14 Generalized Method of Moments

Suppose we have a set of observations on a variable y_t whose probability law depends on an unknown vector of parameters $\boldsymbol{\theta}$. One general approach to estimating $\boldsymbol{\theta}$ is based on the principle of maximum likelihood—we choose as the estimate $\hat{\boldsymbol{\theta}}$ the value for which the data would be most likely to have been observed. A drawback of this approach is that it requires us to specify the form of the likelihood function.

This chapter explores an alternative principle for parameter estimation known as *generalized method of moments* (*GMM*). Although versions of this approach have been used for a long time, the general statement of *GMM* on which this chapter is based was only recently developed by Hansen (1982). The key advantage of *GMM* is that it requires specification only of certain moment conditions rather than the full density. This can also be a drawback, in that *GMM* often does not make efficient use of all the information in the sample.

Section 14.1 introduces the ideas behind *GMM* estimation and derives some of the key results. Section 14.2 shows how various other estimators can be viewed as special cases of *GMM*, including ordinary least squares, instrumental variable estimation, two-stage least squares, estimators for systems of nonlinear simultaneous equations, and estimators for dynamic rational expectations models. Extensions and further discussion are provided in Section 14.3. In many cases, even maximum likelihood estimation can be viewed as a special case of *GMM*. Section 14.4 explores this analogy and uses it to derive some general asymptotic properties of maximum likelihood and quasi-maximum likelihood estimation.

14.1. *Estimation by the Generalized Method of Moments*

Classical Method of Moments

It will be helpful to introduce the ideas behind *GMM* with a concrete example. Consider a random variable Y_t drawn from a standard t distribution with ν degrees of freedom, so that its density is

$$f_{Y_t}(y_t;\,\nu) = \frac{\Gamma[(\nu+1)/2]}{(\pi\nu)^{1/2}\Gamma(\nu/2)}\,[1 + (y_t^2/\nu)]^{-(\nu+1)/2}, \qquad [14.1.1]$$

where $\Gamma(\cdot)$ is the gamma function. Suppose we have an i.i.d. sample of size T (y_1, y_2, . . . , y_T) and want to estimate the degrees of freedom parameter ν. One approach is to estimate ν by maximum likelihood. This approach calculates the sample log likelihood

$$\mathcal{L}(\nu) = \sum_{t=1}^{T} \log f_{Y_t}(y_t; \nu)$$

and chooses as the estimate $\hat{\nu}$ the value for which $\mathcal{L}(\nu)$ is largest.

An alternative principle on which estimation of ν might be based reasons as follows. Provided that $\nu > 2$, a standard t variable has population mean zero and variance given by

$$\mu_2 \equiv E(Y_t^2) = \nu/(\nu - 2). \qquad [14.1.2]$$

As the degrees of freedom parameter (ν) goes to infinity, the variance [14.1.2] approaches unity and the density [14.1.1] approaches that of a standard $N(0, 1)$ variable. Let $\hat{\mu}_{2.T}$ denote the average squared value of y observed in the actual sample:

$$\hat{\mu}_{2.T} \equiv (1/T) \sum_{t=1}^{T} y_t^2. \qquad [14.1.3]$$

For large T, the sample moment ($\hat{\mu}_{2.T}$) should be close to the population moment (μ_2):

$$\hat{\mu}_{2.T} \xrightarrow{p} \mu_2.$$

Recalling [14.1.2], this suggests that a consistent estimate of ν can be obtained by finding a solution to

$$\nu/(\nu - 2) = \hat{\mu}_{2.T} \qquad [14.1.4]$$

or

$$\hat{\nu}_T = \frac{2 \cdot \hat{\mu}_{2.T}}{\hat{\mu}_{2.T} - 1}. \qquad [14.1.5]$$

This estimate exists provided that $\hat{\mu}_{2.T} > 1$, that is, provided that the sample seems to exhibit more variability than the $N(0, 1)$ distribution. If we instead observed $\hat{\mu}_{2.T} \leq 1$, the estimate of the degrees of freedom would be infinity—a $N(0, 1)$ distribution fits the sample second moment better than any member of the t family.

The estimator derived from [14.1.4] is known as a *classical method of moments* estimator. A general description of this approach is as follows. Given an unknown $(a \times 1)$ vector of parameters $\boldsymbol{\theta}$ that characterizes the density of an observed variable y_t, suppose that a distinct population moments of the random variable can be calculated as functions of $\boldsymbol{\theta}$, such as

$$E(Y_t^i) = \mu_i(\boldsymbol{\theta}) \qquad \text{for } i = i_1, i_2, \ldots, i_a. \qquad [14.1.6]$$

The classical method of moments estimate of $\boldsymbol{\theta}$ is the value $\hat{\boldsymbol{\theta}}_T$ for which these population moments are equated to the observed sample moments; that is, $\hat{\boldsymbol{\theta}}_T$ is the value for which

$$\mu_i(\hat{\boldsymbol{\theta}}_T) = (1/T) \sum_{t=1}^{T} y_t^i \qquad \text{for } i = i_1, i_2, \ldots, i_a.$$

An early example of this approach was provided by Pearson (1894).

Generalized Method of Moments

In the example of the t distribution just discussed, a single sample moment ($\hat{\mu}_{2.T}$) was used to estimate a single population parameter (ν). We might also have made use of other moments. For example, if $\nu > 4$, the population fourth moment of a standard t variable is

$$\mu_4 \equiv E(Y_t^4) = \frac{3\nu^2}{(\nu - 2)(\nu - 4)},$$

and we might expect this to be close to the sample fourth moment,

$$\hat{\mu}_{4,T} \equiv (1/T) \sum_{t=1}^{T} y_t^4.$$

We cannot choose the single parameter ν so as to match both the sample second moment and the sample fourth moment. However, we might try to choose ν so as to be as close as possible to both, by minimizing a criterion function such as

$$Q(\nu; y_T, y_{T-1}, \ldots, y_1) \equiv \mathbf{g}'\mathbf{W}\mathbf{g}, \qquad [14.1.7]$$

where

$$\mathbf{g} \equiv \begin{bmatrix} \left\{\hat{\mu}_{2,T} - \dfrac{\nu}{\nu - 2}\right\} \\ \left\{\hat{\mu}_{4,T} - \dfrac{3\nu^2}{(\nu - 2)(\nu - 4)}\right\} \end{bmatrix}. \qquad [14.1.8]$$

Here $\mathbf{W}$ is a (2×2) positive definite symmetric weighting matrix reflecting the importance given to matching each of the moments. The larger is the (1, 1) element of $\mathbf{W}$, the greater is the importance of being as close as possible to satisfying [14.1.4].

An estimate based on minimization of an expression such as [14.1.7] was called a "minimum chi-square" estimator by Cramér (1946, p. 425), Ferguson (1958), and Rothenberg (1973) and a "minimum distance estimator" by Malinvaud (1970). Hansen (1982) provided the most general characterization of this approach and derived the asymptotic properties for serially dependent processes. Most of the results reported in this section were developed by Hansen (1982), who described this as estimation by the "generalized method of moments."

Hansen's formulation of the estimation problem is as follows. Let $\mathbf{w}_t$ be an $(h \times 1)$ vector of variables that are observed at date t, let $\boldsymbol{\theta}$ denote an unknown $(a \times 1)$ vector of coefficients, and let $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$ be an $(r \times 1)$ vector-valued function, $\mathbf{h}$: $(\mathbb{R}^a \times \mathbb{R}^h) \rightarrow \mathbb{R}^r$. Since $\mathbf{w}_t$ is a random variable, so is $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$. Let $\boldsymbol{\theta}_0$ denote the true value of $\boldsymbol{\theta}$, and suppose this true value is characterized by the property that

$$E\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\} = \mathbf{0}. \qquad [14.1.9]$$

The r rows of the vector equation [14.1.9] are sometimes described as *orthogonality conditions*. Let $\mathcal{Y}_T \equiv (\mathbf{w}_T', \mathbf{w}_{T-1}', \ldots, \mathbf{w}_1')'$ be a $(Th \times 1)$ vector containing all the observations in a sample of size T, and let the $(r \times 1)$ vector-valued function $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ denote the sample average of $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$:

$$\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T) \equiv (1/T) \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t). \qquad [14.1.10]$$

Notice that $\mathbf{g}$: $\mathbb{R}^a \rightarrow \mathbb{R}^r$. The idea behind *GMM* is to choose $\boldsymbol{\theta}$ so as to make the sample moment $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ as close as possible to the population moment of zero; that is, the *GMM* estimator $\hat{\boldsymbol{\theta}}_T$ is the value of $\boldsymbol{\theta}$ that minimizes the scalar

$$Q(\boldsymbol{\theta}; \mathcal{Y}_T) = [\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'\mathbf{W}_T[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)], \qquad [14.1.11]$$

where $\{\mathbf{W}_T\}_{T=1}^{\infty}$ is a sequence of $(r \times r)$ positive definite weighting matrices which may be a function of the data $\mathcal{Y}_T$. Often, this minimization is achieved numerically using the methods described in Section 5.7.

The classical method of moments estimator of ν given in [14.1.5] is a special case of this formulation with $\mathbf{w}_t = y_t$, $\boldsymbol{\theta} = \nu$, $\mathbf{W}_T = 1$, and

$$h(\boldsymbol{\theta}, \mathbf{w}_t) = y_t^2 - \nu/(\nu - 2)$$

$$g(\boldsymbol{\theta}; \mathcal{Y}_T) = (1/T) \sum_{t=1}^{T} y_t^2 - \nu/(\nu - 2).$$

Here, $r = a = 1$ and the objective function [14.1.11] becomes

$$Q(\boldsymbol{\theta}; \mathcal{Y}_T) = \left\{(1/T)\sum_{t=1}^{T} y_t^2 - \nu/(\nu - 2)\right\}^2.$$

The smallest value that can be achieved for $Q(\cdot)$ is zero, which obtains when ν is the magnitude given in [14.1.5].

The estimate of ν obtained by minimizing [14.1.7] is also a *GMM* estimator with $r = 2$ and

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \begin{bmatrix} \left\{y_t^2 - \dfrac{\nu}{\nu - 2}\right\} \\ \left\{y_t^4 - \dfrac{3\nu^2}{(\nu - 2)(\nu - 4)}\right\} \end{bmatrix}.$$

Here, $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ and $\mathbf{W}_T$ would be as described in [14.1.7] and [14.1.8].

A variety of other estimators can also be viewed as examples of *GMM*, including ordinary least squares, instrumental variable estimation, two-stage least squares, nonlinear simultaneous equations estimators, estimators for dynamic rational expectations models, and in many cases even maximum likelihood. These applications will be discussed in Sections 14.2 through 14.4.

If the number of parameters to be estimated (a) is the same as the number of orthogonality conditions (r), then typically the objective function [14.1.11] will be minimized by setting

$$\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = \mathbf{0}. \qquad [14.1.12]$$

If $a = r$, then the *GMM* estimator is the value $\hat{\boldsymbol{\theta}}_T$ that satisfies these r equations. If instead there are more orthogonality conditions than parameters to estimate ($r > a$), then [14.1.12] will not hold exactly. How close the ith element of $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ is to zero depends on how much weight the ith orthogonality condition is given by the weighting matrix $\mathbf{W}_T$.

For any value of $\boldsymbol{\theta}$, the magnitude of the $(r \times 1)$ vector $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ is the sample mean of T realizations of the $(r \times 1)$ random vector $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$. If $\mathbf{w}_t$ is strictly stationary and $\mathbf{h}(\cdot)$ is continuous, then it is reasonable to expect the law of large numbers to hold:

$$\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T) \xrightarrow{p} E\{\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\}.$$

The expression $E\{\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\}$ denotes a population magnitude that depends on the value of $\boldsymbol{\theta}$ and on the probability law of $\mathbf{w}_t$. Suppose that this function is continuous in $\boldsymbol{\theta}$ and that $\boldsymbol{\theta}_0$ is the only value of $\boldsymbol{\theta}$ that satisfies [14.1.9]. Then, under fairly general stationarity, continuity, and moment conditions, the value of $\hat{\boldsymbol{\theta}}_T$ that minimizes [14.1.11] offers a consistent estimate of $\boldsymbol{\theta}_0$; see Hansen (1982), Gallant and White (1988), and Andrews and Fair (1988) for details.

