CHAPTER 8 GENERALIZED METHOD OF MOMENTS

Key words: CAPM, GMM, IV Estimation, Model specification test, Moment condition, Moment matching, Optimal estimation, Overidentification, Rational expectations.

Abstract: Many economic theories and hypotheses have implications on and only on a moment condition or a set of moment conditions. A popular method to estimate model parameters contained in the moment condition is the Generalized Method of Moments (GMM). In this chapter, we first provide some economic examples for the moment condition, and define the GMM estimator. We then establish the consistency and asymptotic normality of the GMM estimator. Since the asymptotic variance of a GMM estimator depends on the choice of a weighting matrix, we introduce an asymptotically optimal two-stage GMM estimator with a suitable choice of a weighting matrix. With the construction of a consistent asymptotic variance estimator, we then propose an asymptotically χ^2 Wald test statistic for the hypothesis of interest, and a model specification test for the moment condition.

8.1 Introduction to the Method of Moments Estimation (MME)

To motivate the generalized method of moments (GMM) estimation, we first consider a traditional method in statistics which is called the method of moments estimation (MME).

MME Procedure: Suppose $f(y, \beta^o)$ is the probability density function (pdf) or the probability mass function (pmf) of a univariate random variable Y_t.

Question: How to estimate the unknown parameter β^o using a realization of the random sample $\{Y_t\}_{t=1}^n$?

The basic idea of MME is to match the sample moments with the population moments obtained under the probability distributional model. Specifically, MME can be implemented as follows:

Step 1: Compute population moments $\mu_k(\beta^o) \equiv E(Y_t^k)$ under the model density $f(y, \beta^o)$.

For example, for $k = 1, 2$, we have

$$\begin{aligned} E(Y_t) &= \int_{-\infty}^{\infty} y f(y, \beta^o) dy = \mu_1(\beta^o) \\ E(Y_t^2) &= \int_{-\infty}^{\infty} y^2 f(y, \beta^o) dy \\ &= \sigma^2(\beta^o) + \mu_1^2(\beta^o), \end{aligned}$$

where $\sigma^2(\beta^o)$ is the variance of Y_t.

Step 2: Compute the sample moments from the random sample $Y^n = (Y_1, ..., Y_n)'$:

For example, for $k = 1, 2$, we have

$$\begin{aligned} \hat{m}_1 &= \bar{Y}_n \to^p \mu(\beta^o) \\ \hat{m}_2 &= n^{-1} \sum_{t=1}^{n} Y_t^2 \\ &\to^p E(Y_t^2) = \sigma^2(\beta^o) + \mu_1^2(\beta^o), \end{aligned}$$

where $\sigma^2(\beta^o) = \mu_2(\beta^o) - \mu_1^2(\beta^o)$, and the weak convergence follows by the WLLN.

Step 3 Match the sample moments with the corresponding population moments evaluated at some parameter value $\hat{\beta}$:

For example, for $k = 1, 2$, we set

$$\begin{aligned} \hat{m}_1 &= \mu(\hat{\beta}), \\ \hat{m}_2 &= \sigma^2(\hat{\beta}) + \mu^2(\hat{\beta}). \end{aligned}$$

Step 4: Solve for the system of equations. The solution $\hat{\beta}$ is called the method of moment estimator for β^o.

Remark: In general, if β is a $K \times 1$ parameter vector, we need K equations of matching moments.

Question: Is MME consistent for β^o?

Answer: Because $\mu_k(\hat{\beta}) = \hat{m}_k \to^p \mu_k(\beta^o)$ by the WLLN, we expect that $\hat{\beta} \to^p \beta^o$ as $n \to \infty$.

We now illustrate MME by two simple examples.

Example 1: Suppose the random sample $\{Y_t\}_{t=1}^n \sim$ i.i.d. EXP(λ). Find an estimator for λ using the method of moment estimation.

