13

MINIMUM DISTANCE ESTIMATION AND THE GENERALIZED METHOD OF MOMENTS

13.1 INTRODUCTION

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

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

Example 13.1 Euler Equations and Life Cycle Consumption

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

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

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

total assets A_t not including human capital. In this model, it is claimed that the only source of uncertainty is w_t. No assumption is made about the stochastic properties of w_t except that there exists an expected future earnings, $E_t[w_{t+\tau} | \boldsymbol{\Omega}_t]$. 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

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

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

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

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

Hall assumed that r was constant over time. Other applications of this modeling framework modified the framework so as to involve a forecasted interest rate, r_{t+1}.[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,

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

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.

13.2 CONSISTENT ESTIMATION: THE METHOD OF MOMENTS

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

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

[1]For example, Hansen and Singleton (1982).

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

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 13.4.

13.2.1 RANDOM SAMPLING AND ESTIMATING THE PARAMETERS OF DISTRIBUTIONS

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

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

By Theorem D.4,

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

and

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

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

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

Finally, by the Lindeberg–Levy central limit theorem,

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

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

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

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

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

and

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

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

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

and

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

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

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

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

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

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

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

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

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

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

Example 13.3 Inverse Gaussian (Wald) Distribution

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

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

The mean is μ while the variance is μ^3/λ. The efficient maximum likelihood estimators of the two parameters are based on $(1/n)\sum_{i=1}^{n} y_i$ and $(1/n)\sum_{i=1}^{n}(1/y_i)$. Because the mean and variance are simple functions of the underlying parameters, we can also use the sample mean and sample variance as moment estimators of these functions. Thus, an alternative pair of method of moments estimators for the parameters of the Wald distribution can be based on $(1/n)\sum_{i=1}^{n} y_i$ and $(1/n)\sum_{i=1}^{n} y_i^2$. The precise formulas for this pair of moment estimators are left as an exercise.

Example 13.4 Mixture of Normal Distributions

Quandt and Ramsey (1978) analyzed the problem of estimating the parameters of a mixture of two normal distributions. Suppose that each observation in a random sample is drawn from one of two different normal distributions. The probability that the observation is drawn

from the first distribution, $N[\mu_1, \sigma_1^2]$, is λ and the probability that it is drawn from the second is $(1 - \lambda)$. The density for the observed y is $f(y) = \lambda\, N[\mu_1, {\sigma_1}^2] + (1 - \lambda)N[\mu_2, {\sigma_2}^2]$, $0 < \lambda < 1$. Inserting the definitions gives

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

Before proceeding, we note that this density is precisely the same as the finite mixture model described in Section 14.15.1. Maximum likelihood estimation of the model using the method described there would be simpler than the method of moment generating functions developed here.

The sample mean and second through fifth central moments,

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

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

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

For the mixed normal distribution, the mean and variance are

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

and

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

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

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

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

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

to estimate the parameters. The moment equations are $\overline{M}_t - \Lambda_t(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda) = 0$. They label this procedure the **method of moment generating functions**. (See Section B.6 for the 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 13.1 Exponential Family

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

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

where $a(\cdot)$, $b(\cdot)$, $c_k(\cdot)$, and $s_k(\cdot)$ are functions. The members of the "family" are distinguished by the different parameter values. The normal distribution and the Wald distribution in Example 13.3 are examples.

If the log-likelihood function is of this form, then the functions $c_k(\cdot)$ are called **sufficient statistics**.[2] When sufficient statistics exist, method of moments estimator(s) can be functions of them. In this case, the method of moments estimators will also be the maximum likelihood estimators, so, of course, they will be efficient, at least asymptotically. We emphasize, in this case, the probability distribution is fully specified. Because the normal distribution is an exponential family with sufficient statistics $\overline{m}'_1$ and $\overline{m}'_2$, the estimators described in Example 13.2 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 13.3 is an exponential family. You should be able to show that the sufficient statistics are the ones that are suggested in Example 13.3 as the bases for the MLEs of μ and λ.

Example 13.5 Gamma Distribution

The gamma distribution (see Section B.4.5) is

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

The log-likelihood function for this distribution is

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

This function is an exponential family with $a(\mathbf{data}) = 0$, $b(\boldsymbol{\theta}) = n[P \ln \lambda - \ln \Gamma(P)]$ 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

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

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

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

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

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

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

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

13.2.2 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

[2] 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 17 is that if the statistics, $c_k(\mathbf{data})$, are sufficient statistics, then the conditional density, $f[y_1, \ldots, y_n | c_k(\mathbf{data}), k = 1, \ldots, K]$, is not a function of the parameters.

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 $Var[y]/n$, 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 13.4, the underlying sample statistic is

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

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

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

In general, when the moments are computed as

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

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

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

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

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

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

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 $\overline{\mathbf{G}}_n(\boldsymbol{\theta})$ be the $K \times K$ matrix whose kth row is the vector of partial derivatives,

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

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

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

Therefore,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx -[\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0)]^{-1}\sqrt{n}[\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)]. \tag{13-1}$$

(We have treated this as an approximation because we are not dealing formally with the higher-order term in the Taylor series. We will make this explicit in the treatment of the GMM estimator in Section 13.4.) The argument needed to characterize the large sample behavior of the estimator, $\hat{\boldsymbol{\theta}}$, is discussed in Appendix D. We have from Theorem D.18 (the central limit theorem) that $\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)$ has a limiting normal distribution with mean vector $\mathbf{0}$ and covariance matrix equal to $\boldsymbol{\Phi}$. Assuming that the functions in the moment equation are continuous and functionally independent, we can expect $\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0)$ to converge to a nonsingular matrix of constants, $\Gamma(\boldsymbol{\theta}_0)$. Under general conditions, the limiting distribution of the right-hand side of (13-1) 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 $(1/n) \times \{-[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1}\}\boldsymbol{\Phi}\{-[\boldsymbol{\Gamma}'(\boldsymbol{\theta}_0)]^{-1}\}$. Thus, the asymptotic covariance matrix for the method of moments estimator may be estimated with

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

Example 13.5 (Continued)

Using the estimates $\hat{\boldsymbol{\theta}}(m_1', m_*') = (2.4106, 0.0770702)$,

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

[The function $\Psi'(P)$ is $d^2 \ln \Gamma(P)/dP^2 = (\Gamma\Gamma'' - \Gamma')/\Gamma^2$. With $\hat{P} = 2.4106$, $\hat{\Gamma} = 1.250832$, $\hat{\Psi} = 0.658347$, and $\hat{\Psi}' = 0.512408$].[3] The matrix $\mathbf{F}$ is the sample covariance matrix of y and $\ln y$ (using 19 as the divisor),

$$\mathbf{F} = \begin{bmatrix} 500.68 & 14.31 \\ 14.31 & 0.47746 \end{bmatrix}.$$

The product is

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

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

$$[-\mathbf{H}]^{-1} = \frac{1}{n}\begin{bmatrix} \Psi' & -1/\lambda \\ -1/\lambda & P/\lambda^2 \end{bmatrix}^{-1} = \begin{bmatrix} 0.51243 & 0.01638 \\ 0.01638 & 0.00064654 \end{bmatrix}.$$

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

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

a sign change, also the derivatives of the log-likelihood. The estimates of the two variances are 0.51203 and 0.00064654, respectively, which agrees reasonably well with the method of moments estimates. The difference would be due to sampling variability in a finite sample and the presence of **F** in the first variance estimator.

13.2.3 SUMMARY—THE METHOD OF MOMENTS

In the simplest cases, the method of moments is robust to differences in the specification of the data-generating process (DGP). A sample mean or variance estimates its population counterpart (assuming it exists), regardless of the underlying process. It is this freedom from unnecessary distributional assumptions that has made this method so popular in recent years. However, this comes at a cost. If more is known about the DGP, its specific distribution for example, then the method of moments may not make use of all of the available information. Thus, in Example 13.3, 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.

13.3 MINIMUM DISTANCE ESTIMATION

The preceding analysis has considered **exactly identified cases**. In each example, there were K parameters to estimate and we used K moments to estimate them. In Example 13.5, we examined the gamma distribution, a two-parameter family, and considered different pairs of moments that could be used to estimate the two parameters. The most efficient estimator for the parameters of this distribution will be based on $(1/n)\Sigma_i y_i$ and $(1/n)\Sigma_i \ln y_i$. This does raise a general question: How should we proceed if we have more moments than we need? It would seem counterproductive to simply discard the additional information. In this case, logically, the sample information provides more than one estimate of the model parameters, and it is now necessary to reconcile those competing estimators.

We have encountered this situation in several earlier examples: In Example 11.23, in Passmore et al.'s (2005) study of Fannie Mae, we have four independent estimators of a single parameter, $\hat{\alpha}_j$, each with estimated asymptotic variance $\hat{V}_j$, $j = 1, \ldots, 4$. The estimators were combined using a **criterion function**,

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

The solution to this minimization problem is a minimum distance estimator,

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

In forming the two-stage least squares estimator of the parameters in a dynamic panel data model in Section 11.10.3, we obtained $T - 2$ instrumental variable estimators of the parameter vector $\boldsymbol{\theta}$ by forming different instruments for each period for which we had sufficient data. The $T - 2$ estimators of the same parameter vector are $\hat{\boldsymbol{\theta}}_{\text{IV}(t)}$. The Arellano–Bond estimator of the single parameter vector in this setting is

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

where

$$\mathbf{W}_{(t)} = \left(\hat{\tilde{\mathbf{X}}}_{(t)}'\hat{\tilde{\mathbf{X}}}_{(t)}\right), \mathbf{R}_{(t)} = \left(\sum_{t=3}^{T}\mathbf{W}_{(t)}\right)^{-1}\mathbf{W}_{(t)} \text{ and } \sum_{t=3}^{T}\mathbf{R}_{(t)} = \mathbf{I}.$$

Finally, Carey's (1997) analysis of hospital costs that we examined in Example 11.13 involved a seemingly unrelated regressions model that produced multiple estimates of several of the model parameters. We will revisit this application in Example 13.6.