Optimal Weighting Matrix

Suppose that when evaluated at the true value $\boldsymbol{\theta}_0$, the process $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}_{t=-\infty}^{\infty}$ is strictly stationary with mean zero and vth autocovariance matrix given by

$$\boldsymbol{\Gamma}_v = E\{[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)][\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_{t-v})]'\}. \qquad [14.1.13]$$

Assuming that these autocovariances are absolutely summable, define

$$\mathbf{S} \equiv \sum_{v=-\infty}^{\infty} \boldsymbol{\Gamma}_v. \qquad [14.1.14]$$

Recall from the discussion in Section 10.5 that S is the asymptotic variance of the sample mean of $\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)$:

$$\mathbf{S} = \lim_{T\to\infty} T\cdot E\{[\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)][\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)]'\}.$$

The optimal value for the weighting matrix $\mathbf{W}_T$ in [14.1.11] turns out to be given by $\mathbf{S}^{-1}$, the inverse of the asymptotic variance matrix. That is, the minimum asymptotic variance for the *GMM* estimator $\hat{\boldsymbol{\theta}}_T$ is obtained when $\hat{\boldsymbol{\theta}}_T$ is chosen to minimize

$$Q(\boldsymbol{\theta}; \mathcal{Y}_T) = [\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'\mathbf{S}^{-1}[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]. \qquad [14.1.15]$$

To see the intuition behind this claim, consider a simple linear model in which we have r different observations $(y_1, y_2, \ldots, y_r)$ with a different population mean for each observation $(\mu_1, \mu_2, \ldots, \mu_r)$. For example, y_1 might denote the sample mean in a sample of T_1 observations on some variable, y_2 the sample mean from a second sample, and so on. In the absence of restrictions, the estimates would simply be $\hat{\mu}_i = y_i$ for $i = 1, 2, \ldots, r$. In the presence of linear restrictions across the μ's, the best estimates that are linear functions of the y's would be obtained by generalized least squares. Recall that the *GLS* estimate of $\boldsymbol{\mu}$ is the value that minimizes

$$(\mathbf{y} - \boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu}), \qquad [14.1.16]$$

where $\mathbf{y} = (y_1, y_2, \ldots, y_r)'$, $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_r)'$, and $\boldsymbol{\Omega}$ is the variance-covariance matrix of $\mathbf{y} - \boldsymbol{\mu}$:

$$\boldsymbol{\Omega} = E[(\mathbf{y} - \boldsymbol{\mu})(\mathbf{y} - \boldsymbol{\mu})'].$$

The optimal weighting matrix to use with the quadratic form in [14.1.16] is given by $\boldsymbol{\Omega}^{-1}$. Just as $\boldsymbol{\Omega}$ in [14.1.16] is the variance of $(\mathbf{y} - \boldsymbol{\mu})$, so $\mathbf{S}$ in [14.1.15] is the asymptotic variance of $\sqrt{T}\cdot\mathbf{g}(\cdot)$.

If the vector process $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}_{t=-\infty}^{\infty}$ were serially uncorrelated, then the matrix S could be consistently estimated by

$$\mathbf{S}_T^* = (1/T)\sum_{t=1}^{T} [\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)][\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)]'. \qquad [14.1.17]$$

Calculating this magnitude requires knowledge of $\boldsymbol{\theta}_0$, though it often also turns out that

$$\hat{\mathbf{S}}_T \equiv (1/T)\sum_{t=1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}_T, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}_T, \mathbf{w}_t)]' \xrightarrow{p} \mathbf{S} \qquad [14.1.18]$$

for $\hat{\boldsymbol{\theta}}_T$ any consistent estimate of $\boldsymbol{\theta}_0$, assuming that $\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)$ is serially uncorrelated.

Note that this description of the optimal weighting matrix is somewhat circular—before we can estimate $\boldsymbol{\theta}$, we need an estimate of the matrix S, and before we can estimate the matrix S, we need an estimate of $\boldsymbol{\theta}$. The practical procedure used in *GMM* is as follows. An initial estimate $\hat{\boldsymbol{\theta}}_T^{(0)}$ is obtained by minimizing [14.1.11] with an arbitrary weighting matrix such as $\mathbf{W}_T = \mathbf{I}_r$. This estimate of $\boldsymbol{\theta}$ is then used in [14.1.18] to produce an initial estimate $\hat{\mathbf{S}}_T^{(0)}$. Expression [14.1.11] is then minimized with $\mathbf{W}_T = [\hat{\mathbf{S}}_T^{(0)}]^{-1}$ to arrive at a new *GMM* estimate $\hat{\boldsymbol{\theta}}_T^{(1)}$. This process can be iterated until $\hat{\boldsymbol{\theta}}_T^{(j)} \cong \hat{\boldsymbol{\theta}}_T^{(j+1)}$, though the estimate based on a single iteration $\hat{\boldsymbol{\theta}}_T^{(1)}$ has the same asymptotic distribution as that based on an arbitrarily large number of iterations. Iterating nevertheless offers the practical advantage that the resulting estimates are invariant with respect to the scale of the data and to the initial weighting matrix for $\mathbf{W}_T$.

On the other hand, if the vector process $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}_{t=-\infty}^{\infty}$ is serially correlated, the Newey-West (1987) estimate of $\mathbf{S}$ could be used:

$$\hat{\mathbf{S}}_T = \hat{\mathbf{\Gamma}}_{0,T} + \sum_{v=1}^{q} \{1 - [v/(q + 1)]\}(\hat{\mathbf{\Gamma}}_{v,T} + \hat{\mathbf{\Gamma}}'_{v,T}), \qquad [14.1.19]$$

where

$$\hat{\mathbf{\Gamma}}_{v,T} = (1/T) \sum_{t=v+1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_{t-v})]', \qquad [14.1.20]$$

with $\hat{\boldsymbol{\theta}}$ again an initial consistent estimate of $\boldsymbol{\theta}_0$. Alternatively, the estimators proposed by Gallant (1987), Andrews (1991), or Andrews and Monahan (1992) that were discussed in Section 10.5 could also be applied in this context.

Asymptotic Distribution of the GMM *Estimates*

Let $\hat{\boldsymbol{\theta}}_T$ be the value that minimizes

$$[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'\hat{\mathbf{S}}_T^{-1}[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)], \qquad [14.1.21]$$

with $\hat{\mathbf{S}}_T$ regarded as fixed with respect to $\boldsymbol{\theta}$ and $\hat{\mathbf{S}}_T \xrightarrow{p} \mathbf{S}$. Assuming an interior optimum, this minimization is achieved by setting the derivative of [14.1.21] with respect to $\boldsymbol{\theta}$ to zero. Thus, the *GMM* estimate $\hat{\boldsymbol{\theta}}_T$ is typically a solution to the following system of nonlinear equations:

$$\underbrace{\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}'}_{(a\times r)} \times \underbrace{\hat{\mathbf{S}}_T^{-1}}_{(r\times r)} \times \underbrace{[\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)]}_{(r\times 1)} = \underbrace{\mathbf{0}}_{(a\times 1)}. \qquad [14.1.22]$$

Here $[\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)/\partial \boldsymbol{\theta}']|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}$ denotes the $(r \times a)$ matrix of derivatives of the function $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$, where these derivatives are evaluated at the *GMM* estimate $\hat{\boldsymbol{\theta}}_T$.

Since $\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)$ is the sample mean of a process whose population mean is zero, $\mathbf{g}(\cdot)$ should satisfy the central limit theorem given conditions such as strict stationarity of $\mathbf{w}_t$, continuity of $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$, and restrictions on higher moments. Thus, in many instances it should be the case that

$$\sqrt{T}\cdot \mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T) \xrightarrow{L} N(\mathbf{0}, \mathbf{S}).$$

Not much more than this is needed to conclude that the *GMM* estimator $\hat{\boldsymbol{\theta}}_T$ is asymptotically Gaussian and to calculate its asymptotic variance. The following proposition, adapted from Hansen (1982), is proved in Appendix 14.A at the end of this chapter.

Proposition 14.1: *Let* $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ *be differentiable in* $\boldsymbol{\theta}$ *for all* $\mathcal{Y}_T$, *and let* $\hat{\boldsymbol{\theta}}_T$ *be the GMM estimator satisfying* [14.1.22] *with* $r \geq a$. *Let* $\{\hat{\mathbf{S}}_T\}_{T=1}^{\infty}$ *be a sequence of positive definite* $(r \times r)$ *matrices such that* $\hat{\mathbf{S}}_T \xrightarrow{p} \mathbf{S}$, *with* $\mathbf{S}$ *positive definite. Suppose, further, that the following hold:*

(a) $\hat{\boldsymbol{\theta}}_T \xrightarrow{p} \boldsymbol{\theta}_0$;

(b) $\sqrt{T}\cdot \mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T) \xrightarrow{L} N(\mathbf{0}, \mathbf{S})$; *and*

(c) for any sequence $\{\boldsymbol{\theta}_T^*\}_{T=1}^{\infty}$ *satisfying* $\boldsymbol{\theta}_T^* \xrightarrow{p} \boldsymbol{\theta}_0$, *it is the case that*

$$\operatorname{plim}\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_T^*}\right\} = \operatorname{plim}\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}\right\} \equiv \underset{(r\times a)}{\mathbf{D}'}, \qquad [14.1.23]$$

with the columns of $\mathbf{D}'$ *linearly independent.*
Then

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}), \qquad [14.1.24]$$

where

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

Proposition 14.1 implies that we can treat $\hat{\boldsymbol{\theta}}_T$ approximately as

$$\hat{\boldsymbol{\theta}}_T \approx N(\boldsymbol{\theta}_0, \hat{\mathbf{V}}_T/T), \qquad [14.1.25]$$

where

$$\hat{\mathbf{V}}_T = \{\hat{\mathbf{D}}_T\hat{\mathbf{S}}_T^{-1}\hat{\mathbf{D}}_T'\}^{-1}.$$

The estimate $\hat{\mathbf{S}}_T$ can be constructed as in [14.1.18] or [14.1.19], while

$$\underset{(r\times a)}{\hat{\mathbf{D}}_T'} = \left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}.$$

Testing the Overidentifying Restrictions

When the number of orthogonality conditions exceeds the number of parameters to be estimated ($r > a$), the model is overidentified, in that more orthogonality conditions were used than are needed to estimate $\boldsymbol{\theta}$. In this case, Hansen (1982) suggested a test of whether all of the sample moments represented by $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ are as close to zero as would be expected if the corresponding population moments $E\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}$ were truly zero.

From Proposition 8.1 and condition (b) in Proposition 14.1, notice that if the population orthogonality conditions in [14.1.9] were all true, then

$$[\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)]'\mathbf{S}^{-1}[\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)] \xrightarrow{L} \chi^2(r). \qquad [14.1.26]$$

In [14.1.26], the sample moment function $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$ is evaluated at the true value of $\boldsymbol{\theta}_0$. One's first guess might be that condition [14.1.26] also holds when [14.1.26] is evaluated at the *GMM* estimate $\hat{\boldsymbol{\theta}}_T$. However, this is not the case. The reason is that [14.1.22] implies that a different linear combinations of the ($r \times 1$) vector $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ are identically zero, these being the a linear combinations obtained when $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ is premultiplied by the ($a \times r$) matrix

$$\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1}.$$

For example, when $a = r$, *all* linear combinations of $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ are identically zero, and if $\boldsymbol{\theta}_0$ were replaced by $\hat{\boldsymbol{\theta}}_T$, the magnitude in [14.1.26] would simply equal zero in all samples.

