear instrumental variables model. Some care is needed in applying the two-stage least squares approach to such models. It turns out that there are special cases—models linear in parameters but nonlinear in variables—for which a 2SLS approach will work if the endogenous variables are handled correctly, but the 2SLS approach does not work in general.

The general case

Let the nonlinear specification be $Y_i = h(\mathbf{X}_i, \theta) + \epsilon_i$ and as above, assume that we have a vector $\mathbf{Z}_i$ of $m > k$ valid instruments. Under homoskedasticity, $\mathrm{E}(\epsilon_i^2|\mathbf{Z}_i) = \sigma^2$. The nonlinear IV estimator is defined as the θ that minimizes the quadratic form

$$Q(\theta) = \frac{1}{n}\sum_i \mathbf{Z}_i(Y_i - h_i(\theta))' \cdot \mathbf{M}_n \cdot \frac{1}{n}\sum_i \mathbf{Z}_i(Y_i - h_i(\theta)),$$

[8] That is, identical apart from the constant k_2 by which the numerator of the $\mathcal{F}$ statistic must be divided.

and, leaving out $\hat{\sigma}^2$ as its value does not affect the minimizing $\hat{\theta}$, an asymptotically efficient form of the weighting matrix $\mathbf{M}_n$ is

$$\hat{\mathbf{V}}_n^{-1} = \left(\frac{1}{n} \sum_i \mathbf{Z}_i \mathbf{Z}_i' \right)^{-1}.$$

If the disturbances are heteroskedastic, then for the efficient weighting matrix we would instead employ

$$\hat{\mathbf{V}}_n^{-1} = \left(\frac{1}{n} \sum_i e_i^2 \mathbf{Z}_i \mathbf{Z}_i' \right)^{-1},$$

in which e_i is the IV residual from a first stage of estimation based on a weight matrix that is symmetric positive definite but inefficient. When we study the generalized method of moments in Chapter 28, we will examine the first-order conditions for such problems and will learn how to derive the limiting distribution of the transformed estimator $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

Tests of over-identifying restrictions for the non-linear case are very similar to those for the linear case, see Davidson and MacKinnon (1993, p. 236). Under the assumption of homoskedastic disturbances, we would simply regress the IV residuals $\mathbf{Y} - \hat{\mathbf{g}}$ on the full set of instruments $\mathbf{Z}$, and examine the nR^2 from this regression, which is distributed as χ^2_{m-k} under the null.

Using 2SLS methods in nonlinear problems

Two special cases merit further consideration. The first is a model that is linear in parameters, but nonlinear in respect to the manner in which the explanatory variables enter. For example,

$$Y_i = X_i \beta_1 + X_i^2 \beta_2 + \epsilon_i$$

is such a model if we are concerned that X_i may be correlated with ϵ_i. To handle this kind of situation, we simply treat X_i and X_i^2 as if they were two entirely different endogenous variables. In the first stage of 2SLS, we separately project $\mathbf{X}$ and $\mathbf{X}^2$ onto $\mathbf{Z}$ and employ these separate projections in the second stage. We should expect to encounter some difficulties because $\hat{\mathbf{X}}$ and $\widehat{\mathbf{X}^2}$ are likely to be highly correlated, and collinearity may prevent us from securing precise estimates of β_1 and β_2. Bowden and Turkington (1984, pp. 165–166) advise that care be taken to select instruments that exploit the nonlinear structure of the model, and urge that powers and other non-linear transforms of the basic $\mathbf{Z}$ variables be considered.

Cameron and Trivedi (2005, p. 198) show why we cannot simply project $\mathbf{X}$ onto $\mathbf{Z}$ and then enter the square of the projected value in the second stage of a two-stage least-squares approach. To get a sense of the techniques they use, we first consider a simple linear model with iid data,

$$\begin{aligned} Y_i &= X_i \beta + \epsilon_i \\ X_i &= Z_i \pi + u_i, \end{aligned}$$

and reconfirm that β can be estimated consistently by the method of two-stage least squares. Rewrite the first equation as

$$Y_i = \hat{X}_i \beta + (X_i - \hat{X}_i)\beta + \epsilon_i$$

and, treating $(X_i - \hat{X}_i)\beta + \epsilon_i$ as the composite disturbance, consider the probability limit that is the key to the proof of consistency,

$$\begin{aligned}\operatorname{plim} \frac{1}{n} \sum_i \hat{X}_i(X_i - \hat{X}_i) &= \operatorname{plim} \frac{1}{n} \sum_i \hat{X}_i(Z_i\pi + u_i - Z_i\hat{\pi}) \\ &= \operatorname{plim} \frac{1}{n} \sum_i Z_i\hat{\pi}(Z_i(\pi - \hat{\pi}) + u_i)\end{aligned}$$

This can be rearranged as

$$\operatorname{plim} \hat{\pi}(\pi - \hat{\pi}) \cdot \operatorname{plim} \frac{1}{n} \sum_i Z_i^2 + \operatorname{plim} \hat{\pi} \cdot \operatorname{plim} \frac{1}{n} \sum_i Z_i u_i.$$

The first term converges to zero by the consistency of $\hat{\pi}$, as does the second term given that Z_i and u_i are uncorrelated. In short, as we already knew, the two-stage least-squares estimator of β is consistent.

But when we apply a similar treatment to the nonlinear model,

$$\begin{aligned}Y_i &= X_i^2\beta + \epsilon_i \\ X_i &= Z_i\pi + u_i\end{aligned}$$

we find that simply inserting $(\hat{X}_i)^2$ into the second stage produces an *inconsistent* estimator of our β parameter. Rewrite the equation for Y_i as

$$Y_i = \hat{X}_i^2\beta + (X_i^2 - \hat{X}_i^2)\beta + \epsilon_i.$$

and consider plim $\frac{1}{n}\sum_i \hat{X}_i^2(X_i^2 - \hat{X}_i^2)$ to determine the consistency of the naive two-stage estimator. Substitute $Z_i\pi + u_i$ for X_i and $Z_i\hat{\pi}$ for $\hat{X}_i$, giving

$$X_i^2 - \hat{X}_i^2 = Z_i^2(\pi^2 - \hat{\pi}^2) + 2Z_i\pi u_i + u_i^2.$$

Now examine

$$\operatorname{plim} \frac{1}{n} \sum (\hat{\pi}Z_i)^2 \left(Z_i^2(\pi^2 - \hat{\pi}^2) + 2Z_i\pi u_i + u_i^2\right).$$

The three terms are

$$\operatorname{plim} \hat{\pi}^2(\pi^2 - \hat{\pi}^2) \cdot \operatorname{plim} \frac{1}{n} \sum Z_i^4,$$

which converges to zero by the consistency of $\hat{\pi}$, and

$$2 \operatorname{plim} \hat{\pi}^2\pi \cdot \operatorname{plim} \frac{1}{n} \sum Z_i^3 u_i,$$

which would also converge to zero if Z_i^3 is uncorrelated with u_i, leaving us with a troublesome third term,

$$\operatorname{plim} \hat{\pi}^2 \cdot \operatorname{plim} \frac{1}{n} \sum Z_i^2 u_i^2.$$

Even if Z_i and u_i were fully independent, the probability limit of this last term is non-zero—it would equal $\pi^2 \cdot \mathrm{E}\, Z_i^2 \,\mathrm{E}\, u_i^2$ under the assumption of iid data.

The second special case of the general nonlinear regression model is one in which the **X** variables enter linearly, but the parameters associated with them appear in a nonlinear form. A model of this kind is

$$Y_i = X_{1,i}\beta_1 + X_{2,i}\beta_1^2 + \epsilon_i.$$

A model such as this could be estimated as a linear model, with $Y_i = X_{1,i}\theta_1 + X_{2,i}\theta_2 + \epsilon_i$, subject to the non-linear constraint $\theta_2 = \theta_1^2$.

Note that if the fundamental assumption justifying the instrumental-variables approach holds, $\mathrm{E}(\epsilon_i|\mathbf{Z}_i) = 0$, then functions of $\mathbf{Z}_i$ can be used as additional instruments. For instance, in the structural model

$$Y_i = X_i\beta_1 + X_i^2\beta_2 + \epsilon_i$$

in which both X_i and X_i^2 are correlated with ϵ_i, the possible instruments include not only Z_i but also Z_i^2 and even Z_i^3, the cubed version of the instrument. Theory does not give us much rigorous guidance on which functional forms involving Z_i should be used, other than to check the standard criteria involving the strength of the partial correlation between the additional (excluded) instruments and the endogenous X_i and X_i^2, net of the included instruments (if there are any).

27.9 Control Function (CF) Approaches

An interesting variant on standard instrumental variables methods goes under the name of "control functions", which can be especially helpful in nonlinear models (Wooldridge 2015). The literature on control functions envisions the estimation problem in terms of a two-equation system (or a multi-equation system) in which the main issue is how to secure consistent estimates of the structural parameters in the one equation of principal interest. An issue that comes to the fore has to do with the correct specification of the *other* equation(s) of the system.

If you recall the method-of-moments and GMM approaches outlined earlier in this chapter, you'll remember that these alternative approaches focus only on the equation of interest and ask whether sufficient valid instruments **Z** are available to estimate that equation's structural parameters. In MM and GMM there is no model of the other equation(s) to contemplate (or at least that is not a necessary part of the set-up). The control function approach is therefore quite different in spirit; it takes a 2SLS view of the estimation problem and borrows ideas and terminology from this older econometric tradition.

To appreciate the method, suppose that we have a two-equation linear equation system

$$\begin{aligned} Y_{i,1} =& \mathbf{Z}_i'\pi + \epsilon_{i,1} \\ Y_{i,2} =& \mathbf{Z}_{i,2}'\beta + Y_{i,1}\alpha + \epsilon_{i,2} \end{aligned}$$

in which $\mathbf{Z}_{i,2}$ is a subset of the full vector of instruments $\mathbf{Z}_i$. (In other words, at least one instrument has been excluded from the structural equation of interest, which is the second equation of the pair.) Clearly if $\epsilon_{i,1}$ is correlated with $\epsilon_{i,2}$, then $Y_{i,1}$ is endogenous in the second equation. The essence of the control function idea is to insert into the second equation an observed proxy for $\epsilon_{i,1}$ that can be calculated from the first equation. Might

$$e_{i,1} = Y_{i,1} - \mathbf{Z}_i'\hat{\pi},$$

the residual from estimation of the first equation, do the job? Could this residual "control for" the endogeneity of $Y_{i,1}$ to such an extent that we are able to estimate the structural parameters of the second equation consistently?

The answer is "yes", but only if some additional assumptions are made that go beyond what the instrumental variables method strictly requires. First, rewrite the residual as

$$e_{i,1} = Y_{i,1} - \mathbf{Z}_i'\hat{\pi} = \epsilon_{i,1} - \mathbf{Z}_i'(\hat{\pi} - \pi).$$

Provided that the conditional expectation $\mathrm{E}(\epsilon_{i,1}|\mathbf{Z}) = 0$, the first stage estimator $\hat{\pi}$ is consistent for π and thus $e_{i,1} \stackrel{a}{=} \epsilon_{i,1}$. In other words, the first-equation residual is a valid proxy for the first-equation disturbance term.

Now examine the conditional expectation of $Y_{i,2}$ given the full set of instruments $\mathbf{Z}_i$ and also $Y_{i,1}$. To begin, we see that $\mathrm{E}(Y_{i,2}|\mathbf{Z}_i, Y_{i,1}) = \mathrm{E}(Y_{i,2}|\mathbf{Z}_i, \epsilon_{i,1})$ since $\mathbf{Z}_i$ and $\epsilon_{i,1}$ together determine $Y_{i,1}$. Hence,

$$\mathrm{E}(Y_{i,2}|\mathbf{Z}_i, Y_{i,1}) = \mathbf{Z}_{2,i}'\beta + Y_{i,1}\alpha + \mathrm{E}(\epsilon_{i,2}|\mathbf{Z}, \epsilon_{i,1}).$$

The *first* substantive assumption we need is that $\mathrm{E}(\epsilon_{i,2}|\mathbf{Z}, \epsilon_{i,1}) = \mathrm{E}(\epsilon_{i,2}|\epsilon_{i,1})$. This would not hold in general, but if $(\epsilon_{i,1}, \epsilon_{i,2})$ happened to be independent of $\mathbf{Z}_i$, the assumption would be valid. Imposing the assumption takes us this far,

$$\mathrm{E}(Y_{i,2}|\mathbf{Z}_i, Y_{i,1}) = \mathbf{Z}_{2,i}'\beta + Y_{i,1}\alpha + \mathrm{E}(\epsilon_{i,2}|\epsilon_{i,1}).$$

Now we invoke a *second* substantive assumption, that $\mathrm{E}(\epsilon_{i,2}|\epsilon_{i,1}) = \rho\epsilon_{i,1}$. In other words, we assume that the conditional expectation of $\epsilon_{i,2}$ given $\epsilon_{i,1}$ is *linear* in $\epsilon_{i,1}$. This, too, does not hold as a general rule. If, however, $(\epsilon_{i,1}, \epsilon_{i,2})$ happens to be joint normal as well as fully independent of $\mathbf{Z}_i$, the assumption would hold. (It is unclear how generally the linear conditional expectation assumption applies where other joint distributions are concerned.) As we will see in some detail in Chapter 32, popular maximum-likelihood models of sample selectivity exploit the full implications of disturbance-term joint normality.

If the conditional expectations assumption is valid, then we have a very interesting result,

$$\mathrm{E}(Y_{i,2}|\mathbf{Z}_i, Y_{i,1}) = \mathbf{Z}_{2,i}'\beta + Y_{i,1}\alpha + \rho\epsilon_{i,1}.$$

The parameters of this conditional expectation are precisely the structural parameters of interest, β and α, augmented with ρ times the first-equation disturbance term. As we already know, the first-equation *residual* $e_{i,1}$ is asymptotically equivalent to the $\epsilon_{i,1}$ disturbance. We exploit this to arrive at an equation whose parameters β, α, ρ we can estimate consistently despite the endogeneity of $Y_{i,1}$ in the structural equation,

$$\mathrm{E}(Y_{i,2}|\mathbf{Z}_i, Y_{i,1}) = \mathbf{Z}_{2,i}'\beta + Y_{i,1}\alpha + \rho e_{i,1}.$$

Now, to be sure, because $e_{i,1}$ is a "generated" covariate produced by a first-stage estimator $\hat{\pi}$, which differs from the true π by sampling error, the standard errors of $\hat{\beta}, \hat{\alpha}, \hat{\rho}$ will all need to be adjusted.

You can appreciate the roles played by the assumptions by comparing the result to what taking conditional expectations alone would deliver:

$$\mathrm{E}(Y_{i,2}|\mathbf{Z}_i, Y_{i,1}) = \mathbf{Z}_{2,i}'\beta + Y_{i,1}\alpha + \mathrm{E}(\epsilon_{i,2}|\mathbf{Z}_i, Y_{i,1}).$$

As it stands, there is nothing particularly useful that can be said about the $\mathrm{E}(\epsilon_{i,2}|\mathbf{Z}_i, Y_{i,1})$ term on the right. The additional assumptions allow us to simplify this otherwise intractable term to $\rho\epsilon_{i,1}$. Also, you might wonder how the exclusion of some instruments from the structural equation enters this argument. As Wooldridge (2015) writes, "The exogenous variation induced by excluded instrumental variables provides separate variation in the residuals (or generalized residuals) obtained from a reduced form, and these residuals serve as the control functions."

Now comes the interesting part: Looking back on the proof we've just outlined, we can see that it applies without modification to the *nonlinear structural model*

$$Y_{i,2} = \mathbf{Z}_{2,i}'\beta + Y_{i,1}\alpha_1 + Y_{i,1}^2\alpha_2 + \epsilon_{i,2}.$$

In conditional-expectations form with the first-equation residual inserted in place of $\epsilon_{i,1}$, this equation becomes

$$\mathrm{E}(Y_{i,2}|\mathbf{Z}_i, Y_{i,1}) = \mathbf{Z}_{2,i}'\beta + Y_{i,1}\alpha_1 + Y_{i,1}^2\alpha_2 + \rho e_{i,1}.$$

All of its structural parameters, along with ρ, can be estimated consistently under the same assumptions we've been using for the linear case. So long as the control function assumptions are met, there is no need for additional instruments as would have been the case with conventional instrumental variables approaches. Indeed, the control function method can be extended to cover more complicated nonlinear equation systems such as

$$\begin{aligned} Y_{i,1} =& g(\mathbf{Z}_i, \pi) + \epsilon_{i,1} \\ Y_{i,2} =& \mathbf{Z}_{2,i}'\beta + h(Y_{i,1}, \alpha) + \epsilon_{i,2} \end{aligned}$$

in which $g()$ and $h()$ are nonlinear functions.

27.10 Random-coefficient models: IV and CF approaches

Structural models in which an individual-specific coefficient is attached to a right-hand side endogenous variable cause conventional instrumental-variables models to become inconsistent. As Wooldridge (2015) explains, it is possible to rescue IV consistency with an additional constant-covariance assumption. Control-function methods offer another route for consistent estimation in such random-coefficient models.

Consider the structural model

$$Y_{i,2} = \alpha + \mathbf{X}_i'\beta + \gamma_i Y_{i,1} + \epsilon_i$$

with the endogenous right-hand side $Y_{i,1}$ determined via

$$Y_{i,1} = \mathbf{Z}_i'\pi + u_i.$$

Here, $\mathbf{Z}_i$ is the full set of exogenous instruments, and $\mathbf{X}_i \subset \mathbf{Z}_i$ with at least one variable in $\mathbf{Z}_i$ excluded from $\mathbf{X}_i$. Let the individual-specific coefficient $\gamma_i = \gamma + g_i$. In this context, instrument validity means not only that $\mathrm{E}(\epsilon_i|\mathbf{Z}_i) = 0$ but also $\mathrm{E}(g_i|\mathbf{Z}_i) = 0$.

Rewrite the structural model as

$$Y_{i,2} = \alpha + \mathbf{X}_i'\beta + \gamma Y_{i,1} + g_i Y_{i,1} + \epsilon_i$$

in which $g_i Y_{i,1} + \epsilon_i$ is now the composite disturbance term. It is clear that without additional assumptions, the core assumption of instrumental variables—that instruments be correlated with the right-hand-side endogenous variable $Y_{i,1}$ but not correlated with the composite disturbance—cannot be satisfied: $Y_{i,1}$ is part of that composite disturbance.

Wooldridge (2015) notes that conditional on instruments $\mathbf{Z}_i$, the covariance of g_i and $Y_{i,1}$ is

$$C_i = \mathrm{E}(g_i\, Y_{i,1} \mid \mathbf{Z}_i)$$

since $\mathrm{E}(g_i|\mathbf{Z}_i) = 0$ by assumption. Suppose, however, that this covariance were *constant*, that is, not a function of $\mathbf{Z}_i$,

$$\mathrm{E}(g_i\, Y_{i,1} \mid \mathbf{Z}_i) = \mathrm{E}(g_i\, Y_{i,1}) = C.$$

We could then re-cast the structural model as

$$Y_{i,2} = (\alpha + C) + \mathbf{X}_i'\beta + \gamma Y_{i,1} + (g_i Y_{i,1} - C) + v_i,$$

a reformulation that allows $\mathbf{Z}_i$ to be correlated with $Y_{i,1}$ and yet not correlated with the composite disturbance $(g_i Y_{i,1} - C) + v_i$. All that has been lost is the ability to identify and consistently estimate the α constant term.

In developing an alternative control-function approach to the random-coefficients problem, Wooldridge (2015) makes the familiar linearity assumption about the conditional expectation of the two structural equation disturbances g_i and ϵ_i given the u_i disturbance of the $Y_{i,1}$ equation,

$$\mathrm{E}(g_i|u_i) = a\, u_i \text{ and } \mathrm{E}(\epsilon_i|u_i) = b\, u_i.$$

Additionally, the three disturbances (u_i, g_i, ϵ_i) are assumed to be fully independent of $\mathbf{Z}_i$. Then from

$$\mathrm{E}(Y_{i,2}|\mathbf{Z}_i, Y_{i,1}, u_i) = \alpha + \mathbf{X}_i'\beta + \gamma Y_{i,1} + a\, u_i Y_{i,1} + b\, u_i$$

and using the first-equation residual $\hat{u}_i$ in place of u_i, we can estimate the α, β, γ parameters via the estimating equation

$$Y_{i,2} = \alpha + \mathbf{X}_i'\beta + \gamma Y_{i,1} + a\, \hat{u}_i Y_{i,1} + b\, \hat{u}_i.$$

Wooldridge (2015) recommends bootstrapping to estimate the standard errors.

Chapter 28

Generalized Method-of-Moments Estimators

Students: Supplement this chapter by reading Cameron and Trivedi (2005, Chapter 6).

There is now a large literature on GMM, but to date there are few treatments of the topic that are accessible to beginning graduate students. Hansen (1982) is the seminal reference in the econometrics literature, although of course the method of moments approach was in use long before he generalized it. Hansen is especially concerned with the application of the method to dependent observations (e.g., time-series data) and provides a more extensive and general set of results than we can discuss here. This chapter, based in large part on lecture notes by McFadden (1999), is pitched at an intermediate level with minimal attention to technical regularity conditions. Students seeking more of the technical detail can consult Newey and McFadden (1994) for updates and restatements of Hansen's work; Hayashi (2000) is written in a way that parallels and nicely complements the Newey–McFadden piece.

28.1 The GMM Approach

Consider a sample of size n, with the observed and unobserved random variables being $(Y_i, \mathbf{X}_i, \mathbf{Z}_i, \epsilon_i)$, $i = 1, \ldots, n$. We will often be thinking of Y_i as a dependent variable, $\mathbf{X}_i$ as a set of explanatory variables (possibly endogenous), $\mathbf{Z}_i$ as a set of instruments, and ϵ_i as a disturbance term. The parameter vector of interest is θ, which is of dimension k. An example is provided by the case of a nonlinear regression with instrumental variables, in which $Y_i = \phi(\mathbf{X}_i, \theta) + \epsilon_i$ and the $\mathbf{Z}_i$ are instruments used to handle the endogeneity of some of the $\mathbf{X}_i$ covariates. In other cases, however, there may be no natural way to separate Y_i and $\mathbf{X}_i$ into dependent and explanatory categories. For this handout, we will assume that $(Y_i, \mathbf{X}_i, \mathbf{Z}_i, \epsilon_i)$ is drawn from an i.i.d. or i.n.i.d. sequence, but see Hayashi (2000) for guidance on how to handle dependent observations.

The researcher has specified a vector of $m \geq k$ functions of the observed variables, which we denote by $\mathbf{g}(Y_i, \mathbf{X}_i, \mathbf{Z}_i, \theta)$, or, in a more compact notation, simply by $\mathbf{g}_i(\theta)$. Each

element $j = 1, \ldots, m$ of this vector has the property that $\mathrm{E}\, \mathbf{g}_i^j(\theta_0) = 0$, with θ_0 being the true value of θ. These m functions are termed "generalized moments" and the expectation conditions termed "moment conditions." For the nonlinear instrumental variables example with $Y_i = \phi(\mathbf{X}_i, \theta) + \epsilon_i$ and with m instruments in the set $\mathbf{Z}_i$ (some of which will overlap with variables in $\mathbf{X}_i$), we have

$$\mathbf{g}_i(\theta_0) = \mathbf{Z}_i \left(Y_i - \phi(\mathbf{X}_i, \theta_0)\right) = \mathbf{Z}_i \epsilon_i$$

and, since the $\mathbf{Z}_i$ are exogenous instruments, $\mathrm{E}\, \mathbf{Z}_i \epsilon_i = \mathbf{0}$.

The expression $n^{-1} \sum_{i=1}^n \mathbf{g}_i(\theta_0)$ is the sample analog to the population moment $\mathrm{E}\, \mathbf{g}(Y, \mathbf{X}, \mathbf{Z}, \theta_0)$. As $n \to \infty$, we expect a law of large numbers to guarantee that the sample analogs will converge in probability to zero, the value of the corresponding population moments. For the nonlinear IV case, the sample analog to the population moment is

$$\frac{1}{n} \sum_i \mathbf{Z}_i \left(Y_i - \phi(\mathbf{X}_i, \theta_0)\right),$$

which (since $\mathbf{Z}_i$ is an $m \times 1$ column vector) is a set of m moments.

When $m = k$ we can solve directly for the $\hat{\theta}$ values that set all m sample moments to zero,

$$\frac{1}{n} \sum_i \mathbf{Z}_i \left(Y_i - \phi(\mathbf{X}_i, \hat{\theta})\right) = \mathbf{0}_{k \times 1}$$

typically using numerical methods to obtain the solution. We would then apply asymptotic theory to determine the properties of $\hat{\theta}$ and the limiting distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$. The relevant theory is discussed in treatments of the conventional method of moments approach; it is very similar to what we will outline below for the more general case.

If $m > k$ then not all m sample moments can be set to zero simultaneously and a different approach is needed, which goes under the name of the *generalized* method of moments. To estimate θ with $m > k$, we define a quadratic form $S_1(\theta)$ in which the vectors on the wings are $m \times 1$ and the $m \times m$ matrix $\mathbf{W}_n$ is symmetric and positive definite,

$$S_1(\theta) = \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\theta) \right)' \mathbf{W}_n \left(\frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\theta) \right). \tag{28.1}$$

We proceed to minimize $S_1(\theta)$ by differentiating it with respect to θ to obtain a reduced set of k equations in k unknowns. We then solve these first-order conditions for $\hat{\theta}$. The first-order conditions—see Appendix A for the full details, which are very much worth knowing—can be represented as

$$\left(\frac{1}{n} \sum_{i=1}^n \mathbf{G}_i(\hat{\theta}) \right)' \mathbf{W}_n \left(\frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\hat{\theta}) \right) = \mathbf{0}_{k \times 1}, \tag{28.2}$$

with $\mathbf{G}_i$ being the $m \times k$ matrix of derivatives for the i-th observation. In its (j, l) entry, $\mathbf{G}_i$ has typical element $\partial \mathbf{g}_i^j / \partial \theta_l$, this being the derivative of the j-th function with respect to the l-th parameter. (The Appendix explains why the $1/2$ factor has disappeared.)

Newey and McFadden (1994) show that the estimator $\hat{\theta} \xrightarrow{p} \theta_0$. Their consistency proof, which we will not reproduce here, relies on a uniform law of large numbers, and is a generalization of the proof for maximum likelihood estimators. Given consistency, it is not too hard to see how to derive the asymptotic distribution of $\sqrt{n}(\hat{\theta} - \theta_0)$. First, note that by a uniform law of large numbers and the consistency of $\hat{\theta}$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{G}_i(\hat{\theta}) \xrightarrow{p} \Gamma_{m\times k},$$

where $\Gamma = \mathrm{E}\,\mathbf{G}_i(\theta_0)$ in the case of i.i.d. sequences and $\Gamma = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}\,\mathbf{G}_i(\theta_0)$ in the case of i.n.i.d. sequences. Second, assume that the weighting matrix $\mathbf{W}_n$ is defined in such a way that $\mathbf{W}_n$ converges to a positive definite symmetric matrix $\mathbf{W}$. Then, taking the sample size n to be large, re-express the GMM first-order condition (equation 28.2) as

$$\Gamma'\mathbf{W}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\hat{\theta})\right) \overset{a}{=} \mathbf{0}. \tag{28.3}$$

Now Taylor-expand the factor in parentheses around the true θ_0, giving

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\hat{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\theta_0) + \frac{1}{n}\sum_{i=1}^{n} \mathbf{G}_i(\theta^*)(\hat{\theta} - \theta_0).$$

where θ^* lies between θ_0 and $\hat{\theta}$. Insert this into the first-order conditions (equation 28.3) and multiply through by $\sqrt{n}$, very much as we did in manipulating ML first-order conditions. Making use once again of consistency and the result that $n^{-1}\sum \mathbf{G}_i(\theta^*) \xrightarrow{p} \Gamma$, we obtain an asymptotic version of the first-order conditions,

$$\Gamma'\mathbf{W}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\theta_0)\right) + \Gamma'\mathbf{W}\Gamma \cdot \sqrt{n}(\hat{\theta} - \theta_0) \overset{a}{=} \mathbf{0}.$$

This can also be written as

$$\sqrt{n}(\hat{\theta} - \theta_0) \overset{a}{=} -(\Gamma'\mathbf{W}\Gamma)^{-1}\Gamma'\mathbf{W}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\theta_0)\right). \tag{28.4}$$

A central limit theorem for i.i.d. or i.n.i.d. sequences establishes that

$$\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Delta),$$

where Δ is

$$\Delta = \mathrm{E}\,\mathbf{g}_i(\theta_0)\mathbf{g}_i(\theta_0)'$$

in the case of i.i.d. sequences and for i.n.i.d. sequences, with $\mathbf{V}_i = \mathrm{E}\,\mathbf{g}_i(\theta_0)\mathbf{g}_i(\theta_0)'$, we define $\Delta = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} \mathbf{V}_i$. Hence,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} -(\Gamma'\mathbf{W}\Gamma)^{-1}\Gamma'\mathbf{W} \cdot \mathcal{N}(\mathbf{0}, \Delta) \tag{28.5}$$

$$\xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\Gamma'\mathbf{W}\Gamma)^{-1}\Gamma'\mathbf{W}\Delta\mathbf{W}\Gamma(\Gamma'\mathbf{W}\Gamma)^{-1}\right) \tag{28.6}$$

We can consistently estimate Γ by its sample analog, using

$$\hat{\Gamma} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{G}_i(\hat{\theta}), \tag{28.7}$$

and can consistently estimate Δ using the outer product

$$\hat{\Delta} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\hat{\theta})\mathbf{g}_i(\hat{\theta})'. \tag{28.8}$$

For a given choice of the weighting matrix $\mathbf{W}_n$ and knowledge of the matrix $\mathbf{W}$ to which it converges, we now have all the ingredients needed to implement the GMM estimator.

One choice for $\mathbf{W}_n$ would be $\mathbf{I}$, the identity matrix. This choice would yield a consistent estimator for θ_0 by the arguments made above and we could estimate its asymptotic variance. We would not really expect $\mathbf{W}_n = \mathbf{I}$ to be the best choice. In fact, the specification $\mathbf{W}_n = \mathbf{I}$ is normally used only as the first step in a two-step procedure, to which we now turn.

28.2 Efficient GMM

What, then, *is* the best choice of $\mathbf{W}_n$? For an efficient GMM estimator—one with the smallest possible asymptotic variance in the class of estimators—we want a sequence $\{\mathbf{W}_n\}$ that converges to $\mathbf{W} = \Delta^{-1}$. This result is due to Hansen (1982) and the logic leading to it is reviewed by Newey and McFadden (1994). If Δ were known, then after substituting for $\mathbf{W}$ in equation (28.6), the efficient GMM estimator $\tilde{\theta}$ would be asymptotically distributed as

$$\sqrt{n}(\tilde{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, (\Gamma'\Delta^{-1}\Gamma)^{-1}\right). \tag{28.9}$$

Of course, Δ is not known, but we can use the consistent estimator $\hat{\Delta}$ shown in equation (28.8) in its place.

To implement the efficient estimator, the model is first estimated with $\mathbf{W}_n = \mathbf{I}$, yielding an initial estimator $\hat{\theta}$ and a covariance estimate $\hat{\Delta}$. Second, letting $\mathbf{W}_n = \hat{\Delta}^{-1}$, we set up the new quadratic form

$$S_2(\theta) = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\theta)\right)' \hat{\Delta}^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i(\theta)\right), \tag{28.10}$$

and find its minimum with respect to θ. The minimizing $\tilde{\theta}$ is the feasible version of the efficient GMM estimator.

How do we know that setting $\mathbf{W}_n = \hat{\Delta}^{-1}$ is the efficient way to proceed? Examine the difference in the variance matrices of the inefficient and efficient estimators:

$$(\Gamma'\mathbf{W}\Gamma)^{-1}\Gamma'\mathbf{W}\Delta\mathbf{W}\Gamma(\Gamma'\mathbf{W}\Gamma)^{-1} - (\Gamma'\Delta^{-1}\Gamma)^{-1}.$$

This difference is positive semidefinite if and only if the difference

$$\Gamma'\Delta^{-1}\Gamma - (\Gamma'\mathbf{W}\Gamma)(\Gamma'\mathbf{W}\Delta\mathbf{W}\Gamma)^{-1}(\Gamma'\mathbf{W}\Gamma)$$

is positive semidefinite. Cleverly factoring things, we can write the second expression as

$$\Gamma'\Delta^{-1/2}\left(\mathbf{I} - \Delta^{1/2}\mathbf{W}\Gamma(\Gamma'\mathbf{W}\Delta\mathbf{W}\Gamma)^{-1}\Gamma'\mathbf{W}\Delta^{1/2}\right)\Delta^{-1/2}\Gamma.$$

If we define $\mathbf{Z} = \Delta^{1/2}\mathbf{W}\Gamma$, then the matrix in parentheses can be written as $\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'$, which is symmetric idempotent and therefore positive semidefinite.

28.3 Example: Poisson Count-Data Models with Endogeneity

Back in Chapter 22 we laid out the maximum-likelihood approach to estimating Poisson count-data models with exogenous explanatory variables (assuming *i.n.i.d.* data series). Recall that the probability mass function for $y \in \{0, 1, 2, \ldots\}$ is

$$\Pr(Y_i = y \mid \mathbf{X}_i) = \frac{\lambda_i^y e^{\lambda_i}}{y!}$$

with the $\mathbf{X}_i$ covariates entering via the strictly positive $\lambda_i = e^{\mathbf{X}_i'\theta}$. In this specification, the expected value of Y_i given $\mathbf{X}_i$ is $e^{\mathbf{X}_i'\theta}$, and the conditional variance also equals this same value. Unfortunately, the equality of mean and variance is not commonly supported in count datasets.

Given the evident tension between model and data, one strand of the literature has pursued less restrictive method-of-moments methods to estimate $\mathrm{E}(Y_i|\mathbf{X}_i) = e^{\mathbf{X}_i'\theta}$ without imposing strict equality between mean and variance. Additionally, in the recent literature, moments are formed that allow $Y_i = 0$ counts to be included—this is important given the typically substantial proportion of zero-event cases in a count dataset. Following Mullahy (1997), this approach introduces a "multiplicative" non-negative disturbance term via

$$Y_i = e^{\mathbf{X}_i'\theta}\,\epsilon_i.$$

As is the case even with linear regression models with constant terms—here, $\mathbf{X}_i'\theta$ is assumed to include such a term—some normalization needs to be imposed on $\mathrm{E}\,\epsilon_i$ in order to identify the constant term. The natural normalizing assumption for the multiplicative Poisson is $\mathrm{E}(\epsilon_i|\mathbf{X}_i) = \mathrm{E}(\epsilon_i) = 1$. This assumption also helps with the interpretation of the marginal effects of covariates on the conditional mean, $\partial Y_i/\partial \mathbf{X}_{j,i} = \theta_j\, e^{\mathbf{X}_i'\theta}\,\epsilon_i$, whose expected value conditional on $\mathbf{X}_i$ is just

$$\mathrm{E}\left(\frac{\partial Y_i}{\partial \mathbf{X}_{j,i}} \mid \mathbf{X}_i\right) = \theta_j\, e^{\mathbf{X}_i'\theta}.$$

given that $\mathrm{E}(\epsilon_i|\mathbf{X}_i) = 1$.

Keeping this normalizing assumption in mind, for a standard method-of-moments approach we first express ϵ_i in terms of θ and the observables and then, having isolated ϵ_i in this way, proceed to subtract 1 from it. Thus at the true θ_0 we can write

$$\epsilon_i - 1 = Y_i\, e^{-\mathbf{X}_i'\theta_0} - 1$$

Assuming exogenous $\mathbf{X}_i$ covariates,

$$\mathrm{E}\left((Y_i\, e^{-\mathbf{X}_i'\theta_0} - 1)|\mathbf{X}_i\right) = \mathrm{E}(\epsilon_i - 1 \mid \mathbf{X}_i) = \mathrm{E}(\epsilon_i|\mathbf{X}_i) - 1 = 0.$$

This expression provides the basis for the k sample moments

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i\left(Y_i\,e^{-\mathbf{X}_i'\theta}-1\right),$$

which would be used to estimate the θ parameters in a standard method-of-moments approach.

Note that the discrete-valued, count-data features of Y_i have by now receded well into the background. We really haven't exploited these features. Indeed, these moment equations could be used for any strictly non-negative Y_i variable with conditional mean $\mathrm{E}(Y_i|\mathbf{X}_i) = e^{\mathbf{X}_i\theta}$. One could argue with some justification that the model being estimated should be termed a "quasi-Poisson" model since it violates the strict mean-equals-variance criterion of the Poisson model proper.

For an over-identified *generalized method-of-moments* model with $m > k$ instruments $\mathbf{Z}_i$ satisfying the key assumption $\mathrm{E}(\epsilon_i|\mathbf{Z}_i) = 1$, we would use instead the specification

$$g_i(\theta) = \mathbf{Z}_i\left(Y_i\,e^{-\mathbf{X}_i'\theta}-1\right),$$

and would employ $\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$ in the wings of the GMM quadratic form. In a just-identified case with only $m = k$ valid instruments, $\frac{1}{n}\sum_{i=1}^{n} g_i(\theta)$ would be used to find the method-of-moments estimator of θ allowing for the endogeneity of the $\mathbf{X}_i$ covariates.

28.4 Example: Probit-Like Models with Endogeneity

Chapter 20 outlines the usual maximum-likelihood approach to probit and logit models involving exogenous explanatory variables. We want to develop a comparable appraoch for the case in which at least some of the covariates are endogenous. Consider a "yes-no" binary dependent variable Y_i that depends on covariates $\mathbf{X}_i$, some of which are endogenous. We can write the model in the form of a nonlinear regression with an *additive* disturbance term,

$$Y_i = \Phi(\mathbf{X}_i'\theta) + \epsilon_i,$$

with the endogeneity of $\mathbf{X}_i$ expressed as $\mathrm{E}\,\epsilon_i|\mathbf{X}_i \neq 0$. Note that this is not, strictly speaking, a probit model; rather it is a nonlinear equation that uses the same Φ function that appears in a probit model. Nevertheless, it can be regarded as a probit-like specification that, among other things, has the nice feature that the predicted values $\hat{Y}_i = \Phi(\mathbf{X}_i'\hat{\theta})$ lie between zero and one. In this set-up, the ϵ_i disturbances are heteroskedastic. In principle we could introduce a correction for heteroskedasticity in the first-stage consistent GMM estimator, but in what follows we will make that correction instead in the second-stage efficient GMM estimator.

Suppose that we have m valid instruments $\mathbf{Z}_i$. The vector of m functions that we will use is

$$\mathbf{Z}_i(Y_i - \Phi(\mathbf{X}_i'\theta)),$$

which at the true θ_0 equals $\mathbf{Z}_i\epsilon_i$ and has expectation zero given the valid instruments condition $\mathrm{E}(\epsilon_i \mid \mathbf{Z}_i) = 0$. The sample analog to the population moment is

$$\frac{1}{n}\sum_{i}^{n}\mathbf{Z}_i(Y_i - \Phi(\mathbf{X}_i'\theta)) \equiv \frac{1}{n}\sum_{i}^{n} g_i(\theta).$$

The quadratic form to be minimized is, in its general form,

$$S_1(\theta) = \frac{1}{2}\left(\frac{1}{n}\sum_i g_i(\theta)\right)' \mathbf{W}_n \left(\frac{1}{n}\sum_i g_i(\theta)\right)$$

and, as you will recall, this yields the general form of the GMM first-order condition

$$\left(\frac{1}{n}\sum_i \mathbf{G}_i(\hat{\theta})\right)' \mathbf{W}_n \left(\frac{1}{n}\sum_i g_i(\hat{\theta})\right) = \mathbf{0}, \tag{28.11}$$

where the $\mathbf{G}_i$ matrix of this expression has entries that are the derivatives of the m moments with respect to the k parameters, and $\mathbf{G}_i$ is thus of dimension $m \times k$.

To understand $\mathbf{G}_i$ in the case at hand, consider the j-th instrument Z_{ij} and the l-th element θ_l. The (j, l) entry of $\mathbf{G}_i$ is then

$$\frac{\partial Z_{ij}(Y_i - \Phi(\mathbf{X}_i'\hat{\theta}))}{\partial \theta_l} = -\phi(\mathbf{X}_i'\hat{\theta}) Z_{ij} X_{il},$$

and, writing this out for all j and l,

$$\mathbf{G}_i = -\phi(\mathbf{X}_i'\hat{\theta}) \begin{bmatrix} Z_{i1} \\ \vdots \\ Z_{im} \end{bmatrix} \begin{bmatrix} X_{i1} & \dots & X_{ik} \end{bmatrix} = -\phi(\mathbf{X}_i'\hat{\theta}) \cdot \mathbf{Z}_i \mathbf{X}_i'.$$

Using some positive definite $\mathbf{W}_n$ matrix to secure a first-stage consisent estimator $\hat{\theta}$, and letting the residual $e_i = Y_i - \Phi(\mathbf{X}_i'\hat{\theta})$, we then form the efficient weighting matrix using

$$\hat{\Delta} == \frac{1}{n}\sum_i \mathbf{g}_i(\hat{\theta})\mathbf{g}_i(\hat{\theta})' = \frac{1}{n}\sum_i e_i^2 \mathbf{Z}_i \mathbf{Z}_i'$$

and employ $\hat{\Delta}^{-1}$ as the weighting matrix.

28.5 Two-step Estimation: GMM and ML

The discussion below follows McFadden (1999). Consider a likelihood function that can be factored into marginal and conditional components,

$$f(y_1, y_2, |\mathbf{X}, \alpha, \beta) = f^c(y_2|\mathbf{X}, y_1, \alpha, \beta) \cdot f^m(y_1|\mathbf{X}, \alpha),$$

where $\mathbf{X}$ represents exogenous variables and y_1, y_2 are the endogenous variables. In a two-step approach, $\hat{\alpha}$ is estimated first and in the second step, $\hat{\beta}$ is found taking $\hat{\alpha}$ as given. In a number of cases, this approach much simplifies the estimation task relative to what it would be if the full likelihood were maximized directly. However, there is a price to be paid for this: the two-step approach is less efficient compared with the full-information maximum-likelihood approach, and additional programming is needed to correct the standard errors of $\hat{\beta}$ in the second step of the procedure.

To see what is involved, we will first set up the problem in a GMM framework, using one moment vector $g(y_1, \mathbf{X}, \alpha)$ to estimate α and another moment vector $h(y_1, y_2, \mathbf{X}, \alpha, \beta)$ to estimate β taking $\alpha = \hat{\alpha}$. Then we will see how the covariance matrix corrections are simplified in a maximum likelihood application, taking g to be $\partial \log f^m / \partial \alpha$ and taking $h = \partial \log f^c / \partial \beta$, these being the scores of the marginal and conditional likelihood components with $\hat{\alpha}$ fixed in the latter.

The GMM Approach

Let α be $k \times 1$ and assume that there are k moments in the g vector; likewise, let β be $l \times 1$ and assume that the h vector has l moments. In this simplified setup (which can be generalized) the estimate $\hat{\alpha}$ is found by setting all k elements of g equal to zero and similarly for the h vector.

With $\hat{\alpha}$ in hand, we can expand $n^{-1} \sum \mathbf{g}_i(\hat{\alpha}) = \mathbf{0}$ around the true α_0 and then multiply by $\sqrt{n}$, yielding

$$\sqrt{n} \frac{1}{n} \sum \mathbf{g}_i(\hat{\alpha}) = \sqrt{n} \frac{1}{n} \sum \mathbf{g}_i(\alpha_0) + \frac{1}{n} \sum \frac{\partial \mathbf{g}_i(\alpha^*)}{\partial \alpha} \cdot \sqrt{n}(\hat{\alpha} - \alpha_0) = \mathbf{0}.$$

Let the $k \times k$ matrix $\mathbf{G}$ be the probability limit of $-\frac{1}{n} \sum \frac{\partial \mathbf{g}_i(\alpha^*)}{\partial \alpha}$, and rewrite the above as

$$\sqrt{n} \frac{1}{n} \sum \mathbf{g}_i(\alpha_0) - \mathbf{G} \sqrt{n}(\hat{\alpha} - \alpha_0) \stackrel{a}{=} \mathbf{0},$$

or,

$$\sqrt{n}(\hat{\alpha} - \alpha_0) \stackrel{a}{=} \mathbf{G}^{-1} \sqrt{n} \frac{1}{n} \sum \mathbf{g}_i(\alpha_0).$$

Applying the same techniques to $n^{-1} \sum h_i(\hat{\beta}, \hat{\alpha})$, we obtain

$$\sqrt{n} \frac{1}{n} \sum h_i(\hat{\beta}, \hat{\alpha}) = \sqrt{n} \frac{1}{n} \sum h_i(\beta_0, \alpha_0) + \frac{1}{n} \sum \frac{\partial h_i}{\partial \alpha} \sqrt{n}(\hat{\alpha} - \alpha_0) + \frac{1}{n} \sum \frac{\partial h_i}{\partial \beta} \sqrt{n}(\hat{\beta} - \beta_0) = \mathbf{0},$$

and defining the $l \times k$ matrix $\mathbf{H}_\alpha = -\operatorname{plim} \frac{1}{n} \sum \frac{\partial h_i}{\partial \alpha}$ and the $l \times l$ matrix $\mathbf{H}_\beta = -\operatorname{plim} n^{-1} \sum \frac{\partial h_i}{\partial \beta}$, we arrive at

$$\mathbf{0} \stackrel{a}{=} \sqrt{n} \frac{1}{n} \sum h_i(\beta_0, \alpha_0) - \mathbf{H}_\alpha \sqrt{n}(\hat{\alpha} - \alpha_0) - \mathbf{H}_\beta \sqrt{n}(\hat{\beta} - \beta_0).$$

Then substituting $\mathbf{G}^{-1} \sqrt{n} \frac{1}{n} \sum \mathbf{g}_i(\alpha_0)$ for $\sqrt{n}(\hat{\alpha} - \alpha_0)$, we obtain

$$\mathbf{H}_\beta \sqrt{n}(\hat{\beta} - \beta_0) \stackrel{a}{=} \sqrt{n} \frac{1}{n} \sum h_i(\beta_0, \alpha_0) - \mathbf{H}_\alpha \mathbf{G}^{-1} \sqrt{n} \frac{1}{n} \sum \mathbf{g}_i(\alpha_0). \tag{28.12}$$

The vectors $\sqrt{n} \frac{1}{n} \sum h_i(\beta_0, \alpha_0)$ and $\sqrt{n} \frac{1}{n} \sum \mathbf{g}_i(\alpha_0)$ will, by a central limit theorem, converge in distribution to multivariate $\mathcal{N}(\mathbf{0}, \mathbf{V})$, with

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{hh} & \mathbf{V}_{hg} \\ \mathbf{V}_{gh} & \mathbf{V}_{gg} \end{bmatrix}.$$

The right-hand side of (28.12) will converge in distribution to

$$\mathcal{N}(\mathbf{0}, \mathbf{V}_{hh}) - \mathbf{H}_\alpha \mathbf{G}^{-1} \cdot \mathcal{N}(\mathbf{0}, \mathbf{V}_{gg}),$$

that is, to a linear combination of normally-distributed vectors. The overall variance matrix for the right-hand side is therefore

$$\tilde{\mathbf{V}} = \mathbf{V}_{hh} + \mathbf{H}_\alpha \mathbf{G}^{-1} \mathbf{V}_{gg} (\mathbf{G}^{-1})' \mathbf{H}_\alpha' - \mathbf{H}_\alpha \mathbf{G}^{-1} \mathbf{V}_{gh} - \mathbf{V}_{hg} (\mathbf{G}^{-1})' \mathbf{H}_\alpha'.$$

Returning to equation (28.12) and multiplying through by $\mathbf{H}_\beta^{-1}$, we find that $\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \Omega)$ with the limiting variance matrix being

$$\Omega = \mathbf{H}_\beta^{-1} \tilde{\mathbf{V}} (\mathbf{H}_\beta^{-1})'. \tag{28.13}$$

Despite the truly forbidding appearance of Ω, all of its elements can be estimated in a straightforward way in the iid case. We have

$$\hat{\mathbf{H}}_\alpha = -\frac{1}{n} \sum \frac{\partial h_i(\hat{\beta}, \hat{\alpha})}{\partial \alpha}$$

$$\hat{\mathbf{H}}_\beta = -\frac{1}{n} \sum \frac{\partial h_i(\hat{\beta}, \hat{\alpha})}{\partial \beta}$$

$$\hat{\mathbf{G}} = -\frac{1}{n} \sum \frac{\partial \mathbf{g}_i(\hat{\alpha})}{\partial \alpha}$$

$$\hat{\mathbf{V}}_{hh} = \frac{1}{n} \sum h_i(\hat{\beta}, \hat{\alpha}) h_i(\hat{\beta}, \hat{\alpha})'$$

$$\hat{\mathbf{V}}_{gh} = \frac{1}{n} \sum \mathbf{g}_i(\hat{\beta}, \hat{\alpha}) h_i(\hat{\beta}, \hat{\alpha})'$$

and

$$\hat{\mathbf{V}}_{gg} = \frac{1}{n} \sum \mathbf{g}_i(\hat{\alpha}) \mathbf{g}_i(\hat{\alpha})'.$$

Application to ML Models: Special case

Further simplifications in the limiting covariance matrix of $\sqrt{n}(\hat{\beta} - \beta_0)$ can be achieved when the moment vectors g and h equal the scores of the marginal and conditional likelihood. That is, we let

$$\mathbf{g}_i(\alpha) = \frac{\partial \log f_i^m}{\partial \alpha} \text{ and } h_i(\beta, \alpha) = \frac{\partial \log f_i^c}{\partial \beta}.$$

Then with

$$\mathbf{G}_i \equiv -\frac{\partial \mathbf{g}_i}{\partial \alpha} = -\frac{\partial^2 \log f_i^m}{\partial \alpha \partial \alpha'}$$

and $\operatorname{plim} n^{-1} \sum \mathbf{G}_i = \mathbf{G} = \mathcal{J}^m_{\alpha\alpha}$, the normalized information matrix for the marginal model. Also,

$$\mathbf{H}_{\alpha,i} = -\frac{\partial h_i}{\partial \alpha} = -\frac{\partial^2 \log f_i^c}{\partial \beta \partial \alpha}$$

and $\operatorname{plim} n^{-1} \sum \mathbf{H}_{\alpha,i} \equiv \mathbf{H}_\alpha = \mathcal{J}^c_{\beta\alpha}$, an off-diagonal block of the normalized information matrix for the conditional model. Finally,

$$\mathbf{H}_{\beta,i} = -\frac{\partial h_i}{\partial \beta} = -\frac{\partial^2 \log f_i^c}{\partial \beta \partial \beta'}$$

and $\operatorname{plim} n^{-1} \sum \mathbf{H}_{\beta,i} \equiv \mathbf{H}_\beta = \mathcal{J}^c_{\beta\beta}$. Furthermore, note that

$$\mathbf{V}_{gg} = \mathrm{E}\, \mathbf{g}_i \mathbf{g}_i' = \mathrm{E} \left(\frac{\partial \log f_i^m}{\partial \alpha} \right) \left(\frac{\partial \log f_i^m}{\partial \alpha} \right)' = \mathcal{J}^m_{\alpha\alpha},$$

and by a similar argument $\mathbf{V}_{hh} = \mathcal{J}^c_{\beta\beta}$.