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

$$\text{plim } \overline{m}_{n,l} = g_l(\boldsymbol{\theta}_0), l = 1, \ldots, L,$$

where $\boldsymbol{\theta}_0$ is a vector of $K \leq L$ parameters to be estimated. Arrange these moments and functions in $L \times 1$ vectors $\overline{\mathbf{m}}_n$ and $\mathbf{g}(\boldsymbol{\theta}_0)$ and further assume that the statistics are jointly asymptotically normally distributed with plim $\overline{\mathbf{m}}_n = \mathbf{g}(\boldsymbol{\theta})$ and Asy.Var$[\overline{\mathbf{m}}_n] = (1/n)\boldsymbol{\Phi}$. Define the criterion function

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

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

THEOREM 13.1 Asymptotic Distribution of the Minimum Distance Estimator

Under the assumption that $\sqrt{n}[\overline{\mathbf{m}}_n - \mathbf{g}(\boldsymbol{\theta}_0)] \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}]$, *the asymptotic properties of the minimum distance estimator are as follows:*

$$\text{plim } \hat{\boldsymbol{\theta}}_{\text{MDE}} = \boldsymbol{\theta}_0,$$

$$\text{Asy.Var}[\boldsymbol{\theta}_{\text{MDE}}] = \frac{1}{n}[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1}[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)][\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1} = \frac{1}{n}\mathbf{V},$$

where

$$\boldsymbol{\Gamma}(\boldsymbol{\theta}_0) = \text{plim } \mathbf{G}(\hat{\boldsymbol{\theta}}_{\text{MDE}}) = \text{plim } \frac{\partial \mathbf{g}(\hat{\boldsymbol{\theta}}_{\text{MDE}})}{\partial \hat{\boldsymbol{\theta}}_{\text{MDE}}'},$$

and

$$\hat{\boldsymbol{\theta}}_{\text{MDE}} \xrightarrow{a} N\left[\boldsymbol{\theta}_0, \frac{1}{n}\mathbf{V}\right].$$

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

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

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

$$\text{Asy.Var}[\hat{\boldsymbol{\theta}}_{\text{MDE}}] = \frac{1}{n}[\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}(\boldsymbol{\theta}_0)]^{-1},$$

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

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

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

An important aspect of the MDE arises in the exactly identified case. If K equals L, and if the functions $g_l(\boldsymbol{\theta})$ are functionally independent, that is, $\mathbf{G}(\boldsymbol{\theta})$ has full row rank, K, then it is possible to solve the moment equations exactly. That is, the minimization problem becomes one of simply solving the K moment equations, $\overline{m}_{n,l} = g_l(\boldsymbol{\theta}_0)$ in the K unknowns, $\hat{\boldsymbol{\theta}}_{\text{MDE}}$. This is the method of moments estimator examined in the preceding section. In this instance, the weighting matrix, $\mathbf{W}$, is irrelevant to the solution, because the MDE will now satisfy the moment equations

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

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

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

for Passmore's analysis of Fannie Mae, and

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

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

Example 13.6 Minimum Distance Estimation of a Hospital Cost Function

In Carey's (1997) study of hospital costs in Example 11.13, Chamberlain's (1984) seemingly unrelated regressions (SUR) approach to a panel data model produces five period-specific estimates of a parameter vector, $\boldsymbol{\theta}_t$. Some of the parameters are specific to the year while others (it is hypothesized) are common to all five years. There are two specific parameters of interest, β_D and β_O, that are allowed to vary by year, but are each estimated multiple times by the SUR model. We focus on just these parameters. The model states

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

where

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

DIS_{it} is patient discharges, and OUT_{it} is outpatient visits. (We are changing Carey's notation slightly and suppressing parts of the model that are extraneous to the development here. The terms A_{it} and B_i contain those additional components.) The preceding model is

estimated by inserting the expression for α_i in the main equation, then fitting an unrestricted seemingly unrelated regressions model by FGLS. There are five years of data, hence five sets of estimates. Note, however, with respect to the discharge variable, *DIS*, although each equation provides separate estimates of ($\gamma_{D,1}$, . . . , ($\beta_{D,t} + \gamma_{D,t}$), . . . , $\gamma_{D,5}$), a total of five parameter estimates in each equation (year), there are only 10, not 25 parameters to be estimated in total. The parameters on OUT_{it} are likewise overidentified. Table 13.1 reproduces the estimates in Table 11.7 for the discharge coefficients and adds the estimates for the outpatient variable.

TABLE 13.1a Coefficient Estimates for DIS in SUR Model for Hospital Costs

Coefficient on Variable in the Equation

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

TABLE 13.1b Coefficient Estimates for OUT in SUR Model for Hospital Costs

Coefficient on Variable in the Equation

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

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

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

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

Let $\hat{\boldsymbol{\beta}}_t$ denote the 10×1 period-specific estimator of the model parameters. Unlike the other cases we have examined, the individual estimates here are not uncorrelated. In the SUR model, the estimated asymptotic covariance matrix is the partitioned matrix given in (10-7). For the estimators of two equations,

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

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

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

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

$$\overline{\mathbf{m}}_{n,87} - \mathbf{g}_{87}(\boldsymbol{\theta})$$

$$= \begin{bmatrix} \{\hat{\beta}^{87}_{D,87} - (\beta_{D,87} + \gamma_{D,87})\}, \{\hat{\beta}^{87}_{D,88} - \gamma_{D,88}\}, \{\hat{\beta}^{87}_{D,89} - \gamma_{D,89}\}, \{\hat{\beta}^{87}_{D,90} - \gamma_{D,90}\}, \{\hat{\beta}^{87}_{D,91} - \gamma_{D,90}\}, \\ \{\hat{\beta}^{87}_{O,87} - (\beta_{O,87} + \gamma_{O,87})\}, \{\hat{\beta}^{87}_{O,88} - \gamma_{O,88}\}, \{\hat{\beta}^{87}_{O,89} - \gamma_{O,89}\}, \{\hat{\beta}^{87}_{O,90} - \gamma_{O,90}\}, \{\hat{\beta}^{87}_{O,91} - \gamma_{O,90}\} \end{bmatrix}'$$

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

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

Note there are 50 estimated parameters from the SUR equations (those are listed in Table 13.1) and 20 unknown parameters to be calibrated in the criterion function. The reported minimum distance estimates are shown in the last row of each table.

13.4 THE GENERALIZED METHOD OF MOMENTS (GMM) ESTIMATOR

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

The GMM estimation technique is an extension of the minimum distance estimator described in Section 13.3.[4] 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 13.2.

13.4.1 ESTIMATION BASED ON ORTHOGONALITY CONDITIONS

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

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

The sample analog is

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

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

For the instrumental variables estimator of Chapter 8, we relied on a large sample analog to the moment condition,

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

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

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

where the columns of $\hat{\mathbf{X}}$ are the fitted values in regressions on all the columns of $\mathbf{Z}$ (that is, the projections of these columns of $\mathbf{X}$ into the column space of $\mathbf{Z}$). (See Section 8.3.4 for further details.)

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

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

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

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

[4]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,b); Davidson and MacKinnon (1993); and Newey and McFadden (1994). Useful summaries of GMM estimation are provided by Pagan and Wickens (1989) and Matyas (1999). An application of some of these techniques that contains useful summaries is Pagan and Vella (1989). Some further discussion can be found in Davidson and MacKinnon (2004). Ruud (2000) provides many of the theoretical details. Hayashi (2000) is another extensive treatment of estimation centered on GMM estimators.

where $f(\cdot)$ is the density function and $\boldsymbol{\theta}$ is the parameter vector. For densities that satisfy the regularity conditions [see Section 14.4.1],

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

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

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

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

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

13.4.2 GENERALIZING THE METHOD OF MOMENTS

The preceding examples all have a common aspect. In each case listed, save for the general case of the instrumental variable estimator, there are exactly as many moment equations as there are parameters to be estimated. Thus, each of these are **exactly identified** cases. There will be a single solution to the moment equations, and at that solution, the equations will be exactly satisfied.[5] But there are cases in which there are more moment equations than parameters, so the system is overdetermined.

In Example 13.5, 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 $m_{il}(\boldsymbol{\theta})$ indicates the dependence on $(y_i, \mathbf{x}_i, \mathbf{z}_i)$. Denote the corresponding sample means as

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

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

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

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

will not have a unique solution.[6] For convenience, the moment equations are defined implicitly here as opposed to equalities of moments to functions as in Section 13.3. 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,[7]

$$q = \sum_{l=1}^{L} \bar{m}_l^2 = \bar{\mathbf{m}}(\boldsymbol{\theta})'\bar{\mathbf{m}}(\boldsymbol{\theta}). \tag{13-2}$$

It can be shown that under the assumptions we have made so far, specifically that plim $\bar{\mathbf{m}}(\boldsymbol{\theta}) = E[\bar{\mathbf{m}}(\boldsymbol{\theta})] = \mathbf{0}$, the minimizer of q in (13-2) produces a consistent, though possibly inefficient, estimator of $\boldsymbol{\theta}$.[8] We can 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 (13-2), to produce a consistent estimator of $\boldsymbol{\theta}$.[9] 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 plim $\mathbf{W}_n =$ a positive definite matrix, $\mathbf{W}$.

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

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

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

$$q = \bar{\mathbf{m}}(\boldsymbol{\theta})'\boldsymbol{\Phi}^{-1}\bar{\mathbf{m}}(\boldsymbol{\theta}). \tag{13-3}$$

In general, the L elements of $\bar{\mathbf{m}}$ are freely correlated. In (13-3), we have used a diagonal $\mathbf{W}$ that ignores this correlation. To use generalized least squares, we would define the full matrix,

$$\mathbf{W} = \{\text{Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}]\}^{-1} = \boldsymbol{\Phi}^{-1}. \tag{13-4}$$

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

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

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

[7] This approach is one that Quandt and Ramsey (1978) suggested for the problem in Example 13.4.

[8] See, for example, Hansen (1982).

[9] In principle, the weighting matrix can be a function of the parameters as well. See Hansen, Heaton, and Yaron (1996) for discussion. Whether this provides any benefit in terms of the asymptotic properties of the estimator seems unlikely. The one payoff the authors do note is that certain estimators become invariant to the sort of normalization that is discussed in Example 14.1. In practical terms, this is likely to be a consideration only in a fairly small class of cases.

are minimum distance estimators as defined in Section 13.3. The general result is that if $\mathbf{W}_n$ is a positive definite matrix and if

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

then the minimum distance (GMM) estimator of $\boldsymbol{\theta}$ is consistent.[10] Because the OLS criterion in (13-2) 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 (13-4) 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 (GMM) estimator** is

$$\mathbf{V}_{GMM} = \frac{1}{n}[\boldsymbol{\Gamma}'\mathbf{W}\boldsymbol{\Gamma}]^{-1} = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{\Phi}^{-1}\boldsymbol{\Gamma}]^{-1}, \tag{13-5}$$

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

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

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 13.4.3.

Example 13.7 GMM Estimation of a Nonlinear Regression Model

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

$$\textit{Income} = h(1, \textit{Age}, \textit{Education}, \textit{Female}, \boldsymbol{\gamma}) + \varepsilon,$$

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

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

which, for convenience, we will write $y_i = \exp(\mathbf{x}_i'\boldsymbol{\gamma}) + \varepsilon_i = \mu_i + \varepsilon_i$.[11] The sample consists of the 1988 wave of the panel, less two observations for which *Income* equals zero. The resulting sample contains 4,481 observations. Descriptive statistics for the sample data are given in Table 7.2. We will first consider nonlinear least squares estimation of the parameters. The normal equations for nonlinear least squares will be

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

Note that the orthogonality condition involves the pseudoregressors, $\partial\mu_i/\partial\gamma = \mathbf{x}_i^0 = \mu_i\mathbf{x}_i$. The implied population moment equation is $E[\varepsilon_i(\mu_i\mathbf{x}_i)] = \mathbf{0}$. Computation of the nonlinear least squares estimator is discussed in Section 7.2.8. The estimator of the asymptotic covariance matrix is

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\gamma}}_{\text{NLSQ}}] = \frac{\Sigma_{i=1}^{n}(y_i - \hat{\mu}_i)^2}{(4{,}481 - 4)}\left[\sum_{i=1}^{4{,}481}(\hat{\mu}_i\mathbf{x}_i)(\hat{\mu}_i\mathbf{x}_i)'\right]^{-1}, \quad \text{where } \hat{\mu}_i = \exp(\mathbf{x}_i'\hat{\boldsymbol{\gamma}}).$$

[10] 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).