Since the vector $\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)$ contains ($r - a$) nondegenerate random variables, it turns out that a correct test of the overidentifying restrictions for the case when $r > a$ can be based on the fact that

$$[\sqrt{T}\cdot\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)]'\hat{\mathbf{S}}_T^{-1}[\sqrt{T}\cdot\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)] \xrightarrow{L} \chi^2(r - a). \qquad [14.1.27]$$

Moreover, this test statistic is trivial to calculate, for it is simply the sample size T times the value attained for the objective function [14.1.21] at the *GMM* estimate $\hat{\boldsymbol{\theta}}_T$.

Unfortunately, Hansen's χ^2 test based on [14.1.27] can easily fail to detect a misspecified model (Newey, 1985). It is therefore often advisable to supplement this test with others described in Section 14.3.

14.2. *Examples*

This section shows how properties of a variety of different estimators can be obtained as special cases of Hansen's results for generalized method of moments estimation. To facilitate this discussion, we first summarize the results of the preceding section.

Summary of GMM

The statistical model is assumed to imply a set of r orthogonality conditions of the form

$$\underset{(r\times 1)}{E\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}} = \underset{(r\times 1)}{\mathbf{0}}, \qquad [14.2.1]$$

where $\mathbf{w}_t$ is a strictly stationary vector of variables observed at date t, $\boldsymbol{\theta}_0$ is the true value of an unknown $(a \times 1)$ vector of parameters, and $\mathbf{h}(\cdot)$ is a differentiable r-dimensional vector-valued function with $r \geq a$. The *GMM* estimate $\hat{\boldsymbol{\theta}}_T$ is the value of $\boldsymbol{\theta}$ that minimizes

$$\underset{(1\times r)}{[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'}\underset{(r\times r)}{\hat{\mathbf{S}}_T^{-1}}\underset{(r\times 1)}{[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]}, \qquad [14.2.2]$$

where

$$\underset{(r\times 1)}{\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)} \equiv (1/T)\sum_{t=1}^{T} \underset{(r\times 1)}{\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)} \qquad [14.2.3]$$

and $\hat{\mathbf{S}}_T$ is an estimate of[1]

$$\underset{(r\times r)}{\mathbf{S}} = \lim_{T\to\infty} (1/T)\sum_{t=1}^{T}\sum_{v=-\infty}^{\infty} E\{\underset{(r\times 1)}{[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)]}\underset{(1\times r)}{[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_{t-v})]'}\}. \qquad [14.2.4]$$

The *GMM* estimate can be treated as if

$$\underset{(a\times 1)}{\hat{\boldsymbol{\theta}}_T} \approx N(\underset{(a\times 1)}{\boldsymbol{\theta}_0}, \underset{(a\times a)}{\hat{\mathbf{V}}_T/T}), \qquad [14.2.5]$$

where

$$\underset{(a\times a)}{\hat{\mathbf{V}}_T} = \{\underset{(a\times r)}{\hat{\mathbf{D}}_T}\cdot\underset{(r\times r)}{\hat{\mathbf{S}}_T^{-1}}\cdot\underset{(r\times a)}{\hat{\mathbf{D}}_T')}\}^{-1} \qquad [14.2.6]$$

and

$$\underset{(r\times a)}{\hat{\mathbf{D}}_T'} = \left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}. \qquad [14.2.7]$$

We now explore how these results would be applied in various special cases.

Ordinary Least Squares

Consider the standard linear regression model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t, \qquad [14.2.8]$$

for $\mathbf{x}_t$ a $(k \times 1)$ vector of explanatory variables. The critical assumption needed to justify *OLS* regression is that the regression residual u_t is uncorrelated with the explanatory variables:

$$E(\mathbf{x}_t u_t) = \mathbf{0}. \qquad [14.2.9]$$

[1] Under strict stationarity, the magnitude

$$E\{[\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)][\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_{t-v})]'\} = \boldsymbol{\Gamma}_v$$

does not depend on t. The expression in the text is more general than necessary under the stated assumptions. This expression is appropriate for a characterization of *GMM* that does not assume strict stationarity. The expression in the text is also helpful in suggesting estimates of S that can be used in various special cases described later in this section.

In other words, the true value $\boldsymbol{\beta}_0$ is assumed to satisfy the condition

$$E[\mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}_0)] = \mathbf{0}. \qquad [14.2.10]$$

Expression [14.2.10] describes k orthogonality conditions of the form of [14.2.1], in which $\mathbf{w}_t = (y_t, \mathbf{x}_t')'$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, and

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta}). \qquad [14.2.11]$$

The number of orthogonality conditions is the same as the number of unknown parameters in $\boldsymbol{\beta}$, so that $r = a = k$. Hence, the standard regression model could be viewed as a just-identified *GMM* specification. Since it is just identified, the *GMM* estimate of $\boldsymbol{\beta}$ is the value that sets the sample average value for [14.2.11] equal to zero:

$$\mathbf{0} = \mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = (1/T)\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_T). \qquad [14.2.12]$$

Rearranging [14.2.12] results in

$$\sum_{t=1}^{T} \mathbf{x}_t y_t = \left\{\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}\hat{\boldsymbol{\beta}}_T$$

or

$$\hat{\boldsymbol{\beta}}_T = \left\{\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1}\left\{\sum_{t=1}^{T} \mathbf{x}_t y_t\right\}, \qquad [14.2.13]$$

which is the usual *OLS* estimator. Hence, *OLS* is a special case of *GMM*.

Note that in deriving the *GMM* estimator in [14.2.13] we assumed that the residual u_t was uncorrelated with the explanatory variables, but we did not make any other assumptions about heteroskedasticity or serial correlation of the residuals. In the presence of heteroskedasticity or serial correlation, *OLS* is not as efficient as *GLS*. Because *GMM* uses the *OLS* estimate even in the presence of heteroskedasticity or serial correlation, *GMM* in general is not efficient. However, recall from Section 8.2 that one can still use *OLS* in the presence of heteroskedasticity or serial correlation. As long as condition [14.2.9] is satisfied, *OLS* yields a consistent estimate of $\boldsymbol{\beta}$, though the formulas for standard errors have to be adjusted to take account of the heteroskedasticity or autocorrelation.

The *GMM* expression for the variance of $\hat{\boldsymbol{\beta}}_T$ is given by [14.2.6]. Differentiating [14.2.11], we see that

$$\begin{aligned}\hat{\mathbf{D}}_T' &= \left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T} \\ &= (1/T)\sum_{t=1}^{T} \left.\frac{\partial \mathbf{x}_t(y_t - \mathbf{x}_t'\boldsymbol{\beta})}{\partial \boldsymbol{\beta}'}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}_T} \\ &= -(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'.\end{aligned} \qquad [14.2.14]$$

Substituting [14.2.11] into [14.2.4] results in

$$\mathbf{S} = \lim_{T\to\infty} (1/T)\sum_{t=1}^{T}\sum_{v=-\infty}^{\infty} E\{u_t u_{t-v}\mathbf{x}_t\mathbf{x}_{t-v}'\}. \qquad [14.2.15]$$

Suppose that u_t is regarded as conditionally homoskedastic and serially uncorrelated:

$$E\{u_t u_{t-v}\mathbf{x}_t\mathbf{x}_{t-v}'\} = \begin{cases}\sigma^2 E(\mathbf{x}_t\mathbf{x}_t') & \text{for } v = 0\\ \mathbf{0} & \text{for } v \neq 0.\end{cases}$$

In this case the matrix in [14.2.15] should be consistently estimated by

$$\hat{\mathbf{S}}_T = \hat{\sigma}_T^2 (1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t', \tag{14.2.16}$$

where

$$\hat{\sigma}_T^2 = (1/T) \sum_{t=1}^{T} \hat{u}_t^2$$

for $\hat{u}_t \equiv y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_T$ the *OLS* residual. Substituting [14.2.14] and [14.2.16] into [14.2.6] produces a variance-covariance matrix for the *OLS* estimate $\hat{\boldsymbol{\beta}}_T$ of

$$\begin{aligned}(1/T)\hat{\mathbf{V}}_T &= (1/T)\left\{(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\left[\hat{\sigma}_T^2(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1} (1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1} \\ &= \hat{\sigma}_T^2\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1}.\end{aligned}$$

Apart from the estimate of σ^2, this is the usual expression for the variance of the *OLS* estimator under these conditions.

On the other hand, suppose that u_t is conditionally heteroskedastic and serially correlated. In this case, the estimate of $\mathbf{S}$ proposed in [14.1.19] would be

$$\hat{\mathbf{S}}_T = \hat{\boldsymbol{\Gamma}}_{0,T} + \sum_{v=1}^{q} \{1 - [v/(q + 1)]\}(\hat{\boldsymbol{\Gamma}}_{v,T} + \hat{\boldsymbol{\Gamma}}_{v,T}'),$$

where

$$\hat{\boldsymbol{\Gamma}}_{v,T} = (1/T) \sum_{t=v+1}^{T} \hat{u}_t\hat{u}_{t-v}\mathbf{x}_t\mathbf{x}_{t-v}'.$$

Under these assumptions, the *GMM* approximation for the variance-covariance matrix of $\hat{\boldsymbol{\beta}}_T$ would be

$$\begin{aligned}E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})'] &\cong (1/T)\left\{(1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' \, \hat{\mathbf{S}}_T^{-1} \, (1/T) \sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1} \\ &= T\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1} \hat{\mathbf{S}}_T\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right]^{-1},\end{aligned}$$

which is the expression derived earlier in equation [10.5.21]. White's (1980) heteroskedasticity-consistent standard errors in [8.2.35] are obtained as a special case when $q = 0$.

Instrumental Variable Estimation

Consider again a linear model

$$y_t = \mathbf{z}_t'\boldsymbol{\beta} + u_t, \tag{14.2.17}$$

where $\mathbf{z}_t$ is a $(k \times 1)$ vector of explanatory variables. Suppose now that some of the explanatory variables are endogenous, so that $E(\mathbf{z}_t u_t) \neq \mathbf{0}$. Let $\mathbf{x}_t$ be an $(r \times 1)$ vector of predetermined explanatory variables that are correlated with $\mathbf{z}_t$ but uncorrelated with u_t:

$$E(\mathbf{x}_t u_t) = \mathbf{0}.$$

The r orthogonality conditions are now

$$E[\mathbf{x}_t(y_t - \mathbf{z}_t'\boldsymbol{\beta}_0)] = \mathbf{0}. \tag{14.2.18}$$

This again will be recognized as a special case of the *GMM* framework in which $\mathbf{w}_t = (y_t, \mathbf{z}_t', \mathbf{x}_t')'$, $\boldsymbol{\theta} = \boldsymbol{\beta}$, $a = k$, and

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \mathbf{x}_t(y_t - \mathbf{z}_t'\boldsymbol{\beta}). \tag{14.2.19}$$

Suppose that the number of parameters to be estimated equals the number of orthogonality conditions ($a = k = r$). Then the model is just identified, and the *GMM* estimator satisfies

$$\mathbf{0} = \mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = (1/T) \sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{z}_t' \hat{\boldsymbol{\beta}}_T) \qquad [14.2.20]$$

or

$$\hat{\boldsymbol{\beta}}_T = \left\{ \sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t' \right\}^{-1} \left\{ \sum_{t=1}^{T} \mathbf{x}_t y_t \right\},$$

which is the usual instrumental variable estimator for this model. To calculate the standard errors implied by Hansen's (1982) general results, we differentiate [14.2.19] to find

$$\begin{aligned} \hat{\mathbf{D}}_T' &= \left. \frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T} \\ &= (1/T) \sum_{t=1}^{T} \left. \frac{\partial \mathbf{x}_t(y_t - \mathbf{z}_t' \boldsymbol{\beta})}{\partial \boldsymbol{\beta}'} \right|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_T} \qquad [14.2.21] \\ &= -(1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t'. \end{aligned}$$

