18

THE GENERALIZED METHOD OF MOMENTS

18.1 INTRODUCTION

The **maximum likelihood estimator** 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 which 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 which, as will be clear shortly, encompasses nearly all the familiar estimators discussed in this book. Section 18.2 will introduce the estimation framework with the method of moments. Formalities of the GMM estimator are presented in Section 18.3. Section 18.4 discusses hypothesis testing based on moment equations. A major applications, dynamic panel data models, is described in Section 18.5.

Example 18.1 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:

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

The information available at time t is denoted Ω_t so that E_t denotes the expectation formed at time t based on information set Ω_t. The maximand is the expected discounted stream of future 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, $E_t[w_{t+\tau} \mid \Omega_t]$. 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

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

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. In order 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

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

Hall assumed that r was constant over time. Other applications of this modeling framework [e.g., 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.

18.2 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 argued, 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, $\bar{m}_2' = (1/n)\sum_{i=1}^{n} y_i^2$ will converge in mean square to the variance plus the square of the mean of the distribution of 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, $\bar{m}_1, \ldots, \bar{m}_K$, 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 Lindberg–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 18.3.

18.2.1 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 sample consists of n observations, $y_1, \ldots, y_n$. The kth "raw" or **uncentered moment** is

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

By Theorem D.1,

$$E[\bar{m}'_k] = \mu'_k = E\left[y_i^k\right]$$

and

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

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

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

Finally, by the Lindberg–Levy Central Limit Theorem,

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

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 18.2 Method of Moments Estimator for N[μ, σ^2]

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

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

and

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

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

$$\hat{\mu} = \bar{m}'_1 = \bar{y}$$

and

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

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

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

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

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

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,**

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

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

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

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 18.3 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 $(1/n)\sum_{i=1}^{n}(1/y_i)$. Since 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 18.4 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

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

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

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

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 $\bar{m}_k'$ as an estimator of μ_k' apply as well to $\bar{m}_k$ 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

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

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

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

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.)

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

DEFINITION 18.1 Exponential Family

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

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

where $a(\cdot)$, $b(\cdot)$, $c(\cdot)$, and $s(\cdot)$ 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. Since the normal distribution is an exponential family with sufficient statistics $\bar{m}_1'$ and $\bar{m}_2'$, the estimators described in Example 18.2 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 18.3 is an exponential family. You should be able to show that the sufficient statistics are the ones that are suggested in Example 18.3 as the bases for the MLEs of μ and λ.

Example 18.5 Gamma Distribution

The gamma distribution (see Section C.4.5) is

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

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 $a(\mathbf{data}) = 0$, $b(\boldsymbol{\theta}) = n[P \ln \lambda - \ln \Gamma(P)]$ 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

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

(The functions $\Gamma(P)$ and $\Psi(P) = \mathrm{d} \ln \Gamma(P)/\mathrm{d}P$ are discussed in Section E.5.3.) 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 21 is that if the statistics, $c_k(\mathbf{data})$ are sufficient statistics, then the conditional density $f[y_1, \ldots, y_n \mid c_k(\mathbf{data}), k = 1, \ldots, K]$ is not a function of the parameters.

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

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

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

$$(\hat{P}, \hat{\lambda}) = \begin{bmatrix} & \bar{m}_1' & \bar{m}_2' & \bar{m}_{-1}' \\ \bar{m}_2' & 2.05682, 0.065759 & & \\ \bar{m}_{-1}' & 2.77198, 0.0886239 & 2.60905, 0.0800475 & \\ \bar{m}_*' & 2.4106, 0.0770702 & 2.26450, 0.071304 & 3.03580, 0.1018202 \end{bmatrix}.$$

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

18.2.2 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 18.4, the underlying sample statistic is

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

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

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

In general, when the moments are computed as

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

where $\mathbf{y}_i$ is an observation on a vector of variables, an appropriate estimator of the asymptotic covariance matrix of $[\bar{m}_1, \ldots, \bar{m}_k]$ can be computed using

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

(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 our assumption of iid random sampling from a distribution with finite moments up to $2K$, $\mathbf{F}$ will converge in probability to the appropriate covariance matrix of the normalized vector of moments, $\mathbf{\Phi} = \text{Asy.Var}[\sqrt{n}\bar{\mathbf{m}}_n(\boldsymbol{\theta})]$. Finally, under our assumptions of random sampling, though the precise distribution is likely to be unknown, we can appeal to the Lindberg–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 $\bar{\mathbf{G}}_n(\boldsymbol{\theta})$ be the $K \times K$ matrix whose kth row is the vector of partial derivatives

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

Now, expand the set of solved moment equations around the true values of the parameters $\boldsymbol{\theta}_0$ in a linear **Taylor series.** The linear approximation is

$$\mathbf{0} \approx [\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)] + \bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0).$$

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx -[\bar{\mathbf{G}}'_n(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}[\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)]. \tag{18-1}$$

(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 below.) The argument needed to characterize the large sample behavior of the estimator, $\hat{\boldsymbol{\theta}}$, are discussed in Appendix D. We have from Theorem D.18 (the Central Limit Theorem) that $\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)$ 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 $\bar{\mathbf{G}}_n(\boldsymbol{\theta}_0)$ to converge to a nonsingular matrix of constants, $\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)$. Under general conditions, the limiting distribution of the right hand side of (18-1) will be that of a linear function of a normally distributed vector. Jumping to the conclusion, we expect the asymptotic distribution of $\hat{\boldsymbol{\theta}}$ to be normal with mean vector $\boldsymbol{\theta}_0$ and covariance matrix $(1/n) \times \{-[\boldsymbol{\Gamma}'(\boldsymbol{\theta}_0)]^{-1}\}\mathbf{\Phi}\{-[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1}\}$. Thus, the asymptotic covariance matrix for the method of moments estimator may be estimated with

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

Example 18.5 (Continued)
Using the estimates $\hat{\boldsymbol{\theta}}(m'_1, m'_*) = (2.4106,\ 0.0770702)$,

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

[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 1/19 as the divisor),

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

The product is

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

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

$$\frac{1}{n}[-\mathbf{H}]^{-1} = \frac{1}{n}\begin{bmatrix} \Psi' & -1/\lambda \\ -1/\lambda & P/\lambda^2 \end{bmatrix}^{-1} = \begin{bmatrix} 0.51203 & 0.01637 \\ 0.01637 & 0.00064654 \end{bmatrix}.$$

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 estimates above. The difference would be due to sampling variability in a finite sample and the presence of $\mathbf{F}$ in the first variance estimator.

18.2.3 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. 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 18.3, 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.

18.3 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.

Before continuing, it will be useful for you to read (or reread) the following sections:

1. Consistent Estimation: The Method of Moments: Section 18.2,
2. Correlation Between $\mathbf{x}_i$ and ε_i: Instrumental Variables Estimation, Section 5.4,

[2] Ψ' is the digamma 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.

3. GMM Estimation in the Generalized Regression Model: Sections 10.4, 11.3, and 12.6,
4. Nonlinear Regression Models, Chapter 9,
5. Optimization, Section E.5,
6. **Robust Estimation** of Asymptotic Covariance Matrices, Section 10.3,
7. The Wald Test, Theorem 6.1,
8. GMM Estimation of Dynamic Panel Data Models, Section 13.6.

The GMM estimation technique is an extension of the method of moments technique described in Section 18.2.[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 18.2.