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

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

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

implies four moment equations. The sample counterparts will be

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

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

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

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

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

$$\hat{\boldsymbol{\Phi}} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{4{,}481}\mathbf{m}_i(\hat{\boldsymbol{\gamma}}^0)\mathbf{m}_i'(\hat{\boldsymbol{\gamma}}^0)\right\} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{4{,}481}(\hat{\varepsilon}_i^0\mathbf{x}_i)(\hat{\varepsilon}_i^0\mathbf{x}_i)'\right\}$$

$$\overline{\mathbf{G}} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{n}\mathbf{x}_i(-\mu_i^0\mathbf{x}_i)'\right\}.$$

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

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

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

$$E\left[\varepsilon_i \begin{pmatrix} Health\ Satisfaction_i \\ Marital\ Status_i \end{pmatrix}\right],$$

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

$$\boldsymbol{\Phi} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{4{,}481}\overline{\mathbf{m}}_i(\hat{\boldsymbol{\gamma}})\mathbf{m}_i'(\hat{\boldsymbol{\gamma}})\right\} = \left\{\frac{1}{4{,}481}\sum_{i=1}^{4{,}481}(\hat{\varepsilon}_i\mathbf{z}_i)(\hat{\varepsilon}_i\mathbf{z}_i)'\right\},$$

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

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

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

TABLE 13.2 Nonlinear Regression Estimates (Standard errors in parentheses)

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

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

13.4.3 PROPERTIES OF THE GMM ESTIMATOR

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

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

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

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

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

where

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

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

Example 13.8 Empirical Moment Equation for Instrumental Variables

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

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

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

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

ASSUMPTION 13.1 Convergence of the Empirical Moments

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

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

The laws of large numbers that we examined in Appendix D accommodate cases of independent observations. Cases of dependent or correlated observations can be gathered under the **Ergodic theorem** (20.1). For this more general case, then, we would assume that the sequence of observations $\mathbf{m}(\boldsymbol{\theta})$ constitutes a jointly $(L \times 1)$ stationary and ergodic process.

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

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

converge to a probability limit, say, plim $\overline{\mathbf{G}}_n(\boldsymbol{\theta}_0) = \overline{\mathbf{G}}(\boldsymbol{\theta}_0)$. [See (13-1), (13-5), and Theorem 13.1.] For sets of *independent* observations, the continuity of the functions and the derivatives will allow us to invoke the Slutsky theorem to obtain this result. For the more general case of sequences of *dependent* observations, Theorem 20.2, Ergodicity of Functions, will provide a counterpart to the Slutsky theorem for time-series data. In sum, if the moments themselves obey a law of large numbers, then it is reasonable to assume that the derivatives do as well.

ASSUMPTION 13.2 Identification

For any $n \geq K$, if $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are two different parameter vectors, then there exist data sets such that $\overline{\mathbf{m}}_n(\boldsymbol{\theta}_1) \neq \overline{\mathbf{m}}_n(\boldsymbol{\theta}_2)$. Formally, in Section 12.5.3, identification is defined to imply that the probability limit of the GMM criterion function is uniquely minimized at the true parameters, θ_0.

Assumption 13.2 is a practical prescription for identification. More formal conditions are discussed in Section 12.5.3. We have examined two violations of this crucial assumption. In the linear regression model, one of the assumptions is full rank of the matrix of exogenous variables—the absence of multicollinearity in $\mathbf{X}$. In our discussion of the maximum likelihood estimator, we will encounter a case (Example 14.1) in which a normalization is needed to identify the vector of parameters.[12] Both of these cases are included in this assumption. The identification condition has three important implications:

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

Assumptions 13.1 and 13.2 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 13.3 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 $\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) \xrightarrow{d} N[\mathbf{0}, \boldsymbol{\Phi}]$.

The underlying requirements on the data for this assumption to hold will vary and will be complicated if the observations comprising the empirical moment are not independent. For samples of independent observations, we assume the conditions underlying the Lindeberg–Feller (D.19) or Liapounov central limit theorem (D.20) will suffice. For the more general case, it is once again necessary to make some assumptions about the data. We have assumed that $E[\mathbf{m}_i(\boldsymbol{\theta}_0)] = \mathbf{0}$. If we can go a step further and assume that the functions $\mathbf{m}_i(\boldsymbol{\theta}_0)$ are an ergodic, stationary **martingale difference sequence**, $E[\mathbf{m}_i(\boldsymbol{\theta}_0) \,|\, \mathbf{m}_{i-1}(\boldsymbol{\theta}_0), \mathbf{m}_{i-2}(\boldsymbol{\theta}_0) \ldots] = \mathbf{0}$, then we can invoke Theorem 20.3, the central limit theorem for the 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

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

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

[12] See Hansen et al. (1996) for discussion of this case.

With the assumptions in place, we have

THEOREM 13.2 Asymptotic Distribution of the GMM Estimator
Under the preceding assumptions,

$$\hat{\boldsymbol{\theta}}_{GMM} \xrightarrow{p} \boldsymbol{\theta}_0,$$

$$\hat{\boldsymbol{\theta}}_{GMM} \overset{a}{\sim} N[\boldsymbol{\theta}_0, \mathbf{V}_{GMM}], \tag{13-6}$$

where $\mathbf{V}_{GMM}$ *is defined in* (*13-5*).

We will now sketch a proof of Theorem 13.2. The GMM estimator is obtained by minimizing the criterion function,

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

where $\mathbf{W}_n$ is the weighting matrix used. Consistency of the estimator that minimizes this criterion can be established by the same logic that will be 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 13.1, $q_n(\boldsymbol{\theta}_0)$ 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. Because $\mathbf{W}_n$ is positive definite, for any finite n, we know that

$$0 \leq q_n(\hat{\boldsymbol{\theta}}_{GMM}) \leq q_n(\boldsymbol{\theta}_0). \tag{13-7}$$

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 (13-7) is true, then it must follow that $q_n(\hat{\boldsymbol{\theta}}_{GMM}) \xrightarrow{p} 0$ as well because of the identification assumption, 13.2. 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 $n \to \infty$, $\overline{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) \to \overline{\mathbf{m}}_n(\boldsymbol{\theta}_0)$, because the function is quadratic and $\mathbf{W}$ is positive definite. The identification condition that we assumed earlier now assures that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_{GMM}$ 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

$$\frac{\partial q_n(\hat{\boldsymbol{\theta}}_{GMM})}{\partial \hat{\boldsymbol{\theta}}_{GMM}} = 2\overline{\mathbf{G}}_n(\hat{\boldsymbol{\theta}}_{GMM})'\mathbf{W}_n\overline{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \mathbf{0}. \tag{13-8}$$

(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, $\boldsymbol{\theta}_0$,

$$\overline{\mathbf{m}}_n(\hat{\boldsymbol{\theta}}_{GMM}) = \overline{\mathbf{m}}_n(\boldsymbol{\theta}_0) + \overline{\mathbf{G}}_n(\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0), \tag{13-9}$$

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 (13-9) in (13-8) to obtain

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

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

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

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 $\overline{\boldsymbol{\theta}}$ both converge to $\boldsymbol{\theta}_0$. By the strict continuity assumed, it must also be the case that

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

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 left-hand side must be the same as that on the right-hand side in (13-10),

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{GMM} - \boldsymbol{\theta}_0) \xrightarrow{\ d\ } \{-[\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\}\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\theta}_0). \tag{13-10}$$

We now invoke Assumption 13.3. The matrix in curled brackets is a set of constants. The last term has the normal limiting distribution given in Assumption 13.3. The mean and variance of this limiting distribution are zero and $\boldsymbol{\Phi}$, respectively. Collecting terms, we have the result in Theorem 13.2, where

$$\mathbf{V}_{GMM} = \frac{1}{n}[\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}\,\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\boldsymbol{\Phi}\mathbf{W}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)[\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\mathbf{W}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}. \tag{13-11}$$

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

$$\mathbf{V}_{GMM,optimal} = \frac{1}{n}[\overline{\mathbf{G}}(\boldsymbol{\theta}_0)'\boldsymbol{\Phi}^{-1}\overline{\mathbf{G}}(\boldsymbol{\theta}_0)]^{-1}. \tag{13-12}$$

Returning to (13-11), there is a special case of interest. If we use least squares or instrumental variables with $\mathbf{W} = \mathbf{I}$, then

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

This equation prescribes essentially the White or **Newey–West estimator**, which returns us to our departure point and provides a neat symmetry to the GMM principle. We will formalize this in Section 13.6.1.

13.5 TESTING HYPOTHESES IN THE GMM FRAMEWORK

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

13.5.1 TESTING THE VALIDITY OF THE MOMENT RESTRICTIONS

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

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

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,

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

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 13.9 Overidentifying Restrictions

In Hall's consumption model, two orthogonality conditions noted in Example 13.1 exactly identify the two parameters. But his analysis of the model suggests a way to test the specification. The conclusion, "No information available in time t apart from the level of consumption, c_t, helps predict future consumption, c_{t+1}, in the sense of affecting the expected value of marginal utility. In particular, income or wealth in periods t or earlier are irrelevant once c_t is known," suggests how one might test the model. If lagged values of income (Y_t might equal the ratio of current income to the previous period's income) are added to the set of instruments, then the model is now overidentified by the orthogonality conditions,

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

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

The preceding is a **specification test**, not a test of parametric restrictions. However, there is a symmetry between the moment restrictions and restrictions on the parameter vector. Suppose $\boldsymbol{\theta}$ is subjected to J restrictions (linear or nonlinear) that restrict the number of free parameters from K to $K - J$. (That is, reduce the dimensionality of the parameter space from K to $K - J$.) The nature of the GMM estimation problem we have posed is not changed at all by the restrictions. The constrained problem may be stated in terms of

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

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

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

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

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

The test is attributed to Newey and West (1987b). This provides one method of testing a set of restrictions. (The small-sample properties of this test will be the central focus of the application discussed in Section 13.6.5.) We now consider several alternatives.

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

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

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

where

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

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

$$W = \text{Wald} = [\mathbf{r}(\mathbf{c}_1)]'\{\text{Est.Asy.Var}[\mathbf{r}(\mathbf{c}_1)]\}^{-1}[\mathbf{r}(\mathbf{c}_1)]. \tag{13-14}$$

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

$$\text{Est.Asy.Var}[\mathbf{r}(\mathbf{c}_1)] = \mathbf{R}_1\{\text{Est.Asy.Var}[\mathbf{c}_1]\}\mathbf{R}_1', \tag{13-15}$$

where

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

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

$$\text{LM} = \text{Lagrange multiplier} = \mathbf{g}_1'(\mathbf{c}_0)\{\text{Est.Asy.Var}[\mathbf{g}_1(\mathbf{c}_0)]\}^{-1}\mathbf{g}_1(\mathbf{c}_0), \tag{13-16}$$

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 $\text{Est.Asy.Var}[\mathbf{g}_1(\mathbf{c}_0)]$ is the 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.[13] That is, in (13-14), we would use the unrestricted GMM estimator of $\boldsymbol{\theta}$. The appropriate asymptotic covariance matrix is (13-12). The computation is exactly the same. The counterpart to the LR statistic is the difference in the values of nq in (13-13). It is necessary to use the same weighting matrix, $\mathbf{W}$, in both restricted and unrestricted estimators. Because 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 $\boldsymbol{\Phi}_1^{-1} = \{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}_1(\mathbf{c}_1)]\}^{-1}$. 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 $(nq_0 - nq_1)$.[14] Because we are using the same $\mathbf{W}$ in both cases, this statistic is necessarily nonnegative. (This is the statistic discussed in Section 13.5.1.)