The requirement in Proposition 14.1 that the plim of this matrix have linearly independent columns is the same condition that was needed to establish consistency of the *IV* estimator in Chapter 9, namely, the condition that the rows of $E(\mathbf{z}_t \mathbf{x}_t')$ be linearly independent. The *GMM* variance for $\hat{\boldsymbol{\beta}}_T$ is seen from [14.2.6] to be

$$(1/T)\hat{\mathbf{V}}_T = (1/T)\left\{ \left[(1/T) \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t' \right] \hat{\mathbf{S}}_T^{-1} \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t' \right] \right\}^{-1}, \qquad [14.2.22]$$

where $\hat{\mathbf{S}}_T$ is an estimate of

$$\mathbf{S} = \lim_{T \to \infty} (1/T) \sum_{t=1}^{T} \sum_{v=-\infty}^{\infty} E\{u_t u_{t-v} \mathbf{x}_t \mathbf{x}_{t-v}'\}. \qquad [14.2.23]$$

If the regression residuals $\{u_t\}$ are serially uncorrelated and homoskedastic with variance σ^2, the natural estimate of $\mathbf{S}$ is

$$\hat{\mathbf{S}}_T = \hat{\sigma}_T^2 \cdot (1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \qquad [14.2.24]$$

for $\hat{\sigma}_T^2 = (1/T) \Sigma_{t=1}^{T} (y_t - \mathbf{z}_t' \hat{\boldsymbol{\beta}}_T)^2$. Substituting this estimate into [14.2.22] yields

$$\begin{aligned} E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})'] &\cong \hat{\sigma}_T^2 \left\{ \left[\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t' \right] \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t' \right] \right\}^{-1} \\ &= \hat{\sigma}_T^2 \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t' \right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' \right] \left[\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t' \right]^{-1}, \end{aligned}$$

the same result derived earlier in [9.2.30]. On the other hand, a heteroskedasticity- and autocorrelation-consistent variance-covariance matrix for *IV* estimation is given by

$$E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})'] \cong T \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}_t' \right]^{-1} \hat{\mathbf{S}}_T \left[\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}_t' \right]^{-1}, \qquad [14.2.25]$$

where

$$\hat{\mathbf{S}}_T = \hat{\boldsymbol{\Gamma}}_{0,T} + \sum_{v=1}^{q} \{1 - [v/(q+1)]\}(\hat{\boldsymbol{\Gamma}}_{v,T} + \hat{\boldsymbol{\Gamma}}'_{v,T}), \qquad [14.2.26]$$

$$\hat{\boldsymbol{\Gamma}}_{v,T} = (1/T) \sum_{t=v+1}^{T} \hat{u}_t \hat{u}_{t-v} \mathbf{x}_t \mathbf{x}'_{t-v}$$

$$\hat{u}_t = y_t - \mathbf{z}'_t \hat{\boldsymbol{\beta}}_T.$$

Two-Stage Least Squares

Consider again the linear model of [14.2.17] and [14.2.18], but suppose now that the number of valid instruments r exceeds the number of explanatory variables k. For this overidentified model, *GMM* will no longer set all the sample orthogonality conditions to zero as in [14.2.20], but instead will be the solution to [14.1.22],

$$\begin{aligned} \mathbf{0} &= \left\{ \left. \frac{\partial \mathbf{g}(\boldsymbol{\theta}; \boldsymbol{\mathcal{Y}}_T)}{\partial \boldsymbol{\theta}'} \right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T} \right\}' \times \hat{\mathbf{S}}_T^{-1} \times [\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \boldsymbol{\mathcal{Y}}_T)] \\ &= \left\{ -(1/T) \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right\} \hat{\mathbf{S}}_T^{-1} \left\{ (1/T) \sum_{t=1}^{T} \mathbf{x}_t (y_t - \mathbf{z}'_t \hat{\boldsymbol{\beta}}_T) \right\}, \end{aligned} \qquad [14.2.27]$$

with the last line following from [14.2.21] and [14.2.20]. Again, if u_t is serially uncorrelated and homoskedastic with variance σ^2, a natural estimate of $\mathbf{S}$ is given by [14.2.24]. Using this estimate, [14.2.27] becomes

$$(1/\hat{\sigma}_T^2) \times \left\{ \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right\} \left\{ \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t \right\}^{-1} \left\{ \sum_{t=1}^{T} \mathbf{x}_t (y_t - \mathbf{z}'_t \hat{\boldsymbol{\beta}}_T) \right\} = \mathbf{0}. \qquad [14.2.28]$$

As in expression [9.2.5], define

$$\hat{\boldsymbol{\delta}}' \equiv \left\{ \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right\} \left\{ \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t \right\}^{-1}.$$

Thus, $\hat{\boldsymbol{\delta}}'$ is a $(k \times r)$ matrix whose ith row represents the coefficients from an *OLS* regression of z_{it} on $\mathbf{x}_t$. Let

$$\hat{\mathbf{z}}_t \equiv \hat{\boldsymbol{\delta}}' \mathbf{x}_t$$

be the $(k \times 1)$ vector of fitted values from these regressions of $\mathbf{z}_t$ on $\mathbf{x}_t$. Then [14.2.28] implies that

$$\sum_{t=1}^{T} \hat{\mathbf{z}}_t (y_t - \mathbf{z}'_t \hat{\boldsymbol{\beta}}_T) = \mathbf{0}$$

or

$$\hat{\boldsymbol{\beta}}_T = \left\{ \sum_{t=1}^{T} \hat{\mathbf{z}}_t \mathbf{z}'_t \right\}^{-1} \left\{ \sum_{t=1}^{T} \hat{\mathbf{z}}_t y_t \right\}.$$

Thus, the *GMM* estimator for this case is simply the two-stage least squares estimator as written in [9.2.8]. The variance given in [14.2.6] would be

$$\begin{aligned} (1/T)\hat{\mathbf{V}}_T &= (1/T) \left\{ \left[(1/T) \sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right] \hat{\mathbf{S}}_T^{-1} \left[(1/T) \sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}'_t \right] \right\}^{-1} \\ &= \hat{\sigma}_T^2 \left\{ \left[\sum_{t=1}^{T} \mathbf{z}_t \mathbf{x}'_t \right] \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}'_t \right]^{-1} \left[\sum_{t=1}^{T} \mathbf{x}_t \mathbf{z}'_t \right] \right\}^{-1}, \end{aligned}$$

as earlier derived in expression [9.2.25]. A test of the overidentifying assumptions embodied in the model in [14.2.17] and [14.2.18] is given by

$$
\begin{aligned}
&T[\mathbf{g}(\hat{\boldsymbol{\theta}}_T;\, \mathcal{Y}_T)]'\hat{\mathbf{S}}_T^{-1}[\mathbf{g}(\hat{\boldsymbol{\theta}}_T;\, \mathcal{Y}_T)] \\
&\quad = T\left\{(1/T)\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_T)\right\}'\left\{\hat{\sigma}_T^2\cdot(1/T)\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1} \\
&\qquad \times \left\{(1/T)\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_T)\right\} \\
&\quad = \hat{\sigma}_T^{-2}\left\{\sum_{t=1}^{T} \hat{u}_t\mathbf{x}_t'\right\}\left\{\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t'\right\}^{-1}\left\{\sum_{t=1}^{T} \mathbf{x}_t\hat{u}_t\right\}.
\end{aligned}
$$

This magnitude will have an asymptotic χ^2 distribution with $(r - k)$ degrees of freedom if the model is correctly specified.

Alternatively, to allow for heteroskedasticity and autocorrelation for the residuals u_t, the estimate $\hat{\mathbf{S}}_T$ in [14.2.24] would be replaced by [14.2.26]. Recall the first-order condition [14.2.27]:

$$\left\{(1/T)\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right\}\hat{\mathbf{S}}_T^{-1}\left\{(1/T)\sum_{t=1}^{T} \mathbf{x}_t(y_t - \mathbf{z}_t'\hat{\boldsymbol{\beta}}_T)\right\} = \mathbf{0}. \qquad [14.2.29]$$

If we now define

$$\hat{\mathbf{z}}_t \equiv \hat{\boldsymbol{\delta}}'\mathbf{x}_t$$

$$\hat{\boldsymbol{\delta}}' \equiv \left\{(1/T)\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right\}\hat{\mathbf{S}}_T^{-1},$$

then [14.2.29] implies that the *GMM* estimator for this case is given by

$$\hat{\boldsymbol{\beta}}_T = \left\{\sum_{t=1}^{T} \hat{\mathbf{z}}_t\mathbf{z}_t'\right\}^{-1}\left\{\sum_{t=1}^{T} \hat{\mathbf{z}}_t y_t\right\}.$$

This characterization of $\hat{\boldsymbol{\beta}}_T$ is circular—in order to calculate $\hat{\boldsymbol{\beta}}_T$, we need to know $\hat{\mathbf{z}}_t$ and thus $\hat{\mathbf{S}}_T$, whereas to construct $\hat{\mathbf{S}}_T$ from [14.2.26] we first need to know $\hat{\boldsymbol{\beta}}_T$. The solution is first to estimate $\boldsymbol{\beta}$ using a suboptimal weighting matrix such as $\hat{\mathbf{S}}_T = (1/T)\Sigma_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$, and then to use this estimate of $\mathbf{S}$ to reestimate $\boldsymbol{\beta}$. The asymptotic variance of the *GMM* estimator is given by

$$E[(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}}_T - \boldsymbol{\beta})'] \cong T\left\{\left[\sum_{t=1}^{T} \mathbf{z}_t\mathbf{x}_t'\right]\hat{\mathbf{S}}_T^{-1}\left[\sum_{t=1}^{T} \mathbf{x}_t\mathbf{z}_t'\right]\right\}^{-1}.$$

Nonlinear Systems of Simultaneous Equations

Hansen's (1982) *GMM* also provides a convenient framework for estimating the nonlinear systems of simultaneous equations analyzed by Amemiya (1974), Jorgenson and Laffont (1974), and Gallant (1977). Suppose that the goal is to estimate a system of n nonlinear equations of the form

$$\mathbf{y}_t = \mathbf{f}(\boldsymbol{\theta}, \mathbf{z}_t) + \mathbf{u}_t$$

for $\mathbf{z}_t$ a $(k \times 1)$ vector of explanatory variables and $\boldsymbol{\theta}$ an $(a \times 1)$ vector of unknown parameters. Let $\mathbf{x}_{it}$ denote a vector of instruments that are uncorrelated with the ith element of $\mathbf{u}_t$. The r orthogonality conditions for this model are

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \begin{bmatrix} [y_{1t} - f_1(\boldsymbol{\theta}, \mathbf{z}_t)]\mathbf{x}_{1t} \\ [y_{2t} - f_2(\boldsymbol{\theta}, \mathbf{z}_t)]\mathbf{x}_{2t} \\ \vdots \\ [y_{nt} - f_n(\boldsymbol{\theta}, \mathbf{z}_t)]\mathbf{x}_{nt} \end{bmatrix},$$

where $f_i(\boldsymbol{\theta}, \mathbf{z}_t)$ denotes the ith element of $\mathbf{f}(\boldsymbol{\theta}, \mathbf{z}_t)$ and $\mathbf{w}_t \equiv (\mathbf{y}_t', \mathbf{z}_t', \mathbf{x}_t')'$. The *GMM* estimate of $\boldsymbol{\theta}$ is the value that minimizes

$$Q(\boldsymbol{\theta}; \mathcal{Y}_T) = \left[(1/T)\sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\right]' \hat{\mathbf{S}}_T^{-1}\left[(1/T)\sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\right], \qquad [14.2.30]$$

where an estimate of $\mathbf{S}$ that could be used with heteroskedasticity and serial correlation of u_t is given by

$$\hat{\mathbf{S}}_T = \hat{\boldsymbol{\Gamma}}_{0,T} + \sum_{v=1}^{q} \{1-[v/(q+1)]\}(\hat{\boldsymbol{\Gamma}}_{v,T} + \hat{\boldsymbol{\Gamma}}_{v,T}')$$

$$\hat{\boldsymbol{\Gamma}}_{v,T} = (1/T)\sum_{t=v+1}^{T} [\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_{t-v})]'.$$

Minimization of [14.2.30] can be achieved numerically. Again, in order to evaluate [14.2.30], we first need an initial estimate of $\mathbf{S}$. One approach is to first minimize [14.2.30] with $\mathbf{S}_T = \mathbf{I}_r$, use the resulting estimate $\hat{\boldsymbol{\theta}}$ to construct a better estimate of $\mathbf{S}_T$, and recalculate $\hat{\boldsymbol{\theta}}$; the procedure can be iterated further, if desired. Identification requires an order condition ($r \geq a$) and the rank condition that the columns of the plim of $\hat{\mathbf{D}}_T'$ be linearly independent, where

$$\hat{\mathbf{D}}_T' = (1/T)\sum_{t=1}^{T} \left.\frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}.$$

Standard errors for $\hat{\boldsymbol{\theta}}_T$ are then readily calculated from [14.2.5] and [14.2.6].