18.3.1 ESTIMATION BASED ON ORTHOGONALITY CONDITIONS

Estimation by the method of moments proceeds as follows. The model specified for the random variable y_i implies certain expectations, for example

$$E[y_i] = \mu,$$

where μ is the mean of the distribution of y_i. Estimation of μ then proceeds by forming a sample analog to the population expectation:

$$E[y_i - \mu] = 0.$$

The sample counterpart to this expectation is the **empirical moment equation,**

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu}) = 0.$$

The estimator is the value of $\hat{\mu}$ that satisfies the sample moment equation. The example given is, of course, a trivial one. Example 18.5 describes a more elaborate case of sampling from a gamma distribution. The moment conditions used for estimation in that example (taken two at a time from a set of four) include

$$E[y_i - P/\lambda] = 0$$

and

$$E[\ln y_i - \Psi(P) + \ln\lambda] = 0.$$

(These two coincide with the terms in the likelihood equations for this model.) Inserting the sample data into the sample analogs produces the moment equations for estimation:

$$\frac{1}{n}\sum_{i=1}^{n}[y_i - \hat{P}/\hat{\lambda}] = 0$$

[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 McFadden and Newey (1994). Useful summaries of GMM estimation and other developments in econometrics is 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 (1993). Ruud (2000) provides many of the theoretical details. Hayashi (2000) is another extensive treatment of estimation centered on GMM estimators.

and

$$\frac{1}{n}\sum_{i=1}^{n}[\ln y_i - \Psi(\hat{P}) + \ln \hat{\lambda}] = 0.$$

Example 18.6 Orthogonality Conditions

Assuming that households are forecasting interest rates as well as earnings, Hall's consumption model with the corollary implies the following orthogonality conditions:

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

Now, consider the apparently different case of 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 Section 5.4, 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

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

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 5.4 for further details.)

The nonlinear least squares estimator was defined similarly, though in this case, the normal equations are more complicated since 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

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

All the maximum likelihood estimators that we have looked at thus far and will encounter later are obtained by equating the derivatives of a log-likelihood to zero. The

scaled log-likelihood function is

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

where $f(\cdot)$ is the density function and $\boldsymbol{\theta}$ is the parameter vector. For densities that satisfy the regularity conditions [see Section 17.4.1],

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

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

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

(Dividing by n to make this result comparable with 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 nearly all the estimators defined in this book can be viewed as method of moments estimators.

18.3.2 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 18.5, 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 $\boldsymbol{\theta} = (P, \lambda)$.

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, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, and that the theory provides a set of $L > K$ moment conditions,

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

where y_i, $\mathbf{x}_i$, and $\mathbf{z}_i$ are variables that appear in the model and the subscript i on $m_{il}(\boldsymbol{\theta})$

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

indicates the dependence on $(y_i, \mathbf{x}_i, \mathbf{z}_i)$. Denote the corresponding sample means as

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

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

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

will not have a unique solution.[5] It will be necessary to reconcile the $\binom{L}{K}$ different sets of estimates that can be produced. One possibility is to minimize a criterion function, such as the sum of squares,

$$q = \sum_{l=1}^{L} \bar{m}_l^2 = \bar{\mathbf{m}}(\boldsymbol{\theta})'\bar{\mathbf{m}}(\boldsymbol{\theta}).^{6} \tag{18-2}$$

It can be shown [see, e.g., Hansen (1982)] that under the assumptions we have made so far, specifically that $\operatorname{plim} \bar{\mathbf{m}}(\boldsymbol{\theta}) = E[\bar{\mathbf{m}}(\boldsymbol{\theta})] = \mathbf{0}$, minimizing q in (18-2) produces a consistent (albeit, as we shall see, possibly inefficient) estimator of $\boldsymbol{\theta}$. We can, in fact, use as the criterion a weighted sum of squares,

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

where $\mathbf{W}_n$ is *any* positive definite matrix that may depend on the data but is not a function of $\boldsymbol{\theta}$, such as $\mathbf{I}$ in (18-2), to produce a consistent estimator of $\boldsymbol{\theta}$.[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 $\operatorname{plim} \mathbf{W}_n =$ 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** procedure would minimize

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})'\boldsymbol{\Phi}^{-1}\bar{\mathbf{m}}(\boldsymbol{\theta}). \tag{18-3}$$

[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 18.3.

[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 we discussed in Example 17.1. In practical terms, this is likely to be a consideration only in a fairly small class of cases.

In general, the L elements of $\bar{\mathbf{m}}$ are freely correlated. In (18-3), we have used a diagonal $\mathbf{W}$ that ignores this correlation. To use generalized least squares, we would define the full matrix,

$$\mathbf{W} = \left\{\text{Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}]\right\}^{-1} = \boldsymbol{\Phi}^{-1}. \tag{18-4}$$

The estimators defined by choosing $\boldsymbol{\theta}$ to minimize

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

are **minimum distance estimators.** 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] Since the OLS criterion in (18-2) 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 (18-4) 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

$$\mathbf{V}_{GMM} = \frac{1}{n}[\boldsymbol{\Gamma}'\mathbf{W}\boldsymbol{\Gamma}]^{-1} = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}]^{-1}, \tag{18-5}$$

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 18.3.3.

Example 18.7 GMM Estimation of the Parameters of a Gamma Distribution

Referring once again to our earlier results in Example 18.5, 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

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

[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 the book.

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 $\bar{m}_4(P, \lambda)$.

For our initial set of estimates, we will use ordinary least squares. The optimization problem is

$$\text{Minimize}_{P,\lambda} \sum_{l=1}^{4}\bar{m}_l(P, \lambda)^2 = \sum_{l=1}^{4}[\bar{y}_l - \mu_l(P, \lambda)]^2 = \bar{\mathbf{m}}(P, \lambda)'\bar{\mathbf{m}}(P, \lambda).$$

This estimator will be the **minimum distance estimator** with $\mathbf{W} = \mathbf{I}$. This nonlinear optimization problem must be solved iteratively. As starting values for the iterations, we used the maximum likelihood estimates from Example 18.5, $\hat{P}_{\text{ML}} = 2.4106$ and $\hat{\lambda}_{\text{ML}} = 0.0770702$. The least squares values that result from this procedure are $\hat{P} = 2.0582996$ and $\hat{\lambda} = 0.06579888$. We can now use these to form our estimate of $\mathbf{W}$. GMM estimation usually requires a first-step estimation such as this one to obtain the weighting matrix $\mathbf{W}$. With these new estimates in hand, we obtained

$$\hat{\mathbf{\Phi}} = \left\{ \frac{1}{20}\sum_{i=1}^{20} \begin{bmatrix} y_{i1} - \hat{P}/\hat{\lambda} \\ y_{i2} - \hat{P}(\hat{P}+1)/\hat{\lambda}^2 \\ y_{i3} - \Psi(\hat{P}) + \ln\hat{\lambda} \\ y_{i4} - \hat{\lambda}/(\hat{P}-1) \end{bmatrix} \begin{bmatrix} y_{i1} - \hat{P}/\hat{\lambda} \\ y_{i2} - \hat{P}(\hat{P}+1)/\hat{\lambda}^2 \\ y_{i3} - \Psi(\hat{P}) + \ln\hat{\lambda} \\ y_{i4} - \hat{\lambda}/(\hat{P}-1) \end{bmatrix}' \right\}.$$

(Note, we could have computed $\hat{\mathbf{\Phi}}$ using the maximum likelihood estimates.) The GMM estimator is now obtained by minimizing

$$q = \bar{\mathbf{m}}(P, \lambda)'\hat{\mathbf{\Phi}}^{-1}\bar{\mathbf{m}}(P, \lambda).$$