Finally, the counterpart to the LM statistic would be

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

The logic for this LM statistic is the same as that for the MLE. The derivatives of the minimized criterion q in (13-3) evaluated at the restricted estimator are

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

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

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

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

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

The Wald statistic would be

$$\begin{aligned}\text{Wald} &= \mathbf{g}_1(\mathbf{c}_0)'\{\text{Est.Asy.Var}[\mathbf{g}_1(\mathbf{c}_0)]\}^{-1}\mathbf{g}_1(\mathbf{c}_0)\\ &= n\overline{\mathbf{m}}_1'(\mathbf{c}_0)\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0)\{\overline{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0)\}^{-1}\overline{\mathbf{G}}_1(\mathbf{c}_0)'\hat{\boldsymbol{\Phi}}_1^{-1}\overline{\mathbf{m}}_1(\mathbf{c}_0). \quad \textbf{(13-17)}\end{aligned}$$

13.6 GMM ESTIMATION OF ECONOMETRIC MODELS

The preceding has suggested that the GMM approach to estimation broadly encompasses most of the estimators we will encounter in this book. We have implicitly examined least squares and the general method of instrumental variables in the process. In this section,

[13] 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.

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

we will formalize more specifically the GMM estimators for several of the estimators that appear in the earlier chapters. Section 13.6.1 examines the generalized regression model of Chapter 9. Section 13.6.2 describes a relatively minor extension of the GMM/IV estimator to nonlinear regressions. Section 13.6.3 describes the GMM estimators for our models of systems of seemingly unrelated regression (SUR) model. Finally, in Section 13.6.4, we develop one of the major applications of GMM estimation, the Arellano–Bond–Bover estimator for dynamic panel data models.

13.6.1 SINGLE-EQUATION LINEAR MODELS

It is useful to confine attention to the instrumental variables case, as it is fairly general and we can easily specialize it to the simpler regression models if that is appropriate. Thus, we depart from the usual linear model (8-1), but we no longer require that $E[\varepsilon_i|\mathbf{x}_i] = 0$. Instead, we adopt the instrumental variables formulation in Chapter 8. That is, the model is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$
$$E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$$

for K variables in $\mathbf{x}_i$ and for some set of L instrumental variables, $\mathbf{z}_i$, where $L \geq K$. The earlier case of the generalized regression model arises if $\mathbf{z}_i = \mathbf{x}_i$, and the classical regression results if we add $\boldsymbol{\Omega} = \mathbf{I}$ as well, so this is a convenient encompassing model framework.

In Chapter 9 on generalized least squares estimation, we considered two cases, first one with a known $\boldsymbol{\Omega}$, then one with an unknown $\boldsymbol{\Omega}$ that must be estimated. In estimation by the generalized method of moments, neither of these approaches is relevant because we begin with much less (assumed) knowledge about the data-generating process. We will consider three cases:

- Classical regression: $\text{Var}[\varepsilon_i|\mathbf{X}, \mathbf{Z}] = \sigma^2$,
- Heteroscedasticity: $\text{Var}[\varepsilon_i|\mathbf{X}, \mathbf{Z}] = \sigma_i^2$,
- Generalized model: $\text{Cov}[\varepsilon_t, \varepsilon_s|\mathbf{X}, \mathbf{Z}] = \sigma^2\omega_{ts}$,

where $\mathbf{Z}$ and $\mathbf{X}$ are the $n \times L$ and $n \times K$ observed data matrices, respectively. (We assume, as will often be true, that the fully general case will apply in a time-series setting. Hence the change in the subscripts.) No specific distribution is assumed for the disturbances, conditional or unconditional.

The assumption $E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$ implies the following orthogonality condition,

$$\text{Cov}[\mathbf{z}_i, \varepsilon_i] = \mathbf{0}, \quad \text{or} \quad E[\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})] = \mathbf{0}.$$

By summing the terms, we find that this further implies the **population moment equation**,

$$E\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\boldsymbol{\beta})\right] = E[\overline{\mathbf{m}}(\boldsymbol{\beta})] = \mathbf{0}. \tag{13-18}$$

This relationship suggests how we might now proceed to estimate $\boldsymbol{\beta}$. Note, in fact, that if $\mathbf{z}_i = \mathbf{x}_i$, then this is just the population counterpart to the least squares normal equations. So, as a guide to estimation, this would return us to least squares. Suppose we now translate this population expectation into a sample analog and use that as our guide for estimation. That is, if the population relationship holds for the true parameter vector, $\boldsymbol{\beta}$,

suppose we attempt to mimic this result with a sample counterpart, or empirical moment equation,

$$\left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}})\right] = \left[\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\hat{\boldsymbol{\beta}})\right] = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}. \tag{13-19}$$

In the absence of other information about the data-generating process, we can use the empirical moment equation as the basis of our estimation strategy.

The empirical moment condition is L equations (the number of variables in $\mathbf{Z}$) in K unknowns (the number of parameters we seek to estimate). There are three possibilities to consider:

1. **Underidentified.** $L < K$. If there are fewer moment equations than there are parameters, then it will not be possible to find a solution to the equation system in (13-19). With no other information, such as restrictions that would reduce the number of free parameters, there is no need to proceed any further with this case.

   For the identified cases, it is convenient to write (13-19) as

   $$\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n}\mathbf{Z}'\mathbf{y}\right) - \left(\frac{1}{n}\mathbf{Z}'\mathbf{X}\right)\hat{\boldsymbol{\beta}}. \tag{13-20}$$

2. **Exactly identified.** If $L = K$, then you can easily show (we leave it as an exercise) that the single solution to our equation system is the familiar instrumental variables estimator from Section 8.3.2,

   $$\hat{\boldsymbol{\beta}} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{13-21}$$

3. **Overidentified.** If $L > K$, then there is no unique solution to the equation system $\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. In this instance, we need to formulate some strategy to choose an estimator. One intuitively appealing possibility which has served well thus far is least squares. In this instance, that would mean choosing the estimator based on the criterion function,

   $$\text{Min}_{\boldsymbol{\beta}}\, q = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}})'\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}).$$

   We do keep in mind that we will only be able to minimize this at some positive value; there is no exact solution to (13-19) in the overidentified case. Also, you can verify that if we treat the exactly identified case as if it were overidentified, that is, use least squares anyway, we will still obtain the IV estimator shown in (13-21) for the solution to case (2). For the overidentified case, the first-order conditions are

   $$\begin{aligned}\frac{\partial q}{\partial \hat{\boldsymbol{\beta}}} &= 2\left(\frac{\partial \overline{\mathbf{m}}'(\hat{\boldsymbol{\beta}})}{\partial \hat{\boldsymbol{\beta}}}\right)\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) = 2\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}) \\ &= 2\left(\frac{1}{n}\mathbf{X}'\mathbf{Z}\right)\left(\frac{1}{n}\mathbf{Z}'\mathbf{y} - \frac{1}{n}\mathbf{Z}'\mathbf{X}\hat{\boldsymbol{\beta}}\right) = \mathbf{0}.\end{aligned} \tag{13-22}$$

   We leave as exercise to show that the solution in both cases (2) and (3) is now

   $$\hat{\boldsymbol{\beta}} = [(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{y}). \tag{13-23}$$

   The estimator in (13-23) is a hybrid that we have not encountered before, though if $L = K$, then it does reduce to the earlier one in (13-21). (In the overidentified case, (13-23) is not an IV estimator, it is, as we have sought, a **method of moments estimator**.)

It remains to establish consistency and to obtain the asymptotic distribution and an asymptotic covariance matrix for the estimator. The intermediate results we need are Assumptions 13.1, 13.2, and 13.3 in Section 13.4.3:

- **Convergence of the moments.** The sample moment converges in probability to its population counterpart. That is, $\overline{\mathbf{m}}(\boldsymbol{\beta}) \rightarrow \mathbf{0}$. Different circumstances will produce different kinds of convergence, but we will require it in some form. For the simplest cases, such as a model of heteroscedasticity, this will be convergence in mean square. Certain time-series models that involve correlated observations will necessitate some other form of convergence. But, in any of the cases we consider, we will require the general result: plim $\overline{\mathbf{m}}(\boldsymbol{\beta}) = \mathbf{0}$.
- **Identification.** The parameters are identified in terms of the moment equations. Identification means, essentially, that a large enough sample will contain sufficient information for us actually to estimate $\boldsymbol{\beta}$ consistently using the sample moments. There are two conditions which must be met—an **order condition**, which we have already assumed ($L \geq K$), and a **rank condition**, which states that the moment equations are not redundant. The rank condition implies the order condition, so we need only formalize it:
- **Identification condition for GMM estimation.** The $L \times K$ matrix,

$$\boldsymbol{\Gamma}(\boldsymbol{\beta}) = E[\overline{\mathbf{G}}(\boldsymbol{\beta})] = \text{plim } \overline{\mathbf{G}}(\boldsymbol{\beta}) = \text{plim } \frac{\partial \overline{\mathbf{m}}}{\partial \boldsymbol{\beta}'} = \text{plim } \frac{1}{n}\sum_{i=1}^{n}\frac{\partial \mathbf{m}_i}{\partial \boldsymbol{\beta}'},$$

  must have row rank equal to K.[15] Because this requires $L \geq K$, this implies the order condition. This assumption means that this derivative matrix converges in probability to its expectation. Note that we have assumed, in addition, that the derivatives, like the moments themselves, obey a law of large numbers—they converge in probability to their expectations.
- **Limiting Normal Distribution for the Sample Moments.** The population moment obeys a central limit theorem. Because we are studying a generalized regression model, Lindeberg–Levy (D.18) will be too narrow—the observations will have different variances. Lindeberg–Feller (D.19.A) suffices in the heteroscedasticity case, but in the general case, we will ultimately require something more general. See Section 13.4.3.

It will follow from Assumptions 13.1–13.3 (again, at this point we do this without proof) that the GMM estimators that we obtain are, in fact, consistent. By virtue of the Slutsky theorem, we can transfer our limiting results to the empirical moment equations.

To obtain the asymptotic covariance matrix we will simply invoke the general result for GMM estimators in Section 13.4.3. That is,

$$\text{Asy.Var}[\hat{\boldsymbol{\beta}}] = \frac{1}{n}[\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1}\boldsymbol{\Gamma}'\{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})]\}\boldsymbol{\Gamma}[\boldsymbol{\Gamma}'\boldsymbol{\Gamma}]^{-1}.$$

For the particular model we are studying here,

$$\begin{aligned}\overline{\mathbf{m}}(\boldsymbol{\beta}) &= (1/n)(\mathbf{Z}'\mathbf{y} - \mathbf{Z}'\mathbf{X}\boldsymbol{\beta}),\\ \overline{\mathbf{G}}(\boldsymbol{\beta}) &= (1/n)\mathbf{Z}'\mathbf{X},\\ \boldsymbol{\Gamma}(\boldsymbol{\beta}) &= \mathbf{Q}_{\mathbf{ZX}} \text{ (see Section 8.3.2)}\end{aligned}$$

[15] We require that the row rank be at least as large as K. There could be redundant, that is, functionally dependent, moments, so long as there are at least K that are functionally independent.