Estimation of Dynamic Rational Expectation Models

People's behavior is often influenced by their expectations about the future. Unfortunately, we typically do not have direct observations on these expectations. However, it is still possible to estimate and test behavioral models if people's expectations are formed rationally in the sense that the errors they make in forecasting are uncorrelated with information they had available at the time of the forecast. As long as the econometrician observes a subset of the information people have actually used, the rational expectations hypothesis suggests orthogonality conditions that can be used in the *GMM* framework.

For illustration, we consider the study of portfolio decisions by Hansen and Singleton (1982). Let c_t denote the overall level of spending on consumption goods by a particular stockholder during period t. The satisfaction or utility that the stockholder receives from this spending is represented by a function $u(c_t)$, where it is assumed that

$$\frac{\partial u(c_t)}{\partial c_t} > 0 \qquad \frac{\partial^2 u(c_t)}{\partial c_t^2} < 0.$$

The stockholder is presumed to want to maximize

$$\sum_{\tau=0}^{\infty} \beta^{\tau} E\{u(c_{t+\tau})|\mathbf{x}_t^*\}, \qquad [14.2.31]$$

where $\mathbf{x}_t^*$ is a vector representing all the information available to the stockholder at date t and β is a parameter satisfying $0 < \beta < 1$. Smaller values of β mean that the stockholder places a smaller weight on future events. At date t, the stockholder contemplates purchasing any of m different assets, where a dollar invested in asset i at date t will yield a gross return of $(1 + r_{i,t+1})$ at date $t + 1$; in general this rate of return is not known for certain at date t. Assuming that the stockholder takes a position in each of these m assets, the stockholder's optimal portfolio will satisfy

$$u'(c_t) = \beta E\{(1 + r_{i,t+1})u'(c_{t+1})|\mathbf{x}_t^*\} \qquad \text{for } i = 1, 2, \ldots, m, \qquad [14.2.32]$$

where $u'(c_t) \equiv \partial u/\partial c_t$. To see the intuition behind this claim, suppose that condition [14.2.32] failed to hold. Say, for example, that the left side were smaller than the right. Suppose the stockholder were to save one more dollar at date t and invest the dollar in asset i, using the returns to boost period $t + 1$ consumption. Following this strategy would cause consumption at date t to fall by one dollar (reducing [14.2.31] by an amount given by the left side of [14.2.32]), while consumption at date $t + 1$ would rise by $(1 + r_{i,t+1})$ dollars (increasing [14.2.31] by an amount given by the right side of [14.2.32]). If the left side of [14.2.32] were less than the right side of [14.2.32], then the stockholder's objective [14.2.31] would be improved under this change. Only when [14.2.32] is satisfied is the stockholder as well off as possible.[2]

Suppose that the utility function is parameterized as

$$u(c_t) = \begin{cases} \dfrac{c_t^{1-\gamma}}{1-\gamma} & \text{for } \gamma > 0 \text{ and } \gamma \neq 1 \\ \log c_t & \text{for } \gamma = 1. \end{cases}$$

The parameter γ is known as the *coefficient of relative risk aversion*, which for this class of utility functions is a constant. For this function, [14.2.32] becomes

$$c_t^{-\gamma} = \beta E\{(1 + r_{i,t+1})c_{t+1}^{-\gamma}|\mathbf{x}_t^*\}. \qquad [14.2.33]$$

Dividing both sides of [14.2.33] by $c_t^{-\gamma}$ results in

$$1 = \beta E\{(1 + r_{i,t+1})(c_{t+1}/c_t)^{-\gamma}|\mathbf{x}_t^*\}, \qquad [14.2.34]$$

where c_t could be moved inside the conditional expectation operator, since it represents a decision based solely on the information contained in $\mathbf{x}_t^*$. Expression [14.2.34] requires that the random variable described by

$$1 - \beta\{(1 + r_{i,t+1})(c_{t+1}/c_t)^{-\gamma}\} \qquad [14.2.35]$$

be uncorrelated with any variable contained in the information set $\mathbf{x}_t^*$ for any asset i that the stockholder holds. It should therefore be the case that

$$E\{[1 - \beta\{(1 + r_{i,t+1})(c_{t+1}/c_t)^{-\gamma}\}]\mathbf{x}_t\} = \mathbf{0}, \qquad [14.2.36]$$

where $\mathbf{x}_t$ is any subset of the stockholder's information set $\mathbf{x}_t^*$ that the econometrician is also able to observe.

Let $\boldsymbol{\theta} \equiv (\beta, \gamma)'$ denote the unknown parameters that are to be estimated, and let $\mathbf{w}_t \equiv (r_{1,t+1}, r_{2,t+1}, \ldots, r_{m,t+1}, c_{t+1}/c_t, \mathbf{x}_t')'$ denote the vector of variables that are observed by the econometrician for date t. Stacking the equations in [14.2.36] for $i = 1, 2, \ldots, m$ produces a set of r orthogonality conditions that can be used to estimate $\boldsymbol{\theta}$:

$$\underset{(r\times 1)}{\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)} = \begin{bmatrix} [1 - \beta\{(1 + r_{1,t+1})(c_{t+1}/c_t)^{-\gamma}\}]\mathbf{x}_t \\ [1 - \beta\{(1 + r_{2,t+1})(c_{t+1}/c_t)^{-\gamma}\}]\mathbf{x}_t \\ \vdots \\ [1 - \beta\{(1 + r_{m,t+1})(c_{t+1}/c_t)^{-\gamma}\}]\mathbf{x}_t \end{bmatrix}. \qquad [14.2.37]$$

The sample average value of $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$ is

$$\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T) \equiv (1/T)\sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t),$$

and the *GMM* objective function is

$$Q(\boldsymbol{\theta}) = [\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]'\hat{\mathbf{S}}_T^{-1}[\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)]. \qquad [14.2.38]$$

This expression can then be minimized numerically with respect to $\boldsymbol{\theta}$.

According to the theory, the magnitude in [14.2.35] should be uncorrelated with any information the stockholder has available at time t, which would include

[2] For further details, see Sargent (1987).

lagged values of [14.2.35]. Hence, the vector in [14.2.37] should be uncorrelated with its own lagged values, suggesting that $\mathbf{S}$ can be consistently estimated by

$$\hat{\mathbf{S}}_T = (1/T) \sum_{t=1}^{T} \{[\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathbf{w}_t)]'\},$$

where $\hat{\boldsymbol{\theta}}$ is an initial consistent estimate. This initial estimate $\hat{\boldsymbol{\theta}}$ could be obtained by minimizing [14.2.38] with $\hat{\mathbf{S}}_T = \mathbf{I}_r$.

Hansen and Singleton (1982) estimated such a model using real consumption expenditures for the aggregate United States divided by the U.S. population as their measure of c_t. For r_{1t}, they used the inflation-adjusted return that an investor would earn if one dollar was invested in every stock listed on the New York Stock Exchange, while r_{2t} was a value-weighted inflation-adjusted return corresponding to the return an investor would earn if the investor owned the entire stock of each company listed on the exchange. Hansen and Singleton's instruments consisted of a constant term, lagged consumption growth rates, and lagged rates of return:

$$\mathbf{x}_t = (1, c_t/c_{t-1}, c_{t-1}/c_{t-2}, \ldots, c_{t-\ell+1}/c_{t-\ell}, r_{1t}, r_{1,t-1}, \ldots,$$
$$r_{1,t-\ell+1}, r_{2,t}, r_{2,t-1}, \ldots, r_{2,t-\ell+1})'.$$

When ℓ lags are used, there are $3\ell + 1$ elements in $\mathbf{x}_t$, and thus $r = 2(3\ell + 1)$ separate orthogonality conditions are represented by [14.2.37]. Since $a = 2$ parameters are estimated, the χ^2 statistic in [14.1.27] has 6ℓ degrees of freedom.

14.3. *Extensions*

GMM *with Nonstationary Data*

The maintained assumption throughout this chapter has been that the $(h \times 1)$ vector of observed variables $\mathbf{w}_t$ is strictly stationary. Even if the raw data appear to be trending over time, sometimes the model can be transformed or reparameterized so that stationarity of the transformed system is a reasonable assumption. For example, the consumption series $\{c_t\}$ used in Hansen and Singleton's study (1982) is increasing over time. However, it was possible to write the equation to be estimated [14.2.36] in such a form that only the consumption growth rate (c_{t+1}/c_t) appeared, for which the stationarity assumption is much more plausible. Alternatively, suppose that some of the elements of the observed vector $\mathbf{w}_t$ are presumed to grow deterministically over time according to

$$\mathbf{w}_t = \boldsymbol{\alpha} + \boldsymbol{\delta} \cdot t + \mathbf{w}_t^*, \qquad [14.3.1]$$

where $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ are $(h \times 1)$ vectors of constants and $\mathbf{w}_t^*$ is strictly stationary with mean zero. Suppose that the orthogonality conditions can be expressed in terms of $\mathbf{w}_t^*$ as

$$E\{\mathbf{f}(\boldsymbol{\theta}_0, \mathbf{w}_t^*)\} = \mathbf{0}.$$

Then Ogaki (1993) recommended jointly estimating $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\delta}$ using

$$\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t) = \begin{bmatrix} \mathbf{w}_t - \boldsymbol{\alpha} - \boldsymbol{\delta}t \\ \mathbf{f}(\boldsymbol{\theta}, \mathbf{w}_t - \boldsymbol{\alpha} - \boldsymbol{\delta}t) \end{bmatrix}$$

to construct the moment condition in [14.2.3].