The two estimates are $\hat{P}_{GMM} = 3.35894$ and $\hat{\lambda}_{GMM} = 0.124489$. At these two values, the value of the function is $q = 1.97522$. To obtain an asymptotic covariance matrix for the two estimates, we first recompute $\hat{\mathbf{\Phi}}$ as shown above;

$$\frac{1}{20}\hat{\mathbf{\Phi}} = \begin{bmatrix} 24.7051 & & & \\ 2307.126 & 229{,}609.5 & & \\ 0.6974 & 58.8148 & 0.0230 & \\ -0.0283 & -2.1423 & -0.0011 & 0.000065413 \end{bmatrix}.$$

To complete the computation, we will require the derivatives matrix,

$$\bar{\mathbf{G}}'(\theta) = \begin{bmatrix} \partial\bar{m}_1/\partial P & \partial\bar{m}_2/\partial P & \partial\bar{m}_3/\partial P & \partial\bar{m}_4/\partial P \\ \partial\bar{m}_1/\partial\lambda & \partial\bar{m}_2/\partial\lambda & \partial\bar{m}_3/\partial\lambda & \partial\bar{m}_4/\partial\lambda \end{bmatrix}$$

$$= \begin{bmatrix} -1/\lambda & -(2P+1)/\lambda^2 & -\Psi'(P) & \lambda/(P-1)^2 \\ P/\lambda^2 & 2P(P+1)/\lambda^3 & 1/\lambda & -1/(P-1) \end{bmatrix}.$$

$$\bar{\mathbf{G}}'(\hat{\theta}) = \begin{bmatrix} -8.0328 & -498.01 & -0.34635 & 0.022372 \\ 216.74 & 15178.2 & 8.0328 & -0.42392 \end{bmatrix}.$$

Finally,

$$\frac{1}{20}[\hat{\mathbf{G}}'\hat{\mathbf{\Phi}}^{-1}\hat{\mathbf{G}}]^{-1} = \begin{bmatrix} 0.202201 & 0.0117344 \\ 0.0117344 & 0.000867519 \end{bmatrix}$$

TABLE 18.1 Estimates of the Parameters of a Gamma Distribution

Parameter	*Maximum Likelihood*	*Generalized Method of Moments*
P	2.4106	3.3589
Standard Error	(0.87683)	(0.449667)
λ	0.0770701	0.12449
Standard Error	(0.02707)	(0.029099)

gives the estimated asymptotic covariance matrix for the estimators. Recall that in Example 18.5, we obtained maximum likelihood estimates of the same parameters. Table 18.1 summarizes.

Looking ahead, we should have expected the GMM estimator to improve the standard errors. The fact that it does for P but not for λ might cast some suspicion on the specification of the model. In fact, the data generating process underlying these data is not a gamma population—the values were hand picked by the author. Thus, the findings in Table 18.1 might not be surprising. We will return to this issue in Section 18.4.1.

18.3.3 PROPERTIES OF THE GMM ESTIMATOR

We will now examine the properties of the GMM estimator in some detail. Since the GMM estimator includes other familiar estimators that we have already encountered, including least squares (linear and nonlinear), instrumental variables, and maximum likelihood, 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). This development will continue the analysis begun in Section 10.4 and add some detail to the formal results of Section 16.5.

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 right 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

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

where

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

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 which define the instrumental variable estimator for a linear or nonlinear regression model.

Example 18.8 Empirical Moment Equation for Instrumental Variables
For the IV estimator in the linear or nonlinear regression model, we assume

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

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 18.1. **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*

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

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** (12.1). For this more general case, then, we would assume that the sequence of observations $\mathbf{m}(\boldsymbol{\theta})$ constant 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 example above, this would mean that the conditional mean function, $h(\mathbf{x}_i, \boldsymbol{\beta})$ is a continuous function of $\boldsymbol{\beta}$ (though not necessarily of $\mathbf{x}_i$).

With continuity and differentiability, we also will be able to assume that the derivatives of the moments,

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

converge to a probability limit, say $\operatorname{plim} \bar{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \bar{\mathbf{G}}(\boldsymbol{\theta}_0)$. 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 12.2, 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 18.2. **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 $\bar{\mathbf{m}}_n(\boldsymbol{\theta}_1) \neq \bar{\mathbf{m}}_n(\boldsymbol{\theta}_2)$. Formally, in Section 16.5.3, 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 18.2 is a practical prescription for identification. More formal conditions are discussed in Section 16.5.3. 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 encountered a case (Example 17.2) in which the a normalization was needed to identify the vector of parameters. [See Hansen et al. (1996) for discussion of this case.] Both of these cases are included in this assumption. The identification condition has three important implications:

Order Condition The number of moment conditions is at least as large as the number of parameter; $L \geq K$. This is necessary but not sufficient for identification.

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.)

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 18.1 and 18.2 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 18.3. **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 Lindberg–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 12.3, 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.

With the assumptions in place, we have

THEOREM 18.1 Asymptotic Distribution of the GMM Estimator
Under the preceding assumptions,

$$\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{p} \theta$$

$$\hat{\boldsymbol{\theta}}_{GMM} \overset{a}{\sim} N[\boldsymbol{\theta}, \mathbf{V}_{GMM}], \tag{18-6}$$

where $\mathbf{V}_{GMM}$ *is defined in (18-5).*

We will now sketch a proof of Theorem 18.1. The GMM estimator is obtained by minimizing the criterion function

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

where $\mathbf{W}_n$ is the weighting matrix used. Consistency of the estimator that minimizes this criterion can be established by the same logic we used for the maximum likelihood estimator. It must first be established that $q_n(\boldsymbol{\theta})$ converges to a value $q_0(\boldsymbol{\theta})$. By our assumptions of strict continuity and Assumption 18.1, $q_n(\boldsymbol{\theta}_0)$ converges to 0. (We could apply the Slutsky theorem to obtain this result.) We will assume that $q_n(\boldsymbol{\theta})$ converges to $q_0(\boldsymbol{\theta})$ for other points in the parameter space as well. Since $\mathbf{W}_n$ is positive definite, for any finite n, we know that

$$0 \leq q_n(\hat{\boldsymbol{\theta}}_{GMM}) \leq q_n(\boldsymbol{\theta}_0). \tag{18-7}$$

That is, in the finite sample, $\hat{\boldsymbol{\theta}}_{GMM}$ actually minimizes the function, so the sample value of the criterion is not larger at $\hat{\boldsymbol{\theta}}_{GMM}$ than at any other value, including the true parameters. But, at the true parameter values, $q_n(\boldsymbol{\theta}_0) \xrightarrow{p} 0$. So, if (18-7) is true, then it must follow that $q_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{p} 0$ as well because of the identification assumption, 18.2. As $n \to \infty$, $q_n(\hat{\boldsymbol{\theta}}_{GMM})$ and $q_n(\boldsymbol{\theta})$ converge to the same limit. It must be the case, then, that as $n \to \infty$, $\bar{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) \to \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0)$, since the function is quadratic and $\mathbf{W}$ is positive definite. The identification condition that we assumed earlier now assures that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_{GMM}$ must equal $\boldsymbol{\theta}_0$. 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

$$\frac{\partial q_n(\hat{\boldsymbol{\theta}}_{GMM})}{\partial \hat{\boldsymbol{\theta}}_{GMM}} = 2\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\bar{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \mathbf{0}. \tag{18-8}$$

(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. $\boldsymbol{\theta}$;

$$\bar{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) + \bar{\mathbf{G}}_n(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0), \tag{18-9}$$

where $\bar{\boldsymbol{\theta}}$ is a point between $\hat{\boldsymbol{\theta}}_{GMM}$ and the true parameters, $\boldsymbol{\theta}_0$. 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 (18-9) in (18-8) to obtain

$$\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0) + \bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\bar{\mathbf{G}}_n(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) = \mathbf{0}.$$

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

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) = -[\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\bar{\mathbf{G}}_n(\bar{\boldsymbol{\theta}})]^{-1}\bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0).$$