(You should check in the preceding expression that the dimensions of the particular matrices and the dimensions of the various products produce the correctly configured matrix that we seek.) The remaining detail, which is the crucial one for the model we are examining, is for us to determine,

$$\mathbf{V} = \text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})].$$

Given the form of $\overline{\mathbf{m}}(\boldsymbol{\beta})$,

$$\mathbf{V} = \frac{1}{n}\text{Var}\left[\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right] = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sigma^2\omega_{ij}\mathbf{z}_i\mathbf{z}_j' = \sigma^2\frac{\mathbf{Z}'\boldsymbol{\Omega}\mathbf{Z}}{n}$$

for the most general case. Note that this is precisely the expression that appears in (9-9), so the question that arose there arises here once again. That is, under what conditions will this converge to a constant matrix? We take the discussion there as given. The only remaining detail is how to estimate this matrix. The answer appears in Section 9.2, where we pursued this same question in connection with robust estimation of the asymptotic covariance matrix of the least squares estimator. To review then, what we have achieved to this point is to provide a theoretical foundation for the instrumental variables estimator. As noted earlier, this specializes to the least squares estimator. The estimators of $\mathbf{V}$ for our three cases will be

- Classical regression:

$$\hat{\mathbf{V}} = \frac{(\mathbf{e}'\mathbf{e}/n)}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i' = \frac{(\mathbf{e}'\mathbf{e}/n)}{n}\mathbf{Z}'\mathbf{Z}.$$

- Heteroscedastic regression:

$$\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}e_i^2\mathbf{z}_i\mathbf{z}_i'. \tag{13-24}$$

- Generalized regression:

$$\hat{\mathbf{V}} = \frac{1}{n}\left[\sum_{t=1}^{n}e_t^2\mathbf{z}_t\mathbf{z}_t' + \sum_{\ell=1}^{p}\left(1 - \frac{\ell}{(p+1)}\right)\sum_{t=\ell+1}^{n}e_te_{t-\ell}(\mathbf{z}_t\mathbf{z}_{t-\ell}' + \mathbf{z}_{t-\ell}\mathbf{z}_t')\right].$$

We should observe that in each of these cases, we have actually used some information about the structure of $\boldsymbol{\Omega}$. If it is known only that the terms in $\overline{\mathbf{m}}(\boldsymbol{\beta})$ are uncorrelated, then there is a convenient estimator available, $\hat{\mathbf{V}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i(\hat{\boldsymbol{\beta}})\mathbf{m}_i(\hat{\boldsymbol{\beta}})'$, that is, the natural, empirical variance estimator. Note that this is what is being used in the heteroscedasticity case in (13-24).

Collecting all the terms so far, then, we have

$$\begin{aligned}\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] &= \frac{1}{n}[\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})]^{-1}\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})[\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})]^{-1}\\ &= n[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}(\mathbf{Z}'\mathbf{X})[(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{X})]^{-1}.\end{aligned} \tag{13-25}$$

The preceding might seem to endow the least squares or method of moments estimators with some degree of optimality, but that is not the case. We have only provided them with a different statistical motivation (and established consistency). We now consider the question of whether, because this is the generalized regression model, there is some better (more efficient) means of using the data.

The class of minimum distance estimators for this model is defined by the solutions to the criterion function, $\text{Min}_{\boldsymbol{\beta}}\, q = \overline{\mathbf{m}}(\boldsymbol{\beta})'\mathbf{W}\overline{\mathbf{m}}(\boldsymbol{\beta})$, where $\mathbf{W}$ is *any* positive definite **weighting matrix**. Based on the assumptions just made, we can invoke Theorem 13.1 to obtain

$$\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{MD}] = \frac{1}{n}[\overline{\mathbf{G}}'\mathbf{W}\overline{\mathbf{G}}]^{-1}\overline{\mathbf{G}}'\mathbf{W}\mathbf{V}\mathbf{W}\overline{\mathbf{G}}[\overline{\mathbf{G}}'\mathbf{W}\overline{\mathbf{G}}]^{-1}.$$

Note that our entire preceding analysis was of the simplest minimum distance estimator, which has $\mathbf{W} = \mathbf{I}$. The obvious question now arises, if any $\mathbf{W}$ produces a consistent estimator, is any $\mathbf{W}$ better than any other one, or is it simply arbitrary? There is a firm answer, for which we have to consider two cases separately:

- **Exactly identified case.** If $L = K$; that is, if the number of moment conditions is the same as the number of parameters being estimated, then $\mathbf{W}$ is irrelevant to the solution, so on the basis of simplicity alone, the optimal $\mathbf{W}$ is $\mathbf{I}$.
- **Overidentified case.** In this case, the "optimal" weighting matrix, that is, the $\mathbf{W}$ that produces the most efficient estimator, is $\mathbf{W} = \mathbf{V}^{-1}$. The best weighting matrix is the inverse of the asymptotic covariance of the moment vector. In this case, the MDE will be the GMM estimator with

$$\hat{\boldsymbol{\beta}}_{GMM} = [(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}(\mathbf{X}'\mathbf{Z})\hat{\mathbf{V}}^{-1}(\mathbf{Z}'\mathbf{y}),$$

and

$$\begin{aligned}\text{Asy.Var}[\hat{\boldsymbol{\beta}}_{GMM}] &= \frac{1}{n}[\overline{\mathbf{G}}'\mathbf{V}^{-1}\overline{\mathbf{G}}]^{-1} \\ &= n[(\mathbf{X}'\mathbf{Z})\mathbf{V}^{-1}(\mathbf{Z}'\mathbf{X})]^{-1}.\end{aligned}$$

We conclude this discussion by tying together what should seem to be a loose end. The GMM estimator is computed as the solution to

$$\text{Min}_{\boldsymbol{\beta}}\, q = \overline{\mathbf{m}}(\boldsymbol{\beta})'\{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})\}^{-1}\overline{\mathbf{m}}(\boldsymbol{\beta}),$$

which might suggest that the weighting matrix is a function of the thing we are trying to estimate. The process of GMM estimation will have to proceed in two steps: Step 1 is to obtain an estimate of $\mathbf{V}$; Step 2 will consist of using the inverse of this $\mathbf{V}$ as the weighting matrix in computing the GMM estimator. The following is a common two-step strategy:

Step 1. Use $\mathbf{W} = \mathbf{I}$ to obtain a consistent estimator of $\boldsymbol{\beta}$. Then, in the heteroscedasticity case (i.e., the White estimator), estimate $\mathbf{V}$ with $\hat{\mathbf{V}} = (1/n)\sum_{i=1}^{n} e_i^2\mathbf{z}_i\mathbf{z}_i'$. For the more general case, use the Newey–West estimator.

Step 2. Use $\mathbf{W} = \hat{\mathbf{V}}^{-1}$ to compute the GMM estimator.

By this point, the observant reader should have noticed that in all of the preceding, we have never actually encountered the two-stage least squares estimator that we introduced in Section 8.4.1. To obtain this estimator, we must revert back to the classical, that is, homoscedastic and nonautocorrelated disturbances case. In that instance, the

weighting matrix in Theorem 13.2 will be $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ and we will obtain the apparently missing result.

The **GMM estimator** in the heteroscedastic regression model is produced by the empirical moment equations

$$\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{GMM}) = \frac{1}{n}\mathbf{X}'\hat{\boldsymbol{\varepsilon}}(\hat{\boldsymbol{\beta}}_{GMM}) = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}) = \mathbf{0}. \tag{13-26}$$

The estimator is obtained by minimizing

$$q = \overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM})\mathbf{W}\overline{\mathbf{m}}(\hat{\boldsymbol{\beta}}_{GMM}),$$

where $\mathbf{W}$ is a positive definite weighting matrix. The optimal weighting matrix would be

$$\mathbf{W} = \{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})]\}^{-1},$$

which is the inverse of

$$\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}(\boldsymbol{\beta})] = \text{Asy.Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{x}_i\boldsymbol{\varepsilon}_i\right] = \underset{n\to\infty}{\text{plim}}\frac{1}{n}\sum_{i=1}^{n}\sigma^2\omega_i\mathbf{x}_i\mathbf{x}_i' = \sigma^2\mathbf{Q}^*.$$

(See Section 9.4.1.) The optimal weighting matrix would be $[\sigma^2\mathbf{Q}^*]^{-1}$. But recall that this minimization problem is an exactly identified case, so the weighting matrix is irrelevant to the solution. You can see the result in the moment equation—that equation is simply the normal equation for ordinary least squares. We can solve the moment equations exactly, so there is no need for the weighting matrix. Regardless of the covariance matrix of the moments, the GMM estimator for the heteroscedastic regression model is ordinary least squares. We can use the results we have already obtained to find its asymptotic covariance matrix. The implied estimator is the White estimator in (9-5). (Once again, see Theorem 13.2.) The conclusion to be drawn at this point is that until we make some specific assumptions about the variances, we do not have a more efficient estimator than least squares, but we do have to modify the estimated asymptotic covariance matrix.

13.6.2 SINGLE-EQUATION NONLINEAR 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, because it is possible that $\text{Cov}[\varepsilon_i, h(\mathbf{x}_i, \boldsymbol{\beta})] \neq 0$, or even $\text{Cov}[\varepsilon_i, \mathbf{x}_j] \neq 0$ for all i and j. Consider, for example, a model that contains lagged dependent variables and autocorrelated disturbances. (See Section 20.9.3.) For the present, we assume that $E[\boldsymbol{\varepsilon}|\mathbf{X}] \neq \mathbf{0}$, and $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \sigma^2\boldsymbol{\Omega} = \boldsymbol{\Sigma}$ where Σ 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 $\boldsymbol{\varepsilon}_i$. You will recognize $\mathbf{z}_i$ as a set of **instrumental variables**. The assumptions thus far have implied a set of orthogonality conditions, $E[\mathbf{z}_i\varepsilon_i] = \mathbf{0}$, which may be sufficient to identify (if $L = K$) or even overidentify (if $L > K$) the parameters of the model. (See Section 8.3.4.)

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

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

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

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

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

$$q = \overline{\mathbf{m}}_n(\hat{\boldsymbol{\beta}})'\mathbf{W}\overline{\mathbf{m}}_n(\hat{\boldsymbol{\beta}}) = \left(\frac{1}{n}[\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})'\mathbf{Z}]\right)\mathbf{W}\left(\frac{1}{n}[\mathbf{Z}'\mathbf{e}(\mathbf{X}, \hat{\boldsymbol{\beta}})]\right) \tag{13-27}$$

for some choice of $\mathbf{W}$ that we have yet to determine. The criterion given earlier 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 8.9, where we defined the nonlinear instrumental variables estimator. Apparently (13-27) is more general, because we are not limited to this choice of $\mathbf{W}$. 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. The optimal choice of $\mathbf{W}$ for this estimator is

$$\begin{aligned}\mathbf{W}_{\text{GMM}} &= \{\text{Asy.Var}[\sqrt{n}\,\overline{\mathbf{m}}_n(\boldsymbol{\beta})]\}^{-1} \\ &= \left\{\text{Asy.Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\mathbf{z}_i\varepsilon_i\right]\right\}^{-1} = \left\{\text{Asy.Var}\left[\frac{1}{\sqrt{n}}\mathbf{Z}'\mathbf{e}(\mathbf{X}, \boldsymbol{\beta})\right]\right\}^{-1}.\end{aligned} \tag{13-28}$$

For our model, this is

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

If we insert this result in (13-27), 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 Section 9.3.2. 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-observation terms may be omitted, and what is required is

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

We can use a counterpart to the White (1980) estimator discussed in Section 9.2 for this case,

$$\mathbf{S}_0 = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\mathbf{z}_i'(y_i - h(\mathbf{x}_i, \hat{\boldsymbol{\beta}}))^2. \tag{13-29}$$

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),

$$\mathbf{S} = \left[\mathbf{S}_0 + \frac{1}{n}\sum_{\ell=1}^{p} w(\ell) \sum_{i=\ell+1}^{n} (e_i e_{i-\ell})(\mathbf{z}_i\mathbf{z}_{i-\ell}' + \mathbf{z}_{i-\ell}\mathbf{z}_i')\right] = \sum_{\ell=0}^{p} w(\ell)\mathbf{S}_\ell, \tag{13-30}$$

where $\mathrm{w}(\ell) = 1 - \ell/(p + 1)$. (This is the *Bartlett weight*.) 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 $n^{1/4}$.

