MINIMUM DISTANCE ESTIMATION AND THE GENERALIZED METHOD OF MOMENTS

13.1 INTRODUCTION

The **maximum likelihood estimator** presented in Chapter 14 is fully efficient among consistent and asymptotically normally distributed estimators, *in the context of the specified parametric model.* The possible shortcoming in this result is that to attain that efficiency, it is necessary to make possibly strong, restrictive assumptions about the distribution, or data-generating process. The generalized method of moments (GMM) estimators discussed in this chapter move away from parametric assumptions, toward estimators that are robust to some variations in the underlying data-generating process.

This chapter will present a number of fairly general results on parameter estimation. We begin with perhaps the oldest formalized theory of estimation, the classical theory of the method of moments. This body of results dates to the pioneering work of Fisher (1925). The use of sample moments as the building blocks of estimating equations is fundamental in econometrics. GMM is an extension of this technique that, as will be clear shortly, encompasses nearly all the familiar estimators discussed in this book. Section 13.2 will introduce the estimation framework with the method of moments. The technique of minimum distance estimation is developed in Section 13.3. Formalities of the GMM estimator are presented in Section 13.4. Section 13.5 discusses hypothesis testing based on moment equations. Major applications, including dynamic panel data models, are described in Section 13.6.

Example 13.1 Euler Equations and Life Cycle Consumption

One of the most often-cited applications of the GMM principle for estimating econometric models is Hall's (1978) permanent income model of consumption. The original form of the model (with some small changes in notation) posits a hypothesis about the optimizing behavior of a consumer over the life cycle. Consumers are hypothesized to act according to the model:

$$\text{Maximize } E_t\left[\sum_{\tau=0}^{T-t}\left(\frac{1}{1+\delta}\right)^{\tau} U(c_{t+\tau})\,|\,\Omega_t\right] \text{ subject to } \sum_{\tau=0}^{T-t}\left(\frac{1}{1+r}\right)^{\tau}(c_{t+\tau}-w_{t+\tau}) = A_t.$$

The information available at time t is denoted Ω_t so that E_t denotes the expectation formed at time t based on the information set Ω_t. The maximand is the expected discounted stream of future utility from consumption from time t until the end of life at time T. The individual's subjective rate of time preference is $\beta = 1/(1+\delta)$. The real rate of interest, $r \geq \delta$ is assumed to be constant. The utility function $U(c_t)$ is assumed to be strictly concave and time separable (as shown in the model). One period's consumption is c_t. The intertemporal budget constraint states that the present discounted excess of c_t over earnings, w_t, over the lifetime equals

total assets A_t not including human capital. In this model, it is claimed that the only source of uncertainty is w_t. No assumption is made about the stochastic properties of w_t except that there exists an expected future earnings, $E_t[w_{t+r} \mid \Omega_t]$. Successive values are not assumed to be independent and w_t is not assumed to be stationary.

Hall's major "theorem" in the paper is the solution to the optimization problem, which states

$$E_t[U'(c_{t+1}) \mid \Omega_t] = \frac{1+\delta}{1+r} U'(c_t).$$

For our purposes, the major conclusion of the paper is "Corollary 1" which states "No information available in time t apart from the level of consumption, c_t, helps predict future consumption, c_{t+1}, in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods t or earlier are irrelevant once c_t is known." We can use this as the basis of a model that can be placed in the GMM framework. To proceed, it is necessary to assume a form of the utility function. A common (convenient) form of the utility function is $U(c_t) = c_t^{1-\alpha}/(1-\alpha)$, which is monotonic, $U' = c_t^{-\alpha} > 0$ and concave, $U''/U' = -\alpha/c_t < 0$. Inserting this form into the solution, rearranging the terms, and reparameterizing it for convenience, we have

$$E_t\left[(1+r)\left(\frac{1}{1+\delta}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\alpha} - 1 \mid \Omega_t\right] = E_t\left[\beta(1+r)R_{t+1}^{\lambda} - 1 \mid \Omega_t\right] = 0,$$

where $R_{t+1} = c_{t+1}/c_t$ and $\lambda = -\alpha$.

Hall assumed that r was constant over time. Other applications of this modeling framework [for example, Hansen and Singleton (1982)] have modified the framework so as to involve a forecasted interest rate, r_{t+1}. How one proceeds from here depends on what is in the information set. The unconditional mean does not identify the two parameters. The corollary states that the only relevant information in the information set is c_t. Given the form of the model, the more natural instrument might be R_t. This assumption exactly identifies the two parameters in the model:

$$E_t\left[\left(\beta(1+r_{t+1})R_{t+1}^{\lambda} - 1\right)\begin{pmatrix}1\\R_t\end{pmatrix}\right] = \begin{bmatrix}0\\0\end{bmatrix}.$$

As stated, the model has no testable implications. These two moment equations would exactly identify the two unknown parameters. Hall hypothesized several models involving income and consumption which would overidentify and thus place restrictions on the model.

15.2 CONSISTENT ESTIMATION: THE METHOD OF MOMENTS

Sample statistics such as the mean and variance can be treated as simple descriptive measures. In our discussion of estimation in Appendix C, however, we argue that, in general, sample statistics each have a counterpart in the population, for example, the correspondence between the sample mean and the population expected value. The natural (perhaps obvious) next step in the analysis is to use this analogy to justify using the sample "moments" as estimators of these population parameters. What remains to establish is whether this approach is the best, or even a good way to use the sample data to infer the characteristics of the population.

The basis of the **method of moments** is as follows: In random sampling, under generally benign assumptions, a sample statistic will converge in probability to some constant. For example, with i.i.d. random sampling, $\overline{m}_2' = (1/n)\sum_{i=1}^{n} y_i^2$ will converge in

mean square to the variance plus the square of the mean of the random variable, y_i. This constant will, in turn, be a function of the unknown parameters of the distribution. To estimate K parameters, $\theta_1, \ldots, \theta_K$, we can compute K such statistics, $\overline{m}_1, \ldots, \overline{m}_K$, whose **probability limits** are known functions of the parameters. These K moments are equated to the K functions, and the functions are inverted to express the parameters as functions of the moments. The moments will be consistent by virtue of a law of large numbers (Theorems D.4–D.9). They will be asymptotically normally distributed by virtue of the Lindeberg–Levy **Central Limit theorem** (D.18). The derived parameter estimators will inherit consistency by virtue of the Slutsky theorem (D.12) and asymptotic normality by virtue of the delta method (Theorem D.21).

This section will develop this technique in some detail, partly to present it in its own right and partly as a prelude to the discussion of the generalized method of moments, or GMM, estimation technique, which is treated in Section 15.4.

13.2.1 RANDOM SAMPLING AND ESTIMATING THE PARAMETERS OF DISTRIBUTIONS

Consider independent, identically distributed random sampling from a distribution $f(y \mid \theta_1, \ldots, \theta_K)$ with finite moments up to $E[y^{2K}]$. The **random sample** consists of n observations, $y_1, \ldots, y_n$. The kth "raw" or **uncentered moment** is

$$\overline{m}'_k = \frac{1}{n}\sum_{i=1}^{n} y_i^k.$$

By Theorem D.4,

$$E[\overline{m}'_k] = \mu'_k = E[y_i^k],$$

and

$$\text{Var}[\overline{m}'_k] = \frac{1}{n}\text{Var}[y_i^k] = \frac{1}{n}(\mu'_{2k} - \mu'^2_k).$$

By convention, $\mu'_1 = E[y_i] = \mu$. By the Khinchine theorem, D.5,

$$\text{plim}\, \overline{m}'_k = \mu'_k = E[y_i^k].$$

Finally, by the Lindeberg–Levy Central Limit theorem,

$$\sqrt{n}(\overline{m}'_k - \mu'_k) \xrightarrow{d} N[0,\ \mu'_{2k} - \mu'^2_k].$$

In general, μ'_k will be a function of the underlying parameters. By computing K raw moments and equating them to these functions, we obtain K equations that can (in principle) be solved to provide estimates of the K unknown parameters.

Example 13.2 Method of Moments Estimator for $N[\mu, \sigma^2]$

In random sampling from $N[\mu, \sigma^2]$,

$$\text{plim}\, \frac{1}{n}\sum_{i=1}^{n} y_i = \text{plim}\, \overline{m}'_1 = E[y_i] = \mu,$$

and

$$\text{plim}\,\frac{1}{n}\sum_{i=1}^{n} y_i^2 = \text{plim}\,\overline{m}_2' = \text{Var}[y_i] + \mu^2 = \sigma^2 + \mu^2.$$

Equating the right- and left-hand sides of the probability limits gives moment estimators

$$\hat{\mu} = \overline{m}_1' = \overline{y},$$

and

$$\hat{\sigma}^2 = \overline{m}_2' - \overline{m}_1'^2 = \left(\frac{1}{n}\sum_{i=1}^{n} y_i^2\right) - \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

Note that $\hat{\sigma}^2$ is biased, although both estimators are consistent.

Although the moments based on powers of y provide a natural source of information about the parameters, other functions of the data may also be useful. Let $m_k(\cdot)$ be a continuous and differentiable function not involving the sample size n, and let

$$\overline{m}_k = \frac{1}{n}\sum_{i=1}^{n} m_k(y_i), \quad k = 1, 2, \ldots, K.$$

These are also "moments" of the data. It follows from Theorem D.4 and the corollary, (D-5), that

$$\text{plim}\,\overline{m}_k = E[m_k(y_i)] = \mu_k(\theta_1, \ldots, \theta_K).$$

We assume that $\mu_k(\cdot)$ involves some of or all the parameters of the distribution. With K parameters to be estimated, the K **moment equations,**

$$\begin{aligned}
\overline{m}_1 - \mu_1(\theta_1, \ldots, \theta_K) &= 0,\\
\overline{m}_2 - \mu_2(\theta_1, \ldots, \theta_K) &= 0,\\
&\cdots\\
\overline{m}_K - \mu_K(\theta_1, \ldots, \theta_K) &= 0,
\end{aligned}$$

provide K equations in K unknowns, $\theta_1, \ldots, \theta_K$. If the equations are continuous and functionally independent, then **method of moments estimators** can be obtained by solving the system of equations for

$$\hat{\theta}_k = \hat{\theta}_k[\overline{m}_1, \ldots, \overline{m}_K].$$

As suggested, there may be more than one set of moments that one can use for estimating the parameters, or there may be more moment equations available than are necessary.

Example 15.3 Inverse Gaussian (Wald) Distribution

The inverse Gaussian distribution is used to model survival times, or elapsed times from some beginning time until some kind of transition takes place. The standard form of the density for this random variable is

$$f(y) = \sqrt{\frac{\lambda}{2\pi y^3}}\exp\left[-\frac{\lambda(y-\mu)^2}{2\mu^2 y}\right], \quad y > 0, \lambda > 0, \mu > 0.$$

The mean is μ while the variance is μ^3/λ. The efficient maximum likelihood estimators of the two parameters are based on $(1/n)\sum_{i=1}^{n} y_i$ and $(1/n)\sum_{i=1}^{n}(1/y_i)$. Because the mean and

variance are simple functions of the underlying parameters, we can also use the sample mean and sample variance as moment estimators of these functions. Thus, an alternative pair of method of moments estimators for the parameters of the Wald distribution can be based on $(1/n)\sum_{i=1}^{n} y_i$ and $(1/n)\sum_{i=1}^{n} y_i^2$. The precise formulas for these two pairs of estimators is left as an exercise.