Assuming that they have them, the quantities on the left- and right-hand sides have the same limiting distributions. By the consistency of $\hat{\boldsymbol{\theta}}_{GMM}$ we know that $\hat{\boldsymbol{\theta}}_{GMM}$ and $\bar{\boldsymbol{\theta}}$ both converge to $\boldsymbol{\theta}_0$. By the strict continuity assumed, it must also be the case that

$$\bar{\mathbf{G}}_n(\bar{\boldsymbol{\theta}}) \xrightarrow{p} \bar{\mathbf{G}}(\boldsymbol{\theta}_0) \text{ and } \bar{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{p} \bar{\mathbf{G}}(\boldsymbol{\theta}_0).$$

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 right hand side must be the same as that on the right hand side in (18-10),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \xrightarrow{p} \{[\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\}\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\theta}_0). \tag{18-10}$$

We now invoke Assumption 18.3. The matrix in curled brackets is a set of constants. The last term has the normal limiting distribution given in Assumption 18.3. The mean and variance of this limiting distribution are zero and $\boldsymbol{\Phi}$, respectively. Collecting terms, we have the result in Theorem 18.1, where

$$V_{GMM} = \frac{1}{n}[\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)[\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}. \tag{18-11}$$

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

$$\mathbf{V}_{GMM,optimal} = \frac{1}{n}[\bar{\mathbf{G}}(\boldsymbol{\theta}_0)'\boldsymbol{\Phi}^{-1}\bar{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}. \tag{18-12}$$

Returning to (18-11), there is a special case of interest. If we use least squares or instrumental variables with $\mathbf{W} = \mathbf{I}$, then

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

This equation is essentially (10-23) to (10-24), the White or **Newey-West estimator,** which returns us to our departure point and provides a neat symmetry to the GMM principle.

18.3.4 GMM ESTIMATION OF SOME SPECIFIC ECONOMETRIC MODELS

Suppose that the theory specifies a relationship

$$y_i = h(\mathbf{x}_i, \boldsymbol{\beta}) + \varepsilon_i,$$

where $\boldsymbol{\beta}$ is a $K \times 1$ parameter vector that we wish to estimate. This may not be a regression relationship, since it is possible that

$$\text{Cov}[\varepsilon_i, h(\mathbf{x}_i, \boldsymbol{\beta})] \neq 0,$$

or even

$$\text{Cov}[\varepsilon_i, \mathbf{x}_j] \neq \mathbf{0} \quad \text{for all } i \text{ and } j.$$

Consider, for example, a model that contains lagged dependent variables and autocorrelated disturbances. (See Section 12.9.4.) For the present, we assume that

$$E[\boldsymbol{\varepsilon} \mid \mathbf{X}] \neq \mathbf{0}$$

and

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}' \mid \mathbf{X}] = \sigma^2\boldsymbol{\Omega} = \boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is symmetric and positive definite but otherwise unrestricted. The disturbances may be heteroscedastic and/or autocorrelated. But for the possibility of correlation between regressors and disturbances, this model would be a generalized, possibly nonlinear, regression model. Suppose that at each observation i we observe a vector of L variables, $\mathbf{z}_i$, such that $\mathbf{z}_i$ is uncorrelated with ε_i. You will recognize $\mathbf{z}_i$ as a set of **instrumental variables.** The assumptions thus far have implied a set of **orthogonality conditions,**

$$E[\mathbf{z}_i\varepsilon_i \mid \mathbf{x}_i] = \mathbf{0},$$

which may be sufficient to identify (if $L = K$) or even overidentify (if $L > K$) the parameters of the model.

For convenience, define

$$\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}}) = y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}), \quad i = 1, \ldots, n,$$

and

$$\mathbf{Z} = n \times L \text{ matrix whose ith row is } \mathbf{z}_i'.$$

By a straightforward extension of our earlier results, we can produce a GMM estimator of $\boldsymbol{\beta}$. The sample moments will be

$$\bar{\mathbf{m}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i e(\mathbf{x}_i, \boldsymbol{\beta}) = \frac{1}{n}\mathbf{Z}'\mathbf{e}(\mathbf{X}, \boldsymbol{\beta}).$$

The minimum distance estimator will be the $\hat{\boldsymbol{\beta}}$ that minimizes

$$q = \bar{\mathbf{m}}_n(\hat{\boldsymbol{\beta}})'\mathbf{W}\bar{\mathbf{m}}_n(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n}[\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})'\mathbf{Z}]\right)\mathbf{W}\left(\frac{1}{n}[\mathbf{Z}'\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})]\right) \tag{18-13}$$

for some choice of $\mathbf{W}$ that we have yet to determine. The criterion given above produces the **nonlinear instrumental variable estimator.** If we use $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$, then we have exactly the estimation criterion we used in Section 9.5.1 where we defined the nonlinear instrumental variables estimator. Apparently (18-13) is more general, since we are not limited to this choice of $\mathbf{W}$. The linear IV estimator is a special case. For any given choice of $\mathbf{W}$, as long as there are enough orthogonality conditions to identify the parameters, estimation by minimizing q is, at least in principle, a straightforward problem in nonlinear optimization. Hansen (1982) showed that the optimal choice of $\mathbf{W}$ for this estimator is

$$\begin{aligned}\mathbf{W}_{\text{GMM}} &= \{\text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}_n(\boldsymbol{\beta})]\}^{-1} \\ &= \left\{\text{Asy. Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right]\right\}^{-1} = \left\{\text{Asy. Var}\left[\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{e}(\mathbf{X}, \boldsymbol{\beta})\right]\right\}^{-1}.\end{aligned} \tag{18-14}$$

For our model, this is

$$\mathbf{W} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\text{Cov}[\mathbf{z}_i\varepsilon_i, \mathbf{z}_j\varepsilon_j] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma_{ij}\mathbf{z}_i\mathbf{z}_j' = \frac{\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}}{n}.$$

If we insert this result in (18-13), we obtain the criterion for the GMM estimator:

$$q = \left[\left(\frac{1}{n}\right)\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})'\mathbf{Z}\right]\left(\frac{\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}}{n}\right)^{-1}\left[\left(\frac{1}{n}\right)\mathbf{Z}'\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})\right].$$

There is a possibly difficult detail to be considered. The GMM estimator involves

$$\frac{1}{n}\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{z}_i\mathbf{z}_j'\text{Cov}[\varepsilon_i\varepsilon_j] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\mathbf{z}_i\mathbf{z}_j'\text{Cov}[(y_i - h(\mathbf{x}_i, \boldsymbol{\beta}))(y_j - h(\mathbf{x}_j, \boldsymbol{\beta}))].$$

The conditions under which such a double sum might converge to a positive definite matrix are sketched in Sections 5.3.2 and 12.4.1. Assuming that they do hold, estimation appears to require that an estimate of $\boldsymbol{\beta}$ be in hand already, even though it is the object of estimation. It may be that a consistent but inefficient estimator of $\boldsymbol{\beta}$ is available. Suppose for the present that one is. If observations are uncorrelated, then the cross observations terms may be omitted, and what is required is

$$\frac{1}{n}\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'\text{Var}[(y_i - h(\mathbf{x}_i, \beta))].$$

We can use the White (1980) estimator discussed in Section 11.2.2 and 11.3 for this case:

$$\mathbf{S}_0 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'(y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))^2. \tag{18-15}$$

If the disturbances are autocorrelated but the process is stationary, then Newey and West's (1987a) estimator is available (assuming that the autocorrelations are sufficiently small at a reasonable lag, p):

$$\mathbf{S} = \left[\mathbf{S}_0 + \frac{1}{n}\sum_{\ell=1}^{p}w(\ell)\sum_{i=\ell+1}^{n}e_ie_{i-\ell}(\mathbf{z}_i\mathbf{z}_{i-\ell}' + \mathbf{z}_{i-\ell}\mathbf{z}_i')\right] = \sum_{\ell=0}^{p}w(\ell)\mathbf{S}_\ell, \tag{18-16}$$

where

$$w(\ell) = 1 - \frac{\ell}{p+1}.$$

The maximum lag length p must be determined in advance. We will require that observations that are far apart in time—that is, for which $|i - \ell|$ is large—must have increasingly smaller covariances for us to establish the convergence results that justify OLS, GLS, and now GMM estimation. The choice of p is a reflection of how far back in time one must go to consider the autocorrelation negligible for purposes of estimating $(1/n)\mathbf{Z}'\boldsymbol{\Sigma}\mathbf{Z}$. Current practice suggests using the smallest integer greater than or equal to $T^{1/4}$.