We next consider $\mathbf{V}_{gh}$ and show that this is a matrix of zeroes. To see this, consider the full log-likelihood contribution of observation i, that is, $L_i = \log f_i^c(\alpha, \beta) + \log f_i^m(\alpha)$, and recall that in the full model,

$$\begin{aligned}
\mathcal{J}_{\beta\alpha} &= \operatorname{plim} \frac{1}{n} \sum \left(\frac{\partial L_i}{\partial \beta} \right) \left(\frac{\partial L_i}{\partial \alpha} \right)' \\
&= \operatorname{plim} \frac{1}{n} \sum \frac{\partial \log f_i^c}{\partial \beta} \left(\frac{\partial \log f_i^c}{\partial \alpha} + \frac{\partial \log f_i^m}{\partial \alpha} \right)' \\
&= \operatorname{plim} \frac{1}{n} \sum \frac{\partial \log f_i^c}{\partial \beta} \left(\frac{\partial \log f_i^c}{\partial \alpha} \right)' + \operatorname{plim} \frac{1}{n} \sum \left(\frac{\partial \log f_i^c}{\partial \beta} \right) \left(\frac{\partial \log f_i^m}{\partial \alpha} \right)'.
\end{aligned}$$

But we also know that

$$\mathcal{J}_{\beta\alpha} = -\operatorname{plim} \frac{1}{n} \sum \frac{\partial^2 L_i}{\partial \beta \partial \alpha} = -\operatorname{plim} \frac{1}{n} \sum \frac{\partial^2 \log f_i^c}{\partial \beta \partial \alpha},$$

and that

$$-\operatorname{plim} \frac{1}{n} \sum \frac{\partial^2 \log f_i^c}{\partial \beta \partial \alpha} = \operatorname{plim} \frac{1}{n} \sum \frac{\partial \log f_i^c}{\partial \beta} \left(\frac{\partial \log f_i^c}{\partial \alpha} \right)'.$$

This implies

$$\operatorname{plim} \frac{1}{n} \sum \left(\frac{\partial \log f_i^c}{\partial \beta} \right) \left(\frac{\partial \log f_i^m}{\partial \alpha} \right)' = \mathbf{0}.$$

We conclude that

$$\begin{aligned}
\operatorname{plim} \hat{\mathbf{V}}_{hg} &= \operatorname{plim} \frac{1}{n} \sum h_i(\hat{\alpha}, \hat{\beta}) \mathbf{g}_i(\hat{\alpha})' \\
&= \operatorname{plim} \frac{1}{n} \sum \left(\frac{\partial \log f_i^c}{\partial \beta} \right) \left(\frac{\partial \log f_i^m}{\partial \alpha} \right)' = \mathbf{0}.
\end{aligned}$$

Making use of all these results, we arrive at a simpler representation of Ω, the limiting covariance matrix of the two-step estimator of β. With $\mathbf{V}_{gh} = \mathbf{0}$, it is

$$\Omega = \mathbf{H}_\beta^{-1} \left(\mathbf{V}_{hh} + \mathbf{H}_\alpha \mathbf{G}^{-1} \mathbf{V}_{gg} (\mathbf{G}^{-1})' \mathbf{H}_\alpha' \right) (\mathbf{H}_\beta^{-1})' \tag{28.14}$$

$$= (\mathcal{J}_{\beta\beta}^c)^{-1} \left(\mathcal{J}_{\beta\beta}^c + \mathcal{J}_{\beta\alpha}^c (\mathcal{J}_{\alpha\alpha}^m)^{-1} \mathcal{J}_{\alpha\alpha}^m (\mathcal{J}_{\alpha\alpha}^m)^{-1} \mathcal{J}_{\alpha\beta}^c \right) (\mathcal{J}_{\beta\beta}^c)^{-1} \tag{28.15}$$

$$= (\mathcal{J}_{\beta\beta}^c)^{-1} + (\mathcal{J}_{\beta\beta}^c)^{-1} \mathcal{J}_{\beta\alpha}^c (\mathcal{J}_{\alpha\alpha}^m)^{-1} \mathcal{J}_{\alpha\beta}^c (\mathcal{J}_{\beta\beta}^c)^{-1} \tag{28.16}$$

28.6 Computation

A Gauss–Newton iteration method for finding the GMM estimator proceeds as follows (Hayashi 2000, pp. 497–498). Write the essentials of the GMM minimization problem in the form

$$\min_\theta S(\theta) = \mathbf{g}_n(\theta)' \mathbf{W}_n \mathbf{g}_n(\theta)$$

with $\mathbf{g}_n(\theta) = (1/n) \sum \mathbf{g}_i(\theta)$. Given an initial estimate θ_1, use the Taylor-series approximation

$$\mathbf{g}_n(\theta_2) \approx \mathbf{g}_n(\theta_1) + \mathbf{G}_n(\theta_1)(\theta_2 - \theta_1) = \mathbf{g}_n(\theta_1) - \mathbf{G}_n(\theta_1) \cdot \theta_1 - (-\mathbf{G}_n(\theta_1)\theta_2)$$

in which $\mathbf{G}_n$ is an $m \times k$ matrix of derivatives. This can be written

$$\mathbf{g}_n(\theta_2) = \mathbf{Z} - \mathbf{X}\theta_2$$

with $\mathbf{Z} \equiv \mathbf{g}_n(\theta_1) - \mathbf{G}_n(\theta_1) \cdot \theta_1$ and $\mathbf{X} \equiv -\mathbf{G}_n(\theta_1)$. With these simplifications, the minimization problem is cast into a familiar form,

$$\min_\theta S(\theta_2) = (\mathbf{Z} - \mathbf{X}\theta_2)' \mathbf{W}_n (\mathbf{Z} - \mathbf{X}\theta_2)$$

implying that

$$\theta_2 = -\left(\mathbf{G}_n(\theta_1)' \mathbf{W}_n \mathbf{G}_n(\theta_1) \right)^{-1} \mathbf{G}_n(\theta_1) \mathbf{W}_n \left(\mathbf{g}_n(\theta_1) - \mathbf{G}_n(\theta_1)\theta_1 \right)$$

or

$$\theta_2 = \theta_1 - \left(\mathbf{G}_n(\theta_1)' \mathbf{W}_n \mathbf{G}_n(\theta_1) \right)^{-1} \mathbf{G}_n(\theta_1) \mathbf{W}_n \mathbf{g}_n(\theta_1)$$

Note that $\mathbf{G}_n(\theta_1) \mathbf{W}_n \mathbf{g}_n(\theta_1)$ is the $k \times 1$ vector of derivatives evaluated at θ_1, and it is premultiplied by a positive definite matrix.

Chapter 29

Two GMM-Based Tests

29.1 Tests of Over-Identifying Restrictions

Suppose that we use efficient GMM methods to estimate the θ vector and that we have more moment conditions than we need. The additional $m-k$ conditions can be easily tested, with the null hypothesis being that each of the additional conditions has a zero expectation. The approach outlined below is the GMM counterpart to the Sargan test.

The minimized sum of squares that is produced by the feasible efficient GMM estimator, using the weighting matrix $W_n = \hat{\Delta}^{-1}$, is

$$S_2(\tilde{\theta}) = \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta})\right)' \hat{\Delta}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta})\right). \tag{29.1}$$

Assume that n is large enough that $\hat{\Delta}^{-1} \approx \Delta^{-1}$ and rewrite the sum of squares as

$$S_2(\tilde{\theta}) \stackrel{a}{=} \frac{1}{2}\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta})\right)' \Delta^{-1/2}\cdot\Delta^{-1/2}\left(\frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta})\right).$$

Then multiply each of the vectors in the wings of the quadratic form by $\sqrt{n}$. The result can be written as the dot product

$$2nS_2(\tilde{\theta}) \stackrel{a}{=} \left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta})\right)' \Delta^{-1/2}\cdot\Delta^{-1/2}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta})\right). \tag{29.2}$$

In what follows, we will show that $2nS_2(\tilde{\theta}) \xrightarrow{d} \chi^2_{m-k}$, where $m-k$ is the number of over-identifying restrictions.

Recall from Chapter 28 that

$$\sqrt{n}(\tilde{\theta}-\theta_0) \stackrel{a}{=} -(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\left(\sqrt{n}\frac{1}{n}\sum g_i(\theta_0)\right),$$

and, since $\sqrt{n}\frac{1}{n}\sum g_i(\theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0},\Delta)$, we can insert $\Delta^{1/2}\cdot\Delta^{-1/2}$ in the above, such that

$$\begin{aligned}\sqrt{n}(\tilde{\theta}-\theta_0) &\stackrel{a}{=} -(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1}\Delta^{1/2}\Delta^{-1/2}\left(\sqrt{n}\frac{1}{n}\sum g_i(\theta_0)\right)\\ &= -(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1/2}\cdot\mathbf{U}_n.\end{aligned}$$

with the $m \times 1$ vector $\mathbf{U}_n = \Delta^{-1/2}\sqrt{n}\frac{1}{n}\sum g_i(\theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$. With these elements in hand, consider the vector $\Delta^{-1/2}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta})\right)$ of equation (29.2). Expanding $g_i(\tilde{\theta})$ in a Taylor series and making the appropriate asymptotic substitutions, we have

$$\begin{aligned}
\Delta^{-1/2}\left(\sqrt{n}\frac{1}{n}\sum_{i=1}^{n} g_i(\tilde{\theta})\right) &= \Delta^{-1/2}(\sqrt{n}\frac{1}{n}\sum g_i(\theta_0)) \\
&\quad + \Delta^{-1/2}\left(\frac{1}{n}\sum G_i(\theta^*)\right)\cdot\left(\sqrt{n}(\tilde{\theta}-\theta_0)\right) \\
&= \mathbf{U}_n - \Delta^{-1/2}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1/2}\cdot\mathbf{U}_n \\
&= \left(\mathbf{I} - \Delta^{-1/2}\Gamma(\Gamma'\Delta^{-1}\Gamma)^{-1}\Gamma'\Delta^{-1/2}\right)\cdot\mathbf{U}_n \\
&= \mathbf{M}\cdot\mathcal{N}(\mathbf{0},\mathbf{I}),
\end{aligned}$$

with $\mathbf{M}$ a symmetric idempotent matrix of rank $m-k$, which is the number of over-identifying restrictions. Since

$$2nS_2(\tilde{\theta}) \stackrel{a}{=} \mathcal{N}(\mathbf{0},\mathbf{I})'\,\mathbf{M}\,\mathcal{N}(\mathbf{0},\mathbf{I}),$$

we arrive at the result that we set out to establish, that $2nS_2(\tilde{\theta}) \xrightarrow{d} \chi^2_{m-k}$. Of course, this result holds only under the null hypothesis that all m moment conditions are valid.

29.2 Conditional Moments Tests

In this section, we will be thinking of a situation in which θ is initially estimated by maximum likelihood and an additional set of moment functions is formulated in order to test aspects of the specification. The conditional moments approach has much in common with the Lagrange Multiplier approach—it can be viewed as a natural extension and generalization. Pagan and Vella (1989) provide a splendid discussion of the conditional moments approach, putting emphasis on applications to qualitative and limited dependent variables models.

As before, let $\mathbf{g}(Y_i, \mathbf{X}_i, \mathbf{Z}_i, \theta)$ be an m-vector of functions. Under the null hypothesis, when they are evaluated at the true θ_0, each of these functions has expectation zero. Under the alternative hypothesis, however, the $\mathbf{g}_i$ functions have non-zero expectations. When the sample counterparts to these moments depart from zero, this provides evidence against the null hypothesis.

The matrix $\mathbf{G}_i$, of dimension $m \times k$, contains the derivatives of the functions with respect to θ. It has typical element $\partial \mathbf{g}_i^j / \partial \theta_l$, this being the derivative of the j-th function with respect to the l-th element of the θ vector. In what follows, we let $\hat{\theta}$ denote the maximum likelihood estimate.

The generalized information matrix equality

We begin with a useful digression. Let $f(y, x, z, \theta)$ be the true density of the data.[1] Then the expectations condition for a given function j can be written in the form

$$\mathrm{E}\, g^j(Y, X, Z, \theta) = \int g^j(y, x, z, \theta) \cdot f(y, x, z, \theta)\, dy\, dx\, dz = 0.$$

Differentiate this with respect to θ_l, assuming that the regularity conditions allowing differentiation under the integral sign are satisfied. We obtain, for each j,

$$-\int \frac{\partial g^j(\theta)}{\partial \theta_l} f(\theta)\, dy\, dx\, dz = \int g^j(\theta) \frac{\partial \ln f(\theta)}{\partial \theta_l} f(\theta)\, dy\, dx\, dz.$$

Arranged in matrix form for all j and l—recall that $\mathbf{G}$ is defined as having dimensions $m \times k$—this becomes

$$-\mathrm{E}\, \mathbf{G}(\theta) = \mathrm{E}\, \mathbf{g}(\theta) \left(\frac{\partial \ln f(\theta)}{\partial \theta} \right)'. \tag{29.3}$$

Equation (29.3) is known as the generalized information matrix equality, because it was derived by methods very similar to those used to derive the information matrix.

Recall that in Chapter 28, we denoted the expected value of $\mathbf{G}$ by Γ. In this notation, equation (29.3) shows that the matrix $-\Gamma$ equals the covariance of the $m \times 1$ moment vector $\mathbf{g}$ and the $k \times 1$ score vector $\partial \ln f / \partial \theta$. In this way, we have come upon a second method for estimating Γ, via the sample analog

$$\hat{\Gamma} = -\frac{1}{n} \sum_{i=1}^{n} \mathbf{g}(y_i, x_i, z_i, \hat{\theta}) \left(\frac{\partial \ln f(y_i, x_i, z_i, \hat{\theta})}{\partial \theta} \right)'. \tag{29.4}$$

Maximum likelihood

Denote by d_i the i-th observation's contribution to the full score, assuming a correct specification. That is,

$$d_i(\hat{\theta}) \equiv \frac{\partial \ln f(y_i, x_i, z_i, \hat{\theta})}{\partial \theta}.$$

As we proceed, we will be using some familiar results from maximum likelihood estimation, each of these being a result that holds under the null hypothesis of correct specification. The results are, first, that

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathcal{J}^{-1}), \tag{29.5}$$

where θ_0 is the true value of θ and $\mathcal{J}$ is the probability limit of the normalized information matrix,

$$\mathcal{J} = \operatorname{plim} \frac{1}{n} \sum_{i}^{n} d_i(\theta_0) d_i(\theta_0)'. \tag{29.6}$$

[1] We will not always need to include the instruments z in this density; an alternative is to proceed by conditioning on them. The main complication entailed by conditioning is that we can no longer think of y and x as being independent draws from a given distribution; rather, we must view the y_i, x_i sequence as being possibly independent over i but certainly not identically distributed, given z. We will ignore such complications in what follows.

Recall that $\hat{\mathcal{J}} = \frac{1}{n}\sum_i d_i(\hat{\theta})d_i(\hat{\theta})'$ converges to $\mathcal{J}$. Second,

$$\sqrt{n}(\hat{\theta} - \theta_0) \stackrel{a}{=} \mathcal{J}^{-1}\sqrt{n}\frac{1}{n}\sum_i^n d_i(\theta_0). \tag{29.7}$$

Equation (29.7) also holds with the d_i evaluated at the maximum likelihood estimate $\hat{\theta}$. Last, we have

$$\sqrt{n}\frac{1}{n}\sum_i^n d_i(\theta_0) \stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \mathcal{J}). \tag{29.8}$$

and this, too, is valid if the d_i are evaluated at the maximum likelihood estimate.

The conditional moments tests

If the null hypothesis is correct, the sample moments evaluated at the maximum likelihood estimate $\hat{\theta}$ should be approximately zero. That is, with

$$\hat{\tau} \equiv \frac{1}{n}\sum_i^n g_i(\hat{\theta}),$$

we expect $\hat{\tau}$ to be approximately a zero vector under the null.

Let us form a test statistic using $\sqrt{n}\hat{\tau}$. As we did before, Taylor-expand the sample moments around the true θ_0 and multiply by $\sqrt{n}$. This yields

$$\sqrt{n}\hat{\tau} = \sqrt{n}\frac{1}{n}\sum_{i=1}^n g_i(\hat{\theta}) = \sqrt{n}\frac{1}{n}\sum_{i=1}^n g_i(\theta_0) + \frac{1}{n}\sum_{i=1}^n G_i(\theta^*)\cdot\sqrt{n}(\hat{\theta} - \theta_0).$$

Use equation (29.7) to substitute for $\sqrt{n}(\hat{\theta} - \theta_0)$ and recall that $\frac{1}{n}\sum_{i=1}^n G_i(\theta^*)$ converges to Γ. Now, arrange the results in matrix form,

$$\sqrt{n}\hat{\tau} = \left[\begin{array}{cc} \mathbf{I} & \Gamma\cdot\mathcal{J}^{-1} \end{array}\right]\left[\begin{array}{c} \sqrt{n}\frac{1}{n}\sum_{i=1}^n g_i(\theta_0) \\ \sqrt{n}\frac{1}{n}\sum_i^n d_i(\theta_0) \end{array}\right]. \tag{29.9}$$

This can be written more compactly as $\sqrt{n}\hat{\tau} = \mathbf{Ah}$, with the dimensions of the $\mathbf{A}$ matrix being $m \times (m+k)$ and the dimensions of the $\mathbf{h}$ vector being $(m+k)\times 1$.

Now, $\mathbf{h} \stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{V})$, and in a moment we will present an estimator for the $\mathbf{V}$ matrix. Given asymptotic normality for $\mathbf{h}$, we have $\sqrt{n}\hat{\tau} \stackrel{d}{\to} \mathcal{N}(\mathbf{0}, \mathbf{AVA}')$. This result yields a χ^2 test statistic if $\mathbf{A}$ and $\mathbf{V}$ can be estimated. For the unknown Γ and $\mathcal{J}^{-1}$ elements in $\mathbf{A}$, we can substitute $\hat{\Gamma}$ and $\hat{\mathcal{J}}^{-1}$ respectively. (We have two expressions available for $\hat{\Gamma}$ and likewise for $\hat{\mathcal{J}}$.) As for $\hat{\mathbf{V}}$, we can use the estimate

$$\hat{\mathbf{V}} = \left[\begin{array}{cc} \frac{1}{n}\sum_i g_i g_i' & \frac{1}{n}\sum_i g_i d_i' \\ \frac{1}{n}\sum_i g_i' d_i & \frac{1}{n}\sum_i d_i d_i' \end{array}\right],$$

for iid or inid data series, with each element being evaluated at $\hat{\theta}$. Then the test statistic T is simply

$$T = \sqrt{n}\hat{\tau}'(\hat{\mathbf{A}}\hat{\mathbf{V}}\hat{\mathbf{A}}')^{-1}\sqrt{n}\hat{\tau},$$

and this converges to a central χ^2_m under the null.

Pagan and Vella (1989, S33–S34) and Davidson and MacKinnon (1993, pp. 571–578) show how to further simplify the calculation of the test statistic to the point that it can be calculated with an auxiliary regression. The simplification makes use of the $\hat{\Gamma}$ estimator of equation (29.4). It seems, however, that this can be an unusually noisy estimator of Γ. Monte Carlo evidence cited in Skeels and Vella (1994a), Skeels and Vella (1994b), and Skeels and Vella (1995) suggests that the simplified form of the test statistic should probably be avoided. Indeed, their Monte Carlo evidence—focused on Tobit and probit models—indicates that the small sample properties of the conditional moments tests are not very well approximated by the asymptotic properties, even if the preferred form of the conditional moments tests is used. Hence, some caution is in order when one applies these tests.

Chapter 30

The Hausman Test

Students: Read the first section carefully—be sure to refresh your understanding of the material in Chapter 13—and skim through the second section. Supplement this with Cameron and Trivedi (2005, Sections 8.3, 8.4).

Although it was controversial when first proposed, the Hausman (1978) test has proven to be a remarkably useful tool for specification analysis. The idea of the test (due originally to Durbin, whose work preceded Hausman's by some twenty years) is to assess the statistical significance of a particular vector of contrasts. On the one hand, we have $\hat{\theta}_e$, an estimator that is efficient under the null hypothesis but inconsistent under the alternative hypothesis. On the other hand, we have $\hat{\theta}_c$, another estimator that is consistent under the null (although inefficient) but which maintains its consistency under the alternative hypothesis. The Hausman test is based on the asymptotic behavior of the contrast between these estimators, i.e., on the vector $\sqrt{n}(\hat{\theta}_c - \hat{\theta}_e)$.

For this test to be applied, it must be the case that both $\sqrt{n}(\hat{\theta}_e - \theta)$ and $\sqrt{n}(\hat{\theta}_c - \theta)$ converge under the null to multivariate normal distributions with covariance matrices $\mathbf{V}_e$ and $\mathbf{V}_c$ respectively. The key to formulating the test statistic is to see that under the null, the variance matrix of the contrast vector is simply $\mathbf{V}_c - \mathbf{V}_e$. This very important result was proven for the small-sample case in Chapter 13, and the large-sample argument is essentially the same. Taking the result as given, it is immediate that

$$\sqrt{n}(\hat{\theta}_c - \hat{\theta}_e)' \left[\mathbf{V}_c - \mathbf{V}_e\right]^- \sqrt{n}(\hat{\theta}_c - \hat{\theta}_e) \xrightarrow{d} \chi^2_k$$

in which the inverse of $\mathbf{V}_c - \mathbf{V}_e$ may be a generalized inverse and the degrees of freedom k are typically equal to, but are sometimes less than, the dimension of θ. The reason for these caveats will be made clear below in the context of one of the most important applications, testing the exogeneity of the explanatory variables.

30.1 Examples

The Hausman test is especially useful when $\hat{\theta}_e$ is the maximum likelihood estimator and $\hat{\theta}_c$ is another estimator that is consistent under less restrictive conditions than the ML estimator.

It has many additional uses, however, and in the next subsections, we will review a few of the most important.

Exogeneity tests

Consider a regression model $\mathbf{Y} = \mathbf{X}\beta + \epsilon$ in which one or more of the $\mathbf{X}$ covariates may be correlated with ϵ. We can formulate a Hausman test for the presence of such correlation, with the "null hypothesis" being that no $\mathbf{X}$ variable is correlated with the disturbance. More precisely, when we have a model with $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$, we would assume that the $\mathbf{X}_1$ variables are known not to be correlated with ϵ, whereas the $\mathbf{X}_2$ variables (of which there are k_2) might be correlated. In this case, assuming that the disturbances are homoskedastic, the least squares estimator is consistent and efficient under the null hypothesis, and the instrumental variables estimator of β is consistent under both the null and the alternative, but is less efficient than OLS under the null. That is, the contrast vector $\hat{\beta}_{IV} - \hat{\beta}_{OLS}$ has a probability limit of zero under the null hypothesis but something other than zero under the alternative. The contrast should therefore be informative about the presence of correlation.

The asymptotic variance of the OLS estimator is $\sigma^2 \operatorname{plim}(n^{-1}\mathbf{X}'\mathbf{X})^{-1}$ under the null, whereas for the IV estimator it is $\sigma^2 \operatorname{plim}(n^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1}$ with $\mathbf{Z}$ being the matrix of instruments. In this case, then,

$$\mathbf{V}_c - \mathbf{V}_e \stackrel{a}{=} \sigma^2 \left[(n^{-1}\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X})^{-1} - (n^{-1}\mathbf{X}'\mathbf{X})^{-1} \right],$$

a difference that is positive semidefinite under the null.[1] The matrices in this expression can usually be retrieved from the statistical package used to estimate the OLS and IV regressions.

What are the appropriate degrees of freedom for this test? With $\mathbf{X}_1$ assumed to be exogenous, we have k_2 variables $\mathbf{X}_2$ whose exogeneity is in doubt. The number of degrees of freedom for the Hausman test is k_2. Indeed, the proper way to form the $\chi^2_{k_2}$ test is to use only the contrast between the IV and the OLS estimates of β_2 in the wings of the quadratic form, and in the matrix $\mathbf{V}_c - \mathbf{V}_e$, to use only the $k_2 \times k_2$ sub-matrices that correspond to β_2.

This argument may well be confusing at first encounter. We will give a more formal rationale in the next section, but in what immediately follows we will attempt an informal justification. This informal presentation shows that the test itself can be derived from a simple diagnostic regression; you don't need to assemble the quadratic form.

Davidson and MacKinnon (2004, pp. 338–40) clarify an argument originally put forward by Hausman (1978), by beginning with $\hat{\beta}_{IV} - \hat{\beta} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{Y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$. The difference in these estimators can be written as

$$(\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\left(\mathbf{X}'\mathbf{P}_Z - (\mathbf{X}'\mathbf{P}_Z\mathbf{X})(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{Y} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\left(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)\mathbf{Y}$$

giving

$$\hat{\beta}_{IV} - \hat{\beta} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{M}_X\mathbf{Y}.$$

[1] The result follows from the fact that $(\mathbf{X}'\mathbf{X}) - (\mathbf{X}'\mathbf{P}_\mathbf{Z}\mathbf{X}) = \mathbf{X}'\mathbf{M}_\mathbf{Z}\mathbf{X}$ is positive semidefinite. The semidefinite aspect comes from the fact that $\mathbf{M}_\mathbf{Z} \cdot \begin{bmatrix}\mathbf{X}_1 & \mathbf{X}_2\end{bmatrix} = \begin{bmatrix}\mathbf{0} & \mathbf{M}_\mathbf{Z}\mathbf{X}_2\end{bmatrix}$ given that the $\mathbf{X}_1$ variables serve as their own instruments.

Since $\mathbf{P}_Z\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 & \hat{\mathbf{X}}_2\end{bmatrix}$ and $\mathbf{M}_X\mathbf{Y}$ is a vector of OLS residuals from a regression of $\mathbf{Y}$ on $\mathbf{X}_1$ and $\mathbf{X}_2$, this becomes

$$\hat{\beta}_{IV} - \hat{\beta} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\begin{bmatrix}\mathbf{X}_1' \\ \hat{\mathbf{X}}_2'\end{bmatrix}\mathbf{M}_X\mathbf{Y} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\begin{bmatrix}\mathbf{0} \\ \hat{\mathbf{X}}_2'\mathbf{M}_X\mathbf{Y}\end{bmatrix}.$$

The key to the expression—the essence of the difference in the two estimators—is the $\hat{\mathbf{X}}_2'\mathbf{M}_X\mathbf{Y}$ term. A test of whether the difference in the estimators is significant rests on this term.

Consider a simple diagnostic regression of $\mathbf{Y}$ on $\mathbf{X}_1, \mathbf{X}_2$ and $\hat{\mathbf{X}}_2$, and focus on the $\hat{d}$ coefficient attached to $\hat{\mathbf{X}}_2$. By the FWL theorem, this diagnostic coefficient is

$$\hat{d} = \left(\hat{\mathbf{X}}_2'\mathbf{M}_X\hat{\mathbf{X}}_2\right)^{-1}\hat{\mathbf{X}}_2'\mathbf{M}_X\mathbf{Y}$$

and you can test the significance of the difference between the IV and OLS estimators by running the diagnostic regression and testing whether $\hat{d}$ is significantly different from zero. The number of degrees of freedom to use for the test is obvious: it is clearly k_2, the number of $\mathbf{X}_2$ variables that are potentially correlated with the disturbance term. So, although the contrast vector $\hat{\beta}_{IV} - \hat{\beta}$ is of length $k > k_2$, the test rightly focuses on the k_2 logical propositions of the null hypothesis.

Another interesting approach, which also leads to a simple diagnostic regression, is due to Nakamura and Nakamura (1981). Write the model in a form that distinguishes between the exogenous $\mathbf{X}_1$ and the potentially endogenous $\mathbf{X}_2$. To keep things simple, suppose that there is only one $\mathbf{X}_2$ variable, and write the structural model as

$$Y_i = \mathbf{X}_{i,1}'\beta_1 + \beta_2 X_{i,2} + \epsilon_i.$$

Think of the first stage of a 2SLS procedure in terms of a linear equation relating $X_{i,2}$ to the instruments $\mathbf{Z}_i$,

$$X_{i,2} = \mathbf{Z}_i'\pi + u_i$$

Clearly $X_{i,2}$ can only be correlated with ϵ_i of the structural model if u_i is correlated with ϵ_i, as we see if we consider

$$\operatorname{plim}\frac{1}{n}\sum_i X_{i,2}'\epsilon_i = \pi' \operatorname{plim}\frac{1}{n}\sum_i \mathbf{Z}_i\epsilon_i + \operatorname{plim}\frac{1}{n}\sum_i u_i\epsilon_i,$$

in which the first of the probability limits must be zero given valid instruments.

Let $\hat{\mathbf{u}}$ be the residual vector from the first stage projection of $\mathbf{X}_2$ on $\mathbf{Z}$,

$$\mathbf{X}_2 = \mathbf{Z}\hat{\pi} + \hat{\mathbf{u}}$$

and take note of the facts that $\hat{\mathbf{u}} = \mathbf{M}_\mathbf{Z}\mathbf{u}$ is orthogonal to each column of $\mathbf{Z}$ and to the linear combination of these columns, $\mathbf{Z}\hat{\pi}$, as well.

Letting $\hat{\mathbf{X}}_2 = \mathbf{Z}\hat{\pi}$ be the projection of $\mathbf{X}_2$ onto the instruments, which include $\mathbf{X}_1$, we can rewrite the structural equation as

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \beta_2\hat{\mathbf{X}}_2 + \beta_2\hat{\mathbf{u}} + \epsilon.$$

When it is expressed in this way, we can think of the structural equation as having the form of a diagnostic regression. Why? As we just observed, $\hat{\mathbf{u}}$ is orthogonal to the columns of $\mathbf{X}_1$ and to $\hat{\mathbf{X}}_2$. If we relabel the coefficients as

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \beta_2\hat{\mathbf{X}}_2 + b_2\hat{\mathbf{u}} + \epsilon,$$

and apply ordinary least squares to this diagnostic equation, the estimates of β_1 and β_2 will be numerically identical to 2SLS estimates and thus consistent for their respective parameters. Evidently, the diagnostic information we seek from the regression must somehow be located in the $\hat{b}_2$ estimate.

Because $\hat{\mathbf{u}}$ is orthogonal to the other explanatory variables, the large-sample behavior of $\hat{b}_2$ depends on whether $\hat{\mathbf{u}}$ is correlated with the ϵ disturbance. Under the null hypothesis, $\hat{\mathbf{u}}$ and ϵ are uncorrelated. To fully appreciate this point, consider

$$\operatorname{plim}\frac{1}{n}\hat{\mathbf{u}}'\epsilon = \operatorname{plim}\frac{1}{n}\mathbf{u}'\mathbf{M}_{\mathbf{Z}}\epsilon,$$

which equals

$$\operatorname{plim}\frac{1}{n}\mathbf{u}'\epsilon - \operatorname{plim}\frac{1}{n}\mathbf{u}'\mathbf{Z}\operatorname{plim}\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\operatorname{plim}\frac{1}{n}\mathbf{Z}'\epsilon.$$

With $\operatorname{plim}\frac{1}{n}\mathbf{Z}'\epsilon = \mathbf{0}$ by the assumption that the instruments are valid, and $\operatorname{plim}\frac{1}{n}\mathbf{Z}'\mathbf{Z} = \mathbf{W}_{\mathbf{ZZ}}$, the probability limit equals $\operatorname{plim}\frac{1}{n}\mathbf{u}'\epsilon$ even if $\operatorname{plim}\frac{1}{n}\mathbf{u}'\mathbf{Z} \neq \mathbf{0}$ so long as this probability limit exists. Hence, under the null hypothesis, $\operatorname{plim}\frac{1}{n}\hat{\mathbf{u}}'\epsilon = 0$ whereas under the alternative hypothesis $\operatorname{plim}\frac{1}{n}\hat{\mathbf{u}}'\epsilon \neq 0$ and $\hat{\mathbf{u}}$ is correlated with ϵ.

The implications for testing exogeneity as are follows. Under the null, the estimators $\hat{\beta}_2$ and $\hat{b}_2$ both converge in probability to β_2. Under the alternative, however, $\hat{\beta}_2 \xrightarrow{p} \beta_2$ but $\hat{b}_2$ does not. Hence, one easy way to test for the exogeneity of $\mathbf{X}_2$ is to test the restriction that $\beta_2 = b_2$ in the diagnostic regression. What could be simpler?

Yet another perspective on these issues can be achieved by reconsidering the vector of contrasts between the ordinary least squares and IV estimators. Write the OLS estimator as

$$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\hat{\beta}_{IV} + \mathbf{e}_{IV}) = \hat{\beta}_{IV} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_{IV}.$$

The difference between the two estimators, $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_{IV}$, is a set of regression coefficients from a diagnostic regression in which the instrumental-variable residuals are regressed on all of the $\mathbf{X}$ explanatory variables. Non-zero coefficients constitute evidence against the null hypothesis of no association between $\mathbf{X}$ and the true disturbance ϵ, for which $\mathbf{e}_{IV}$ is the nearest available proxy.

If the structural model is set out as $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$, with the $\mathbf{X}_1$ assumed to be exogenous, then $\mathbf{X}_1$ is included in the set of instruments. We have previously seen that the Sargan test of the validity of the instruments can be formed by regressing $\mathbf{e}_{IV}$ on all of the instruments, including $\mathbf{X}_1$ but *not* including the problematic $\mathbf{X}_2$ variables. The Hausman and Sargan tests are therefore fundamentally different. Another difference worth mentioning is that the Sargan test requires an over-identified model whereas the Hausman test can be performed even if the model is just-identified.

Fixed versus random effects

The Hausman approach can also be applied to panel or clustered data models, which we discuss in more detail in our next chapter. Let the model be written as

$$Y_{it} = \mathbf{X}'_{it}\beta + \mathbf{Z}'_i\gamma + u_i + \epsilon_{it}.$$

We can efficiently estimate both β and γ by the random effects method under the assumption that u_i is uncorrelated with $\mathbf{X}_{it}$ and $\mathbf{Z}_i$. If this assumption is correct, the random effects estimator is both consistent and efficient. However, if the assumption is incorrect, the random effects estimator is inconsistent. The fixed effects estimator provides a consistent estimator of β in the presence of correlation, but of course this method does not permit γ to be estimated at all.

In other words, the Hausman test can be applied, but only to the contrast vector $(\hat{\beta}_{FE} - \hat{\beta}_{RE})$. To form the χ^2 test statistic, one would need to extract the estimated coefficients $\hat{\beta}_{RE}$ and their associated variance sub-matrix from the random effects output.

The same idea can be applied in testing the null model

$$Y_{it} = \mathbf{X}'_{it}\beta + \mathbf{Z}'_i\gamma + \epsilon_{it},$$

in which ϵ_{it} is assumed to be uncorrelated with both $\mathbf{X}_{it}$ and $\mathbf{Z}_i$, against the alternative

$$Y_{it} = \mathbf{X}'_{it}\beta + \mathbf{Z}'_i\gamma + u_i + \epsilon_{it},$$

in which u_i may be correlated with either $\mathbf{X}_{it}$ or $\mathbf{Z}_i$. In this instance, the appropriate contrast vector is the difference between the fixed effects and ordinary least squares estimators of β.

The Hausman test is not useful, however, in distinguishing a model in which ϵ_{it} is the sole disturbance from an alternative random effects specification in which the composite disturbance is $u_i + \epsilon_{it}$. The reason is that random effects disturbances do not cause the least squares estimator to be inconsistent. Hence the probability limit of the contrast vector is zero.

30.2 Revisiting the Regression Applications

Davidson and MacKinnon (1993, pp. 237–242, 389–93) have shown that for two important tests bearing on linear regressions, the Hausman approach can be reformulated as an artificial regression. We have already seen one example of this above, and will now reproduce Davidson and MacKinnon's illuminating argument justifying the approach. A side benefit is that the argument helps to clarify the basis for the test's degrees of freedom.

Let the two estimators be the OLS estimator $\hat{\beta}_e = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, which is the efficient estimator under the null hypothesis, and $\hat{\beta}_c = (\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}\mathbf{Y}$, which is the consistent estimator under both the null and alternative hypotheses. (The matrix $\mathbf{A}$ will be specified in a moment.) Normally one would proceed by forming the vector of contrasts $\hat{\beta}_c - \hat{\beta}_e$, estimating its asymptotic covariance matrix, inverting this matrix and then using a chi-square test to assess the null. In doing this, one has to be mindful of the possibility that the covariance matrix may not be of full rank, and Hausman has advocated the use of generalized inverses in such circumstances.

Davidson and MacKinnon propose a much simpler approach. Their idea is to form the regression model under the null hypothesis, and then add to the regression equation a set of artificial test variables whose statistical significance would cause the null to be rejected. When this approach can be applied, one does not need to use generalized inverses.

We consider two tests. The first is a standard omitted variables test where the null hypothesis is $\mathbf{Y} = \mathbf{X}_1\beta_1 + \epsilon$ and the alternative is $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$. For the omitted variables test, we have the matrix $\mathbf{A} = \mathbf{M}_2$. The second test focuses on the exogeneity of $\mathbf{X}_2$ in the model $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$, and here the comparison involves the OLS and IV estimators. In this case, the matrix $\mathbf{A} = \mathbf{P}_\mathbf{Z}$ with $\mathbf{Z}$ being the set of instruments. In both cases, $\mathbf{A}$ is symmetric.

The vector of contrasts is, for general $\mathbf{A}$,

$$\begin{aligned}\hat{\beta}_c - \hat{\beta}_e &= (\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}\mathbf{Y} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}\mathbf{X}'\mathbf{A}\mathbf{M}_\mathbf{X}\mathbf{Y}.\end{aligned}$$

If the null hypothesis is incorrect, we expect $\hat{\beta}_c - \hat{\beta}_e$ to converge in probability to a non-zero $k \times 1$ vector. The matrix $(\mathbf{X}'\mathbf{A}\mathbf{X})^{-1}$, when suitably normalized, will converge to a collection of constants, and thus will have no vital role to play in the behavior of the test statistic.

Consider the remaining factor, a $k \times 1$ vector $\mathbf{X}'\mathbf{A}\mathbf{M}_\mathbf{X}\mathbf{Y}$. In some circumstances not all elements of this vector will be stochastic; in particular, the vector may contain entries of 0 in some rows. This will occur if $\mathbf{M}_\mathbf{X}$ annihilates some of the columns of $\mathbf{A}\mathbf{X}$. Let $\mathbf{X}^*$ denote the columns that are not annihilated by $\mathbf{M}_\mathbf{X}$. We see that the essence of the Hausman test is found in the vector $\mathbf{X}^{*\prime}\mathbf{A}\mathbf{M}_\mathbf{X}\mathbf{Y}$.

Now consider the following artificial regression, which is an augmented version of the regression specified by the null hypothesis,

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{A}\mathbf{X}^*\delta + \text{residuals}, \tag{30.1}$$

and apply to this equation the FWL theorem. This yields

$$\hat{\delta} = (\mathbf{X}^{*\prime}\mathbf{A}\mathbf{M}_\mathbf{X}\mathbf{A}\mathbf{X}^*)^{-1}\mathbf{X}^{*\prime}\mathbf{A}\mathbf{M}_\mathbf{X}\mathbf{Y}$$

and plim $\hat{\delta} = \mathbf{0}$ implies that plim $\mathbf{X}^{*\prime}\mathbf{A}\mathbf{M}_\mathbf{X}\mathbf{Y} = \mathbf{0}$. An $\mathcal{F}$ test can be employed to test $\delta = \mathbf{0}$.

We now show how this general approach can be applied to tests for omitted variables and exogeneity.

Omitted variables

In the omitted variables example, the null hypothesis is $\mathbf{Y} = \mathbf{X}_1\beta_1 + \epsilon$, and we augment the null with a set of test variables $\mathbf{A}\mathbf{X}^*$ chosen to test the null $\beta_2 = \mathbf{0}$ against the alternative $\beta_2 \neq \mathbf{0}$. Here (recall the FWL theorem) we take $\mathbf{A} = \mathbf{M}_2$. In this case, no variables are annihilated, that is, $\mathbf{X}_1'\mathbf{M}_2\mathbf{M}_1$ does not contain zeroes.

The artificial regression corresponding to (2) is thus

$$\begin{aligned}\mathbf{Y} &= \mathbf{X}_1\beta_1 + \mathbf{M}_2\mathbf{X}_1\delta + \text{residuals} \\ &= \mathbf{X}_1\beta_1 + \mathbf{E}_1\delta + \text{residuals}\end{aligned} \tag{30.2}$$

where $\mathbf{E}_1$ is a $n \times k_1$ matrix of residuals from the regression of $\mathbf{X}_1$ on $\mathbf{X}_2$. The result can be rewritten as

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}_1\beta_1 + (\mathbf{I} - \mathbf{P}_2)\mathbf{X}_1\delta + \text{residuals} \\ &= \mathbf{X}_1(\beta_1 + \delta) - \mathbf{P}_2\mathbf{X}_1\delta + \text{residuals} \\ &= \mathbf{X}_1(\beta_1 + \delta) - \hat{\mathbf{X}}_1\delta + \text{residuals} \end{aligned} \tag{30.3}$$

to put it in a predicted values form.

An intriguing question is whether in the omitted variables case, the Hausman test is essentially identical to a standard $\mathcal{F}$ test. Davidson and MacKinnon (1993, p. 392) find that in general these tests are not identical, and go on to explore the conditions under which they coincide. Their discussion is highly recommended.

Exogeneity tests

Consider an exogeneity test for the model $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$, in which $\mathbf{X}_2$ may be endogenous. We have a set of instruments $\mathbf{Z} = (\mathbf{X}_1, \mathbf{W})$. Our approach is again to add a set of test variables $\mathbf{A}\mathbf{X}^*$ to the null model $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \epsilon$. In this case $\mathbf{A} = \mathbf{P}_\mathbf{Z}$, and since $\mathbf{P}_\mathbf{Z}\mathbf{X}_1 = \mathbf{X}_1$ and $\mathbf{M}_\mathbf{X}\mathbf{X}_1 = \mathbf{0}$, the exogenous variables $\mathbf{X}_1$ are annihilated and only the $\mathbf{X}_2$ variables will remain to affect the test statistic. Thus, in this instance we have $\mathbf{A}\mathbf{X}^* = \mathbf{P}_\mathbf{Z}\mathbf{X}_2$.

The artificial regression corresponding to (2) is simply

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{P}_\mathbf{Z}\mathbf{X}_2\delta + \text{residuals} \tag{30.4}$$

$$= \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \hat{\mathbf{X}}_2\delta + \text{residuals} \tag{30.5}$$

$$= \mathbf{X}_1\beta_1 + \mathbf{X}_2(\beta_2 + \delta) - \mathbf{M}_\mathbf{Z}\mathbf{X}_2\delta + \text{residuals} \tag{30.6}$$

where we have presented both the "predicted values" and the "residuals" forms of the artificial regression. Either one of these can be used to test $\delta = 0$ by means of an $\mathcal{F}$ test.

A variant of this test can be used when some of the $\mathbf{X}$ variables are known to be endogenous, but the exogeneity of others is to be tested. Let the model be $\mathbf{Y} = \mathbf{X}_1\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{X}_3\beta_3 + \epsilon$, and suppose that $\mathbf{X}_1$ is known to be exogenous, $\mathbf{X}_3$ is known to be endogenous, and we are uncertain of the status of $\mathbf{X}_2$. There are two instrument sets: $\mathbf{Z}_1 = (\mathbf{X}_1, \mathbf{W})$ corresponds to the case in which $\mathbf{X}_2$ is presumed endogenous, and $\mathbf{Z}_2 = (\mathbf{X}_1, \mathbf{X}_2, \mathbf{W})$ corresponds to the case in which $\mathbf{X}_2$ is taken to be exogenous and so can serve as its own instrument. The projection matrices associated with these instrument sets are denoted $\mathbf{P}_1$ and $\mathbf{P}_2$; note that $\mathbf{P}_1\mathbf{P}_2 = \mathbf{P}_1$.

The vector of contrasts $\hat{\beta}_1 - \hat{\beta}_2$ is as follows:

$$\begin{aligned} \hat{\beta}_1 - \hat{\beta}_2 &= (\mathbf{X}'\mathbf{P}_1\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_1\mathbf{Y} - (\mathbf{X}'\mathbf{P}_2\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_2\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{P}_1\mathbf{X})^{-1}\left(\mathbf{X}'\mathbf{P}_1(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{P}_2\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_2)\right)\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{P}_1\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_1(\mathbf{I} - \mathbf{P}_2\mathbf{X}(\mathbf{X}'\mathbf{P}_2\mathbf{X}) - 1\mathbf{X}'\mathbf{P}_2)\mathbf{Y} \\ &= (\mathbf{X}'\mathbf{P}_1\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_1\mathbf{M}_{\mathbf{P}_2,\mathbf{X}}\mathbf{Y}. \end{aligned}$$

where we have used $\mathbf{P}_1\mathbf{P}_2 = \mathbf{P}_1$. As before, the matrix $(\mathbf{X}'\mathbf{P}_1\mathbf{X})^{-1}$ will converge to a matrix of constants when suitably normalized, and the essence of the test is located in the $k \times 1$ vector

$\mathbf{X}'\mathbf{P}_1\mathbf{M}_{\mathbf{P}_2,\mathbf{X}}\mathbf{Y}$. This vector contains some non-stochastic elements, namely those associated with $\mathbf{X}_1$, which is assumed exogenous.

As Davidson and MacKinnon (1993, pp. 241–2) show, we can assess whether the probability limit of this vector is zero by applying two-stage least squares to the artificial regression

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{P}_1\mathbf{X}^*\delta + \text{residuals}$$

using $\mathbf{Z}_2$ as the first-stage set of instruments, and applying an $\mathcal{F}$ test to $\hat{\delta}$. Here $\mathbf{X}^*$ denotes the columns of $\mathbf{X}$ other than $\mathbf{X}_1$.

30.3 "Hausman Tests" using Inefficient Estimators

A fundamental problem with the Hausman test is the requirement that one of the estimators be demonstrably *efficient* under the null hypothesis. This requirement, you will recall, provides the basis for proving the key result—that the variance of the limiting distribution of the contrast between the two estimators, $\sqrt{n}(\hat{\theta}_c - \hat{\theta}_e)$, is simply $\mathbf{V}_c - \mathbf{V}_e$. But what can be done when *neither* of the two estimators is efficient, and therefore the variance of the contrast vector has to include a covariance term, which in its turn must be estimated?

In cases such as these, then the Hausman approach no longer applies in any strict sense. Instead, we must turn to an alternative generalized method of moments (GMM) approach, whereby the first-order conditions of one estimator form one set of moment conditions, and the first-order conditions for the other estimator form a second set of conditions for the same parameters. Under the null hypothesis, both sets of moment conditions have zero means; under the alternative hypothesis, however, one of these sets (whose estimator of θ is inconsistent) has non-zero means. This is the route explored by Creel (2004), who studies several GMM models and evaluates their test performance using simulations. Cameron and Trivedi (2005, p. 273) briefly review this approach. We will defer further discussion of the issues until we have had the chance to develop the GMM method in more detail.

Chapter 31

Panel and Clustered Data

We begin by considering once again the linear structural model

$$Y_{it} = \mathbf{X}'_{it}\beta + \epsilon_{it} \tag{31.1}$$

in which two subscripts, i and t, are employed to identify an observation, with $i = 1, \ldots, N$ and $t = 1, \ldots, T$. The i, t notation is appropriate for a *panel data* model in which an individual (alternatively a firm, or some other unit) is followed over time. By attaching a different meaning to these subscripts, we obtain the *clustered data* model, for which i typically denotes a geographic unit or "sampling cluster" and t denotes individuals within that cluster. When each of the N units contributes the same number (T) of observations in a panel dataset, the panel is said to be *balanced*. Apart from the complication of notation, nothing of great significance is entailed in generalizing things to the unbalanced case in which unit i provides T_i observations. This is, in fact, the usual situation encountered in clustered-data samples, it being uncommon for every geographic cluster to hold exactly the same number of individuals. In this chapter we will mainly refer to the panel data model, but it should be understood that, in general, our methods of analysis and major conclusions apply equally well to clustered data.

Before we proceed, one important caveat needs to be mentioned here at the outset. A key assumption in this chapter is that T in the balanced-panel case and $\{T_i\}$ in the unbalanced case, are *exogenous* in the sense that their values are either not random variables at all or, if random, are determined by forces entirely outside the structural model. In particular, we will assume that the $\{T_i\}$ values are unrelated to the disturbance terms of the model. But especially when individual economic units (people, households, firms) are followed over time, *sample attrition* takes place, causing some units to drop out of the sample. If attrition is in part due to the same features that figure into the disturbance terms of the structural model, then $\{T_i\}$ *cannot* be taken to be exogenously determined. Endogenous sample attrition greatly complicates the analysis of panel data, and may require adaptations to panel-data settings of the sample-selection estimators considered in Chapter 32.

31.1 Modern OLS with Pooled Panel Data

For the moment, we will maintain the core assumption that justifies ordinary least squares regression, that $\mathrm{E}(\epsilon_{it}|\mathbf{X}_{it}) = 0$. However, we will allow the ϵ_{it} to exhibit essentially any pat-

tern of heteroskedasticity. Although we will assume that the disturbances are independent over i, we will also allow them to be freely correlated within the $t = 1, \ldots, T$ records for the i-th unit. (The only restriction we need to place on the patterns of heteroskedasticity and correlation is that whatever they are, OLS must remain a consistent estimator.) We view such correlations as stemming from unmeasured variables that exert persistent influence on the i-th unit that is followed over time in a panel; similarly, unmeasured features of the i-th geographic area induce a correlation among the disturbances of the individuals who reside in that area.

Now that we have two subscripts to consider, how do we organize our data? The convention is to *stack* the data as follows. For $\{Y_{it}\}$ and $\{\mathbf{X}_{it}\}$, we sort the data first on i and then on t within i's block of records, giving for the full $NT \times 1$ vector of dependent variables, $\mathbf{Y}$, and $\mathbf{X}$, the $NT \times k$ matrix of covariates,

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1T} \\ \vdots \\ Y_{N1} \\ \vdots \\ Y_{NT} \end{bmatrix} \text{ and } \mathbf{X} = \begin{bmatrix} \mathbf{X}'_{11} \\ \vdots \\ \mathbf{X}'_{1T} \\ \vdots \\ \mathbf{X}'_{N1} \\ \vdots \\ \mathbf{X}'_{NT} \end{bmatrix}.$$

The ordinary least squares estimator of β can be written compactly as $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ and expressed in summation form as

$$\hat{\beta} = \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}'_{it} \right)^{-1} \cdot \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{X}_{it}Y_{it} = \beta + \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}'_{it} \right)^{-1} \cdot \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{X}_{it}\epsilon_{it}.$$

As discussed in the footnote, we can devise even more compact expressions.[1]

[1] Let $\mathbf{X}_i\, \epsilon_i = \sum_{t=1}^{T} \mathbf{X}_{it}\epsilon_{it}$ in which $\mathbf{X}_i$ is a $k \times T$ matrix and ϵ_i a $T \times 1$ vector; that is,

$$\mathbf{X}_i = \left[\mathbf{X}_{i,1}, \mathbf{X}_{i,2}, \ldots, \mathbf{X}_{i,T}\right] \text{ and } \epsilon_i = \begin{bmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \vdots \\ \epsilon_{i,T} \end{bmatrix}.$$

Given the definition of $\mathbf{X}_i$, it is easy to see that

$$\mathbf{X}_i\mathbf{X}'_i = \sum_{t=1}^{T} \mathbf{X}_{it}\mathbf{X}'_{it}.$$

Hence, the OLS estimator for pooled panel data can be expressed as

$$\hat{\beta} = \beta + \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i\mathbf{X}'_i \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i\epsilon_i,$$

which looks very much like the familiar summation version of the OLS formula. Only the dimensions of the objects in the formula are different. Here $\mathbf{X}_i$ is no longer a $k \times 1$ vector but is rather a $k \times T$ matrix; and ϵ_i is no longer a scalar but a $T \times 1$ vector. Under an inid assumption for the data-generating process, by which blocks of data for observations i and $j \neq i$ are independent, this simplified notation can be helpful.

31.1.1 Robust standard errors for OLS

Assuming that the disturbances ϵ_{it} are heteroskedastic and freely correlated *within* unit i's block of records, although they are independent over i, what is the variance of the limiting distribution of the ordinary least squares estimator? To address this question we first have to say what kind of limit we are imagining. Are we thinking of $N \to \infty$ while T is held fixed (this is the usual microeconometric case), or possibly $T \to \infty$ with N fixed, or both $N \to \infty$ and $T \to \infty$?

What are the essential differences among these? As we'll see, with T fixed and $N \to \infty$, large-sample analysis hardly differs from what we have done with cross-sections. Admittedly, there is a little more notation to cope with, but once you are able to see past the notation, you will discover that the basics are much the same. By contrast, with N fixed and $T \to \infty$, we have what amounts to N distinct time-series driven by common β parameters. The laws of large numbers and central limit theorems you need are mainly those applicable to conventional time-series analysis. When both $N \to \infty$ and $T \to \infty$, we must apply an exotic combination of cross-section and time-series asymptotics. Later in this chapter, we will return to the implications of using OLS in the panel-data context, and will then look more closely at how the nature of asymptotic analysis can affect conclusions about estimator consistency.

Let's consider the usual case encountered in micro-econometric research, in which $N \to \infty$ with T fixed. We have

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{X}_{it} \mathbf{X}_{it}' \right)^{-1} \cdot \sqrt{N} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{X}_{it} \epsilon_{it}.$$

(Since T is fixed by assumption, we would have gained nothing by inserting $1/T$ in these expressions.) Now, $\mathrm{E}\, \mathbf{X}_{it}\epsilon_{it} = \mathbf{0}$ and this implies $\mathrm{E} \sum_{t=1}^{T} \mathbf{X}_{it}\epsilon_{it} = \mathbf{0}$ as well. By the assumption of independence over i,

$$\mathrm{Var} \left(\sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{X}_{it} \epsilon_{it} \right) = \sum_{i=1}^{N} \mathrm{Var} \left(\sum_{t=1}^{T} \mathbf{X}_{it} \epsilon_{it} \right),$$

and

$$\mathrm{Var} \left(\sum_{t=1}^{T} \mathbf{X}_{it} \epsilon_{it} \right) = \sum_{t=1}^{T} \sum_{s=1}^{T} \mathrm{E} \left(\mathbf{X}_{it} \mathbf{X}_{is}' \epsilon_{it} \epsilon_{is} \right) \equiv \mathbf{V}_i,$$

with $\mathbf{V}_i$ being a $k \times k$ matrix. If we assume that

$$\lim_{N \to \infty} \frac{1}{N} \sum_{i=1}^{N} \mathbf{V}_i = \mathbf{V},$$

and assume that the other conditions of the Lindeberg central limit theorem are met, then we obtain the result that

$$\sqrt{N} \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \mathbf{X}_{it} \epsilon_{it} \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{V}).$$

If we further assume that

$$\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{X}_{it}\mathbf{X}_{it}' \xrightarrow{p} \mathbf{W},$$

we arrive at the limiting result

$$\sqrt{N}(\hat{\beta}-\beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathbf{W}^{-1}\mathbf{V}\mathbf{W}^{-1}).$$

We would estimate $\mathbf{W}$ in the usual way, as

$$\hat{\mathbf{W}} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\mathbf{X}_{it}\mathbf{X}_{it}'.$$

Under conditions very similar to those we saw in White's robust standard errors approach to OLS cross-sectional models with heteroskedastic disturbances, we can estimate $\mathbf{V}_i$ using OLS residuals e_{it}, e_{is} in place of the corresponding disturbances $\epsilon_{it}, \epsilon_{is}$. That is,

$$\hat{\mathbf{V}}_i = \sum_{t=1}^{T}\sum_{s=1}^{T}\left(\mathbf{X}_{it}\mathbf{X}_{is}'e_{it}e_{is}\right),$$

and then $\hat{\mathbf{V}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{V}}_i$. In this way, White's cross-sectional approach extends easily to the panel data and clustered data case.

I believe that Arellano (1987) was the first to emphasize this approach in an article on robust standard errors for fixed-effects models. But the White approach is more general than that. It can be applied to create robust standard errors in OLS, fixed-effects, and first-differences models, as we now discuss.

31.2 Error Components Models

We will now consider a family of specifications in which the disturbance term has what is termed an *error components* structure. The disturbance term of the equation is a composite, $u_i + \epsilon_{it}$, containing one component that is specific to unit i and which stays fixed over time. The linear model is

$$Y_{it} = \alpha + \mathbf{X}_{it}'\beta + \mathbf{Z}_i'\gamma + u_i + \epsilon_{it} \tag{31.2}$$

Notice that we distinguish between explanatory covariates $\mathbf{X}_{it}$ that vary over both i and t, and covariates $\mathbf{Z}_i$ that vary only across i, the individual units.