Example 15.4 Mixtures of Normal Distributions

Quandt and Ramsey (1978) analyzed the problem of estimating the parameters of a mixture of normal distributions. Suppose that each observation in a random sample is drawn from one of two different normal distributions. The probability that the observation is drawn from the first distribution, $N[\mu_1, \sigma_1^2]$, is λ, and the probability that it is drawn from the second is $(1-\lambda)$. The density for the observed y is

$$f(y) = \lambda N[\mu_1, \sigma_1^2] + (1-\lambda)N[\mu_2, \sigma_2^2], \quad 0 \le \lambda \le 1$$
$$= \frac{\lambda}{(2\pi\sigma_1^2)^{1/2}} e^{-1/2[(y-\mu_1)/\sigma_1]^2} + \frac{1-\lambda}{(2\pi\sigma_2^2)^{1/2}} e^{-1/2[(y-\mu_2)/\sigma_2]^2}.$$

insert next page (A) msp 13-6

new paragraph

The sample mean and second through fifth **central moments**,

$$m_k = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^k, \quad k = 2, 3, 4, 5,$$

provide five equations in five unknowns that can be solved (via a ninth-order polynomial) for consistent estimators of the five parameters. Because $\bar{y}$ converges in probability to $E[y_i] = \mu$, the theorems given earlier for m_k' as an estimator of μ_k' apply as well to m_k as an estimator of

$$\mu_k = E[(y_i - \mu)^k].$$

For the mixed normal distribution, the mean and variance are

$$\mu = E[y_i] = \lambda\mu_1 + (1-\lambda)\mu_2,$$

and

$$\sigma^2 = \text{Var}[y_i] = \lambda\sigma_1^2 + (1-\lambda)\sigma_2^2 + \lambda(1-\lambda)(\mu_1 - \mu_2)^2,$$

which suggests how complicated the familiar method of moments is likely to become. An alternative method of estimation proposed by the authors is based on

$$E[e^{ty}] = \lambda e^{t\mu_1 + t^2\sigma_1^2/2} + (1-\lambda)e^{t\mu_2 + t^2\sigma_2^2/2} = \Lambda_t,$$

where t is any value not necessarily an integer. Quandt and Ramsey (1978) suggest choosing five values of t that are not too close together and using the statistics

$$\bar{M}_t = \frac{1}{n}\sum_{i=1}^{n} e^{ty_i}$$

to estimate the parameters. The moment equations are $\bar{M}_t - \Lambda_t(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = 0$. They label this procedure the **method of moment generating functions.** (See Section B.6 for definition of the moment generating function.)

KT

In most cases, method of moments estimators are not efficient. The exception is in random sampling from **exponential families** of distributions.

KT

Insert on msp 13-5 where indicated

A f1 Before proceeding, we note that this density is precisely the same as the finite ~~mixture~~ mixture model described in Section 14.9.7.d. Maximum likelihood estimation of the model using the method described there would be simpler than the method of moment generating functions developed here.

end of insert

DEFINITION 15.1 Exponential Family

An exponential (parametric) family of distributions is one whose log-likelihood is of the form

$$\ln L(\theta \mid \text{data}) = a(\text{data}) + b(\theta) + \sum_{k=1}^{K} c_k(\text{data}) s_k(\theta),$$

where $a(\cdot)$, $b(\cdot)$, $c_k(\cdot)$, and $s_k(\cdot)$ are functions. The members of the "family" are distinguished by the different parameter values.

If the log-likelihood function is of this form, then the functions $c_k(\cdot)$ are called **sufficient statistics.**[1] When sufficient statistics exist, method of moments estimator(s) can be functions of them. In this case, the method of moments estimators will also be the maximum likelihood estimators, so, of course, they will be efficient, at least asymptotically. We emphasize, in this case, the probability distribution is fully specified. Because the normal distribution is an exponential family with sufficient statistics $\overline{m}'_1$ and $\overline{m}'_2$, the estimators described in Example 15.2 are fully efficient. (They are the maximum likelihood estimators.) The mixed normal distribution is not an exponential family. We leave it as an exercise to show that the Wald distribution in Example 15.3 is an exponential family. You should be able to show that the sufficient statistics are the ones that are suggested in Example 15.3 as the bases for the MLEs of μ and λ.

Example 15.5 Gamma Distribution

The gamma distribution (see Section B.4.5) is

$$f(y) = \frac{\lambda^P}{\Gamma(P)} e^{-\lambda y} y^{P-1}, \quad y \geq 0, P > 0, \lambda > 0.$$

The log-likelihood function for this distribution is

$$\frac{1}{n} \ln L = [P \ln \lambda - \ln \Gamma(P)] - \lambda \frac{1}{n} \sum_{i=1}^{n} y_i + (P-1) \frac{1}{n} \sum_{i=1}^{n} \ln y_i.$$

This function is an exponential family with $a(\text{data}) = 0$, $b(\theta) = n[P \ln \lambda - \ln \Gamma(P)]$ and two sufficient statistics, $\frac{1}{n}\sum_{i=1}^{n} y_i$ and $\frac{1}{n}\sum_{i=1}^{n} \ln y_i$. The method of moments estimators based on $\frac{1}{n}\sum_{i=1}^{n} y_i$ and $\frac{1}{n}\sum_{i=1}^{n} \ln y_i$ would be the maximum likelihood estimators. But, we also have

$$\text{plim} \frac{1}{n} \sum_{i=1}^{n} \begin{bmatrix} y_i \\ y_i^2 \\ \ln y_i \\ 1/y_i \end{bmatrix} = \begin{bmatrix} P/\lambda \\ P(P+1)/\lambda^2 \\ \Psi(P) - \ln \lambda \\ \lambda/(P-1) \end{bmatrix}.$$

(The functions $\Gamma(P)$ and $\Psi(P) = \text{d} \ln \Gamma(P)/\text{d}P$ are discussed in Section E.2.3.) Any two of these can be used to estimate λ and P.

[1] Stuart and Ord (1989, pp. 1–29) give a discussion of sufficient statistics and exponential families of distributions. A result that we will use in Chapter 23 is that if the statistics, $c_k(\text{data})$ are sufficient statistics, then the conditional density $f[y_1, \ldots, y_n \mid c_k(\text{data}), k = 1, \ldots, K]$ is not a function of the parameters.

For the income data in Example C.1, the four moments listed earlier are

$$(\overline{m}'_1, \overline{m}'_2, \overline{m}'_*, \overline{m}'_{-1}) = \frac{1}{n}\sum_{i=1}^{n}\left[y_i, y_i^2, \ln y_i, \frac{1}{y_i}\right] = [31.278, 1453.96, 3.22139, 0.050014].$$

The method of moments estimators of $\theta = (P, \lambda)$ based on the six possible pairs of these moments are as follows:

$$(\hat{P}, \hat{\lambda}) = \begin{bmatrix} & \overline{m}'_1 & \overline{m}'_2 & \overline{m}'_{-1} \\ \overline{m}'_2 & 2.05682, 0.065759 & & \\ \overline{m}'_{-1} & 2.77198, 0.0886239 & 2.60905, 0.080475 & \\ \overline{m}'_* & 2.4106, 0.0770702 & 2.26450, 0.071304 & 3.03580, 0.1018202 \end{bmatrix}.$$

The maximum likelihood estimates are $\hat{\theta}(\overline{m}'_1, \overline{m}'_*) = (2.4106, 0.0770702)$.

15.2.2 ASYMPTOTIC PROPERTIES OF THE METHOD OF MOMENTS ESTIMATOR

In a few cases, we can obtain the exact distribution of the method of moments estimator. For example, in sampling from the normal distribution, $\hat{\mu}$ has mean μ and variance σ^2/n and is normally distributed, while $\hat{\sigma}^2$ has mean $[(n-1)/n]\sigma^2$ and variance $[(n-1)/n]^2 2\sigma^4/(n-1)$ and is exactly distributed as a multiple of a chi-squared variate with $(n-1)$ degrees of freedom. If sampling is not from the normal distribution, the exact variance of the sample mean will still be $\text{Var}[y]/n$, whereas an asymptotic variance for the moment estimator of the population variance could be based on the leading term in (D-27), in Example D.10, but the precise distribution may be intractable.

There are cases in which no explicit expression is available for the variance of the underlying sample moment. For instance, in Example 15.4, the underlying sample statistic is

$$\overline{M}_t = \frac{1}{n}\sum_{i=1}^{n} e^{ty_i} = \frac{1}{n}\sum_{i=1}^{n} M_{it}.$$

The exact variance of $\overline{M}_t$ is known only if t is an integer. But if sampling is random, and if $\overline{M}_t$ is a sample mean: we can estimate its variance with $1/n$ times the sample variance of the observations on M_{it}. We can also construct an estimator of the covariance of $\overline{M}_t$ and $\overline{M}_s$:

$$\text{Est. Asy. Cov}[\overline{M}_t, \overline{M}_s] = \frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}[(e^{ty_i} - \overline{M}_t)(e^{sy_i} - \overline{M}_s)]\right\}.$$

In general, when the moments are computed as

$$\overline{m}_{n,k} = \frac{1}{n}\sum_{i=1}^{n} m_k(\mathbf{y}_i), \quad k = 1, \ldots, K,$$

where $\mathbf{y}_i$ is an observation on a vector of variables, an appropriate estimator of the asymptotic covariance matrix of $\overline{\mathbf{m}}_n = [\overline{m}_{n,1}, \ldots, \overline{m}_{n,k}]$ can be computed using

$$\frac{1}{n}\mathbf{F}_{jk} = \frac{1}{n}\left\{\frac{1}{n}\sum_{i=1}^{n}[(m_j(\mathbf{y}_i) - \overline{m}_j)(m_k(\mathbf{y}_i) - \overline{m}_k)]\right\}, \quad j, k = 1, \ldots, K.$$

(One might divide the inner sum by $n - 1$ rather than n. Asymptotically it is the same.) This estimator provides the asymptotic covariance matrix for the moments used in computing the estimated parameters. Under the assumption of i.i.d. random sampling from a distribution with finite moments, $n\mathbf{F}$ will converge in probability to the appropriate covariance matrix of the normalized vector of moments, $\mathbf{\Phi} = \text{Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}_n(\theta)]$. Finally, under our assumptions of random sampling, although the precise distribution is likely to be unknown, we can appeal to the Lindeberg–Levy Central Limit theorem (D.18) to obtain an asymptotic approximation.