Solution: In our application, $\beta = \lambda$. Because the exponential pdf

$$f(y, \lambda) = \lambda e^{-\lambda y} \text{ for } y > 0,$$

it can be shown that

$$\begin{aligned}\mu(\lambda) &= E(Y_t) = \int_0^\infty y f(y,\lambda) dy \\ &= \int_0^\infty y\lambda e^{-\lambda y} dy \\ &= \frac{1}{\lambda}.\end{aligned}$$

On the other hand, the first sample moment is the sample mean:

$$\hat{m}_1 = \bar{Y}_n.$$

Matching the sample mean with the population mean evaluated at $\hat{\lambda}$:

$$\hat{m}_1 = \mu(\hat{\lambda}) = \frac{1}{\hat{\lambda}},$$

we obtain the method of moment estimator

$$\hat{\lambda} = \frac{1}{\hat{m}_1} = \frac{1}{\bar{Y}_n}.$$

Example 2: Suppose the random sample $\{Y_t\}_{t=1}^n \sim$ i.i.d.$N(\mu, \sigma^2)$. Find MME for $\beta^o = (\mu, \sigma^2)'$.

Solution: The first two population moments are

$$\begin{aligned}E(Y_t) &= \mu, \\ E(Y_t^2) &= \sigma^2 + \mu^2.\end{aligned}$$

The first two sample moments are

$$\begin{aligned}\hat{m}_1 &= \bar{Y}_n, \\ \hat{m}_2 &= \frac{1}{n}\sum_{t=1}^n Y_t^2.\end{aligned}$$

Matching the first two moments, we have

$$\begin{aligned}\bar{Y}_n &= \hat{\mu}, \\ \frac{1}{n}\sum_{t=1}^n Y_t^2 &= \hat{\sigma}^2 + \hat{\mu}^2.\end{aligned}$$

It follows that the MME

$$\begin{aligned}\hat{\mu} &= \bar{Y}_n, \\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{t=1}^{n} Y_t^2 - \bar{Y}_n^2 \\ &= \frac{1}{n}\sum_{t=1}^{n}(Y_t - \bar{Y}_n)^2.\end{aligned}$$

It is well-known that $\hat{\mu} \to^p \mu$ and $\hat{\sigma}^2 \to^p \sigma^2$ as $n \to \infty$.

8.2 Generalized Method of Moments

Suppose β is a $K \times 1$ unknown parameter vector, and there exists a $l \times 1$ moment function $m_t(\beta)$ such that

$$E[m_t(\beta^o)] = 0,$$

where sub-index t denotes that $m_t(\beta)$ is a function of both β and some random variables indexed by t. For example, we may have

$$m_t(\beta) = X_t(Y_t - X_t'\beta)$$

in the OLS estimation, or

$$m_t(\beta) = Z_t(Y_t - X_t'\beta)$$

in the 2SLS estimation, or more generally in the instrumental variable (IV) estimation, where Z_t is a $l \times 1$ instrument vector.

If $l = K$, that is, if the number of moment conditions is the same as the number of unknown parameters, the model $E[m_t(\beta^o)] = 0$ is called exactly identified. If $l > K$, that is, if the number of moment conditions is more than the number of unknown parameters, the model is called overidentified.

The moment condition $E[m_t(\beta^o)] = 0$ may follow from economic and financial theory (e.g. rational expectations and correct asset pricing). We now illustrate this by the following example.

Example 1 [Capital Asset Pricing Model (CAPM)]: Define Y_t as an $L \times 1$ vector of excess returns for L assets (or portfolios of assets) in period t. For these L assets, the excess returns can be described using the excess-return market model:

$$\begin{aligned}Y_t &= \beta_0^o + \beta_1^o R_{mt} + \varepsilon_t \\ &= \beta^{o\prime} X_t + \varepsilon_t,\end{aligned}$$

where $X_t = (1, R_{mt})'$ is a bivariate vector, R_{mt} is the excess market portfolio return, β^o is a $2 \times L$ parameter matrix, and ε_t is an $L \times 1$ disturbance, with $E(\varepsilon_t | X_t) = 0$.