There are two error component specifications that differ in what is assumed about the u_i component. The *random effects* model is obtained if we assume $\mathrm{E}(u_i \mid \mathbf{X}_i, \mathbf{Z}_i) = 0$, where $\mathbf{X}_i$, the $k \times T$ matrix we introduced previously, holds T observations on k time-varying explanatory variables for unit i. The disturbance term correlations can be viewed as a special case of the freely-correlated model we just examined, but in which there is sufficient structure for generalized least squares methods to be applied for greater (asymptotic) efficiency, as we'll see later in this chapter.

The *fixed-effects* model represents a more important departure from the pooled OLS and GLS models—it results if we assume that $\mathrm{E}(u_i \mid \mathbf{X}_i, \mathbf{Z}_i) \neq 0$. (Here, u_i is termed the "fixed

effect", at least by economists. Elsewhere in the statistical world, the phrase "fixed effects" often refers to what we in economics would call "parameters".) A fundamental assumption justifying ordinary least squares is thus violated: the (composite) disturbance term is correlated with one or more of the explanatory variables. In the fixed-effects approach, we do not usually posit a model of the correlation between u_i and the right-hand side variables $(\mathbf{X}_i, \mathbf{Z}_i)$. That is, we assume that some correlation exists but take no position as to its precise form. Note that without a model of the correlation, we have no basis for separating the explanatory variables into a "safe" set and a "problematic" set as we did with instrumental variables. We would not even know how many instruments are required!

It is common practice in both the random and the fixed-effects approaches to focus attention on u_i by assuming that the other disturbance term component ϵ_{it} is innocuous in the sense that $\mathrm{E}(\epsilon_{it} \mid \mathbf{X}_i, \mathbf{Z}_i) = 0$. Typically the ϵ_{it} component of the overall disturbance is additionally assumed to be homoskedastic and uncorrelated in both the i and t dimensions. But rather than impose these specialized assumptions from the outset, we will first see what more general assumptions on $\{\epsilon_{it}, t = 1, \ldots, T\}$ might entail.

31.3 Solution by Subtraction?

Evidently, the fixed-effects estimation problem has mainly to do with the u_i error component. If we could somehow find a way to make it disappear, surely that would open up possibilities for consistent estimation of the β parameters. There are two ways to make u_i disappear; they are termed the *deviations from means* and the *first-differencing* approaches. Both approaches begin with

$$Y_{it} = \alpha + \mathbf{X}'_{it}\beta + \mathbf{Z}'_i\gamma + u_i + \epsilon_{it}$$

and transform the series of observations for unit i in such a way that u_i is removed from the scene. For $T = 2$ the two methods are algebraically identical, but they differ in fundamental ways for $T \geq 3$.

31.3.1 Deviations from means

In the deviations from means method, we form the mean value $\bar{Y}_i$ by taking averages over unit i's records,

$$\bar{Y}_i = \alpha + \bar{\mathbf{X}}'_i\beta + \mathbf{Z}'_i\gamma + u_i + \bar{\epsilon}_i$$

and subtract this from the equation above, yielding

$$Y_{it} - \bar{Y}_i = (\mathbf{X}_{it} - \bar{\mathbf{X}}_i)'\beta + \epsilon_{it} - \bar{\epsilon}_i.$$

Note first that the transformation to deviations from unit means, which has the desirable effect of removing u_i, also has an unfortunate and deeply undesirable side-effect: it removes *all* explanatory variables that are constant for unit i. Once the data are transformed in this way, neither the constant α nor the γ coefficients associated with $\mathbf{Z}_i$ can be estimated.

Transforming the data imposes a heavy penalty when these coefficients are of substantive interest, as they usually are.[2]

Representing the transformed data as

$$\tilde{Y}_{it} = \tilde{\mathbf{X}}'_{it}\beta + \tilde{\epsilon}_{it},$$

we now need to investigate whether with u_i out of the way, we can estimate β consistently by applying ordinary least squares to the transformed data. The consistency issue concerns the conditional expectation

$$\mathrm{E}\left(\tilde{\epsilon}_{it} \mid \tilde{\mathbf{X}}_{it}\right) = \mathrm{E}\left(\epsilon_{it} - \bar{\epsilon}_i \mid \mathbf{X}_{it} - \bar{\mathbf{X}}_i\right),$$

which involves *all* of the $\{\epsilon_{i,1}, \epsilon_{i,2}, \ldots, \epsilon_{i,T}\}$ disturbances for unit i and also *all* of the $\mathbf{X}_{i,t}$ vectors of explanatory variables for that unit. If

$$\mathrm{E}\left(\epsilon_{it} - \bar{\epsilon}_i \mid \mathbf{X}_{it} - \bar{\mathbf{X}}_i\right) = 0 \qquad \forall t, i$$

then we say that $\{\mathbf{X}_{i,t}\}$ is *strongly exogenous* to $\{\epsilon_{it}\}$: there is no correlation between any pair of $\epsilon_{i,t}$ and $\mathbf{X}_{i,s}$. In particular, no past value of $\mathbf{X}_{i,s}$ can be linked to a future $\epsilon_{i,t}$; nor can a past $\epsilon_{i,s}$ be connected to a future $\mathbf{X}_{i,t}$. If strong exogeneity holds, then only mild additional assumptions are needed to prove that applying OLS to the transformed data gives a consistent estimator of the β parameters.

Deviations from means is also termed *within* regression, as it uses only within-unit variation around the unit means. It may be helpful to see it expressed differently. Let $\mathbf{M} = \mathbf{I}_T - \frac{1}{T}\boldsymbol{\iota}\boldsymbol{\iota}'$. The transformation to deviations from means amounts to multiplying the full $\mathbf{Y}$ vector and $\mathbf{X}$ matrix by the $NT \times NT$ matrix $\mathbf{Q} = \mathbf{I}_N \otimes \mathbf{M}$. Note that $\mathbf{Q}$ is symmetric and idempotent. We write the transformed model as

$$\mathbf{QY} = \mathbf{QX}\beta + \mathbf{Qu} + \mathbf{Q}\epsilon = \mathbf{QX}\beta + \mathbf{Q}\epsilon$$

and apply least squares to these transformed data. We thus obtain the fixed-effects (within) estimator of β,

$$\tilde{\beta} = (\mathbf{X}'\mathbf{QX})^{-1}\mathbf{X}'\mathbf{QY} = \beta + (\mathbf{X}'\mathbf{QX})^{-1}\mathbf{X}'\mathbf{Q}\epsilon. \tag{31.3}$$

There is one remaining question about applying OLS to the deviations-from-means data: Can the usual OLS form of the $\hat{\beta}$ covariance matrix be used to test hypotheses? The problem is that even if the original $\epsilon_{i,t}$ disturbances are uncorrelated over t and homoskedastic, the transformed disturbances $\tilde{\epsilon}_{it} = \epsilon_{it} - \bar{\epsilon}_i$ will be inter-correlated. Of course, if the original $\epsilon_{i,t}$ disturbances are serially correlated or heteroskedastic or both, we would hardly expect to conduct hypothesis tests without first modifying the OLS covariance matrix. We'll return shortly to explore these issues.

[2]The cost is due in part to our professed ignorance about how u_i and $(\mathbf{X}_i, \mathbf{Z}_i)$ might be correlated. If we have specific hypotheses about the nature of the correlation, see Hausman and Taylor (1981), a possibility exists for recovering information about the coefficients of $\mathbf{Z}_i$.

31.3.2 First differences

Another way of removing u_i is to take *first differences*. Begin with

$$Y_{it} = \alpha + \mathbf{X}'_{it}\beta + \mathbf{Z}'_i\gamma + u_i + \epsilon_{it}$$

and, writing the equation for period $t-1$,

$$Y_{it-1} = \alpha + \mathbf{X}'_{it-1}\beta + \mathbf{Z}'_i\gamma + u_i + \epsilon_{it-1},$$

subtract Y_{it-1} from Y_{it} to obtain

$$Y_{it} - Y_{it-1} = (\mathbf{X}_{it} - \mathbf{X}_{it-1})'\beta + \epsilon_{it} - \epsilon_{it-1}.$$

We can represent the first-differenced model as

$$\overset{*}{Y}_{it} = \overset{*}{\mathbf{X}}{}'_{it}\beta + \overset{*}{\epsilon}_{it}.$$

Just as with the deviation-from-means approach, the first-differences approach sweeps away the constant term and other time-invariant explanatory variables along with the u_i error component. Note, too, that the transformed disturbance term $\overset{*}{\epsilon}_{it} = \epsilon_{it} - \epsilon_{it-1}$ is serially correlated.

The consistency issue with the first-differenced data has to do with

$$\mathrm{E}\left(\epsilon_{it} - \epsilon_{it-1} \mid \mathbf{X}_{it} - \mathbf{X}_{it-1}\right),$$

which concerns only the disturbances and explanatory variables for periods t and $t-1$. In assuming

$$\mathrm{E}\left(\epsilon_{it} - \epsilon_{it-1} \mid \mathbf{X}_{it} - \mathbf{X}_{it-1}\right) = 0,$$

we are assuming *weak exogeneity* in the relationship. This is, in essence, the claim that $\mathbf{X}_{it-1}$ has no correlation with ϵ_{it} and likewise that ϵ_{it-1} is uncorrelated with $\mathbf{X}_{it}$. It is often a more palatable assumption than the strong exogeneity assumption needed for the deviations-from-means model.

To put this differently, *first-differences gives consistency under weaker conditions than fixed-effects*. This theoretical advantage needs to be weighed against some potential disadvantages. In the data-generating process that you are modelling, it may be the case that year-to-year changes in the dependent and explanatory variables tend to be small, implying that parameters of first-differences models may not be well-identified in empirical terms. In deviations-from-means models estimated on time-trended data, by contrast, the deviations may exhibit more variability over the full range of $t \in \{1, 2, \ldots, T\}$, allowing the parameters to be estimated more precisely. But these are loose, non-rigorous arguments; other factors need consideration as well, such as the existence and serial correlation of measurement error in the explanatory variables (see Griliches and Hausman 1986).

31.3.3 Hypothesis testing with transformed data

Let's now consider the conditions that permit us to test hypotheses using the deviations-from-means method. We have already discussed the *strong exogeneity* assumption that delivers consistency, and so let us pass on to the question of the covariance matrix of the OLS estimator applied to such transformed data. Write the OLS estimator using the simplified notation we introduced earlier in the chapter,

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{X}}_i \tilde{\mathbf{X}}_i' \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^{N} \tilde{\mathbf{X}}_i \tilde{\epsilon}_i,$$

in which

$$\tilde{\mathbf{X}}_i = \left[\tilde{\mathbf{X}}_{i,1}, \tilde{\mathbf{X}}_{i,2}, \ldots, \tilde{\mathbf{X}}_{i,T} \right] \text{ and } \tilde{\epsilon}_i = \begin{bmatrix} \tilde{\epsilon}_{i,1} \\ \tilde{\epsilon}_{i,2} \\ \vdots \\ \tilde{\epsilon}_{i,T} \end{bmatrix}$$

Note that

$$\tilde{\epsilon}_i = \mathbf{M} \epsilon_i$$

for $\mathbf{M} = \mathbf{I}_T - \frac{1}{T} \boldsymbol{\iota} \boldsymbol{\iota}'$. Similarly,

$$\tilde{\mathbf{X}}_i' = \mathbf{M} \mathbf{X}_i'$$

Hence, since $\mathbf{M}$ is symmetric idempotent,

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i \mathbf{M} \mathbf{X}_i' \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i \mathbf{M} \epsilon_i,$$

By the strong exogeneity assumption, $\mathrm{E}\, \mathbf{X}_i \mathbf{M} \epsilon_i = \mathbf{0}_k$ and the variance of this vector is therefore

$$\Omega_i = \mathrm{E}\, \mathbf{X}_i \mathbf{M} \epsilon_i \epsilon_i' \mathbf{M} \mathbf{X}_i'.$$

If the ϵ_i are serially correlated, heteroskedastic, or both, we could continue along these lines to develop a White-type estimator of *robust standard errors* as outlined earlier in this chapter for the pooled OLS estimator. Apart from the need to use the transformed $\tilde{\mathbf{X}}$ variables, the approach is very much the same as in the OLS application (Arellano 1987).

Even today, however, much of the literature concentrates on the simplest textbook case in which strong exogeneity holds *and* the untransformed disturbances are not only serially uncorrelated within i's record but also homoskedastic with variance σ_ϵ^2. The variance simplifies to

$$\sigma_\epsilon^2 \mathrm{E}\, \mathbf{X}_i \mathbf{M} \mathbf{X}_i' \equiv \sigma_\epsilon^2 \mathbf{V}_i.$$

Assuming now, as we usually do when applying central limit theorems to inid data, that

$$\lim_{N \to \infty} \frac{1}{N} \sum_i \mathbf{V}_i = \mathbf{V},$$

then

$$\sqrt{N} \frac{1}{N} \sum_{i=1}^{N} \mathbf{X}_i \mathbf{M} \epsilon_i \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{V})$$

and in consequence,

$$\sqrt{N}(\hat{\beta} - \beta) \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{V}^{-1})$$

This is the usual large-sample formula for the covariance of the limiting distribution. At least under strong exogeneity, if the ϵ_i disturbances are serially uncorrelated and homoskedastic, OLS can be applied to deviations from means data and hypothesis tests can be conducted using the standard OLS formula if a consistent estimator of σ_ϵ^2 is available.

The combination of the strong exogeneity assumption and inid data series gives us a surprising gift: not only is the fixed-effect estimator consistent, it is actually *unbiased*! The reason is that the strong exogeneity assumption is essentially an assumption that for all i, we have $\mathrm{E}(\epsilon_i \mid \mathbf{X}_i) = \mathbf{0}_{T\times 1}$ and with independence assumed over i, we have effectively assumed $\mathrm{E}(\epsilon \mid \mathbf{X}) = \mathbf{0}_{NT\times 1}$, which is an assumption that we have rarely seen since our earliest investigations into the simple OLS regression model.

The fact that $\mathrm{E}(\epsilon \mid \mathbf{X}) = \mathbf{0}$ helps us to estimate σ_ϵ^2. We can work directly with the residuals from the transformed data, $\tilde{\mathbf{e}} = \tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\hat{\beta} = \tilde{\mathbf{M}}\epsilon$, where $\tilde{\mathbf{M}} = \mathbf{Q} - \mathbf{Q}\mathbf{X}(\mathbf{X}'\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}'\mathbf{Q}$, a symmetric and idempotent matrix. We have $\tilde{\mathbf{e}}'\tilde{\mathbf{e}} = \epsilon'\tilde{\mathbf{M}}\epsilon$ and it is readily verified that $\mathrm{trace}(\tilde{\mathbf{M}}) = \mathrm{trace}(\mathbf{Q}) - k$, where k is the number of variables in $\mathbf{X}$. Recall that the $NT \times NT$ matrix $\mathbf{Q}$ is block diagonal, with the $T \times T$ matrix $\mathbf{M}_\iota$ on each diagonal. Given that $\mathrm{trace}(\mathbf{M}_\iota) = T - 1$, it follows that $\mathrm{trace}(\mathbf{Q}) = N(T-1)$. Thus, an unbiased estimator of σ_ϵ^2 is found to be

$$\tilde{\sigma}_\epsilon^2 = \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{N(T-1) - k}. \tag{31.4}$$

Note that the degrees of freedom are $N(T-1) - k$, not $NT - k$; the latter is the expression that would be used by a simple OLS program applied to the transformed data. More generally, in an unbalanced panel in which unit i supplies T_i observations, the degrees of freedom are $\sum_{i=1}^{N} T_i - N - k$, as can be seen by replacing $\mathbf{M}$ with $\mathbf{M}_i$ in the above.

31.4 The LSDV method

One often hears the acronym LSDV—for "least squares dummy variables"—used as a synonym for the deviations from means method. The basic idea of LSDV is to introduce into the structural specification N dummy variables, one for each i unit, that identify each such unit. Obviously a model with N such dummy variables cannot be estimated if it also has time-invariant variables; the combination of the dummy variables and the invariant terms of the model (the α and $\mathbf{Z}_i'\delta$ terms in the structural model $Y_{it} = \alpha + \mathbf{X}_{it}'\beta + \mathbf{Z}_i'\delta + \tilde{u}_i + \epsilon_{it}$) would exhibit perfect collinearity. So to pursue the LSDV approach, we must first subsume all time-invariant terms in the u_i fixed effect, that is, we respecify the model in the form $u_i = \alpha + \mathbf{Z}_i'\delta + \tilde{u}_i$ and adjust our interpretation of u_i accordingly.

With these modifications, the model set out in equation (31.2) can be rewritten as

$$\mathbf{Y} = \mathbf{X}\beta + (\mathbf{I}_N \otimes \boldsymbol{\iota}_T)\mathbf{u} + \epsilon \tag{31.5}$$

where $\mathbf{Y}$ is of dimension $NT \times 1$ and $I_N \otimes \iota_T$ is a matrix of dimension $NT \times N$,

$$(\mathbf{I}_N \otimes \boldsymbol{\iota}_T)\mathbf{u} = \begin{bmatrix} \boldsymbol{\iota}_T & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\iota}_T & \mathbf{0} & \mathbf{0} \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \boldsymbol{\iota}_T \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{bmatrix} \equiv \mathbf{Du},$$

which has the effect in equation (31.5) of repeating u_i some T times for each unit i.

When the model is written out in this way, we are led to think of the u_i not as unobserved random variables that need to be subtracted away, but rather as N additional parameters that might be estimated by least squares, yielding a set of $\tilde{u}_i$. That is, we might regard a given u_i as a realization of an unobserved random variable U_i for the i-th unit. Since the data provide us with T observations on each such unit, we are in a position to estimate each realized value.[3]

To see this more clearly, let's relabel the N dummy-variable columns of the $\mathbf{D}$ matrix, writing the full model as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{D}_1 u_1 + \mathbf{D}_2 u_2 + \cdots + \mathbf{D}_N u_N + \epsilon.$$

We could estimate this equation by ordinary least squares. For large N, the LSDV approach might not seem to be computationally feasible, but applying the FWL Theorem simplifies the calculations. The theorem suggests that to estimate β we can first project $\mathbf{Y}$ and $\mathbf{X}$ onto the matrix of dummy variables and then work with the residuals from these projections. We can then return to the untransformed data to find estimates of the u_i.

Considering the nature of the dummy variables, it is clear that the projection of $\mathbf{Y}$ onto them must take the form

$$[\underbrace{\bar{Y}_1, \ldots, \bar{Y}_1}_{T}, \ldots, \underbrace{\bar{Y}_N, \ldots, \bar{Y}_N}_{T}]'$$

in which the mean value of $\mathbf{Y}_i$, denoted by $\bar{Y}_i$, is repeated T times for each i. A similar expression applies to each column of $\mathbf{X}$. For a given i, the residuals from the projection of $\mathbf{Y}$ can be written as

$$\begin{bmatrix} Y_{i1} - \bar{Y}_i \\ \vdots \\ Y_{iT} - \bar{Y}_i \end{bmatrix} = \left[\mathbf{I}_T - \frac{\boldsymbol{\iota}_T \boldsymbol{\iota}_T'}{T}\right] \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{iT} \end{bmatrix} = \mathbf{M}_\iota \mathbf{Y}_i,$$

and all columns of $\mathbf{X}$ can be treated likewise. In other words, we simply transform the $\mathbf{Y}_i$ and $\mathbf{X}_i$ data into deviations from the i-specific means. The LSDV estimator of β is numerically identical to the fixed-effect estimator!

Given $\tilde{\beta}$, we can consider estimating the u_i via the formula

$$\tilde{u}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} - \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_{it}'\tilde{\beta} = \frac{1}{T}\sum_{t=1}^{T} \mathbf{X}_{it}'(\beta - \tilde{\beta}) + u_i + \frac{1}{T}\sum_{t=1}^{T} \epsilon_{it}. \tag{31.6}$$

[3] Alternatively—and this is the point of view we will adopt in the random effects model below—we might continue to think of the u_i as realizations of random variables U_i but attempt to estimate only the parameters of U_i's distribution. This is possible, in general, only if we assert that U_i and $\mathbf{X}_i$ are independent, as otherwise the distribution of U_i will depend in some manner on $\mathbf{X}_i$. Although we could imagine formulating a model of the relationship between U_i and $\mathbf{X}_i$, to do so would take us outside the conventional fixed-effects framework.

The estimator is exactly what would be produced by applying least squares directly to equation (31.5). To see this, let $\mathbf{e}$ be the full $NT \times 1$ vector of LSDV residuals, with

$$e_{it} = Y_{it} - \mathbf{X}'_{it}\tilde{\beta} - \sum_{j=1}^{N} \mathbf{D}_j \tilde{u}_j,$$

and study the implications of the orthogonality condition $\mathbf{D}'_i\mathbf{e} = 0$.

Interestingly, even if the fixed-effect (LSDV) estimator of β is consistent, $\tilde{u}_i$ is *not* consistent for u_i. The reason is that $\tilde{u}_i$ depends directly only on the T observations available for unit i. (It also depends indirectly on other data through $\tilde{\beta}$, but we can assume that $\tilde{\beta} \xrightarrow{p} \beta$.) With T fixed, $\tilde{u}_i$ does not converge to u_i, but rather converges in probability to a random variable. That is,

$$\tilde{u}_i \stackrel{a}{=} u_i + T^{-1} \sum_{t=1}^{T} \epsilon_{it},$$

which is the quantity u_i plus a random variable with a non-degenerate distribution.

This is a curious result indeed! For one set of parameters, the β vector and σ^2_ϵ, we have obtained consistent estimators. For the other set, the u_i parameters, we have derived informative but inconsistent estimators.

Would there be any point to applying the $\tilde{u}_i$ formula if the structural model contains a constant term α and $\mathbf{Z}_i$? In this case

$$\tilde{u}_i = \bar{Y}_i - \bar{\mathbf{X}}'_i\tilde{\beta},$$

converges in probability to the random variable $u_i + \alpha + \mathbf{Z}'_i\gamma + \bar{\epsilon}_i$. This obviously affects the interpretation that we place on $\tilde{u}_i$. Moreover, as we shall see shortly, it also rules out one possible use of the $\tilde{u}_i$ as inputs into the random effects model.

31.5 Two-way and multi-way fixed effects

The deviations-from-means method is the simplest example of a transformation that allows the β parameters of the structural model $Y_{it} = \mathbf{X}'_{it}\beta + u_i + \epsilon_{it}$ to be estimated consistently despite the presence of the N "nusiance parameters" u_i, the number of which obviously goes to infinity as N itself does. Even for very large N, the transformation of Y_{it} to $Y_{it} - \bar{Y}_i$ is computationally inexpensive. With $\tilde{\beta}$ in hand, we can then generate estimates of $\tilde{u}_i$ of the fixed effects, although as we've just seen, these estimators (although useful and informative) are not consistent in the fixed-T, $N \to \infty$ case.

What complications would be introduced by including a second set of fixed effects in addition to the $\{u_i\}$? For example, could fixed-effects τ_t for each time-period $t \in \{1, 2, \ldots, T\}$ be added to the structural model? If T is not large, the most straightforward approach is simply to add dummy variables for each period t to the structural model, and then transform the data by taking deviations from i-specific means exactly as before. Any second group of fixed effects, provided that the number of additional dummy variables is not large, can be handled in this way.

But when the number of additional dummies needed grows large, this approach becomes computationally infeasible. In economics, such problems arise in long longitudinal worker–firm datasets in which hundreds of thousands of workers are followed over time along with the thousands of firms in which they are employed. Models with both worker and firm fixed effects requires new computational techniques. One approach—see Guimarães and Portugal (2010) for a very readable account—exploits the OLS first-order conditions for the β parameters and each one of the many fixed effects (also treated as parameters), and solves these equations in an iterative manner (zig-zagging between β and the large block of fixed-effect equations) starting from assumed initial values—this is an application of the famous *Gauss-Seidel* algorithm. Although ultimately effective, the algorithm is undeniably slow: Typically hundreds of iterations are required before the parameter estimates converge. Other approaches—some of which are quite mathematically sophisticated— estimate β taking the FWL theorem as their starting point, and successively approximate the projection used in this theorem (Correia 2016).

There are further practical difficulties to consider, having to do with identification. For example, as discussed in the influential paper by Abowd, Kramarz, and Margolis (1999) that did much to launch the economic literature on two-way fixed effects, in a matched worker-firm longitudinal dataset with effects for workers and for individual firms, imagine a world in which all worker are attached for the life of the panel study to the firm at which they were employed when the panel began, and firms end the panel with the same workers they started with. In an imaginary world like this, separate worker and firm fixed effects could not be identified. In more realistic settings, we would need, at a minimum, to separate out workers who switch firms from those who do not, and make provision in the model specification to handle the identification issues.

31.6 Models of the correlation

If you are willing to specify a model of the way in which u_i is associated with the variables in $\mathbf{X}_i$, it is at least possible to learn more about some covariate effects than can be learned from conventional fixed-effects models. Consider the specification

$$u_i = f(\mathbf{Z}_i, \mathbf{X}_{i,1}, \mathbf{X}_{i,2}, \ldots, \mathbf{X}_{i,T}, \theta) + v_i$$

in which you have in mind a particular functional form for $f()$ and are comfortable assuming that v_i in this auxiliary model is *uncorrelated* with all of the explanatory variables. Inserting this correlation model into the structural model yields

$$Y_{i,t} = \mathbf{X}'_{i,t}\beta + \mathbf{Z}'_i\delta + f(\mathbf{Z}_i, \mathbf{X}_{i,1}, \mathbf{X}_{i,2}, \ldots, \mathbf{X}_{i,T}, \theta) + v_i + \epsilon_{i,t}.$$

The error components v_i and ϵ_{it} can now be regarded as "random effects" in the sense that neither one (by assumption) is correlated with the explanatory variables. That sounds like progress of a sort, since as we will see, random effects models can be estimated as they stand without any need to first remove the v_i error component by a "within" transformation.

However, we are left with the difficult question of *identification*. The time-constant variables Z_i enter both the structural model and the model of correlation—it is not obvious that these two distinct effects on $Y_{i,t}$ can be disentangled. Indeed, in the most famous

specification of this type due to Mundlak (1978), which is $u_i = \mathbf{Z}_i'\theta_Z + \bar{\mathbf{X}}_i'\theta_X + v_i$, the δ coefficients of the structural model clearly cannot be separated from the θ_Z coefficients of the correlation model. There is somewhat more potential for distinguishing β from θ_X since only $\mathbf{X}_{i,t}$ enters the structural model whereas past and future $\mathbf{X}_{i,s}, s \neq t$, enter the $f()$ correlation model. But in a practical application, there may be relatively little variation in $\mathbf{X}_{i,t}$ that is empirically distinguishable from variation in $\mathbf{X}_{i,s}$ and note that $\mathbf{X}_{i,t}$ itself enters $f()$ in the general case. In short, it would appear that the payoffs from specifying a model of correlation will be situation-dependent, sensitive to the functional form chosen for $f()$ and to the nature of the variation in the data.

31.7 The Random-Effects Model

The random effects model is applicable provided that $\mathrm{E}(u_i \mid \mathbf{X}_i, \mathbf{Z}_i) = 0$. In this model, we attempt to estimate not the u_i values themselves, but rather the variance σ_u^2 of the distribution from which the u_i are drawn. There is no need to be concerned about the presence of a constant term α or time-invariant $\mathbf{Z}_i$ variables, because in the random effects case these variables present no serious difficulties.

Properties of the OLS estimator

The random effects model is little more than a special case of the generalized least squares model, and as we did earlier in this chapter and in more detail in Chapter 24, we should begin by examining the consequences if ordinary least squares is applied to such a model. As we have come to expect, in the random effects case the OLS estimator of equation (31.2) is consistent for α, β and γ under fairly weak conditions provided that $N \to \infty$. However, its estimated standard errors will be incorrect. And we have not really investigated what might happen if $T \to \infty$ with N fixed.

To understand the issues, consider the simplest case in which $Y_{it} = \alpha + u_i + \epsilon_{it}$. The least squares estimator of α is the sample mean of Y_{it}, or

$$\hat{\alpha} = \alpha + \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}(u_i + \epsilon_{it}).$$

From this expression it is obvious that $\hat{\alpha}$ is unbiased. Since the data are uncorrelated over i, the variance of $\hat{\alpha}$ is

$$\begin{aligned}
\mathrm{Var}(\hat{\alpha}) &= \frac{1}{(NT)^2}\sum_{i=1}^{N}\mathrm{Var}\left(\sum_{t=1}^{T}(u_i + \epsilon_{it})\right) \\
&= \frac{1}{(NT)^2}\sum_{i=1}^{N}\mathrm{Var}\left(Tu_i + \sum_{t=1}^{T}\epsilon_{it}\right) \\
&= \frac{1}{(NT)^2}\sum_{i=1}^{N}\left(T^2\sigma_u^2 + T\sigma_\epsilon^2\right) \\
&= \frac{1}{N}\sigma_u^2 + \frac{1}{NT}\sigma_\epsilon^2.
\end{aligned}$$

Evidently $\hat{\alpha}$ is not consistent unless $N \to \infty$, since with N fixed and $T \to \infty$ only the second term in the expression goes to zero.

This curious result arises because of the following. For a given unit i, there are limited returns to increasing T, because each Y_{it} that is added to the sample will be correlated with earlier values $Y_{is}, s < t$, and the extent of correlation will remain the same no matter how far apart in time are Y_{it} and Y_{is}. Information about the mean α does not accumulate as rapidly with increasing T as it would in a simple random sample. What is true for a given i remains true when N is held fixed: increases in T *alone* fail to contribute enough new information for $\hat{\alpha}$ to converge to α.

To continue the example, let us abandon for a moment the assumption that the panel is balanced. Let T vary by unit, with T_i being the number of observations for the i-th unit. This is the so-called "unbalanced" case, a sample design that is commonly encountered in cluster samples.

In the cluster sample context, we might ask how the finite-sample variance of $\hat{\alpha}$ is affected by within-cluster correlation due to u_i. We anticipate that the variance will be inflated by comparison to the case of independent random sampling, but by how much? Deaton (1997, pp. 53–59) explores the statistical literature on clustered samples from the viewpoint of the applied econometrician; his discussion of these matters is highly recommended.

To see the issues clearly, let us compare the random effects model $Y_{it} = \alpha + u_i + \epsilon_{it}$ to an alternative model, $Y_{it} = \alpha + w_{it}$, in which the disturbance w_{it} has a mean of zero and is uncorrelated over both i and t. Since the random effects disturbance term $u_i + \epsilon_{it}$ has variance $\sigma_u^2 + \sigma_\epsilon^2$, the most instructive way to compare these two cases is to assume that w_{it} also has a variance equal to $\sigma_u^2 + \sigma_\epsilon^2$.

In the unbalanced random effects model, the use of ordinary least squares to estimate α gives

$$\hat{\alpha} = \alpha + \frac{1}{(\sum_{i=1}^{N} T_i)} \sum_{i=1}^{N} \sum_{t=1}^{T_i} (u_i + \epsilon_{it}),$$

and, after some manipulation, the variance of $\hat{\alpha}$ can be shown to be

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma_\epsilon^2}{\sum_i T_i} + \left(\frac{\sigma_u^2}{\sum_i T_i} \right) \left(\frac{\sum_i T_i^2}{\sum_i T_i} \right).$$

The last expression in parentheses on the right is recognizable as a weighted average: it is the average cluster size with the cluster size itself serving as a weight. That is, with

$$\tilde{T} \equiv \sum_i T_i \left(\frac{T_i}{\sum_j T_j} \right),$$

we add and subtract $\sigma_u^2 / \sum_i T_i$ to obtain

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma_u^2 + \sigma_\epsilon^2}{\sum_i T_i} + \frac{\sigma_u^2 (\tilde{T} - 1)}{\sum T_i}.$$

Because we have assumed the variance of w_{it} to be $\sigma_u^2 + \sigma_\epsilon^2$, the first term above represents the variance of $\hat{\alpha}$ in the case of uncorrelated disturbances. The second term represents the inflation of the variance that comes about because of the inter-correlation of the disturbances.

This expression is closely related to what statisticians term the *design effect*. The terminology refers to the fact that for reasons of cost, surveys are commonly designed to be fielded in two steps, with the geographic sampling clusters selected first and a set of T_i individuals then sampled from within each cluster. In view of the travel and related logistical costs of interviewing, this design is much less expensive to implement than simple random sampling. If there exist cluster-specific unmeasured effects akin to u_i, however, the design causes the variances of estimators to be inflated relative to what they would be in a simple random sample.

To sum up, if ordinary least squares is applied to a random effects model, the estimators may not be consistent (as we've seen, this depends on whether N or T goes to infinity) and the covariance matrix calculated by the regression program will be incorrect, causing the estimated standard errors of the regression coefficients to be too small, on average. A researcher might therefore be misled into thinking that a result is statistically significant when, in fact, it is not.

The random effects model

Having explored the consequences of applying ordinary least squares, we now develop the appropriate generalized least squares estimator. Let $\mathbf{V}_i$ denote the covariance matrix of the composite disturbance for unit i, where $\mathbf{V}_i = \sigma_u^2\mathbf{J} + \sigma_\epsilon^2\mathbf{I}$, with $\mathbf{J}$ being $\boldsymbol{\iota}_T\boldsymbol{\iota}_T'$, a $T \times T$ matrix of ones. It can be shown, see Baltagi (2005), that $\mathbf{V}_i$ has only two distinct eigenvalues, $T\sigma_u^2 + \sigma_\epsilon^2$ and σ_ϵ^2, the former having multiplicity 1 and the latter having multiplicity $T-1$. The determinant of $\mathbf{V}_i$ is the product of its eigenvalues, i.e.,

$$\mid \mathbf{V}_i \mid = (T\sigma_u^2 + \sigma_\epsilon^2)(\sigma_\epsilon^2)^{T-1},$$

a result that we will use shortly in the context of maximum likelihood estimation. The inverse of $\mathbf{V}_i$ is

$$\begin{aligned} \mathbf{V}_i^{-1} &= \frac{\boldsymbol{\iota}\boldsymbol{\iota}'}{T}\frac{1}{(T\sigma_u^2 + \sigma_\epsilon^2)} + \left(\mathbf{I} - \frac{\boldsymbol{\iota}\boldsymbol{\iota}'}{T}\right)\frac{1}{\sigma_\epsilon^2} \\ &= \frac{1}{\sigma_\epsilon^2}\left(\mathbf{I} - \frac{\sigma_u^2}{T\sigma_u^2 + \sigma_\epsilon^2}\boldsymbol{\iota}\boldsymbol{\iota}'\right) \end{aligned} \tag{31.7}$$

Furthermore, the matrix $\mathbf{P}_i$ such that $\mathbf{P}_i'\mathbf{P}_i = \mathbf{V}_i^{-1}$ is

$$\begin{aligned} \mathbf{P}_i &= \mathbf{I} - \left(1 - \left(\frac{\sigma_\epsilon^2}{T\sigma_u^2 + \sigma_\epsilon^2}\right)^{1/2}\right)\frac{\boldsymbol{\iota}\boldsymbol{\iota}'}{T} \\ &= \mathbf{I} - \theta\frac{\boldsymbol{\iota}\boldsymbol{\iota}'}{T} \end{aligned}$$

The EGLS approach

To estimate the generalized least squares model, we first want to derive an estimate of θ, call it $\tilde{\theta}$, and then transform the data using $\tilde{\mathbf{P}}_i$. Note that

$$\tilde{\mathbf{P}}_i \mathbf{Y}_i = \begin{bmatrix} Y_{i1} - \tilde{\theta}\bar{Y}_i \\ \vdots \\ Y_{iT} - \tilde{\theta}\bar{Y}_i \end{bmatrix},$$

and similarly for each column vector of $\mathbf{X}_i$ and for the time-invariant columns $\boldsymbol{\iota}$ (attached to the constant term) and $\boldsymbol{\iota} \otimes \mathbf{Z}_i'$. Transforming i's data by $\tilde{\mathbf{P}}_i$ is equivalent to quasi-differencing from means. To make this transformation, however, we must already have in hand estimates of the variances σ_ϵ^2 and σ_u^2.

The fixed-effects transformation provides us with an estimate of σ_ϵ^2 and this remains valid whether or not the model includes a constant term and time-invariant $\mathbf{Z}_i$. Hence, it is standard practice to begin the estimation of a random-effects model by calculating the within regression in order to determine the σ_ϵ^2 parameter. Securing an estimate of σ_u^2 is somewhat more difficult. One approach (applicable only if the structural model has neither a constant term nor $\mathbf{Z}_i$) is simply to use the sample variance of the $\tilde{u}_i$, the fixed-effects estimates. Another approach is to collapse the data into N means,

$$\bar{Y}_i = \alpha + \bar{\mathbf{X}}_i'\beta + \mathbf{Z}_i'\gamma + u_i + \bar{\epsilon}_i$$

and to apply least squares to what is termed the *between* regression. The squared residuals from the between regression provide a consistent estimate of the composite quantity

$$\sigma_*^2 = \sigma_u^2 + \frac{\sigma_\epsilon^2}{T}$$

and from this we can derive

$$\tilde{\sigma}_u^2 = \sigma_*^2 - \frac{\tilde{\sigma}_\epsilon^2}{T}$$

using the estimator of $\tilde{\sigma}_\epsilon^2$ taken from the within regression. Unfortunately, when $\tilde{\sigma}_u^2$ is calculated in this way it can be and often is negative. See Hsiao (1986) and Baltagi (2005) for references and a discussion of alternative procedures.

Maximum likelihood estimation

When u_i and ϵ_{it} are independent and normally distributed, the log-likelihood contribution made by unit i's data is

$$L_i = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\ln \mid \mathbf{V}_i \mid -\frac{1}{2}(\mathbf{Y}_i - \alpha - \mathbf{X}_i\beta - \mathbf{Z}_i'\gamma)'\mathbf{V}_i^{-1}(\mathbf{Y}_i - \alpha - \mathbf{X}_i\beta - \mathbf{Z}_i'\gamma).$$

Substituting $(T-1)\ln\sigma_\epsilon^2 + \ln(T\sigma_u^2 + \sigma_\epsilon^2)$ for $\log \mid \mathbf{V}_i \mid$, using the result given above for $\mathbf{V}_i^{-1}$, and adding across i gives the full log-likelihood function. Hsiao (1986, pp. 38–41) derives the score and the asymptotic covariance matrix of the maximum-likelihood estimator.

31.8 Nonlinear Panel-Data Models

It is not conceptually difficult to expand the scope of panel-data models to include nonlinear equations, provided that the disturbance terms enter the specification either in a multiplicative or additive fashion. We'll first develop the case of a quasi-Poisson count-data model (see Chapter 28 on GMM modelling) with a multiplicative fixed effect u_i such that $\mathrm{E}(u_i|\mathbf{X}_{it}) \neq 0$,

$$Y_{it} = \phi(\mathbf{X}'_{it}\boldsymbol{\theta})u_i + \epsilon_{it}.$$

To keep the notation simple, let $\phi_{it}(\boldsymbol{\theta}) \equiv \phi(\mathbf{X}'_{it}\boldsymbol{\theta})$. A Poisson-like structure would be expressed in the functional form $\phi_{it}(\boldsymbol{\theta}) = e^{\mathbf{X}'_{it}\boldsymbol{\theta}}$ as explained in the GMM chapter. Although they will eventually need some consideration, the functional-form particulars are not important to the main lines of the argument.

First differences For a first-differences approach, let's use the lagged equation $Y_{it-1} = \phi_{it-1}(\boldsymbol{\theta})u_i + \epsilon_{it-1}$ to isolate the u_i fixed effect,

$$u_i = \frac{1}{\phi_{it-1}(\boldsymbol{\theta})}Y_{it-1} - \frac{1}{\phi_{it-1}(\boldsymbol{\theta})}\epsilon_{it-1}$$

and substitute this into the period-t equation. This substitution yields a "quasi-differenced" expression

$$Y_{it} - \frac{\phi_{it}(\boldsymbol{\theta})}{\phi_{it-1}(\boldsymbol{\theta})}Y_{it-1} = \epsilon_{it} - \frac{\phi_{it}(\boldsymbol{\theta})}{\phi_{it-1}(\boldsymbol{\theta})}\epsilon_{it-1}$$

Now that the fixed effect u_i has been eliminated, we have the possibility of using $\mathbf{X}_{it}, \mathbf{X}_{it-1}$ as instruments in a moment specification

$$g_{it}(\boldsymbol{\theta}) = \begin{bmatrix}\mathbf{X}_{it}\\ \mathbf{X}_{it-1}\end{bmatrix}\left(Y_{it} - \frac{\phi_{it}(\boldsymbol{\theta})}{\phi_{it-1}(\boldsymbol{\theta})}Y_{it-1}\right) = \begin{bmatrix}\mathbf{X}_{it}\\ \mathbf{X}_{it-1}\end{bmatrix}\left(\epsilon_{it} - \frac{\phi_{it}(\boldsymbol{\theta})}{\phi_{it-1}(\boldsymbol{\theta})}\epsilon_{it-1}\right).$$

This would be valid assuming that two "weak exoneneity" conditions $\mathrm{E}(\epsilon_{it}|\mathbf{X}_{it}, \mathbf{X}_{it-1}) = 0$ and $\mathrm{E}(\epsilon_{it-1}|\mathbf{X}_{it}, \mathbf{X}_{it-1}) = 0$ both hold. (Conditioning on $\mathbf{X}_{it}$ and $\mathbf{X}_{it-1}$ renders $\phi_{it}(\boldsymbol{\theta})/\phi_{it-1}(\boldsymbol{\theta})$ akin to a constant.) Assuming weak exogeneity thus ensures that the crucial GMM moment condition, $\mathrm{E}\, g_{it}(\boldsymbol{\theta}_0) = \mathbf{0}$ with $\boldsymbol{\theta}_0$ being the true value of $\boldsymbol{\theta}$, will be satisfied.

Deviations from means Another way of eliminating the fixed effect u_i is through subtraction of the i-specific sample means. From

$$\bar{Y}_i = \bar{\phi}_i u_i + \bar{\epsilon}_i$$

we have

$$u_i = \frac{\bar{Y}_i}{\bar{\phi}_i} - \frac{\bar{\epsilon}_i}{\bar{\phi}_i}.$$

Substituting for the u_i fixed effect then yields

$$Y_{it} - \frac{\phi_{it}(\boldsymbol{\theta})}{\bar{\phi}_i(\boldsymbol{\theta})}\bar{Y}_i = \epsilon_{it} - \frac{\phi_{it}(\boldsymbol{\theta})}{\bar{\phi}_i(\boldsymbol{\theta})}\bar{\epsilon}_i$$

The moment condition requirements are quite a bit more demanding in this case. Since all of i's explanatory $\mathbf{X}_{is}$ variables enter $\bar{\phi}_i(\boldsymbol{\theta})$, and likewise all of the ϵ_{is} disturbances enter $\bar{\epsilon}_i$, we would need to invoke the "strong exogeneity" assumption that

$$\mathrm{E}(\epsilon_{is}|\mathbf{X}_{i1}, \mathbf{X}_{i2}, \ldots, \mathbf{X}_{iT}) = 0 \,\forall s \in \{1, 2, \ldots, T_i\}.$$

If this condition holds, then the full set of $\mathbf{X}_{i,s}$ could in principle serve as the instruments.

Of course, most economic models would include in $\mathbf{X}_{i,t}$ some i-specific but time-invariant explanatory variables, which would reduce the total number of instruments available for both the first-differenced and deviations from means approaches. Furthermore, depending on the functional form of $\phi_{it}(\boldsymbol{\theta})$, the coefficients associated with such time-invariant variables might not be identifiable. For example, in the Poisson-like specification $\phi_{it}(\boldsymbol{\theta}) = e^{\mathbf{X}'_{it}\boldsymbol{\theta}}$, the coefficients of the invariant variables would cancel out of the $\phi_{it}(\boldsymbol{\theta})/\phi_{it-1}(\boldsymbol{\theta})$ and $\phi_{it}(\boldsymbol{\theta})/\bar{\phi}_i(\boldsymbol{\theta})$ ratios. For more complicated nonlinear functional forms, however, some or even all such time-invariant variables could conceivably remain in the picture.

The multiplicative models have one disconcerting feature. Although their $\boldsymbol{\theta}$ parameters can be estimated consistently using a GMM algorithm, the impacts of the associated covariates on Y_{it} are not identified in an absolute sense because $Y_{i,t} = \phi_{it}(\boldsymbol{\theta})u_i + \epsilon_{it}$ and the multiplicative fixed effect u_i is never observed. The covariate effects are identified in a relative sense—in the ratio of partial derivatives of $Y_{i,t}$ with respect to two different covariates, the u_i factor would cancel out—and we could therefore tell which of the two has the stronger effect.

For nonlinear models with additive fixed effects, such as the panel-data quasi-probit model

$$Y_{it} = \Phi(\mathbf{X}'_{it}\boldsymbol{\theta}) + u_i + \epsilon_{it}$$

we would proceed very much as in the case of linear panel-data models. With modern software, only minor computational challenges (having to do in this case with the nonlinearities of the $\Phi()$ function) would need to be addressed.

What if neither multiplicative nor additive effects make sense? There is a third approach, termed by Cameron and Trivedi (2005) the "single-index" specification, in which u_i is embedded in the functional form, but it is by far the most difficult. To see the problem, consider this specification:

$$Y_{it} = h(\mathbf{X}'_{i,t}\boldsymbol{\theta} + \mathbf{Z}'_i\gamma + u_i) + \epsilon_{i,t}.$$

Here u_i enters the h function in the same manner as the explanatory covariates. But in this context, u_i cannot be eliminated by either division or subtraction. One approach would be to introduce N dummy variables to stand in for the combination $\mathbf{Z}'_i\gamma + u_i$ as in a linear fixed-effects model. Unfortunately, in marked contrast to the linear case, this approach does not in general allow $\boldsymbol{\theta}$ to be estimated consistently. This is the so-called "incidental parameters" problem, in which as $N \to \infty$, so does the total number of u_i parameters to be estimated. In such pathological cases, inconsistency generally afflicts *all* parameter estimates. (And a number of Monte Carlo studies suggest that the large-sample biases of the dummy-variable approach can be serious.) Also, if u_i is associated with the $\mathbf{X}_{i,t}$ and $\mathbf{Z}_i$ covariates, it cannot be "integrated out" of the model as in random-effects specifications

using Gaussian quadrature (Chapter 36), unless the researcher is somehow able to specify the conditional density $g(u_i \mid \mathbf{X}_{it}, \mathbf{Z}_i)$, which is unlikely.

There are special-case solutions to the incidental parameters problem already available for panel-data logit and Poisson models—we will discuss them in a moment—but no general solutions have yet been devised. However, efforts to move beyond these special cases are being actively pursued. Dhaene and Jochmans (2015) gives an informative review of developments, a number of which (like this paper) follow the theoretical breakthoughs of Hahn and Newey (2004) in striving to find methods to compensate for the large-sample biases. Also see Bergé, Krantz, and McDermott (2021) for methods based on maximum likelihood as well as linear fixed-effect models. This is an exciting and fast-breaking area of current econometric research.

31.9 Fixed-Effects Logit and Poisson Models for Panel Data

As mentioned above, progress has been made on the inclusion of fixed-effects in a few nonlinear models for panel data. Among these, the fixed-effect logit model is most often used, although the techniques also apply to Poisson and negative binomial models. Unfortunately the methods do not extend to probit models, because as we've just mentioned, the probit functional form does not allow the fixed effect to be cancelled out. Hamerle and Ronning (1995) give a very clear exposition of the fixed-effect logit case, which was originally developed by statisticians working in the field of psychology.

Let $\mathbf{y}_i$ be a vector of yes–no outcomes, coded as ones and zeros, for the T periods over which individual i contributes data. Let $\pi_{i,t}$ be the probability of a "yes" outcome for period t, with the dependence on exogenous covariates and coefficients suppressed for the moment. The probability of the sequence coded in $\mathbf{y}_i$ can be written as

$$\begin{aligned}\Pr(\mathbf{y}_i) &= \exp\left(\sum_{t=1}^{T} y_{i,t} \ln \pi_{i,t} + (1 - y_{i,t}) \ln(1 - \pi_{i,t})\right) \\ &= \exp\left(\sum_{t=1}^{T} y_{i,t} \ln\left(\frac{\pi_{i,t}}{(1 - \pi_{i,t})}\right) + \sum_{t=1}^{T} \ln(1 - \pi_{i,t})\right).\end{aligned}$$

In a logit specification,

$$\ln\left(\frac{\pi_{i,t}}{(1 - \pi_{i,t})}\right) = \mathbf{X}'_{i,t}\beta + u_i,$$

where u_i is the fixed effect, and inserting this yields

$$\Pr(\mathbf{y}_i) = \exp\left(\sum_{t=1}^{T} y_{i,t}\mathbf{X}'_{i,t}\beta + u_i \sum_{t=1}^{T} y_{i,t} - \sum_{t=1}^{T} \ln(1 + e^{\mathbf{X}'_{i,t}\beta + u_i})\right). \tag{31.8}$$

Note that the third of the summation terms involves u_i but does not involve $\mathbf{y}_i$.

From this last result, we recognize that $\sum_{t=1}^{T} y_{i,t}$ is a *sufficient statistic* for u_i, and thus by definition, the conditional probability for observation i,

$$\Pr\left(\mathbf{y}_i \mid \sum_{t=1}^{T} y_{i,t}\right),$$

does not depend on the u_i fixed effect. In other words, if we form a *conditional likelihood function* using these conditional probabilities, the fixed effects will not appear and, as it turns out, we can estimate β consistently.

To see the essentials of the argument, consider a two-period example. For a two-period model, conditioning on $\sum_{t=1}^{2} y_{i,t} = 1$ is the only case that we need to consider. This is because conditioning on sums equalling 0 or 2 gives conditional probabilities of one—the $\mathbf{y}_i$ vector would be completely determined by the sufficient statistic. In a practical sense, this means that individuals who always respond "yes" or always respond "no" drop out of the sample used to estimate the β parameters. The fixed-effects method can entail a serious loss in the number of observations, especially in short panels (with small T) in which most individuals do not change their responses.

To begin our analysis of the two-period case, note that the unconditional probability of the sequence $Y_{i,1} = 0, Y_{i,2} = 1$ is

$$\Pr(0,1) = \frac{1}{1 + e^{\mathbf{X}'_{i,1}\beta + u_i}} \cdot \frac{e^{\mathbf{X}'_{i,2}\beta + u_i}}{1 + e^{\mathbf{X}'_{i,2}\beta + u_i}},$$

and similarly,

$$\Pr(1,0) = \frac{e^{\mathbf{X}'_{i,1}\beta + u_i}}{1 + e^{\mathbf{X}'_{i,1}\beta + u_i}} \cdot \frac{1}{1 + e^{\mathbf{X}'_{i,2}\beta + u_i}},$$

From these expressions, we can derive the conditional probability $\Pr(Y_{i,1} = 0, Y_{i,2} = 1 \mid Y_{i,1} + Y_{i,2} = 1)$. Let $\bar{P}_{0,1}$ denote the conditional probability we seek, and write it in terms of the unconditional probabilities

$$\bar{P}_{0,1} = \frac{\Pr(0,1)}{\Pr(0,1) + \Pr(1,0)} = \frac{1}{1 + \frac{\Pr(1,0)}{\Pr(0,1)}}.$$

From above, we have $\Pr(1,0) / \Pr(0,1) = e^{(\mathbf{X}_{i,1} - \mathbf{X}_{i,2})'\beta}$, and substituting yields

$$\bar{P}_{0,1} = \frac{e^{\mathbf{X}'_{i,2}\beta}}{e^{\mathbf{X}'_{i,1}\beta} + e^{\mathbf{X}'_{i,2}\beta}}.$$

Applying the same reasoning gives us the other conditional probability

$$\bar{P}_{1,0} = \frac{e^{\mathbf{X}'_{i,1}\beta}}{e^{\mathbf{X}'_{i,1}\beta} + e^{\mathbf{X}'_{i,2}\beta}}.$$

Remarkably, the β parameter can be estimated using standard logit software.

In the general case of $T > 2$, we condition on $\sum_t y_{i,t} = s$ for all values of s except $s = 0$ and $s = T$. We find the probability that $\sum_{t=1}^{T} y_{i,t} = s$ by summing the probabilities given in equation (31.8) over all possible $\mathbf{y}_i$ vectors that would contain a total of s "yes" values. (As we saw in the $T = 2$ case, observation i provides only one of the vectors that meet this criterion; there are $\binom{T}{s}$ such vectors in total.) We denote by $\{\mathbf{y}_i^* : \sum_t y_{i,t}^* = s\}$ the set of such vectors. To obtain the conditional probability for the $\mathbf{y}_i$ vector actually given in the dataset,

we then divide the unconditional probability (31.8) by this sum, thereby eliminating the fixed effect, and obtain the conditional probability

$$\bar{P}(\mathbf{y}_i \mid \sum_t y_{i,t} = s) = \frac{e^{\sum_t y_{i,t}\mathbf{x}'_{i,t}\beta}}{\sum_{\{\mathbf{y}_i^*:\sum_t y_{i,t}^*=s\}} e^{\sum_t y_{i,t}^*\mathbf{x}'_{i,t}\beta}}.$$

Until fairly recently it was not obvious how to calculate the denominator when T is large and therefore the number of terms given in $\binom{T}{s}$ is also large, but as the STATA manual explains in its discussion of the `clogit` command, a recursive formula has been discovered that greatly eases the computational burden. The fixed-effect logit model for $T > 2$ can be estimated by using an algorithm for McFadden's conditional logit model—discussed in Chapter 21—in which each individual is allowed to have a different choice set.

Essentially the same treatment is available for the Poisson count data model, whose parameters can also be estimated by specifying a conditional likelihood function that does not depend on the u_i fixed effects. The STATA manual gives a concise account of the method in the discussion of the `xtpoisson` command. The negative binomial count data model is another model in which fixed effects can be eliminated via the conditional likelihood approach.

It is important to note that in each of these three cases, the observations $\{Y_{i,t}\}$ are assumed to be independent over t for a given individual i conditional on u_i and $\{\mathbf{X}_{i,t}\}$. The conditional fixed-effect methods are *not* applicable to dynamic specifications. This limitation is being addressed in much current research.