To formalize the remainder of this derivation, refer back to the moment equations, which we will now write

$$\bar{m}_{n,k}(\theta_1, \theta_2, \ldots, \theta_K) = 0, \quad k = 1, \ldots, K.$$

The subscript n indicates the dependence on a data set of n observations. We have also combined the sample statistic (sum) and function of parameters, $\mu(\theta_1, \ldots, \theta_K)$ in this general form of the moment equation. Let $\bar{\mathbf{G}}_n(\theta)$ be the $K \times K$ matrix whose kth row is the vector of partial derivatives

$$\bar{\mathbf{G}}'_{n,k} = \frac{\partial \bar{m}_{n,k}}{\partial \theta'}.$$

Now, expand the set of solved moment equations around the true values of the parameters θ_0 in a linear Taylor series. The linear approximation is

$$\mathbf{0} \approx [\bar{\mathbf{m}}_n(\theta_0)] + \bar{\mathbf{G}}'_n(\theta_0)(\hat{\theta} - \theta_0).$$

Therefore,

$$\sqrt{n}(\hat{\theta} - \theta_0) \approx -[\bar{\mathbf{G}}_n(\theta_0)]^{-1}\sqrt{n}[\bar{\mathbf{m}}_n(\theta_0)]. \tag{15-1}$$

(We have treated this as an approximation because we are not dealing formally with the higher order term in the Taylor series. We will make this explicit in the treatment of the GMM estimator in Section 15.4.) The argument needed to characterize the large sample behavior of the estimator, $\hat{\theta}$, is discussed in Appendix D. We have from Theorem D.18 (the Central Limit theorem) that $\sqrt{n}\,\bar{\mathbf{m}}_n(\theta_0)$ has a limiting normal distribution with mean vector $\mathbf{0}$ and covariance matrix equal to $\mathbf{\Phi}$. Assuming that the functions in the moment equation are continuous and functionally independent, we can expect $\bar{\mathbf{G}}_n(\theta_0)$ to converge to a nonsingular matrix of constants, $\mathbf{\Gamma}(\theta_0)$. Under general conditions, the limiting distribution of the right-hand side of (15-1) will be that of a linear function of a normally distributed vector. Jumping to the conclusion, we expect the asymptotic distribution of $\hat{\theta}$ to be normal with mean vector θ_0 and covariance matrix $(1/n) \times \{-[\mathbf{\Gamma}(\theta_0)]^{-1}\}\mathbf{\Phi}\{-[\mathbf{\Gamma}'(\theta_0)]^{-1}\}$. Thus, the asymptotic covariance matrix for the method of moments estimator may be estimated with

$$\text{Est. Asy. Var}[\hat{\theta}] = \frac{1}{n}[\bar{\mathbf{G}}'_n(\hat{\theta})\mathbf{F}^{-1}\bar{\mathbf{G}}_n(\hat{\theta})]^{-1}.$$

Example 15.5 (Continued)

Using the estimates $\hat{\theta}(m'_1, m'_2) = (2.4106,\ 0.0770702)$,

$$\hat{\bar{\mathbf{G}}} = \begin{bmatrix} -1/\hat{\lambda} & \hat{P}/\hat{\lambda}^2 \\ -\hat{\Psi}' & 1/\hat{\lambda} \end{bmatrix} = \begin{bmatrix} -12.97515 & 405.8353 \\ -0.51241 & 12.97515 \end{bmatrix}.$$

[The function Ψ' is $d^2 \ln \Gamma(P)/dP^2 = (\Gamma\Gamma'' - \Gamma'^2)/\Gamma^2$. With $\hat{P} = 2.4106$, $\hat{\Gamma} = 1.250832$, $\hat{\Psi} = 0.658347$, and $\hat{\Psi}' = 0.512408$].[2] The matrix $\mathbf{F}$ is the sample covariance matrix of y and $\ln y$ (using 19 as the divisor),

$$\mathbf{F} = \begin{bmatrix} 25.034 & 0.7155 \\ 0.7155 & 0.023873 \end{bmatrix}.$$

The product is

$$\frac{1}{n}\left[\hat{\mathbf{G}}'\mathbf{F}^{-1}\hat{\mathbf{G}}\right]^{-1} = \begin{bmatrix} 0.38978 & 0.014605 \\ 0.014605 & 0.00068747 \end{bmatrix}.$$

For the maximum likelihood estimator, the estimate of the asymptotic covariance matrix based on the expected (and actual) Hessian is

$$\hat{\mathbf{V}} = [-\mathbf{H}]^{-1} = \frac{1}{n}\begin{bmatrix} \Psi' & -1/\lambda \\ -1/\lambda & P/\lambda^2 \end{bmatrix}^{-1} = \begin{bmatrix} 0.51203 & 0.016377 \\ 0.016377 & 0.00064654 \end{bmatrix}.$$

The Hessian has the same elements as $\mathbf{G}$ because we chose to use the sufficient statistics for the moment estimators, so the moment equations that we differentiated are, apart from a sign change, also the derivatives of the log-likelihood. The estimates of the two variances are 0.51203 and 0.00064654, respectively, which agrees reasonably well with the method of moments estimates. The difference would be due to sampling variability in a finite sample and the presence of $\mathbf{F}$ in the first variance estimator.

15.2.3 SUMMARY—THE METHOD OF MOMENTS

In the simplest cases, the method of moments is robust to differences in the specification of the data generating process (DGP). A sample mean or variance estimates its population counterpart (assuming it exists), regardless of the underlying process. It is this freedom from unnecessary distributional assumptions that has made this method so popular in recent years. However, this comes at a cost. If more is known about the DGP, its specific distribution for example, then the method of moments may not make use of all of the available information. Thus, in Example 15.3, the natural estimators of the parameters of the distribution based on the sample mean and variance turn out to be inefficient. The method of maximum likelihood, which remains the foundation of much work in econometrics, is an alternative approach which utilizes this out of sample information and is, therefore, more efficient.

15.3 MINIMUM DISTANCE ESTIMATION

The preceding analysis has considered **exactly identified cases.** In each example, there were K parameters to estimate and we used K moments to estimate them. In Example 15.5, we examined the gamma distribution, a two-parameter family, and considered different pairs of moments that could be used to estimate the two parameters. (The most efficient estimator for the parameters of this distribution will be based on $(1/n)\Sigma_i y_i$ and $(1/n)\Sigma_i \ln y_i$. This does raise a general question: How should we proceed if we have more moments than we need? It would seem counterproductive to simply discard the

AU: Confirm 13.5 is correct x-ref example number

[2] Ψ' is the trigamma function. Values for $\Gamma(P)$, $\Psi(P)$, and $\Psi'(P)$ are tabulated in Abramovitz and Stegun (1971). The values given were obtained using the IMSL computer program library.

additional information. In this case, logically, the sample information provides more than one estimate of the model parameters, and it is now necessary to reconcile those competing estimators.

We have encountered this situation in several earlier examples: In Example 9.13, in Passmore's (2005) study of Fannie Mae, we have four independent estimators of a single parameter, $\hat{\alpha}_j$, with estimated asymptotic variance $\hat{V}_j$, $j = 1, \ldots, 4$. The estimators were combined using a **criterion function:**

$$\text{minimize with respect to } \alpha : q = \sum_{j=i}^{4} \frac{(\hat{\alpha}_j - \alpha)^2}{\hat{V}_j}.$$

The solution to this minimization problem is

$$\hat{\alpha}_{\text{MDE}} = \sum_{j=1}^{4} w_j \hat{\alpha}_j, \; w_j = \frac{1/\hat{V}_j}{\sum_{s=1}^{4}(1/\hat{V}_s)}, \; j = 1, \ldots, 4 \text{ and } \sum_{j=1}^{4} w_j = 1.$$

In forming the two-stage least squares estimator of the parameters in a dynamic panel data model in Section 12.8.2, we obtained $T - 2$ instrumental variable estimators of the parameter vector θ by forming different instruments for each period for which we had sufficient data. The $T - 2$ estimators of the same parameter vector are $\hat{\theta}_{\text{IV}(t)}$. The Arellano–Bond estimator of the single parameter vector in this setting is

$$\hat{\theta}_{\text{IV}} = \left(\sum_{t=3}^{T} \mathbf{W}_{(t)}\right)^{-1} \left(\sum_{t=3}^{T} \mathbf{W}_{(t)} \hat{\theta}_{\text{IV}(t)}\right)$$

$$= \sum_{t=3}^{T} \mathbf{R}_{(t)} \hat{\theta}_{\text{IV}(t)},$$

where

$$\mathbf{W}_{(t)} = \left(\hat{\mathbf{X}}'_{(t)} \hat{\mathbf{X}}_{(t)}\right)$$

and

$$\mathbf{R}_{(t)} = \left(\sum_{t=3}^{T} \mathbf{W}_{(t)}\right)^{-1} \mathbf{W}_{(t)} \text{ and } \sum_{t=3}^{T} \mathbf{R}_{(t)} = \mathbf{I}.$$

Finally, Carey's (1997) analysis of hospital costs that we examined in Example 10.6 involved a seemingly unrelated regressions model that produced multiple estimates of several of the model parameters. We will revisit this application in Example 15.6.

A minimum distance estimator (MDE) is defined as follows: Let $\overline{m}_{n,l}$ denote a sample statistic based on n observations such that

$$\text{plim } \overline{m}_{n,l} = g_l(\theta_0), \; l = 1, \ldots, L,$$

where θ_0 is a vector of $K \leq L$ parameters to be estimated. Arrange these moments and functions in $L \times 1$ vectors $\overline{\mathbf{m}}_n$ and $\mathbf{g}(\theta_0)$ and further assume that the statistics are jointly asymptotically normally distributed with plim $\overline{\mathbf{m}}_n = \mathbf{g}(\theta)$ and Asy. Var$[\overline{\mathbf{m}}_n] = (1/n)\boldsymbol{\Phi}$.

Define the criterion function

$$q = [\overline{\mathbf{m}}_n - \mathbf{g}(\theta)]'\mathbf{W}[\overline{\mathbf{m}}_n - \mathbf{g}(\theta)]$$

for a positive definite **weighting matrix, W**. The minimum distance estimator is the $\hat{\theta}_{\text{MDE}}$ that minimizes q. Different choices of $\mathbf{W}$ will produce different estimators, but the estimator has the following properties for any $\mathbf{W}$:

THEOREM 13.1 Asymptotic Distribution of the Minimum Distance Estimator

Under the assumption that $\sqrt{n}[\overline{\mathbf{m}}_n - \mathbf{g}(\theta_0)] \xrightarrow{d} N[\mathbf{0}, \mathbf{\Phi}]$, *the asymptotic properties of the minimum distance estimator are as follows:*

$$\text{plim}\,\hat{\theta}_{\text{MDE}} = \theta_0,$$

$$\text{Asy. Var}\,[\hat{\theta}_{\text{MDE}}] = \frac{1}{n}[\mathbf{\Gamma}(\theta_0)'\mathbf{W}(\mathbf{\Gamma}\theta_0)]^{-1}[\mathbf{\Gamma}(\theta_0)'\mathbf{W}\mathbf{\Phi}\mathbf{W}\mathbf{\Gamma}(\theta_0)][\mathbf{\Gamma}(\theta_0)'\mathbf{W}\mathbf{\Gamma}(\theta_0)]^{-1}$$

$$= \frac{1}{n}\mathbf{V},$$

where

$$\mathbf{\Gamma}(\theta_0) = \text{plim}\,\mathbf{G}(\hat{\theta}_{\text{MDE}}) = \text{plim}\frac{\partial \mathbf{g}(\hat{\theta}_{\text{MDE}})}{\partial \hat{\theta}'_{\text{MDE}}},$$

and

$$\hat{\theta}_{\text{MDE}} \xrightarrow{a} N\left[\theta_0, \frac{1}{n}\mathbf{V}\right].$$

Proofs may be found in Malinvaud (1970) and Amemiya (1985). For our purposes, we can note that the MDE is an extension of the method of moments presented in the preceding section. One implication is that the estimator is consistent for any $\mathbf{W}$, but the asymptotic covariance matrix is a function of $\mathbf{W}$. This suggests that the choice of $\mathbf{W}$ might be made with an eye toward the size of the covariance matrix and that there might be an optimal choice. That does indeed turn out to be the case. For minimum distance estimation, the weighting matrix that produces the smallest variance is

$$\textbf{optimal weighting matrix: } \mathbf{W}^* = [\textit{Asy. Var.}\sqrt{n}\{\overline{\mathbf{m}}_n - \mathbf{g}(\theta)\}]^{-1}$$

$$= \mathbf{\Phi}^{-1}.$$

[See Hansen (1982) for discussion.] With this choice of $\mathbf{W}$,

$$\text{Asy. Var}\,[\hat{\theta}_{\text{MDE}}] = \frac{1}{n}\left[\mathbf{\Gamma}(\theta_0)'\mathbf{\Phi}^{-1}\mathbf{\Gamma}(\theta_0)\right]^{-1},$$

which is the result we had earlier for the method of moments estimator.