Define the $l \times 1$ moment function

$$m_t(\beta) = X_t \otimes (Y_t - \beta' X_t),$$

where $l = 2L$ and $\otimes$ denotes the Kronecker product. When CAPM holds, we have

$$E[m_t(\beta^o)] = 0.$$

These $l \times 1$ moment conditions form a basis to estimate and test the CAPM.

In fact, for any measurable function $h : R^2 \to R^l$, CAPM implies

$$E[h(X_t)(Y_t - \beta' X_t)] = 0.$$

This can also be used to estimate the CAPM model.

Question: How to choose the instruments $h(X_t)$?

Example 2 [Hansen and Singleton (1982, Econometrica) Dynamic Capital Asset Pricing Model]:

Suppose a representative economic agent has a constant relative risk aversion utility over his lifetime

$$U = \sum_{t=0}^{n} \delta^t u(C_t) = \sum_{t=0}^{n} \delta^t \frac{C_t^\gamma - 1}{\gamma},$$

where $u(\cdot)$ is the time-invariant utility function of the economic agent in each time period (here we assume $u(c) = (c^\gamma - 1)/\gamma$), δ is the agent's time discount factor, γ is the economic agent's risk aversion parameter, and C_t is the consumption during period t. Let the information available to the agent at time $t-1$ be represented by the sigma-algebra I_{t-1}– in the sense that any variable whose value is known at time $t-1$ is presumed to be I_{t-1}-measurable, and let

$$R_t = \frac{P_t}{P_{t-1}} = 1 + \frac{P_t - P_{t-1}}{P_{t-1}}$$

be the gross return to an asset acquired at time $t-1$ at the price of P_{t-1} (we assume no dividend on the asset). The agent's optimization problem is to

$$\max_{\{C_t\}} E(U)$$

subject to the intertemporal budget constraint

$$C_t + P_t q_t = Y_t + P_t q_{t-1},$$

where q_t is the quantity of the asset purchased at time t and Y_t is the agent's labor income during period t. Define the marginal rate of intertemporal substitution

$$\text{MRS}_t(\gamma) = \frac{\frac{\partial u(C_t)}{\partial C_t}}{\frac{\partial u(C_{t-1})}{\partial C_{t-1}}} = \left(\frac{C_t}{C_{t-1}}\right)^{\gamma-1}.$$

The first order conditions of the agent optimization problem are characterized by the Euler equation:

$$E\left[\delta^o \text{MRS}_t(\gamma^o) R_t | I_{t-1}\right] = 1 \text{ for some } \beta^o = (\delta^o, \gamma^o)'.$$

That is, the marginal rate of intertemporal substitution discounts gross returns to unity.

Remark: Any dynamic asset pricing model is equivalent to a specification of MRS_t.

We may write the Euler equation as follows:

$$E\left[\{\delta^o \text{MRS}_t(\gamma^o) R_t - 1\} | I_{t-1}\right] = 0.$$

Thus, one may view that $\{\delta \text{MRS}_t(\gamma) R_t - 1\}$ is a generalized model residual which has the MDS property when evaluated at the true structural parameters $\beta^o = (\delta^o, \gamma^o)'$.

Question: How to estimate the unknown parameter β^o in an asset pricing model?

More generally, how to estimate β^o from any linear or nonlinear econometric model which can be formulated as a set of moment conditions? Note that the joint distribution of the random sample is not given or implied by economic theory; only a set of conditional moments is given.

From the Euler equation, we can induce the following conditional moment restrictions:

$$\begin{aligned} E\left(\delta^o \text{MRS}_t(\gamma^o) R_t - 1\right) &= 0, \\ E\left[\frac{C_{t-1}}{C_{t-2}}\left(\delta^o \text{MRS}_t(\gamma^o) R_t - 1\right)\right] &= 0, \\ E\left[R_{t-1}\left(\delta^o \text{MRS}_t(\gamma^o) R_t - 1\right)\right] &= 0. \end{aligned}$$

Therefore, we can consider the 3×1 sample moments

$$\hat{m}(\beta) = \frac{1}{n}\sum_{t=1}^{n} m_t(\beta),$$

where

$$m_t(\beta) = [\delta \text{MRS}_t(\gamma) R_t - 1] \left(1, \frac{C_{t-1}}{C_{t-2}}, R_{t-1}\right)'$$

can serve as the basis for estimation. The elements of the vector

$$Z_t \equiv \left(1, \frac{C_{t-1}}{C_{t-2}}, R_{t-1}\right)'$$

are called instrumental variables which are a subset of information set I_{t-1}.