31.10 Dynamic Panel-Data Models

Consider a linear panel data model with a lagged dependent variable and an error-components disturbance term,

$$Y_{it} = \alpha Y_{it-1} + \mathbf{X}'_{it}\beta + u_i + \epsilon_{it}. \tag{31.9}$$

Since the u_i effect is persistent, it is correlated with Y_{it-1}, and the model cannot be estimated consistently by ordinary least squares. What moments can be constructed in a GMM approach that would allow the parameters of this equation to be estimated? It is conventional to assume—we will revisit this important point shortly—that the ϵ_{it} component of the disturbance exhibits no serial correlation, so that the persistence of the composite disturbance is due entirely to its u_i component.

Students: Read Cameron and Trivedi (2005, p. 22.5).

The model has been examined by Anderson and Hsiao (1981), Holtz-Eakin, Newey, and Rosen (1988), Arellano and Bond (1991), and Blundell and Bond (1998). As Roodman (2006) explains in his review of methods, there are two main routes that have been pursued to estimate the structural model, each being guided by the desire to find moment conditions that are *internal* to the model, that is, conditions that follow logically from the model's structure and assumptions about serial correlation. The better-known of the two approaches, due to Arellano and Bond (1991), takes first differences of the structural equation and then makes use of lagged *levels* of Y_{it} in the moment conditions. The lesser-known approach,

developed by Blundell and Bond (1998), returns to the structural model (31.9), which includes the u_i fixed effect, and uses lagged *differences* such as $Y_{it-1} - Y_{it-2}$ as instruments for the right-hand side variable Y_{it-1} of the model.

Some insights into the nature of the identifying moments can be gained by re-writing the structural model as

$$Y_{it} - Y_{it-1} = (\alpha - 1)Y_{it-1} + \mathbf{X}'_{it}\beta + u_i + \epsilon_{it}. \tag{31.10}$$

Equation (31.10) shows how the *level* of Y_i in period $t-1$ affects the *change* in Y_i between that period and the next. This connection will figure prominently in the specification of moment conditions.

The Arellano–Bond method begins by first-differencing the structural equation (31.9), which eliminates the u_i term and gives

$$Y_{it} - Y_{it-1} = \alpha(Y_{it-1} - Y_{it-2}) + (\mathbf{X}_{it} - \mathbf{X}_{it-1})'\beta + \epsilon_{it} - \epsilon_{it-1}. \tag{31.11}$$

Note that the ϵ_{it-1} term of the transformed disturbance is correlated with Y_{it-1}, and so without further manipulation, this equation cannot yet be consistently estimated. However, the moments needed for a GMM approach can be constructed from lagged values of the $\mathbf{X}_{it}$ explanatory variables and also, under certain conditions, from lagged values of Y_{it}, the dependent variable. We should also note that the transformed disturbance term, $\epsilon_{it} - \epsilon_{it-1}$, whose variance is

$$\mathrm{E}(\epsilon_{it} - \epsilon_{it-1})^2 = 2\sigma^2,$$

has negative first-order serial correlation,

$$\mathrm{E}(\epsilon_{it} - \epsilon_{it-1})(\epsilon_{it-1} - \epsilon_{it-2}) = -\sigma^2.$$

No correlation exists in the differenced disturbances beyond this first-order lag.

We will begin our treatment of these models by asking about the role of lagged values of the $\mathbf{X}_{it}$ explanatory variables. Assume that there are k such variables, and further assume that they are weakly exogenous in the sense that $\mathrm{E}\,\mathbf{X}_{it-s}\epsilon_{it} = \mathbf{0}$ for $s \geq 0$. This assumption allows the $\mathbf{X}$ variables to be predetermined; that is, $\mathbf{X}_{it}$ can be influenced by *past* values of the ϵ_{it-s} disturbances, but the assumption rules out contemporaneous correlation as well as correlation between current $\mathbf{X}_{it}$ variables and *future* disturbances.

For the i-th unit of the panel, we can define k moments as follows:

$$g_i^1(\theta) = \sum_{t=2}^{T} \mathbf{X}_{it-1} \cdot \left(Y_{it} - Y_{it-1} - \alpha(Y_{it-1} - Y_{it-2}) - (\mathbf{X}_{it} - \mathbf{X}_{it-1})'\beta\right)$$

with $\theta = (\alpha, \beta)$. These are valid moments because at $\theta = \theta_0$,

$$g_i^1(\theta_0) = \sum_{t=2}^{T} \mathbf{X}_{it-1} \cdot (\epsilon_{it} - \epsilon_{it-1})$$

and $\mathrm{E}\,\mathbf{X}_{it-1} \cdot (\epsilon_{it} - \epsilon_{it-1}) = \mathbf{0}_{k\times 1}$ by the weak exogeneity assumption. The moment vector is expressed in what Roodman terms a "collapsed form," which means that for unit i, the moment is constructed by summing across all relevant t values for that unit. As usual, we

would make use of the sample average $n^{-1}\sum_i g_i^1(\theta)$ in constructing the GMM quadratic form and estimating the θ parameters.

Note that a second vector of k moments is provided by

$$g_i^2(\theta) = \sum_{t=3}^{T} \mathbf{X}_{it-2} \cdot \left(Y_{it} - Y_{it-1} - \alpha(Y_{it-1} - Y_{it-2}) - (\mathbf{X}_{it} - \mathbf{X}_{it-1})'\beta\right)$$

which is also in collapsed form but with the summation index beginning at $t = 3$ in this case. Obviously, even more sets of such moments can be generated with third-order and deeper lags of the $\mathbf{X}$ variables.

Much of the interest in the Arellano–Bond approach has been stimulated by the possibility of using *lagged dependent variables* to construct moment conditions. For example, consider Y_{it-2}, which would enter a "collapsed" moment condition as

$$g_i^3(\theta) = \sum_{t=3}^{T} Y_{it-2} \cdot \left(Y_{it} - Y_{it-1} - \alpha(Y_{it-1} - Y_{it-2}) - (\mathbf{X}_{it} - \mathbf{X}_{it-1})'\beta\right).$$

As written, this expression provides only one moment condition. But as Holtz-Eakin and Rosen point out, we could instead write the conditions in an expanded form:

$$g_i^3(\theta) = Y_{i1} \cdot \left(Y_{i3} - Y_{i2} - \alpha(Y_{i2} - Y_{i1}) - (\mathbf{X}_{i3} - \mathbf{X}_{i2})'\beta\right),$$

$$g_i^4(\theta) = Y_{i2} \cdot \left(Y_{i4} - Y_{i3} - \alpha(Y_{i3} - Y_{i2}) - (\mathbf{X}_{i4} - \mathbf{X}_{i3})'\beta\right),$$

and so on, to

$$g_i^T(\theta) = Y_{iT-2} \cdot \left(Y_{iT} - Y_{iT-1} - \alpha(Y_{iT-1} - Y_{iT-2}) - (\mathbf{X}_{iT} - \mathbf{X}_{iT-1})'\beta\right),$$

this being a total of $T - 2$ moments. As with the $\mathbf{X}$ variables, additional moments can be created using third-order and deeper lags of the Y variables.

Note, however, that whether expressed in collapsed or expanded form, lagged Y values provide valid moment conditions only if the ϵ_{it} disturbances are serially uncorrelated. For example, if ϵ_{it} follows an $AR(1)$ process, then Y_{it-2} will not generally be uncorrelated with the $\epsilon_{it} - \epsilon_{it-1}$ differences. Using lagged Ys in the moment conditions is therefore riskier than using lagged weakly exogenous $\mathbf{X}$ variables.

The alternative approach that we mentioned, which was developed by Blundell and Bond (1998), considers whether the structural equation (31.9) might be estimated using lagged *differences* $Y_{it-1} - Y_{it-2}$ as instruments for the endogenous Y_{it} variable. Is this approach valid? On the one hand, equation (31.11) provides some encouragement, in that it suggests that u_i is eliminated from the $Y_{it-1} - Y_{it-2}$ first difference. But on the other, equation (31.10) generates some uneasiness about the procedure, as it shows that the first difference can also be written in a form in which u_i appears. In seeking to reconcile the two equations and determine conditions under which lagged differences in Y can serve as instruments for Y_{it-1}, Blundell and Bond are led to examine the initial conditions of the $\{Y_{it}\}$ time-series, and work out restrictions under which lagged differences can be used in this way. The details are a bit complicated—they involve assumptions about stationarity of the time-series—and the reader is referred to the papers for further discussion.

31.11 Evaluating Programs with Panel Data

Panel data have become more widely available, and now serve as an important tool in evaluations of the effects of program interventions. In addition to the fixed-effects and first-differencing methods, it is increasingly common for researchers to employ the *difference-in-differences* method for isolating program effects. This method shares some features with the first-differences and fixed-effects approaches, but is at once simpler, more powerful, and more limited in scope than these approaches. To understand the difference-in-differences terminology, let's consider the following illustrative case.

A job-training program is offered to a large group of eligible workers. Some of the workers choose to enroll in the program and others do not. Imagine that we have panel data on all workers that span the period before the program was offered and afterwards. We have in mind a simple structural model of wages, which for worker i at time t is

$$Y_{it} = \alpha + t \cdot \beta + D_{it}\delta + u_i + \epsilon_{it}.$$

This model contains a constant, a time trend, a dummy variable D_{it} which takes the value 1 if the worker has participated in the training program on or before time t, and an error-components disturbance term. If we think of the u_i component as representing the worker's motivation, among other things, we would then suspect that more motivated workers (those with higher values of u_i) are also the kind of people who are more likely to take advantage of opportunities for job training. That is, it seems likely that u_i and D_{it} will be positively correlated.

What about the transitory components ϵ_{it}—are these likely to be correlated with D_{it} as well? They could be correlated if workers experiencing a period or two of unusually low wages (due to low ϵ_{it}) might be motivated to enroll in a training program. In studies of program participation, it is common to see that workers who participate often do so after several rounds of below-normal wages, a phenomenon that has been dubbed the "Ashenfelter dip" after the Princeton labor economist Orley Ashenfelter who was among the first to notice it. In what follows, we will assume that there is no correlation between $\{\epsilon_{it}\}$ and $\{D_{it}\}$, but in a real evaluation of training effects, this would be considered a highly suspect assumption.

We now examine two groups of workers—those who participated in the program and those who did not—and two time periods, defined as values of t before the program was mounted and values during and after the program. If worker i participated in the program, her average wages before the program was offered were

$$\bar{Y}_{i,B} = \alpha + \beta\bar{t}_B + u_i + \bar{\epsilon}_{i,B}$$

with the "B" subscript indicating "before", and her average wages after the program (when, for this worker, $D_{it} = 1$) are

$$\bar{Y}_{i,A} = \alpha + \beta\bar{t}_A + \delta + u_i + \bar{\epsilon}_{i,A},$$

with the "A" subscript indicating "after". The difference in these two averages for a worker who participates in the program is

$$\beta(\bar{t}_A - \bar{t}_B) + \delta + \bar{\epsilon}_{i,A} - \bar{\epsilon}_{i,B}.$$

Note that u_i has disappeared and, given our earlier assumption ruling out "Ashenfelter dips", the expected value of $\bar{\epsilon}_{i,A} - \bar{\epsilon}_{i,B}$ conditional on participation is zero. Even so, the after–before difference clearly does not identify the program effect δ as such but rather the combined effects of the time trend and the program. However, if we now look to the workers who did not participate in job training and for them construct the difference in average wages between the period after and the period before the program was offered, we obtain

$$\beta(\bar{t}_A - \bar{t}_B) + \bar{\epsilon}_{j,A} - \bar{\epsilon}_{j,B}.$$

Assume as before that the expected value of the differenced average disturbances, conditional this time on non-participation, is zero. Then if we subtract this expression for non-participants from its counterpart for participants, we see that we obtain an unbiased estimate of δ, the effect of the program. This way of thinking about the problem explains the "difference-in-differences" terminology.

The method is disarmingly simple. Note that the time trend specification could have been expressed in very general terms as a $\phi(t, \theta)$ function dependent on unknown θ parameters—in fact, we could have inserted two trend functions $\phi_B(t, \theta)$ and $\phi_A(t, \theta)$ for the "before" and "after" periods without any difficulty. The difference-in-differences method is narrowly focused on the program effect δ, and the final difference eliminates time trends altogether, leaving no particular need to estimate them.

However—and this is an important provision and limitation of the method—the difference-in-differences method requires the pre-program and post-program averages of $\phi(t, \theta)$ to be *the same for program participants and non-participants*, as otherwise they would not be subtracted away in the final step. Likewise, *the method does not handle time-varying covariates whose before-program and after-program averages differ for the two groups*. Much research is currently underway to understand how to incorporate such covariates without sacrificing the method's attractive simplicity.

If you have an opportunity to apply this interesting method, please remember the warning about naively ignoring the "Ashenfelter dip". The unadorned difference-in-differences method will not give us consistent estimates of δ if there is correlation between the transitory error components ϵ_{it} and D_{it}. In such cases, differencing would need to be combined with the use of instrumental variables to identify the program effect.

Regression discontinuity designs

Let i index individuals, c sampling clusters, and t time, and denote by $Y_{i,c,t}$ an outcome measure specific to an individual in a given cluster at a point in time. There are two time points under consideration, with $t = 0$ representing the baseline and $t = 1$ the endline of the study period. The program indicator takes the value $T_{c,1} = 1$ at endline for the clusters that receive the treatment, and $T_{c,t} = 0$ otherwise.

Consider a situation in which a program is placed in a cluster according to P_c, the percentage of the cluster's population that is poor. The design is such that clusters with $P_c < \bar{P}$ receive no program whereas clusters with $P_c \geq \bar{P}$ are assigned a program. The researcher is assumed to know each cluster's percentage poor and also knows the level of the $\bar{P}$ threshold that determines program placement. This is what is termed a *sharp regression*

discontinuity design, a method discussed by Cameron and Trivedi (2005, p. 25.6) in their illuminating chapter on evaluating treatment effects.

We specify the structural model of interest at the level of individuals, as

$$Y_{i,c,t} = X'_{i,c,t}\beta + T_{c,t}\delta + u_c + \epsilon_{i,c,t}, \tag{31.12}$$

in which $X_{i,c,t}$ includes a set of explanatory variables measured at the individual and cluster levels at time t. The unobserved disturbances of the model are expressed in an error-components form, with u_c representing all unmeasured time-invariant features of the cluster and $\epsilon_{i,c,t}$ representing other unmeasured characteristics that differ by individual, cluster, and time period. To secure consistent estimates of the parameters of the structural model—including δ, the treatment effect—we must take into account the possibility of correlation between $T_{c,t}$ and the composite disturbance. This correlation may stem from any number of direct or indirect connections between these unobserved disturbances and the composition of the cluster's population as indicated in its below-poverty percentage P_c. Thinking of the problem in terms of the joint distribution of (P_c, u_c), we expect that $\mathrm{E}(u_c \mid P_c > \bar{P}) \neq 0$. There may also be connections between P_c and the $\epsilon_{i,c,t}$ disturbances.

There are several approaches that can be pursued in an effort to address the problem of correlation. A simple implementation of the regression discontinuity (RD) design would assume that $\mathrm{E}\left(u_c + \epsilon_{i,c,t} \mid X_{i,t,c}, P_c\right) = \phi(P_c)$ for clusters with P_c values near the selection thresholds. The exact form of the $\phi(P_c)$ function is unknown, although it is assumed to be continuous. For districts that are near the $\bar{P}$ threshold, $\phi(\cdot)$ can be approximated by a *control function* $\hat{\phi}(P_c)$; in the literature this approximation is usually expressed as the polynomial (hence continuous) produced by a Taylor expansion of the unknown ϕ around the eligibility threshold $\bar{P}$, giving a series of terms and their associated α parameters that we can abbreviate here as $\hat{\phi}(\alpha)$. In this way we would arrive at the augmented RD specification

$$Y_{i,c,t} = X'_{i,c,t}\beta + T_{c,t}\delta + \hat{\phi}(\alpha) + w_{i,c,t}, \tag{31.13}$$

in which $w_{i,c,t}$ denotes the original disturbance now purged of correlation with the program treatment $T_{c,t}$ through the introduction of the $\hat{\phi}(\alpha)$ control function into the model.

Note carefully the role of one key assumption, that $\mathrm{E}\left(u_c + \epsilon_{i,c,t} \mid X_{i,t,c}, P_c\right) = \phi(P_c)$ is not a function of the $X_{i,t,c}$ variables. If it were a function of them, and we proceeded to approximate the control function by Taylor-expanding with respect to $X_{i,t,c}$ as well as P_c, we would not be able to distinguish the direct effects of $X_{i,t,c}$ that operate through β from the indirect effects expressed through the control function. The β parameters would not be identified. The introduction of a control function does not by itself magically solve the problem of correlation of the program $T_{c,t}$ and the composite disturbance—it helps us only if we are comfortable making additional assumptions about which variables can be excluded from the control function.

Other approaches to evaluating treatment effects build on these basic ideas. Suppose that at endline, the researchers revisit the clusters that were sampled at baseline, a design that provides a panel dataset at the cluster level (although not at the individual level). Armed with data such as these, we can re-express the $Y_{i,c,t}$, $X_{i,c,t}$, and $T_{c,t}$ variables of the structural model in terms of deviations from cluster averages, yielding an equation

$$Y_{i,c,t} - \bar{Y}_c = \left(X_{i,c,t} - \bar{X}_c\right)'\beta + (T_{c,t} - \bar{T}_c)\delta + \epsilon_{i,c,t} - \bar{\epsilon}_c \tag{31.14}$$

from which the cluster effect u_c has disappeared (as have any other time-invariant $X_{i,c,t}$ variables along with their associated β coefficients). The great advantage of this data transformation is that it removes from the scene the cluster error component u_c that is arguably most likely to be correlated with the P_c areal poverty measures that decide program placement. If u_c is indeed the only component of the composite disturbance that is so correlated, the parameters of equation (31.14) could estimated consistently by applying ordinary least squares to the transformed data. Econometricians would usually describe this as a model with cluster-specific fixed effects, whose parameters can be consistently estimated by transforming the data into deviations from cluster means.

In fact is not necessary to assume that $\epsilon_{i,c,t}$ is uncorrelated with P_c. Letting

$$\lambda(P_c) = \mathrm{E}\left(\epsilon_{i,c,t} - \bar{\epsilon}_c \mid X_{i,c,t} - \bar{X}_c, P_c\right),$$

we can insert the control function $\hat{\lambda}(\gamma)$, a polynomial approximation to $\lambda(\cdot)$ with associated parameters γ, into the transformed equation,

$$Y_{i,c,t} - \bar{Y}_c = \left(X_{i,c,t} - \bar{X}_c\right)' \beta + (T_{c,t} - \bar{T}_c)\delta + \hat{\lambda}(\gamma) + v_{i,c,t}. \tag{31.15}$$

Equation (31.15) provides us with a general and potentially very useful evaluation tool.

Chapter 32

Censoring and Sample Selection Models

This chapter presents results for regression-like models in which the dependent variable is subject to censoring or is generated by a sample selection mechanism such as often seen in labor and health economics. We begin our discussion with an examination of endogenous explanatory variables in probit models. The techniques used in this case are of interest in their own right, and are also of use in the censored-data and sample-selection contexts.

32.1 Probit with Endogenous Variables

It is not uncommon for a probit (or logit) equation to be specified with endogenous right-hand side variables, especially when the equation is part of a larger equation system. If it is possible to specify the likelihood function for the full system, then the probit equation parameters and their standard errors can be estimated consistently by the maximum likelihood method. But if the system is large or complex, specifying and estimating the full model can present daunting difficulties. It can be helpful to have a simple approach that provides estimates of the probit parameters.

Rivers and Vuong (1988) develop a two-step estimator for the probit case, following the lead of Smith and Blundell (1986) for the Tobit model. They take what is essentially a "control function" approach. The Rivers–Vuong method is applicable when Y_1, the right-hand side endogenous variable of the probit equation, is generated by a simple linear regression model. A full treatment of the approach requires results that we will introduce in Chapter 28, and here we will set out only the basics. Recall that in previous chapters we have explored GMM alternatives for probit-like instrumental variables models; those model also deserve consideration when the probit equation has endogenous right-hand side explanatory variables.

Consider a two-equation system in which one of the dependent variables, Y_1, is potentially correlated with ϵ, the disturbance term of the probit equation. Writing out the system

for one observation, and omitting the i subscript, we have

$$Y_1 = \mathbf{Z}'\pi + u \tag{32.1}$$
$$Y_2^{**} = \mathbf{X}'\tilde{\boldsymbol{\beta}} + Y_1\tilde{\delta} + \epsilon. \tag{32.2}$$

Here Y_2^{**} is a latent dependent variable with Y_2 the binary (yes/no) observed outcome. The $\mathbf{Z}$ covariates must include at least one variable that is excluded from $\mathbf{X}$ and which therefore functions as an instrument. We assume that u and ϵ are mean zero and jointly normally distributed. In STATA, this model is available through the `ivprobit` command. R has several implementations as well.

If the disturbances are assumed to be joint normal, this implies (recall Chapter 2) that

$$\epsilon = \frac{\sigma_{u\epsilon}}{\sigma_{uu}}u + w,$$

with $w \sim \mathcal{N}(0, \sigma_{ww})$ and $\sigma_{ww} = \sigma_{\epsilon\epsilon} - \sigma_{u\epsilon}^2/\sigma_{uu}$. (Sometimes you will see the equivalent expression $\sigma_{ww} = \sigma_{\epsilon\epsilon}(1) - \rho^2$ in which ρ is the correlation of the u and ϵ disturbances.) Also, w is independent of u. Substituting into equation (32.2) and dividing through by σ_w yields

$$Y_2^* = \mathbf{X}'\beta + Y_1\delta + \lambda u + v,$$

in which $v = w/\sigma_w$ is standard normal and independent of u, and $\lambda = \sigma_{u\epsilon}/(\sigma_{uu}\sigma_w)$. The original two-equation system is thus re-expressed as

$$Y_1 = \mathbf{Z}'\pi + u \tag{32.3}$$
$$Y_2^* = \mathbf{X}'\beta + Y_1\delta + \lambda u + v. \tag{32.4}$$

The parameters of the first equation are π and σ_{uu}, and (β, δ, λ) are the parameters of the latent-variable probit equation. Please take note of the re-scaling of the probit parameters relative to the original specification.

We might now proceed to estimate the system as a whole. Write the joint density of the disturbances as $f(u, v) = g(u)h(v|u)$ and, exploiting the linearity of the system's first equation, substitute $Y_1 - \mathbf{Z}'\pi$ for u. The likelihood contribution made by an observation is then

$$g(Y_1 - \mathbf{Z}'\pi) \cdot \big(\Phi(\mathbf{X}'\beta + Y_1\delta + \lambda(Y_1 - \mathbf{Z}'\pi))\big)^{Y_2} \cdot \big(1 - \Phi(\mathbf{X}'\beta + Y_1\delta + \lambda(Y_1 - \mathbf{Z}'\pi))\big)^{1-Y_2}, \tag{32.5}$$

where the first factor $g(\cdot)$ corresponds to the marginal $\mathcal{N}(0, \sigma_{uu})$ density for the Y_1 equation and the second two factors derive from the conditional density $h(v|u)$. Owing to the conditioning on u and the independence of v from u and thus Y_1, the second and third factors closely resemble what we would see in an ordinary probit likelihood function.

To find the parameters that maximize the full-system log-likelihood, we would calculate derivatives with respect to σ_{uu} (given the definition of λ, this derivative would involve only $g(\cdot)$, the first factor), the π vector (all three factors would figure into these derivatives), and finally the derivatives with respect to $\boldsymbol{\beta}$, δ and λ. Both R and STATA provide means

for specifying non-standard likelihood functions, and have options to accept code for the derivatives as well. In my view this is the preferred approach.

Rivers and Vuong (1988) suggest an alternative estimation procedure that can be executed in two easy steps. In their method, equation (32.3) is estimated separately and the residual from this equation, $\hat{u} = Y_1 - \mathbf{Z}'\hat{\pi}$, is inserted into equation (32.4) where u appears. The $\boldsymbol{\beta}$, δ and λ parameters are then estimated by an ordinary probit ML algorithm, just as if $\hat{u}$ were an explanatory variable. The estimators from this second step, ($\hat{\boldsymbol{\beta}}, \hat{\delta}$ and $\hat{\lambda}$), can be shown to be consistent. Hence, the estimation task is far simpler than what would be confronted in full system estimation.

Unfortunately, this simplification comes at considerable cost. In most applications we will want estimates of standard errors in addition to consistent parameter estimates, and the two-step method does not give a correctly-formed variance matrix for $\hat{\boldsymbol{\beta}}, \hat{\delta},$ and $\hat{\lambda}$. To test hypotheses, therefore, some extra programming is needed to correct the standard errors. The modifications required are spelled out in Chapter 28; to implement them you would need access to software (such as R and STATA) that allows easy multiplication and inversion of matrices.[1] Although the standard error corrections are in principle straightforward, the additional programming they entail means that the computational burden of the two-step approach is not much lighter than that of the full-system ML approach. In the end, the simplicity promised by the two-step approach proves to be something of an illusion.[2]

Wooldridge (2010, pp. 586–587) has a nice treatment of the Rivers–Vuong method, and he draws out one additional feature of this method that is worth mentioning here. Although the estimated standard error of the second-step estimator $\hat{\lambda}$ needs to be corrected for most purposes, it is valid as it stands for the null hypothesis $H_0 : \lambda = 0$. In essence, this null hypothesis posits that Y_1 is independent of the disturbance term ϵ of the original probit equation, so that a test of the null is a test of the exogeneity of Y_1. The result of this easy test should indicate whether there is any need to implement the full-system model.

Wooldridge (2010) goes on to consider a more elaborate case in which both equations of the system take a probit form but with the observed Y_1 entering the second equation as an endogenous binary covariate. The probit form of the first equation disallows the substitution of $Y_1 - \mathbf{Z}'\pi$ for u that we used earlier. For the case of two probit equations, Wooldridge therefore advocates estimation of the full system.

32.2 The Tobit Model

The Tobit model is often used with partially-observed dependent variables. Recall that in the probit case, the magnitude of the latent dependent variable is never known; only its sign is observed. In the Tobit case more information than this is available. Given the latent variable structural equation

$$Y_i^* = \mathbf{X}_i'\beta + \epsilon_i \tag{32.6}$$

[1] To apply the formulas presented in Chapter 28 to the case at hand, take the α parameter of that chapter's discussion to be (π, σ_{uu}) and take the β parameter to denote the vector of parameters $(\boldsymbol{\beta}, \delta, \lambda)$ in the case we have been discussing.

[2] Cameron and Trivedi (2005) suggest finding the standard errors of the second-step model by the method of bootstrapping, which might well be easier than applying the analytic corrections we have discussed.

with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, we observe $Y_i = Y_i^*$ if $Y_i^* > 0$ although $Y_i = 0$ otherwise. The logic surrounding this model is very little changed if the observability threshold is something other than 0, so long as the threshold point is known, and modifications to handle the case $Y_i = Y_i^*$ if $Y_i^* < \tau$ are also straightforward. We often encounter the latter case with top-coded income or wealth data, where the actual value of the dependent variable is suppressed when it exceeds the threshold point τ and only the threshold value is reported in the dataset.

Partially-observed dependent variables such as these are said to be *censored*. What problems stem from this censoring? To understand the issues, we will need the following results. Let $X \sim \mathcal{N}(0, 1)$. Simple integration shows that

$$\begin{aligned} \mathrm{E}(X \mid X > k) &= \frac{\phi(k)}{1 - \Phi(k)} > 0, \\ \mathrm{E}(X \mid X < k) &= -\frac{\phi(k)}{\Phi(k)} < 0. \end{aligned}$$

Moreover, see Johnson and Kotz (1970a, Chapter 13, pages 81–83) or Maddala (1983, pages 365–366),

$$\begin{aligned} \mathrm{Var}(X \mid X > k) &= 1 + k\,\frac{\phi(k)}{1 - \Phi(k)} - \left(\frac{\phi(k)}{1 - \Phi(k)}\right)^2, \\ \mathrm{Var}(X \mid X < k) &= 1 - k\,\frac{\phi(k)}{\Phi(k)} - \left(-\frac{\phi(k)}{\Phi(k)}\right)^2. \end{aligned}$$

These expressions can be compactly written as

$$\mathrm{Var}(X \mid k) = 1 + k\lambda(k) - \lambda(k)^2$$

with $\lambda(k) \equiv \phi(k)/(1 - \Phi(k))$ for $X > k$ and $-\phi(k)/\Phi(k)$ for $X < k$. With $X \sim \mathcal{N}(\mu, \sigma^2)$ the expressions above become

$$\begin{aligned} \mathrm{E}(X \mid X > k) &= \mu + \sigma\,\frac{\phi(k^*)}{1 - \Phi(k^*)} \\ \mathrm{E}(X \mid X < k) &= \mu - \sigma\,\frac{\phi(k^*)}{\Phi(k^*)} \\ \mathrm{Var}(X \mid X > k) &= \sigma^2\left[1 + k^*\lambda(k^*) - \lambda(k^*)^2\right] \end{aligned}$$

in which $k^* = (k - \mu)/\sigma$ and $\lambda(k^*)$ is defined as in the $\mathcal{N}(0, 1)$ case.

With this as background, we now return to the structural equation (32.6). Where estimation is concerned, why can't we do the common-sense thing with this equation, using the cases with $Y_i = Y_i^*$ to estimate the β parameters? These are, after all, the non-censored cases for which the actual values of Y_i^* are known.

To appreciate the difficulties that would ensue, consider the expected value of Y_i in such a sub-sample:

$$\begin{aligned} \mathrm{E}[Y_i \mid Y_i^* > 0] &= \mathbf{X}_i'\beta + \mathrm{E}[\epsilon_i \mid Y_i^* > 0] \\ &= \mathbf{X}_i'\beta + \mathrm{E}[\epsilon_i \mid \epsilon_i > -\mathbf{X}_i'\beta]. \end{aligned}$$

Note that $\mathrm{E}[\epsilon_i \mid \epsilon_i > -\mathbf{X}_i'\beta] > 0$, and furthermore the value of this expectation varies systematically with $\mathbf{X}_i$. Suppose, for instance, that $\beta_k > 0$. Then an increase in X_{ik} implies that $-\mathbf{X}_i'\beta$ becomes more negative; this, in turn, implies that the expected value of ϵ_i in the sub-sample decreases—i.e., ϵ_i and X_{ik} are negatively correlated. Obviously, when a correlation such as this is present, least-square estimates of β have no desirable statistical properties.

Now return to our regression problem and consider $\mathrm{E}[\epsilon_i \mid \epsilon_i > -\mathbf{X}_i'\beta]$ in the case where $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Convert ϵ to standard normal using $\sigma \cdot \frac{\epsilon}{\sigma}$; then

$$\begin{aligned}
\mathrm{E}[\epsilon_i \mid \epsilon_i > -\mathbf{X}_i'\beta] &= \sigma \cdot \mathrm{E}\left[\frac{\epsilon_i}{\sigma} \mid \frac{\epsilon_i}{\sigma} > \frac{-\mathbf{X}_i'\beta}{\sigma}\right] \\
&= \sigma \cdot \frac{\phi\left(\frac{-\mathbf{X}_i'\beta}{\sigma}\right)}{1 - \Phi\left(\frac{-\mathbf{X}_i'\beta}{\sigma}\right)} \\
&= \sigma \cdot \lambda\left(\frac{-\mathbf{X}_i'\beta}{\sigma}\right).
\end{aligned}$$

The $\lambda()$ function is often called the "inverse Mills ratio." Also

$$\begin{aligned}
\mathrm{Var}[\epsilon_i \mid \epsilon_i \geq -\mathbf{X}_i'\beta] &= \sigma^2 \,\mathrm{Var}\left[\frac{\epsilon_i}{\sigma} \mid \frac{\epsilon_i}{\sigma} \geq \frac{-\mathbf{X}_i'\beta}{\sigma}\right] \\
&= \sigma^2\left[1 + \left(\frac{-\mathbf{X}_i'\beta}{\sigma}\right)\lambda\left(\frac{-\mathbf{X}_i'\beta}{\sigma}\right) - \lambda\left(\frac{-\mathbf{X}_i'\beta}{\sigma}\right)^2\right].
\end{aligned}$$

One implication is that in the sub-sample with $Y_i = Y_i^* > 0$, the appropriate regression model would be *nonlinear* in $\mathbf{X}$:

$$Y_i = \mathbf{X}_i'\beta + \sigma \cdot \lambda\left(\frac{-\mathbf{X}_i'\beta}{\sigma}\right) + u_i.$$

The disturbance term u_i of this regression is mean zero and uncorrelated with $\mathbf{X}_i$ by construction. However, it is also heteroskedastic.

One can estimate the model using nonlinear regression methods, taking the heteroskedasticity of u_i into account, but a full maximum-likelihood approach is preferable. Consider the i-th person's contribution to the log-likelihood. Let $D_i = 1$ if $Y_i = Y_i^*$, and let $D_i = 0$ otherwise. Then the i-th person's contribution to the log-likelihood is

$$L_i = (1 - D_i) \cdot \log\left[\Phi\left(\frac{-\mathbf{X}_i'\beta}{\sigma}\right)\right] + D_i \cdot \log\left[\frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}(Y_i - \mathbf{X}_i'\beta)^2\right)\right]$$

See Greene (2003) and especially Cameron and Trivedi (2005) for further discussion.

32.3 Heckman Sample Selection Models

Let the structural equations of interest be

$$Y_1^* = \mathbf{Z}'\gamma + \epsilon_1 \tag{32.7}$$

$$Y_2 = \mathbf{X}'\beta + \epsilon_2 \tag{32.8}$$

and let Y_2 be observed only on the condition that $Y_1^* > 0$. Further, let

$$\epsilon_1,\ \epsilon_2 \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right).$$

As we discuss below, a system like this might arise in a study of women's labor market participation and wages, in which the first equation is a probit-like specification with a latent variable indexing the net utility from participating in the labor market, and the second equation provides a specification of wage rates among those women who participate.

Note that

$$\mathrm{E}(Y_2 \mid Y_1^* > 0) = \mathbf{X}'\beta + \mathrm{E}(\epsilon_2 \mid \epsilon_1 > -\mathbf{Z}'\gamma),$$

and since

$$\epsilon_2 = \frac{\sigma_{12}}{\sigma_{11}}\,\epsilon_1 + v,$$

we see that

$$\begin{aligned}
\mathrm{E}(\epsilon_2 \mid \epsilon_1 > -\mathbf{Z}'\gamma) &= \frac{\sigma_{12}}{\sigma_{11}} \cdot \mathrm{E}(\epsilon_1 \mid \epsilon_1 > -\mathbf{Z}'\gamma) \\
&= \frac{\sigma_{12}}{\sigma_{11}} \cdot \sigma_1 \, \frac{\phi(-\mathbf{Z}'\gamma/\sigma_1)}{1 - \Phi(-\mathbf{Z}'\gamma/\sigma_1)} \\
&= \frac{\sigma_{12}}{\sigma_1} \cdot \lambda\left(\frac{-\mathbf{Z}'\gamma}{\sigma_1}\right) \\
&= \rho\sigma_2 \cdot \lambda\left(\frac{-\mathbf{Z}'\gamma}{\sigma_1}\right).
\end{aligned}$$

Evidently, if the β parameters were estimated without reference to the selection rule governing the observability of Y_2, the degree of bias would depend on the σ_{12} parameter (there is no bias if $\sigma_{12} = 0$) and the direction of the bias in β would depend on the correlation between $\mathbf{X}$ and $\mathbf{Z}$.

Much as in the Tobit case, we could estimate this model as a nonlinear regression using the sub-sample in which Y_2 is observed:

$$Y_2 = \mathbf{X}'\beta + \frac{\sigma_{12}}{\sigma_1} \cdot \lambda\left(\frac{-\mathbf{Z}'\gamma}{\sigma_1}\right) + \epsilon^*$$

or two-step methods could be used to estimate $\lambda()$. These techniques dominated the early literature on sample selection methods. Today, however, the full maximum-likelihood approach is generally preferred.[3]

Deriving the likelihood function

There are two types of events to which we want to attach probability expressions. If $Y_1^* \leq 0$ then Y_2 is not observed; the relevant probability is simply $\Phi(-\mathbf{Z}'\gamma/\sigma_1)$. If $Y_1^* > 0$ and Y_2 is

[3]Even if it is not used as a basis for estimation, the nonlinear regression representation is useful in showing how parameter identification is achieved in these models, whether by exclusion restrictions involving $\mathbf{X}$ and $\mathbf{Z}$, or via the nonlinearities implicit in the $\lambda()$ function.

observed, however, we face a more complicated derivation. We want to find an expression for $\Pr(Y_1^* > 0 \text{ and } Y_2 = y_2 \mid \mathbf{X}, \mathbf{Z})$ or what is equivalent,

$$\Pr(\epsilon_1 > -\mathbf{Z}'\gamma \text{ and } \epsilon_2 = y_2 - \mathbf{X}'\beta).$$

Let $f(\epsilon_1, \epsilon_2)$ be the joint normal density of (ϵ_1, ϵ_2). The probability expression we seek can be represented as

$$\int_{-\mathbf{Z}'\gamma}^{\infty} f(\epsilon_1, y_2 - \mathbf{X}'\beta)\, d\epsilon_1.$$

Much as in our earlier treatment of probit models with endogenous variables, it is helpful here to factor the joint density f into the product of a conditional and a marginal density, $f(\epsilon_1, \epsilon_2) = h(\epsilon_1 \mid \epsilon_2)g(\epsilon_2)$. This simplifies the probability expression to

$$g(\epsilon_2) \int_{-\mathbf{Z}'\gamma}^{\infty} h(\epsilon_1 \mid \epsilon_2)\, d\epsilon_1.$$

Now, given that

$$\epsilon_1 \mid \epsilon_2 \sim \mathcal{N}\left(\frac{\sigma_{12}}{\sigma_{22}}\epsilon_2,\ \sigma_{11}\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right)\right),$$

we have

$$\begin{aligned}
\Pr(\epsilon_1 > -\mathbf{Z}'\gamma \mid \epsilon_2) &= \Pr\left(\frac{\epsilon_1 - \frac{\sigma_{12}}{\sigma_{22}}\epsilon_2}{\sqrt{\sigma_{11}\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right)}} > \frac{-\mathbf{Z}'\gamma - \frac{\sigma_{12}}{\sigma_{22}}\epsilon_2}{\sqrt{\sigma_{11}\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right)}}\right) \\
&= 1 - \Phi\left(\frac{-\mathbf{Z}'\gamma - \frac{\sigma_{12}}{\sigma_{22}}\epsilon_2}{\sqrt{\sigma_{11}\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right)}}\right) \\
&= 1 - \Phi\left(\frac{-\mathbf{Z}'\gamma - \frac{\sigma_{12}}{\sigma_{22}}(y_2 - \mathbf{X}'\beta)}{\sqrt{\sigma_{11}\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right)}}\right).
\end{aligned}$$

The likelihood contribution when Y_2 is observed, that is, when $\epsilon_1 > -\mathbf{Z}'\gamma$ and $\epsilon_2 = y_2 - \mathbf{X}'\beta$, is therefore

$$1 - \Phi\left(\frac{-\mathbf{Z}'\gamma - \frac{\sigma_{12}}{\sigma_{22}}(y_2 - \mathbf{X}'\beta)}{\sqrt{\sigma_{11}\left(1 - \frac{\sigma_{12}^2}{\sigma_{11}\sigma_{22}}\right)}}\right) \cdot \phi\left(\frac{y_2 - \mathbf{X}'\beta}{\sigma_2}\right)$$

Note that in some models σ_{11} is not identified and can be normalized to unity.

Example: Estimating Wage Equations

Let

$$W_O = \mathbf{X}_O\beta_O + \epsilon_O$$

represent the offered wage (market wage) as function of individual covariates X, and let

$$W_R = \mathbf{X}_R\beta_R + \epsilon_R$$

represent the reservation wage, which is the minimum wage level necessary to induce participation. W_R is *not* observed.

The labor market participation rule is to participate if $W_O > W_R$, that is, if

$$Y_1^* \equiv W_O - W_R = \mathbf{X}_O\beta_O - \mathbf{X}_R\beta_R + \epsilon_O - \epsilon_R > 0$$

Let $\epsilon_O - \epsilon_R \equiv \epsilon_1$, and let $\epsilon_O \equiv \epsilon_2$. Then we have

$$\begin{aligned}\sigma_{11} &= \sigma_{OO} - 2\,\sigma_{OR} + \sigma_{RR}\\ \sigma_{12} &= \sigma_{OO} - \sigma_{OR}\\ \sigma_{22} &= \sigma_{OO}\end{aligned}$$

Given this, we use the framework outlined above to derive the likelihood function.

32.4 Extensions for Panel Data

Although the methods are somewhat more complicated than those described above, it is possible to extend the sample-selection models to the panel-data context. Vella and Verbeek (1999) explain how to proceed, and Jensen, Rosholm, and Verner (2001) provide a highly informative review and Monte Carlo study of the various panel data estimators.

Chapter 33

Program Evaluation: Selection Models, RCTs, and Propensity Scores

With the aid of the techniques explained in earlier chapters, you should now be well-equipped to explore program evaluation. The sections that follow begin by situating an evaluation problem in a Heckman switching regression model, a full maximum-likelihood approach to evaluation that requires the kind of strong distributional assumptions that have fallen somewhat out of fashion. Since the 1980s when the Heckman approach was last dominant, the econometrics literature has looked with increasing favor on the use of randomized experiments—often termed "randomized control trials" or RCTs—to evaluate the effects of programs. An enormous literature, both in and outside economics, has addressed the pros and cons of true experiments, with skeptics emphasizing how difficult it can be to achieve randomization in the field. Where randomization is not possible, a third approach, mainly involving the use of *propensity scores* with or without *matching*, has gained many adherents. This approach is less general in important ways than the Heckman approach—it relies on the strong assumption of *ignorability* in the way that individuals are selected into programs—but in other ways is both more general and (depending on how it is implemented) less difficult computationally. The founding paper of this literature is Rosenbaum and Rubin (1983), although Heckman and his colleagues have also made (and continue to make) important contributions. The propensity-score methods have spawned a literature that by now rivals the experiments literature in size and scope. I can offer here no more than an introduction to this vast body of material; see Cameron and Trivedi (2005, Chapter 25) and Wooldridge (2010, Chapter 21) for deeper treatments.

The program evaluation literature has its own specialized vocabulary and core concepts, which have arisen from the effort to re-think the problem of causal inference starting from first principles. To see the issues, we will consider a program intervention in the labor market aimed at re-training workers so that they can command higher wages. If person i participates in the intervention, this is recorded in the dummy variable $D_i = 1$ with non-participants having $D_i = 0$. Let $Y_{i,0}$ represent the value of the outcome variable (i.e., future wages) for persons who do not enroll in the program, and let $Y_{i,1}$ represent the outcome variable for those who do enroll (i.e., participate). Since any given person either does or does not enter the program, only one of the pair of random variables $(Y_{i,0}, Y_{i,1})$ is ever observed:

$Y_{i,0}$ for those with $D_i = 0$ and $Y_{i,1}$ for those with $D_i = 1$. The other, unobserved variable is sometimes referred to as the "counterfactual," indicating what wages a non-participant would have earned had that person instead participated in the program, or the wages a participant would have earned had he or she not participated. So the pair of random variables $(Y_{i,0}, Y_{i,1})$ involves one random variable that always goes unobserved for person i.

At a conceptual level, we might regard the *program effect for person i* as being $Y_{i,1} - Y_{i,0}$. The fundamental challenge for statistical inference is to find conditions under which the *expected program effect* $\mathrm{E}(Y_{i,1} - Y_{i,0})$ can be estimated, despite the fact that the effect for person i can never be measured. Since all we can observe is $Y_i = D_i \cdot Y_{1,i} + (1 - D_i) \cdot Y_{0,i}$, we must somehow use this single observable outcome variable along with $\mathbf{X}_i$ to estimate $\mathrm{E}(Y_{i,1} - Y_{i,0})$, the difference between two means. Often it will be necessary to first estimate program effects *conditional on covariates* $\mathbf{X}_i$ and then integrate over $\mathbf{X}$ (or do the equivalent empirically) to determine the unconditional expectation.

In this chapter, I will let

$$\tau(\mathbf{x}) = \mathrm{E}\left((Y_{i,1} - Y_{i,0}) \mid \mathbf{X}_i = \mathbf{x}\right)$$

denote the expected program effect conditional on the value of $\mathbf{X}_i$, and let

$$\tau = \mathrm{E}_{\mathbf{X}}\, \tau(\mathbf{X})$$

be the unconditional expected effect. Note that both $\tau(\mathbf{X})$ and τ refer only to the expected program effect—the actual person-level program effect is allowed to vary from one person to the next even with covariates $\mathbf{X}_i$ held constant.

33.1 Heckman-type selection models

Let the equation $Y_{0,i} = \mathbf{X}_i'\beta_0 + \epsilon_{0,i}$ indicate how observed covariates $\mathbf{X}_i$, parameters β_0, and a disturbance $\epsilon_{0,i}$ would affect the person-specific outcome variable if there is no program participation, and let $Y_{1,i} = \mathbf{X}_i'\beta_1 + \epsilon_{1,i}$ give the person-specific effect with participation. We'll assume that the worker considering training knows both $Y_{0,i}$ and $Y_{1,i}$ and chooses to participate in the program if $Y_{1,i} \geq Y_{0,i}$. Note that in a real application, $Y_{0,i}$ and $Y_{1,i}$ would represent worker-specific forecasts of the outcomes and we would set actual outcomes equal to those that were forecast plus a forecast error. But for the sake of exposition, we will continue with the example as it is.

In this simplified model, the unobservable person-specific program effect is $Y_{1,i} - Y_{0,i}$. Assuming that the disturbances are mean zero conditional on $\mathbf{X}_i$, the expected program effect conditional on these covariates is

$$\tau(\mathbf{x}) = \mathrm{E}\left(Y_{1,i} - Y_{0,i} \mid \mathbf{X}_i = \mathbf{x}\right) = \mathbf{x}'(\beta_1 - \beta_0).$$

If we can find a way to estimate β_1 and β_0 from the observed data, then we can estimate $\tau(\mathbf{x})$. The challenge is how to carry out the estimation when program participation is the result of self-selection on the part of the workers.

The rule by which workers select themselves into the program—that is, the program participation rule—is

$$D_i = 1 \text{ if } Y_{1,i} \geq Y_{0,i}$$

or equivalently

$$D_i = 1 \text{ if } \epsilon_{1,i} - \epsilon_{0,i} \geq -\mathbf{X}_i'(\beta_1 - \beta_0).$$

For a worker i who chooses to participate in the program ($D_i = 1$, $Y_i = Y_{1,i}$), the expected outcome is

$$\begin{aligned} \mathrm{E}(Y_{1,i} \mid \mathbf{X}_i, D_i = 1) &= \mathbf{X}_i'\beta_1 + \mathrm{E}(\epsilon_{1,i} \mid \mathbf{X}_i, D_i = 1) \\ &= \mathbf{X}_i'\beta_1 + \mathrm{E}(\epsilon_{1,i} \mid \mathbf{X}_i, \epsilon_{1,i} - \epsilon_{0,i} \geq -\mathbf{X}_i'(\beta_1 - \beta_0)). \end{aligned}$$

A worker i who chooses not to participate in the program ($D_i = 0$, $Y_i = Y_{0,i}$), has expected outcome

$$\begin{aligned} \mathrm{E}(Y_{0,i} \mid \mathbf{X}_i, D_i = 0) &= \mathbf{X}_i'\beta_0 + \mathrm{E}(\epsilon_{0,i} \mid \mathbf{X}_i, D_i = 0) \\ &= \mathbf{X}_i'\beta_0 + \mathrm{E}(\epsilon_{0,i} \mid \mathbf{X}_i, \epsilon_{1,i} - \epsilon_{0,i} < -\mathbf{X}_i'(\beta_1 - \beta_0)). \end{aligned}$$

Let $v_i = \epsilon_{1,i} - \epsilon_{0,i}$ and let $\Delta = \beta_1 - \beta_0$. We know from earlier work with Heckman-type models how to calculate the selectivity terms under the assumption of joint normal disturbances. For the program participants, for example,

$$\mathrm{E}(\epsilon_{1,i} \mid \mathbf{X}_i, v_i \geq -\mathbf{X}_i'\Delta) = \frac{\sigma_{v1}}{\sigma_v} \cdot \frac{\phi(-\mathbf{X}_i'\Delta/\sigma_v)}{1 - \Phi(-\mathbf{X}_i'\Delta/\sigma_v)},$$

which we would expect to be positive. Subject to identification conditions, a full two-equation maximum-likelihood approach can be applied to consistently estimate β_1 and β_0 and in this way estimate the expected program effect $\tau(\mathbf{X})$ given the covariates. Therefore, even though the *worker-specific* program effect is never observed, we can recover consistent estimates of the *expected* conditional program effect and test hypotheses about it using relatively standard maximum-likelihood techniques. These results come at the cost of imposing strong assumptions.

33.2 Fully randomized assignment

Since the heyday of the Heckman-type models in the mid-1980s, economists have become increasingly uneasy about the assumptions such models require (including joint normal disturbances, a distributional assumption that is difficult to test) and have sought simpler, less assumption-laden alternatives. One important alternative is that of randomized experimental interventions, whereby (in our example) workers *do not self-select* into programs but rather are *randomly assigned* to the participation or non-participation groups. When it is done correctly, randomization allows expected program effects to be estimated with simple techniques.

In the usual case, assignment to the participation ("treatment") and non-participation ("control") groups is carried out independently of the $\mathbf{X}_i$ covariates and the (hypothetical) outcomes. (In this section and the next, we will suppress the i subscript to simplify notation.) Random assignment is often expressed in the experiments literature using the following notation,

$$(Y_0, Y_1, \mathbf{X}) \perp D,$$

in which "$\perp$" means "independent". In more conventional notation, we would write the (mixed, since D is discrete) joint density of $Y_0, Y_1, \mathbf{X}$ and D in the factored form $f(y_0, y_1, \mathbf{x}) \cdot g(d)$.

Since the observed $Y = DY_1 + (1 - D)Y_0$, the implications of independence for estimating expected program effects are immediate:

$$\begin{aligned} \mathrm{E}(Y \mid \mathbf{X}, D) &= D \cdot \mathrm{E}(Y_1 \mid \mathbf{X}, D) + (1 - D) \cdot \mathrm{E}(Y_0 \mid \mathbf{X}, D) \\ &= D \cdot \mathrm{E}(Y_1 \mid \mathbf{X}) + (1 - D) \cdot \mathrm{E}(Y_0 \mid \mathbf{X}) \\ &= D\mu_1(\mathbf{X}) + (1 - D)\mu_0(\mathbf{X}) \end{aligned}$$

The difference between the conditional expected values of Y for the treatment and control groups is

$$\mathrm{E}(Y \mid \mathbf{X}, D = 1) - \mathrm{E}(Y \mid \mathbf{X}, D = 0) = \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) = \tau(\mathbf{X}),$$

which is the program effect conditional on $\mathbf{X}$. This effect can be estimated for each $\mathbf{X} = \mathbf{x}$ simply by subtracting the sample mean of Y for the control group (more precisely, the subset of controls with $\mathbf{X} = \mathbf{x}$) from the mean of the counterpart treatment group, giving

$$\hat{\tau}(\mathbf{x}) = \hat{\mu}_1(\mathbf{x}) - \hat{\mu}_0(\mathbf{x}).$$

Note that estimates of the program effect require that *both* the treatment and control groups have enough observations with $\mathbf{X} = \mathbf{x}$ to allow the two sample means to be estimated. This can be an issue when there are many $\mathbf{X}$ covariates, or when some categorical covariates have relatively few cases in a category. A number of less stringent matching methods have been proposed to address this common practical problem.

If only the overall program effect is of interest, then there is no need for conditioning on the covariates. From

$$\begin{aligned} \mathrm{E}(Y \mid D) &= D \cdot \mathrm{E}(Y_1 \mid D) + (1 - D) \cdot \mathrm{E}(Y_0 \mid D) \\ &= D \cdot \mathrm{E}(Y_1) + (1 - D) \cdot \mathrm{E}(Y_0) \\ &= D\mu_1 + (1 - D)\mu_0, \end{aligned}$$

with the overall program effect estimated by the difference in the sample means of the treatment and control groups.

33.2.1 In what sense are "experiments" the gold standard?

Having seen the basic econometrics of randomization, we should now be more critical about how it operates in the field, especially in economic and other social-science experiments. In the medical field, the usual steps in implementing a randomized experiment are as follows:

Identification of study population Researchers specify the population that might benefit from a new treatment and assemble *a sample of potential participants*.

Informed consent Members of the sample are given a description of the study and are offered an opportunity to participate. *Those who agree* form the actual sample for the experiment.

Random assignment Some random mechanism (e.g., a coin flip) is used to separate participants into a *treatment group* and a *control group*. The *randomness is critical*: done properly, it ensures that there is *no systematic difference* between the two groups.

Application of treatment At this stage, the treatment is given to the treatment group, and something else (nothing; a placebo; an older well-established treatment against which the new treatment is being compared) is given to the control group.

Evaluation of results Outcomes in the treatment group are compared statistically with outcomes in the control group.

Even in medical contexts, there are reasons to doubt whether the assumptions underlying randomized experiments are met in any strict sense. Here are some potential weak points to consider, which typically apply with even more force to social science settings:

Weak Point (1): Identifying the Study Population For health interventions, we now know that there are *significant differences* between (obviously) the genetic makeups of women and men, which extend to their *immune systems and thus their responses to treatment*. Likewise there appear to be significant differences by "race" (not a scientific term), with Blacks often showing different responses. Also, there are likely to be significant differences in the prevalence and severity of *pre-existing conditions* by these and other socioeconomic categories.

Researchers therefore *need to anticipate different responses among sub-groups of their participants*. If they don't bear this in mind, the study may not be interpretable—it may suggest no positive effect overall even if there are substantial benefits for some sub-groups, or may suggest important positive effects overall that really occur only in one particular sub-group. *Ideally, analysis of outcomes should be sub-group-specific*. Unfortunately, it may be *too costly* to recruit enough participants from all of the relevant sub-groups to carry out a convincing sub-group analysis.