AU: Is "Asy. Var." ok in Italics?

The solution to the MDE estimation problem is found by locating the $\hat{\theta}_{\text{MDE}}$ such that

$$\frac{\partial q}{\partial \hat{\theta}_{\text{MDE}}} = -\mathbf{G}(\hat{\theta}_{\text{MDE}})'\mathbf{W}\left[\overline{\mathbf{m}}_n - \mathbf{g}(\hat{\theta}_{\text{MDE}})\right] = \mathbf{0}.$$

An important aspect of the MDE arises in the exactly identified case. If K equals L, and if the functions $g_l(\theta)$ are functionally independent, that is, $\mathbf{G}(\theta)$ has full row rank, K, then it is possible to solve the moment equations exactly. That is, the minimization problem becomes one of simply solving the K moment equations, $\overline{m}_{n,l} = g_l(\theta_0)$ in the K unknowns, $\hat{\theta}_{\text{MDE}}$. This is the method of moments estimator examined in the preceding section. In this instance, the weighting matrix, $\mathbf{W}$, is irrelevant to the solution, because the MDE will now satisfy the moment equations

$$\left[\overline{\mathbf{m}}_n - \mathbf{g}(\hat{\theta}_{\text{MDE}})\right] = \mathbf{0}.$$

For the examples listed earlier, which are all for **overidentified cases**, the minimum distance estimators are defined by

$$q = \begin{pmatrix}(\hat{\alpha}_1 - \alpha) & (\hat{\alpha}_2 - \alpha) & (\hat{\alpha}_3 - \alpha) & (\hat{\alpha}_4 - \alpha)\end{pmatrix}\begin{bmatrix}\hat{V}_1 & 0 & 0 & 0\\ 0 & \hat{V}_2 & 0 & 0\\ 0 & 0 & \hat{V}_3 & 0\\ 0 & 0 & 0 & \hat{V}_4\end{bmatrix}^{-1}\begin{pmatrix}(\hat{\alpha}_1 - \alpha)\\ (\hat{\alpha}_2 - \alpha)\\ (\hat{\alpha}_3 - \alpha)\\ (\hat{\alpha}_4 - \alpha)\end{pmatrix}$$

for Passmore's analysis of Fannie Mae, and

$$q = \begin{pmatrix}(\mathbf{b}_{\text{IV}(3)} - \theta) & \cdots & (\mathbf{b}_{\text{IV}(T)} - \theta)\end{pmatrix}'\begin{bmatrix}\left(\hat{\mathbf{X}}'_{(3)}\hat{\mathbf{X}}_{(3)}\right) & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & \left(\hat{\mathbf{X}}'_{(T)}\hat{\mathbf{X}}_{(T)}\right)\end{bmatrix}^{-1}\begin{pmatrix}(\mathbf{b}_{\text{IV}(3)} - \theta)\\ \vdots\\ (\mathbf{b}_{\text{IV}(T)} - \theta)\end{pmatrix}$$

for the Arellano–Bond estimator of the dynamic panel data model.

Example 15.6 Minimum Distance Estimation of a Hospital Cost Function

In Carey's (1997) study of hospital costs in Example 10.8, Chamberlain's (1984) seemingly unrelated regressions approach to a panel data model produces five period-specific estimates of a parameter vector, θ_t. Some of the parameters are specific to the year while others (it is hypothesized) are common to all five years. There are two specific parameters of interest, β_D and β_O, that are allowed to vary by year, but are each estimated multiple times by the SUR model. We focus on just these parameters. The model states

$$y_{it} = \alpha_i + A_{it} + \beta_{D,t}\,DIS_{it} + \beta_{O,t}\,OUT_{it} + \varepsilon_{it},$$

where

$$\alpha_i = B_i + \Sigma_t \gamma_{D,t}\,DIS_{it} + \Sigma_t \gamma_{O,t}\,OUT_{it} + u_i, t = 1987, \ldots, 1991,$$

DIS_{it} is patient discharges, and OUT_{it} is outpatient visits. (We are changing Carey's notation slightly and suppressing parts of the model that are extraneous to the development here. The terms A_{it} and B_i contain those additional components.) The preceding model is estimated by inserting the expression for α_i in the main equation, then fitting an unrestricted seemingly unrelated regressions model by FGLS. There are five years of data, hence five sets of estimates. Note, however, with respect to the discharge variable, DIS, although each equation provides separate estimates of $(\gamma_{D,1}, \ldots, (\beta_{D,t} + \gamma_{D,t}), \ldots, \gamma_{D,5})$, a total of five parameter estimates in each each equation (year), there are only 10, not 25 parameters to be estimated in total.

TABLE 15.1a Coefficient Estimates for DIS in SUR Model for Hospital Costs

	Coefficient on Variable in the Equation				
Equation	*DIS87*	*DIS88*	*DIS89*	*DIS90*	*DIS91*
SUR87	$\beta_{D,87} + \gamma_{D,87}$ 1.76	$\gamma_{D,88}$ 0.116	$\gamma_{D,89}$ −0.0881	$\gamma_{D,90}$ 0.0570	$\gamma_{D,91}$ −0.0617
SUR88	$\gamma_{D,87}$ 0.254	$\beta_{D,88} + \gamma_{D,88}$ 1.61	$\gamma_{D,89}$ −0.0934	$\gamma_{D,90}$ 0.0610	$\gamma_{D,91}$ −0.0514
SUR89	$\gamma_{D,87}$ 0.217	$\gamma_{D,88}$ 0.0846	$\beta_{D,89} + \gamma_{D,89}$ 1.51	$\gamma_{D,90}$ 0.0454	$\gamma_{D,91}$ −0.0253
SUR90	$\gamma_{D,87}$ 0.179	$\gamma_{D,88}$ 0.0822	$\gamma_{D,89}$ 0.0295	$\beta_{D,90} + \gamma_{D,90}$ 1.57	$\gamma_{D,91}$ 0.0244
SUR91	$\gamma_{D,87}$ 0.153	$\gamma_{D,88}$ 0.0363	$\gamma_{D,89}$ −0.0422	$\gamma_{D,90}$ 0.0813	$\beta_{D,91} + \gamma_{D,91}$ 1.70
MDE	$\beta = 1.50$ $\gamma = 0.219$	$\beta = 1.58$ $\gamma = 0.0666$	$\beta = 1.54$ $\gamma = -0.0539$	$\beta = 1.57$ $\gamma = 0.0690$	$\beta = 1.63$ $\gamma = -0.0213$

TABLE 15.1b Coefficient Estimates for OUT in SUR Model for Hospital Costs

	Coefficient on Variable in the Equation				
Equation	*OUT87*	*OUT88*	*OUT89*	*OUT90*	*OUT91*
SUR87	$\beta_{O,87} + \gamma_{D,87}$ 0.0139	$\gamma_{O,88}$ 0.00292	$\gamma_{O,89}$ 0.00157	$\gamma_{O,90}$ 0.000951	$\gamma_{O,91}$ 0.000678
SUR88	$\gamma_{O,87}$ 0.00347	$\beta_{O,88} + \gamma_{O,88}$ 0.0125	$\gamma_{O,89}$ 0.00501	$\gamma_{O,90}$ 0.00550	$\gamma_{O,91}$ 0.00503
SUR89	$\gamma_{O,87}$ 0.00118	$\gamma_{O,88}$ 0.00159	$\beta_{O,89} + \gamma_{O,89}$ 0.00832	$\gamma_{O,90}$ −0.00220	$\gamma_{O,91}$ −0.00156
SUR90	$\gamma_{O,87}$ −0.00226	$\gamma_{O,88}$ −0.00155	$\gamma_{O,89}$ 0.000401	$\beta_{O,90} + \gamma_{O,90}$ 0.00897	$\gamma_{O,91}$ 0.000450
SUR91	$\gamma_{O,87}$ 0.00278	$\gamma_{O,88}$ 0.00255	$\gamma_{O,89}$ 0.00233	$\gamma_{O,90}$ 0.00305	$\beta_{O,91} + \gamma_{O,91}$ 0.0105
MDE	$\beta = 0.0112$ $\gamma = 0.00177$	$\beta = 0.00999$ $\gamma = 0.00408$	$\beta = 0.0100$ $\gamma = -0.00011$	$\beta = 0.00915$ $\gamma = -0.00073$	$\beta = 0.00793$ $\gamma = 0.00267$

TB 13.1

The parameters on OUT_{it} are likewise overidentified. Table 15.1 reproduces the estimates in Table 10.2 for the discharge coefficients and adds the estimates for the outpatient variable.

Looking at the tables we see that the SUR model provides four direct estimates of $\gamma_{D,87}$, based on the 1988–1991 equations. It also implicitly provides four estimates of $\beta_{D,87}$ since any of the four estimates of $\gamma_{D,87}$ from the last four equations can be subtracted from the coefficient on DIS in the 1987 equation to estimate $\beta_{D,87}$. There are 50 parameter estimates of different functions of the 20 underlying parameters

$$\theta = (\beta_{D,87}, \ldots, \beta_{D,91}), (\gamma_{D,87}, \ldots, \gamma_{D,91}), (\beta_{O,87}, \ldots, \beta_{O,91}), (\gamma_{O,87}, \ldots, \gamma_{O,91}),$$

and, therefore, 30 constraints to impose in finding a common, restricted estimator. An MDE was used to reconcile the competing estimators.