Still left open is the question of where the initial consistent estimator should be obtained. One possibility is to obtain an inefficient but consistent GMM estimator by

using $\mathbf{W} = \mathbf{I}$ in (18-13). That is, use a nonlinear (or linear, if the equation is linear) instrumental variables estimator. This first-step estimator can then be used to construct $\mathbf{W}$, which, in turn, can then be used in the GMM estimator. Another possibility is that $\boldsymbol{\beta}$ may be consistently estimable by some straightforward procedure other than GMM.

Once the GMM estimator has been computed, its asymptotic covariance matrix and asymptotic distribution can be estimated based on (18-11) and (18-12). Recall that

$$\bar{\mathbf{m}}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \varepsilon_i,$$

which is a sum of $L \times 1$ vectors. The derivative, $\partial \bar{\mathbf{m}}_n(\boldsymbol{\beta})/\partial \boldsymbol{\beta}'$, is a sum of $L \times K$ matrices, so

$$\bar{\mathbf{G}}(\boldsymbol{\beta}) = \partial \bar{\mathbf{m}}(\boldsymbol{\beta})/\partial \boldsymbol{\beta}' = \frac{1}{n}\sum_{i=1}^{n} \mathbf{G}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{z}_i \left[\frac{\partial \varepsilon_i}{\partial \boldsymbol{\beta}'}\right]. \tag{18-17}$$

In the model we are considering here,

$$\frac{\partial \varepsilon_i}{\partial \boldsymbol{\beta}'} = \frac{-\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}.$$

The derivatives are the pseudoregressors in the linearized regression model that we examined in Section 9.2.3. Using the notation defined there,

$$\frac{\partial \varepsilon_i}{\partial \boldsymbol{\beta}} = -\mathbf{x}_{i0},$$

so

$$\bar{\mathbf{G}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{G}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} -\mathbf{z}_i \mathbf{x}_{i0}' = -\frac{1}{n}\mathbf{Z}'\mathbf{X}_0. \tag{18-18}$$

With this matrix in hand, the estimated asymptotic covariance matrix for the GMM estimator is

$$\text{Est.Asy. Var}[\hat{\boldsymbol{\beta}}] = \left[\mathbf{G}(\hat{\boldsymbol{\beta}})'\left(\frac{1}{n}\mathbf{Z}'\hat{\boldsymbol{\Sigma}}\mathbf{Z}\right)^{-1}\mathbf{G}(\hat{\boldsymbol{\beta}})\right]^{-1} = [(\mathbf{X}_0'\mathbf{Z})(\mathbf{Z}'\hat{\boldsymbol{\Sigma}}\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X}_0)]^{-1}. \tag{18-19}$$

(The two minus signs, a $1/n^2$ and an n^2, all fall out of the result.)

If the $\boldsymbol{\Sigma}$ that appears in (18-19) were $\sigma^2\mathbf{I}$, then (18-19) would be precisely the asymptotic covariance matrix that appears in Theorem 5.4 for linear models and Theorem 9.3 for nonlinear models. But there is an interesting distinction between this estimator and the IV estimators discussed earlier. In the earlier cases, when there were more instrumental variables than parameters, we resolved the overidentification by specifically choosing a set of K instruments, the K projections of the columns of $\mathbf{X}$ or $\mathbf{X}_0$ into the column space of $\mathbf{Z}$. Here, in contrast, we do not attempt to resolve the overidentification; we simply use all the instruments and minimize the GMM criterion. Now you should be able to show that when $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ *and we use this information,* when all is said and done, the same parameter estimates will be obtained. But, if we use a weighting matrix that differs from $\mathbf{W} = (\mathbf{Z}'\mathbf{Z}/n)^{-1}$, then they are not.

18.4 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 that we have examined at several points in the preceding chapters. 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.

18.4.1 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,

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

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 18.9 Overidentifying Restrictions

In Hall's consumption model with the corollary the two orthogonality conditions noted in Example 18.6 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 model. 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. His test is more involved than the one suggested above Hansen and Singleton (1982) operated directly on this form of the model. Other studies, for example, Campbell and Mankiw (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, though 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

$$(nq_R - nq) \xrightarrow{d} \chi^2[J]. \tag{18-20}$$

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 18.5.) We now consider several alternatives.

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

Section 17.5 described 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

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

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

$$W = \text{Wald} = [\mathbf{r}(\mathbf{c}_1)]'\{\text{Est.Asy. Var}[\mathbf{r}(\mathbf{c}_1)]\}^{-1}[\mathbf{r}(\mathbf{c}_1)]. \tag{18-21}$$

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

$$\text{Est.Asy. Var}[\mathbf{r}(\mathbf{c}_1)] = \mathbf{A}_1\{\text{Est.Asy. Var}[\mathbf{c}_1]\}\mathbf{A}_1', \tag{18-22}$$

where

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

The Wald statistic can be computed using only the unrestricted estimate. The LM statistic is

$$\text{LM} = \text{Lagrange multiplier} = \mathbf{g}_1'(\mathbf{c}_0)\{\text{Est.Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)]\}^{-1}\mathbf{g}_1(\mathbf{c}_0), \tag{18-23}$$

where

$$\mathbf{g}_1(\mathbf{c}_0) = \partial \ln L_1(\mathbf{c}_0)/\partial \mathbf{c}_0,$$

that is, the first derivatives of the *unconstrained* log-likelihood computed at the *restricted* estimates. The term Est.Asy. Var$[\mathbf{g}_1(\mathbf{c}_0)]$ is inverse of any of the usual estimators of the asymptotic covariance matrix of the maximum likelihood estimators of the parameters, computed using the restricted estimates. The most convenient choice is usually the BHHH estimator. The LM statistic is based on the restricted estimates.

Newey and West (1987b) have devised counterparts to these test statistics for the GMM estimator. The Wald statistic is computed identically, using the results of GMM estimation rather than maximum likelihood.[10] That is, in (18-21), we would use the unrestricted GMM estimator of $\boldsymbol{\theta}$. The appropriate asymptotic covariance matrix is (18-12). The computation is exactly the same. The counterpart to the LR statistic is the difference in the values of nq in (18-20). It is necessary to use the same weighting matrix, $\mathbf{W}$, in both restricted and unrestricted estimators. Since the unrestricted estimator is consistent under both H_0 and H_1, a consistent, unrestricted estimator of $\boldsymbol{\theta}$ is used to compute $\mathbf{W}$. Label this $\boldsymbol{\Phi}_1^{-1} = \{\text{Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}_1(\mathbf{c}_1)]\}^{-1}$. In each occurrence, the subscript 1 indicates reference to the unrestricted estimator. Then q is minimized without restrictions to obtain q_1 and then subject to the restrictions to obtain q_0. The statistic is then $(nq_0 - nq_1)$.[11] Since we are using the same $\mathbf{W}$ in both cases, this statistic is necessarily nonnegative. (This is the statistic discussed in Section 18.4.1.)