Testing for Structural Stability

Suppose we want to test the hypothesis that the $(a \times 1)$ parameter vector $\boldsymbol{\theta}$ that characterizes the first T_0 observations in the sample is different from the value that characterizes the last $T - T_0$ observations, where T_0 is a known change point. One approach is to obtain an estimate $\hat{\boldsymbol{\theta}}_{1.T_0}$ based solely on the first T_0 observations, minimizing

$$\begin{aligned} &Q(\boldsymbol{\theta}_1; \mathbf{w}_{T_0}, \mathbf{w}_{T_0-1}, \ldots, \mathbf{w}_1) \\ &= \left[(1/T_0)\sum_{t=1}^{T_0} \mathbf{h}(\boldsymbol{\theta}_1, \mathbf{w}_t)\right]' \hat{\mathbf{S}}_{1.T_0}^{-1} \left[(1/T_0)\sum_{t=1}^{T_0} \mathbf{h}(\boldsymbol{\theta}_1, \mathbf{w}_t)\right], \end{aligned} \qquad [14.3.2]$$

where, for example, if $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathbf{w}_t)\}$ is serially uncorrelated,

$$\hat{\mathbf{S}}_{1.T_0} = (1/T_0)\sum_{t=1}^{T_0} [h(\hat{\boldsymbol{\theta}}_{1.T_0}, \mathbf{w}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}_{1.T_0}, \mathbf{w}_t)]'.$$

Proposition 14.1 implies that

$$\sqrt{T_0}(\hat{\boldsymbol{\theta}}_{1.T_0} - \boldsymbol{\theta}_1) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}_1) \qquad [14.3.3]$$

as $T_0 \rightarrow \infty$, where $\mathbf{V}_1$ can be estimated from

$$\hat{\mathbf{V}}_{1.T_0} = \{\hat{\mathbf{D}}_{1.T_0}\hat{\mathbf{S}}_{1.T_0}^{-1}\hat{\mathbf{D}}_{1.T_0}'\}^{-1}$$

for

$$\hat{\mathbf{D}}_{1.T_0}' \equiv (1/T_0)\sum_{t=1}^{T_0} \left.\frac{\partial \mathbf{h}(\boldsymbol{\theta}_1, \mathbf{w}_t)}{\partial \boldsymbol{\theta}_1'}\right|_{\boldsymbol{\theta}_1 = \hat{\boldsymbol{\theta}}_{1.T_0}}.$$

Similarly, a separate estimate $\hat{\boldsymbol{\theta}}_{2.T-T_0}$ can be based on the last $T - T_0$ observations, with analogous measures $\hat{\mathbf{S}}_{2.T-T_0}$, $\hat{\mathbf{V}}_{2.T-T_0}$, $\hat{\mathbf{D}}_{2.T-T_0}$, and

$$\sqrt{T - T_0}(\hat{\boldsymbol{\theta}}_{2.T-T_0} - \boldsymbol{\theta}_2) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}_2) \qquad [14.3.4]$$

as $T - T_0 \rightarrow \infty$. Let $\pi \equiv T_0/T$ denote the fraction of observations contained in the first subsample. Then [14.3.3] and [14.3.4] state that

$$\begin{aligned} \sqrt{T}(\hat{\boldsymbol{\theta}}_{1.T_0} - \boldsymbol{\theta}_1) &\xrightarrow{L} N(\mathbf{0}, \mathbf{V}_1/\pi) \\ \sqrt{T}(\hat{\boldsymbol{\theta}}_{2.T-T_0} - \boldsymbol{\theta}_2) &\xrightarrow{L} N(\mathbf{0}, \mathbf{V}_2/(1 - \pi)) \end{aligned}$$

as $T \rightarrow \infty$. Andrews and Fair (1988) suggested using a Wald test of the null hypothesis that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$, exploiting the fact that under the stationarity conditions needed to justify Proposition 14.1, $\hat{\boldsymbol{\theta}}_1$ is asymptotically independent of $\hat{\boldsymbol{\theta}}_2$:

$$\begin{aligned} \lambda_T = {} & T(\hat{\boldsymbol{\theta}}_{1.T_0} - \hat{\boldsymbol{\theta}}_{2.T-T_0})' \\ & \times \{\pi^{-1}\cdot\hat{\mathbf{V}}_{1.T_0} + (1 - \pi)^{-1}\cdot\hat{\mathbf{V}}_{2.T-T_0}\}^{-1}(\hat{\boldsymbol{\theta}}_{1.T_0} - \hat{\boldsymbol{\theta}}_{2.T-T_0}). \end{aligned}$$

Then $\lambda_T \xrightarrow{L} \chi^2(a)$ under the null hypothesis that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

One can further test for structural change at a variety of different possible dates, repeating the foregoing test for all T_0 between, say, $0.15T$ and $0.85T$ and choosing the largest value for the resulting test statistic λ_T. Andrews (1993) described the asymptotic distribution of such a test.

Another simple test associates separate moment conditions with the observations before and after T_0 and uses the χ^2 test suggested in [14.1.27] to test the validity of the separate sets of conditions. Specifically, let

$$d_{1t} = \begin{cases} 1 & \text{for } t \leq T_0 \\ 0 & \text{for } t > T_0. \end{cases}$$

If $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$ is an $(r \times 1)$ vector whose population mean is zero at $\boldsymbol{\theta}_0$, define

$$\underset{(2r\times 1)}{\mathbf{h}^*(\boldsymbol{\theta}, \mathbf{w}_t, d_{1t})} \equiv \begin{bmatrix} \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\cdot d_{1t} \\ \mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)\cdot(1 - d_{1t}) \end{bmatrix}.$$

The a elements of $\boldsymbol{\theta}$ can then be estimated by using the $2r$ orthogonality conditions given by $E\{\mathbf{h}^*(\boldsymbol{\theta}_0, \mathbf{w}_t, d_{1t})\} = \mathbf{0}$ for $t = 1, 2, \ldots, T$, by simply replacing $\mathbf{h}(\boldsymbol{\theta}, \mathbf{w}_t)$

in [14.2.3] with $\mathbf{h}^*(\boldsymbol{\theta}, \mathbf{w}_t, d_{1t})$ and minimizing [14.2.2] in the usual way. Hansen's χ^2 test statistic described in [14.1.27] based on the $\mathbf{h}^*(\cdot)$ moment conditions could then be compared with a $\chi^2(2r - a)$ critical value to provide a test of the hypothesis that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

A number of other tests for structural change have been proposed by Andrews and Fair (1988) and Ghysels and Hall (1990a, b).

GMM *and Econometric Identification*

For the portfolio decision model [14.2.34], it was argued that any variable would be valid to include in the instrument vector $\mathbf{x}_t$, as long as that variable was known to investors at date t and their expectations were formed rationally. Essentially, [14.2.34] represents an asset demand curve. In the light of the discussion of simultaneous equations bias in Section 9.1, one might be troubled by the claim that it is possible to estimate a demand curve without needing to think about the way that variables may affect the demand and supply of assets in different ways.

As stressed by Garber and King (1984), the portfolio choice model avoids simultaneous equations bias because it postulates that equation [14.2.32] holds *exactly*, with no error term. The model as written claims that if the econometrician had the same information $\mathbf{x}_t^*$ used by investors, then investors' behavior could be predicted with an R^2 of unity. If there were no error term in the demand for oranges equation [9.1.1], or if the error in the demand for oranges equation were negligible compared with the error term in the supply equation, then we would not have had to worry about simultaneous equations bias in that example, either.

It is hard to take seriously the suggestion that the observed data are exactly described by [14.2.32] with no error. There are substantial difficulties in measuring aggregate consumption, population, and rates of return on assets. Even if these aggregates could in some sense be measured perfectly, it is questionable whether they are the appropriate values to be using to test a theory about individual investor preferences. And even if we had available a perfect measure of the consumption of an individual investor, the notion that the investor's utility could be represented by a function of this precise parametric form with γ constant across time is surely hard to defend.

Once we acknowledge that an error term reasonably ought to be included in [14.2.32], then it is no longer satisfactory to say that any variable dated t or earlier is a valid instrument. The difficulties with estimation are compounded by the nonlinearity of the equations of interest. If one wants to take seriously the possibility of an error term in [14.2.32] and its correlation with other variables, the best approach currently available appears to be to linearize the dynamic rational expectations model. Any variable uncorrelated with both the forecast error people make and the specification error in the model could then be used as a valid instrument for traditional instrumental variable estimation; see Sill (1992) for an illustration of this approach.

Optimal Choice of Instruments

If one does subscribe to the view that any variable dated t or earlier is a valid instrument for estimation of [14.2.32], this suggests a virtually infinite set of possible variables that could be used. One's first thought might be that, the more orthogonality conditions used, the better the resulting estimates might be. However, Monte Carlo simulations by Tauchen (1986) and Kocherlakota (1990) strongly suggest that one should be quite parsimonious in the selection of $\mathbf{x}_t$. Nelson and

Startz (1990) in particular stress that in the linear simultaneous equations model $y_t = \mathbf{z}_t'\boldsymbol{\beta} + u_t$, a good instrument not only must be uncorrelated with u_t, but must also be strongly correlated with $\mathbf{z}_t$. See Bates and White (1988), Hall (1993), and Gallant and Tauchen (1992) for further discussion on instrument selection.

14.4. GMM *and Maximum Likelihood Estimation*

In many cases the maximum likelihood estimate of $\boldsymbol{\theta}$ can also be viewed as a *GMM* estimate. This section explores this analogy and shows how asymptotic properties of maximum likelihood estimation and quasi-maximum likelihood can be obtained from the previous general results about *GMM* estimation.

The Score and Its Population Properties

Let $\mathbf{y}_t$ denote an $(n \times 1)$ vector of variables observed at date t, and let $\mathcal{Y}_t \equiv (\mathbf{y}_t', \mathbf{y}_{t-1}', \ldots, \mathbf{y}_1')'$ denote the full set of data observed through date t. Suppose that the conditional density of the tth observation is given by

$$f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta}). \qquad [14.4.1]$$

Since [14.4.1] is a density, it must integrate to unity:

$$\int_{\mathcal{A}} f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = 1, \qquad [14.4.2]$$

where $\mathcal{A}$ denotes the set of possible values that $\mathbf{y}_t$ could take on and $\int d\mathbf{y}_t$ denotes multiple integration:

$$\int h(\mathbf{y}_t)\, d\mathbf{y}_t \equiv \int \int \cdots \int h(y_{1t}, y_{2t}, \ldots, y_{nt})\, dy_{1t}\, dy_{2t} \cdots dy_{nt}.$$

Since [14.4.2] holds for all admissible values of $\boldsymbol{\theta}$, we can differentiate both sides with respect to $\boldsymbol{\theta}$ to conclude that

$$\int_{\mathcal{A}} \frac{\partial f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\, d\mathbf{y}_t = \mathbf{0}. \qquad [14.4.3]$$

The conditions under which the order of differentiation and integration can be reversed as assumed in arriving at [14.4.3] and the equations to follow are known as "regularity conditions" and are detailed in Cramér (1946). Assuming that these hold, we can multiply and divide the integrand in [14.4.3] by the conditional density of $\mathbf{y}_t$:

$$\int_{\mathcal{A}} \frac{\partial f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{1}{f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})} f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = \mathbf{0},$$

or

$$\int_{\mathcal{A}} \frac{\partial \log f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = \mathbf{0}. \qquad [14.4.4]$$

Let $\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)$ denote the derivative of the log of the conditional density of the tth observation:

$$\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) = \frac{\partial \log f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}. \qquad [14.4.5]$$

If there are a elements in $\boldsymbol{\theta}$, then [14.4.5] describes an $(a \times 1)$ vector for each date t that is known as the *score* of the tth observation. Since the score is a function of $\mathcal{Y}_t$, it is a random variable. Moreover, substitution of [14.4.5] into [14.4.4] reveals that

$$\int_{\mathcal{A}} \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = \mathbf{0}. \qquad [14.4.6]$$

Equation [14.4.6] indicates that if the data were really generated by the density [14.4.1], then the expected value of the score conditional on information observed through date $t - 1$ should be zero:

$$E\{\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) | \mathcal{Y}_{t-1}\} = \mathbf{0}. \qquad [14.4.7]$$

In other words, the score vectors $\{\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)\}_{t=1}^{\infty}$ should form a martingale difference sequence. This observation prompted White (1987) to suggest a general specification test for models estimated by maximum likelihood based on whether the sample scores appear to be serially correlated. Expression [14.4.7] further implies that the score has unconditional expectation of zero, provided that the unconditional first moment exists:

$$E\{\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)\} = \mathbf{0}. \qquad [14.4.8]$$

Maximum Likelihood and GMM

Expression [14.4.8] can be viewed as a set of a orthogonality conditions that could be used to estimate the a unknown elements of $\boldsymbol{\theta}$. The *GMM* principle suggests using as an estimate of $\boldsymbol{\theta}$ the solution to

$$\mathbf{0} = (1/T) \sum_{t=1}^{T} \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t). \qquad [14.4.9]$$

But this is also the characterization of the maximum likelihood estimate, which is based on maximization of

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta}),$$

the first-order conditions for which are

$$\sum_{t=1}^{T} \frac{\partial \log f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \qquad [14.4.10]$$

assuming an interior maximum. Recalling [14.4.5], observe that [14.4.10] and [14.4.9] are identical conditions—the *MLE* is the same as the *GMM* estimator based on the orthogonality conditions in [14.4.8].