Let $\hat{\beta}_t$ denote the 10×1 period-specific estimator of the model parameters. Unlike the other cases we have examined, the individual estimates here are not uncorrelated. In the SUR model, the estimated asymptotic covariance matrix is the partitioned matrix given in

(10-7). For the estimators of two equations,

$$\text{Est. Asy. Cov}\left[\hat{\beta}_t, \hat{\beta}_s\right] = \text{the } t,s \text{ block of } \begin{bmatrix} \hat{\sigma}^{11}\mathbf{X}_1'\mathbf{X}_1 & \hat{\sigma}^{12}\mathbf{X}_1'\mathbf{X}_2 & \cdots & \hat{\sigma}^{15}\mathbf{X}_1'\mathbf{X}_5 \\ \hat{\sigma}^{21}\mathbf{X}_2'\mathbf{X}_1 & \hat{\sigma}^{22}\mathbf{X}_2'\mathbf{X}_2 & \cdots & \hat{\sigma}^{25}\mathbf{X}_2'\mathbf{X}_5 \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\sigma}^{51}\mathbf{X}_5'\mathbf{X}_1 & \hat{\sigma}^{52}\mathbf{X}_5'\mathbf{X}_2 & \cdots & \hat{\sigma}^{55}\mathbf{X}_5'\mathbf{X}_5 \end{bmatrix}^{-1} = \hat{\mathbf{V}}_{ts}$$

where $\hat{\sigma}^{ts}$ is the t,s element of $\hat{\boldsymbol{\Sigma}}^{-1}$. (We are extracting a submatrix of the relevant matrices here since Carey's SUR model contained 26 other variables in each equation in addition to the five periods of DIS and OUT). The 50×50 weighting matrix for the MDE is

$$\mathbf{W} = \begin{bmatrix} \hat{\mathbf{V}}_{87,87} & \hat{\mathbf{V}}_{87,88} & \hat{\mathbf{V}}_{87,89} & \hat{\mathbf{V}}_{87,90} & \hat{\mathbf{V}}_{87,91} \\ \hat{\mathbf{V}}_{88,87} & \hat{\mathbf{V}}_{88,88} & \hat{\mathbf{V}}_{88,89} & \hat{\mathbf{V}}_{88,90} & \hat{\mathbf{V}}_{88,91} \\ \hat{\mathbf{V}}_{89,87} & \hat{\mathbf{V}}_{89,88} & \hat{\mathbf{V}}_{89,89} & \hat{\mathbf{V}}_{89,90} & \hat{\mathbf{V}}_{89,91} \\ \hat{\mathbf{V}}_{90,87} & \hat{\mathbf{V}}_{90,88} & \hat{\mathbf{V}}_{90,89} & \hat{\mathbf{V}}_{90,90} & \hat{\mathbf{V}}_{90,91} \\ \hat{\mathbf{V}}_{91,87} & \hat{\mathbf{V}}_{91,88} & \hat{\mathbf{V}}_{91,89} & \hat{\mathbf{V}}_{91,90} & \hat{\mathbf{V}}_{91,91} \end{bmatrix}^{-1} = \left[\hat{\mathbf{V}}^{ts}\right].$$

The vector of the quadratic form is a stack of five 10×1 vectors; the first is

$$\bar{\mathbf{m}}_{n,87} - \mathbf{g}_{87}(\theta) = \begin{bmatrix} \left\{\hat{\beta}^{87}_{D,87} - (\beta_{D,87} + \gamma_{D,87})\right\}, \left\{\hat{\beta}^{87}_{D,88} - \gamma_{D,88}\right\}, \left\{\hat{\beta}^{87}_{D,89} - \gamma_{D,89}\right\}, \left\{\hat{\beta}^{87}_{D,90} - \gamma_{D,90}\right\}, \left\{\hat{\beta}^{87}_{D,91} - \gamma_{D,90}\right\}, \\ \left\{\hat{\beta}^{87}_{O,87} - (\beta_{O,87} + \gamma_{O,87})\right\}, \left\{\hat{\beta}^{87}_{O,88} - \gamma_{O,88}\right\}, \left\{\hat{\beta}^{87}_{O,89} - \gamma_{D,89}\right\}, \left\{\hat{\beta}^{87}_{O,90} - \gamma_{O,90}\right\}, \left\{\hat{\beta}^{87}_{O,91} - \gamma_{O,90}\right\} \end{bmatrix}$$

for the 1987 equation and likewise for the other four equations. The MDE criterion function for this model is

$$q = \sum_{t=1987}^{1991} \sum_{s=1987}^{1981} [\bar{\mathbf{m}}_t - \mathbf{g}_t(\theta)]' \hat{\mathbf{V}}^{ts} [\bar{\mathbf{m}}_s - \mathbf{g}_s(\theta)].$$

Note, there are 50 estimated parameters from the SUR equations (those are listed in Table 15.1) and 20 unknown parameters to be calibrated in the criterion function. The reported minimum distance estimates are shown in the last row of each table.

15.4 THE GENERALIZED METHOD OF MOMENTS (GMM) ESTIMATOR

A large proportion of the recent empirical work in econometrics, particularly in macroeconomics and finance, has employed GMM estimators. As we shall see, this broad class of estimators, in fact, includes most of the estimators discussed elsewhere in this book.

The GMM estimation technique is an extension of the minimum distance technique described in Section 15.3.[3] In the following, we will extend the generalized method of moments to other models beyond the generalized linear regression, and we will fill in some gaps in the derivation in Section 15.2.

[3] Formal presentation of the results required for this analysis are given by Hansen (1982); Hansen and Singleton (1988); Chamberlain (1987); Cumby, Huizinga, and Obstfeld (1983); Newey (1984, 1985a, 1985b); Davidson and MacKinnon (1993); and Newey and McFadden (1994). Useful summaries of GMM estimation and other developments in econometrics are provided by Pagan and Wickens (1989) and Matyas (1999). An application of some of these techniques that contains useful summaries is Pagan and Vella (1989). Some further discussion can be found in Davidson and MacKinnon (2004). Ruud (2000) provides many of the theoretical details. Hayashi (2000) is another extensive treatment of estimation centered on GMM estimators.

15.4.1 ESTIMATION BASED ON ORTHOGONALITY CONDITIONS

Consider the least squares estimator of the parameters in the classical linear regression model. An important assumption of the model is

$$E[\mathbf{x}_i\varepsilon_i] = E[\mathbf{x}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

The sample analog is

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i\hat{\varepsilon}_i = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}) = \mathbf{0}.$$

The estimator of $\boldsymbol{\beta}$ is the one that satisfies these moment equations, which are just the normal equations for the least squares estimator. So, we see that the OLS estimator is a method of moments estimator.

For the instrumental variables estimator of Chapter 12, we relied on a large sample analog to the moment condition,

$$\operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right) = \operatorname{plim}\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right) = \mathbf{0}.$$

We resolved the problem of having more instruments than parameters by solving the equations

$$\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{Z}\right)^{-1}\left(\frac{1}{n}\mathbf{Z}'\hat{\boldsymbol{\varepsilon}}\right) = \frac{1}{n}\hat{\mathbf{X}}'\hat{\boldsymbol{\varepsilon}} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{x}}_i\hat{\varepsilon}_i = \mathbf{0},$$

where the columns of $\hat{\mathbf{X}}$ are the fitted values in regressions on all the columns of $\mathbf{Z}$ (that is, the projections of these columns of $\mathbf{X}$ into the column space of $\mathbf{Z}$). (See Section 12.3.3 for further details.)

The nonlinear least squares estimator was defined similarly, although in this case, the normal equations are more complicated because the estimator is only implicit. The population **orthogonality condition** for the nonlinear regression model is $E[\mathbf{x}_i^0\varepsilon_i] = \mathbf{0}$. The **empirical moment equation** is

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\partial E[y_i \mid \mathbf{x}_i, \boldsymbol{\beta}]}{\partial \boldsymbol{\beta}}\right)(y_i - E[y_i \mid \mathbf{x}_i, \boldsymbol{\beta}]) = \mathbf{0}.$$

Maximum likelihood estimators are obtained by equating the derivatives of a log-likelihood to zero. The scaled log-likelihood function is

$$\frac{1}{n}\ln L = \frac{1}{n}\sum_{i=1}^{n}\ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta}),$$

where $f(\cdot)$ is the density function and $\boldsymbol{\theta}$ is the parameter vector. For densities that satisfy the regularity conditions [see Chapter 16],

$$E\left[\frac{\partial \ln f(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \mathbf{0}.$$

The maximum likelihood estimator is obtained by equating the sample analog to zero:

$$\frac{1}{n}\frac{\partial \ln L}{\partial \hat{\theta}} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \ln f(y_i \mid \mathbf{x}_i, \hat{\theta})}{\partial \hat{\theta}} = \mathbf{0}.$$

(Dividing by n to make this result comparable to our earlier ones does not change the solution.) The upshot is that nearly all the estimators we have discussed and will encounter later can be construed as method of moments estimators. [Manski's (1992) treatment of **analog estimation** provides some interesting extensions and methodological discourse.]

As we extend this line of reasoning, it will emerge that most of the estimators defined in this book can be viewed as generalized method of moments estimators.

15.4.2 GENERALIZING THE METHOD OF MOMENTS

The preceding examples all have a common aspect. In each case listed, save for the general case of the instrumental variable estimator, there are exactly as many moment equations as there are parameters to be estimated. Thus, each of these are **exactly identified** cases. There will be a single solution to the moment equations, and at that solution, the equations will be exactly satisfied.[4] But there are cases in which there are more moment equations than parameters, so the system is overdetermined. In Example 15.5, we defined four sample moments,

$$\bar{\mathbf{g}} = \frac{1}{n}\sum_{i=1}^{n}\left[y_i, y_i^2, \frac{1}{y_i}, \ln y_i\right]$$

with probability limits P/λ, $P(P+1)/\lambda^2$, $\lambda/(P-1)$, and $\psi(P) - \ln\lambda$, respectively. Any pair could be used to estimate the two parameters, but as shown in the earlier example, the six pairs produce six somewhat different estimates of $\theta = (P, \lambda)$.

In such a case, to use all the information in the sample it is necessary to devise a way to reconcile the conflicting estimates that may emerge from the overdetermined system. More generally, suppose that the model involves K parameters, $\theta = (\theta_1, \theta_2, \ldots, \theta_K)'$, and that the theory provides a set of $L > K$ moment conditions,

$$E[m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \theta)] = E[m_{il}(\theta)] = 0,$$

where y_i, $\mathbf{x}_i$, and $\mathbf{z}_i$ are variables that appear in the model and the subscript i on $m_{il}(\theta)$ indicates the dependence on $(y_i, \mathbf{x}_i, \mathbf{z}_i)$. Denote the corresponding sample means as

$$\bar{m}_l(\mathbf{y}, \mathbf{X}, \mathbf{Z}, \theta) = \frac{1}{n}\sum_{i=1}^{n} m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \theta) = \frac{1}{n}\sum_{i=1}^{n} m_{il}(\theta).$$

Unless the equations are functionally dependent, the system of L equations in K unknown parameters,

$$\bar{m}_l(\theta) = \frac{1}{n}\sum_{i=1}^{n} m_l(y_i, \mathbf{x}_i, \mathbf{z}_i, \theta) = 0, \quad l = 1, \ldots, L,$$

[4] That is, of course if there is *any* solution. In the regression model with multicollinearity, there are K parameters but fewer than K independent moment equations.

AU: Confirm renumbered pages are in correct order mspp 13-18–13-24

13-23 (13-18)

will not have a unique solution.[5] For convenience, the moment equations are defined implicitly here as opposed to equalities of moments to functions as in Section 15.3. It will be necessary to reconcile the $(\frac{L}{K})$ different sets of estimates that can be produced. One possibility is to minimize a criterion function, such as the sum of squares,[6]

$$q = \sum_{l=1}^{L} \bar{m}_l^2 = \bar{\mathbf{m}}(\theta)'\bar{\mathbf{m}}(\theta). \tag{15-2}$$

It can be shown [see, e.g., Hansen (1982)] that under the assumptions we have made so far, specifically that $\text{plim}\,\bar{\mathbf{m}}(\theta) = E[\bar{\mathbf{m}}(\theta)] = \mathbf{0}$, the minimizer of q in (15-2) produces a consistent (albeit, as we shall see, possibly inefficient) estimator of θ. We can, in fact, use as the criterion a weighted sum of squares,

$$q = \bar{\mathbf{m}}(\theta)'\mathbf{W}_n\bar{\mathbf{m}}(\theta),$$

where $\mathbf{W}_n$ is *any* positive definite matrix that may depend on the data but is not a function of θ, such as $\mathbf{I}$ in (15-2), to produce a consistent estimator of θ.[7] For example, we might use a diagonal matrix of weights if some information were available about the importance (by some measure) of the different moments. We do make the additional assumption that $\text{plim}\,\mathbf{W}_n =$ a positive definite matrix, $\mathbf{W}$.