Weak Point (2): The Informed Consent Stage Potential participants weigh the advantages and disadvantages of participating, drawing on the description of the study given to them and keeping their own circumstances, constraints, and priorities in mind. *They make a choice.*

At an early moment in the fight against COVID-19, it became difficult to recruit participants to an experimental evaluation of the benefits of the *convalescent plasma* treatment, because that treatment could be obtained elsewhere without the hassle of experimental protocols. Not surprisingly, people preferred to get the treatment for sure from an outside source, rather than sign up for a study in which they might be given a placebo. Difficulties of recruitment were further magnified once effective vaccines began to appear.

Those who agree to take part in an experiment may differ in many respects from those who decline—by sex, "race", income level, distance from the facility where the study takes place, and the like. The resulting *sample of participants may not be representative of the original study population*. We can't be sure that their responses to treatment will generalize to the population at large.

Weak Point (3): Application of Treatment and Control Activities

Placebos In the simplest medical experiments, the treatment group is given the new treatment and the control group is given a "placebo"—for instance, a sugar pill that looks identical to the pill containing the new drug that goes to the treatment group. Neither treatment nor control patients know which pill they've received. This *"single blind"* approach should ensure that there are no systematic differences in subsequent patient behavior (or attentiveness to possible symptoms) afterward.

"Double blind" Those *administering* the treatment and placebo should also not know which one they've provided—this should ensure that the medical staff do not provide systematically different care to the treatment and control groups.

But single-blind/double-blind approaches can be infeasible In more complex medical and nearly all social science experiments, it may be impossible to follow these approaches. If the treatment requires the patient to follow a strict regimen of medication and follow-up, and nothing identical can be devised for controls, both patients and medical staff will likely become aware of who's in which group.

In social science experiments, the treatment may involve (for example) a new type of job-training program that is quite different from the program undergone by the controls—the differences may well become apparent to both treatment and control-group members and of course their trainers will know.

Drop-out When the treatment and control approaches play out over time, it is inevitable that some participants will drop out before completing the study. *Dropping out is a choice*, made by participants who perceive continuing in the experiment to be less advantageous than stopping. People have many demands on their time and widely differing abilities to stick with a program. Some people may be more apt to experience side-effects from health treatments, or worry more about them, causing drop-out. Researchers may not be able to identify and measure all the factors that lead to drop-out, and are thus often unable to adjust their estimates of treatment effect with these factors in mind.

The sample of people who end up taking the full course of treatment and control activities may be unrepresentative of those who started. Differential drop-out is a serious problem across all types of experiments, in medicine as well as social science.

So, are randomized experiments really the "gold standard" of research? *It depends*. No blanket statement of this sort should be made, despite what you may have heard. You have to examine critically all the steps in a randomized experiment to know whether its outcome can be trusted. Keep a sharp eye out for junctures where *choices—on the part of study participants, or those administering the study*—can gum up the works.

An additional point to consider is the extent to which the new treatments provided in an experiment are identical to those that will be rolled out later on when the treatment is actually adopted at scale. Experimental treatments could have been hindered by logistical glitches and other unanticipated snags, which might well be smoothed out once the treatment is provided in the larger population. Or, the fact that the treatment is new

and promising at the time of the experiment may mean that its implementers tend to be unusually enthusiastic promoters and problem-solvers within their treatment teams; but once treatment delivery is standardized and the health/labor bureaucracy takes charge, such "champions" may no longer exert influence. For many reasons, therefore, successful "pioneer" programs may ultimately fail to replicate their results at scale. In the real world, the "scaling-up" problem can present truly formidable difficulties.

33.3 Non-experimental data and ignorability

As the preceding discussion suggests, the textbook notion that "program participation" can be summarized in a binary, "does or doesn't" variable is something of a convenient fiction. Even for those who get as far as enrolling in an experiment, participation is often better viewed as a matter of degree, an endogenous position chosen along the continuum from "never really participated in the treatment" to "participated to the fullest possible extent". How much of the experimental approach survives when workers select themselves into programs or choose the level of effort they invest in adhering to experimental protocols?

If selection is *ignorable*, then quite a lot. Ignorability is expressed as

$$(Y_0, Y_1) \perp D \mid \mathbf{X},$$

meaning that conditional on covariates $\mathbf{X}$, the (hypothetical) outcomes Y_0, Y_1 are independent of participation. Written more conventionally in terms of the conditional joint (mixed) densities of Y_0, Y_1 and D given $\mathbf{X}$, ignorability implies that the factored form of the conditional density is $f(y_0, y_1 \mid \mathbf{x}) \cdot g(d \mid \mathbf{x})$. (Despite having just argued against this, we will continue with the textbook tradition of treating D as binary variable.)

Under the ignorability assumption, it is conceptually straightforward to find the expected program effect conditional on covariates $\mathbf{X}$. Unlike the case of randomized experimental assignment, here is it essential to begin the analysis in conditional terms, because it is only in that way that we can exploit the implications of ignorability.

Again denoting the observed outcome by Y, with

$$Y = DY_1 + (1 - D)Y_0,$$

we have

$$\begin{aligned} \mathrm{E}(Y \mid D, \mathbf{X}) &= D\,\mathrm{E}(Y_1 \mid D, \mathbf{X}) + (1 - D)\,\mathrm{E}(Y_0 \mid D, \mathbf{X}) \\ &= D\,\mathrm{E}(Y_1 \mid \mathbf{X}) + (1 - D)\,\mathrm{E}(Y_0 \mid \mathbf{X}) \\ &= D\mu_1(\mathbf{X}) + (1 - D)\mu_0(\mathbf{X}) \end{aligned}$$

Therefore,

$$\mathrm{E}(Y \mid D = 1, \mathbf{X}) - \mathrm{E}(Y \mid D = 0, \mathbf{X}) = \mu_1(\mathbf{X}) - \mu_0(\mathbf{X}) \equiv \tau(\mathbf{X}).$$

That is, for any given $\mathbf{X} = \mathbf{x}$, the difference between the expected value of Y for the participants (the "treatment group") and its expected value for non-participants (the "comparison group") is the expected program effect.

This seems almost too easy. What kinds of self-selection does ignorability rule out? Looking back at the Heckman-type model, in which workers decide whether to enter

training on the basis of whether $Y_{1,i} \geq Y_{0,i}$, the participation rule is $\epsilon_{1,i} - \epsilon_{0,i} \geq -\mathbf{X}_i'(\beta_1 - \beta_0)$. In this model, participation does not depend on $\mathbf{X}_i$ alone, but rather on both $\mathbf{X}_i$ and the disturbances $\epsilon_{1,i}, \epsilon_{0,i}$, with the disturbances also figuring into the outcomes $Y_{0,i}$ and $Y_{1,i}$. Hence, the conditional mixed density for the Heckman model *does not factor* into one component specific to $Y_{0,i}, Y_{1,i}$ given $\mathbf{X}_i$ alone, and another component specific to D_i given $\mathbf{X}_i$ alone. In the Heckman model, therefore, selection is *not ignorable*.

In an applied problem, the number of variables in $\mathbf{X}$ would generally need to be large—encompassing essentially all factors governing self-selection—if the ignorability assumption is to be at all persuasive. Even in cases in which ignorability can be safely assumed, there can be difficulties in estimating expected program effects. Obviously, if a particular $\mathbf{x}$ combination of covariates is found in the treatment (participation) group but not in the comparison group (non-participants), or vice–versa, then the conditional expected effect $\tau(\mathbf{x})$ cannot be estimated. This might well happen in an empirical application, especially since the number of variables in $\mathbf{X}$ needs to be large.

33.3.1 Propensity scores

As a practical matter, therefore, it would be nice if a function of $\mathbf{X}$ could be found that, when applied, reduced the multiple dimensions of the covariates to a single summary index. The leading candidate for such a function is the so-called *propensity score* $p(\mathbf{X})$, defined as the conditional probability of program participation given $\mathbf{X}$:

$$p(\mathbf{X}) = \Pr\left(D = 1 \mid \mathbf{X}\right).$$

Since Rosenbaum and Rubin (1983), an enormous and in places highly technical literature has explored the possibilities of basing program evaluation on propensity scores. This outpouring of research has all but obscured the fundamental limitation of the method: *ignorability is its core assumption*. For those who think of program participation in terms of the Heckman-type selection models sketched above, propensity score approaches are less than satisfying. Even so, every student of econometrics needs to know at least the the basics of the approach.

The Rosenbaum–Rubin approach The paper that effectively launched the propensity score method is Rosenbaum and Rubin (1983), who showed how an analysis of expected program effects conditional on $\mathbf{X}$ could be replaced with an analysis of expected effects given the level of the propensity score. Since $\mathbf{X}$ is multidimensional whereas the propensity score is a single index derived from $\mathbf{X}$, this approach seems to offer a simpler alternative.

The main result of the paper is that ignorability expressed in terms of $\mathbf{X}$ implies ignorability in terms of the propensity score $p(\mathbf{X})$; that is

$$(Y_0, Y_1) \perp D \mid \mathbf{X} \Rightarrow (Y_0, Y_1) \perp D \mid p(\mathbf{X})$$

provided that $0 < p(\mathbf{X}) < 1$. The proof of this somewhat surprising result is developed in Angrist and Pischke (2009, pp. 80–81) as an exercise in iterated expectations. The aim is to show that

$$\Pr(D = 1 \mid (Y_1, Y_0), p(\mathbf{X}))$$

is not, in fact, a function of the potential outcomes (Y_1, Y_0). That is, we aim to show that given $p(\mathbf{X})$, the D variable is independent of (Y_1, Y_0).

It is helpful to begin by re-expressing things in terms of conditional expectations, exploiting the fact that D is a binary random variable,

$$\Pr(D = 1 \mid (Y_1, Y_0), p(\mathbf{X})) = \mathrm{E}(D \mid (Y_1, Y_0), p(\mathbf{X})).$$

We now apply the method of iterated expectations to analyze the right-hand side. The specific technique we'll use was explained in Chapter 5. The key is to add $\mathbf{X}$ itself to the inner conditioning set in the first step of iterated expectations, using

$$\mathrm{E}(D \mid (Y_1, Y_0), p(\mathbf{X})) = \mathrm{E}_{\mathbf{X} \mid (Y_1, Y_0), p(\mathbf{X})} \left(\mathrm{E}\left(D \mid (Y_1, Y_0), p(\mathbf{X}), \mathbf{X}\right) \right)$$

Almost all the action of the proof happens in the first step of iterated expectations.

So, let's extract the "inner" expectation and work with it. Since the conditioning set has been expanded to include $\mathbf{X}$, we recognize that

$$\mathrm{E}\left(D \mid (Y_1, Y_0), p(\mathbf{X}), \mathbf{X}\right) = \mathrm{E}\left(D \mid (Y_1, Y_0), \mathbf{X}\right)$$

because $\mathbf{X}$ determines $p(\mathbf{X})$ and thus $p(\mathbf{X})$ is redundant in the conditioning set. Now we make use of the strong ignorability assumption with respect to $\mathbf{X}$, which yields

$$\mathrm{E}\left(D \mid (Y_1, Y_0), \mathbf{X}\right) = \mathrm{E}\left(D \mid \mathbf{X}\right) = \Pr(D = 1 | \mathbf{X}) = p(\mathbf{X}).$$

We can now bring this result back into the "outer" expectation, which represents the second step of iterated expectations. I'll make the conditioning set for this second step more explicit,

$$\mathrm{E}\left(p(\mathbf{X}) \;\middle|\; (Y_1, Y_0), p(\mathbf{X}) \right)$$

But look at how simple this second step is: We are taking the expectation of $p(\mathbf{X})$ conditional on $p(\mathbf{X})$ itself—nothing relevant is added by the presence of (Y_1, Y_0) in the conditioning set. Hence, the result of the second step of iterated expectations is simply

$$\mathrm{E}\left(p(\mathbf{X}) \;\middle|\; (Y_1, Y_0), p(\mathbf{X}) \right) = p(\mathbf{X})$$

Linking the start and end of the proof, we see that

$$\Pr(D = 1 \mid (Y_1, Y_0), p(\mathbf{X})) = p(\mathbf{X})$$

which means that in fact the distribution of D conditional on (Y_1, Y_0) and $p(\mathbf{X})$ is, in fact, not a function of (Y_1, Y_0). In other words, D is independent of (Y_1, Y_0) conditional on $p(\mathbf{X})$, which is what we set out to prove.

In the propensity score approach, "treatment" and "comparison" groups are no longer defined in terms of common $\mathbf{X} = \mathbf{x}$ values, but rather in terms of common (ideally, equal or nearly equal) propensity scores $p(\mathbf{X}) = p$. The essence of the matching idea is to pair up a program participant with propensity score $\tilde{p}$ with a non-participant having the same score,

and then examine the difference in their respective Y values, doing so for a large enough collection of paired individuals to obtain a credible estimate.

Given hypothetical outcomes (Y_0, Y_1) as before, we proceed to write their conditional means given $p(\mathbf{X}) = \tilde{p}$ as

$$\mathrm{E}(Y_0 \mid p(\mathbf{X}) = \tilde{p}) = m_0(\tilde{p}) \text{ and } \mathrm{E}(Y_1 \mid p(\mathbf{X}) = \tilde{p}) = m_1(\tilde{p})$$

and likewise redefine the conditional program effect given the score as $t(\tilde{p}) = m_1(\tilde{p}) - m_0(\tilde{p})$. That is, in a propensity score analysis we no longer speak of $\tau(\mathbf{x})$, the expected program effect given $\mathbf{X} = \mathbf{x}$, but instead focus attention on $t(\tilde{p})$, the expected program effect given $p(\mathbf{X}) = \tilde{p}$. Other than the focus on groups defined by $p(\mathbf{X}) = \tilde{p}$ rather than by $\mathbf{X} = \mathbf{x}$, the logic is the same as that laid out just above.

It is certainly nicer computationally to work with a single index $p(\mathbf{X})$ than to grapple with $\mathbf{X}$ itself, a complex multivariate object. The down-side of the single index method is that if we find that conditional program effects differ between one treatment–comparison group with propensity score p' and another group with score p'', it can be hard to understand precisely *why* those effects differ. To determine the composition of the two groups, you would have to return to the full set of $\mathbf{X}$ covariates and summarize them in some fashion, hoping to spot some compositional difference between the groups that might produce different responses to the program. In this sense, the simplicity seemingly offered by the propensity score method is a bit of an illusion.

Much of the literature on propensity-score methods is given over to the details of exactly how to match program participants to non-participants on the basis of propensity scores. There are many possibilities:

- For each participant with score $\tilde{p}$ find a non-participant with exactly that score;
- For each participant with score $\tilde{p}$ find a set of non-participants with scores in the range $(\tilde{p} - \Delta, \tilde{p} + \Delta)$ for a relatively small value $\Delta > 0$;
- Match each participant with score $\tilde{p}$ to a weighted average of all non-participants, using weights that increasingly down-weight non-participants the further the non-participant score is from $\tilde{p}$;

and there are yet further variations along these lines.

Within the econometric community, opinion is still divided on whether any of these matching methods is really needed. Both $m_0(\tilde{p})$ and $m_1(\tilde{p})$ can be estimated using flexible functional forms, involving for instance the squares and cubes of $\tilde{p}$, or by application of similar nearly non-parametric techniques. Wooldridge (2010) provides a detailed account of such regression-related approaches.

33.4 Behavioral responses to randomization: The LATE estimator

In this section we study the "local average treatment effect" or LATE estimator, which is less an estimator as such than a method of interpreting the coefficient from a standard instrumental variables estimator when there are person-specific responses to the program (Angrist and Pischke 2009).

The set-up for our discussion is as follows. Let the variable Z_i represents the (fully randomized) assignment of subject i either to the program ($Z_i = 1$) or to the control group ($Z_i = 0$). When presented with a randomly-generated assignment, a subject decides whether or not to comply with it. In other words, there is self-selection into the program or control group. The exact mechanism by which self-selection takes place is left unspecified, so that non-ignorable selection (i.e., selection on unobservables) is permitted. Since the assignment Z_i surely exerts *some* influence on whether people enter the program or the control group, and is furthermore randomly generated, it would seem to have good potential to serve as an instrumental variable.

If the observed outcome variable Y_i can be modeled in terms of the observed participation D_i as

$$Y_i = \alpha + \tau D_i + \epsilon_i$$

with a program effect coefficient τ that is *the same for all who participate*, then τ can be estimated consistently by the method of instrumental variables. In this case, with a binary instrument Z_i and a binary endogenous variable D_i, the estimator is easily derived from

$$\begin{aligned} \mathrm{E}(Y_i|Z_i = 1) &= \alpha + \mathrm{E}(D_i|Z_i = 1)\,\tau \\ \mathrm{E}(Y_i|Z_i = 0) &= \alpha + \mathrm{E}(D_i|Z_i = 0)\,\tau \end{aligned}$$

since $\mathrm{E}(\epsilon_i|Z_i) = 0$ by the randomness of the assignment. Hence, the true value of τ is

$$\tau = \frac{\mathrm{E}(Y_i|Z_i = 1) - \mathrm{E}(Y_i|Z_i = 0)}{\mathrm{E}(D_i|Z_i = 1) - \mathrm{E}(D_i|Z_i = 0)}$$

and this true value can be consistently estimated by using sample means in place of the expectations.

If the treatment effect is person-specific, however, such that the effect for person i is τ_i, then as discussed in an earlier chapter [MM: to be written up when time permits], standard instrumental-variables methods *do not estimate even its population mean* $\mathrm{E}\,\tau_i$ *consistently*. Rather, as we will see, under one critical assumption the simple IV estimator converges to *the mean of* τ_i *in a particular population sub-group known as "always-compliers"*, those persons who would agree to participate if that is their assignment (via $Z_i = 1$), but who would also agree to join the control group if so directed (by $Z_i = 0$). To show this, we will re-examine the right-hand side of the equation above for the case of person-specific program effects and show that it converges not to $\mathrm{E}\,\tau$, the overall average program effect, but rather to the mean program effect in this special sub-group. The following exposition spells out the logic, developed using the concepts of hypothetical (or "potential") participation and outcomes that we've seen earlier in this chapter. The details of the argument are taken from Schroeder (2010).

We can think of the conceptual model as having three stages. In the first stage, subject i receives a fully randomized assignment Z_i. In the second stage, she or he decides how to respond to it, by participating or not. A pair of variables $D_{i,0}$ and $D_{i,1}$, each of them being binary, represent these potential responses. The $D_{i,0}$ variable refers to subjects assigned to the controls; it takes the value 0 if person i would agree to join the controls when assigned $Z_i = 0$, but takes the value 1 if, when assigned $Z_i = 0$, that person instead chooses to participate in the program. The $D_{i,1}$ variable refers to subjects assigned to the program;

it takes the value 0 if person i decides instead to join the controls, but takes the value 1 if that person agrees to participate in the program in line with the assignment. These binary response variables have the joint distribution shown in the following table:

		$D_{i,1}$	
		0	1
$D_{i,0}$	0	$\pi_{0,0}$	$\pi_{0,1}$
	1	$\pi_{1,0}$	$\pi_{1,1}$

(The critical assumption mentioned above has to do with a restriction placed on this [hypothetical] joint distribution.) The third stage of the conceptual model is the outcome stage: If person i participates in the program, he or she receives outcome $Y_{i,1}$; but if she or he does not participate, the outcome is $Y_{i,0}$. The difference $Y_{i,1} - Y_{i,0}$ is the person-specific program effect.

We begin by rewriting the numerator of the equation

$$\mathrm{E}(Y_i|Z_i = 1) - \mathrm{E}(Y_i|Z_i = 0)$$

in terms of potential participation and outcomes, as

$$\begin{aligned}
&\mathrm{E}(Y_i|Z_i = 1) - \mathrm{E}(Y_i|Z_i = 0) \\
&\qquad = \mathrm{E}(D_i Y_{i,1} + (1 - D_i)Y_{i,0} \mid Z_i = 1) - \mathrm{E}(D_i Y_{i,1} + (1 - D_i)Y_{i,0} \mid Z_i = 0) \\
&\qquad = \mathrm{E}(D_{i,1} Y_{i,1} + (1 - D_{i,1})Y_{i,0} \mid Z_i = 1) - \mathrm{E}(D_{i,0} Y_{i,1} + (1 - D_{i,0})Y_{i,0} \mid Z_i = 0).
\end{aligned}$$

Because of the randomization in assignment, we have $(Y_{i,1}, Y_{i,0}, D_{i,1}, D_{i,0}) \perp Z_i$, so that expectations conditional on Z_i are the same as the unconditional expectations. The last line above therefore simplifies to

$$\mathrm{E}(D_{i,1} Y_{i,1} + (1 - D_{i,1})Y_{i,0}) - \mathrm{E}(D_{i,0} Y_{i,1} + (1 - D_{i,0})Y_{i,0}).$$

We can now collect terms and rearrange things as follows,

$$\begin{aligned}
\mathrm{E}(Y_i|Z_i = 1) - \mathrm{E}(Y_i|Z_i = 0) &= \mathrm{E}(D_{i,1} - D_{i,0})Y_{i,1} + \mathrm{E}((1 - D_{i,1}) - (1 - D_{i,0}))Y_{i,0} \\
&= \mathrm{E}(D_{i,1} - D_{i,0})Y_{i,1} - \mathrm{E}(D_{i,1} - D_{i,0})Y_{i,0} \\
&= \mathrm{E}\left(D_{i,1} - D_{i,0}\right)\left(Y_{i,1} - Y_{i,0}\right).
\end{aligned}$$

Whenever $D_{i,1} = D_{i,0}$ the product of the two terms in parentheses is zero. Hence the only cases that matter in this expectation are those in which the two variables differ: $D_{i,1} - D_{i,0} = 1$ and $D_{i,1} - D_{i,0} = -1$. Having restricted attention to these two cases, we can then write the unconditional expectation as the weighted average of conditional expectations,

$$\begin{aligned}
&\mathrm{E}\left(D_{i,1} - D_{i,0}\right)\left(Y_{i,1} - Y_{i,0}\right) = \\
&\qquad \mathrm{E}\left(\left(D_{i,1} - D_{i,0}\right)\left(Y_{i,1} - Y_{i,0}\right) \mid D_{i,1} - D_{i,0} = 1\right) \Pr(D_{i,1} - D_{i,0} = 1) \\
&\qquad + \mathrm{E}\left(\left(D_{i,1} - D_{i,0}\right)\left(Y_{i,1} - Y_{i,0}\right) \mid D_{i,1} - D_{i,0} = -1\right) \Pr(D_{i,1} - D_{i,0} = -1)
\end{aligned}$$

which can be further simplified to

$$\begin{aligned}\mathrm{E}\,(D_{i,1} - D_{i,0})\,(Y_{i,1} - Y_{i,0}) = \\ \mathrm{E}\,\big((Y_{i,1} - Y_{i,0}) \mid D_{i,1} - D_{i,0} = 1\big)\,\Pr(D_{i,1} - D_{i,0} = 1) \\ -\,\mathrm{E}\,\big((Y_{i,1} - Y_{i,0}) \mid D_{i,1} - D_{i,0} = -1\big)\,\Pr(D_{i,1} - D_{i,0} = -1).\end{aligned}$$

This is the difference between the expected program effects for two sub-groups, each weighted by its proportion in the overall subject population—the "always-compliers" who would participate or not in accordance with their random assignment to the program or control group, and the "always-defiers" who would always do the opposite of what their random assignment would direct that they do.

Interestingly—and this is the critical point—even if all individuals in the population would benefit from the program, that is, even if program effects $Y_{i,1} - Y_{i,0} > 0$ for all i, this weighted difference in expected program effects between the two groups *need not be positive*. Its sign would depend on the distribution of $Y_{i,1}$ and $Y_{i,0}$ in these two groups as well as their relative sizes.

At this point in the development of the LATE perspective, a large and potentially controversial assumption is introduced: *The probability of being an aways-defier is set to zero*. In other words, the joint distribution of $D_{i,0}$ and $D_{i,1}$ is assumed to be triangular,

		$D_{i,1}$	
		0	1
$D_{i,0}$	0	$\pi_{0,0}$	$\pi_{0,1}$
	1		$\pi_{1,1}$

Can such defiant behavior really be ruled out? What are the substantive issues that we should think about? The conceptual model as set out so far provides us with little guidance. Obviously, the demands that programs place on participants and controls probably vary a great deal across programs, but these restrictions and costs are not specified in the model. In particular, nothing has been said up to now on whether it is possible for an individual to switch from program to control group, or vice–versa, or whether individuals who dislike their assignment would simply abandon the experiment altogether and be lost to follow-up. This lack of specific detail is often viewed as a modelling strength by adherents of the LATE method; but when it comes to interpretation, the lack of detail is also a weakness.

Note that the no-defiers assumption still allows individuals assigned to the program to join the controls, and allows those assigned to the control group to join the program—it only disallows those who are "always-defiers", those who always disobey their assignments. The model retains the "always-compliers", however, those who always obey their assignments. This asymmetric treatment of defiers and compliers is perhaps not completely unreasonable, but needs some substantive justification. The model itself provides none, so that justification must be sought elsewhere.

Assuming that the always-defiers are indeed absent from the overall population of subjects, we obtain this simplification,

$$\begin{aligned}\mathrm{E}(Y_i|Z_i = 1) - \mathrm{E}(Y_i|Z_i = 0) = \\ \mathrm{E}\,\big((Y_{i,1} - Y_{i,0}) \mid D_{i,1} - D_{i,0} = 1\big)\,\Pr(D_{i,1} - D_{i,0} = 1)\end{aligned}$$

which is the expected program benefit for "always-compliers" multiplied by their proportion in the population.

We'll now show that the proportion of such compliers in the population—the factor on the right—can be made to cancel out of an appropriately-defined ratio. In terms of the joint distribution shown in the table above, $\Pr(D_{i,1} - D_{i,0} = 1) = \pi_{0,1}$. Now, given that observed participation

$$D_i = D_{i,1}Z_i + D_{i,0}(1 - Z_i)$$

and that

$$\begin{aligned}
\mathrm{E}(D_i|Z_i = 1) &= \mathrm{E}(D_{i,1}|Z_i = 1) = \mathrm{E}(D_{i,1}) = \pi_{0,1} + \pi_{1,1} \\
\mathrm{E}(D_i|Z_i = 0) &= \mathrm{E}(D_{i,0}|Z_i = 0) = \mathrm{E}(D_{i,0}) = \pi_{1,1},
\end{aligned}$$

it follows that

$$\Pr(D_{i,1} - D_{i,0} = 1) = \pi_{0,1} = \mathrm{E}(D_i|Z_i = 1) - \mathrm{E}(D_i|Z_i = 0).$$

Hence, the ratio that is the population value of the IV estimator with Z_i treated as an instrument for D_i, or

$$\frac{\mathrm{E}(Y_i|Z_i = 1) - \mathrm{E}(Y_i|Z_i = 0)}{\mathrm{E}(D_i|Z_i = 1) - \mathrm{E}(D_i|Z_i = 0)}$$

is, by the analysis above,

$$\frac{\mathrm{E}\left((Y_{i,1} - Y_{i,0}) \mid D_{i,1} - D_{i,0} = 1\right) \Pr(D_{i,1} - D_{i,0} = 1)}{\mathrm{E}(D_i|Z_i = 1) - \mathrm{E}(D_i|Z_i = 0)},$$

and after the common term in numerator and denominator is cancelled, this is revealed to be

$$\mathrm{E}\left((Y_{i,1} - Y_{i,0}) \mid D_{i,1} - D_{i,0} = 1\right),$$

the average program effect in the population of always-compliers.

It is not obvious that these compliers—people who always obey their assignments, whether to the program or to the control group—are necessarily an interesting sub-group in economic terms or in terms of the program's welfare-improving objectives. Insight into these issues would require a specification of program details, perhaps accompanied by a conceptual model of the decision to accept or evade random assignment. But the LATE estimator at least extracts some information of potential value from the probability limit of the standard IV estimator.

Chapter 34

Measurement Error: The Basics

This chapter discusses the econometric properties of regression models that include mis-measured explanatory variables. We present the standard model of measurement error first, and then relax some of its simplifying assumptions. The next chapter takes a broader view of the problem, situating the standard errors-in-variables model in the context of missing and proxy variables.

Our discussion will be organized around the linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\delta + \epsilon \tag{34.1}$$

in which the dependent variable $\mathbf{Y}$ is $n \times 1$, the $\mathbf{X}$ matrix of explanatory variables is $n \times k$ and the $n \times 1$ vector $\mathbf{Z}$ represents an explanatory variable whose true value is not reported in the dataset. We assume that the disturbance term ϵ has mean zero and covariance matrix $\sigma^2\mathbf{I}$. We will make the standard assumption about the relationships between ϵ, $\mathbf{X}$ and $\mathbf{Z}$, namely, that ϵ is weakly exogenous to both $\mathbf{X}$ and $\mathbf{Z}$. By "weakly exogenous" we mean that $\text{plim}\, n^{-1}\mathbf{X}'\epsilon = \mathbf{0}$, and $\text{plim}\, n^{-1}\mathbf{Z}'\epsilon = 0$. With mild additional assumptions, this would guarantee that least-squares estimators $\hat{\beta}$ and $\hat{\delta}$ would be consistent if the true values of $\mathbf{Z}$ were available.

34.1 Classical Errors-in-Variables

Suppose that in our dataset we have $\mathbf{P}$, a mis-measured version of $\mathbf{Z}$, and that $\mathbf{P} = \mathbf{Z} + \mathbf{m}$, with the measurement error $\mathbf{m}$ assumed to be uncorrelated with $\mathbf{Z}$ as well as uncorrelated with $\mathbf{X}$ and ϵ. This is sometimes termed the classical errors-in-variables specification (Fuller 1987). It is straightforward to show that in this instance, the OLS estimator $\hat{\delta}$ is biased toward zero.

The proof begins as follows:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \mathbf{P}\delta + \epsilon - \mathbf{m}\delta \\ &= \mathbf{X}\beta + \mathbf{P}\delta + \mathbf{v} \end{aligned}$$

with $\mathbf{v} = \epsilon - \mathbf{m}\delta$. The proxy $\mathbf{P}$ and the composite error $\mathbf{v}$ are correlated, with $\text{plim}\, n^{-1}\mathbf{P}'\mathbf{v} =$

$-\delta\sigma_m^2$. The FWL theorem tells us that

$$\hat{\delta} = \delta + \left(\frac{1}{n}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P}\right)^{-1}\frac{1}{n}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{v}.$$

Expanding $\mathbf{M}_\mathbf{X}$ and using the assumption $\operatorname{plim} n^{-1}\mathbf{X}'\mathbf{v} = \mathbf{0}$, we have

$$\frac{1}{n}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{v} \stackrel{a}{=} \frac{1}{n}\mathbf{P}'\mathbf{v} \xrightarrow{p} -\delta\sigma_m^2.$$

Now substituting $\mathbf{Z} + \mathbf{m}$ for $\mathbf{P}$ and expanding $\mathbf{M}_\mathbf{X}$, we can show that $\operatorname{plim} n^{-1}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P} = \sigma_m^2 + \operatorname{plim} n^{-1}\mathbf{Z}'\mathbf{M}_\mathbf{X}\mathbf{Z}$. Drawing all this together,

$$\operatorname{plim}\hat{\delta} = \delta\left(\frac{\operatorname{plim}\frac{1}{n}\mathbf{Z}'\mathbf{M}_\mathbf{X}\mathbf{Z}}{\sigma_m^2 + \operatorname{plim}\frac{1}{n}\mathbf{Z}'\mathbf{M}_\mathbf{X}\mathbf{Z}}\right).$$

Because the expression in parentheses lies between 0 and 1, we obtain an important result: under the assumptions of the classical errors-in-variables model, using a proxy $\mathbf{P}$ for the true $\mathbf{Z}$ causes the least-squares estimator of δ to be inconsistent, with the probability limit of $\hat{\delta}$ being closer to zero than the true parameter. Some refer to this result using the terms *shrinkage* or *attenuation* to describe how the estimator is drawn away from the true coefficient and toward zero.

Note that the larger is the variance of the measurement error σ_m^2, the greater is the degree of shrinkage. Also, the empirical association between $\mathbf{Z}$ and $\mathbf{X}$ figures into the relationship, with more shrinkage taking place when $\mathbf{X}$ is a good predictor of $\mathbf{Z}$, that is, when the normalized sum of squared residuals $n^{-1}\mathbf{Z}'\mathbf{M}_\mathbf{X}\mathbf{Z}$ from a regression of $\mathbf{Z}$ on $\mathbf{X}$ is relatively small. Furthermore, because

$$\sigma_m^2 + \frac{1}{n}\mathbf{Z}'\mathbf{M}_\mathbf{X}\mathbf{Z} \stackrel{a}{=} \frac{1}{n}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P},$$

if we are prepared to specify a value for the measurement error variance σ_m^2, we can estimate the shrinkage factor and thereby adjust the OLS estimator for inconsistency.

34.2 Reverse Regression

Another technique is to find a second estimator of δ whose probability limit is *further* from zero than the true parameter. Since the standard OLS estimator has a plim that is *closer* to zero than the true parameter, taking the two estimators together would give us a range in which (at least asymptotically) the true value of the parameter would be found.

The method is termed *reverse regression* and is easiest to explain if we strip down the structural model to the essential features, which can be seen in this simplified version,

$$Y_i = Z_i\delta + \epsilon_i.$$

We can think of Z_i as being measured in terms of deviations from its own mean, so that $(1/n)\sum_i Z_i^2 \stackrel{a}{=} \sigma_Z^2$. Assume as we have above that $\mathrm{E}(\epsilon_i|Z_i) = 0$, which with the usual mild

additional assumptions would mean that if Z_i were actually available, the δ parameter could be consistently estimated.

The key step in reverse regression is to rewrite the model by rearranging things so that Z_i is on the left-hand side just as if it were a dependent variable,

$$Z_i = \frac{1}{\delta} Y_i - \frac{1}{\delta} \epsilon_i.$$

By adding the measurement error m_i to both sides, we obtain

$$P_i = \frac{1}{\delta} Y_i - \frac{1}{\delta} \epsilon_i + m_i.$$

Relabel $1/\delta$ as "d" for convenience, giving

$$P_i = dY_i - d\epsilon_i + m_i.$$

Now consider $\hat{d}$, the OLS estimator of d:

$$\hat{d} = d + (\mathbf{Y}'\mathbf{Y})^{-1} \cdot \mathbf{Y}'(-d\boldsymbol{\epsilon} + \mathbf{m}).$$

Substituting $Z\delta + \epsilon$ for Y gives

$$\mathbf{Y}'\mathbf{Y} \stackrel{a}{=} \delta^2 \frac{1}{n} \mathbf{Z}'\mathbf{Z} + \frac{1}{n} \boldsymbol{\epsilon}'\boldsymbol{\epsilon} \equiv \delta^2 \sigma_Z^2 + \sigma_\epsilon^2.$$

Also,

$$\frac{1}{n} \left(\mathbf{Y}'(-d\boldsymbol{\epsilon} + \mathbf{m}) \right) \stackrel{a}{=} -d\sigma_\epsilon^2.$$

Therefore,

$$\hat{d} \stackrel{a}{=} d - d \left(\frac{\sigma_\epsilon^2}{\delta^2 \sigma_Z^2 + \sigma_\epsilon^2} \right) = d \cdot \left(\frac{\delta^2 \sigma_Z^2}{\delta^2 \sigma_Z^2 + \sigma_\epsilon^2} \right).$$

Recalling that $d \equiv 1/\delta$, and inverting both sides, we arrive at the probability limit of our second estimator of δ,

$$\tilde{\delta} \stackrel{a}{=} \delta \left(\frac{\delta^2 \sigma_Z^2 + \sigma_\epsilon^2}{\delta^2 \sigma_Z^2} \right)$$

which is clearly further from zero than the true value of the δ parameter.

34.3 Departures from the Classical Error Assumptions

How robust is the "attenuation" result to variations in the assumptions of the classical error model? As it turns out, the result is rather fragile.

Multiple error-ridden covariates

What happens if we have *two* $\mathbf{Z}$ variables, each measured with error? Is it still the case that attenuation (shrinkage) takes place? From the FWL theorem, with $\hat{\delta}$ being 2×1,

$$\hat{\delta} = \delta + (\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{v},$$

with $\mathbf{v} = \epsilon - \mathbf{m}_1\delta_1 - \mathbf{m}_2\delta_2$. Then

$$\frac{1}{n}\mathbf{P}'\mathbf{v} \stackrel{a}{=} \begin{bmatrix} \frac{1}{n}(\mathbf{Z}_1 + \mathbf{m}_1)' \\ \frac{1}{n}(\mathbf{Z}_2 + \mathbf{m}_2)' \end{bmatrix} \cdot (\epsilon - \mathbf{m}_1\delta_1 - \mathbf{m}_2\delta_2)$$

or

$$\frac{1}{n}\mathbf{P}'\mathbf{v} \stackrel{a}{=} -\begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}.$$

Using this, we find

$$\operatorname{plim} \hat{\delta} = \delta - \left(\operatorname{plim} \frac{1}{n}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P} \right)^{-1} \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix} \begin{bmatrix} \delta_1 \\ \delta_2 \end{bmatrix}.$$

The sign and nature of the inconsistency is difficult to determine in this case. Evidently, the "shrinkage" result does not readily generalize to cover the case of two variables measured with error.

A robustness check

If you are willing to make assumptions about the variances and covariances of the measurement errors, you can construct a corrected version of the OLS estimator that (if your assumptions are correct) would be consistent for the regression slope parameters. Generally this approach is conducted as a kind of sensitivity or robustness check, the aim being to assess how sensitive the elements of the OLS estimator are to measurement errors in selected explanatory variables.

Suppose that for observation i, the k-vector of measurement errors for all k explanatory variables is $\mathbf{m}_i$, allowing some variables to be measured without error. (Note that we have changed notation from the preceding section.) Let $\Sigma_m \equiv \mathrm{E}\,\mathbf{m}_i\mathbf{m}_i'$ be the covariance matrix of the vector of measurement errors. To implement the approach, you would fill in the various variances and covariances that are the elements of Σ_m. Since Σ_m is $k \times k$, most researchers would not attempt to fully populate Σ_m with assumed values, but instead would focus on only one or two variables likely to be afflicted with measurement errors, taking the remaining variables to be error-free.

For observation i, with $\mathbf{P}_i = \mathbf{Z}_i + \mathbf{m}_i$, we have

$$Y_i = \mathbf{P}_i'\boldsymbol{\beta} + \epsilon_i - \mathbf{m}_i'\boldsymbol{\beta}$$

and for all n observations,

$$\mathbf{Y} = \mathbf{P}\boldsymbol{\beta} + \boldsymbol{\epsilon} - \mathbf{M}\boldsymbol{\beta}$$

with $\mathbf{M} = [\mathbf{m}_1, \mathbf{m}_2, \ldots, \mathbf{m}_n]'$ as usual. Hence,

$$\hat{\boldsymbol{\beta}} = (\mathbf{P}'\mathbf{P})^{-1}\left(\mathbf{P}'\mathbf{P}\boldsymbol{\beta} + \mathbf{P}'\boldsymbol{\epsilon} - \mathbf{P}'\mathbf{M}\boldsymbol{\beta}\right)$$

Using $\mathbf{P} = \mathbf{Z} + \mathbf{M}$ and assuming no correlation between $\mathbf{Z}$ and the measurement errors, we have

$$\operatorname{plim} \frac{1}{n}\mathbf{P}'\mathbf{M} = \operatorname{plim} \frac{1}{n}(\mathbf{Z} + \mathbf{M})'\mathbf{M} = \Sigma_m.$$

Then

$$\hat{\boldsymbol{\beta}} \stackrel{a}{=} \left(\frac{1}{n}\mathbf{P}'\mathbf{P}\right)^{-1}\left(\frac{1}{n}\mathbf{P}'\mathbf{P}\boldsymbol{\beta} - \Sigma_m\boldsymbol{\beta}\right) = \left(\frac{1}{n}\mathbf{P}'\mathbf{P}\right)^{-1}\left(\frac{1}{n}\mathbf{P}'\mathbf{P} - \Sigma_m\right)\boldsymbol{\beta}$$

Given the values you have assumed for the elements of Σ_m, most of which are likely to be zero, you can form the corrected estimator

$$\tilde{\boldsymbol{\beta}} = \left(\frac{1}{n}\mathbf{P}'\mathbf{P} - \Sigma_m\right)^{-1} \cdot \frac{1}{n}\mathbf{P}'\mathbf{P} \cdot \hat{\boldsymbol{\beta}}.$$

Even if only one variable is measured with error—that would correspond to a Σ_m matrix that is all zeroes except for one error variance on the diagonal—this correction can be used to assess the sensitivity of the estimators of the *other* slope parameters.

Correlated measurement errors

Again let $\mathbf{P} = \mathbf{Z} + \mathbf{m}$ but now allow $\operatorname{plim} n^{-1}\mathbf{Z}'\mathbf{m} = \mathrm{E}\, Z_i m_i = \rho \neq 0$. To keep the exposition simple, let's suppress the role of the $\mathbf{X}$ covariates, thereby avoiding the need to use the FWL theorem.

The structural model is now $\mathbf{Y} = \mathbf{Z}\delta + \epsilon$ or, expressed in terms of $\mathbf{P}$,

$$\mathbf{Y} = \mathbf{P}\delta + \epsilon - \delta\mathbf{m} = \mathbf{P}\delta + \mathbf{v}.$$

The OLS estimator is

$$\hat{\delta} = (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{Y} = \delta + (\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'\mathbf{v}.$$

The association between $\mathbf{P}$ and the composite disturbance is now

$$\frac{1}{n}\mathbf{P}'\mathbf{v} \stackrel{a}{=} -\delta(\rho + \sigma_m^2),$$

and also,

$$\frac{1}{n}\mathbf{P}'\mathbf{P} \stackrel{a}{=} \frac{1}{n}\mathbf{Z}'\mathbf{Z} + 2\rho + \sigma_m^2.$$

Hence, letting $\frac{1}{n}\mathbf{Z}'\mathbf{Z} \xrightarrow{p} \Omega$,

$$\hat{\delta} \xrightarrow{p} \delta\left(\frac{\Omega + \rho}{\sigma_m^2 + \Omega + 2\rho}\right).$$

Evidently if $\rho > 0$ we still have shrinkage, but if $\rho < 0$ the nature of the inconsistency is uncertain.

An example is provided by the case of a dummy variable $\mathbf{Z}_i$ whose value is mis-classified by $\mathbf{P}_i$, which is also a dummy variable. Writing $\mathbf{P}_i = \mathbf{Z}_i + m_i$, we have three cases to consider. When $m_i = 0$, there is no measurement error; let this case occur with probability p_0. When $m_i = 1$, then the only possible value for $\mathbf{Z}_i = 0$, and this case occurs with probability p_1. Finally, with $m_i = -1$ we can have $\mathbf{Z}_i = 1$, which occurs with probability p_{-1}. Calculating $\rho = \mathrm{E}\,\mathbf{Z}_i m_i$, we find that $\rho = -p_{-1} \leq 0$. So in this example, we cannot say for sure that $\hat{\delta}$ will shrink toward zero.

Errors in dependent and explanatory variables

In models of labor supply with the number of hours supplied depending on the wage rate, we often face the problem that measurement error in reported hours spills over to contaminate the calculated wage rate. Let H_i^* be the correct value of hours worked and let $H_i = H_i^* e^{m_i}$ be the mis-reported number of hours. The dependent variable for the labor supply regression is $\ln H_i = \ln H_i^* + m_i$. Let w_i^* be the true value of the wage rate and $w_i^* H_i^*$ the true value of labor earnings. We assume that total labor earnings are correctly reported (this assumption is easily generalized). The wage rate we would calculate by dividing total earnings by (mis-reported) hours worked is then

$$w_i = \frac{w_i^* H_i^*}{H_i} = \frac{w_i^*}{e^{m_i}}.$$

Hence, $\ln w_i = \ln w_i^* - m_i$, so that the caculated wage equals the true wage less the measurement error in reported hours worked.

Suppose for simplicity that the labor supply equation is linear in the logs of hours and wages,

$$\ln H_i^* = \delta \ln w_i^* + \epsilon_i.$$

After substituting, we obtain

$$\ln H_i = \delta \ln w_i + \epsilon_i + m_i(1+\delta) = \delta \ln w_i + v_i.$$

Hence,

$$\operatorname{plim} \frac{1}{n} \sum_i \ln w_i \cdot v_i = -\sigma_m^2 (1+\delta).$$

With $\operatorname{plim} n^{-1} \sum_i (\ln w_i)^2 = \Omega + \sigma_m^2$, using $\Omega = \operatorname{plim} n^{-1} \sum_i (\ln w_i^*)^2$, we obtain

$$\operatorname{plim} \hat{\delta} = \delta - \frac{\sigma_m^2(1+\delta)}{\sigma_m^2 + \Omega} = \delta \left(\frac{\Omega}{\sigma_m^2 + \Omega} \right) - \frac{\sigma_m^2}{\sigma_m^2 + \Omega}.$$

The last term on the right (which is negative in sign) may or may not counteract the shrinkage implied by the term that precedes it.

Griliches (1986) provides results for related cases, such as errors in nonlinear specifications such as $Y_i = \mathbf{X}_i' \beta + \delta_1 Z_i + \delta_2 Z_i^2 + \epsilon_i$.

Chapter 35

Missing, Proxy, and Predicted Explanatory Variables

Extending the analysis of the previous chapter, here we discuss the econometric properties of regression models that rely in one way or another on proxy variables. The discussion is couched in general terms, but we have in mind an application in which a demographic survey gathers data on various indicators that are collectively meant to represent the standard of living of the surveyed household. For cost or other reasons, data on household incomes and expenditures, the measures thought by most economists to be theoretically appropriate indicators of living standards, are not obtained. How is the researcher to proceed?

As before, our discussion is organized around the linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\delta + \epsilon \tag{35.1}$$

in which the dependent variable $\mathbf{Y}$ is $n \times 1$, the $\mathbf{X}$ matrix of explanatory variables is $n \times k$ and the $n \times 1$ vector $\mathbf{Z}$ represents the missing explanatory variable (e.g., income or consumption). We assume that the disturbance term ϵ has mean zero and covariance matrix $\sigma^2\mathbf{I}$. We will make the standard assumption about the relationships between ϵ, $\mathbf{X}$ and $\mathbf{Z}$, namely, that ϵ is weakly exogenous to both $\mathbf{X}$ and $\mathbf{Z}$. By "weakly exogenous" we mean that $\operatorname{plim} n^{-1}\mathbf{X}'\epsilon = \mathbf{0}$, and $\operatorname{plim} n^{-1}\mathbf{Z}'\epsilon = 0$. With mild additional assumptions, this will guarantee that least-squares estimators $\hat{\beta}$ and $\hat{\delta}$ will be consistent. Unfortunately, equation (35.1) cannot be estimated as it stands because the variable $\mathbf{Z}$ is unavailable.

35.1 Using the Determinants of the Missing Variable

One approach is to specify a model of the missing $\mathbf{Z}$ variable, say,

$$\mathbf{Z} = \mathbf{W}\theta + \mathbf{v}, \tag{35.2}$$

and substitute this model into equation (35.1). We obtain

$$\mathbf{Y} = \mathbf{X}\beta + (\mathbf{W}\theta + \mathbf{v})\delta + \epsilon = \mathbf{X}\beta + \mathbf{W}\theta\delta + \mathbf{v}\delta + \epsilon, \tag{35.3}$$

in which we have a composite disturbance term $\mathbf{v}\delta + \epsilon$. If the covariates $\mathbf{X}$ and $\mathbf{W}$ are both uncorrelated with $\mathbf{v}$ and $\boldsymbol{\epsilon}$, then equation (35.3) can be estimated without difficulty. Unfortunately, substitution gives us no means of identifying the effect of $\mathbf{Z}$ itself—that is, we cannot estimate the δ coefficient. In addition, $\mathbf{W}$ is likely to overlap with $\mathbf{X}$, and for any covariate that appears in both $\mathbf{X}$ and $\mathbf{W}$, only the sum of the associated $\boldsymbol{\beta}$ and $\theta\delta$ parameters is identified.

For an example of this approach, consider a model in which $\mathbf{Y}$ is a measure of individual health and the missing $\mathbf{Z}$ is a measure of individual income. Suppose that (as in the United States) we have from census data an estimate of the average income in the area (e.g., the census tract) in which the individual lives. Let the average income in the area be denoted by $\mathbf{Z}_a$ and imagine an equation that relates individual income $\mathbf{Z}$ to the areal average, $\mathbf{Z} = \theta_1 + \theta_2\mathbf{Z}_a + \mathbf{v}$. After substituting this equation into the health equation, the new constant term is $\beta_1 + \theta_1\delta$ and the new disturbance is the composite $\mathbf{v}\delta + \boldsymbol{\epsilon}$. If the structural health equation (the counterpart to (35.1)) does not include the areal average $\mathbf{Z}_a$, then the estimated coefficient on $\mathbf{Z}_a$ can be interpreted as an estimate of $\theta_2\delta$. However, there might be good reasons to consider an alternative structural model in which both individual income $\mathbf{Z}$ *and* areal average income $\mathbf{Z}_a$ appear in the health equation. If this is the correct specification of the health equation, then the coefficient on $\mathbf{Z}_a$ should be interpreted as $\beta_j + \theta_2\delta$ where β_j is the parameter attached to areal average income in the original structural model.

Clearly the substitution approach gives results that are less than satisfactory in general. In some cases, the researcher might discover independent estimates of the θ coefficients in the literature or be able to estimate them from other samples of data, and these estimates could be used to resolve the identification problem. We will return to this possibility in a moment. Otherwise, when there is no overlap between $\mathbf{X}$ and $\mathbf{W}$, then the substitution technique at least provides a "control" for the missing $\mathbf{Z}$ and allows the $\boldsymbol{\beta}$ coefficients to be estimated consistently. But when there is substantial overlap between $\mathbf{X}$ and $\mathbf{W}$, even this rationale is undermined.

Much of the problem with the substitution method is that it lacks enough structure to permit δ to be distinguished from θ. In some cases, helpful additional structure can supplied by factor-analytic specifications in which $\mathbf{Z}$ is modelled as an unobserved latent variable. Perhaps the most promising of these approaches—to be discussed later in this handout—is the MIMIC method (an acronym for "multiple indicator, multiple cause") which is applicable when we have both multiple proxies for the unobserved $\mathbf{Z}$ and multiple determinants of $\mathbf{Z}$. Such methods are in widespread use in quantitative sociology and psychology, but to date have seldom been seen in econometrics.

35.2 Using Predicted Values

Suppose that we are in the fortunate position of having access to a predicted version of the missing $\mathbf{Z}$ variable. For instance, in developing countries few general-purpose data sets include information on household income or consumption, but specialized surveys can be used to predict the values of these variables. What are the econometric consequences of using $\hat{\mathbf{Z}}$, a predicted version of $\mathbf{Z}$, in the regression equation? This problem has been

analyzed in some detail by Elbers, Lanjouw, and Lanjouw (2003) and Elbers, Lanjouw, and Lanjouw (2005), and there is now a large literature in what is termed "small area poverty mapping" to draw upon for methods and applications.

To be specific about the set-up, we assume that although our main dataset does not contain $\mathbf{Z}$, it does contain the determinants $\mathbf{W}$ of $\mathbf{Z}$ as specified in equation (35.2) above. Moreover, we have access to another data set that holds both $\mathbf{Z}$ and $\mathbf{W}$, and can use this extra dataset to obtain estimates of the θ parameters. Our ability to combine data in this way depends crucially on what we assume about the data-generating process for $\mathbf{Z}$. If we can say that the specification $\mathbf{Z} = \mathbf{W}\theta + \mathbf{v}$ holds for both datasets—this would be a questionable assertion if, for instance, the survey with both $\mathbf{Z}$ and $\mathbf{W}$ was administered to a selected sub-set of the population, or was fielded long before or long after the other dataset—and if we can further assume that θ can be estimated consistently by ordinary least squares given data on $\mathbf{Z}$ and $\mathbf{W}$, then a sound basis exists for combining the data.

For the dataset with both $\mathbf{Z}$ and $\mathbf{W}$, we write the $\mathbf{Z}$ equation as

$$\mathbf{Z}_1 = \mathbf{W}_1\theta + \mathbf{v}_1, \tag{35.4}$$

using the subscript "1" to indicate that these data are found in Dataset 1 of our two sets of data. Under standard assumptions, in particular that plim $\frac{1}{n_1}\mathbf{W}_1'\mathbf{v}_1 = \mathbf{0}$, with n_1 being the size of Dataset 1, then $\hat{\theta}_1$ is consistent for θ.

Now we turn to Dataset 2, the main sample of data. We write the structural equation as

$$\mathbf{Y}_2 = \mathbf{X}_2\beta + \mathbf{Z}_2\delta + \epsilon_2 \tag{35.5}$$

with subscript "2" indicating that the model applies to the second dataset. In Dataset 2, we have explanatory variables $\mathbf{W}_2$ available as well as $\mathbf{Y}_2$ and $\mathbf{X}_2$; only $\mathbf{Z}_2$ is lacking.