We now define the GMM estimator.

Definition [GMM Estimator] The generalized method of moments (GMM) estimator is

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta),$$

where

$$\hat{m}(\beta) = n^{-1} \sum_{t=1}^{n} m_t(\beta)$$

is a $l \times 1$ sample moment vector, $\hat{W}$ is a $l \times l$ symmetric nonsingular matrix which is possibly data-dependent, and β is a $K \times 1$ unknown parameter vector, and Θ is a K-dimensional parameter space. Here, we assume $l \geq K$, i.e., the number of moments may be larger than or at least equal to the number of parameters.

Question: Why do we require $l \geq K$ in GMM estimation?

Question: Why is the GMM estimator $\hat{\beta}$ not defined by setting the $l \times 1$ sample moments to zero jointly, namely

$$\hat{m}(\hat{\beta}) = 0?$$

Remark: When $l > K$, i.e., when the number of equations is larger than the number of unknown parameters, we generally cannot find a $\hat{\beta}$ such that $\hat{m}(\hat{\beta}) = 0$. However, we can find a $\hat{\beta}$ which makes $\hat{m}(\hat{\beta})$ as close to a $l \times 1$ zero vector as possible by minimizing the quadratic form

$$\hat{m}(\beta)' \hat{m}(\beta) = \sum_{i=1}^{l} \hat{m}_i^2(\beta),$$

where $\hat{m}_i(\beta) = n^{-1} \sum_{t=1}^{n} m_{it}(\beta)$, $i = 1, ..., l$. Since each sample moment conponent $\hat{m}_i(\beta)$ has a different variance, and $\hat{m}_i(\beta)$ and $\hat{m}_j(\beta)$ may be correlated, we can introduce a weighting matrix $\hat{W}$ and choose $\hat{\beta}$ to minimize a weighted quadratic form in $\hat{m}(\hat{\beta})$, namely

$$\hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta).$$

Question: What is the role of $\hat{W}$?

When $\hat{W} = I$, an identity matrix, each of the l component sample moments is weighted equally. If $\hat{W} \neq I$, then the l sample moment components are weighted differently. A suitable choice of weighting matrix $\hat{W}$ can improve the efficiency of the resulting estimator. Here, a natural question is: what is the optimal weighting function for the choice of $\hat{W}$?

Intuitively, the sample moment components which have large sampling variations should be discounted. This is an idea similar to GLS, which discounts noisy observations by dividing by the conditional standard deviation of the disturbance term and differencing out serial correlations.

Special Case: Linear IV Estimation

Question: Does the GMM estimator have a closed form expression?

In general, when the moment function $m_t(\beta)$ is nonlinear in parameter β, there is no closed form solution for $\hat{\beta}$. However, there is an important special case where the GMM estimator $\hat{\beta}$ has a closed form. This is the case of so-called linear IV estimation where we have

$$m_t(\beta) = Z_t(Y_t - X_t'\beta)$$

and

$$E[Z_t(Y_t - X_t'\beta^o)] = 0 \text{ for some } \beta^o,$$

where Y_t is a scalar, X_t is a $K \times 1$ vector, and Z_t is $l \times 1$ vector, with $l \geq K$.