By the same logic that makes generalized least squares preferable to ordinary least squares, it should be beneficial to use a weighted criterion in which the weights are inversely proportional to the variances of the moments. Let $\mathbf{W}$ be a diagonal matrix whose diagonal elements are the reciprocals of the variances of the individual moments,

$$w_{ll} = \frac{1}{\text{Asy. Var}[\sqrt{n}\,\bar{m}_l]} = \frac{1}{\phi_{ll}}.$$

(We have written it in this form to emphasize that the right-hand side involves the variance of a sample mean which is of order $(1/n)$.) Then, a **weighted least squares** estimator would minimize

$$q = \bar{\mathbf{m}}(\theta)'\mathbf{\Phi}^{-1}\bar{\mathbf{m}}(\theta). \tag{15-3}$$

In general, the L elements of $\bar{\mathbf{m}}$ are freely correlated. In (15-3), we have used a diagonal $\mathbf{W}$ that ignores this correlation. To use generalized least squares, we would define the full matrix,

$$\mathbf{W} = \{\text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}]\}^{-1} = \mathbf{\Phi}^{-1}. \tag{15-4}$$

The estimators defined by choosing θ to minimize

$$q = \bar{\mathbf{m}}(\theta)'\mathbf{W}_n\bar{\mathbf{m}}(\theta)$$

[5] It may if L is greater than the sample size, n. We assume that L is strictly less than n.

[6] This approach is one that Quandt and Ramsey (1978) suggested for the problem in Example 15.4.

[7] In principle, the weighting matrix can be a function of the parameters as well. See Hansen, Heaton, and Yaron (1996) for discussion. Whether this provides any benefit in terms of the asymptotic properties of the estimator seems unlikely. The one payoff the authors do note is that certain estimators become invariant to the sort of normalization that is discussed in Example 16.1. In practical terms, this is likely to be a consideration only in a fairly small class of cases.

are minimum distance estimators as defined in Section 15.3. The general result is that if $\mathbf{W}_n$ is a positive definite matrix and if

$$\text{plim } \bar{\mathbf{m}}(\boldsymbol{\theta}) = \mathbf{0},$$

then the minimum distance (generalized method of moments, or GMM) estimator of $\boldsymbol{\theta}$ is consistent.[8] Because the OLS criterion in (15-2) uses $\mathbf{I}$, this method produces a consistent estimator, as does the weighted least squares estimator and the full GLS estimator. What remains to be decided is the best $\mathbf{W}$ to use. Intuition might suggest (correctly) that the one defined in (15-4) would be optimal, once again based on the logic that motivates generalized least squares. This result is the now-celebrated one of Hansen (1982).

The asymptotic covariance matrix of this **generalized method of moments estimator** is

$$\mathbf{V}_{GMM} = \frac{1}{n}[\boldsymbol{\Gamma}'\mathbf{W}\boldsymbol{\Gamma}]^{-1} = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}]^{-1}, \tag{15-5}$$

where $\boldsymbol{\Gamma}$ is the matrix of derivatives with jth row equal to

$$\boldsymbol{\Gamma}^j = \text{plim } \frac{\partial \bar{m}_j(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'},$$

and $\boldsymbol{\Phi} = \text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}]$. Finally, by virtue of the central limit theorem applied to the sample moments and the **Slutsky theorem** applied to this manipulation, we can expect the estimator to be asymptotically normally distributed. We will revisit the asymptotic properties of the estimator in Section 15.4.3.

Example 15.7 GMM Estimation of the Parameters of a Gamma Distribution

Referring once again to our earlier results in Example 15.5, we consider how to use all four of our sample moments to estimate the parameters of the gamma distribution.[9] The four moment equations are

$$E\begin{bmatrix} y_i - P/\lambda \\ y_i^2 - P(P+1)/\lambda^2 \\ \ln y_i - \Psi(P) + \ln\lambda \\ 1/y_i - \lambda/(P-1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

The sample means of these will provide the moment equations for estimation. Let $y_1 = y$, $y_2 = y^2$, $y_3 = \ln y$, and $y_4 = 1/y$. Then

$$\bar{m}_1(P, \lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_{i1} - P/\lambda) = \frac{1}{n}\sum_{i=1}^{n}[y_{i1} - \mu_1(P, \lambda)] = \bar{y}_1 - \mu_1(P, \lambda),$$

and likewise for $\bar{m}_2(P, \lambda)$, $\bar{m}_3(P, \lambda)$, and $\bar{m}_4(P, \lambda)$.

[8] In the most general cases, a number of other subtle conditions must be met so as to assert consistency and the other properties we discuss. For our purposes, the conditions given will suffice. Minimum distance estimators are discussed in Malinvaud (1970), Hansen (1982), and Amemiya (1985).

[9] We emphasize that this example is constructed only to illustrate the computation of a GMM estimator. The gamma model is fully specified by the likelihood function, and the MLE is fully efficient. We will examine other cases that involve less detailed specifications later in this chapter.

Example 13.7 GMM Estimation of a Nonlinear Regression Model

In Example 7.6, we examined a nonlinear regression model for income using the German Socioeconomic Panel Data set. The regression model was

$$Income = h(1, Age, Education, Female, \gamma) + \varepsilon,$$

where $h(.)$ is an exponential function of the variables. In the example, we used several interaction terms. In this application, we will simplify the conditional mean function somewhat, and use

$$Income = \exp(\gamma_1 + \gamma_2 Age + \gamma_3 Education + \gamma_4 Female) + \varepsilon,$$

which, for convenience, we will write

$$\begin{aligned} y_i &= \exp(\mathbf{x}_i'\boldsymbol{\gamma}) + \varepsilon_i \\ &= \mu_i + \varepsilon_i. \end{aligned}$$

FN 9

The sample consists of the 1988 wave of the panel, less two observations for which *Income* equals zero. The resulting sample contains 4481 observations. Descriptive statistics for the sample data are given in Table 7.2.

We will first consider nonlinear least squares estimation of the parameters. The normal equations for nonlinear least squares will be

$$(1/n)\,\Sigma_i\,[(y_i - \mu_i)\,\mu_i\mathbf{x}_i] = (1/n)\,\Sigma_i\,[\varepsilon_i\,\mu_i\,\mathbf{x}_i] = \mathbf{0}.$$

Note that the orthogonality condition involves the pseudoregressors, $\partial\mu_i/\partial\boldsymbol{\gamma} = \mathbf{x}_i^0 = \mu_i\mathbf{x}_i$. The implied population moment equation is

$$E[\,\varepsilon_i\,(\mu_i\mathbf{x}_i)] = \mathbf{0}.$$

Computation of the nonlinear least squares estimator is discussed in Section 7.2.6. The estimator of the asymptotic covariance matrix is

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}_{\text{NLSQ}}] = \frac{\Sigma_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{(4481-4)}\left[\sum_{i=1}^{4481}(\hat{\mu}_i\mathbf{x}_i)(\hat{\mu}_i\mathbf{x}_i)'\right]^{-1}, \text{ where } \hat{\mu}_i = \exp(\mathbf{x}_i'\hat{\boldsymbol{\gamma}}).$$

A simple method of moments estimator might be constructed from the hypothesis that $\mathbf{x}_i$ (not $\mathbf{x}_i^0$) is orthogonal to ε_i. Then,

$$E[\varepsilon_i\,\mathbf{x}_i] = E\left[\varepsilon_i\begin{pmatrix}1\\ Age_i\\ Education_i\\ Female_i\end{pmatrix}\right] = \mathbf{0}$$

[9] We note that in this model, it is likely that *Education* is endogenous. It would be straightforward to accommodate that in the GMM estimator. However, for purposes of a straightforward numerical example, we will proceed assuming that *Education* is exogenous.

implies four moment equations. The sample counterparts will be

$$\bar{m}_k(\boldsymbol{\gamma}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu_i)x_{ik} = \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i x_{ik}.$$

In order to compute the method of moments estimator, we will minimize the sum of squares,

$$\bar{\mathbf{m}}'(\boldsymbol{\gamma})\bar{\mathbf{m}}(\boldsymbol{\gamma}) = \sum_{k=1}^{4}\bar{m}_k^2(\boldsymbol{\gamma})$$

This is a nonlinear optimization problem that must be solved iteratively using the methods described in Section E.3.

With the first step estimated parameters, $\hat{\boldsymbol{\gamma}}^0$, in hand, the covariance matrix is estimated using (13-5).

$$\hat{\boldsymbol{\Phi}} = \left\{\frac{1}{4481}\sum_{i=1}^{4481}\mathbf{m}_i(\hat{\boldsymbol{\gamma}}^0)\mathbf{m}_i'(\hat{\boldsymbol{\gamma}}^0)\right\} = \left\{\frac{1}{4481}\sum_{i=1}^{4481}\left(\hat{\varepsilon}_i^0\mathbf{x}_i\right)\left(\hat{\varepsilon}_i^0\mathbf{x}_i\right)'\right\}$$

$$\bar{\mathbf{G}} = \left\{\frac{1}{4481}\sum_{i=1}^{n}\left(\hat{\varepsilon}_i^0\mathbf{x}_i\right)\left(-\hat{\mu}_i^0\mathbf{x}_i\right)'\right\}.$$

The asymptotic covariance matrix for the MOM estimator is computed using (13-5),

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}_{\text{MOM}}] = \frac{1}{n}\left[\bar{\mathbf{G}}\hat{\boldsymbol{\Phi}}^{-1}\bar{\mathbf{G}}'\right]^{-1}.$$

Suppose we have in hand additional variables, *Health Satisfaction* and *Marital Status* such that although the conditional mean function remains as given above, we will use them to form a GMM estimator. This provides two additional moment equations,

$$E\left[\varepsilon_i\begin{pmatrix}\textit{Health Satisfaction}_i\\ \textit{Marital Status}_i\end{pmatrix}\right]$$

for a total of six moment equations for estimating the four parameters. We constuct the generalized method of moments estimator as follows: The initial step is the same as before, except the sum of squared moments, $\bar{\mathbf{m}}'(\boldsymbol{\gamma})\bar{\mathbf{m}}(\boldsymbol{\gamma})$, is summed over six rather than four terms. We then construct

$$\hat{\boldsymbol{\Phi}} = \left\{\frac{1}{4481}\sum_{i=1}^{4481}\mathbf{m}_i(\hat{\boldsymbol{\gamma}})\mathbf{m}_i'(\hat{\boldsymbol{\gamma}})\right\} = \left\{\frac{1}{4481}\sum_{i=1}^{4481}(\hat{\varepsilon}_i\mathbf{z}_i)(\hat{\varepsilon}_i\mathbf{z}_i)'\right\}$$

where now, $\mathbf{z}_i$ in the second term is the six exogenous variables, rather than the original four (including the constant term). Thus, $\hat{\boldsymbol{\Phi}}$ is now a 6×6 moment matrix. The optimal weighting matrix for estimation (developed in the next section) is $\hat{\boldsymbol{\Phi}}^{-1}$. The GMM estimator is computed by minimizing with respect to $\boldsymbol{\gamma}$

$$q = \bar{\mathbf{m}}'(\boldsymbol{\gamma})\hat{\boldsymbol{\Phi}}^{-1}\bar{\mathbf{m}}(\boldsymbol{\gamma})$$

The asymptotic covariance matrix is computed using (13-5) as it was for the simple method of moments estimator.