Finally, the counterpart to the LM statistic would be

$$\text{LM}_{GMM} = n[\bar{\mathbf{m}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{G}}_1(\mathbf{c}_0)]\,[\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{G}}_1(\mathbf{c}_0)]^{-1}\,[\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\bar{\mathbf{m}}_1(\mathbf{c}_0)].$$

[10] See Burnside and Eichenbaum (1996) for some small-sample results on this procedure. Newey and McFadden (1994) have shown the asymptotic equivalence of the three procedures.

[11] Newey and West label this test the D test.

The logic for this LM statistic is the same as that for the MLE. The derivatives of the minimized criterion q in (18-3) are

$$\mathbf{g}_1(\mathbf{c}_0) = \frac{\partial q}{\partial \mathbf{c}_0} = 2\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\mathbf{\Phi}}_1^{-1}\bar{\mathbf{m}}(\mathbf{c}_0).$$

The **LM statistic,** LM_{GMM}, is a Wald statistic for testing the hypothesis that this vector equals zero under the restrictions of the null hypothesis. From our earlier results, we would have

$$\text{Est.Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)] = \frac{4}{n}\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\mathbf{\Phi}}_1^{-1}\{\text{Est.Asy. Var}[\sqrt{n}\,\bar{\mathbf{m}}(\mathbf{c}_0)]\}\hat{\mathbf{\Phi}}_1^{-1}\bar{\mathbf{G}}_1(\mathbf{c}_0).$$

The estimated asymptotic variance of $\sqrt{n}\,\bar{\mathbf{m}}(\mathbf{c}_0)$ is $\hat{\mathbf{\Phi}}_1$, so

$$\text{Est.Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)] = \frac{4}{n}\bar{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\mathbf{\Phi}}_1^{-1}\bar{\mathbf{G}}_1(\mathbf{c}_0).$$

The Wald statistic would be

$$\begin{aligned}\text{Wald} &= \mathbf{g}_1(\mathbf{c}_0)'\{\text{Est.Asy. Var}[\mathbf{g}_1(\mathbf{c}_0)]\}^{-1}\mathbf{g}_1(\mathbf{c}_0)\\ &= n\,\bar{\mathbf{m}}_1'(\mathbf{c}_0)\hat{\mathbf{\Phi}}_1^{-1}\bar{\mathbf{G}}(\mathbf{c}_0)\{\bar{\mathbf{G}}(\mathbf{c}_0)'\hat{\mathbf{\Phi}}_1^{-1}\bar{\mathbf{G}}(\mathbf{c}_0)\}^{-1}\bar{\mathbf{G}}(\mathbf{c}_0)'\hat{\mathbf{\Phi}}_1^{-1}\bar{\mathbf{m}}_1(\mathbf{c}_0).\end{aligned} \tag{18-24}$$

8.5 APPLICATION: GMM ESTIMATION OF A DYNAMIC PANEL DATA MODEL OF LOCAL GOVERNMENT EXPENDITURES

(This example continues the analysis begun in Example 13.7.) Dahlberg and Johansson (2000) estimated a model for the local government expenditure of several hundred municipalities in Sweden observed over the 9-year period $t = 1979$ to 1987. The equation of interest is

$$S_{i,t} = \alpha_t + \sum_{j=1}^{m}\beta_j S_{i,t-j} + \sum_{j=1}^{m}\gamma_j R_{i,t-j} + \sum_{j=1}^{m}\delta_j G_{i,t-j} + f_i + \varepsilon_{it}$$

for $i = 1, \ldots, N = 265$ and $t = m+1, \ldots, 9$. (We have changed their notation slightly to make it more convenient.) $S_{i,t}$, $R_{i,t}$ and $G_{i,t}$ are municipal spending, receipts (taxes and fees) and central government grants, respectively. Analogous equations are specified for the current values of $R_{i,t}$ and $G_{i,t}$. The appropriate lag length, m, is one of the features of interest to be determined by the empirical study. The model contains a municipality specific effect, f_i, which is not specified as being either "fixed" or "random." In order to eliminate the individual effect, the model is converted to first differences. The resulting equation is

$$\Delta S_{i,t} = \lambda_t + \sum_{j=1}^{m}\beta_j \Delta S_{i,t-j} + \sum_{j=1}^{m}\gamma_j \Delta R_{i,t-j} + \sum_{j=1}^{m}\delta_j \Delta G_{i,t-j} + u_{it}$$

or

$$y_{i,t} = \mathbf{x}_{i,t}'\boldsymbol{\theta} + u_{i,t},$$

where $\Delta S_{i,t} = S_{i,t} - S_{i,t-1}$ and so on and $u_{i,t} = \varepsilon_{i,t} - \varepsilon_{i,t-1}$. This removes the group effect and leaves the time effect. Since the time effect was unrestricted to begin with,

$\Delta\alpha_t = \lambda_t$ remains an unrestricted time effect, which is treated as "fixed" and modeled with a time-specific dummy variable. The maximum lag length is set at $m = 3$. With 9 years of data, this leaves useable observations from 1983 to 1987 for estimation, that is, $t = m+2, \ldots, 9$. Similar equations were fit for $R_{i,t}$ and $G_{i,t}$.

The orthogonality conditions claimed by the authors are

$$E[S_{i,s}u_{i,t}] = E[R_{i,s}u_{i,t}] = E[G_{i,s}u_{i,t}] = 0, \quad s = 1, \ldots, t-2.$$

The orthogonality conditions are stated in terms of the levels of the financial variables and the differences of the disturbances. The issue of this formulation as opposed to, for example, $E[\Delta S_{i,s}\Delta\varepsilon_{i,t}] = 0$ (which is implied) is discussed by Ahn and Schmidt (1995). As we shall see, this set of orthogonality conditions implies a total of 80 instrumental variables. The authors use only the first of the three sets listed above, which produces a total of 30. For the five observations, using the formulation developed in Section 13.6, we have the following matrix of instrumental variables for the orthogonality conditions

$$\mathbf{Z}_i = \begin{bmatrix} S_{81-79} & d_{83} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & S_{82-79} & d_{84} & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{83-79} & d_{85} & \mathbf{0}' & 0 & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{84-79} & d_{86} & \mathbf{0}' & 0 \\ \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & \mathbf{0}' & 0 & S_{85-79} & d_{87} \end{bmatrix} \begin{matrix} 1983 \\ 1984 \\ 1985 \\ 1986 \\ 1987 \end{matrix}$$

where the notation E_{t1-t0} indicates the range of years for that variable. For example, S_{83-79} denotes $[S_{i,1983}, S_{i,1982}, S_{i,1981}, S_{i,1980}, S_{i,1979}]$ and d_{year} denotes the year specific dummy variable. Counting columns in $\mathbf{Z}_i$ we see that using only the lagged values of the dependent variable and the time dummy variables, we have $(3+1)+(4+1)+(5+1)+(6+1)+(7+1) = 30$ instrumental variables. Using the lagged values of the other two variables in each equation would add 50 more, for a total of 80 if all the orthogonality conditions suggested above were employed. Given the construction above, the orthogonality conditions are now

$$E[\mathbf{Z}_i'\mathbf{u}_i] = \mathbf{0},$$

where $\mathbf{u}_i = [u_{i,1987}, u_{i,1986}, u_{i,1985}, u_{i,1984}, u_{i,1983}]'$. The empirical moment equation is

$$\text{plim}\left[\frac{1}{n}\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{u}_i\right] = \text{plim}\,\bar{\mathbf{m}}(\boldsymbol{\theta}) = \mathbf{0}.$$