The *GMM* formula [14.2.6] suggests that the variance-covariance matrix of the *MLE* can be approximated by

$$E[(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0)'] \cong (1/T)\{\hat{\mathbf{D}}_T \hat{\mathbf{S}}_T^{-1} \hat{\mathbf{D}}_T'\}^{-1}, \qquad [14.4.11]$$

where

$$\begin{aligned}
\underset{(a \times a)}{\hat{\mathbf{D}}_T'} &= \left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T} \\
&= (1/T) \sum_{t=1}^{T} \left.\frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T} \qquad [14.4.12] \\
&= (1/T) \sum_{t=1}^{T} \left.\frac{\partial^2 \log f(\mathbf{y}_t | \mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_T}.
\end{aligned}$$

Moreover, the observation in [14.4.7] that the scores are serially uncorrelated suggests estimating $\mathbf{S}$ by

$$\hat{\mathbf{S}}_T = (1/T)\sum_{t=1}^{T}[\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)][\mathbf{h}(\hat{\boldsymbol{\theta}}, \mathcal{Y}_t)]'. \qquad [14.4.13]$$

The Information Matrix Equality

Expression [14.4.12] will be recognized as -1 times the second derivative estimate of the information matrix. Similarly, expression [14.4.13] is the outer-product estimate of the information matrix. That these two expressions are indeed estimating the same matrix if the model is correctly specified can be seen from calculations similar to those that produced [14.4.6]. Differentiating both sides of [14.4.6] with respect to $\boldsymbol{\theta}'$ reveals that

$$\begin{aligned}\mathbf{0} &= \int_{\mathcal{A}} \frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)}{\partial \boldsymbol{\theta}'} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t + \int_{\mathcal{A}} \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) \frac{\partial f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\, d\mathbf{y}_t \\ &= \int_{\mathcal{A}} \frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)}{\partial \boldsymbol{\theta}'} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t \\ &\quad + \int_{\mathcal{A}} \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) \frac{\partial \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t\end{aligned}$$

or

$$\int_{\mathcal{A}} [\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)][\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)]' f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t = -\int_{\mathcal{A}} \frac{\partial \mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t)}{\partial \boldsymbol{\theta}'} f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})\, d\mathbf{y}_t.$$

This equation implies that if the model is correctly specified, the expected value of the outer product of the vector of first derivatives of the log likelihood is equal to the negative of the expected value of the matrix of second derivatives:

$$\begin{aligned}&E\left\{\left[\frac{\partial \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]\left[\frac{\partial \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right]\middle|\,\mathcal{Y}_{t-1}\right\} \\ &\quad = -E\left\{\frac{\partial^2 \log f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\middle|\,\mathcal{Y}_{t-1}\right\} \qquad [14.4.14] \\ &\quad \equiv \boldsymbol{\mathcal{J}}_t.\end{aligned}$$

Expression [14.4.14] is known as the *information matrix equality*. Assuming that $(1/T)\sum_{t=1}^{T} \boldsymbol{\mathcal{J}}_t \xrightarrow{p} \boldsymbol{\mathcal{J}}$, a positive definite matrix, we can reasonably expect that for many models, the estimate $\hat{\mathbf{S}}_T$ in [14.4.13] converges in probability to the information matrix $\boldsymbol{\mathcal{J}}$ and the estimate $\hat{\mathbf{D}}_T'$ in [14.4.12] converges in probability to $-\boldsymbol{\mathcal{J}}$. Thus, result [14.4.11] suggests that if the data are stationary and the estimates do not fall on the boundaries of the allowable parameter space, it will often be the case that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, \boldsymbol{\mathcal{J}}^{-1}), \qquad [14.4.15]$$

where the information matrix $\boldsymbol{\mathcal{J}}$ can be estimated consistently from either $-\hat{\mathbf{D}}_T'$ in [14.4.12] or $\hat{\mathbf{S}}_T$ in [14.4.13].

In small samples, the estimates $-\hat{\mathbf{D}}_T'$ and $\hat{\mathbf{S}}_T$ will differ, though if they differ too greatly this suggests that the model may be misspecified. White (1982) developed an alternative specification test based on comparing these two magnitudes.

The Wald Test for Maximum Likelihood Estimates

Result [14.4.15] suggests a general approach to testing hypotheses about the value of a parameter vector $\boldsymbol{\theta}$ that has been estimated by maximum likelihood.

Consider a null hypothesis involving m restrictions on $\boldsymbol{\theta}$ represented as $\mathbf{g}(\boldsymbol{\theta}) = \mathbf{0}$ where $\mathbf{g}\colon \mathbb{R}^a \to \mathbb{R}^m$ is a known differentiable function. The Wald test of this hypothesis is given by

$$\underset{(1\times m)}{T[\mathbf{g}(\hat{\boldsymbol{\theta}}_T)]'}\left\{\underset{(m\times a)}{\left[\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right]}\underset{(a\times a)}{\hat{\mathcal{J}}_T^{-1}}\underset{(a\times m)}{\left[\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right]'}\right\}^{-1}\underset{(m\times 1)}{[\mathbf{g}(\hat{\boldsymbol{\theta}}_T)]}, \qquad [14.4.16]$$

which converges in distribution to a $\chi^2(m)$ variable under the null hypothesis. Again, the estimate of the information matrix $\hat{\mathcal{J}}_T$ could be based on either $-\hat{\mathbf{D}}_T'$ in [14.4.12] or $\hat{\mathbf{S}}_T$ in [14.4.13].

The Lagrange Multiplier Test

We have seen that if the model is correctly specified, the scores $\{\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t)\}_{t=1}^{\infty}$ often form a martingale difference sequence. Expression [14.4.14] indicates that the conditional variance-covariance matrix of the tth score is given by $\mathcal{J}_t$. Hence, typically,

$$T\left[(1/T)\sum_{t=1}^{T}\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t)\right]'\hat{\mathcal{J}}_T^{-1}\left[(1/T)\sum_{t=1}^{T}\mathbf{h}(\boldsymbol{\theta}_0, \mathcal{Y}_t)\right] \xrightarrow{L} \chi^2(a). \qquad [14.4.17]$$

Expression [14.4.17] does not hold when $\boldsymbol{\theta}_0$ is replaced by $\hat{\boldsymbol{\theta}}_T$, since, from [14.4.9], this would cause [14.4.17] to be identically zero.

Suppose, however, that the likelihood function is maximized subject to m constraints on $\boldsymbol{\theta}$, and let $\tilde{\boldsymbol{\theta}}_T$ denote the restricted estimate of $\boldsymbol{\theta}$. Then, as in the *GMM* test for overidentifying restrictions [14.1.27], we would expect that

$$T\left[(1/T)\sum_{t=1}^{T}\mathbf{h}(\tilde{\boldsymbol{\theta}}_T, \mathcal{Y}_t)\right]'\hat{\mathcal{J}}_T^{-1}\left[(1/T)\sum_{t=1}^{T}\mathbf{h}(\tilde{\boldsymbol{\theta}}_T, \mathcal{Y}_t)\right] \xrightarrow{L} \chi^2(m). \qquad [14.4.18]$$

The magnitude in [14.4.18] was called the *efficient score* statistic by Rao (1948) and the *Lagrange multiplier test* by Aitchison and Silvey (1958). It provides an extremely useful class of diagnostic tests, enabling one to estimate a restricted model and test it against a more general specification without having to estimate the more general model. Breusch and Pagan (1980), Engle (1984), and Godfrey (1988) illustrated applications of the usefulness of the Lagrange multiplier principle.

Quasi-Maximum Likelihood Estimation

Even if the data were not generated by the density $f(\mathbf{y}_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta})$, the orthogonality conditions [14.4.8] might still provide a useful description of the parameter vector of interest. For example, suppose that we incorrectly specified that a scalar series y_t came from a Gaussian $AR(1)$ process:

$$\log f(y_t|\mathcal{Y}_{t-1}; \boldsymbol{\theta}) = -\tfrac{1}{2}\log(2\pi) - \tfrac{1}{2}\log(\sigma^2) - (y_t - \phi y_{t-1})^2/(2\sigma^2),$$

with $\boldsymbol{\theta} \equiv (\phi, \sigma^2)'$. The score vector is then

$$\mathbf{h}(\boldsymbol{\theta}, \mathcal{Y}_t) = \begin{bmatrix} (y_t - \phi y_{t-1})y_{t-1}/\sigma^2 \\ -1/(2\sigma^2) + (y_t - \phi y_{t-1})^2/(2\sigma^4) \end{bmatrix},$$

which has expectation zero whenever

$$E[(y_t - \phi y_{t-1})y_{t-1}] = 0 \qquad [14.4.19]$$

$$E[(y_t - \phi y_{t-1})^2] = \sigma^2. \qquad [14.4.20]$$

The value of the parameter ϕ that satisfies [14.4.19] corresponds to the coefficient of a linear projection of y_t on y_{t-1} regardless of the time series process followed by y_t, while σ^2 in [14.4.20] is a general characterization of the mean squared error of this linear projection. Hence, the moment conditions in [14.4.8] hold for a broad class of possible processes, and the estimates obtained by maximizing a Gaussian likelihood function (that is, the values satisfying [14.4.9]) should give reasonable estimates of the linear projection coefficient and its mean squared error for a fairly general class of possible data-generating mechanisms.

However, if the data were not generated by a Gaussian $AR(1)$ process, then the information matrix equality no longer need hold. As long as the score vector is serially uncorrelated, the variance-covariance matrix of the resulting estimates could be obtained from [14.4.11]. Proceeding in this fashion—maximizing the likelihood function in the usual way, but using [14.4.11] rather than [14.4.15] to calculate standard errors—was first proposed by White (1982), who described this approach as *quasi-maximum likelihood estimation.*[3]

APPENDIX 14.A. *Proof of Chapter 14 Proposition*

■ **Proof of Proposition 14.1.** Let $g_i(\boldsymbol{\theta}; \mathcal{Y}_T)$ denote the ith element of $\mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)$, so that g_i: $\mathbb{R}^a \rightarrow \mathbb{R}^1$. By the mean-value theorem,

$$g_i(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = g_i(\boldsymbol{\theta}_0; \mathcal{Y}_T) + [\mathbf{d}_i(\boldsymbol{\theta}^*_{i,T}; \mathcal{Y}_T)]'(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0), \qquad [14.A.1]$$

where

$$\underset{(a\times 1)}{\mathbf{d}_i(\boldsymbol{\theta}^*_{i,T}; \mathcal{Y}_T)} = \left.\frac{\partial g_i(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*_{i,T}}$$

for some $\boldsymbol{\theta}^*_{i,T}$ between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_T$; notice that $\mathbf{d}_i$: $\mathbb{R}^a \rightarrow \mathbb{R}^a$. Define

$$\underset{(r\times a)}{\mathbf{D}'_T} \equiv \begin{bmatrix} [\mathbf{d}_1(\boldsymbol{\theta}^*_{1,T}; \mathcal{Y}_T)]' \\ [\mathbf{d}_2(\boldsymbol{\theta}^*_{2,T}; \mathcal{Y}_T)]' \\ \vdots \\ [\mathbf{d}_r(\boldsymbol{\theta}^*_{r,T}; \mathcal{Y}_T)]' \end{bmatrix}. \qquad [14.A.2]$$

Stacking the equations in [14.A.1] in an $(r \times 1)$ vector produces

$$\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T) = \mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T) + \mathbf{D}'_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0). \qquad [14.A.3]$$

If both sides of [14.A.3] are premultiplied by the $(a \times r)$ matrix

$$\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1},$$

the result is

$$\begin{aligned} &\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times [\mathbf{g}(\hat{\boldsymbol{\theta}}_T; \mathcal{Y}_T)] \\ &\quad = \left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times [\mathbf{g}(\boldsymbol{\theta}_0; \mathcal{Y}_T)] \qquad [14.A.4] \\ &\quad\quad + \left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta}; \mathcal{Y}_T)}{\partial \boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times \mathbf{D}'_T(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0). \end{aligned}$$

[3] For further discussion, see Gourieroux, Monfort, and Trognon (1984), Gallant and White (1988), and Wooldridge (1991a, b).