Table 13.2 presents four sets of estimates, nonlinear least squares, method of moments, first step GMM and and GMM using the opptimal weighting matrix. Two comparisons are noted. The method of moments slightly different results from the nonlinear least squares estimator. This is to be expected, since they are different criteria. Judging by the standard errors, the GMM estimator seems to provide a very slight improvement over the nonlinear least squares and method of moments estimators. The conclusion, though, would seem to be that the two additional moments (variables) do not provide very much additional information for estimation of the parameters.

Table 13.2 Nonlinear Regression Estimates (Standard Errors in Parentheses)

Estimate	*Nonlinear Least Squares*	*Method of Moments*	*First Step GMM*	*GMM*
Constant	-1.69331	-1.62969	-1.45551	-1.61192
	(0.04408)	(0.04214)	(0.10102)	(0.04163)
Age	0.00207	0.00178	-0.00028	0.00092
	(0.00061)	(0.00057)	(0.00100)	(0.00056)
Education	0.04792	0.04861	0.03731	0.04647
	(0.00247)	(0.00262)	(0.00518)	(0.00262)
Female	-0.00658	0.00070	-0.02205	-0.01517
	(0.01373)	(0.01384)	(0.01445)	(0.01357)

TABLE 15.2 Estimates of the Parameters of a Gamma Distribution

Parameter	*Maximum Likelihood*	*Generalized Method of Moments*
P	2.4106	3.35894
Standard Error	(0.87682)	(0.449538)
λ	0.0770702	0.124489
Standard Error	(0.02707)	(0.029095)

15.4.3 PROPERTIES OF THE GMM ESTIMATOR

We will now examine the properties of the GMM estimator in some detail. Because the GMM estimator includes other familiar estimators that we have already encountered, including least squares (linear and nonlinear), and instrumental variables, these results will extend to those cases. The discussion given here will only sketch the elements of the formal proofs. The assumptions we make here are somewhat narrower than a fully general treatment might allow, but they are broad enough to include the situations likely to arise in practice. More detailed and rigorous treatments may be found in, for example, Newey and McFadden (1994), White (2001), Hayashi (2000), Mittelhammer et al. (2000), or Davidson (2000).

The GMM estimator is based on the set of population orthogonality conditions,

$$E[\mathbf{m}_i(\boldsymbol{\theta}_0)] = \mathbf{0},$$

where we denote the true parameter vector by $\boldsymbol{\theta}_0$. The subscript i on the term on the left-hand side indicates dependence on the observed data, $(y_i, \mathbf{x}_i, \mathbf{z}_i)$. Averaging this over the sample observations produces the sample moment equation

$$E[\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)] = \mathbf{0},$$

where

$$\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\boldsymbol{\theta}_0).$$

This moment is a set of L equations involving the K parameters. We will assume that this expectation exists and that the sample counterpart converges to it. The definitions are cast in terms of the population parameters and are indexed by the sample size. To fix the ideas, consider, once again, the empirical moment equations that define the instrumental variable estimator for a linear or nonlinear regression model.

Example 15.8 Empirical Moment Equation for Instrumental Variables

For the IV estimator in the linear or nonlinear regression model, we assume

$$E[\overline{\mathbf{m}}_n(\boldsymbol{\beta})] = E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i[y_i - h(\mathbf{x}_i, \boldsymbol{\beta})]\right] = \mathbf{0}.$$

There are L instrumental variables in $\mathbf{z}_i$ and K parameters in $\boldsymbol{\beta}$. This statement defines L moment equations, one for each instrumental variable.

We make the following assumptions about the model and these empirical moments:

ASSUMPTION 13.1. **Convergence of the Empirical Moments:** *The data generating process is assumed to meet the conditions for a law of large numbers to apply, so that we may assume that the empirical moments converge in probability to their expectation. Appendix D lists several different laws of large numbers that increase in generality. What is required for this assumption is that*

$$\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{0}.$$

The laws of large numbers that we examined in Appendix D accommodate cases of independent observations. Cases of dependent or correlated observations can be gathered under the **Ergodic theorem** (19.1). For this more general case, then, we would assume that the sequence of observations $\mathbf{m}(\boldsymbol{\theta})$ constitutes a jointly $(L \times 1)$ stationary and ergodic process.

The empirical moments are assumed to be continuous and continuously differentiable functions of the parameters. For our earlier example, this would mean that the conditional mean function, $h(\mathbf{x}_i, \boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ (although not necessarily of $\mathbf{x}_i$). With continuity and differentiability, we will also be able to assume that the derivatives of the moments,

$$\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \frac{\partial \overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'} = \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{m}_{i,n}(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'},$$

converge to a probability limit, say, $\operatorname{plim} \overline{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \overline{\mathbf{G}}(\boldsymbol{\theta}_0)$. [See (15-1), (15-5), and Theorem 15.1.] For sets of *independent* observations, the continuity of the functions and the derivatives will allow us to invoke the Slutsky theorem to obtain this result. For the more general case of sequences of *dependent* observations, Theorem 19.2, Ergodicity of Functions, will provide a counterpart to the Slutsky theorem for time series data. In sum, if the moments themselves obey a law of large numbers, then it is reasonable to assume that the derivatives do as well.

ASSUMPTION 13.2. **Identification:** *For any $n \geq K$, if $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two different parameter vectors, then there exist data sets such that $\overline{\mathbf{m}}_n(\boldsymbol{\theta}_1) \neq \overline{\mathbf{m}}_n(\boldsymbol{\theta}_2)$. Formally, in Section 14.5.3, identification is defined to imply that the probability limit of the GMM criterion function is uniquely minimized at the true parameters, $\boldsymbol{\theta}_0$.*

Assumption 13.2 is a practical prescription for identification. More formal conditions are discussed in Section 14.5.3. We have examined two violations of this crucial assumption. In the linear regression model, one of the assumptions is full rank of the matrix of exogenous variables—the absence of multicollinearity in $\mathbf{X}$. In our discussion of the maximum likelihood estimator, we will encounter a case (Example 16.1) in which a normalization is needed to identify the vector of parameters. [See Hansen et al.

14.1

(1996) for discussion of this case.] Both of these cases are included in this assumption. The identification condition has three important implications:

1. **Order condition.** The number of moment conditions is at least as large as the number of parameters, $L \geq K$. This is necessary but not sufficient for identification.
2. **Rank condition.** The $L \times K$ matrix of derivatives, $\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)$ will have row rank equal to K. (Again, note that the number of rows must equal or exceed the number of columns.)
3. **Uniqueness.** With the continuity assumption, the identification assumption implies that the parameter vector that satisfies the population moment condition is unique. We know that at the true parameter vector, $\operatorname{plim} \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) = \mathbf{0}$. If $\boldsymbol{\theta}_1$ is any parameter vector that satisfies this condition, then $\boldsymbol{\theta}_1$ must equal $\boldsymbol{\theta}_0$.

Assumptions 15.1 and 15.2 characterize the parameterization of the model. Together they establish that the parameter vector will be estimable. We now make the statistical assumption that will allow us to establish the properties of the GMM estimator.

ASSUMPTION 15.3. **Asymptotic Distribution of Empirical Moments:** *We assume that the empirical moments obey a central limit theorem. This assumes that the moments have a finite asymptotic covariance matrix,* $(1/n)\boldsymbol{\Phi}$, *so that*

$$\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}].$$

The underlying requirements on the data for this assumption to hold will vary and will be complicated if the observations comprising the empirical moment are not independent. For samples of independent observations, we assume the conditions underlying the Lindeberg–Feller (D.19) or Liapounov Central Limit theorem (D.20) will suffice. For the more general case, it is once again necessary to make some assumptions about the data. We have assumed that

$$E[\mathbf{m}_i(\boldsymbol{\theta}_0)] = \mathbf{0}.$$

If we can go a step further and assume that the functions $\mathbf{m}_i(\boldsymbol{\theta}_0)$ are an ergodic, stationary **martingale difference series,**

$$E[\mathbf{m}_i(\boldsymbol{\theta}_0) \mid \mathbf{m}_{i-1}(\boldsymbol{\theta}_0), \mathbf{m}_{i-2}(\boldsymbol{\theta}_0) \ldots] = \mathbf{0},$$

then we can invoke Theorem 19.3, the Central Limit Theorem for Martingale Difference Series. It will generally be fairly complicated to verify this assumption for nonlinear models, so it will usually be assumed outright. On the other hand, the assumptions are likely to be fairly benign in a typical application. For regression models, the assumption takes the form

$$E[\mathbf{z}_i \varepsilon_i \mid \mathbf{z}_{i-1}\varepsilon_{i-1}, \ldots] = \mathbf{0}$$

which will often be part of the central structure of the model.

13-26

Q: In text (13.6) through (13.7) are subs GMM ok italic?

With the assumptions in place, we have

THEOREM 15.2 Asymptotic Distribution of the GMM Estimator
Under the preceding assumptions,

$$\hat{\theta}_{GMM} \xrightarrow{p} \theta_0,$$

$$\hat{\theta}_{GMM} \overset{a}{\sim} N[\theta_0, \mathbf{V}_{GMM}], \tag{15-6}$$

where $\mathbf{V}_{GMM}$ *is defined in* (15-5).

We will now sketch a proof of Theorem 15.2. The GMM estimator is obtained by minimizing the criterion function

$$q_n(\theta) = \overline{\mathbf{m}}_n(\theta)'\mathbf{W}_n\overline{\mathbf{m}}_n(\theta),$$

where $\mathbf{W}_n$ is the weighting matrix used. Consistency of the estimator that minimizes this criterion can be established by the same logic that will be used for the maximum likelihood estimator. It must first be established that $q_n(\theta)$ converges to a value $q_0(\theta)$. By our assumptions of strict continuity and Assumption 15.1, $q_n(\theta_0)$ converges to 0. (We could apply the Slutsky theorem to obtain this result.) We will assume that $q_n(\theta)$ converges to $q_0(\theta)$ for other points in the parameter space as well. Because $\mathbf{W}_n$ is positive definite, for any finite n, we know that

$$0 \leq q_n(\hat{\theta}_{GMM}) \leq q_n(\theta_0). \tag{15-7}$$

That is, in the finite sample, $\hat{\theta}_{GMM}$ actually minimizes the function, so the sample value of the criterion is not larger at $\hat{\theta}_{GMM}$ than at any other value, including the true parameters. But, at the true parameter values, $q_n(\theta_0) \xrightarrow{p} 0$. So, if (15-7) is true, then it must follow that $q_n(\hat{\theta}_{GMM}) \xrightarrow{p} 0$ as well because of the identification assumption, 15.2. As $n \to \infty$, $q_n(\hat{\theta}_{GMM})$ and $q_n(\theta)$ converge to the same limit. It must be the case, then, that as $n \to \infty$, $\overline{\mathbf{m}}_n(\hat{\theta}_{GMM}) \to \overline{\mathbf{m}}_n(\theta_0)$, because the function is quadratic and $\mathbf{W}$ is positive definite. The identification condition that we assumed earlier now assures that as $n \to \infty$, $\hat{\theta}_{GMM}$ must equal θ_0. This establishes consistency of the estimator.