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 (13-27). 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 Theorem 13.2. 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

$$\bar{\mathbf{G}}(\boldsymbol{\beta}) = \partial\bar{\mathbf{m}}(\boldsymbol{\beta})/\partial\boldsymbol{\beta}' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{G}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{z}_i\left[\frac{\partial\varepsilon_i}{\partial\boldsymbol{\beta}'}\right]. \tag{13-31}$$

In the model we are considering here, $\dfrac{\partial\varepsilon_i}{\partial\boldsymbol{\beta}'} = \dfrac{-\partial h(\mathbf{x}_i, \boldsymbol{\beta})}{\partial\boldsymbol{\beta}'}$. The derivatives are the pseudoregressors in the linearized regression model that we examined in Section 7.2.3. Using the notation defined there, $\dfrac{\partial\varepsilon_i}{\partial\boldsymbol{\beta}} = -\mathbf{x}_i^0$, so

$$\bar{\mathbf{G}}(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{G}_i(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} -\mathbf{z}_i\mathbf{x}_i^{0\prime} = -\frac{1}{n}\mathbf{Z}'\mathbf{X}^0. \tag{13-32}$$

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

$$\text{Est.Asy.Var}[\hat{\boldsymbol{\beta}}] = \frac{1}{n}\left[\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})'\left(\frac{1}{n}\mathbf{Z}'\hat{\boldsymbol{\Sigma}}\mathbf{Z}\right)^{-1}\overline{\mathbf{G}}(\hat{\boldsymbol{\beta}})\right]^{-1} = [(\mathbf{X}^{0\prime}\mathbf{Z})(\mathbf{Z}'\hat{\boldsymbol{\Sigma}}\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X}^{0})]^{-1}. \tag{13-33}$$

(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 (13-33) were $\sigma^2\mathbf{I}$, then (13-33) would be precisely the asymptotic covariance matrix that appears in Theorem 8.1 for linear models and Theorem 8.2 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. You should be able to show that when $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ *and we use this information,* the same parameter estimates will be obtained when all is said and done. But, if we use a weighting matrix that differs from $\mathbf{W} = (\mathbf{Z}'\mathbf{Z}/n)^{-1}$, then they are not.

13.6.3 SEEMINGLY UNRELATED REGRESSION EQUATIONS

In Section 10.2.3, we considered FGLS estimation of the equation system

$$\begin{aligned}\mathbf{y}_1 &= \mathbf{h}_1(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_1,\\ \mathbf{y}_2 &= \mathbf{h}_2(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_2,\\ &\vdots\\ \mathbf{y}_M &= \mathbf{h}_M(\mathbf{X}, \boldsymbol{\beta}) + \boldsymbol{\varepsilon}_M.\end{aligned}$$

The development there extends backward to the linear system as well. However, none of the estimators considered is consistent if the pseudoregressors, $\mathbf{x}_{tm}^0$, or the actual regressors, $\mathbf{x}_{tm}$, for the linear model, are correlated with the disturbances, ε_{tm}. Suppose we allow for this correlation both within and across equations. (If it is, in fact, absent, then the GMM estimator developed here will remain consistent.) For simplicity in this section, we will denote observations with subscript t and equations with subscripts i and j. Suppose, as well, that there are a set of instrumental variables, $\mathbf{z}_t$, such that

$$E[\mathbf{z}_t\varepsilon_{tm}] = \mathbf{0}, t = 1, \ldots, T \text{ and } m = 1, \ldots, M. \tag{13-34}$$

(We could allow a separate set of instrumental variables for each equation, but it would needlessly complicate the presentation.)

Under these assumptions, the nonlinear FGLS and ML estimators given earlier will be inconsistent. But a relatively minor extension of the instrumental variables technique developed for the single-equation case in Section 8.4 can be used instead. The sample analog to (13-34) is

$$\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t[y_{ti} - h_i(\mathbf{X}_t, \boldsymbol{\beta})] = \mathbf{0}, \quad i = 1, \ldots, M.$$

If we use this result for each equation in the system, one at a time, then we obtain exactly the GMM estimator discussed in Section 13.6.2. But, in addition to the efficiency loss

that results from not imposing the cross-equation constraints in $\boldsymbol{\beta}$, we would also neglect the correlation between the disturbances. Let

$$\frac{1}{T}\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z} = E\left[\frac{\mathbf{Z}'\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}_j'\mathbf{Z}}{T}\right]. \tag{13-35}$$

The GMM criterion for estimation in this setting is

$$\begin{aligned} q &= \sum_{i=1}^{M}\sum_{j=1}^{M}[(\mathbf{y}_i - \mathbf{h}_i(\mathbf{X},\boldsymbol{\beta}))'\mathbf{Z}/T][\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'(\mathbf{y}_j - \mathbf{h}_j(\mathbf{X},\boldsymbol{\beta}))/T] \\ &= \sum_{i=1}^{M}\sum_{j=1}^{M}[\boldsymbol{\varepsilon}_i(\boldsymbol{\beta})'\mathbf{Z}/T][\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'\boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T], \end{aligned} \tag{13-36}$$

where $[\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T]^{ij}$ denotes the ijth block of the inverse of the matrix with the ijth block equal to $\mathbf{Z}'\boldsymbol{\Omega}_{ij}\mathbf{Z}/T$.

GMM estimation would proceed in several passes. To compute any of the variance parameters, we will require an initial consistent estimator of $\boldsymbol{\beta}$. This step can be done with equation-by-equation nonlinear instrumental variables—see Section 8.9—although if equations have parameters in common, then a choice must be made as to which to use. At the next step, the familiar White or Newey–West technique is used to compute, block by block, the matrix in (13-35). Because it is based on a consistent estimator of $\boldsymbol{\beta}$ (we assume), this matrix need not be recomputed. Now, with this result in hand, an iterative solution to the maximization problem in (13-36) can be sought, for example, using the methods of Appendix E. The first-order conditions are

$$\frac{\partial q}{\partial \boldsymbol{\beta}} = -2\sum_{i=1}^{M}\sum_{j=1}^{M}[\mathbf{X}_i^0(\boldsymbol{\beta})'\mathbf{Z}/T][\mathbf{Z}'\mathbf{W}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'\boldsymbol{\varepsilon}_j(\boldsymbol{\beta})/T] = \mathbf{0}. \tag{13-37}$$

Note again that the blocks of the inverse matrix in the center are extracted from the larger constructed matrix *after inversion*.[16] At completion, the asymptotic covariance matrix for the GMM estimator is estimated with

$$\mathbf{V}_{\text{GMM}} = \frac{1}{T}\left[\sum_{i=1}^{M}\sum_{j=1}^{M}[\mathbf{X}_i^0(\boldsymbol{\beta})'\mathbf{Z}/T][\mathbf{Z}'\mathbf{W}_{ij}\mathbf{Z}/T]^{ij}[\mathbf{Z}'\mathbf{X}_j^0(\boldsymbol{\beta})/T]\right]^{-1}.$$

13.6.4 GMM ESTIMATION OF DYNAMIC PANEL DATA MODELS

Panel data are well suited for examining dynamic effects, as in the first-order model,

$$\begin{aligned} y_{it} &= \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + c_i + \varepsilon_{it} \\ &= \mathbf{w}_{it}'\boldsymbol{\theta} + \alpha_i + \varepsilon_{it}, \end{aligned}$$

where the set of right-hand-side variables, $\mathbf{w}_{it}$, now includes the lagged dependent variable, $y_{i,t-1}$. Adding dynamics to a model in this fashion creates a major change in the interpretation of the equation. Without the lagged variable, the independent variables represent the full set of information that produce observed outcome y_{it}. With the lagged variable, we now have in the equation the entire history of the right-hand-side variables, so that any measured influence is conditioned on this history; in this case, any impact of $\mathbf{x}_{it}$

[16] This brief discussion might understate the complexity of the optimization problem in (13-36), but that is inherent in the procedure.

represents the effect of *new* information. Substantial complications arise in estimation of such a model. In both the fixed and random effects settings, the difficulty is that the lagged dependent variable is correlated with the disturbance, even if it is assumed that ε_{it} is not itself autocorrelated. For the moment, consider the fixed effects model as an ordinary regression with a lagged dependent variable that is dependent across observations. In that dynamic regression model, the estimator based on T observations is biased in finite samples, but it is consistent in T. The finite sample bias is of order $1/T$. The same result applies here, but the difference is that whereas before we obtained our large sample results by allowing T to grow large, in this setting, T is assumed to be small and fixed, and large-sample results are obtained with respect to n growing large, not T. The fixed effects estimator of $\boldsymbol{\theta} = [\boldsymbol{\beta}, \delta]$ can be viewed as an average of n such estimators. Assume for now that $T \geq K + 1$ where K is the number of variables in $\mathbf{x}_{it}$. Then, from (11-14),

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\right]^{-1}\left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{y}_i\right] \\ &= \left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\right]^{-1}\left[\sum_{i=1}^{n} \mathbf{W}_i'\mathbf{M}^0\mathbf{W}_i\mathbf{d}_i\right] \\ &= \sum_{i=1}^{n} \mathbf{F}_i\mathbf{d}_i,\end{aligned}$$

where the rows of the $T \times (K + 1)$ matrix $\mathbf{W}_i$ are $\mathbf{w}_{it}'$ and $\mathbf{M}^0$ is the $T \times T$ matrix that creates deviations from group means [see (11-14)]. Each group-specific estimator, $\mathbf{d}_i$, is inconsistent, as it is biased in finite samples and its variance does not go to zero as n increases. This matrix weighted average of n inconsistent estimators will also be inconsistent. (This analysis is only heuristic. If $T < K + 1$, then the individual coefficient vectors cannot be computed.[17])

The problem is more transparent in the random effects model. In the model

$$y_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \delta y_{i,t-1} + u_i + \varepsilon_{it},$$

the lagged dependent variable is correlated with the compound disturbance in the model because the same u_i enters the equation for every observation in group i.

Neither of these results renders the model inestimable, but they do make necessary some technique other than our familiar LSDV or FGLS estimators. The general approach, which has been developed in several stages in the literature,[18] relies on instrumental variables estimators and, most recently, on a GMM estimator. For example, in either the fixed or random effects cases, the heterogeneity can be swept from the model by taking first differences, which produces

$$y_{it} - y_{i,t-1} = (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta} + \delta(y_{i,t-1} - y_{i,t-2}) + (\varepsilon_{it} - \varepsilon_{i,t-1}).$$

This model is still complicated by correlation between the lagged dependent variable and the disturbance (and by its first-order moving average disturbance). But without the

[17]Further discussion is given by Nickell (1981), Ridder and Wansbeek (1990), and Kiviet (1995).

[18]The model was first proposed in this form by Balestra and Nerlove (1966). See, for example, Anderson and Hsiao (1981, 1982), Bhargava and Sargan (1983), Arellano (1989), Arellano and Bond (1991), Arellano and Bover (1995), Ahn and Schmidt (1995), and Nerlove (1971a,b).

group effects, there is a simple instrumental variables estimator available. Assuming that the time series is long enough, one could use the lagged differences, $(y_{i,t-2} - y_{i,t-3})$, or the lagged levels, $y_{i,t-2}$ and $y_{i,t-3}$, as one or two instrumental variables for $(y_{i,t-1} - y_{i,t-2})$. (The other variables can serve as their own instruments.) This is the Anderson and Hsiao estimator developed for this model in Section 11.8.3. By this construction, then, the treatment of this model is a standard application of the instrumental variables technique that we developed in Section 11.8.[19] This illustrates the flavor of an instrumental variables approach to estimation. But, as Arellano et al. and Ahn and Schmidt (1995) have shown, there is still more information in the sample that can be brought to bear on estimation, in the context of a GMM estimator, which we now consider.