The parameters are vastly overidentified. Using only the lagged values of the dependent variable in each of the three equations estimated, there are 30 moment conditions and 14 parameters being estimated when $m = 3$, 11 when $m = 2$, 8 when $m = 1$ and 5 when $m = 0$. (As we do our estimation of each of these, we will retain the same matrix of instrumental variables in each case.) GMM estimation proceeds in two steps. In the first step, basic, unweighted instrumental variables is computed using

$$\hat{\boldsymbol{\theta}}_{IV}' = \left[\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1}\left(\sum_{i=1}^{N}\mathbf{X}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{N}\mathbf{Z}_i'\mathbf{y}_i\right)$$

where

$$\mathbf{y}_i' = (\Delta S_{83} \quad \Delta S_{84} \quad \Delta S_{85} \quad \Delta S_{86} \quad \Delta S_{87})$$

and

$$\mathbf{X}_i = \begin{bmatrix} \Delta S_{82} & \Delta S_{81} & \Delta S_{80} & \Delta R_{82} & \Delta R_{81} & \Delta R_{80} & \Delta G_{82} & \Delta G_{81} & \Delta G_{80} & 1 & 0 & 0 & 0 & 0 \\ \Delta S_{83} & \Delta S_{82} & \Delta S_{81} & \Delta R_{83} & \Delta R_{82} & \Delta R_{81} & \Delta G_{83} & \Delta G_{82} & \Delta G_{81} & 0 & 1 & 0 & 0 & 0 \\ \Delta S_{84} & \Delta S_{83} & \Delta S_{82} & \Delta R_{84} & \Delta R_{83} & \Delta R_{82} & \Delta G_{84} & \Delta G_{83} & \Delta G_{82} & 0 & 0 & 1 & 0 & 0 \\ \Delta S_{85} & \Delta S_{84} & \Delta S_{83} & \Delta R_{85} & \Delta R_{84} & \Delta R_{83} & \Delta G_{85} & \Delta G_{84} & \Delta G_{83} & 0 & 0 & 0 & 1 & 0 \\ \Delta S_{86} & \Delta S_{85} & \Delta S_{84} & \Delta R_{86} & \Delta R_{85} & \Delta R_{84} & \Delta G_{86} & \Delta G_{85} & \Delta G_{84} & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The second step begins with the computation of the new weighting matrix,

$$\hat{\mathbf{\Phi}} = \text{Est.Asy. Var}[\sqrt{N}\bar{\mathbf{m}}] = \frac{1}{N}\sum_{i=1}^{N} \mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i.$$

After multiplying and dividing by the implicit $(1/N)$ in the outside matrices, we obtain the estimator,

$$\begin{aligned} \boldsymbol{\theta}'_{GMM} &= \left[\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{N} \mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1} \\ &\quad \times \left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Z}_i\right)\left(\sum_{i=1}^{N} \mathbf{Z}_i'\hat{\mathbf{u}}_i\hat{\mathbf{u}}_i'\mathbf{Z}_i\right)^{-1}\left(\sum_{i=1}^{N} \mathbf{Z}_i'\mathbf{y}_i\right) \\ &= \left[\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Z}_i\right)\mathbf{W}\left(\sum_{i=1}^{N} \mathbf{Z}_i'\mathbf{X}_i\right)\right]^{-1}\left(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{Z}_i\right)\mathbf{W}\left(\sum_{i=1}^{N} \mathbf{Z}_i'\mathbf{y}_i\right). \end{aligned}$$

The estimator of the asymptotic covariance matrix for the estimator is the matrix in square brackets in the first line of the result.

The primary focus of interest in the study was not the estimator itself, but the lag length and whether certain lagged values of the independent variables appeared in each equation. These restrictions would be tested by using the GMM criterion function, which in this formulation would be (based on recomputing the residuals after GMM estimation)

$$q = \left(\sum_{i=1}^{n} \hat{\mathbf{u}}_i'\mathbf{Z}_i\right)\mathbf{W}\left(\sum_{i=1}^{n} \mathbf{Z}_i'\hat{\mathbf{u}}_i\right).$$

Note that the weighting matrix is not (necessarily) recomputed. For purposes of testing hypotheses, the same weighting matrix should be used.

At this point, we will consider the appropriate lag length, m. The specification can be reduced simply by redefining $\mathbf{X}$ to change the lag length. In order to test the specification, the weighting matrix must be kept constant for all restricted versions ($m = 2$ and $m = 1$) of the model.

The Dahlberg and Johansson data may be downloaded from the *Journal of Applied Econometrics* website—See Appendix Table F18.1. The authors provide the summary statistics for the raw data that are given in Table 18.2. The data used in the study

TABLE 18.2 Descriptive Statistics for Local Expenditure Data

Variable	*Mean*	*Std. Deviation*	*Minimum*	*Maximum*
Spending	18478.51	3174.36	12225.68	33883.25
Revenues	13422.56	3004.16	6228.54	29141.62
Grants	5236.03	1260.97	1570.64	12589.14

TABLE 18.3 Estimated Spending Equation

Variable	*Estimate*	*Standard Error*	*t Ratio*
Year 1983	−0.0036578	0.0002969	−12.32
Year 1984	−0.00049670	0.0004128	−1.20
Year 1985	0.00038085	0.0003094	1.23
Year 1986	0.00031469	0.0003282	0.96
Year 1987	0.00086878	0.0001480	5.87
Spending ($t-1$)	1.15493	0.34409	3.36
Revenues ($t-1$)	−1.23801	0.36171	−3.42
Grants ($t-1$)	0.016310	0.82419	0.02
Spending ($t-2$)	−0.0376625	0.22676	−0.17
Revenues ($t-2$)	0.0770075	0.27179	0.28
Grants ($t-2$)	1.55379	0.75841	2.05
Spending ($t-3$)	−0.56441	0.21796	−2.59
Revenues ($t-3$)	0.64978	0.26930	2.41
Grants ($t-3$)	1.78918	0.69297	2.58

and provided in the internet source are nominal values in Swedish Kroner, deflated by a municipality specific price index then converted to per capita values. Descriptive statistics for the raw and transformed data appear in Table 18.2.[12] Equations were estimated for all three variables, with maximum lag lengths of $m = 1$, 2, and 3. (The authors did not provide the actual estimates.) Estimation is done using the methods developed by Ahn and Schmidt (1995), Arellano and Bover (1995) and Holtz-Eakin, Newey, and Rosen (1988), as described above. The estimates of the first specification given above are given in Table 18.3.

Table 18.4 contains estimates of the model parameters for each of the three equations, and for the three lag lengths, as well as the value of the GMM criterion function for each model estimated. The base case for each model has $m = 3$. There are three restrictions implied by each reduction in the lag length. The critical chi-squared value for three degrees of freedom is 7.81 for 95 percent significance, so at this level, we find that the two-level model is just barely accepted for the spending equation, but clearly appropriate for the other two—the difference between the two criteria is 7.62. Conditioned on $m = 2$, only the revenue model rejects the restriction of $m = 1$. As a final test, we might ask whether the data suggest that perhaps no lag structure at all is necessary. The GMM criterion value for the three equations with only the time dummy variables are 45.840, 57.908, and 62.042, respectively. Therefore, all three zero lag models are rejected.

[12]The data provided on the website and used in our computations were further transformed by dividing by 100,000.