However, the missing $\mathbf{Z}_2$ variable is generated by the equation $\mathbf{Z}_2 = \mathbf{W}_2\theta + \mathbf{v}_2$. Denote by $\hat{\mathbf{Z}}_2$ the predicted value of $\mathbf{Z}_2$ given $\mathbf{W}_2$,

$$\hat{\mathbf{Z}}_2 = \mathbf{W}_2\hat{\theta}_1 = \mathbf{W}_2\theta + \mathbf{W}_2(\mathbf{W}_1'\mathbf{W}_1)^{-1}\mathbf{W}_1'\mathbf{v}_1 = \mathbf{W}_2\theta + \mathbf{W}_2(\hat{\theta}_1 - \theta).$$

Subtracting $\hat{\mathbf{Z}}_2$ from $\mathbf{Z}_2$ and rearranging, we have

$$\mathbf{Z}_2 = \hat{\mathbf{Z}}_2 + \mathbf{v}_2 - \mathbf{W}_2(\hat{\theta}_1 - \theta).$$

We now rewrite the structural equation in terms of the predicted welfare measure $\hat{\mathbf{Z}}_2$, as

$$\mathbf{Y}_2 = \mathbf{X}_2\beta + \hat{\mathbf{Z}}_2\delta + (\mathbf{v}_2 - \mathbf{W}_2(\hat{\theta}_1 - \theta)) \cdot \delta + \epsilon_2 = \mathbf{X}_2\beta + \hat{\mathbf{Z}}_2\delta + \mathbf{u}_2.$$

We can easily show that $\hat{\mathbf{Z}}_2$ is asymptotically uncorrelated with the composite disturbance term $\mathbf{u}_2$. The key to the result is to think of both $n_1 \to \infty$ and $n_2 \to \infty$, and then to see that

$$\text{plim}\,\frac{1}{n_2}\hat{\mathbf{Z}}_2'(\mathbf{v}_2 - \mathbf{W}_2(\hat{\theta}_1 - \theta))$$

is

$$\text{plim}\,\hat{\theta}_1'\,\text{plim}\,\frac{1}{n_2}\mathbf{W}_2'\mathbf{v}_2 - \text{plim}\,\hat{\theta}_1'\,\text{plim}\,\frac{1}{n_2}\mathbf{W}_2'\mathbf{W}_2\,\text{plim}(\hat{\theta}_1 - \theta)).$$

The (assumed) consistency of $\hat{\theta}_1$, and the key assumption plim $\frac{1}{n_2}\mathbf{W}_2'\mathbf{v}_2 = \mathbf{0}$ that is linked to consistency, give us the result. Note that we must also assume that $\mathbf{W}_2$ is uncorrelated with ϵ_2, the structural disturbance

It follows that both δ and β are consistently estimated in the second sample even though we have used the predicted welfare variable $\hat{\mathbf{Z}}_2$ instead of the true $\mathbf{Z}_2$. This is potentially a delightful result. Unfortunately, the composite disturbance term has a decidedly non-scalar covariance matrix, which we could think of as

$$\Omega = \text{Var}\left((\mathbf{v}_2 - \mathbf{W}_2(\hat{\theta}_1 - \theta)) \cdot \delta + \epsilon_2\right) = (\delta^2\sigma_v^2 + \sigma_\epsilon^2)\mathbf{I} + \delta^2\mathbf{W}_2\Sigma_1\mathbf{W}_2'.$$

Hence, if we estimate the structural equation by ordinary least squares, the standard errors of its regression coefficients will certainly need to be corrected.

35.3 Using Proxy Variables

In the remainder of this chapter, we proceed on the assumption that $\mathbf{Z}$ is missing and that we have either mis-measured or proxy variables in the dataset that can be used in its place. In contrast to the approach just explored above, we do not assume here that we have articulated a model of $\mathbf{Z}$ and are in the fortunate position of being able to estimate its parameters using a second dataset. Instead, we simply assume that one or more proxy variables $\mathbf{P}$ are available to stand in for the unobserved $\mathbf{Z}$.

What guidance is available from theory about the advantages to be gained from using such proxies? Should they be inserted in the model in place of $\mathbf{Z}$, or would it be better to ignore $\mathbf{P}$ altogether? Intuition suggests that so long as $\mathbf{P}$ contains useful information about $\mathbf{Z}$, the better course of action must be to include $\mathbf{P}$ in the model being estimated. This is a sensible-sounding argument, and yet it is surprisingly difficult to establish its statistical foundation. In what follows we will study several particular and rather special cases that lend support to the use of proxies. In general, however, there can be no guarantee that making use of proxy variables must improve the situation.

Suppose that a proxy variable $\mathbf{P}$, or a set of l such proxies, is available for $\mathbf{Z}$, the missing covariate. The proxy $\mathbf{P}$ has dimension $n \times l$. In using $\mathbf{P}$ in place of $\mathbf{Z}$, we estimate the misspecified equation

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{P}\mathbf{d} + \text{ residuals}, \tag{35.6}$$

$\mathbf{d}$ being $l \times 1$, where in writing "residuals" we mean to emphasize the difference between the true disturbances ϵ that appear in equation (35.1) and the messy composite of true disturbances and specification errors that appears in equation (35.6).

We now explore the statistical properties of the least-squares estimators $\hat{\mathbf{b}}$ and $\hat{\mathbf{d}}$ in relation to the parameters of interest, β and δ. In the ensuing analyses, we will maintain the assumption that like $\mathbf{X}$ and $\mathbf{Z}$, the proxies $\mathbf{P}$ are also weakly exogenous to the disturbance ϵ.

Estimates of β

Using the FWL theorem we obtain the expression below for $\hat{\mathbf{b}}$:

$$\begin{aligned}\hat{\mathbf{b}} &= (\mathbf{X}'\mathbf{M}_\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{P}\mathbf{Y} \\ &= \beta + (\mathbf{X}'\mathbf{M}_\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{P}\mathbf{Z}\delta + (\mathbf{X}'\mathbf{M}_\mathbf{P}\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_\mathbf{P}\epsilon \end{aligned} \tag{35.7}$$

where $\mathbf{M_P} = \mathbf{I} - \mathbf{P}(\mathbf{P}'\mathbf{P})^{-1}\mathbf{P}'$. Given that $\mathbf{P}$ is weakly exogenous to ϵ, it is straightforward to show that

$$\hat{\mathbf{b}} \xrightarrow{p} \beta + \text{plim}(n^{-1}\mathbf{X}'\mathbf{M_P}\mathbf{X})^{-1} \text{plim}(n^{-1}\mathbf{X}'\mathbf{M_P}\mathbf{Z})\delta. \tag{35.8}$$

Evidently $\hat{\mathbf{b}}$ is inconsistent for β in general. However, recall that in the previous sections, we examined a special case in which $\mathbf{P}$ happens to be the same set of variables as contained in $\mathbf{W}$, the determinants of the missing $\mathbf{Z}$, and those results are relevant here. (Note that in defining $\mathbf{P}$ to be "proxies" for $\mathbf{Z}$, meaning that $\mathbf{P}$ would have no role to play in the equation if $\mathbf{Z}$ were available, we rule out the possibility of overlap between $\mathbf{P}$ and $\mathbf{X}$.) It is at least possible, therefore, for $\hat{\mathbf{b}}$ to be consistent.

Estimates of δ

Applying the same techniques to the estimator $\hat{\mathbf{d}}$, we obtain

$$\begin{aligned}
\hat{\mathbf{d}} &= (\mathbf{P}'\mathbf{M_X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{M_X}\mathbf{Y} \\
&= (\mathbf{P}'\mathbf{M_X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{M_X}\mathbf{Z}\,\delta + (\mathbf{P}'\mathbf{M_X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{M_X}\epsilon \qquad (35.9) \\
&= \hat{\theta}_n\,\delta + (\mathbf{P}'\mathbf{M_X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{M_X}\epsilon \qquad (35.10)
\end{aligned}$$

where we have used the fact that $\mathbf{M_X}\mathbf{X} = \mathbf{0}$ and labelled as $\hat{\theta}_n$ what can be recognized as the coefficients (via FWL) on the proxy variables $\mathbf{P}$ coming from a (hypothetical) regression of the missing $\mathbf{Z}$ on both $\mathbf{P}$ and $\mathbf{X}$. This is one measure of the "strength" of the proxies in explaining the missing $\mathbf{Z}$ net of the contribution of the $\mathbf{X}$ variables.

Upon taking probability limits, we obtain

$$\hat{\mathbf{d}} \xrightarrow{p} \text{plim}\,\hat{\theta}_n\,\delta = \theta\,\delta, \tag{35.11}$$

which is analogous to the relationship one sees in analyses of the consequences of missing variables, which is more or less the case that we have in front of us. The large-sample bias of $\hat{d}$, so to speak, thus depends on how the true $\mathbf{Z}$ is related to its proxy $\mathbf{P}$ net of the $\mathbf{X}$ variables. Now, $\hat{\mathbf{d}}$ and δ will generally be of different dimensions, this occurring when we employ a set of $l > 1$ proxies for the single unobserved variable $\mathbf{Z}$. Even if $\hat{d}$ and δ are both of dimension one, however, equation (35.11) shows that $\hat{d}$ is inconsistent for δ.

Estimating σ^2

We should consider how the introduction of proxies $\mathbf{P}$ affects the estimator of σ_ϵ^2, the variance of the true disturbance term ϵ. If we carry out the regression $\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{P}\mathbf{d} + \text{residuals}$, the least-squares residual vector is $\mathbf{e} = \mathbf{M}\mathbf{Y}$, where $\mathbf{M}$ is the symmetric idempotent matrix analogous to $\mathbf{M_X}$ and $\mathbf{M_P}$ above but which involves both $\mathbf{X}$ and $\mathbf{P}$. As before, $\mathbf{M}\mathbf{X} = \mathbf{0}$, and therefore $\mathbf{e}$ can be expressed as

$$\mathbf{e} = \mathbf{M}\mathbf{Y} = \mathbf{M}\mathbf{Z}\delta + \mathbf{M}\epsilon.$$

Maintaining the assumption that $\mathbf{P}$, $\mathbf{Z}$ and $\mathbf{X}$ are all weakly exogenous to ϵ, we have

$$\begin{aligned}
\text{plim}\,\frac{\mathbf{e}'\mathbf{e}}{n} &= \text{plim}\, n^{-1}\delta'\mathbf{Z}'\mathbf{M}\mathbf{Z}\delta + \text{plim}\, n^{-1}\epsilon'\mathbf{M}\epsilon \\
&= \text{plim}\, n^{-1}\delta'\mathbf{Z}'\mathbf{M}\mathbf{Z}\delta + \sigma_\epsilon^2 \qquad (35.12)
\end{aligned}$$

Since $\mathbf{e}'\mathbf{e}/n$ is the standard estimator of the variance, we see that it will converge to a value that exceeds the true variance σ_ϵ^2.

Summary

One conclusion seems inescapable: If $\mathbf{Z}$ is not itself available, the parameters of a regression model estimated with $\mathbf{P}$ as a proxy for $\mathbf{Z}$ will be inconsistent, at least in general. What we will be examining in the next few sections is the *degree* of inconsistency. As a prelude to this analysis, we turn now to the question of hypothesis testing.

35.4 Testing Hypotheses about δ

When estimators are inconsistent, this usually invalidates hypothesis tests. However, for a certain kind of hypothesis one can formulate a perfectly valid test by making use of proxy variables $\mathbf{P}$. We refer to a test that is focused on the relevance of the omitted variables $\mathbf{Z}$, that is, a test for $\delta = 0$. To study this test, we explore the properties of the estimator $\hat{\mathbf{d}}$ derived above.

A χ^2 test for $\delta = 0$

Under the null hypothesis $\delta = 0$, the estimator $\hat{\mathbf{d}}$ reduces to

$$\hat{\mathbf{d}} = (\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{M}_\mathbf{X}\epsilon.$$

It is then easy to establish that

$$\hat{\mathbf{d}} \xrightarrow{p} \mathbf{0}.$$

On making the additional assumption that $\sqrt{n}\frac{1}{n}\mathbf{P}'\mathbf{M}_\mathbf{X}\epsilon \xrightarrow{d} \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q})$ with $\mathbf{Q} = \text{plim } n^{-1}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P}$, we obtain

$$\sqrt{n}\hat{\mathbf{d}} \xrightarrow{d} \mathcal{N}\left(\mathbf{0}, \sigma^2\mathbf{Q}^{-1}\right).$$

This last result provides the ingredients for a χ^2 test statistic T,

$$T = \sqrt{n}\hat{\mathbf{d}}'\left(\sigma^2\mathbf{Q}^{-1}\right)^{-1}\sqrt{n}\hat{\mathbf{d}} \xrightarrow{d} \chi_l^2 \tag{35.13}$$

where l, the number of degrees of freedom, is the number of proxy variables used in the test. The larger is l—the more proxy variables one uses—the greater is the value of T required to reject the null hypothesis $\delta = 0$.

Better proxies, better tests

For the χ_l^2 test in (35.13) to be worth considering, it must possess reasonable power. In other words, if the truth is that $\delta \neq 0$, the test should reject the null hypothesis $\delta = 0$ with high probability. The power of this test is derived in large part from the degree of correlation between the unobserved $\mathbf{Z}$ and the proxy variables $\mathbf{P}$. MacKinnon (1992), Davidson and

MacKinnon (1993) and Pagan and Hall (1983) present analyses of the power function. The discussion below follows MacKinnon (1992).

Our aim is to derive the distribution of the test statistic T when $\delta \neq 0$. In an asymptotic analysis of test power, we typically make use of a device known as "Pitman drift," whereby if the null hypothesis is $\delta = 0$, the alternative hypothesis is expressed in terms that are dependent on sample size, $\delta_n = \frac{1}{\sqrt{n}}\tau$, where τ is a scalar chosen to be positive or negative according to the direction of departure from the null that we want to investigate. Inserting δ_n in place of δ, we have

$$\hat{\mathbf{d}} = (\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P})^{-1}\mathbf{P}'\mathbf{M}_\mathbf{X}(\mathbf{Z}\sqrt{n}\frac{1}{n}\tau + \epsilon)$$

and then

$$\sqrt{n}\hat{\mathbf{d}} = \hat{\theta}_n\ \tau + (n^{-1}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P})^{-1}\sqrt{n}\frac{1}{n}\mathbf{P}'\mathbf{M}_\mathbf{X}\epsilon.$$

Let $\mathbf{Q} = \text{plim}\, n^{-1}\mathbf{P}'\mathbf{M}_\mathbf{X}\mathbf{P}$. Then $\sqrt{n}\hat{\mathbf{d}}$ converges in distribution to a vector of constants, $\theta\,\tau$, plus a random variable that converges to $\mathcal{N}(\mathbf{0}, \sigma^2\mathbf{Q}^{-1})$. Thus $\sqrt{n}\hat{d}$ is itself $\mathcal{N}(\theta\tau, \sigma^2\mathbf{Q}^{-1})$ in the limit.

Returning to equation (35.13), we see that with $\tau \neq 0$ the quadratic form defining T is distributed as *non-central* χ^2_l with non-centrality parameter λ, where

$$\lambda \stackrel{a}{=} \frac{1}{2}\tau'\theta'\ \left(\sigma^2\mathbf{Q}^{-1}\right)^{-1}\theta\tau$$

This indicates that the larger the magnitude of the θ coefficients, the larger is the non-centrality parameter λ and thus the greater is the power of the test. In other words: Better proxies produce better tests. See Montgomery et al. (2000) for more discussion and an alternative development of the non-centrality parameter.

How many proxies?

The number of proxies used in constructing the χ^2_l test, that is, l, has two opposing effects on the power of this test. We expect the R^2 of regression (**??**) to increase as the number of proxies increases, in that with more $\mathbf{P}$ variables, $\mathbf{M}_\mathbf{X}\mathbf{P}$ should better explain $\mathbf{M}_\mathbf{X}\mathbf{Z}$. Taken by itself, this will increase the power of the test. However, the degrees of freedom of the test will also increase with l, so that as l rises the test statistic T must be compared to higher and higher critical values. This will reduce the power of the test. The net effect of increases in the number of proxy variables is therefore ambiguous.

The MacKinnon (1992) view is that in small samples, in which test power would be low in any case, it would be best to use one or only a few proxy variables $\mathbf{P}$ so as to keep the degrees of freedom low. In a small sample, one might well want to choose the best single proxy for $\mathbf{Z}$, where by "best" we mean the proxy that arguably has the highest R^2 in regression (**??**). As the sample size grows, however, one can begin to consider using more proxy variables in the hope that the cost in degrees of freedom will be more than offset by the improvement in the R^2.

To sum up, the analysis above shows that for some purposes, namely tests of the hypothesis $\delta = 0$, it makes sense to use proxy variables provided that they are reasonably

well correlated, net of $\mathbf{X}$, with the unobserved $\mathbf{Z}$. The reasoning we have used *does not* extend to other hypotheses about δ, however. We really cannot hope to test more general hypotheses, such as $\delta = \delta_0$, unless $\delta_0 = 0$.

35.5 Why Use a Proxy at All?

If one's substantive interest centers on β rather than δ, it is always possible to estimate the β parameter by ignoring altogether $\mathbf{Z}$ and its proxies. That is, one could apply least squares to the misspecified regression equation

$$\mathbf{Y} = \mathbf{X}\mathbf{b}_2 + \text{ residuals.} \tag{35.14}$$

The estimator implied by this approach, $\hat{\mathbf{b}}_2$, is in general inconsistent for β. It is easily shown that

$$\hat{\mathbf{b}}_2 - \beta \xrightarrow{p} \mathbf{Q}_{\mathbf{XX}}^{-1}\mathbf{Q}_{\mathbf{XZ}}\delta, \tag{35.15}$$

where $\mathbf{Q}_{\mathbf{XX}} = \operatorname{plim} n^{-1}\mathbf{X}'\mathbf{X}$ and $\mathbf{Q}_{\mathbf{XZ}} = \operatorname{plim} n^{-1}\mathbf{X}'\mathbf{Z}$, the former being a $k \times k$ matrix and the latter a $k \times 1$ vector. An interesting question is whether the large-sample bias in $\hat{\mathbf{b}}_2$ exceeds the bias in $\hat{\mathbf{b}}$ that was shown in equation (35.7).

The answer is not immediately obvious from a comparison of equations (35.7) and (35.15). We can see that the large-sample biases of $\hat{\mathbf{b}}$ and $\hat{\mathbf{b}}_2$ will generally differ, but from the expressions presented in these equations it cannot be determined which of the two has smaller bias. The cross-product matrices implicit in (35.7) involve $\mathbf{X}$, $\mathbf{P}$ and $\mathbf{Z}$, whereas in (35.15) only products involving $\mathbf{X}$ and $\mathbf{Z}$ appear. In what follows we attempt to simplify (35.7) by specifying various auxiliary models that link $\mathbf{P}$ to $\mathbf{Z}$.

Case 1

Wickens (1972) has explored a special case in which there is a single proxy variable $\mathbf{P}$ whose relationship to $\mathbf{Z}$ is described by

$$\mathbf{P} = \mathbf{Z}\theta + \mathbf{u}. \tag{35.16}$$

To analyze this case, Wickens assumes that $\mathbf{u}$ is uncorrelated with $\mathbf{Z}$ as well as being uncorrelated with $\mathbf{X}$ and ϵ. With these (strong) assumptions, equation (35.16) resembles the classical measurement error model, which would be produced by the further condition $\theta = 1$. Our treatment here is a slight generalization of the approach pursued earlier.

The true model, equation (35.1), can be re-expressed in terms of $\mathbf{X}$ and $\mathbf{P}$ as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{P}(\delta/\theta) + \epsilon - \mathbf{u}(\delta/\theta) \tag{35.17}$$

or more compactly as

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{P}d + \mathbf{v} \tag{35.18}$$

where $d = \delta/\theta$ and $\mathbf{v} = \epsilon - \mathbf{u}d$. Note that $\mathbf{P}$ is correlated with $\mathbf{v}$.

As earlier, we consider the standard decomposition of the least-squares estimator,

$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} \beta \\ d \end{bmatrix} + \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{P} \\ \mathbf{P}'\mathbf{X} & \mathbf{P}'\mathbf{P} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}'\mathbf{v} \\ \mathbf{P}'\mathbf{v} \end{bmatrix},$$

and from this derive the probability limit of $\hat{\mathbf{b}}$. We already know that both $\hat{\mathbf{b}}$ and $\hat{d}$ are inconsistent for their respective parameters. Our goal is to find a simple expression for the bias of $\hat{\mathbf{b}}$ that can be compared to the bias of $\hat{\mathbf{b}}_2$.

From $\operatorname{plim} n^{-1}\mathbf{X}'\mathbf{v} = \mathbf{0}$ and $\operatorname{plim} n^{-1}\mathbf{P}'\mathbf{v} = -d\sigma_u^2$, we have

$$\operatorname{plim} \begin{bmatrix} \hat{\mathbf{b}} - \beta \\ \hat{d} - d \end{bmatrix} = \begin{bmatrix} \mathbf{Q_{XX}} & \mathbf{Q_{XP}} \\ \mathbf{Q_{PX}} & \mathbf{Q_{PP}} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -d\sigma_u^2 \end{bmatrix}$$

Using a partitioned inversion formula[1] and recalling that $d = \delta/\theta$, we find that

$$\hat{\mathbf{b}} - \beta \xrightarrow{p} \mathbf{Q_{XX}^{-1}Q_{XP}} \left(\mathbf{Q_{PP}} - \mathbf{Q_{PX}Q_{XX}^{-1}Q_{XP}}\right)^{-1} (\delta/\theta)\sigma_u^2.$$

This is in itself a useful result. It shows that the *direction* of bias in $\hat{\mathbf{b}}$ can be estimated if the researcher is willing to make assumptions about the signs of δ and θ. The other quantities that appear in the expression—$\mathbf{Q_{PP}}$, $\mathbf{Q_{PX}}$ and $\mathbf{Q_{XX}}$—can all be estimated from their sample counterparts.

Returning to the main line of argument, we now substitute the right-hand side of equation (35.16) for $\mathbf{P}$ and obtain $\mathbf{Q_{XP}} = \mathbf{Q_{XZ}}\theta$. We make this substitution for the first occurrence of $\mathbf{Q_{XP}}$ above, which yields

$$\hat{\mathbf{b}} - \beta \xrightarrow{p} \mathbf{Q_{XX}^{-1}Q_{XZ}}\delta \left(\mathbf{Q_{PP}} - \mathbf{Q_{PX}Q_{XX}^{-1}Q_{XP}}\right)^{-1} \sigma_u^2.$$

This expression is the product of two factors. The first, a $k \times 1$ vector $\mathbf{Q_{XX}^{-1}Q_{XZ}}\delta$, is the large-sample bias of $\hat{\mathbf{b}}_2$, the estimator that does not make use of $\mathbf{P}$. To know whether including the proxy variable $\mathbf{P}$ reduces large-sample bias relative to this benchmark, we must consider the size of the second factor, which is the scalar

$$\left(\mathbf{Q_{PP}} - \mathbf{Q_{PX}Q_{XX}^{-1}Q_{XP}}\right)^{-1} \sigma_u^2.$$

Again using equation (35.16) for $\mathbf{P}$, we find that

$$\mathbf{Q_{PP}} - \mathbf{Q_{PX}Q_{XX}^{-1}Q_{XP}} = \sigma_u^2 + \theta^2 \left(\mathbf{Q_{ZZ}} - \mathbf{Q_{ZX}Q_{XX}^{-1}Q_{XZ}}\right),$$

and it is helpful to write the term in parentheses as $\mathbf{Q_{ZZ}}(1 - R^2_{\mathbf{Z},\mathbf{X}})$, where by $R^2_{\mathbf{Z},\mathbf{X}}$ we mean the probability limit of the R^2 from a regression of the true variable $\mathbf{Z}$ on $\mathbf{X}$. Using this notation, we obtain

$$\hat{\mathbf{b}} - \beta = \mathbf{Q_{XX}^{-1}Q_{XZ}}\delta \left(\frac{\sigma_u^2}{\sigma_u^2 + \theta^2\mathbf{Q_{ZZ}}(1 - R^2_{\mathbf{Z},\mathbf{X}})}\right).$$

Since the scalar in large parentheses is less than unity, the large-sample bias of $\hat{\mathbf{b}}$ is clearly smaller than $\mathbf{Q_{XX}^{-1}Q_{XZ}}\delta$, the bias of $\hat{\mathbf{b}}_2$. Thus, under the assumptions we have employed, it is unambiguously better to use the proxy variable $\mathbf{P}$ than to ignore it. Although $\mathbf{P}$ is only a proxy for $\mathbf{Z}$, it adds useful information that serves to reduce—if not eliminate—the bias in the least-squares estimator of β.

[1] If the matrix $\mathbf{Q}$ is partitioned as

$$\mathbf{Q} = \begin{bmatrix} \mathbf{Q_{XX}} & \mathbf{Q_{XP}} \\ \mathbf{Q_{PX}} & \mathbf{Q_{PP}} \end{bmatrix}$$

the upper right element of its inverse is $-\mathbf{Q_{XX}^{-1}Q_{XP}}(\mathbf{Q_{PP}} - \mathbf{Q_{PX}Q_{XX}^{-1}Q_{XP}})^{-1}$.

Case 2

The case analyzed by Wickens (1972) can be generalized to allow the proxy variable to be influenced by the $\mathbf{X}$ covariates as well as the unobservable $\mathbf{Z}$. We rewrite equation (35.16) as

$$\mathbf{P} = \mathbf{X}\alpha + \mathbf{Z}\theta + \mathbf{u}. \tag{35.19}$$

Our revised model can now be expressed as

$$\mathbf{Y} = \mathbf{X}\left(\beta - \alpha(\delta/\theta)\right) + \mathbf{P}(\delta/\theta) + \epsilon - \mathbf{u}(\delta/\theta) \tag{35.20}$$

or more compactly as

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{P}d + \mathbf{v} \tag{35.21}$$

where $\mathbf{b} = \beta - \alpha(\delta/\theta)$, $d = \delta/\theta$ and $\mathbf{v} = \epsilon - \mathbf{u}(\delta/\theta)$.

We derive $\text{plim}(\hat{\mathbf{b}} - \mathbf{b})$ following the same steps as above. First, write

$$\text{plim}\begin{bmatrix} \hat{\mathbf{b}} - \mathbf{b} \\ \hat{d} - d \end{bmatrix} = \begin{bmatrix} \mathbf{Q_{XX}} & \mathbf{Q_{XP}} \\ \mathbf{Q_{PX}} & \mathbf{Q_{PP}} \end{bmatrix}^{-1} \begin{bmatrix} 0 \\ -d\sigma_u^2 \end{bmatrix}.$$

Again using partitioned inversion, we find that

$$\hat{\mathbf{b}} - \mathbf{b} \xrightarrow{p} \mathbf{Q_{XX}^{-1}Q_{XP}}\left(\mathbf{Q_{PP}} - \mathbf{Q_{PX}Q_{XX}^{-1}Q_{XP}}\right)^{-1}(\delta/\theta)\sigma_u^2.$$

We now substitute the right side of the equation above for the proxy $\mathbf{P}$ to obtain $\mathbf{Q_{XP}} = \mathbf{Q_{XX}}\alpha + \mathbf{Q_{XZ}}\theta$. With some manipulation, we find for the scalar in parentheses

$$\mathbf{Q_{PP}} - \mathbf{Q_{PX}Q_{XX}^{-1}Q_{XP}} = \sigma_u^2 + \theta^2\left(\mathbf{Q_{ZZ}} - \mathbf{Q_{ZX}Q_{XX}^{-1}Q_{XZ}}\right)$$

and as before this can be written as $\mathbf{Q_{ZZ}}(1 - R^2_{\mathbf{Z},\mathbf{X}})$. We finally obtain

$$\begin{aligned} \hat{\mathbf{b}} - \mathbf{b} \quad \xrightarrow{p} \alpha(\delta/\theta) \quad & \left(\frac{\sigma_u^2}{\sigma_u^2 + \theta^2\mathbf{Q_{ZZ}}(1 - R^2_{\mathbf{Z},\mathbf{X}})}\right) \\ + \quad & \mathbf{Q_{XX}^{-1}Q_{XZ}}\delta\left(\frac{\sigma_u^2}{\sigma_u^2 + \theta^2\mathbf{Q_{ZZ}}(1 - R^2_{\mathbf{Z},\mathbf{X}})}\right). \end{aligned}$$

Recall that $\mathbf{b} = \beta - \alpha(\delta/\theta)$; therefore we can write

$$\begin{aligned} \hat{\mathbf{b}} - \beta \quad \xrightarrow{p} \alpha(\delta/\theta) \quad & \left(\frac{\sigma_u^2}{\sigma_u^2 + \theta^2\mathbf{Q_{ZZ}}(1 - R^2_{\mathbf{Z},\mathbf{X}})} - 1\right) \\ + \quad & \mathbf{Q_{XX}^{-1}Q_{XZ}}\delta\left(\frac{\sigma_u^2}{\sigma_u^2 + \theta^2\mathbf{Q_{ZZ}}(1 - R^2_{\mathbf{Z},\mathbf{X}})}\right). \end{aligned}$$

Note if we were to set α equal to zero we would obtain the probability limit found by Wickens (1972). In this more general case, however, the decision of whether to use of a proxy or omit it is no longer as straightfoward The decision now turns on the sign and size of the vector $\alpha(\delta/\theta)$.

Part VI

Advanced Topics

Chapter 36

Gaussian Quadrature

36.1 Overview

The method of Gaussian quadrature is often applied when one desires a good numerical approximation to an integral of the type

$$\int_{-\infty}^{\infty} e^{-\epsilon^2} P(\epsilon) d\epsilon$$

where the function $P(\epsilon) > 0$ and the integral in question cannot be represented in a closed form. The quadrature method approximates this integral by a weighted summation over a pre-selected number of quadrature points. The method is explained in illuminating detail by Press et al. (1992) and Press et al. (1996), who provide additional references as well as programming subroutines that will calculate the relevant quadrature points and weights.

36.2 A Single Random Effect

We will be using quadrature methods in a panel or clustered data context, in which a persistent unobserved random variable u, termed an *error component* or *random effect*, influences the probability associated with a discrete-valued dependent variable. This note will provide some necessary introductory material, so as to lay a foundation for the more detailed discussions to follow.

We are typically led to integrals of the form above by consideration of normally-distributed random effects. Let u be distributed as a normal, mean zero, variance σ_{11} variable. The integral we seek to approximate is then

$$\int_{-\infty}^{\infty} (2\pi)^{-1/2} \sigma_{11}^{-1/2} e^{-\frac{1}{2\sigma_{11}} u^2} P(u) du.$$

To put this in the form above, we need only make a change of variables

$$\epsilon = \frac{u}{\sqrt{2}\sigma_1}$$

with σ_1 being the standard deviation of u. The transformation implies $u^2 = 2\sigma_{11}\epsilon^2$ and

$$\frac{du}{d\epsilon} = \sqrt{2}\sigma_1.$$

On making the change of variables, we obtain

$$\pi^{-1/2} \int e^{-\epsilon^2} P(\sqrt{2}\sigma_1 \epsilon) d\epsilon,$$

which is in the required form apart from constants. The quadrature method approximates the integral by the sum

$$\sum_{j=1}^{n_1} w_{1,j} P(\sqrt{2}\sigma_1 e_{1,j})$$

over $n_1 > 1$ quadrature points. The quadrature points $e_{1,j}$ are symmetric about zero, as are the weights $w_{1,j}$ with which they are associated. The number of points n_1 is under the control of the researcher, but the approximation improves as the number of points increases.

36.3 Two Independent Random Effects

The use of Gaussian quadrature with two random effects, so long as these are normally-distributed and independent, presents few additional complications. Let u_1 and u_2 be the random effects in question, each having mean zero and variances σ_{11} and σ_{22} respectively. The integral that we will wish to approximate is

$$\int_{u_1} \int_{u_2} (2\pi)^{-1} (\sigma_{11}\sigma_{22})^{-1/2} e^{-\frac{1}{2\sigma_{11}} u_1^2} e^{-\frac{1}{2\sigma_{22}} u_2^2} P(u_1, u_2) du_1 du_2.$$

To concentrate attention first on the u_2 dimension, we set off the relevant terms in braces,

$$(2\pi)^{-1} \sigma_1^{-1} \int_{u_1} e^{-\frac{1}{2\sigma_{11}} u_1^2} \left\{ \sigma_2^{-1} \int_{u_2} e^{-\frac{1}{2\sigma_{22}} u_2^2} P(u_1, u_2) du_2 \right\} du_1.$$

We must first reshape this expression into a form compatible with quadrature, by making a change of variables for u_2,

$$\epsilon_2 = \frac{u_2}{\sqrt{2}\sigma_2}.$$

This gives $u_2^2 = 2\sigma_{22}\epsilon_2^2$, and

$$\frac{du_2}{d\epsilon_2} = \sqrt{2}\sigma_2.$$

With this change of variables, the full integral becomes

$$2^{-1/2} \pi^{-1} \sigma_1^{-1} \int_{u_1} e^{-\frac{1}{2\sigma_{11}} u_1^2} \left\{ \int_{\epsilon_2} e^{-\epsilon_2^2} P(u_1, \sqrt{2}\sigma_2 \epsilon_2) d\epsilon_2 \right\} du_1.$$

As was seen earlier, the rightmost integral can be approximated using a set of quadrature weights $\{w_{2,k}\}$ and points $\{e_{2,k}\}$,

$$\sum_{k=1}^{n_2} w_{2,k} P(u_1, \sqrt{2}\sigma_2 e_{2,k})$$

where n_2 is the number of quadrature points chosen for this dimension. The approximation leaves us with an expression

$$2^{-1/2}\pi^{-1}\sigma_1^{-1}\int_{u_1} e^{-\frac{1}{2}\frac{1}{\sigma_{11}}u_1^2}\left\{\sum_k^{n_2} w_{2,k}P(u_1, \sqrt{2}\sigma_2 e_{2,k})\right\}du_1$$

that requires one further round of manipulation. Making a second change of variables,

$$\epsilon_1 = \frac{u_1}{\sqrt{2}\sigma_1},$$

and applying the same logic as before yields

$$\pi^{-1}\int e^{-\epsilon_1^2}(\sum_k^{n_2} w_{2,k}P(\sqrt{2}\sigma_1\epsilon_1, \sqrt{2}\sigma_2 e_{2,k})d\epsilon_1.$$

Approximating once again, we arrive at an expression that applies to the full double integral,

$$\pi^{-1}\sum_j^{n_1} w_{1,j}(\sum_k^{n_2} w_{2,k}P(\sqrt{2}\sigma_1 e_{1,j}, \sqrt{2}\sigma_2 e_{2,k})$$

or

$$\pi^{-1}\sum_j^{n_1}\sum_k^{n_2} w_{1,j}w_{2,k}P(\sqrt{2}\sigma_1 e_{1,j}, \sqrt{2}\sigma_2 e_{2,k}).$$

It is a relatively straightforward task to generalize the approach described above to the case of two correlated random effects.

36.4 Estimation: General Principles

When we discuss maximum-likelihood estimation in this note, we will let l_i^* represent the contribution made by individual i to the sample likelihood function and l_i the contribution to the log-likelihood.[1] This contribution depends on covariates specific to individual i and on a set of parameters θ, one of which is the variance σ_{11} of the random effect. To display these dependencies more explicitly than we yet have, we now write

$$l_i^* = \pi^{-1/2}\sum_{j=1}^{n_1} w_{1,j}P_i(\tilde{\theta}, \sqrt{2}\sigma_1 e_{1,j})$$

for the single random effect case. In this notation, $\tilde{\theta}$ comprises all parameters save σ_{11}, and we will let the full set of parameters be denoted by $\theta = (\tilde{\theta}, \sigma_{11})$.

[1] We have not yet precisely defined what we mean by "individual" in this context. If we are considering a set of children within a given family, who are linked by a common dependence on a family random effect u_1, then in this instance "individual" means "individual family." However, if these families are themselves grouped into villages or clusters, and such families are further linked by a dependence on a common cluster effect u_2, then in this case "individual" means "individual cluster".

Estimation of the parameters θ proceeds by maximizing the log-likelihood function $L = \sum_i \ln l_i^* = \sum_i l_i$. A key step is to derive the *scores*, which are the derivatives $\partial L/\partial\theta$, with

$$\frac{\partial L}{\partial \theta} = \sum_i \frac{\partial l_i}{\partial \theta}.$$

Note that individual i's contribution to the score is

$$\frac{\partial l_i}{\partial \theta} = \frac{\sum_j w_{1,j} \frac{\partial P_{i,j}}{\partial \theta}}{\sum_j w_{1,j} P_{i,j}}.$$

As we will see later, it will prove useful to re-express this derivative in the form

$$\frac{\partial l_i}{\partial \theta} = \frac{\sum_j w_{1,j} P_{i,j} \frac{\partial \ln P_{i,j}}{\partial \theta}}{\sum_j w_{1,j} P_{i,j}}$$

because the derivatives of $\ln P_{i,j}$ with respect to θ are generally similar to their counterparts in models without random effects. This point will be clarified by numerous examples in later handouts.

Chapter 37

Probit Models for Panel or Clustered Data

37.1 The Random Effect Model

Consider the multiperiod probit model with a single random effect. For a given individual i and time period t, the latent variable representation of the model is

$$Y_{it}^* = X_{it}'\beta + u_i + v_{it}$$

in which X_{it} is a column vector of K fully exogenous explanatory variables and β is a vector of coefficients. Assume that u_i, the random effect, is normally distributed with mean zero and variance σ_{uu}. The remaining error component v_{it} is also assumed normal with mean zero and variance σ_{vv}. We take u_i and v_{it} to be independent of each other and of the $X_{it}, t = 1, \ldots, T_i$ explanatory variables.

In a probit model, the error variances are not identified and some normalization of scale must be imposed. Following Heckman (1981, p. 129), we choose the normalization rule to be $\sigma_{uu} + \sigma_{vv} = 1$. This is a convenient rule to apply if one begins with $\hat{\beta}$ estimates from a standard probit model, since they are based on an assumed error variance of unity. The random-effects maximization program will not then have to re-scale the $\hat{\beta}$ to accommodate a different (composite) error variance. Given the normlization rule, the correlation between any two disturbance terms for individual i is simply

$$E(u_i + v_{it})(u_i + v_{it'}) \equiv \rho = \frac{\sigma_{uu}}{\sigma_{uu} + \sigma_{vv}} = \sigma_{uu}.$$

In this notation, ρ is the variance of the random effect u and the variance of v_{it} is therefore $1 - \rho$. The value of ρ must lie between 0 and 1.

The equation defining the latent variable Y_{it}^* may now be divided through by $\sqrt{1-\rho}$ so that the result is in the usual probit form. Letting $r = (1-\rho)^{-1/2}$, we can see that

$$rY_{it}^* = r(X_{it}'\beta + u_i) + rv_{it}$$

is in the desired form since rv_{it} is standard normal. The probability associated with the observed dependent variable Y_{it}, conditional on the random effect u_i, is then

$$\Pr(Y_{it} = y_{it} | X_{it}, u_i) = \Phi\left(y_{it} r(X_{it}'\beta + u_i)\right).$$

The product of such probabilities over the data sequence for person i is

$$P(u_i) = \prod_{t=1}^{T_i} \Phi\left(y_{it} r(X'_{it}\beta + u_i)\right).$$

To integrate out the unobservable random effect u, we want the quadrature approximation to an integral of the general form,

$$\int_{-\infty}^{\infty} (2\pi)^{-1/2} \rho^{-1/2} e^{-\frac{1}{2\rho}u^2} P(u) du.$$

Applying the change of variables and approximating with n_1 quadrature points, we obtain

$$L_i^* = \pi^{-1/2} \sum_{j=1}^{n_1} w_j P_i(\sqrt{2}\rho^{1/2} e_j),$$

an expression in which the roles of the covariates and coefficients are left somewhat implicit. When we need to see their roles more clearly, we write out the expression for P_i in full, as

$$P_i(\sqrt{2}\rho^{1/2} e_j) = \prod_{t=1}^{T_i} \Phi\left(y_{it} r(X'_{it}\beta + \sqrt{2}\rho^{1/2} e_j)\right).$$

We will refer to this expression as $P_{i,j}(\theta)$, a notation in which $\theta = (\beta, \rho)'$.

The scores

Recall from our earlier handout on quadrature that individual i's contribution to the full score vector is

$$\frac{\partial L_i}{\partial \theta} = \frac{\sum_j w_j \frac{\partial P_{i,j}}{\partial \theta}}{\sum_j w_j P_{i,j}}.$$

It is helpful to re-express this derivative in the form

$$\frac{\partial L_i}{\partial \theta} = \frac{\sum_j w_j P_{i,j} \frac{\partial \ln P_{i,j}}{\partial \theta}}{\sum_j w_j P_{i,j}}.$$

The benefit of rewriting the derivative is that $\ln P_{i,j}(\theta)$ is itself the sum of the logs of the probabilities $p_{it,j}(\theta)$ that are specific to individual i at time t. The derivatives of $\ln P_{i,j}$ with respect to β are quite similar to those based on models without random effects. The derivative with respect to ρ is also fairly easy to calculate.

To develop these derivatives, we begin with

$$p_{it,j} = \Phi\left(y_{it} r(X'_{it}\beta + \sqrt{2}\rho^{1/2} e_j)\right).$$

The $p_{it,j}$ derivative with respect to β_k is already generally familiar, being

$$\frac{\partial \ln p_{it,j}}{\partial \beta_k} = \frac{\phi_{it,j}}{\Phi_{it,j}} y_{it} r X_{it,k} = \frac{\partial \ln p_{it,j}}{\partial \beta_1} X_{it,k}.$$

As for the derivative with respect to ρ, if we recall that r is a function of ρ, a little manipulation gives

$$\frac{\partial \ln p_{it,j}}{\partial \rho} = \frac{\partial \ln p_{it,j}}{\partial \beta_1} \frac{1}{2} \left(r^2 Z_{it} + \sqrt{2} \rho^{-1/2} e_j \right)$$

with $Z_{it} = X'_{it}\beta + \sqrt{2}\rho^{1/2} e_j$. These results provide all the ingredients we need to estimate the model.

37.2 Models with Two Random Effects

To illustrate this more general structure, we consider a model of children (subscript i) grouped within families (subscript f), which are then grouped within clusters (subscript c), giving the latent model

$$Y^*_{ifc} = X'_{ifc}\beta + u_f + u_c + v_{ifc}.$$

The error components u_f, u_c and v_{ifc} are assumed to be normal and mutually independent, with variances $\sigma_{ff}, \sigma_{cc}, \sigma_{vv}$. The overall variance is normalized by imposing the rule $\sigma_{ff} + \sigma_{cc} + \sigma_{vv} = 1$.

Note that for two children in the same family and cluster, the error correlation is $\sigma_{ff} + \sigma_{cc}$, whereas for a pair of children in different families within the same cluster, the correlation is σ_{cc}. Were we to observe two children from the same family who live in different clusters, their error correlation would be σ_{ff}, although we might or might not encounter such cases depending on the nature of the data set.

To maintain some consistency of notation with the single random-effect model developed earlier, let $\rho_f = \sigma_{ff}$ and $\rho_c = \sigma_{cc}$ and then define $r = (1 - \rho_f - \rho_c)^{-1/2}$. Under the nomalization rule, we can express the probability associated with an observed Y_{ifc} as

$$\Phi \left(y_{ifc} r (X'_{ifc}\beta + u_f + u_c) \right)$$

conditional on the u_f and u_c random effects. Invoking the change-of-variables procedure required to implement quadrature, we obtain

$$\Phi \left(y_{ifc} r (X'_{ifc}\beta + \sqrt{2}\rho_f^{1/2} e_{f,j} + \sqrt{2}\rho_c^{1/2} e_{c,k}) \right)$$

as the key factor. The quadrature approximation will now be a sum over the double indices j (corresponding to families) and k (corresponding to clusters) with associated weights $w_{1,j}$ and $w_{2,k}$.

The scores

To focus only on the essential aspects of this case, let us suppress these j and k indices and write the probability expression for a given (i, f, c) as

$$p = \Phi \left(y_{ifc} r (X'_{ifc}\beta + \sqrt{2}\rho_f^{1/2} e_f + \sqrt{2}\rho_c^{1/2} e_c) \right).$$

Let $Z_{ifc} = X'_{ifc}\beta + \sqrt{2}\rho_f^{1/2}e_f + \sqrt{2}\rho_c^{1/2}e_c$. With this notation, we have $p = \Phi(y_{ifc}rZ_{ifc})$ and the score contributions for a given (i, f, c) can be written as

$$\begin{aligned}
\frac{\partial \ln p}{\partial \beta_1} &= \frac{\phi}{\Phi} y_{ifc} r \\
\frac{\partial \ln p}{\partial \beta_k} &= \frac{\partial \ln p}{\partial \beta_1} X_{ifc,k} \\
\frac{\partial \ln p}{\partial \rho_f} &= \frac{\partial \ln p}{\partial \beta_1} \frac{1}{2} \left(r^2 Z_{ifc} + \sqrt{2}\rho_f^{-1/2} e_f \right) \\
\frac{\partial \ln p}{\partial \rho_c} &= \frac{\partial \ln p}{\partial \beta_1} \frac{1}{2} \left(r^2 Z_{ifc} + \sqrt{2}\rho_c^{-1/2} e_c \right)
\end{aligned}$$

Chapter 38

Ordered-Probit Models

38.1 The Standard Ordered-Probit Model

The ordered probit model has a latent variable representation identical to that of the probit model, but differs in that successively higher ranges for the latent variable Y_{it}^* are mapped to ordered, discrete values for its observed counterpart Y_{it}. (We include two subscripts, i and t, because in the next section, we will be examining a panel data version of this model.) The latent structural model is specfied as

$$Y_{it}^* = X_{it}\beta + v_{it}$$

in which v_{it} is normally distributed with zero mean and unit variance and is independent of the explantory variables. The observed Y_{it} is then defined by the mapping

$$\begin{aligned} Y_{it} = &\quad 1 &\quad & Y_{it}^* \leq c_1 \\ = &\quad 2 &\quad & c_1 < Y_{it}^* \leq c_2 \\ \vdots &\quad \vdots &\quad & \vdots \\ = &\quad nc &\quad & c_{nc-1} < Y_{it}^* \leq c_{nc} \\ = &\quad nc+1 &\quad & c_{nc} < Y_{it}^* \end{aligned}$$

In this formulation, Y_{it} lies in a set of integers, $1, 2, \ldots, nc+1$, and in addition to the β vector, some nc cut-point parameters $\{c_j\}$ are needed to complete the specification of the model.[1] For the mapping to make sense, these cut-points must also be ordered, that is, it must be the case that $c_j < c_{j+1}$. No constant term should be included among the X_{it}, because the cut-points take over its role.

We will adopt a parameterization of the cut points $\{c_j\}$ in which they are expressed in

[1] The $\{c_j\}$ are not normally made functions of explanatory variables, although the model might be generalized to allow for this.

terms of parameters $\{d_j\}$, such that

$$\begin{aligned} c_1 &= d_1 \\ c_2 &= d_1 + d_2^2 \\ c_3 &= d_1 + d_2^2 + d_3^2 \\ \vdots &\ \vdots\ \vdots \\ c_{nc} &= d_1 + \sum_{j=2}^{nc} d_j^2 \end{aligned}$$

The gain from this respecification is that it imposes the necessary order on the cut-points.

Four cases

We now develop expressions for the log-likelihood contribution and the score contribution for four distinct cases differentiated according to the value taken by Y_{it}.

1. $Y_{it} = 1$

   The individual contribution to the log-likelihood is

   $$\ln p_{it} = \ln\left(\Phi\left(d_1 - X_{it}\beta\right)\right) \equiv \ln \Phi_1.$$

   The score contributions are

   $$\begin{aligned} \frac{\partial \ln p_{it}}{\partial d_1} &= \frac{\phi(d_1 - X_{it}\beta)}{\Phi(d_1 - X_{it}\beta)} = \frac{\phi_1}{\Phi_1}, \qquad \frac{\partial \ln p_{it}}{\partial d_j} = 0, \qquad j > 1 \\ \frac{\partial \ln p_{it}}{\partial \beta_k} &= -\frac{\phi_1}{\Phi_1} X_{it,k} = -\frac{\partial \ln p_{it}}{\partial d_1} X_{it,k}. \end{aligned}$$

2. $Y_{it} = 2$

   The likelihood contribution is

   $$\ln p_{it} = \ln\left(\Phi\left(d_1 + d_2^2 - X_{it}\beta\right) - \Phi\left(d_1 - X_{it}\beta\right)\right) \equiv \ln\left(\Phi_2 - \Phi_1\right).$$

   The score contributions are

   $$\begin{aligned} \frac{\partial \ln p_{it}}{\partial d_1} &= \frac{\phi_2 - \phi_1}{\Phi_2 - \Phi_1} \\ \frac{\partial \ln p_{it}}{\partial d_2} &= \frac{\phi_2}{\Phi_2 - \Phi_1} 2d_2 \\ \frac{\partial \ln p_{it}}{\partial \beta_k} &= \frac{\phi_1 - \phi_2}{\Phi_2 - \Phi_1} X_{it,k} = -\frac{\partial \ln p_{it}}{\partial d_1} X_{it,k}. \end{aligned}$$

3. $Y_{it} \geq 3$ and $Y_{it} \leq nc$

   The contribution to the likelihood is

   $$\ln p_{it} = \ln\left(\Phi\left(d_1 + \sum_{j=2}^{y} d_j^2 - X_{it}\beta\right) - \Phi\left(d_1 + \sum_{j=2}^{y-1} d_j^2 - X_{it}\beta\right)\right) \equiv \ln\left(\Phi_y - \Phi_{y-1}\right),$$

and the scores are

$$\begin{aligned}
\frac{\partial \ln p_{it}}{\partial d_1} &= \frac{(\phi_y - \phi_{y-1})}{(\Phi_y - \Phi_{y-1})} \\
\frac{\partial \ln p_{it}}{\partial d_2} &= \frac{\phi_y - \phi_{y-1}}{(\Phi_y - \Phi_{y-1})} 2d_2 = \frac{\partial \ln p_{it}}{\partial d_1} 2d_2 \\
&\vdots \quad \vdots \quad \vdots \\
\frac{\partial \ln p_{it}}{\partial d_y} &= \frac{\phi_y}{(\Phi_y - \Phi_{y-1})} 2d_y \\
\frac{\partial \ln p_{it}}{\partial \beta_k} &= \frac{(\phi_{y-1} - \phi_y)}{(\Phi_y - \Phi_{y-1})} X_{it,k} = -\frac{\partial \ln p_{it}}{\partial d_1} X_{it,k}.
\end{aligned}$$

4. $Y_{it} = nc + 1$

   In this last case, the likelihood contribution is

$$\ln p_{it} = \ln \left(1 - \Phi \left(d_1 + \sum_{j=2}^{nc} d_j^2 - X_{it}\beta \right) \right) \equiv \ln \left(1 - \Phi_{nc} \right),$$

   and the scores are

$$\begin{aligned}
\frac{\partial \ln p_{it}}{\partial d_1} &= \frac{-\phi_{nc}}{1 - \Phi_{nc}} \\
\frac{\partial \ln p_{it}}{\partial d_2} &= \frac{-\phi_{nc}}{1 - \Phi_{nc}} 2d_2 = \frac{\partial \ln p_{it}}{\partial d_1} 2d_2 \\
&\vdots \quad \vdots \quad \vdots \\
\frac{\partial \ln p_{it}}{\partial d_{nc}} &= \frac{-\phi_{nc}}{1 - \Phi_{nc}} 2d_{nc} = \frac{\partial \ln p_{it}}{\partial d_1} 2d_{nc} \\
\frac{\partial \ln p_{it}}{\partial \beta_k} &= \frac{\phi_{nc}}{1 - \Phi_{nc}} X_{it,k} = -\frac{\partial \ln p_{it}}{\partial d_1} X_{it,k}.
\end{aligned}$$

To understand the ordered-probit model, one must appreciate the assumption of monotonicity that is embedded in it. In this model, if a given $\beta_k > 0$, then an increase in the associated $X_{it,k}$ will reduce the probabilities of low values of Y_{it} and increase the probabilities associated with high values. Think of the bars of a histogram all shifting with $X_{it,k}$, some going down while others go up. The absolute amount by which these probabilities (bar heights) are changed will vary, however, and one should examine closely the derivatives of the probabilities with respect to $X_{it,k}$ to get a sense of what can happen.

Generally, to know whether a change in a given covariate has a appreciable overall effect, you will find that you need to generate a vector of predicted probabilities having some $nc + 1$ entries (one for each bar of the histogram) and then examine the associated mean or median. Without such a translation, the raw coefficients and cut-points of the ordered-probit model will prove to be difficult to interpret. In STATA, one can easily produce and manipulate this vector of predicted probabilities.

38.2 Incorporating a Random Effect

In the context of panel or clustered data, the latent variable Y_{it}^* can include a random effect u_i,

$$Y_{it}^* = X_{it}\beta + u_i + v_{it}$$

in which u_i is assumed to be normally-distributed with mean zero and variance ρ and independent of both v_{it} and the explanatory variables. As with the random effects probit model, we must impose some normalization rule on the composite variance, and adopt the rule $\rho + \sigma_{vv} = 1$.

Each of the equations linking Y_{it}^* to Y_{it} can now be re-expressed by dividing through by $\sqrt{1-\rho}$. Thus, for the case of $Y_{it} = 1$, we would have

$$\ln p_{it} = \ln\left(\Phi(r(d_1 - X_{it}\beta - u_i))\right)$$

with $r = (1-\rho)^{-1/2}$. When the quadrature approximation is made, this expression becomes

$$\ln p_{it,j} = \ln\left(\Phi(r(d_1 - X_{it}\beta - \sqrt{2}\rho^{1/2}e_j))\right)$$

for the j-th quadrature point. The other $\ln p_{it,j}$ contributions for $Y_{it} = 2, \ldots, nc+1$ are altered in a similar fashion. The score contributions for the $\{d_j\}$ and β parameters are essentially what we derived for the conventional case, with the exception that each expression must now be multiplied by r. (In fact, if the score contributions for d_j and β are defined with reference to $\partial \ln p_{it,j}/\partial d_1$, as might well be computationally convenient, then these score contributions can be written just as they were above.)

The new element of the score has to do with ρ, the variance of the random effect. For the case of $Y_{it} = 1$, let $Z_{it,j}^1 = d_1 - X_{it}\beta - \sqrt{2}\rho^{1/2}e_j$, and let the other $Z_{it,j}^y$ be defined similarly. We then have, for the case of $Y_{it} = 1$,

$$\frac{\partial p_{it,j}}{\partial \rho} = \phi_1 \frac{r}{2}\left(r^2 Z_{it,j}^1 - \sqrt{2}\rho^{-1/2}e_j\right);$$

for $Y_{it} \in \{2, \ldots, nc\}$,

$$\frac{\partial p_{it,j}}{\partial \rho} = \phi_y \frac{r}{2}\left(r^2 Z_{it,j}^y - \sqrt{2}\rho^{-1/2}e_j\right) - \phi_{y-1}\frac{r}{2}\left(r^2 Z_{it,j}^{y-1} - \sqrt{2}\rho^{-1/2}e_j\right);$$

and for $Y_{it} = nc + 1$,

$$\frac{\partial p_{it,j}}{\partial \rho} = -\phi_{nc}\frac{r}{2}\left(r^2 Z_{it,j}^{nc} - \sqrt{2}\rho^{-1/2}e_j\right).$$

Each of these expressions should be divided by the appropriate $p_{it,j}$ to obtain the log contributions to the score.