We can extend the Hausman and Taylor (HT) formulation of the random effects model in Section 11.8.2 to include the lagged dependent variable,

$$\begin{aligned} y_{it} &= \delta y_{i,t-1} + \mathbf{x}_{1it}'\boldsymbol{\beta}_1 + \mathbf{x}_{2it}'\boldsymbol{\beta}_2 + \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 + \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 + \varepsilon_{it} + u_i \\ &= \boldsymbol{\theta}'\mathbf{w}_{it} + \varepsilon_{it} + u_i \\ &= \boldsymbol{\theta}'\mathbf{w}_{it} + \eta_{it}, \end{aligned}$$

where

$$\mathbf{w}_{it} = [y_{i,t-1}, \mathbf{x}_{1it}', \mathbf{x}_{2it}', \mathbf{z}_{1i}', \mathbf{z}_{2i}']'$$

is now a $(1 + K_1 + K_2 + L_1 + L_2) \times 1$ vector. The terms in the equation are the same as in the Hausman and Taylor model. Instrumental variables estimation of the model without the lagged dependent variable is discussed in Section 11.8.1 on the HT estimator. Moreover, by just including $y_{i,t-1}$ in $\mathbf{x}_{2it}$, we see that the HT approach extends to this setting as well, essentially without modification. Arellano et al. suggest a GMM estimator and show that efficiency gains are available by using a larger set of moment conditions. In the previous treatment, we used a GMM estimator constructed as follows: the set of moment conditions we used to formulate the instrumental variables were

$$E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{it} - \bar{\eta}_i)\right] = E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\varepsilon_{it} - \bar{\varepsilon}_i)\right] = \mathbf{0}.$$

This moment condition is used to produce the instrumental variable estimator. We could ignore the nonscalar variance of η_{it} and use simple instrumental variables at this point. However, by accounting for the random effects formulation and using the counterpart to feasible GLS, we obtain the more efficient estimator in Section 11.8.4. As usual, this can be done in two steps. The inefficient estimator is computed to obtain the residuals needed to estimate the variance components. This is Hausman and Taylor's steps 1 and 2. Steps 3 and 4 are the GMM estimator based on these estimated variance components.

[19]There is a question as to whether one should use differences or levels as instruments. Arellano (1989) and Kiviet (1995) give evidence that the latter is preferable.

Arellano et al. suggest that the preceding does not exploit all the information in the sample. In simple terms, within the T observations in group i, we have not used the fact that

$$E\left[\begin{pmatrix}\mathbf{x}_{1it}\\ \mathbf{x}_{2it}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{is}-\bar{\eta}_i)\right]=\mathbf{0}\quad\text{for some } s\neq t.$$

Thus, for example, not only are disturbances at time t uncorrelated with these variables at time t, arguably, they are uncorrelated with the same variables at time $t-1$, $t-2$, possibly $t+1$, and so on. In principle, the number of valid instruments is potentially enormous. Suppose, for example, that the set of instruments listed above is strictly exogenous with respect to η_{it} in every period including current, lagged, and future. Then, there are a total of $[T(K_1+K_2)+L_1+K_1)]$ moment conditions for every observation. Consider, for example, a panel with two periods. We would have for the two periods,

$$E\left[\begin{pmatrix}\mathbf{x}_{1i1}\\ \mathbf{x}_{2i1}\\ \mathbf{x}_{1i2}\\ \mathbf{x}_{2i2}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{i1}-\bar{\eta}_i)\right]=\mathbf{0}\qquad\text{and}\qquad E\left[\begin{pmatrix}\mathbf{x}_{1i1}\\ \mathbf{x}_{2i1}\\ \mathbf{x}_{1i2}\\ \mathbf{x}_{2i2}\\ \mathbf{z}_{1i}\\ \bar{\mathbf{x}}_{1i.}\end{pmatrix}(\eta_{i2}-\bar{\eta}_i)\right]=\mathbf{0}. \tag{13-38}$$

How much useful information is brought to bear on estimation of the parameters is uncertain, as it depends on the correlation of the instruments with the included exogenous variables in the equation. The farther apart in time these sets of variables become, the less information is likely to be present. (The literature on this subject contains reference to *strong* versus *weak* instrumental variables.[20]) To proceed, as noted, we can include the lagged dependent variable in $\mathbf{x}_{2i}$. This set of instrumental variables can be used to construct the estimator, actually whether the lagged variable is present or not. We note, at this point, that on this basis, Hausman and Taylor's estimator did not actually use all the information available in the sample. We now have the elements of the Arellano et al. estimator in hand; what remains is essentially the (unfortunately, fairly involved) algebra, which we now develop.

Let

$$\mathbf{W}_i=\begin{bmatrix}\mathbf{w}_{i1}'\\ \mathbf{w}_{i2}'\\ \vdots\\ \mathbf{w}_{iT}'\end{bmatrix}=\text{the full set of rhs data for group } i,\quad\text{and}\quad\mathbf{y}_i=\begin{bmatrix}y_{i1}\\ y_{i2}\\ \vdots\\ y_{iT}\end{bmatrix}.$$

Note that $\mathbf{W}_i$ is assumed to be a $T\times(1+K_1+K_2+L_1+L_2)$ matrix. Because there is a lagged dependent variable in the model, it must be assumed that there are actually $T+1$ observations available on y_{it}. To avoid cumbersome, cluttered notation, we will leave this distinction embedded in the notation for the moment. Later, when necessary,

[20]See West (2001).

we will make it explicit. It will reappear in the formulation of the instrumental variables. A total of T observations will be available for constructing the IV estimators. We now form a matrix of instrumental variables.[21] We will form a matrix $\mathbf{V_i}$ consisting of $T_i - 1$ rows constructed the same way for $T_i - 1$ observations and a final row that will be different, as discussed later.[22] The matrix will be of the form

$$\mathbf{V}_i = \begin{bmatrix} \mathbf{v}_{i1}' & \mathbf{0}' & \cdots & \mathbf{0}' \\ \mathbf{0}' & \mathbf{v}_{i2}' & \cdots & \mathbf{0}' \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0}' & \mathbf{0}' & \cdots & \mathbf{a}_i' \end{bmatrix}. \tag{13-39}$$

The instrumental variable sets contained in $\mathbf{v}_{it}'$ which have been suggested might include the following from within the model:

$\mathbf{x}_{it}$ and $\mathbf{x}_{i,t-1}$ (i.e., current and one lag of all the time-varying variables),
$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}$ (i.e., all current, past, and future values of all the time-varying variables),
$\mathbf{x}_{i1}, \ldots, \mathbf{x}_{it}$ (i.e., all current and past values of all the time-varying variables).

The time-invariant variables that are uncorrelated with u_i, that is, $\mathbf{z}_{1i}$, are appended at the end of the nonzero part of each of the first $T - 1$ rows. It may seem that including $\mathbf{x}_2$ in the instruments would be invalid. However, we will be converting the disturbances to deviations from group means which are free of the latent effects—that is, this set of moment conditions will ultimately be converted to what appears in (13-38). While the variables are correlated with u_i by construction, they are not correlated with $\varepsilon_{it} - \bar{\varepsilon}_i$. The final row of $\mathbf{V}_i$ is important to the construction. Two possibilities have been suggested:

$\mathbf{a}_i' = [\mathbf{z}_{1i}' \quad \bar{\mathbf{x}}_{i1}]$ (produces the Hausman and Taylor estimator),
$\mathbf{a}_i' = [\mathbf{z}_{1i}' \quad \mathbf{x}_{1i1}', \mathbf{x}_{1i2}', \ldots, x_{1iT}]$ (produces Amemiya and MaCurdy's estimator).

Note that the **a** variables are exogenous time-invariant variables, $\mathbf{z}_{1i}$ and the exogenous time-varying variables, either condensed into the single group mean or in the raw form, with the full set of T observations.

To construct the estimator, we will require a transformation matrix, **H**, constructed as follows. Let $\mathbf{M}^{01}$ denote the first $T - 1$ rows of $\mathbf{M}^0$, the matrix that creates deviations from group means. Then,

$$\mathbf{H} = \begin{bmatrix} \mathbf{M}^{01} \\ \frac{1}{T}\mathbf{i}_T' \end{bmatrix}.$$

Thus, **H** replaces the last row of $\mathbf{M}^0$ with a row of $1/T$. The effect is as follows: if **q** is T observations on a variable, then **Hq** produces **q*** in which the first $T - 1$ observations are converted to deviations from group means and the last observation is the group mean. In particular, let the $T \times 1$ column vector of disturbances,

$$\eta_i = [\eta_{i1}, \eta_{i2}, \ldots, \eta_{iT}] = [(\varepsilon_{i1} + u_i), (\varepsilon_{i2} + u_i), \ldots, (\varepsilon_{i\mathrm{T}} + u_i)]',$$

[21] Different approaches to this have been considered by Hausman and Taylor (1981), Arellano et al. (1991, 1995, 1999), Ahn and Schmidt (1995), and Amemiya and MaCurdy (1986), among others.

[22] This is to exploit a useful algebraic result discussed by Arellano and Bover (1995).

then

$$\mathbf{H}\boldsymbol{\eta} = \begin{bmatrix} \eta_{i1} - \overline{\eta}_i \\ \vdots \\ \eta_{i,T-1} - \overline{\eta}_i \\ \overline{\eta}_i \end{bmatrix}.$$

We can now construct the moment conditions. With all this machinery in place, we have the result that appears in (13-40), that is,

$$E[\mathbf{V}_i'\mathbf{H}\boldsymbol{\eta}_i] = E[\mathbf{g}_i] = \mathbf{0}.$$

It is useful to expand this for a particular case. Suppose $T = 3$ and we use as instruments the current values in period 1, and the current and previous values in period 2 and the Hausman and Taylor form for the invariant variables. Then the preceding is

$$E\left[\begin{pmatrix} \mathbf{x}_{1i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{x}_{2i1} & \mathbf{0} & \mathbf{0} \\ \mathbf{z}_{1i} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i1} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{1i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{x}_{2i2} & \mathbf{0} \\ \mathbf{0} & \mathbf{z}_{1i} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{z}_{1i} \\ \mathbf{0} & \mathbf{0} & \overline{\mathbf{x}}_{1i} \end{pmatrix} \begin{pmatrix} \eta_{i1} - \overline{\eta}_i \\ \eta_{i2} - \overline{\eta}_i \\ \overline{\eta}_i \end{pmatrix}\right] = \mathbf{0}. \tag{13-40}$$

This is the same as (13-38).[23] The empirical moment condition that follows from this is

$$\begin{aligned} &\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\boldsymbol{\eta}_i \\ &\quad = \text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\begin{pmatrix} y_{i1} - \delta y_{i0} - \mathbf{x}_{1i1}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i1}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \\ y_{i2} - \delta y_{i1} - \mathbf{x}_{1i2}'\boldsymbol{\beta}_1 - \mathbf{x}_{2i2}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \\ \vdots \\ y_{iT} - \delta y_{i,T-1} - \mathbf{x}_{1iT}'\boldsymbol{\beta}_1 - \mathbf{x}_{2iT}'\boldsymbol{\beta}_2 - \mathbf{z}_{1i}'\boldsymbol{\alpha}_1 - \mathbf{z}_{2i}'\boldsymbol{\alpha}_2 \end{pmatrix} = \mathbf{0}. \end{aligned}$$