But equation [14.1.22] implies that the left side of [14.A.4] is zero, so that

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) = -&\left[\left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta};\mathcal{Y}_T)}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times \mathbf{D}_T'\right]^{-1} \\ &\times \left\{\left.\frac{\partial \mathbf{g}(\boldsymbol{\theta};\mathcal{Y}_T)}{\partial\boldsymbol{\theta}'}\right|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}_T}\right\}' \times \hat{\mathbf{S}}_T^{-1} \times [\mathbf{g}(\boldsymbol{\theta}_0;\mathcal{Y}_T)].\end{aligned} \qquad [14.A.5]$$

Now, $\boldsymbol{\theta}_{i,T}^*$ in [14.A.1] is between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}_T$, so that $\boldsymbol{\theta}_{i,T}^* \xrightarrow{p} \boldsymbol{\theta}_0$ for each i. Thus, condition (c) ensures that each row of $\mathbf{D}_T'$ converges in probability to the corresponding row of $\mathbf{D}'$. Then [14.A.5] implies that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{p} -\{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1} \times \{\mathbf{D}\mathbf{S}^{-1}\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0;\mathcal{Y}_T)\}. \qquad [14.A.6]$$

Define

$$\mathbf{C} \equiv -\{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1} \times \mathbf{D}\mathbf{S}^{-1},$$

so that [14.A.6] becomes

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{p} \mathbf{C}\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0;\mathcal{Y}_T).$$

Recall from condition (b) of the proposition that

$$\sqrt{T}\cdot\mathbf{g}(\boldsymbol{\theta}_0;\mathcal{Y}_T) \xrightarrow{L} N(\mathbf{0}, \mathbf{S}).$$

It follows as in Example 7.5 of Chapter 7 that

$$\sqrt{T}(\hat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{L} N(\mathbf{0}, \mathbf{V}), \qquad [14.A.7]$$

where

$$\mathbf{V} = \mathbf{C}\mathbf{S}\mathbf{C}' = \{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1}\mathbf{D}\mathbf{S}^{-1} \times \mathbf{S} \times \mathbf{S}^{-1}\mathbf{D}'\{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1} = \{\mathbf{D}\mathbf{S}^{-1}\mathbf{D}'\}^{-1},$$

as claimed. ■

Chapter 14 Exercise

14.1. Consider the Gaussian linear regression model,

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + u_t,$$

with $u_t \sim$ i.i.d. $N(0, \sigma^2)$ and u_t independent of $\mathbf{x}_\tau$ for all t and τ. Define $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}', \sigma^2)'$. The log of the likelihood of $(y_1, y_2, \ldots, y_T)$ conditional on $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T)$ is given by

$$\mathcal{L}(\boldsymbol{\theta}) = -(T/2)\log(2\pi) - (T/2)\log(\sigma^2) - \sum_{t=1}^{T}(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2/(2\sigma^2).$$

(a) Show that the estimate $\hat{\mathbf{D}}_T'$ in [14.4.12] is given by

$$\hat{\mathbf{D}}_T' = \begin{bmatrix} -\dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'/\hat{\sigma}_T^2 & \mathbf{0} \\ \mathbf{0}' & \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\left\{\dfrac{1}{2\hat{\sigma}_T^4} - \dfrac{\hat{u}_t^2}{\hat{\sigma}_T^6}\right\}\end{bmatrix},$$

where $\hat{u}_t \equiv (y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}}_T)$ and $\hat{\boldsymbol{\beta}}_T$ and $\hat{\sigma}_T^2$ denote the maximum likelihood estimates.

(b) Show that the estimate $\hat{\mathbf{S}}_T$ in [14.4.13] is given by

$$\hat{\mathbf{S}}_T = \begin{bmatrix} \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\hat{u}_t^2\mathbf{x}_t\mathbf{x}_t'/\hat{\sigma}_T^4 & \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\left\{\dfrac{\hat{u}_t^3\mathbf{x}_t}{2\hat{\sigma}_T^6}\right\} \\ \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\left\{\dfrac{\hat{u}_t^3\mathbf{x}_t'}{2\hat{\sigma}_T^6}\right\} & \dfrac{1}{T}\displaystyle\sum_{t=1}^{T}\left\{\dfrac{\hat{u}_t^2}{2\hat{\sigma}_T^4} - \dfrac{1}{2\hat{\sigma}_T^2}\right\}^2\end{bmatrix}.$$

(c) Show that $\text{plim}(\hat{\mathbf{S}}_T) = -\text{plim}(\hat{\mathbf{D}}_T) = \boldsymbol{\mathcal{I}}$, where

$$\boldsymbol{\mathcal{I}} = \begin{bmatrix}\mathbf{Q}/\sigma^2 & \mathbf{0} \\ \mathbf{0}' & 1/(2\sigma^4)\end{bmatrix}$$

for $\mathbf{Q} = \text{plim}(1/T)\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$.

(d) Consider a set of m linear restrictions on $\boldsymbol{\beta}$ of the form $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ for $\mathbf{R}$ a known $(m \times k)$ matrix and $\mathbf{r}$ a known $(m \times 1)$ vector. Show that for $\hat{\mathscr{I}}_T = -\hat{\mathbf{D}}_T$, the Wald test statistic given in [14.4.16] is identical to the Wald form of the *OLS* χ^2 test in [8.2.23] with the *OLS* estimate of the variance s_T^2 in [8.2.23] replaced by the *MLE* $\hat{\sigma}_T^2$.

(e) Show that when the lower left and upper right blocks of $\hat{\mathbf{S}}_T$ are set to their plim of zero, then the quasi-maximum likelihood Wald test of $\mathbf{R}\boldsymbol{\beta} = \mathbf{r}$ is identical to the heteroskedasticity-consistent form of the *OLS* χ^2 test given in [8.2.36].

Chapter 14 References

Aitchison, J., and S. D. Silvey. 1958. "Maximum Likelihood Estimation of Parameters Subject to Restraints." *Annals of Mathematical Statistics* 29:813–28.

Amemiya, Takeshi. 1974. "The Nonlinear Two-Stage Least-Squares Estimator." *Journal of Econometrics* 2:105–10.

Andrews, Donald W. K. 1991. "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation." *Econometrica* 59:817–58.

———. 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point." *Econometrica* 61:821–56.

——— and Ray C. Fair. 1988. "Inference in Nonlinear Econometric Models with Structural Change." *Review of Economic Studies* 55:615–40.

——— and J. Christopher Monahan. 1992. "An Improved Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimator." *Econometrica* 60:953–66.

Bates, Charles, and Halbert White. 1988. "Efficient Instrumental Variables Estimation of Systems of Implicit Heterogeneous Nonlinear Dynamic Equations with Nonspherical Errors," in William A. Barnett, Ernst R. Berndt, and Halbert White, eds., *Dynamic Econometric Modeling*. Cambridge, England: Cambridge University Press.

Breusch, T. S., and A. R. Pagan. 1980. "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics." *Review of Economic Studies* 47:239–53.

Cramér, H. 1946. *Mathematical Methods of Statistics*. Princeton, N.J.: Princeton University Press.

Engle, Robert F. 1984. "Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics," in Zvi Griliches and Michael D. Intriligator, eds., *Handbook of Econometrics*, Vol. 2. Amsterdam: North-Holland.

Ferguson, T. S. 1958. "A Method of Generating Best Asymptotically Normal Estimates with Application to the Estimation of Bacterial Densities." *Annals of Mathematical Statistics* 29:1046–62.

Gallant, A. Ronald. 1977. "Three-Stage Least-Squares Estimation for a System of Simultaneous, Nonlinear, Implicit Equations." *Journal of Econometrics* 5:71–88.

———. 1987. *Nonlinear Statistical Models*. New York: Wiley.

——— and George Tauchen. 1992. "Which Moments to Match?" Duke University. Mimeo.

——— and Halbert White. 1988. *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*. Oxford: Blackwell.

Garber, Peter M., and Robert G. King. 1984. "Deep Structural Excavation? A Critique of Euler Equation Methods." University of Rochester. Mimeo.

Ghysels, Eric, and Alastair Hall. 1990a. "A Test for Structural Stability of Euler Conditions Parameters Estimated via the Generalized Method of Moments Estimator." *International Economic Review* 31:355–64.

——— and ———. 1990b. "Are Consumption-Based Intertemporal Capital Asset Pricing Models Structural?" *Journal of Econometrics* 45:121–39.

Godfrey, L. G. 1988. *Misspecification Tests in Econometrics: The Lagrange Multiplier Principle and Other Approaches*. Cambridge, England: Cambridge University Press.

Gourieroux, C., A. Monfort, and A. Trognon. 1984. "Pseudo Maximum Likelihood Methods: Theory." *Econometrica* 52:681–700.

Hall, Alastair. 1993. "Some Aspects of Generalized Method of Moments Estimation," in C. R. Rao, G. S. Maddala, and H. D. Vinod, eds., *Handbook of Statistics*, Vol. 11, *Econometrics*. Amsterdam: North-Holland.

Hansen, Lars P. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica* 50:1029–54.

——— and Kenneth J. Singleton. 1982. "Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models." *Econometrica* 50:1269–86. Errata: *Econometrica* 52:267–68.

Jorgenson, D. W., and J. Laffont. 1974. "Efficient Estimation of Nonlinear Simultaneous Equations with Additive Disturbances." *Annals of Economic and Social Measurement* 3:615–40.

Kocherlakota, Narayana R. 1990. "On Tests of Representative Consumer Asset Pricing Models." *Journal of Monetary Economics* 26:285–304.

Malinvaud, E. 1970. *Statistical Methods of Econometrics*. Amsterdam: North-Holland.

Nelson, Charles R., and Richard Startz. 1990. "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58:967–76.

Newey, Whitney K. 1985. "Generalized Method of Moments Specification Testing." *Journal of Econometrics* 29:229–56.

——— and Kenneth D. West. 1987. "A Simple Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix." *Econometrica* 55:703–8.

Ogaki, Masao. 1993. "Generalized Method of Moments: Econometric Applications," in G. S. Maddala, C. R. Rao, and H. D. Vinod, eds., *Handbook of Statistics*, Vol. 11, *Econometrics*. Amsterdam: North-Holland.

Pearson, Karl. 1894. "Contribution to the Mathematical Theory of Evolution." *Philosophical Transactions of the Royal Society of London*, Series A 185:71–110.

Rao, C. R. 1948. "Large Sample Tests of Statistical Hypotheses Concerning Several Parameters with Application to Problems of Estimation." *Proceedings of the Cambridge Philosophical Society* 44:50–57.

Rothenberg, Thomas J. 1973. *Efficient Estimation with A Priori Information*. New Haven, Conn.: Yale University Press.

Sargent, Thomas J. 1987. *Dynamic Macroeconomic Theory*. Cambridge, Mass.: Harvard University Press.

Sill, Keith. 1992. *Money in the Cash-in-Advance Model: An Empirical Implementation*. Unpublished Ph.D. dissertation, University of Virginia.

Tauchen, George. 1986. "Statistical Properties of Generalized Method-of-Moments Estimators of Structural Parameters Obtained from Financial Market Data." *Journal of Business and Economic Statistics* 4:397–416.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48:817–38.

———. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica* 50:1–25.

———. 1987. "Specification Testing in Dynamic Models," in Truman F. Bewley, ed., *Advances in Econometrics, Fifth World Congress*, Vol. II. Cambridge, England: Cambridge University Press.

Wooldridge, Jeffrey M. 1991a. "On the Application of Robust, Regression-Based Diagnostics to Models of Conditional Means and Conditional Variances." *Journal of Econometrics* 47:5–46.

———. 1991b. "Specification Testing and Quasi-Maximum Likelihood Estimation." *Journal of Econometrics* 48:29–55.