38.3 Allowing for Censored Observations

This model is similar to the standard ordered-probit model in that values for a continuous latent variable Y_i^* are mapped into discrete values for an observed variable Y_i taking values

$1, 2, \ldots, nc + 1$. The only difference is that a censoring variable C_i is present. An example of such a model is provided by the case of children's schooling, where the observed variable Y_i represents the number of grades completed by child i and the censoring variable C_i indicates whether that child is still enrolled in school ($C_i = 1$) or has finished schooling ($C_i = 0$).

In such a model, the case $Y_i = 1, C_i = 1$ is uninformative, because this pair of values simply indicates that the child has not yet completed any grades of school but is currently enrolled. One would normally process the data so as to eliminate such cases from consideration.

For the non-censored cases denoted by $C_i = 0$, the log-likelihood contributions and score contributions are precisely those derived for the conventional ordered-probit model. For censored cases, which have $C_i = 1$, the log-likelihood contribution for $Y_i = y$ is

$$\ln p_y = \ln \left(1 - \Phi_{y-1}\right).$$

The score contributions for the censored cases are

$$\begin{aligned}
\frac{\partial \ln p_y}{\partial d_1} &= -\frac{\phi_{y-1}}{1 - \Phi_{y-1}} \; y \geq 2 \\
\frac{\partial \ln p_y}{\partial d_j} &= \left(\frac{\partial \ln p_y}{\partial d_1}\right) 2 d_j \; y \geq 3, \; j = 2, \ldots, y - 1 \\
\frac{\partial \ln p_y}{\partial \beta_k} &= -\left(\frac{\partial \ln p_y}{\partial d_1}\right) X_{i,k}.
\end{aligned}$$

Chapter 39

Weibull Hazard Rate Models

39.1 The Standard Model

We will begin by presenting the model in its conventional form, without random effects, and then show how it may be generalized to include such effects. For the conventional model, we will simply refer to observation i, without taking account of the fact that these observations are grouped into families or clusters. An additional subscript f, which will denote family or cluster, will appear when we generalize the model to include random effects.

The following specification of the Weibull is adapted from Lancaster (1990, p. 169). Let the hazard rate be represented as

$$r(t \mid X_i) = \theta_i \alpha t^{\alpha-1}$$

with $\theta_i = e^{X_i'\beta}$ and with X_i a vector of covariates for individual i. The implied survivor function is then

$$S(t \mid X_i) = e^{-\theta_i \alpha \int_0^t v^{\alpha-1} dv} = e^{-\theta_i t^\alpha}.$$

The density function is the product of the hazard and survivor functions,

$$f(t \mid X_i) = \theta_i \alpha t^{\alpha-1} e^{-\theta_i t^\alpha}.$$

Likelihood contributions

The contributions made by censored and uncensored cases to the full log-likelihood are, respectively,

$$\ln S(t \mid X_i) = -\theta_i t^\alpha = -e^{X_i'\beta} t^\alpha$$

and

$$\ln f(t \mid X_i) = X_i'\beta + \ln \alpha + (\alpha - 1) \ln t + \ln S(t \mid X_i).$$

Score contributions

The contributions made by observation i to the full scores $\partial L/\partial\alpha$ and $\partial L/\partial\beta$, are as follows.

$$\begin{aligned}
\frac{\partial \ln S_i}{\partial \alpha} &= \ln S_i \ \ln t_i \\
\frac{\partial \ln S_i}{\partial \beta_k} &= \ln S_i \ X_{i,k} \\
\frac{\partial \ln f_i}{\partial \alpha} &= \alpha^{-1} + \ln t_i + \ln S_i \ \ln t_i \\
\frac{\partial \ln f_i}{\partial \beta_k} &= (\ln S_i + 1) \ X_{i,k}.
\end{aligned}$$

Second derivatives

For those desiring to form the information matrix using second derivatives rather than the outer product of the scores—this is always a good idea given the "noise" inherent in the outer product estimator—the derivatives are

$$\begin{aligned}
\frac{\partial^2 \ln S_i}{\partial \alpha^2} &= \ln S_i \, (\ln t_i)^2 \\
\frac{\partial^2 \ln S_i}{\partial \beta_k \partial \beta_j} &= \ln S_i \ X_{i,j} X_{i,k} \\
\frac{\partial^2 \ln S_i}{\partial \alpha \partial \beta_k} &= \ln S_i \ \ln t_i \ X_{i,k} \\
\frac{\partial^2 \ln f_i}{\partial \alpha^2} &= -\alpha^{-2} + \ln S_i \, (\ln t_i)^2 \\
\frac{\partial^2 \ln f_i}{\partial \beta_k \partial \beta_j} &= \ln S_i \ X_{i,j} X_{i,k} \\
\frac{\partial^2 \ln f_i}{\partial \alpha \partial \beta_k} &= \ln S_i \ \ln t_i \ X_{i,k}.
\end{aligned}$$

39.2 Incorporating a Random Effect

Imagine that our observations are grouped into families, so that the subscript i refers to (say) child i and the subscript f refers to family f. Let the random effect u for the family enter the hazard function $r\left(t \mid X_{if}, u\right)$ in a proportional manner, such that

$$\theta_{if} = e^{X'_{if}\beta + u}$$

and assume that $u \sim \mathcal{N}(0, \rho)$. In a quadrature application, this expression will be altered to

$$\theta_{if,j} = e^{X_{if}\beta + \sqrt{2}\rho^{1/2} e_j}$$

for the j-th quadrature point. Let $p^c_{if,j}$ and $p^u_{if,j}$ represent the probabilities attached to censored and uncensored observations, respectively. The derivatives $\partial \ln p_{if,j}/\partial\alpha$ and

$\partial \ln p_{if,j}/\partial\beta$ are identical to those presented above in both the censored and uncensored cases, provided that one substitutes the expression $X'_{if}\beta + \sqrt{2}\rho^{1/2}e_j$ for $X'_i\beta$.

The new derivative to consider is the derivative with respect to the variance of the random effect, ρ. For the censored case, this derivative is

$$\frac{\partial \ln p^c_{if,j}}{\partial\rho} = \frac{1}{2}\ln p^c_{if,j} \cdot \sqrt{2}\rho^{-1/2}e_j.$$

For the uncensored case, we have

$$\frac{\partial \ln p^u_{if,j}}{\partial\rho} = \frac{1}{2}\left(1 + \ln p^c_{if,j}\right)\sqrt{2}\rho^{-1/2}e_j.$$

39.3 Grouped Data

Returning to the model without random effects, suppose that all that is known is that a transition occurred before t_2. Then one takes the log of $P_i(t_2) \equiv 1 - S_i(t_2)$ as the observation's contribution to the full log likelihood. The first derivatives are as follows.

$$\begin{aligned}
\frac{\partial \ln P_i(t_2)}{\partial\beta_k} &= -\frac{S_i(t_2)}{1 - S_i(t_2)}\ln S_i(t_2)X_{i,k} \\
\frac{\partial \ln P_i(t_2)}{\partial\alpha} &= -\frac{S_i(t_2)}{1 - S_i(t_2)}\ln S_i(t_2)\ln t_2
\end{aligned}$$

For cases in which the age range is $t_2 < t < t_3$ the log likelihood contribution is the log of $P_i(t_2, t_3) \equiv S_i(t_2) - S_i(t_3)$. The first derivatives are

$$\begin{aligned}
\frac{\partial \ln P_i}{\partial\beta_k} &= \frac{S_i(t_2)\ln S_i(t_2) - S_i(t_3)\ln S_i(t_3)}{S_i(t_2) - S_i(t_3)}X_{i,k} \\
\frac{\partial \ln P_i}{\partial\alpha} &= \frac{S_i(t_2)\ln S_i(t_2)\ln t_2 - S_i(t_3)\ln S_i(t_3)\ln t_3}{S_i(t_2) - S_i(t_3)}
\end{aligned}$$

Chapter 40

Factor Analysis and Related Methods

In the economic development literature, it is common to see a household's standard of living represented by an index created from the consumer durables that the household owns. In the usual case, these durables include ownership of a television, radio, automobile, bicycle, refrigerator, and the like. They are often termed *indicators* or *proxy variables* for consumption proper. The need for an index arises when the measures of living standards that economists prefer—consumption expenditures or full income—are not available. In developing countries, the effort to gather reliable information on expenditures and incomes requires a survey that is dedicated to this task. Relatively few adults earn salaries; there are high levels of self-employment; production may or may not be channeled through formal markets; seasonal variation exists in both production and consumption—for these and many other reasons, obtaining accurate data on incomes and expenditures is a demanding and time-consuming task (Deaton 2001). As a result, the surveys that have collected such data in developing countries—notably the World Bank's Living Standards Measurement Surveys—have often been forced to limit their inquiries into other socioeconomic areas.

Until recently, the literature could offer little by way of guidance on how living standards indices should be constructed from the proxy variables for expenditure that are available in the typical survey (Montgomery et al. 2000). Over the past few years, however, three reasonably promising approaches have emerged. Filmer and Pritchett (1999) argue strongly that the principal components method should be used to reduce a set of consumption indicators to an index. Sahn and Stifel (2000) prefer the method of confirmatory factor analysis, arguing that it provides a firmer theoretical and statistical foundation for index construction. A third variant, explored in Montgomery and Hewett (2005), generalizes the factor-analytic method to allow for multiple binary indicators of consumption in the context of a MIMIC (multiple indicator, multiple cause) model. None of these approaches is satisfactory on all counts, and much remains to be learned about their relative strengths and weaknesses.[1]

Principal components and confirmatory factor analysis are familiar tools in the fields of psychology and sociology, but they have not been seen as often in economics. In the first two

[1] All the methods share one awkward feature: It is not obvious what concept of living standards—whether transitory or permanent, expressed in per household terms or in per capita, per adult, or per adult equivalent terms—is actually represented in the consumption index (Montgomery et al. 2000). Surprisingly little research has been done on this central issue.

sections of the chapter, we present the statistical foundations of these conventional methods, giving attention to their common features as well as the differences. Upon concluding our discussion of the conventional methods, we proceed to examine MIMIC models for binary indicators in the last section of the chapter.

Note that although this chapter's discussion is framed in terms of consumption indicators, many other applications of the methods come to mind. For example, given the logistical difficulties entailed in establishing the state of an individual's health with physical examinations, a health index could be calculated from readily accessible measures such as self-reports of health and reported limitations on daily activities.

40.1 Principal Components

The principal components method is decidedly more computational than statistical in spirit. Its starting-point is an exact decomposition of the covariance matrix of a set of consumption indicators, with the matrix being expressed as a sum of component matrices. The appeal of the method lies in the hope that, upon inspection, one or two of these components will be found interpretable in economic terms and capable of summarizing most of the variability exhibited by the original variables. Principal components is often viewed as a data reduction device, in that it allows a potentially large number of indicators to be represented by a much smaller number of components (Dunteman 1989).

To understand the method in the present context, consider a set of K indicators of household consumption $\mathbf{Z}_i$, where $\mathbf{Z}_i$ is a column vector and the i subscript serves to index observations. The $\mathbf{Z}_i$ vectors are assumed to have been centered so as to have zero mean in each of their K elements. The estimated covariance matrix of the $\mathbf{Z}_i$ vectors is then

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{Z}_i \mathbf{Z}_i'$$

where n is the sample size. An implicit assumption is that the $\mathbf{Z}_i$ vectors are uncorrelated across households (the i subscript) and homoskedastic.

The principal components method provides a solution to the following problem. How should we choose a K-vector of weights $\mathbf{w}$ to be attached to the indicators in the linear combination $\mathbf{w}'\mathbf{Z}_i$, in such a way as to maximize the variance of this linear combination? The variance of $\mathbf{w}'\mathbf{Z}_i$ is estimated by $\mathbf{w}'\hat{\mathbf{V}}\mathbf{w}$. For this problem to be well-defined, we must impose a normalization that keeps $\mathbf{w}$ within bounds, and therefore restrict the admissible $\mathbf{w}$ vectors to those that satisfy $\mathbf{w}'\mathbf{w} = 1$. The problem is thus posed as

$$\max_{\mathbf{w}:\mathbf{w}'\mathbf{w}=1} \mathbf{w}'\hat{\mathbf{V}}\mathbf{w}.$$

The vector $\mathbf{w}_1$ that solves this problem is the eigenvector of $\hat{\mathbf{V}}$ that is associated with λ_1, the largest of $\hat{\mathbf{V}}$'s eigenvalues. It can be shown that λ_1 is also the maximum variance attained, that is, $\lambda_1 = \mathbf{w}_1'\hat{\mathbf{V}}\mathbf{w}_1$. Succinct proofs of these propositions are given in Schott (1997, 104–110, Theorem 3.15).

With $\mathbf{w}_1$ in hand, we can then address a second problem, how to choose another $\mathbf{w}$ vector, orthogonal to the first, that maximizes $\mathbf{w}'\hat{\mathbf{V}}\mathbf{w}$. That is, we aim to choose $\mathbf{w}$ so as to

$$\max_{\mathbf{w}:\mathbf{w}'\mathbf{w}=1,\mathbf{w}'\mathbf{w}_1=0} \mathbf{w}'\hat{\mathbf{V}}\mathbf{w}.$$

The solution to this second problem is $\mathbf{w}_2$, the eigenvector associated with the second largest of $\hat{\mathbf{V}}$'s eigenvalues (Schott 1997, Theorem 3.16). We can proceed in this fashion, choosing new vectors $\mathbf{w}_j$ from a set orthogonal to all vectors $[\mathbf{w}_1, \ldots, \mathbf{w}_{j-1}]$ chosen previously, until we have finally exhausted the full set of $\hat{\mathbf{V}}$'s eigenvalues and eigenvectors. Because $\hat{\mathbf{V}}$ is symmetric and positive definite, all of the eigenvalues λ_j, $j = 1, \ldots, K$, are positive.

Gathering the $\mathbf{w}_j$ eigenvectors in a $K \times K$ matrix $\mathbf{W} = (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_K)$, and arranging their eigenvalues in a diagonal matrix $\mathbf{D}$, we obtain

$$\hat{\mathbf{V}}\mathbf{W} = \mathbf{W}\mathbf{D}$$

and can then decompose the $\hat{\mathbf{V}}$ matrix as

$$\hat{\mathbf{V}} = \mathbf{W}\mathbf{D}\mathbf{W}' = \sum_{j=1}^{K} \lambda_j \mathbf{w}_j \mathbf{w}_j',$$

owing to the fact that $\mathbf{W}$ is an orthonormal matrix whose transpose is its inverse. The decomposition of $\hat{\mathbf{V}}$ is exact provided that all K of its eigenvalues and eigenvectors are used.

Since $\mathbf{D} = \mathbf{W}'\hat{\mathbf{V}}\mathbf{W}$, we can write

$$\sum_{j=1}^{K} \lambda_j = \text{trace}\, \mathbf{D} = \text{trace}\, \hat{\mathbf{V}}\mathbf{W}\mathbf{W}' = \text{trace}\, \hat{\mathbf{V}} = \sum_{j=1}^{K} \hat{\mathbf{V}}_{j,j},$$

a result that shows that the sum of the eigenvalues equals the sum of the variances of the variables in $\mathbf{Z}$. If $\hat{\mathbf{V}}$ were to be defined from the outset as a correlation matrix, rather than as a covariance matrix, the sum of the eigenvalues would equal K, the number of variables. In this case, the average eigenvalue would be unity.[2]

Using these ingredients, we now define and manipulate the *principal components*. The j-th principal component for observation i is defined to be the linear combination

$$p_{i,j} = \mathbf{w}_j' \mathbf{Z}_i. \tag{40.1}$$

All K principal components for this observation can be placed into a column vector, such that

$$\mathbf{p}_i = \begin{bmatrix} p_{i,1} \\ p_{i,2} \\ \vdots \\ p_{i,K} \end{bmatrix} = \mathbf{W}'\mathbf{Z}_i, \tag{40.2}$$

and the variance matrix of $\mathbf{p}_i$ is then easily seen to be $\mathbf{D}$, the diagonal matrix with the eigenvalues on the diagonal. Equation (40.2) can be inverted by multiplying by $\mathbf{W}$, yielding

$$\mathbf{Z}_i = \mathbf{W}\mathbf{p}_i, \tag{40.3}$$

[2]See Dunteman (1989, p. 22) for a discussion of whether $\lambda_j = 1.0$ is a statistically sensible benchmark. Schott (1997, pp. 404–6) describes how, as with formal test statistics, the asymptotic distributions of eigenvalues and eigenvectors (suitably normalized) might be employed to establish appropriate benchmark values.

in which the original $\mathbf{Z}_i$ vector is here re-expressed as a weighted combination of K principal components.

What have we learned so far? It may seem that all these machinations have accomplished very little. We began by considering K consumption indicators $\mathbf{Z}_i$ with covariance matrix $\hat{\mathbf{V}}$, and have found that they can be represented by an equal number of principal component vectors $\mathbf{p}_i$ having the diagonal matrix $\mathbf{D}$ as their covariance matrix. This is interesting in mathematical terms, but we are still some distance from our objective, which is to form a sensible index of consumption from the principal components.

To reach this goal, we hope to find that the elements of the most important eigenvectors (i.e., $\mathbf{w}_1$ and $\mathbf{w}_2$) have some substantive interpretation and that they adequately represent the full set of $\mathbf{Z}_i$ consumption indicators. Nothing in the method guarantees that this will happen. But in practice, it is sometimes the case that the first and second eigenvectors can be interpreted in socioeconomic terms, and these eigenvectors may account for a reasonably high percentage of the "total variability" of $\mathbf{Z}_i$, with total variability being defined as the sum of the variances.[3]

In the studies that have used this technique (e.g., Filmer and Pritchett 1999; Filmer and Pritchett 2001), the first principal component $p_{i,1}$ has been taken to represent the socioeconomic status of the i-th household, and has been shown to perform much as one would expect a living standards measure to perform in models of schooling and health. Not much attention has been given to the possibility that more than one component might be needed to adequately represent living standards. The need for additional components depends in part on the degree of correlation among the observed $\mathbf{Z}_i$ indicators. When the $\mathbf{Z}_i$ are closely associated, one or two principal components can account for most of their total variability. But when the $\mathbf{Z}_i$ are not so tightly bound together, the number of principal components required for an adequate summary can be large.

40.2 Factor Analysis

Factor analysis is a method not unlike principal components in that it has mainly been regarded as a means of expressing a set of K observed indicators as a function of a smaller number of common factors. Like principal components, factor analysis focuses on explaining the covariances among the indicators. The method departs from principal components in two ways: first, the linkages between factors and indicators are specified in a structural model that can be tested and elaborated as necessary; second, the factors themselves are taken to be unobserved.

In this approach, the indicators $\mathbf{Z}_i$ (again centered to have mean zero, with a $K \times K$ variance matrix $\mathbf{V}$ as above) are specified as

$$\mathbf{Z}_i = \Theta \mathbf{f}_i + \epsilon_i. \tag{40.4}$$

Here, $\mathbf{f}_i$ represents a column vector of m unobserved factors for the i-th observation, Θ is a $K \times m$ matrix of factor "loadings," and ϵ_i is a vector of K disturbance terms that is assumed

[3]That is, total variability is $\sum_{j=1}^{K} \hat{\mathbf{V}}_{j,j}$. As noted above, this sum equals $\sum_{j=1}^{K} \lambda_j$, the sum of the eigenvalues. Hence, $\lambda_j / \sum_{l=1}^{K} \lambda_l$ is the proportion of the total variability attributable to the j-th component.

to be uncorrelated with $\mathbf{f}_i$ and whose covariance matrix $\Omega = \mathrm{E}\,\epsilon_i\epsilon_i'$ is usually taken to be diagonal.

In a one-factor model ($m = 1$), the Θ matrix reduces to a column vector, and the equation for $\mathbf{Z}_i$ can be written as

$$\begin{bmatrix} Z_{i,1} \\ Z_{i,2} \\ \vdots \\ Z_{i,K} \end{bmatrix} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_K \end{bmatrix} f_i + \begin{bmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \\ \vdots \\ \epsilon_{i,K} \end{bmatrix}. \tag{40.5}$$

Returning to the general m-factor case with Θ being $K \times m$ and letting the $m \times m$ covariance matrix of the factors $\Phi = \mathrm{E}\,\mathbf{f}_i\mathbf{f}_i'$, the model implies that the covariance matrix of the $\mathbf{Z}_i$ indicators is

$$\mathbf{V} = \mathrm{E}\,\mathbf{Z}_i\mathbf{Z}_i' = \Theta\Phi\Theta' + \Omega. \tag{40.6}$$

To estimate the unknown parameters on the right-hand side of equation (40.6), the $\mathbf{V}$ implied by a particular set of parameters (the elements of $\Theta, \Phi,$ and Ω) is compared with its empirical counterpart $\hat{\mathbf{V}}$, and parameter values are chosen to minimize the differences between them, using nonlinear least squares, minimum distance, or related maximum-likelihood methods.

To compare this specification with its analog in a principal components analysis, we can return to equation (40.3), which expressed the vector $\mathbf{Z}_i = \mathbf{W}\mathbf{p}_i$ as a linear function of all K principal components. Unpacking this expression gives

$$\begin{bmatrix} Z_{i,1} \\ Z_{i,2} \\ \vdots \\ Z_{i,K} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1 & \mathbf{w}_2 & \dots & \mathbf{w}_K \end{bmatrix} \begin{bmatrix} p_{i,1} \\ p_{i,2} \\ \vdots \\ p_{i,K} \end{bmatrix} = \mathbf{w}_1 p_{i,1} + \sum_{j=2}^{K} \mathbf{w}_j p_{i,j}.$$

Letting $\mathbf{u}_i = \sum_{j=2}^{K} \mathbf{w}_j p_{i,j}$ denote all terms in the sum except the first, we have

$$\begin{bmatrix} Z_{i,1} \\ Z_{i,2} \\ \vdots \\ Z_{i,K} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_{1,1} \\ \mathbf{w}_{1,2} \\ \vdots \\ \mathbf{w}_{1,K} \end{bmatrix} p_{i,1} + \begin{bmatrix} u_{i,1} \\ u_{i,2} \\ \vdots \\ u_{i,K} \end{bmatrix}. \tag{40.7}$$

Equation (40.7) closely resembles the specification given in equation (40.5) for the one-factor model. There are, however, two crucial differences. In equation (40.7) the principal component $p_{i,1}$ is a computable quantity (as are $\mathbf{w}_1$ and the remainder $\mathbf{u}_i$) whereas the factor f_i of equation (40.5) is unobservable. Furthermore, in a factor-analytic model the disturbance terms $\{\epsilon_{i,j},\ j = 1, \dots, k\}$ are assumed to be mutually uncorrelated. But as Jöreskog (1979, p. 15) shows, in a principal components approach the composite remainder terms $\{u_{i,j},\ j = 1, \dots, K\}$ are linked by linear dependencies and therefore cannot be uncorrelated.

Comparison of these methods is further complicated by the need to estimate $\mathbf{f}_i$, the vector of unobserved factors. Although there exists a variety of estimation methods, the most common is a regression in which an estimated value for $\mathbf{f}_i$ is computed from a (hypothetical) regression of $\mathbf{f}_i$ on $\mathbf{Z}_i$ (Lawley and Maxwell 1962; Bollen 1989; Jöreskog 2000). The logic can

be seen most easily in the one-factor case. The predicted value of f_i from the hypothetical regression of f on $\mathbf{Z}$ would be

$$\hat{f}_i = \mathbf{Z}_i' \left(n^{-1} \sum_{j=1}^{n} \mathbf{Z}_j \mathbf{Z}_j' \right)^{-1} \left(n^{-1} \sum_{j=1}^{n} \mathbf{Z}_j f_j \right)$$

For a given i, as $n \to \infty$ the terms in parentheses converge to $\mathbf{V}$ and $\mathrm{E}(\Theta f_j^2 + \epsilon_j f_j) = \Theta \sigma_f^2$, respectively. Estimates of these quantities are available in the estimated covariance matrix of the model; see equation (40.6).

There are further issues to be faced when the estimated factors are used as explanatory variables in structural models. As Bollen (1989) points out, the estimated factor score $\hat{f}_i$ is not itself an error-free measure of f_i, the true factor. The sampling error in $\hat{f}_i$ would not normally affect the consistency of estimators with $\hat{f}_i$ used as one of the explanatory variables, but the standard errors of the structural equation estimators would be affected. Hence, if estimated factor scores are used as inputs into a structural model, robust estimates of standard errors must be employed.

40.3 The Multiple Indicators Model

To generalize the approaches spelled out above, we continue to assume that every household i in the dataset provides a vector $\mathbf{Z}_i$ containing K observed indicators, with each of these being denoted by Z_{ik}. We depart from the models above in assuming that the $\mathbf{Z}_i$ are *binary* indicators, and will work with models that are specifically designed for such indicators.[4]

We first describe a model in which the indicators are expressions of a single unobserved factor $F_i = u_i$, which we take to represent household i's standard of living. Many of the estimation details are discussed in this simple context, as are the procedures used to estimate u_i given the observed values of the indicators. With the basics thus laid out, the last section of the appendix explores an expanded model in which the household's standard of living $F_i = \mathbf{X}_i'\gamma + u_i$, allowing exogenous covariates $\mathbf{X}_i$ such as education and age of the household head to play a role (Montgomery and Hewett 2005). The expanded model is the so-called MIMIC specification, this being an acronym for "multiple indicators, multiple causes."

In this specification each element of the indicator vector $\mathbf{Z}_i$ is assumed to depend on an unobserved factor $F_i = u_i$ through a latent variable. Consider Z_{ik}, one of the K indicators. This observed indicator is linked to its latent counterpart Z_{ik}^* via two equations, the first being

$$\begin{aligned} Z_{ik}^* &= \alpha_k + \beta_k F_i + v_{ik} \\ &= \alpha_k + \beta_k u_i + v_{ik}. \end{aligned} \tag{40.8}$$

In equation (40.8), α_k is a cut-point parameter and β_k is a coefficient indicating how the unobserved factor u_i takes expression through the k-th indicator. The latent variable Z_{ik}^* is

[4] In particular, we no longer need to assume that the indicators are centered to have zero means.

then linked with its observed counterpart Z_{ik} through the mapping

$$Z_{ik} = \begin{cases} 1 & \text{if } Z_{ik}^* > 0, \\ -1 & \text{if } Z_{ik}^* \leq 0. \end{cases}$$

Although unconventional, this $\{1, -1\}$ coding scheme simplifies both the analysis and the programming.

In what follows, we will indicate the dependence of the vector $\mathbf{Z}_i$ on u_i using the notation $P_i(u_i)$, with P_i being the joint probability distribution associated with $\mathbf{Z}_i$ conditional on the (unknown) value of u_i. The unconditional probability associated with $\mathbf{Z}_i$ is derived by "integrating out" the unobserved random factor. We will assume that the factor u_i is normally distributed with mean zero and variance ρ. Given this, the unconditional probability is expressed by the integral

$$\int_{-\infty}^{\infty} (2\pi)^{-1/2} \rho^{-1/2} e^{-\frac{1}{2\rho} u_i^2} P_i(u_i)\, du_i. \tag{40.9}$$

Unfortunately, the integral is not available in a closed form, and numerical approximation methods are required to evaluate it.

Background on quadrature

Recall that the method of Gaussian quadrature is applied when one desires a good approximation to an integral of the type

$$\int_{-\infty}^{\infty} e^{-\epsilon^2} P(\epsilon) d\epsilon$$

where the function $P(\epsilon) > 0$ and the integral in question cannot be represented in a closed form. (Note that, for the moment, the i subscript has been suppressed.) The quadrature method approximates this integral by a weighted summation over a pre-selected number of quadrature points. The method is explained in illuminating detail by Press et al. (1992) and Press et al. (1996), who provide additional references as well as programming subroutines that calculate the quadrature points and the weights associated with them.

To put equation (40.9) in this form, we need only make a change of variables

$$\epsilon = \frac{u}{\sqrt{2}\rho^{1/2}},$$

with $\rho^{1/2}$ being the standard deviation of u. The transformation implies $u^2 = 2\rho\epsilon^2$ and

$$\frac{du}{d\epsilon} = \sqrt{2}\rho^{1/2}.$$

Upon making the change of variables, we obtain

$$\pi^{-1/2} \int e^{-\epsilon^2} P(\sqrt{2}\rho^{1/2}\epsilon) d\epsilon,$$

which is in the required form apart from the constant $\pi^{-1/2}$. The quadrature method approximates the integral by the sum

$$\sum_{j=1}^{n_q} w_j P(\sqrt{2}\rho^{1/2} e_j),$$

whose index j ranges over $n_q > 1$ quadrature points. The quadrature points e_j are symmetric about zero, as are the weights w_j with which they are associated. The number of points n_q is under the control of the researcher, but the quality of the approximation generally improves as the number of points increases.

Maximum likelihood estimation: General approach

Let L_i^* represent the contribution made by household i to the sample likelihood function and let L_i be the contribution to the sample log-likelihood. The contribution depends on covariates specific to household i and on a set of parameters θ, one of which is the variance ρ of the random factor. (The other parameters will be discussed shortly.) To display these dependencies more explicitly than we have thus far, we write

$$L_i^* = \pi^{-1/2} \sum_{j=1}^{n_q} w_j P_i(\theta_0, \sqrt{2}\rho^{1/2} e_j).$$

In this notation, θ_0 contains all unknown parameters save ρ, and we let the full set of parameters be denoted by $\theta = (\theta_0, \rho)'$.

Estimation of the parameters θ proceeds by maximizing the full log-likelihood function $L = \sum_i \ln L_i^* = \sum_i L_i$. A key step is to derive the *score vector*, which is the vector of derivatives $\partial L / \partial \theta$, with

$$\frac{\partial L}{\partial \theta} = \sum_i \frac{\partial L_i}{\partial \theta}.$$

Note that household i's contribution to the score is

$$\frac{\partial L_i}{\partial \theta} = \frac{\sum_j w_j \frac{\partial P_{ij}}{\partial \theta}}{\sum_j w_j P_{ij}}.$$

It will prove helpful to re-express this derivative in the form

$$\frac{\partial L_i}{\partial \theta} = \frac{\sum_j w_j P_{ij} \frac{\partial \ln P_{ij}}{\partial \theta}}{\sum_j w_j P_{ij}},$$

because the derivatives of $\ln P_{ij}$ with respect to θ are generally similar to their counterparts in models without random factors.

Estimation of the model

For convenience, we repeat here the latent variable equation (40.8),

$$Z_{ik}^* = \alpha_k + \beta_k u_i + v_{ik}. \tag{40.8}$$

In constructing probability expressions for the observed indicators Z_{ik}, we assume that the disturbance term v_{ik} of equation (40.8) is normally distributed with mean zero and variance $\sigma_{v_k}^2$. We take u_i and v_{ik} to be independent of each other for all i and k, and assume that the elements of $\{v_{ik},\ k = 1, \ldots, K\}$ are mutually independent. Hence, although the various Z_{ik}^* are inter-correlated, their correlations stem from a common dependence on the u_i factor. Conditional on u_i, the latent variables Z_{ik}^* are independent, as are their observable Z_{ik} counterparts.

In probit-like structures such as these, the sizes of the disturbance variances are not identified and some normalization rule must be imposed. Following in the spirit of Heckman (1981, p. 129), we choose to normalize things so that the variance of $\beta_k u_i + v_{ik}$ equals one, that is, $\beta_k^2 \rho + \sigma_{v_k}^2 = 1$. This is a convenient rule to apply if one begins with $\hat{\alpha}_k$ estimates from standard probit models, since those estimates are based on an assumed disturbance variance of unity. Note that under the normalization rule, the variance of v_{ik} is $1 - \beta_k^2 \rho$. We also define $\beta_1 \equiv 1$ for reasons to be explained below.

Equation (40.8), which defines the latent indicator Z_{ik}^*, may now be multiplied through by $r_k = (1 - \beta_k^2 \rho)^{-1/2}$ to give a result expressed in the usual probit form. We can see that

$$r_k Z_{ik}^* = r_k(\alpha_k + \beta_k u_i) + r_k v_{ik}$$

is in the desired form since $r_k v_{ik}$ is standard normal. The probability associated with the observed dependent variable Z_{ik}, conditional on the random factor u_i, is then

$$\Pr\big(Z_{ik} = z_{ik} | u_i\big) = \Phi\big(z_{ik} r_k \cdot (\alpha_k + \beta_k u_i)\big),$$

where Φ is the standard normal cumulative distribution function, and we have made use of our unconventional $\{1, -1\}$ coding scheme for Z_{ik} and the symmetry of the normal distribution. The product of such probabilities over all indicators for household i is

$$P(u_i) = \prod_{k=1}^{K} \Phi\big(z_{ik} r_k \cdot (\alpha_k + \beta_k u_i)\big).$$

Recall that to integrate out the unobservable random effect u, we need the quadrature approximation to an integral of the general form,

$$\int_{-\infty}^{\infty} (2\pi)^{-1/2} \rho^{-1/2} e^{-\frac{1}{2\rho}u^2} P(u) du.$$

Applying the change of variables and using n_q quadrature points, we obtain

$$L_i^* = \pi^{-1/2} \sum_{j=1}^{n_q} w_j P_i(\theta_0, \sqrt{2}\rho^{1/2} e_j) = \pi^{-1/2} \sum_{j=1}^{n_q} w_j P_{ij},$$

in which $\theta_0 = (\alpha, \beta)'$, this being a vector of length $2K - 1$ containing all unknown parameters except for ρ, the variance of the factor (recall that $\beta_1 \equiv 1$). When we need to see the roles of the parameters more clearly, we write out the expression for P_{ij} in full, as

$$P_{ij} = P_i(\theta_0, \sqrt{2}\rho^{1/2}e_j) = \prod_{k=1}^{K} \Phi\big(z_{ik}r_k \cdot (\alpha_k + \beta_k\sqrt{2}\rho^{1/2}e_j)\big).$$

Below we will refer to this expression as $P_{ij}(\theta)$, a notation in which the vector $\theta = (\alpha, \beta, \rho)'$, of length $2K$, contains all of the model's unknown parameters.

The scores

Recall that household i's contribution to the full score vector is

$$\frac{\partial L_i}{\partial \theta} = \frac{\sum_j w_j P_{ij} \frac{\partial \ln P_{ij}}{\partial \theta}}{\sum_j w_j P_{ij}}.$$

Now, $\ln P_{ij}(\theta)$ is itself the sum over k of the logs of the probabilities specific to indicator k:

$$\ln P_{ij}(\theta) = \sum_{k=1}^{K} \ln \Phi\big(z_{ik}r_k \cdot (\alpha_k + \beta_k\sqrt{2}\rho^{1/2}e_j)\big). \tag{40.10}$$

Hence, for the α parameters we take derivatives of equation (40.10) to obtain

$$\frac{\partial \ln P_{ij}}{\partial \alpha_k} = \frac{\phi_{ik,j}}{\Phi_{ik,j}} z_{ik}r_k, \tag{40.11}$$

with $\phi_{ik,j}$ being the derivative of $\Phi_{ik,j}$ with respect to its argument. Both $\phi_{ik,j}$ and $\Phi_{ik,j}$ are evaluated at the point $z_{ik}r_kW_{kj}$, with $W_{kj} = \alpha_k + \beta_k\sqrt{2}\rho^{1/2}e_j$. Note that the expression involves only parameters specific to the k-th indicator.

For the β parameters, we face a more complicated derivation because r_k depends on β_k. For $k \geq 2$ (again recall that $\beta_1 \equiv 1$), the result is

$$\frac{\partial \ln P_{ij}}{\partial \beta_k} = \frac{\partial \ln P_{ij}}{\partial \alpha_k} \cdot \left(W_{kj}r_k^2\beta_k\rho + \sqrt{2}\rho^{1/2}e_j.\right) \tag{40.12}$$

As for the derivative with respect to ρ, a parameter that enters all of the indicator equations, if we recall that r_k is also a function of ρ we obtain

$$\frac{\partial \ln P_{ij}}{\partial \rho} = \sum_{k=1}^{K} \frac{\partial \ln P_{ij}}{\partial \alpha_k} \cdot \frac{\beta_k}{2}\left(W_{kj}r_k^2\beta_k + \sqrt{2}\rho^{-1/2}e_j\right). \tag{40.13}$$

These results provide all the ingredients needed to estimate the model.

Notes on identification

In setting out the multiple-indicator model, we have imposed a number of restrictions, and some comment is in order on why these are needed and how the restrictions help to identify the parameters. Note first that the restriction $\beta_1 = 1$ is something more than a trivial normalization. Consider a model with a given set of $\{\beta_k\}$ parameters. Because the unobserved factor u_i is symmetrically distributed about zero, given normality, a second model that is observationally equivalent to the first can be constructed by reversing the signs of all of the β_k parameters while leaving their magnitudes untouched. Fixing $\beta_1 = 1$ eliminates this possibility. However, in choosing to set the first of the β_k parameters to unity, we are making the assumption that the first indicator Z_{i1} is known to be positively associated with the unmeasured factor. If there is any doubt about this assumption, another indicator should be used in its place.

A second point to note is that the variances of the composite disturbance terms—by "composite" we mean $u_i + v_{i1}$ for the first indicator and $\beta_k u_i + v_{ik}$ for the k-th—are not identified in latent variable models with binary indicators. By setting each of the composite variances to unity, we are imposing restrictions that acknowledge this fact.

Consider, then, a two-indicator model. The unknown parameters of this model are α_1, α_2, β_2, the factor's variance ρ, and the disturbance variances $\sigma^2_{v_1}$ and $\sigma^2_{v_2}$, giving a total of six parameters. Two restrictions are imposed via the unit variance assumptions, and this reduces the number of unknowns to four. However, the data at hand provide us with only three quantities that can be calculated: conventional single-equation probit models supply consistent estimates of α_1 and α_2, and the covariance between $Z^*_{i,1}$ and Z^*_{i2} can be estimated consistently by a bivariate probit.[5] Unless further assumptions can be made, the two-indicator model is clearly under-identified.

Counting up parameters and calculable quantities for a three-indicator model shows that this model is just-identified. After imposing variance restrictions, we are given six parameters to estimate. Three conventional probits identify the α_k parameters, and three

[5]This is the approach taken in the first step of LISREL's two step estimation method, and it is no doubt used in other software as well. The two-step approach generally proceeds as follows. Consider the k-th indicator equation, written out in a latent form comparable to that of equation (40.8) and also (see below) the MIMIC specification of (40.19)

$$Z^*_k = \alpha_k + \mathbf{X}' g_k + \omega_k,$$

where we have taken $g_k = \beta_k \gamma$ and $\omega_k = \beta_k u + v_k$. We assume that for each k, the variance of the disturbance term ω_k is normalized to unity as in a conventional probit model. Let $\omega = (\omega_1, \ldots, \omega_K)'$ be a vector containing all K of these disturbances, and let Ω, a $K \times K$ matrix, represent the correlation matrix of ω.

The first step of the two-step approach uses bivariate probit methods to estimate each of the correlations $\hat{\sigma}^2_{w,jk}$ that appear in the off-diagonal entries of Ω, with the bivariate models being estimated separately for each distinct pair of indicators. The approach generates some unwanted byproducts—multiple estimates of each α_k parameter and a set of $\hat{g}_k$ estimates that are not decomposed into separate $\hat{\beta}_k$ and $\hat{\gamma}$ components. Further manipulations, not described here, are needed to extract estimates of these parameters.

With the unrestricted $\hat{\Omega}$ in hand, one then imposes the specification

$$\omega_k = u\beta_k + v_k \quad \text{for } k = 1, \ldots, K.$$

and given this structure, obtains a functional form for each off-diagonal correlation. In this way, any trial set of parameter estimates $\{\tilde{\beta}_k\}$ and $\tilde{\sigma}^2_u$ will imply a trial covariance matrix $\tilde{\Omega}$, which may be compared with the unrestricted $\hat{\Omega}$ matrix. In the course of estimation, values for $\{\hat{\beta}_k\}$ and $\tilde{\sigma}^2_u$ are found that minimize the difference between these matrices.

applications of bivariate probit supply estimates of the three cross-equation covariances. By the same logic, models with four indicators or more are over-identified. Each additional indicator adds a new pair of α_k, β_k parameters to estimate, to be sure, but each indicator also makes available a new set of cross-equation covariances that help in estimating all of the β_k parameters and the ρ parameter.

If many indicators are available, some of the assumptions made above can be relaxed. For instance, if the model is over-identified given the assumption of zero covariance between disturbance components v_{ij} and v_{ik}, $j \neq k$, then additional parameters can be introduced to allow for a limited number of non-zero covariances.

Numerical optimization issues

Our experience in estimating these models suggests that on occasion they present numerical difficulties. In particular, we have encountered cases in which one of the normalizing factors $r_k = (1 - \beta_k^2 \rho)^{-1/2}$ behaves badly as the result of a steady drift upward in its β_k. We have not been able to diagnose the root cause of the problem; fortunately, it is generally easy to correct. To arrest the tendency for one or more of the β_k to drift upward, we have programmed special checks that are applied during the course of optimization, which temporarily reduce the absolute amount of change permitted in the parameters once such drift is detected. Slowing things down in this way generally allows the optimization to regain its footing and things usually proceed smoothly thereafter. As a further safeguard, we have estimated the models using an initial grid search over ρ, estimating all other parameters for each ρ value in the grid. The best estimates $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\rho}$ emerging from this grid search are presented as starting values to a full maximum likelihood estimation routine.

Estimating the unobserved factor

Even though the factor u_i is unobserved, we can estimate its value from the values of the observed indicators $\mathbf{Z}_i$ for that observation. The procedure is little more than an application of Bayes' Rule. We seek the conditional expectation

$$E(u_i|\mathbf{Z}_i) = \int u\, P(u|\mathbf{Z}_i)\, du, \tag{40.14}$$

in which the conditional density $P(u|\mathbf{Z}_i)$ is the density for the factor u given the indicator vector $\mathbf{Z}_i$ for the i-th household. By Bayes' Rule,

$$P(u|\mathbf{Z}_i) = \frac{P(u, \mathbf{Z}_i)}{P(\mathbf{Z}_i)} = \frac{P(\mathbf{Z}_i|u)\phi(u)}{P(\mathbf{Z}_i)}, \tag{40.15}$$

with $\phi(u)$ being the normal density function for a factor with mean 0 and variance ρ. Note that $P(\mathbf{Z}_i)$ is the contribution made by observation i to the sample likelihood.

Given realized values $\mathbf{Z}_i = \mathbf{z}_i$, the numerator of $P(u|\mathbf{Z}_i)$, as it is expressed on the right-hand side of equation (40.15), can be written as

$$\prod_{k=1}^{K} \Phi\big(z_{ik} r_k \cdot (\alpha_k + \beta_k u)\big) \cdot \phi(u), \tag{40.16}$$

and the denominator of equation (40.15) is the integral of (40.16) over u.

To calculate the conditional expectation of u, we start with the quadrature approximation to $\int u\, P(\mathbf{Z}_i|u)\phi(u)\, du$, which is

$$\pi^{-1/2} \sum_{j=1}^{n_q} w_j (\sqrt{2}\rho^{1/2} e_j) \cdot \prod_{k=1}^{K} \Phi\left(z_{ik} r_k \cdot (\alpha_k + \beta_k \sqrt{2}\rho^{1/2} e_j)\right). \tag{40.17}$$

In this expression, the first component in parentheses, $\sqrt{2}\rho^{1/2}e_j$, stands in for u. To complete the quadrature approximation to equation (40.14), we divide equation (40.17) by the approximation to $P(\mathbf{Z}_i)$, which is

$$\pi^{-1/2} \sum_{j=1}^{n_q} w_j \prod_{k=1}^{K} \Phi\left(z_{ik} r_k \cdot (\alpha_k + \beta_k \sqrt{2}\rho^{1/2} e_j)\right). \tag{40.18}$$

These calculations are carried out using the estimated $\hat{\alpha}_k$, $\hat{\beta}_k$, and $\hat{\rho}$ parameters.[6]

40.4 Multiple Indicators, Multiple Causes (MIMIC)

With all of this as background, we now generalize things by allowing the unobserved factor to be determined by a set of observed exogenous variables $\mathbf{X}_i$ as well as an unobserved component u_i. This MIMIC model ("multiple indicator, multiple cause") may be represented as $F_i = \mathbf{X}_i'\gamma + u_i$, where F_i is the latent factor, the $\mathbf{X}_i$ covariates are its observed determinants, and u_i is its unobserved determinant, assumed to be independent of $\mathbf{X}_i$. In this approach, the latent indicator Z_{ik}^* is written out as

$$\begin{aligned} Z_{ik}^* &= \alpha_k + \beta_k F_i + v_{ik} \\ &= \alpha_k + \beta_k \mathbf{X}_i'\gamma + \beta_k u_i + v_{ik}. \end{aligned} \tag{40.19}$$

We apply the unit variance restrictions as before,

$$r_k Z_{ik}^* = r_k \left(\alpha_k + \beta_k \mathbf{X}_i'\gamma + \beta_k u_i\right) + r_k v_{ik}, \tag{40.20}$$

and obtain

$$\ln P_{ij} = \sum_{k=1}^{K} \ln \Phi\left(z_{ik} r_k (\alpha_k + \beta_k \mathbf{X}_i'\gamma + \beta_k \sqrt{2}\rho^{1/2} e_j)\right), \tag{40.21}$$

also much as before.

The forms of the scores in the α_k and β_k dimensions are essentially unchanged. For the other parameters, however, we have

$$\frac{\partial \ln P_{ij}}{\partial \gamma} = \sum_{k=1}^{K} \frac{\partial \ln P_{ij}}{\partial \alpha_k} \cdot \beta_k \cdot \mathbf{X}_i, \tag{40.22}$$

[6]The factor u, being normally distributed, takes on negative as well as positive values. It may be that quadrature approximations to the conditional expectation of u perform poorly unless the integrand $uP(u|Z_i)$ is positive. An easy solution is to add a large positive constant to u (i.e., to its proxy $\sqrt{2}\rho^{1/2}e_j$ that appears immediately after the summation sign in equation (40.17)) and then subtract that constant after the integral has been calculated.

a vector of the same dimension as the $\mathbf{X}_i$ vector, and

$$\frac{\partial \ln P_{ij}}{\partial \rho} = \frac{\partial \ln P_{ij}}{\partial \alpha_k} \cdot \left(W_{kj} r_k^2 \beta_k \rho + \mathbf{X}_i' \gamma + \sqrt{2} \rho^{1/2} e_j \right) \tag{40.23}$$

with $W_{kj} = \alpha_k + \beta_k \mathbf{X}_i' \gamma + \beta_k \sqrt{2} \rho^{1/2} e_j$. This definition of W_{kj} would also be used in the modified versions of equations (40.11) and (40.12).

As for estimating the unobserved factor, there is little to distinguish the MIMIC model from the standard model. In this case we aim to estimate $F_i = \mathbf{X}_i' \gamma + u_i$ conditional on $\mathbf{Z}_i$ and $\mathbf{X}_i$. We employ $\hat{\gamma}$ for the first term and then apply the procedures that were outlined above to predict u_i.

Chapter 41

Dynamic Discrete-Choice Models

41.1 Models with Observed States and Controls

Let us consider a dynamic problem set in discrete time with a finite horizon. The *states* of the problem are discrete-valued and are represented by x_t, a random vector. The *controls* are c_t, and these are also discrete-valued.[1] The control variables can be viewed as a set of distinct choice options. The "equations of motion" of the problem are summarized in the conditional probability distribution $P(x_{t+1} \mid x_t, c_t)$, which takes a first-order Markov form. For problems with longer-lived state dependence, the states x_t can be defined to include past values, so that the first-order nature of the transition probabilities is less restrictive than it may appear to be.

The decision-maker has a terminal value function $U(x_T)$ and a period-specific utility function $u(x_t, c_t)$. We could generalize $u(x_t, c_t)$ to $u_t(x_t, c_t)$ and likewise generalize $P(x_{t+1} \mid x_t, c_t)$ to $P_t(x_{t+1} \mid x_t, c_t)$ if the problem requires it. The set of admissible control variables can depend on the state and period, if that is needed. A discount factor can also be brought into the problem.

Bellman's equation

Given an initial state x_0, we have

$$V_0(x_0) = \max_{c_1, c_2, \ldots, c_{T-1}} E_0 \left(\sum_{t=1}^{T-1} u(x_t, c_t) + U(x_T) \right),$$

where E_0 denotes expectations formed as of time $t = 0$. Bellman's equation provides a recursive representation of the value function, such that

$$V_t(x_t) = \max_{c_t} u(x_t, c_t) + E_t \left(V_{t+1}(x_{t+1}) \mid x_t, c_t \right).$$

[1] We should use upper-case notation for both states and controls, following the convention in which random variables are represented by upper-case letters and realizations of these variables by lower case. But such rigorous notation goes beyond our needs in this handout and we will leave all the random variables in lower case.

Now, for a given x_t and c_t,

$$E_t\left(V_{t+1}(x_{t+1}) \mid x_t, c_t\right) = \sum_{x_{t+1}} V_{t+1}(x_{t+1}) P(x_{t+1} \mid x_t, c_t) \equiv \bar{V}_{t+1}(x_t, c_t).$$

The $\bar{V}_{t+1}$ notation helps to emphasize how, as of period t, the conditional expectation of the value function for period $t+1$ depends on the current state x_t and the value c_t for the current control. Using this notation, we can re-express Bellman's equation as

$$V_t(x_t) = \max_{c_t} u(x_t, c_t) + \bar{V}_{t+1}(x_t, c_t). \tag{41.1}$$

The value of c_t that maximizes the right-hand side is the optimal value for period t, because it is chosen with reference both to its implications for current utility $u(x_t, c_t)$ and future utility, the latter taking effect through the $\bar{V}_{t+1}(x_t, c_t)$ function.

41.2 Unobserved States: Rust's Approach

In the above, there was no counterpart to the disturbance term of a conventional econometric model. Rust (1987) worked out a method of including such unobservables, and we describe his important contribution below. Let ϵ_t be a vector of unobserved state variables in period t, and assume that its j-th element is associated with the j-th choice option. Denote the j-th element of ϵ_t by $\epsilon_t(j)$. We will let ϵ_t be continuously distributed and will further assume that it enters the period-specific utility function in a linear fashion, giving

$$u(x_t, \epsilon_t, c_t) = u(x_t, c_t) + \epsilon_t(c_t)$$

as the new period-specific utility function.

Proceeding as above, we can write Bellman's equation with the new state variables as follows,

$$V_t(x_t, \epsilon_t) = \max_{c_t} u(x_t, c_t) + \epsilon_t(c_t) + \bar{V}_{t+1}(x_t, \epsilon_t, c_t), \tag{41.2}$$

where

$$\bar{V}_{t+1}(x_t, \epsilon_t, c_t) = \sum_{x_{t+1}} \int_{\epsilon_{t+1}} V_{t+1}(x_{t+1}, \epsilon_{t+1}) P(x_{t+1}, \epsilon_{t+1} \mid x_t, \epsilon_t, c_t)\, d\epsilon_{t+1},$$

and $P(\cdot \mid \cdot)$ is a mixed probability mass and probability density function.

To render things tractable, a key assumption is now invoked. Suppose that the equation of motion can be factored as follows,

$$P(x_{t+1}, \epsilon_{t+1} \mid x_t, \epsilon_t, c_t) = q(\epsilon_{t+1} \mid x_{t+1}) \cdot p(x_{t+1} \mid x_t, c_t).$$

To say this is to make a strong assumption—ruling out any serial correlation in ϵ_t that is not the result of correlation in x_t—but the assumption has a considerable payoff.

Inserting the reformulated transition probability in the expression for $\bar{V}_{t+1}(\cdot)$ yields

$$\begin{aligned} \bar{V}_{t+1}(x_t, \epsilon_t, c_t) &= \sum_{x_{t+1}} \left(\int_{\epsilon_{t+1}} V_{t+1}(x_{t+1}, \epsilon_{t+1}) q(\epsilon_{t+1} \mid x_{t+1})\, d\epsilon_{t+1} \right) p(x_{t+1} \mid x_t, c_t) \\ &= \bar{V}_{t+1}(x_t, c_t). \end{aligned} \tag{41.3}$$

That is, the conditional expected value function does not depend on the current value of ϵ_t, the unobserved state vector. We can, therefore, simplify equation 41.2 to

$$V_t(x_t, \epsilon_t) = \max_{c_t} u(x_t, c_t) + \epsilon_t(c_t) + \bar{V}_{t+1}(x_t, c_t), \tag{41.4}$$

noting that the unobserved state variables $\epsilon_t(c_t)$ enter (directly) only the middle term.

We substitute the value function $V_{t+1}(x_{t+1}, \epsilon_{t+1})$ into equation 41.3, the definition of the conditional expected value function $\bar{V}_{t+1}(x_t, c_t)$, and obtain

$$\begin{aligned} \bar{V}_{t+1}(x_t, c_t) \quad = \quad & \sum_{x_{t+1}} \left(\int_{\epsilon_{t+1}} \{ \max_{c_{t+1}} u(x_{t+1}, c_{t+1}) + \epsilon_{t+1}(c_{t+1}) + \bar{V}_{t+2}(x_{t+1}, c_{t+1}) \} q(\epsilon_{t+1} \mid x_{t+1}) d\epsilon_{t+1} \right) \\ & \times p(x_{t+1} \mid x_t, c_t) \end{aligned} \tag{41.5}$$

If ϵ_t is a vector of i.i.d. extreme value disturbances, then we have a simple representation for the expression in large parentheses. It is

$$\ln \left(\sum_c e^{u(x_{t+1}, c) + \bar{V}_{t+2}(x_{t+1}, c)} \right) \equiv G_{t+1}(x_{t+1})$$

We have seen this expression before in connection with static conditional logit choice models.

Drawing all this together, we obtain a simplified expression for equation 41.5, the conditional expected value function,

$$\bar{V}_{t+1}(x_t, c_t) = \sum_{x_{t+1}} G_{t+1}(x_{t+1}) p(x_{t+1} \mid x_t, c_t) \tag{41.6}$$

Returning, to equation 41.4 with $\bar{V}_{t+1}(x_t, c_t)$ in hand, we see that the decision problem for c_t is much like that for a standard, static, conditional logit problem. Indeed, we obtain the familiar logit choice probability,

$$\mathrm{P}\left(c_t = c \mid x_t\right) = \frac{e^{u(x_t, c) + \bar{V}_{t+1}(x_t, c)}}{\sum_j e^{u(x_t, j) + \bar{V}_{t+1}(x_t, j)}}, \tag{41.7}$$

which differs from its static counterpart in the inclusion of $\bar{V}_{t+1}(x_t, j)$, a term that accounts for the influence of expected future utility on current choices.