TABLE 18.4 Estimated Lag Equations for Spending, Revenue, and Grants

	Expenditure Model			*Revenue Model*			*Grant Model*		
	$m = 3$	$m = 2$	$m = 1$	$m = 3$	$m = 2$	$m = 1$	$m = 3$	$m = 2$	$m = 1$
S_{t-1}	1.155	0.8742	0.5562	−0.1715	−0.3117	−0.1242	−0.1675	−0.1461	−0.1958
S_{t-2}	−0.0377	0.2493	—	0.1621	−0.0773	—	−0.0303	−0.0304	—
S_{t-3}	−0.5644	—	—	−0.1772	—	—	−0.0955	—	—
R_{t-1}	−1.2380	−0.8745	−0.5328	−0.0176	0.1863	−0.0245	0.1578	0.1453	0.2343
R_{t-2}	0.0770	−0.2776	—	−0.0309	0.1368	—	0.0485	0.0175	—
R_{t-3}	0.6497	—	—	0.0034	—	—	0.0319	—	—
G_{t-1}	0.0163	−0.4203	0.1275	−0.3683	0.5425	−0.0808	−0.2381	−0.2066	−0.0559
G_{t-2}	1.5538	0.1866	—	−2.7152	2.4621	—	−0.0492	−0.0804	—
G_{t-3}	1.7892	—	—	0.0948	—	—	0.0598	—	—
q	22.8287	30.4526	34.4986	30.5398	34.2590	53.2506	17.5810	20.5416	27.5927

Among the interests in this study were the appropriate critical values to use for the specification test of the moment restriction. With 16 degrees of freedom, the critical chi-squared value for 95 percent significance is 26.3, which would suggest that the revenues equation is misspecified. Using a bootstrap technique, the authors find that a more appropriate critical value leaves the specification intact. Finally, note that the three-equation model in the $m = 3$ columns of Table 18.4 imply a **vector autoregression** of the form

$$\mathbf{y}_t = \mathbf{\Gamma}_1\mathbf{y}_{t-1} + \mathbf{\Gamma}_2\mathbf{y}_{t-2} + \mathbf{\Gamma}_3\mathbf{y}_{t-3} + \mathbf{v}_t$$

where $\mathbf{y}_t = (\Delta S_t, \Delta R_t, \Delta G_t)'$. We will explore the properties and characteristics of equation systems such as this in our discussion of time series models in Chapter 20.

18.6 SUMMARY AND CONCLUSIONS

The generalized method of moments provides an estimation framework that includes least squares, nonlinear least squares, instrumental variables, and maximum likelihood, and a general class of estimators that extends beyond these. But it is more than just a theoretical umbrella. The GMM provides a method of formulating models and implied estimators without making strong distributional assumptions. Hall's model of household consumption is a useful example that shows how the optimization conditions of an underlying economic theory produce a set of distribution free estimating equations. In this chapter, we first examined the classical method of moments. GMM as an estimator is an extension of this strategy that allows the analyst to use additional information beyond that necessary to identify the model, in an optimal fashion. After defining and establishing the properties of the estimator, we then turned to inference procedures. It is convenient that the GMM procedure provides counterparts to the familiar trio of test statistics, Wald, LM, and LR. In the final section, we developed an example that appears at many points in the recent applied literature, the dynamic panel data model with individual specific effects, and lagged values of the dependent variable.

This chapter concludes our survey of estimation techniques and methods in econometrics. In the remaining chapters of the book, we will examine a variety of applications

and modeling tools, first in time series and macroeconometrics in Chapters 19 and 20, then in discrete choice models and limited dependent variables, the staples of microeconometrics, in Chapters 21 and 22.

Key Terms and Concepts

- Analog estimation
- Asymptotic properties
- Central limit theorem
- Central moments
- Consistent estimator
- Dynamic panel data model
- Empirical moment equation
- Ergodic theorem
- Euler equation
- Exactly identified
- Exponential family
- Generalized method of moments
- Identification
- Instrumental variables
- LM statistic
- LR statistic
- Martingale difference sequence
- Maximum likelihood estimator
- Mean value theorem
- Method of moment generating functions
- Method of moments
- Method of moments estimators
- Minimum distance estimator
- Moment equation
- Newey–West estimator
- Nonlinear instrumental variable estimator
- Order condition
- Orthogonality conditions
- Overidentifying restrictions
- Probability limit
- Random sample
- Rank condition
- Robust estimation
- Slutsky Theorem
- Specification test statistic
- Sufficient statistic
- Taylor series
- Uncentered moment
- Wald statistic
- Weighted least squares

Exercises

1. For the normal distribution $\mu_{2k} = \sigma^{2k}(2k)!/(k!2^k)$ and $\mu_{2k+1} = 0, k = 0, 1, \ldots$. Use this result to analyze the two estimators

$$\sqrt{b_1} = \frac{m_3}{m_2^{3/2}} \quad \text{and} \quad b_2 = \frac{m_4}{m_2^2}.$$

where $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$. The following result will be useful:

$$\text{Asy.Cov}[\sqrt{n}m_j, \sqrt{n}m_k] = \mu_{j+k} - \mu_j\mu_k + jk\mu_2\mu_{j-1}\mu_{k-1} - j\mu_{j-1}\mu_{k+1} - k\mu_{k-1}\mu_{j+1}.$$

Use the delta method to obtain the asymptotic variances and covariance of these two functions assuming the data are drawn from a normal distribution with mean μ and variance σ^2. (Hint: Under the assumptions, the sample mean is a consistent estimator of μ, so for purposes of deriving asymptotic results, the difference between $\bar{x}$ and μ may be ignored. As such, no generality is lost by assuming the mean is zero, and proceeding from there. Obtain $\mathbf{V}$, the 3×3 covariance matrix for the three moments, then use the delta method to show that the covariance matrix for the two estimators is

$$\mathbf{JVJ}' = \begin{bmatrix} 6 & 0 \\ 0 & 24 \end{bmatrix}$$

where $\mathbf{J}$ is the 2×3 matrix of derivatives.

2. Using the results in Example 18.7, estimate the asymptotic covariance matrix of the method of moments estimators of P and λ based on m_1' and m_2' [Note: You will need to use the data in Example C.1 to estimate $\mathbf{V}$.]

3. **Exponential Families of Distributions.** For each of the following distributions, determine whether it is an exponential family by examining the log-likelihood function. Then, identify the sufficient statistics.
   a. Normal distribution with mean μ and variance σ^2.
   b. The Weibull distribution in Exercise 4 in Chapter 17.
   c. The mixture distribution in Exercise 3 in Chapter 17.
4. In the classical regression model with heteroscedasticity, which is more efficient, ordinary least squares or GMM? Obtain the two estimators and their respective asymptotic covariance matrices, then prove your assertion.
5. Consider the probit model analyzed in Section 17.8. The model states that for given vector of independent variables,

$$\text{Prob}[y_i = 1 \mid \mathbf{x}_i] = \Phi[\mathbf{x}_i'\boldsymbol{\beta}], \quad \text{Prob}[y_i = 0 \mid \mathbf{x}_i] = 1 - \text{Prob}[y_i = 1 \mid \mathbf{x}_i].$$

   We have considered maximum likelihood estimation of the parameters of this model at several points. Consider, instead, a GMM estimator based on the result that

$$E[y_i \mid \mathbf{x}_i] = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$$

   This suggests that we might base estimation on the orthogonality conditions

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

   Construct a GMM estimator based on these results. Note that this is not the nonlinear least squares estimator. Explain—what would the orthogonality conditions be for nonlinear least squares estimation of this model?
6. Consider GMM estimation of a regression model as shown at the beginning of Example 18.8. Let $\mathbf{W}_1$ be the optimal weighting matrix based on the moment equations. Let $\mathbf{W}_2$ be some other positive definite matrix. Compare the asymptotic covariance matrices of the two proposed estimators. Show conclusively that the asymptotic covariance matrix of the estimator based on $\mathbf{W}_1$ is not larger than that based on $\mathbf{W}_2$.