Write this as

$$\text{plim}\,\frac{1}{n}\sum_{i=1}^{n}\mathbf{m}_i = \text{plim}\,\overline{\mathbf{m}} = \mathbf{0}.$$

The GMM estimator $\hat{\boldsymbol{\theta}}$ is then obtained by minimizing $q = \overline{\mathbf{m}}'\mathbf{A}\overline{\mathbf{m}}$ with an appropriate choice of the weighting matrix, $\mathbf{A}$. The optimal weighting matrix will be the inverse of

[23] In some treatments—for example, Blundell and Bond (1998)—an additional condition is assumed for the initial value, y_{i0}, namely $E[y_{i0}|\,\textit{exogenous data}] = \mu_0$. This would add a row at the top of the matrix in (13-40) containing $[(y_{i0} - \mu_0), 0, 0]$.

the asymptotic covariance matrix of $\sqrt{n}\,\bar{\mathbf{m}}$. With a consistent estimator of $\boldsymbol{\theta}$ in hand, this can be estimated empirically using

$$\text{Est.Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{m}}_i\hat{\mathbf{m}}_i' = \frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}\hat{\boldsymbol{\eta}}_i\hat{\boldsymbol{\eta}}_i'\mathbf{H}'\mathbf{V}_i.$$

This is a robust estimator that allows an unrestricted $T \times T$ covariance matrix for the T disturbances, $\varepsilon_{it} + u_i$. But we have assumed that this covariance matrix is the $\boldsymbol{\Sigma}$ defined in (11-31) for the random effects model. To use this information we would, instead, use the residuals in

$$\hat{\boldsymbol{\eta}}_i = \mathbf{y}_i - \mathbf{W}_i\hat{\boldsymbol{\theta}}$$

to estimate σ_u^2 and σ_ε^2 and then $\boldsymbol{\Sigma}$, which produces

$$\text{Est.Asy.Var}[\sqrt{n}\,\bar{\mathbf{m}}] = \frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}'\mathbf{V}_i.$$

We now have the full set of results needed to compute the GMM estimator. The solution to the optimization problem of minimizing q with respect to the parameter vector $\boldsymbol{\theta}$ is

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{GMM} = &\left[\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}\mathbf{V}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{W}_i\right)\right]^{-1}\\ &\times\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\hat{\boldsymbol{\Sigma}}\mathbf{H}\mathbf{V}_i\right)^{-1}\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{y}_i\right).\end{aligned} \qquad \textbf{(13-41)}$$

The estimator of the asymptotic covariance matrix for $\hat{\boldsymbol{\theta}}_{GMM}$ is the inverse matrix in brackets.

The remaining loose end is how to obtain the consistent estimator of $\hat{\boldsymbol{\theta}}$ to compute $\hat{\boldsymbol{\Sigma}}$ Recall that the GMM estimator is consistent with any positive definite weighting matrix, $\mathbf{A}$, in our preceding expression. Therefore, for an initial estimator, we can set $\mathbf{A} = \mathbf{I}$ and use the simple instrumental variables estimator,

$$\hat{\boldsymbol{\theta}}_{\text{IV}} = \left[\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{W}_i\right)\right]^{-1}\left(\sum_{i=1}^{n}\mathbf{W}_i'\mathbf{H}\mathbf{V}_i\right)\left(\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{y}_i\right).$$

It is more common to proceed directly to the 2SLS estimator (see Sections 8.3.4 and 11.8.2), which uses

$$\mathbf{A} = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{V}_i'\mathbf{H}'\mathbf{H}\mathbf{V}_i\right)^{-1}.$$

The estimator is, then, the one given earlier in (13-41) with $\hat{\boldsymbol{\Sigma}}$ replaced by $\mathbf{I}_T$. Either estimator is a function of the sample data only and provides the initial estimator we need.

Ahn and Schmidt (among others) observed that the IV estimator proposed here, as extensive as it is, still neglects quite a lot of information and is therefore (relatively) inefficient. For example, in the first differenced model,

$$E[y_{is}(\varepsilon_{it} - \varepsilon_{i,t-1})] = 0, \quad s = 0, \ldots, t-2, \quad t = 2, \ldots, T.$$

That is, the *level* of y_{is} is uncorrelated with the differences of disturbances that are at least two periods subsequent.[24] (The differencing transformation, as the transformation to deviations from group means, removes the individual effect.) The corresponding moment equations that can enter the construction of a GMM estimator are

$$\frac{1}{n}\sum_{i=1}^{n} y_{is}[(y_{it} - y_{i,t-1}) - \delta(y_{i,t-1} - y_{i,t-2}) - (\mathbf{x}_{it} - \mathbf{x}_{i,t-1})'\boldsymbol{\beta}] = 0$$
$$s = 0, \ldots, t-2, \quad t = 2, \ldots, T.$$

Altogether, Ahn and Schmidt identify $T(T-1)/2 + T - 2$ such equations that involve mixtures of the levels and differences of the variables. The main conclusion that they demonstrate is that in the dynamic model, there is a large amount of information to be gleaned not only from the familiar relationships among the levels of the variables, but also from the implied relationships between the levels and the first differences. The issue of correlation between the transformed y_{it} and the deviations of ε_{it} is discussed in the papers cited.[25]

The number of orthogonality conditions (instrumental variables) used to estimate the parameters of the model is determined by the number of variables in $\mathbf{v}_{it}$ and $\mathbf{a}_i$ in (13-39). In most cases, the model is vastly overidentified—there are far more orthogonality conditions than parameters. As usual in GMM estimation, a test of the overidentifying restrictions can be based on q, the estimation criterion. At its minimum, the limiting distribution of nq is chi squared with degrees of freedom equal to the number of instrumental variables in total minus

$$(1 + K_1 + K_2 + L_1 + L_2).$$[26]

Example 13.10 GMM Estimation of a Dynamic Panel Data Model of Local Government Expenditures

Dahlberg and Johansson (2000) estimated a model for the local government expenditure of several hundred municipalities in Sweden observed over the nine-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,

[24]This is the approach suggested by Holtz-Eakin (1988) and Holtz-Eakin, Newey, and Rosen (1988).

[25]As Ahn and Schmidt show, there are potentially huge numbers of additional orthogonality conditions in this model owing to the relationship between first differences and second moments. We do not consider those. The matrix $\mathbf{V}_i$ could be huge. Consider a model with 10 time-varying, right-hand-side variables and suppose T_i is 15. Then, there are 15 rows and roughly $15 \times (10 \times 15)$ or 2,250 columns. The Ahn and Schmidt estimator, which involves potentially thousands of instruments in a model containing only a handful of parameters may become a bit impractical at this point. The common approach is to use only a small subset of the available instrumental variables. The order of the computation grows as the number of parameters times the square of T.

[26]This is true generally in GMM estimation. It was proposed for the dynamic panel data model by Bhargava and Sargan (1983).

which is not specified as being either *fixed* or *random*. 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. Because 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 nine years of data, this leaves usable 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, which produces a total of 30. For the five observations, using the formulation developed in Section 13.6.5, we have the following matrix of instrumental variables for the orthogonality conditions,

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

where the notation $\boldsymbol{S}_{t1-t0}$ indicates the range of years for that variable. For example, $\boldsymbol{S}_{83-79}$ 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 earlier were employed. Given the preceding construction, the orthogonality conditions are now $E[\mathbf{Z}_i'\mathbf{u}_i] = \mathbf{0}$, where $\mathbf{u}_i = [u_{i,1983}, u_{i,1984}, u_{i,1985}, u_{i,1986}, u_{i,1987}]'$. The empirical moment equation is

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

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

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

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 inverse 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

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

based on recomputing the residuals after GMM estimation. 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. 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* Web site—see Appendix Table F13.1. The authors provide the summary statistics for the raw data that are given in Table 13.1. Kroner, deflated by a municipality-specific price index, then converted to per capita values. Descriptive statistics for the raw data appear in Table 13.3.[27] 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. The estimates of the first specification provided are given in Table 13.4.

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

TABLE 13.3 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 13.5 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% 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.

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% 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 13.5 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 $\mathbf{y}_t = (\Delta S_t, \Delta R_t, \Delta G_t)'$.

TABLE 13.4 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

TABLE 13.5 Estimated Lag Equations for Spending, Revenue, and Grants

	Expenditure Model			*Revenue Model*			*Grant Model*		
	$m = 3$	$m = 2$	$m = 2$	$m = 3$	$m = 2$	$m = 1$	$m = 3$	$m = 2$	$m = 1$
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}	−0.2380	−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	—	—
nq	22.8287	30.4526	34.4986	30.5398	34.2590	53.2506	17.5810	20.5416	27.5927

13.7 SUMMARY AND CONCLUSIONS

The generalized method of moments provides an estimation framework that includes least squares, nonlinear least squares, instrumental variables, 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 specialized the GMM estimator for linear and nonlinear equations and multiple-equation models. We then 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.

Key Terms and Concepts

- Analog estimation
- Central limit theorem
- Criterion function
- Empirical moment equation
- Ergodic theorem
- Exactly identified cases
- Exponential family
- Generalized method of moments (GMM) estimator
- Instrumental variables
- Likelihood ratio statistic
- LM statistic
- Martingale difference series
- Maximum likelihood estimator
- Mean value theorem
- Method of moment generating functions
- Method of moments
- Method of moments estimators
- Minimum distance estimator (MDE)
- Moment equation
- Newey–West estimator
- Nonlinear instrumental variable estimator
- Order condition
- Orthogonality conditions

- Overidentified cases
- Overidentifying restrictions
- Population moment equation
- Probability limit
- Random sample
- Rank condition
- Slutsky theorem
- Specification test
- Sufficient statistic
- Taylor series
- Uncentered moment
- Wald statistic
- Weighted least squares
- Weighting matrix

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, and then use the delta method to show that the covariance matrix for the two estimators is

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

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

2. Using the results in Example 13.5, 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 14.
   c. The mixture distribution in Exercise 3 in Chapter 14.
4. For the Wald distribution discussed in Example 13.3,

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

we have the following results: $E[y] = \mu$, $\text{Var}[y] = \sigma^2 = \mu^3/\lambda$, $E[1/y] = 1/\mu + 1/\lambda$, $\text{Var}[1/y] = 1/(\lambda\mu) + 2/\lambda^2$, $E[y3] = \mu_3 = E[(y-\mu)^3/\sigma^3 = 3\mu^5/\lambda^2$.

   a. Derive the maximum likelihood estimators of μ and λ and an estimator of the asymptotic variances of the MLEs. (*Hint:* Expand the quadratic in the exponent and use the three terms in the derivation.)

b. Derive the method of moments estimators using the three different pairs of moments listed above, $E[y]$, $E[1/y]$ and $E[y3]$.

c. Using a random number generator, I generated a sample of 1,000 draws from the inverse Gaussian population with parameters μ and λ. I computed the following statistics:

	Mean	***Standard Deviation***
y	1.039892	1.438691
$1/y$	2.903571	2.976183
$y3 = (y - \mu)^3/\sigma^3$	4.158523	38.01372

[For the third variable, I used the known (to me) true values of the parameters.] Using the sample data, compute the maximum likelihood estimators of μ and λ and the estimates of the asymptotic standard errors. Compute the method of moments estimators using the means of $1/y$ and $y3$.

5. 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.

6. Consider the probit model analyzed in Chapter 17. The model states that for given vector of independent variables,

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

Consider a GMM estimator based on the result that

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

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?

7. Consider GMM estimation of a regression model as shown at the beginning of Example 13.8. 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 $\mathbf{W}_2$.