Chapter 42

Using Sampling Weights

What *are* sampling weights, anyway? Until recently, it was very difficult to find any discussion of these weights in the econometrics literature, and even today, the leading textbooks manage to ignore the subject. Fortunately, Deaton (1997) devotes two superb chapters to the topic, and to give an overview we need do little more than repeat his analyses.

We will define sampling weights more precisely below, but here it may be useful to draw distinctions among three general types of weights, only one of which should be termed a sampling weight. A sampling weight for person (unit) i, or w_i, can be defined as the inverse of the probability that this person i is selected for inclusion in a sample survey from a list of those eligible to be included.

This use of the term "weight" is to be distinguished from two other uses. Weights are sometimes employed in estimation because some observations are thought to be more reliable than others, and we want to down-weight the less reliable observations in our estimation procedure. A leading example would be the case in which the dependent variable for observation i, call it Y_i, is an average based upon n_i more finely disaggregated data points. This case would arise if our data are state-level averages Y_i, each Y_i being based on n_i individual observations for that state. The sampling variation of Y_i then depends inversely on the number of data points n_i that contributed to it; when n_i is small, Y_i is less reliable than when n_i is large. In linear models, we think of such variations in reliability as a form of heteroskedasticity, and use weights to restore homoskedasticity.

Yet a third use of the term "weight" refers to complex survey designs that deliberately generate sample selection biases, such as in the case-control approach used in studies of rare health conditions. Statisticians will devise weights for these surveys that are meant to be used to counteract the selection biases. Econometricians should not dismiss such weights out of hand. Rather, the method used to formulate the weights should be scrutinized to determine if it is consistent with the econometrician's own view of the structural model. If the procedure used to formulate the weights is inconsistent with this model, then the problem of how best to deal with selectivity is left with the econometrician.

In what follows, we restrict ourselves to the consideration of sampling weights. We begin by imagining a survey whose purpose is to estimate national income from data provided by the respondents on their individual incomes. The respondents are selected from a "sampling frame," which for simplicity we take to be a full listing of the population.

Let π_i be the probability that person i is selected in a single draw from this sampling frame, and let $n\pi_i$ be an approximation to the probability that person i is selected in n such draws, where n is the sample size. We assume that the chance is negligible that person i might be selected more than once, and assume that sampling occurs with replacement.

In a simple random sample, $\pi_i = 1/N$, where N is the population size, and $n\pi_i = n/N$. The expression $(n\pi_i)^{-1} = N/n$ can be viewed as an "inflation factor" that expands person i's data to represent that of N/n persons. In general—that is, in cases more complicated than simple random samples—we let the sampling weight $w_i = (n\pi_i)^{-1}$ represent the inflation factor for the i-th person.

42.1 Estimating the Total Population Size

To understand these sampling weights $\{w_i,\ i = 1, \ldots, n\}$, let us take them to be given for a sample of size n and consider the problem of how to estimate the total population size N from the weights. This must be an artificial problem, since at least some knowledge of the size of the total population would be required to formulate the weights. The estimator of the total population size is

$$\hat{N} = \sum_{i=1}^{n} w_i,$$

and we will show that $\hat{N}$ is unbiased for N, that is, $\mathrm{E}\ \hat{N} = N$. In the proof, we change the range of the summation to extend over the full population and compensate by inserting in the expression a random variable t_i, which takes the value of one if observation i is selected for the sample and is zero otherwise. This reformulation yields

$$\hat{N} = \sum_{i=1}^{N} t_i w_i,$$

and, since $\mathrm{E}\ t_i = n\pi_i$, we have

$$\mathrm{E}\ \hat{N} = \sum_{i=1}^{N} n\pi_i (n\pi_i)^{-1} = N.$$

42.2 Estimating Totals

Let us apply this general method of proof to the problem of estimating a total quantity Y, which we take to be total national income, from a sample of values Y_i of individual incomes. The quantity that we hope to estimate is $Y = \sum_{i=1}^{N} Y_i$, but we must estimate this using only the Y_i values that happen to be in our sample, along with the sampling weights.

To understand what follows, it is important to understand how the Y_i are viewed. Rather than viewing the Y_i (and hence Y) as random variables—the natural approach in an econometric analysis—here we take them to be fixed quantities; from this perspective, they are not unlike individual-specific parameters. We are interested in the sum of such parameters or quantities over the full population. In short, our aim is not to engage in

econometric *modeling* as such, but rather to achieve a statistically rigorous *description* of population totals.

Consider, then, the estimator

$$\hat{Y} = \sum_{i=1}^{n} w_i Y_i = \sum_{i=1}^{N} t_i w_i Y_i.$$

Taking its expected value, we obtain

$$\mathrm{E}\ \hat{Y} = \sum_{i=1}^{N} n\pi_i (n\pi_i)^{-1} Y_i = \sum_{i=1}^{N} Y_i = Y,$$

that is, the weighted sum of the Y_i in the sample is an unbiased estimator of Y, the population total.

Suppose, however, that we had decided not to use the sampling weights in this analysis, and had instead formed the estimator

$$\bar{Y} = \sum_{i=1}^{n} Y_i = \sum_{i=1}^{N} t_i Y_i.$$

The expected value of $\bar{Y}$ is

$$\mathrm{E}\ \bar{Y} = n \sum_{i=1}^{N} \pi_i Y_i \neq Y.$$

This unweighted estimator yields biased estimates of the population total.

42.3 Estimating Population Means

Suppose that our aim is to estimate the population mean, which is denoted by $y = (1/N)\sum_{i=1}^{N} Y_i$. The population mean is *not* viewed as the expected value of a random variable, but is rather a simple arithmetic average of the fixed quantities Y_i. We consider the weighted estimator

$$\hat{y} = \frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{N} t_i w_i Y_i}{\sum_{i=1}^{N} t_i w_i} = \frac{\hat{Y}}{\hat{N}}.$$

Because this estimator is a ratio of random variables, we cannot evaluate its expectation directly (unless N happens to be known, in which case only the numerator depends on random variables), but we can derive its probability limit.[1] Applying a line of reasoning similar to that used above, we find that $\hat{y} \xrightarrow{p} y$.

If we ignore the weights and form the alternative, unweighted estimator

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} Y_i,$$

[1] To apply such limiting concepts, we must assume that N is very large in relation to the sample size n, or think of both N and n increasing.

we find that it converges to

$$\frac{1}{n}\sum_{i}^{N} n\pi_i Y_i = \sum_{i}^{N} \pi_i Y_i,$$

which is not equal to the population average y in general. However, in the special case of simple random sampling, we have $\pi_i = 1/N$, so that in this case, the unweighted sample mean is consistent for the population mean.

To summarize: for general sampling schemes, weighted estimators are required for consistent estimation of population totals and means. In the special case of simple random sampling, however, the population mean y can be consistently estimated without the weights, although the weights are still required to estimate Y, the population total.

42.4 Weights and Econometric Modeling

We have been emphasizing their role in description—but should sampling weights also be used in modeling? One important use of weights is to summarize the results of an estimated model in a way that refers to the full population to which the results apply. For instance, in the case of a probit model, in which the estimated probability of a "yes" is given by $\hat{P}_i = \Phi(X_i'\hat{\beta})$ with X_i being a vector of covariates for observation i and $\hat{\beta}$ the estimated parameters, one could summarize the average predicted probability over the population using

$$\hat{p} = \frac{\sum_{i=1}^{n} w_i \hat{P}_i}{\sum_{i=1}^{n} w_i}$$

much as was done above for average income. Likewise, the derivatives of the probit model, which are measures of the extent to which the predicted probabilities change with X_i, can be summarized at the population level with the aid of sampling weights.

But this is akin to description, in which weights are brought on board in the final stage of summarizing results. To see whether weights need to be considered earlier on, in estimation, we need to consider three questions: (1) What are the parameters of interest? (2) Do these parameters vary over the population? (3) If the parameters vary, is this variation associated with the explanatory variables?

In some disciplines, sampling weights are employed as a safeguard against certain forms of population heterogeneity. To see the argument, consider a population that is divided into S sectors. Suppose that a linear model

$$Y_s = X_s\beta_s + \epsilon_s$$

describes each sector, and let there be n_s observations in the s-th sector. Assume that the disturbance vector ϵ_s is homoskedastic with covariance matrix $\sigma_s^2 I$ and is uncorrelated with the explanatory variables.

Now, under these assumptions, ordinary least squares is the best linear unbiased estimator for the β_s parameters. To use sampling weights could only increase the variance of the $\hat{\beta}_s$ estimator. Evidently, then, if there is a rationale for using weights in estimation, it must be found elsewhere. Heterogeneity alone is an insufficient justification.

Suppose that we are interested not only in the individual sector β_s parameters, but also in their population-weighted sum. That is, we seek to estimate the quantity

$$\beta = \sum_{s=1}^{S} \frac{N_s}{N} \beta_s.$$

How might we proceed?

One approach is to estimate the β_s coefficients by ordinary least squares (or seemingly-unrelated regressions if there are efficiency gains to be secured through that method) and then employ the sampling weights to form the population-weighted average. Given $\hat{\beta}_s$ for each sector, we would form

$$\hat{\beta} = \sum_{s=1}^{S} \frac{\hat{N}_s}{\hat{N}} \hat{\beta}_s,$$

with $\hat{N}_s = \sum_{i=1}^{n_s} w_{is}$ and $\hat{N} = \sum_s \hat{N}_s$. This approach is similar to what we described above, where we estimated a model without weights but then used the weights to summarize the results to the population level. We have still not seen an argument for using the weights in estimation.

To understand why weights might be taken into account, let us consider a particular sampling scheme whereby simple random sampling is undertaken within each sector s, but some sectors are over-sampled relative to their population shares and others under-sampled. In such a design, the sample shares n_s/n need not coincide with the population shares N_s/N, and the inflation factor for the i-th person in the s-th sector is simply $w_{is} = N_s/n_s$, a constant for all persons in that sector.

The pooled OLS estimator

Because the β_s vary across sectors, we would not expect a pooled estimator, which ignores this variation, to have any desirable statistical properties. Nevertheless, it is useful to see how the pooled estimator goes wrong.

The pooled estimator is

$$\bar{\beta} = \left(\sum_s X_s' X_s \right)^{-1} \left(\sum_s X_s' Y_s \right).$$

What is the probability limit of this estimator? We will use the usual trick of dividing and multiplying by the sample size. Letting $n_s \to \infty$ for all s, we have

$$\begin{aligned} \operatorname{plim} \frac{1}{n_s} X_s' X_s &= Q_s \\ \operatorname{plim} \frac{1}{n_s} X_s' Y_s &= Q_s \beta_s. \end{aligned}$$

We let the sample shares n_s/n stay constant as the overall sample size n goes to infinity. Multiplying and dividing by both n_s and n, we obtain

$$\bar{\beta} \stackrel{a}{=} \left(\sum_s \frac{n_s}{n} Q_s \right)^{-1} \left(\sum_s \frac{n_s}{n} Q_s \beta_s \right).$$

Now, if it happens to be the case that all $Q_s = Q$, this expression would simplify to

$$\bar{\beta} \stackrel{a}{=} \sum_s \frac{n_s}{n} \beta_s.$$

However, even in this case $\bar{\beta}$ would not be equal to the population-weighted average of the β_s, because the sample shares n_s/n are not equal to the population shares N_s/N, at least not in general. In a simple random sample, of course, these shares *are* equal and $\bar{\beta}$ will then be asymptotically equivalent to β, the population-weighted average. But this combination of a simple random sample and common Q_s seems most unlikely.

The weighted pooled estimator

The weighted pooled estimator can be written as

$$\tilde{\beta} = \left(\sum_s \frac{N_s}{n_s} X_s' X_s \right)^{-1} \left(\sum_s \frac{N_s}{n_s} X_s' Y_s \right),$$

since the weights $w_{is} = N_s/n_s$. Moving to probability limits, we obtain

$$\tilde{\beta} \stackrel{a}{=} \left(\sum_s \frac{N_s}{N} Q_s \right)^{-1} \left(\sum_s \frac{N_s}{N} Q_s \beta_s \right),$$

where we have multiplied and divided only by N.

Now, in this case, and in contrast to the simple pooled OLS case, if it happens that $Q_s = Q$, the weighted estimator $\tilde{\beta}$ will converge to β, the desired population-averaged quantity. Are we ever likely to have $Q_s = Q$? Yes, if we are interested only in the mean, this being a case in which X_s contains only a constant term. Otherwise, the case of common Q_s still seems unlikely.

But consider another way of writing the probability limit of the weighted pooled estimator, as

$$\tilde{\beta} \stackrel{a}{=} \left(\sum_s \frac{N_s}{N} Q_s \right)^{-1} \left(\sum_s \frac{N_s}{N} Q_s (\beta + (\beta_s - \beta)) \right).$$

This can be reexpressed as

$$\tilde{\beta} \stackrel{a}{=} \beta + \left(\sum_s \frac{N_s}{N} Q_s \right)^{-1} \left(\sum_s \frac{N_s}{N} Q_s (\beta_s - \beta) \right).$$

Consider the last expression on the right. If the variation in the β_s parameters across sectors is unrelated to the variation in Q_s, averaging out to zero when weighted by the population shares N_s/N, then the weighted pooled estimator is consistent for β, the average parameter value that we seek. So, at least under certain forms of parameter heterogeneity, weighting the data with sampling weights will indeed produce consistent estimates of the population-averaged β parameter.

Appendix A

Calculus on vectors and matrices

There are two conventions about how to represent derivatives involving vectors and matrices: the gradient approach and what I'll term the Jacobian approach. To illustrate, if we differentiate the scalar function $y = c(\mathbf{x})$, with $\mathbf{x}$ a vector of length m, then the gradient representation of the derivative is an $m \times 1$ column vector

$$\nabla c(\mathbf{x}) \equiv \frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \partial c/\partial x_1 \\ \partial c/\partial x_2 \\ \vdots \\ \partial c/\partial x_m \end{bmatrix}$$

In the Jacobian representation, however,

$$\mathcal{J} c(\mathbf{x}) = \begin{bmatrix} \partial c/\partial x_1 & \partial c/\partial x_2 & \cdots \partial c/\partial x_m \end{bmatrix},$$

which is a $1 \times m$ row vector. Mathematicians tend to favor this latter approach, especially in the case of differentiation of a *vector* of functions with respect to a *vector of arguments*. Let $\mathbf{y} = \mathbf{c}(\mathbf{x})$ in which $\mathbf{y}$ is an n-vector and $\mathbf{c}(\mathbf{x})$ is a vector of n functions, each of which depends on the m-vector $\mathbf{x}$:

$$\mathbf{c}(\mathbf{x}) = \begin{bmatrix} c_1(\mathbf{x}) \\ c_2(\mathbf{x}) \\ \vdots \\ c_n(\mathbf{x}) \end{bmatrix}$$

Then the Jacobian representation of the derivatives is an $n \times m$ matrix

$$\mathcal{J}\mathbf{c}(\mathbf{x}) = \begin{bmatrix} \partial c_1/\partial x_1 & \partial c_1/\partial x_2 & \cdots & \partial c_1/\partial x_m \\ \partial c_2/\partial x_1 & \partial c_2/\partial x_2 & \cdots & \partial c_2/\partial x_m \\ \vdots & \vdots & \vdots & \vdots \\ \partial c_n/\partial x_1 & \partial c_n/\partial x_2 & \cdots & \partial c_n/\partial x_m \end{bmatrix}$$

In the alternative gradient convention, the result would be expressed as an $m \times n$ matrix whose column vectors are the various gradient vectors:

$$\nabla \mathbf{c}(\mathbf{x}) = \begin{bmatrix} \partial c_1/\partial x_1 & \partial c_2/\partial x_1 & \cdots & \partial c_n/\partial x_1 \\ \partial c_1/\partial x_2 & \partial c_2/\partial x_2 & \cdots & \partial c_n/\partial x_2 \\ \vdots & \vdots & \vdots & \vdots \\ \partial c_1/\partial x_m & \partial c_2/\partial x_m & \cdots & \partial c_n/\partial x_m \end{bmatrix} = \begin{bmatrix} \nabla c_1(\mathbf{x}) & \nabla c_2(\mathbf{x}) & \cdots & \nabla c_n(\mathbf{x}) \end{bmatrix}_{m \times n}$$

Clearly, these are not much more than notational differences, one being the transpose of the other. Even so, when you read the literature, be on the alert for a different representations of the derivatives of vectors with respect to vector arguments. Otherwise you are likely to get hopelessly lost.

A.1 Vector–vector cases

Letting $\mathbf{a}$ and $\mathbf{x}$ be column vectors of the same length m, consider the scalar function $y = \mathbf{a}'\mathbf{x}$. Because

$$y = \mathbf{a}'\mathbf{x} = \sum_{i=1}^{m} a_i x_i,$$

we have $\partial y/\partial x_i = a_i$. Collecting all such derivatives in an $m \times 1$ column vector, we obtain the gradient

$$\nabla y(\mathbf{x}) = \mathbf{a}_{m \times 1}$$

and for the Jacobian form,

$$\mathcal{J} y(\mathbf{x}) = \mathbf{a}_{1 \times m}.$$

The same results apply—as of course they must, since y is a scalar!—when we differentiate $y = \mathbf{x}'\mathbf{a}$ with respect to $\mathbf{x}$.

Now suppose that in $y = \mathbf{a}'\mathbf{x}$ and the m-vector is $\mathbf{x} = \mathbf{x}(\boldsymbol{\theta})$, with $\boldsymbol{\theta}$ being a vector of k parameters,

$$y(\boldsymbol{\theta}) = \sum_{i=1}^{m} a_i x_i(\boldsymbol{\theta}).$$

We could write the derivative as a scalar-weighted sum of the gradient vectors, producing a $k \times 1$ vector,

$$\nabla y(\boldsymbol{\theta}) = \sum_{i=1}^{m} a_i \nabla x_i(\boldsymbol{\theta}) = \begin{bmatrix} \nabla x_1(\boldsymbol{\theta}) & \nabla x_2(\boldsymbol{\theta}) & \cdots & \nabla x_m(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix}$$

or write it in the $1 \times k$ Jacobian form

$$\mathcal{J} y(\boldsymbol{\theta}) = \begin{bmatrix} a_1 & a_2 & \cdots & a_m \end{bmatrix} \begin{bmatrix} \nabla x_1(\boldsymbol{\theta})' \\ \nabla x_2(\boldsymbol{\theta})' \\ \vdots \\ \nabla x_m(\boldsymbol{\theta})' \end{bmatrix}.$$

Finally, for both $\mathbf{a} = \mathbf{a}(\boldsymbol{\theta})$ and $\mathbf{x} = \mathbf{x}(\boldsymbol{\theta})$, we have

$$y = \mathbf{a}(\boldsymbol{\theta})'\mathbf{x}(\boldsymbol{\theta}) = \sum_{i=1}^{m} a_i(\boldsymbol{\theta}) x_i(\boldsymbol{\theta}),$$

and thereby obtain the $k \times 1$ gradient result,

$$\nabla y(\boldsymbol{\theta}) = \begin{bmatrix} \nabla x_1(\boldsymbol{\theta}) & \nabla x_2(\boldsymbol{\theta}) & \cdots & \nabla x_m(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} a_1(\boldsymbol{\theta}) \\ a_2(\boldsymbol{\theta}) \\ \vdots \\ a_m(\boldsymbol{\theta}) \end{bmatrix} + \begin{bmatrix} \nabla a_1(\boldsymbol{\theta}) & \nabla a_2(\boldsymbol{\theta}) & \cdots & \nabla a_m(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} x_1(\boldsymbol{\theta}) \\ x_2(\boldsymbol{\theta}) \\ \vdots \\ x_m(\boldsymbol{\theta}) \end{bmatrix}$$

Take the transpose to obtain the $1 \times k$ Jacobian version.

A.2 Matrix–vector cases

Now consider $\mathbf{y} = \mathbf{A}\mathbf{x}$, in which $\mathbf{y}$ is an $n \times 1$ column vector and $\mathbf{A}$ is a matrix of dimension $n \times m$, with $\mathbf{x}$ again being an m-vector. Then

$$\mathbf{A}\mathbf{x} = \begin{bmatrix} \mathbf{A}_{1.} \\ \mathbf{A}_{2.} \\ \vdots \\ \mathbf{A}_{n.} \end{bmatrix} \mathbf{x} = \begin{bmatrix} \sum_{j=1}^{m} a_{1j}\, x_j \\ \sum_{j=1}^{m} a_{2j}\, x_j \\ \vdots \\ \sum_{j=1}^{m} a_{nj}\, x_j \end{bmatrix}$$

in which $\mathbf{A}_{i.}$ is the i-th *row* of the $\mathbf{A}$ matrix. Hence,

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{m} a_{1j}\, x_j \\ \sum_{j=1}^{m} a_{2j}\, x_j \\ \vdots \\ \sum_{j=1}^{m} a_{nj}\, x_j \end{bmatrix}.$$

Focus on the y_1 component of $\mathbf{y}$, and organize its derivatives as a row vector in the Jacobian fashion,

$$\partial y_1/\partial \mathbf{x} = \begin{bmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 & \cdots & \partial y_1/\partial x_m \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \end{bmatrix} = \mathbf{A}_{1.}$$

The same pattern holds for all y_k components of $\mathbf{y}$, and therefore

$$\mathcal{J}\mathbf{y}(\mathbf{x}) = \begin{bmatrix} \mathbf{A}_{1.} \\ \mathbf{A}_{2.} \\ \vdots \\ \mathbf{A}_{n.} \end{bmatrix} = \mathbf{A}.$$

Note that the result is an $n \times m$ matrix. If we wanted instead to express things in the gradient-like form (which here will be an $m \times n$ matrix), we would simply take the transpose of the Jacobian-form result, giving

$$\nabla \mathbf{y}(\mathbf{x}) = \mathbf{A}' = \begin{bmatrix} \nabla y_1(\mathbf{x}) & \nabla y_2(\mathbf{x}) & \cdots \nabla y_n(\mathbf{x}) \end{bmatrix}.$$

To follow proofs in your econometric textbooks, you'll obviously need to know which form of result the author prefers. To learn this, you will have to follow closely the dimensions of the book's matrices and vectors.

A.3 Quadratic forms

Consider the scalar function $y = \mathbf{x}'\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is a square $m \times m$ matrix and $\mathbf{x}$ is an $m \times 1$ vector. To take derivatives with respect to the elements of $\mathbf{x}$, it is easiest to begin by expressing $y = \mathbf{x}'\mathbf{A}\mathbf{x}$ in summation form,

$$y = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{m}\sum_{j=1}^{m} x_i a_{ij} x_j = \sum_{i=1}^{m} x_i \sum_{j=1}^{m} a_{ij} x_j = \sum_{i=1}^{m} x_i \, \mathbf{A}_{i.}\mathbf{x}$$

in which $\mathbf{A}_{i.}$ is the i-th *row* of the $\mathbf{A}$ matrix. Now examine the components involving x_1, the first element of the $\mathbf{x}$ vector,

$$x_1 \, \mathbf{A}_{1.}\mathbf{x} + x_2 \, \mathbf{A}_{2.}\mathbf{x} + \cdots + x_m \, \mathbf{A}_{m.}\mathbf{x}.$$

Differentiating the first term in this sum with respect to x_1 yields $\mathbf{A}_{1.}\mathbf{x} + x_1 a_{11}$, and as for the remaining terms, we get $\sum_{i=2}^{m} x_i a_{i1}$. Putting both parts of the derivative together,

$$\partial y / \partial x_1 = \mathbf{A}_{1.}\mathbf{x} + \sum_{i=1}^{m} x_i \, a_{i1}.$$

Note that the sum on the right proceeds down the rows of the first column of the $\mathbf{A}$ matrix. Had we instead chosen to differentiate with respect to x_k, the same general pattern would have emerged,

$$\partial y / \partial x_k = \mathbf{A}_{k.}\mathbf{x} + \sum_{i=1}^{m} x_i \, a_{ik},$$

although in this case the summation proceeds down the k-th column of $\mathbf{A}$. Stacking the results obtained so far, the gradient can be expressed as

$$\nabla y(\mathbf{x}) = \begin{bmatrix} \partial y/\partial x_1 \\ \partial y/\partial x_2 \\ \vdots \\ \partial y/\partial x_m \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1.}\mathbf{x} \\ \mathbf{A}_{2.}\mathbf{x} \\ \vdots \\ \mathbf{A}_{m.}\mathbf{x} \end{bmatrix} + \begin{bmatrix} \sum_{i=1}^{m} x_i \, a_{i1} \\ \sum_{i=1}^{m} x_i \, a_{i2} \\ \vdots \\ \sum_{i=1}^{m} x_i \, a_{im} \end{bmatrix} = \mathbf{A}\mathbf{x} + \begin{bmatrix} \sum_{i=1}^{m} x_i \, a_{i1} \\ \sum_{i=1}^{m} x_i \, a_{i2} \\ \vdots \\ \sum_{i=1}^{m} x_i \, a_{im} \end{bmatrix}.$$

Recognizing that the k-th *column* of $\mathbf{A}$ is the same as the k-th *row* of its transpose $\mathbf{A}'$, we can represent the vector of summations on the far right even more compactly as $\mathbf{A}'\mathbf{x}$. This simplification delivers the final result

$$\nabla y(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{A}'\mathbf{x} = (\mathbf{A} + \mathbf{A}')\,\mathbf{x}.$$

If $\mathbf{A}$ happens to be a *symmetric* matrix, implying $\mathbf{A} = \mathbf{A}'$, the result is

$$\nabla y(\mathbf{x}) = 2\mathbf{A}\mathbf{x}.$$

The symmetric $\mathbf{A}$ case is worth some additional thought. Since y is a scalar and thus equals its transpose, we know that $y = \mathbf{x}'\mathbf{A}\mathbf{x} = \mathbf{x}'\mathbf{A}'\mathbf{x}$ and therefore for any square $\mathbf{A}$ matrix, whether symmetric or not,

$$y = \mathbf{x}' \left(\frac{1}{2}(\mathbf{A} + \mathbf{A}') \right) \mathbf{x} = \mathbf{x}' \, \tilde{\mathbf{A}} \, \mathbf{x}$$

in which $\tilde{\mathbf{A}}$ is symmetric.

A.4 GMM-type quadratic forms

In nonlinear problems—such as nonlinear least squares (NLS) and the generalized method of moments (GMM)—we are faced with the task of minimizing quadratic forms such as

$$y = \mathbf{x}(\boldsymbol{\theta})' \, \mathbf{A} \, \mathbf{x}(\boldsymbol{\theta}) = \sum_{i=1}^{m} \sum_{j=1}^{m} x_i(\boldsymbol{\theta}) \, a_{ij} \, x_j(\boldsymbol{\theta}),$$

in which $\mathbf{x}(\boldsymbol{\theta})$ is an $m \times 1$ vector of functions of $\boldsymbol{\theta}$ and $\boldsymbol{\theta}$ itself is a k-vector of parameters. We assume that $\mathbf{A}$ is an $m \times m$ symmetric matrix whose elements are not functions of $\boldsymbol{\theta}$.

Since y is a scalar function of $\boldsymbol{\theta}$, we know that ultimately, the derivatives of y with respect to $\boldsymbol{\theta}$ must be organized as a $k \times 1$ column vector if we take the gradient approach, and as a $1 \times k$ row vector if we follow the Jacobian approach. But how do we get there? To differentiate y with respect to $\boldsymbol{\theta}$, we'll need to examine the derivative vector in some detail. In gradient form it is

$$\begin{aligned} \nabla y(\boldsymbol{\theta}) &= \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\nabla x_i(\boldsymbol{\theta}) \, a_{ij} x_j(\boldsymbol{\theta}) + x_i(\boldsymbol{\theta}) a_{ij} \, \nabla x_j(\boldsymbol{\theta}) \right) \\ &= \sum_{i=1}^{m} \nabla x_i(\boldsymbol{\theta}) \sum_{j=1}^{m} a_{ij} x_j(\boldsymbol{\theta}) + \sum_{j=1}^{m} \nabla x_j(\boldsymbol{\theta}) \sum_{i=1}^{m} a_{ij} x_i(\boldsymbol{\theta}) \end{aligned}$$

in which $\nabla x_i(\boldsymbol{\theta})$ is a $k \times 1$ column vector.

Examining the first of the two parts of the sum, we see that it can be written as

$$\sum_{i=1}^{m} \nabla x_i(\boldsymbol{\theta}) \sum_{j=1}^{m} a_{ij} x_j(\boldsymbol{\theta}) = \begin{bmatrix} \nabla x_1(\boldsymbol{\theta}) & \nabla x_2(\boldsymbol{\theta}) & \cdots & \nabla x_m(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} \mathbf{A}_{1.} \\ \mathbf{A}_{2.} \\ \vdots \\ \mathbf{A}_{m.} \end{bmatrix} \mathbf{x}(\boldsymbol{\theta})$$

in which $\mathbf{A}_{i.}$ is again the i-th row of the $\mathbf{A}$ matrix. Expressed more compactly, this is

$$\begin{bmatrix} \nabla x_1(\boldsymbol{\theta}) & \nabla x_2(\boldsymbol{\theta}) & \cdots & \nabla x_m(\boldsymbol{\theta}) \end{bmatrix} \mathbf{A} \, \mathbf{x}(\boldsymbol{\theta}).$$

To remind you, the matrix of gradients is of dimension $k \times m$, the matrix $\mathbf{A}$ is $m \times m$, and $\mathbf{x}(\boldsymbol{\theta})$ is an $m \times 1$ column vector. Hence this part of $\nabla y(\boldsymbol{\theta})$ is of dimension $k \times 1$ just as you would have expected.

The second of the two parts of the sum can be treated in the same way, once we use the symmetry of $\mathbf{A}$ to substitute a_{ji} in place of a_{ij}, giving

$$\sum_{j=1}^{m} \nabla x_j(\boldsymbol{\theta}) \sum_{i=1}^{m} a_{ji} x_i(\boldsymbol{\theta}) = \begin{bmatrix} \nabla x_1(\boldsymbol{\theta}) & \nabla x_2(\boldsymbol{\theta}) & \cdots & \nabla x_m(\boldsymbol{\theta}) \end{bmatrix} \begin{bmatrix} \mathbf{A}_{1.} \\ \mathbf{A}_{2.} \\ \vdots \\ \mathbf{A}_{m.} \end{bmatrix} \mathbf{x}(\boldsymbol{\theta})$$

and exactly as with the first part, this is

$$\begin{bmatrix} \nabla x_1(\boldsymbol{\theta}) & \nabla x_2(\boldsymbol{\theta}) & \cdots & \nabla x_m(\boldsymbol{\theta}) \end{bmatrix} \mathbf{A}\mathbf{x}(\boldsymbol{\theta}).$$

Since the two parts of the overall sum are identical, we have arrived at the result:

$$\nabla y(\boldsymbol{\theta}) = 2 \begin{bmatrix} \nabla x_1(\boldsymbol{\theta}) & \nabla x_2(\boldsymbol{\theta}) & \cdots & \nabla x_m(\boldsymbol{\theta}) \end{bmatrix} \mathbf{A}\mathbf{x}(\boldsymbol{\theta}).$$

This is the $k \times 1$ gradient-vector form in which the first-order conditions for nonlinear least squares and GMM estimators will typically appear. In careful treatments of the GMM first-order conditions $\nabla y(\hat{\boldsymbol{\theta}}_{GMM}) = \mathbf{0}_k$, the 2 is eliminated from the scene by re-specifying the y function as $\tilde{y} = (1/2)y$.

A.5 Miscellaneous

Here are some miscellaneous results that are occasionally of use. Sometimes we want to differentiate the quadratic form $\mathbf{x}'\mathbf{A}\mathbf{x}$ with respect to the elements of $\mathbf{A}$. The expression is

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{A}} = \mathbf{x}\mathbf{x}',$$

as can also be seen by considering the double summation. As McFadden (2000) notes, occasionally we need to consider partial derivatives of expressions $\mathbf{x}'\mathbf{A}\mathbf{y}$ with respect to the elements of $\mathbf{A}$. The result in this case is $\mathbf{x}\mathbf{y}'$, an outer product (the result holds for $\mathbf{x}$ being of dimension $n \times 1$, $\mathbf{y}$ $m \times 1$ and $\mathbf{A}$ $n \times m$). If $\mathbf{A}$ is square and nonsingular, then the partial derivatives of $\mathbf{x}'\mathbf{A}^{-1}\mathbf{y}$ with respect to the elements of $\mathbf{A}$ is $-\mathbf{A}^{-1}\mathbf{x}\mathbf{y}'\mathbf{A}^{-1}$. The partial derivative of trace $\mathbf{A}$ with respect to the elements of $\mathbf{A}$ is $\mathbf{I}$ (obviously, in this instance $\mathbf{A}$ must be a square matrix). Another useful result involves the derivative of the (natural) log of the determinant of $\mathbf{A}$,

$$\frac{\partial \ln |\mathbf{A}|}{\partial \mathbf{A}} = \mathbf{A}^{(-1)'},$$

for $|\ \mathbf{A}\ | > 0$, an expression that appears in the first-order conditions for multivariate normal likelihood functions.

Bibliography

[1] Karim M. Abadir and Jan R. Magnus. *Matrix Algebra*. Cambridge University Press, 2005.

[2] John M. Abowd, Francis Kramarz, and David N. Margolis. "High Wage Workers and High Wage Firms". *Econometrica* 67.2 (1999), pp. 251–333. DOI: https://doi.org/10.1111/1468-0262.00020. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1468-0262.00020. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/1468-0262.00020.

[3] Takeshi Amemiya. "A Note on a Heteroskedastic Model". *Journal of Econometrics* 6 (1977), pp. 365–370.

[4] Takeshi Amemiya. *Advanced Econometrics*. Cambridge, Massachusetts: Harvard University Press, 1985.

[5] E. Anderson et al. *LAPACK Users' Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1999.

[6] T. W. Anderson and C. Hsiao. "Estimation of Dynamic Models with Error Components". *Journal of the American Statistical Association* 76 (1981), pp. 598–606.

[7] Joshua D. Angrist and Jörn-Steffen Pischke. *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press, 2009.

[8] Luc Anselin. *Spatial Econometrics: Methods and Models*. Dordrecht, The Netherlands: Kluwer Academic Publishers, 1988.

[9] Luc Anselin and Anil K. Bera. "Spatial Dependence in Linear Regression Models with an Introduction to Spatial Econometrics". *Handbook of Applied Economic Statistics*. Ed. by Aman Ullah and David E. A. Giles. New York: Marcel Dekker, 1998, pp. 237–289.

[10] M. Arellano. "Computing Robust Standard Errors for Within-Groups Estimators". *Oxford Bulletin of Economics and Statistics* 49.4 (1987), pp. 431–434.

[11] M. Arellano and S. Bond. "Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations". *Review of Economic Studies* 58 (1991), pp. 277–298.

[12] Badi H. Baltagi. *Econometric Analysis of Panel Data*. Third. Chichester, UK: John Wiley & Sons, 2005.

[13] V. A. Barker et al. *LAPACK95 Users' Guide*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2001.

[14] Ronald Barry and R. Kelley Pace. "A Monte Carlo Estimator of the Log Determinant of Large Sparse Matrices". *Linear Algebra and Its Applications* 289.1–3 (1999), pp. 41–54.

[15] D. Belsley, E. Kuh, and R. Welsch. *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: John Wiley, 1980.

[16] Laurent Bergé, Sebastian Krantz, and Grand McDermott. Description of R package `fixest`, published on the Comprehensive R Archive Network (CRAN). 2021.

[17] P. Bickel and K. Doksum. *Mathematical Statistics: Basic Ideas and Selected Topics*. San Francisco: Holden-Day, 1977.

[18] Herman J. Bierens. *Introduction to the Mathematical and Statistical Foundations of Econometrics*. Cambridge, UK: Cambridge University Press, 2004.

[19] R. Blundell and S. Bond. "Initial Conditions and Moment Restrictions in Dynamic Panel Data Models". *Journal of Econometrics* 87 (1998), pp. 115–143.

[20] Colin R. Blyth. "Minimizing the Sum of Absolute Deviations". *The American Statistician* 44.4 (1990), p. 329.

[21] Kenneth Bollen. *Structural Equations with Latent Variables*. New York: John Wiley & Sons, 1989.

[22] Roger J. Bowden and Darrell A. Turkington. *Instrumental Variables*. Cambridge: Cambridge University Press, 1984.

[23] T. Breusch and A. Pagan. "The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics". *Review of Economic Studies* 47 (1980), pp. 239–53.

[24] Moshe Buchinsky. "Recent Advances in Quantile Regression Models: A Practical Guideline for Empirical Research". *Journal of Human Resources* 33 (1998), pp. 88–126.

[25] P. Burridge. "On the Cliff-Ord Test for Spatial Correlation". *Journal of the Royal Statistical Society, Series B (Methodological)* 42.1 (1980), pp. 107–108.

[26] A. Colin Cameron and Pravin K. Trivedi. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press, 2005.

[27] Sergio Correia. *A Feasible Estimator for Linear Models with Multi-Way Fixed Effects*. Working paper, Duke University. 2016.

[28] Michael Creel. "Modified Hausman Tests for Inefficient Estimators". *Applied Economics* 36.21 (2004), pp. 2373–2376. DOI: 10.1080/0003684042000291281.

[29] James Davidson. *Stochastic Limit Theory: An Introduction for Econometricians*. Oxford, UK: Oxford University Press, 1994.

[30] Russell Davidson and James G. MacKinnon. *Econometric Theory and Methods*. New York: Oxford University Press, 2004.

[31] Russell Davidson and James G. MacKinnon. *Estimation and Inference in Econometrics*. New York: Oxford University Press, 1993.

[32] Angus Deaton. "Counting the World's Poor: Problems and Possible Solutions". *The World Bank Research Observer* 16.2 (2001), pp. 125–48.

[33] Angus Deaton. *The Analysis of Household Surveys: Microeconometric Analysis for Development Policy*. Washington, D.C.: Johns Hopkins University Press, 1997.

[34] Geert Dhaene and Koen Jochmans. "Split-panel Jackknife Estimation of Fixed-effect Models". *The Review of Economic Studies* 82.3 (Feb. 2015), pp. 991–1030. ISSN: 0034-6527. DOI: 10.1093/restud/rdv007. eprint: https://academic.oup.com/restud/article-pdf/82/3/991/5371725/rdv007.pdf. URL: https://doi.org/10.1093/restud/rdv007.

[35] George H. Dunteman. *Principal Components Analysis*. Vol. 69. Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage Publications, 1989.

[36] Chris Elbers, Jean O. Lanjouw, and Peter Lanjouw. "Imputed Welfare Estimates in Regression Analysis". *Journal of Economic Geography* 5.1 (2005), pp. 101–118.

[37] Chris Elbers, Jean O. Lanjouw, and Peter Lanjouw. "Micro-level Estimation of Poverty and Inequality". *Econometrica* 71 (2003), pp. 355–364.

[38] Deon Filmer and Lant Pritchett. "Estimating Wealth Effects Without Expenditure Data—Or Tears: An Application to Educational Enrollments in States of India". *Demography* 38.1 (2001), pp. 115–132.

[39] Deon Filmer and Lant Pritchett. "The Effect of Household Wealth on Educational Attainment: Evidence from 35 Countries". *Population and Development Review* 25.1 (1999), pp. 85–120.

[40] Angel de la Fuente. *Mathematical Methods and Models for Economists*. Cambridge University Press, 2000.

[41] Wayne A. Fuller. *Introduction to Statistical Time Series*. New York: John Wiley & Sons, 1976.

[42] Wayne A. Fuller. *Measurement Error Models*. New York: John Wiley, 1987.

[43] Arthur S. Goldberger. "Best Linear Unbiased Prediction in the Generalized Linear Regression Model". *Journal of the American Statistical Association* 57.298 (1962), pp. 369–375.

[44] Franklin Graybill. *An Introduction to Linear Statistical Models, Vol. 1*. New York: McGraw-Hill, 1961.

[45] E. Greenberg and C. Webster. *Advanced Econometrics:A Bridge to the Literature*. New York: John Wiley, 1983.

[46] William H. Greene. *Econometric Analysis*. Upper Saddle River, New Jersey: Prentice-Hall, 2003.

[47] William H. Greene. *Econometric Analysis*. 6th. New York: Prentice Hall, 2008.

[48] Zvi Griliches. "Economic Data Issues". *The Handbook of Econometrics*. Ed. by Zvi Griliches and Michael Intriligator. Vol. 3. Amsterdam: Elsevier Science, 1986. Chap. 25.

[49] Zvi Griliches and Jerry A. Hausman. "Errors in variables in panel data". *Journal of Econometrics* 31.1 (1986), pp. 93–118. ISSN: 0304-4076. DOI: https://doi.org/10.1016/0304-4076(86)90058-8. URL: https://www.sciencedirect.com/science/article/pii/0304407686900588.

[50] Paulo Guimarães and Pedro Portugal. "A Simple Feasible Procedure to fit Models with High-Dimensional Fixed Effects". *The Stata Journal* 10.4 (2010), pp. 628–649.

[51] Jinyong Hahn and Whitney Newey. "Jackknife and Analytical Bias Reduction for Nonlinear Panel Models". *Econometrica* 72.4 (2004), pp. 1295–1319. DOI: https://doi.org/10.1111/j.1468-0262.2004.00533.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0262.2004.00533.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2004.00533.x.

[52] Alfred Hamerle and Gerd Ronning. "Panel Analysis for Qualitative Variables". *Handbook of Statistical Modeling for the Social and Behavioral Sciences*. Ed. by Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel. New York: Plenum Press, 1995. Chap. 8, pp. 401–451.

[53] Lars Peter Hansen. "Large Sample Properties of Generalized Method of Moments Estimators". *Econometrica* 50 (1982), pp. 1029–1054.

[54] Lingxin Hao and Daniel Q. Naiman. *Quantile Regression*. Quantitative Applications in the Social Sciences Series no. 07–149. Thousand Oaks, CA: Sage Publications. 2007.

[55] Jerry A. Hausman. "Specification Tests in Econometrics". *Econometrica* 46 (1978), pp. 1251–1271.

[56] Jerry A. Hausman and William Taylor. "Panel Data and Unobservable Individual Effects". *Econometrica* 49 (1981), pp. 1377–1398.

[57] Fumio Hayashi. *Econometrics*. Princeton, NJ: Princeton University Press, 2000.

[58] James J. Heckman. "Statistical Models for Discrete Panel Data". *Structural Analysis of Discrete Data with Econometric Applications*. Ed. by Charles F. Manski and Daniel McFadden. Cambridge, MA: MIT Press, 1981, pp. 114–178.

[59] Douglas Holtz-Eakin, Whitney Newey, and Harvey S. Rosen. "Estimating Vector Autoregressions with Panel Data". *Econometrica* 56 (1988), pp. 1371–1395.

[60] C. Hsiao. *Analysis of Panel Data*. Cambridge: Cambridge University Press, 1986.

[61] Peter Jensen, Michael Rosholm, and Mette Verner. *A Comparison of Different Estimators for Panel Data Sample Selection Models*. Working Paper no. 2002-1, Department of Economics, Aarhus School of Business, University of Aarhus, Denmark. 2001.

[62] N. Johnson and S. Kotz. *Continuous Univariate Distributions—1*. New York: John Wiley, 1970.

[63] N. Johnson and S. Kotz. *Continuous Univariate Distributions—2*. New York: John Wiley, 1970.

[64] Karl G. Jöreskog. "Basic Ideas of Factor and Component Analysis". *Factor Analysis and Structural Equation Models*. Ed. by Karl G. Jöreskog and Dag Sörbom. Cambridge, MA: ABT Books, 1979, pp. 5–20.

[65] Karl G. Jöreskog. *Latent Variable Scores and Their Uses*. Unpublished paper, available at http://www.ssicentral.com/lisrel.column6.pdf. 2000.

[66] G. Judge et al. *The Theory and Practice of Econometrics*. New York: John Wiley, 1985.

[67] Harry H. Kelejian and Ingmar R. Prucha. "A Generalized Moments Estimator for the Autoregressive Parameter in a Spatial Model". *International Economic Review* 40.2 (1999), pp. 509–533.

[68] André I. Khuri. *Advanced Calculus with Applications in Statistics*. New York: John Wiley and Sons, 1993.

[69] Tony Lancaster. *The Econometric Analysis of Transition Data*. New York: Cambridge University Press, 1990.

[70] R. Larsen and M. Marx. *An Introduction to Mathematical Statistics and Its Applications*. Englewood Cliffs, New Jersey: Prentice-Hall, 1986.

[71] D. N. Lawley and A. E. Maxwell. "Factor Analysis as a Statistical Method". *Statistician* 12.3 (1962), pp. 209–229.

[72] Yoong-Sin Lee. "Graphical Demonstration of an Optimality Property of the Median". *The American Statistician* 49.4 (1995), pp. 369–372.

[73] James P. LeSage. *The Theory and Practice of Spatial Econometrics*. Documentation distributed with his Econometric Toolbox for MATLAB, Department of Economics, University of Toledo, Toledo, OH. 1999.

[74] D. Luenberger. *Optimization by Vector Space Methods*. New York: John Wiley, 1969.

[75] James G. MacKinnon. "Model Specification Tests and Artificial Regressions". *Journal of Economic Literature* 30.1 (1992), pp. 102–146.

[76] G. S. Maddala. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press, 1983.

[77] Daniel McFadden. "Econometrics 250B: Second Half Reader". Lecture notes, Department of Economics, University of California, Berkeley. 1999.

[78] Daniel McFadden. "Introduction to Statistical Analysis in Large Samples". Lecture notes, Department of Economics, MIT. 1988.

[79] Daniel McFadden. "Statistical Tools". Lecture notes, Department of Economics, University of California, Berkeley. 2000.

[80] Mario J. Miranda and Paul L. Fackler. *Applied Computational Economics and Finance*. Cambridge, MA: MIT Press, 2002.

[81] Ron Mittelhammer. *Mathematical Statistics for Economics and Business*. First. New York: Springer-Verlag, 1996.

[82] Ron C. Mittelhammer. *Mathematical Statistics for Economics and Business*. Second. New York: Springer, 2013.

[83] Ron C. Mittelhammer, George G. Judge, and Douglas J. Miller. *Econometric Foundations*. Cambridge, UK: Cambridge University Press, 2000.

[84] Mark R. Montgomery and Paul C. Hewett. "Urban Poverty and Health in Developing Countries: Household and Neighborhood Effects". *Demography* 42.3 (2005), pp. 397–425.

[85] Mark R. Montgomery et al. "Measuring Living Standards With Proxy Variables". *Demography* 37.2 (2000), pp. 155–174.

[86] John Mullahy. *Estimation of Multivariate Probit Models via Bivariate Probit*. Working Paper 21593, National Bureau of Economic Research (NBER), Cambridge MA. 2015.

[87] John Mullahy. *Marginal Effects in Multivariate Probit and Kindred Discrete and Count Outcome Models, with Applications in Health Economics*. Working Paper 17588, National Bureau of Economic Research (NBER), Cambridge MA. 2011.

[88] Yair Mundlak. "On the Pooling of Time Series and Cross Section Data". *Econometrica* 46.1 (1978), pp. 69–85.

[89] A. Nakamura and M. Nakamura. "On the Relationship Among Several Specification Error Tests Presented by Durbin, Wu, and Hausman". *Econometrica* 49.6 (1981), pp. 1583–1588.

[90] Whitney K. Newey and Daniel L. McFadden. "Large Sample Estimation and Hypothesis Testing". *The Handbook of Econometrics*. Ed. by Robert F. Engle and Daniel L. McFadden. Vol. 4. Amsterdam: Elsevier Science, 1994. Chap. 36.

[91] Dianne P. O'Leary. "Iterative Methods for Linear Systems: Following the Meandering Way". *Computing in Science and Engineering* July/August (2006), pp. 74–78.

[92] Dianne P. O'Leary. "Solving Sparse Linear Systems: Taking the Direct Approach". *Computing in Science and Engineering* September/October (2005), pp. 62–67.

[93] Keith Ord. "Estimation Methods for Models of Spatial Interaction". *Journal of the American Statistical Association* 70.349 (1975), pp. 120–126.

[94] R. Kelley Pace. *SPACESTATPACK: A Spatial Statistics Package in Fortran 90 1.0 (with PC executable code)*. Department of Finance, Louisiana State University, Baton Rouge, LA. 2000.

[95] R. Kelley Pace. *Spatial Statistics Toolbox*. Documentation distributed with his toolbox for MATLAB, Department of Finance, Louisiana State University, Baton Rouge, LA. 2000.

[96] R. Kelley Pace and James P. LeSage. "Chebyshev Approximation of Log-Determinants of Spatial Weight Matrices". *Computational Statistics and Data Analysis* 45 (2004), pp. 179–196.

[97] R. Kelley Pace and James P. LeSage. *Closed-Form Maximum Likelihood Estimates for Spatial Problems*. Department of Finance, Louisiana State University, Baton Rouge, LA. 2000.

[98] Adrian Pagan and Robert Hall. "Diagnostic Tests as Residual Analysis". *Econometric Reviews* 2.2 (1983), pp. 159–218.

[99] Adrian Pagan and Frank Vella. "Diagnostic Tests for Models Based on Individual Data: A Survey". *Journal of Applied Econometrics* 4 (1989), S29–S59.

[100] S. Pollock. *The Algebra of Econometrics*. New York: John Wiley, 1979.

[101] William H. Press et al. *Numerical Recipes in Fortran 90: The Art of Parallel Scientific Computing*. Vol. 2. New York: Cambridge University Press, 1996.

[102] William H. Press et al. *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Second. New York: Cambridge University Press, 1992.

[103] R. Ramanathan. *Statistical Methods in Econometrics*. San Diego: Academic Press, 1993.

[104] Douglas Rivers and Quang H. Vuong. "Limited Information Estimators and Exogeneity Tests for Simulataneous Probit Models". *Journal of Econometrics* 39 (1988), pp. 347–366.

[105] Peter Robinson. "Root-N-Consistent Semiparametric Regression". *Econometrica* 56.4 (1988), pp. 931–54. URL: https://EconPapers.repec.org/RePEc:ecm:emetrp:v:56:y:1988:i:4:p:931-54.

[106] David Roodman. *How to Do `xtabond2`: An Introduction to "Difference" and "System" GMM in* STATA. Working Paper No. 103, Center for Global Development, Washington DC. 2006.

[107] Paul R. Rosenbaum and Donald B. Rubin. "The Central Role of the Propensity Score in Observational Studies for Causal Effects". *Biometrika* 70 (1983), pp. 41–55.

[108] John Rust. "Optimal Replacement of GMC Bus Engines: An Empirical Model of Harold Zurcher". *Econometrica* 55.5 (1987), pp. 999–1033.

[109] Paul A. Ruud. *An Introduction to Classical Econometric Theory*. New York: Oxford University Press, 2000.

[110] David E. Sahn and David C. Stifel. "Poverty Comparisons Over Time and Across Countries in Africa". *World Development* 28.12 (2000), pp. 2123–2155.

[111] James R. Schott. *Matrix Analysis for Statistics*. New York: John Wiley and Sons, 1997.

[112] Douglas A. Schroeder. *Accounting and Causal Effects: Econometric Challenges*. Springer, 2010.

[113] G. Seber. *The Linear Hypothesis: A General Theory*. New York: MacMillan, 1980.

[114] R. Serfling. *Approximation Theorems of Mathematical Statistics*. New York: John Wiley, 1980.

[115] Christopher L. Skeels and Francis Vella. *Monte Carlo Evidence on the Robustness of Conditional Moment Tests in Tobit and Probit Models*. Manuscript, Department of Statistics, Australian National University. 1995.

[116] Christopher L. Skeels and Francis Vella. "The Effect of Estimator Behavior on Conditional Moment Tests in Probit and Tobit Models". *Pakistan Journal of Statistics* 10.2 (1994), pp. 487–516.

[117] Christopher L. Skeels and Francis Vella. *The Performance of Conditional Moment Tests in Tobit and Probit Models*. Manuscript, Department of Statistics, Australian National University. 1994.

[118] Oleg Smirnov and Luc Anselin. "Fast Maximum Likelihood Estimation of Very Large Spatial Autoregressive Models: A Characteristic Polynomial Approach". *Computational Statistics & Data Analysis* 35 (2001), pp. 301–319.

[119] Richard J. Smith and Richard W. Blundell. "An Exogeneity Test for a Simultaneous Equation Tobit Model with an Application to Labor Supply". *Econometrica* 54.3 (1986), pp. 679–685.

[120] A. Spanos. *Statistical Foundations of Econometric Modelling*. New York: Cambridge University Press, 1986.

[121] Lester D. Taylor. *Probability and Mathematical Statistics*. New York: Harper and Row, 1974.

[122] Frank Vella and Marno Verbeek. "Two-step estimation of panel data models with censored endogenous variables and selection bias". *Journal of Econometrics* 90 (1999), pp. 239–263.

[123] Halbert White. "A Heteroskedasticity-consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity". *Econometrica* 48 (1980), pp. 817–838.

[124] Michael R. Wickens. "A Note on the Use of Proxy Variables". *Econometrica* 40.4 (1972), pp. 759–761.

[125] Michael Woodroofe. *Probability With Applications*. New York: McGraw-Hill, 1975.

[126] Jeffrey M. Wooldridge. "Control Function Methods in Applied Econometrics". *The Journal of Human Resources* 50.2 (2015), pp. 420–445. ISSN: 0022166X. URL: http://www.jstor.org.proxy.library.stonybrook.edu/stable/24735991.

[127] Jeffrey M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. Second. Cambridge, MA: The MIT Press, 2010.

[128] Adonis Yatchew and Zvi Griliches. "Specification Error in Probit Models". *Review of Economics and Statistics* 66 (1984), pp. 134–139.

[129] A. Zaman. *Statistical Foundations for Econometric Techniques*. San Diego: Academic Press, 1996.