We will now sketch a proof of the asymptotic normality of the estimator: The first-order conditions for the GMM estimator are

$$\frac{\partial q_n(\hat{\theta}_{GMM})}{\partial \hat{\theta}_{GMM}} = 2\overline{\mathbf{G}}_n(\hat{\theta}_{GMM})'\mathbf{W}_n\overline{\mathbf{m}}_n(\hat{\theta}_{GMM}) = \mathbf{0}. \tag{15-8}$$

(The leading 2 is irrelevant to the solution, so it will be dropped at this point.) The orthogonality equations are assumed to be continuous and continuously differentiable. This allows us to employ the **mean value theorem** as we expand the empirical moments in a linear Taylor series around the true value, θ_0

$$\overline{\mathbf{m}}_n(\hat{\theta}_{GMM}) = \overline{\mathbf{m}}_n(\theta_0) + \overline{\mathbf{G}}_n(\bar{\theta})(\hat{\theta}_{GMM} - \theta_0), \tag{15-9}$$

where $\bar{\theta}$ is a point between $\hat{\theta}_{GMM}$ and the true parameters, θ_0. Thus, for each element $\bar{\theta}_k = w_k\hat{\theta}_{k,GMM} + (1 - w_k)\theta_{0,k}$ for some w_k such that $0 < w_k < 1$. Insert (15-9) in (15-8)

AU: Check subs in equations. OK italic?

to obtain

$$\bar{\mathbf{G}}_n(\hat{\theta}_{GMM})'\mathbf{W}_n\bar{\mathbf{m}}_n(\theta_0) + \bar{\mathbf{G}}_n(\hat{\theta}_{GMM})'\mathbf{W}_n\bar{\mathbf{G}}_n(\bar{\theta})(\hat{\theta}_{GMM} - \theta_0) = \mathbf{0}.$$

Solve this equation for the estimation error and multiply by $\sqrt{n}$. This produces

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) = -[\bar{\mathbf{G}}_n(\hat{\theta}_{GMM})'\mathbf{W}_n\bar{\mathbf{G}}_n(\bar{\theta})]^{-1}\bar{\mathbf{G}}_n(\hat{\theta}_{GMM})'\mathbf{W}_n\sqrt{n}\,\bar{\mathbf{m}}_n(\theta_0).$$

Assuming that they have them, the quantities on the left- and right-hand sides have the same limiting distributions. By the consistency of $\hat{\theta}_{GMM}$, we know that $\hat{\theta}_{GMM}$ and $\bar{\theta}$ both converge to θ_0. By the strict continuity assumed, it must also be the case that

$$\bar{\mathbf{G}}_n(\bar{\theta}) \xrightarrow{p} \bar{\mathbf{G}}(\theta_0) \text{ and } \bar{\mathbf{G}}_n(\hat{\theta}_{GMM}) \xrightarrow{p} \bar{\mathbf{G}}(\theta_0).$$

We have also assumed that the weighting matrix, $\mathbf{W}_n$, converges to a matrix of constants, $\mathbf{W}$. Collecting terms, we find that the limiting distribution of the vector on the left-hand side must be the same as that on the right-hand side in (15-10),

13C

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{p} \{[\bar{\mathbf{G}}(\theta_0)'\mathbf{W}\bar{\mathbf{G}}(\theta_0)]^{-1}\bar{\mathbf{G}}(\theta_0)'\mathbf{W}\}\sqrt{n}\,\bar{\mathbf{m}}_n(\theta_0). \qquad (15\text{-}10)$$

We now invoke Assumption 15.3. The matrix in curled brackets is a set of constants. The last term has the normal limiting distribution given in Assumption 15.3. The mean and variance of this limiting distribution are zero and $\mathbf{\Phi}$, respectively. Collecting terms, we have the result in Theorem 15.2, where

13C

$$\mathbf{V}_{GMM} = \frac{1}{n}[\bar{\mathbf{G}}(\theta_0)'\mathbf{W}\bar{\mathbf{G}}(\theta_0)]^{-1}\bar{\mathbf{G}}(\theta_0)'\mathbf{W}\mathbf{\Phi}\mathbf{W}\bar{\mathbf{G}}(\theta_0)[\bar{\mathbf{G}}(\theta_0)'\mathbf{W}\bar{\mathbf{G}}(\theta_0)]^{-1}. \qquad (15\text{-}11)$$

The final result is a function of the choice of weighting matrix, $\mathbf{W}$. If the optimal weighting matrix, $\mathbf{W} = \mathbf{\Phi}^{-1}$, is used, then the expression collapses to

$$\mathbf{V}_{GMM,optimal} = \frac{1}{n}[\bar{\mathbf{G}}(\theta_0)'\mathbf{\Phi}^{-1}\bar{\mathbf{G}}(\theta_0)]^{-1}. \qquad (15\text{-}12)$$

Returning to (15-11), there is a special case of interest. If we use least squares or instrumental variables with $\mathbf{W} = \mathbf{I}$, then

$$\mathbf{V}_{GMM} = \frac{1}{n}(\bar{\mathbf{G}}'\bar{\mathbf{G}})^{-1}\bar{\mathbf{G}}'\mathbf{\Phi}\bar{\mathbf{G}}(\bar{\mathbf{G}}'\bar{\mathbf{G}})^{-1}.$$

This equation prescribes essentially the White or **Newey-West estimator,** which returns us to our departure point and provides a neat symmetry to the GMM principle. We will formalize this in Section 15.6.1.

15.5 TESTING HYPOTHESES IN THE GMM FRAMEWORK

The estimation framework developed in the previous section provides the basis for a convenient set of statistics for testing hypotheses. We will consider three groups of tests. The first is a pair of statistics that is used for testing the validity of the restrictions that produce the moment equations. The second is a trio of tests that correspond to the familiar Wald, LM, and LR tests. The third is a class of tests based on the theoretical underpinnings of the conditional moments that we used earlier to devise the GMM estimator.

15.5.1 TESTING THE VALIDITY OF THE MOMENT RESTRICTIONS

In the exactly identified cases we examined earlier (least squares, instrumental variables, maximum likelihood), the criterion for GMM estimation

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})'\mathbf{W}\bar{\mathbf{m}}(\boldsymbol{\theta})$$

would be exactly zero because we can find a set of estimates for which $\bar{\mathbf{m}}(\boldsymbol{\theta})$ is exactly zero. Thus in the exactly identified case when there are the same number of moment equations as there are parameters to estimate, the weighting matrix $\mathbf{W}$ is irrelevant to the solution. But if the parameters are overidentified by the moment equations, then these equations imply substantive restrictions. As such, if the hypothesis of the model that led to the moment equations in the first place is incorrect, at least some of the sample moment restrictions will be systematically violated. This conclusion provides the basis for a test of the **overidentifying restrictions.** By construction, when the optimal weighting matrix is used,

$$nq = \left[\sqrt{n}\,\bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})'\right]\left\{\text{Est. Asy. Var}\left[\sqrt{n}\,\bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})\right]\right\}^{-1}\left[\sqrt{n}\,\bar{\mathbf{m}}(\hat{\boldsymbol{\theta}})\right],$$

so nq is a Wald statistic. Therefore, under the hypothesis of the model,

$$nq \xrightarrow{d} \chi^2[L-K].$$

(For the exactly identified case, there are zero degrees of freedom and $q = 0$.)

Example 15.9 Overidentifying Restrictions

In Hall's consumption model, two orthogonality conditions noted in Example 18.1 exactly identify the two parameters. But his analysis of the model suggests a way to test the specification. The conclusion, "No information available in time t apart from the level of consumption, c_t, helps predict future consumption, c_{t+1}, in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods t or earlier are irrelevant once c_t is known" suggests how one might test the model. If lagged values of income (Y_t might equal the ratio of current income to the previous period's income) are added to the set of instruments, then the model is now overidentified by the orthogonality conditions;

$$E_t\left[\left(\beta(1+r_{t+1})R_{t+1}^{\lambda}-1\right)\times\begin{pmatrix}1\\ R_t\\ Y_{t-1}\\ Y_{t-2}\end{pmatrix}\right]=\begin{bmatrix}0\\0\end{bmatrix}.$$

A simple test of the overidentifying restrictions would be suggestive of the validity of the corollary. Rejecting the restrictions casts doubt on the original model. Hall's proposed tests to distinguish the life cycle–permanent income model from other theories of consumption involved adding two lags of income to the information set. Hansen and Singleton (1982) operated directly on this form of the model. Other studies, for example, Campbell and Mankiw's (1989) as well as Hall's, used the model's implications to formulate more conventional instrumental variable regression models.

The preceding is a **specification test,** not a test of parametric restrictions. However, there is a symmetry between the moment restrictions and restrictions on the parameter vector. Suppose $\boldsymbol{\theta}$ is subjected to J restrictions (linear or nonlinear) which restrict the number of free parameters from K to $K-J$. (That is, reduce the dimensionality of the parameter space from K to $K-J$.) The nature of the GMM estimation problem

we have posed is not changed at all by the restrictions. The constrained problem may be stated in terms of

$$q_R = \bar{\mathbf{m}}(\boldsymbol{\theta}_R)'\mathbf{W}\bar{\mathbf{m}}(\boldsymbol{\theta}_R).$$

Note that the weighting matrix, $\mathbf{W}$, is unchanged. The precise nature of the solution method may be changed—the restrictions mandate a constrained optimization. However, the criterion is essentially unchanged. It follows then that

$$nq_R \xrightarrow{d} \chi^2[L-(K-J)].$$

This result suggests a method of testing the restrictions, although the distribution theory is not obvious. The weighted sum of squares with the restrictions imposed, nq_R, must be larger than the weighted sum of squares obtained without the restrictions, nq. The difference is

$$(nq_R - nq) \xrightarrow{d} \chi^2[J]. \tag{15-13}$$

The test is attributed to Newey and West (1987b). This provides one method of testing a set of restrictions. (The small-sample properties of this test will be the central focus of the application discussed in Section 15.6.5.) We now consider several alternatives.

15.5.2 GMM COUNTERPARTS TO THE WALD, LM, AND LR TESTS

Section 14.6 describes a trio of testing procedures that can be applied to a hypothesis in the context of maximum likelihood estimation. To reiterate, let the hypothesis to be tested be a set of J possibly nonlinear restrictions on K parameters $\boldsymbol{\theta}$ in the form $H_0: \mathbf{r}(\boldsymbol{\theta}) = \mathbf{0}$. Let $\mathbf{c}_1$ be the maximum likelihood estimates of $\boldsymbol{\theta}$ estimated without the restrictions, and let $\mathbf{c}_0$ denote the restricted maximum likelihood estimates, that is, the estimates obtained while imposing the null hypothesis. The three statistics, which are asymptotically equivalent, are obtained as follows:

$$\text{LR} = \text{likelihood ratio} = -2(\ln L_0 - \ln L_1),$$

where

$$\ln L_j = \log \text{ likelihood function evaluated at } \mathbf{c}_j, \quad j = 0, 1.$$

The **likelihood ratio statistic** requires that both estimates be computed. The Wald statistic is

$$W = \text{Wald} = [\mathbf{r}(\mathbf{c}_1)]'\{\text{Est. Asy. Var}[\mathbf{r}(\mathbf{c}_1)]\}^{-1}[\mathbf{r}(\mathbf{c}_1)]. \tag{15-14}$$

The **Wald statistic** is the distance measure for the degree to which the unrestricted estimator fails to satisfy the restrictions. The usual estimator for the asymptotic covariance matrix would be

$$\text{Est. Asy. Var}[\mathbf{r}(\mathbf{c}_1)] = \mathbf{R}_1\{\text{Est. Asy. Var}[\mathbf{c}_1]\}\mathbf{R}_1', \tag{15-15}$$

where

$$\mathbf{R}_1 = \partial \mathbf{r}(\mathbf{c}_1)/\partial \mathbf{c}_1' \quad (\mathbf{R}_1 \text{ is a } J \times K \text{ matrix}).$$