In this case, the GMM estimator, or more precisely, the linear IV estimator, $\hat{\beta}$, solves the following minimization problem:

$$\min_{\beta \in R^K} \hat{m}(\beta)'\hat{W}^{-1}\hat{m}(\beta) = n^{-2} \min_{\beta \in R^K} (Y - \mathbf{X}\beta)'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'(Y - \mathbf{X}\beta),$$

where

$$\hat{m}(\beta) = \frac{\mathbf{Z}'(Y - \mathbf{X}\beta)}{n} = \frac{1}{n}\sum_{t=1}^{n} Z_t(Y_t - X_t'\beta).$$

The FOC is given by

$$\begin{aligned} & \frac{\partial}{\partial \beta}\left[(Y - \mathbf{X}\beta)'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'(Y - \mathbf{X}\beta)\right]_{\beta=\hat{\beta}} \\ = \; & -2\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'(Y - \mathbf{X}\hat{\beta}) = 0. \end{aligned}$$

It follows that

$$\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'\mathbf{X}\hat{\beta} = \mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'Y.$$

When the $K \times l$ matrix $Q_{xz} = E(X_t Z_t')$ is of full rank of K, the $K \times K$ matrix $Q_{xz} W Q_{zx}$ is nonsingular. Therefore, $\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'\mathbf{X}$ is not singular at least for large samples, and consequently the GMM estimator $\hat{\beta}$ has the closed form expression:

$$\hat{\beta} = (\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'Y.$$

This is called a linear IV estimator because it estimates the parameter β^o in the linear model $Y_t = X_t'\beta^o + \varepsilon_t$ with $E(\varepsilon_t|Z_t) = 0$.

Interestingly, the 2SLS estimator $\hat{\beta}_{2sls}$ considered in Chapter 7 is a special case of the IV estimator by choosing

$$\hat{W} = (\mathbf{Z}'\mathbf{Z})^{-1}.$$

or more generally, by choosing $\hat{W} = c(\mathbf{Z}'\mathbf{Z})^{-1}$ for any constant $c \neq 0$.

Question: Is the choice of $\hat{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ optimal? In other words, is the 2SLS estimator $\hat{\beta}_{2sls}$ asymptotically efficient in estimating β^o?

When $l = K$ such that $Q_{xz} = E(X_t Z_t')$ is nonsingular, the $K \times K$ matrix $\mathbf{X}'\mathbf{Z}$ is nonsingular at least for large samples. Consequently,

$$\hat{\beta} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'Y.$$

Theorem: *Suppose $m_t(\beta) = Z_t(Y_t - X_t'\beta)$, where Y_t is a scalar, Z_t is a $l \times 1$ vector, X_t is $K \times 1$ vector, with $l \geq K$. Also, the $K \times l$ matrix $\mathbf{X}$ is of full rank K and the $l \times l$ weighting matrx $\hat{W}$ is nonsingular. Then the resulting GMM estimator $\hat{\beta}$ is called a linear IV estimator and has the closed form expression*

$$\hat{\beta} = (\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'Y.$$

When $l = K$, and Q_{xz} is nonsignular,

$$\hat{\beta} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'Y.$$

Note that the IV estimator $\hat{\beta}$ generally depends on the choice of instruments Z_t and weighting matrix $\hat{W}$. However, when $l = K$, the exact identification case, the IV estimator $\hat{\beta}$ does not depend on the choice of $\hat{W}$. This is because in this case the FOC that $\mathbf{X}'\mathbf{Z}\hat{W}^{-1}\mathbf{Z}'(Y - \mathbf{X}\hat{\beta}) = 0$ becomes

$$\begin{aligned} \mathbf{Z}'(Y - \mathbf{X}\hat{\beta}) &= 0 \\ (K \times n)(n \times 1) &= K \times 1 \end{aligned}$$

given $\mathbf{X}'\mathbf{Z}$ and $\hat{W}$ are nonsingular at least for large samples. Obviously, the OLS estimator

$\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Y$ is a special case of the linear IV estimator by choosing $Z_t = X_t$.

8.3 Consistency

Question: What are the statistical properties of GMM $\hat{\beta}$?

To investigate the asymptotic properties of the GMM estimator $\hat{\beta}$, we first provide a set of regularity conditions.

Assumptions

Assumption 8.1 [Compactness]: The parameter space Θ is compact (closed and bounded);

Assumption 8.2 [Uniform convergence]: (i) The moment function $m_t(\beta)$ is a measurable function of a random vector indexed by t for each $\beta \in \Theta$, and given each t, $m_t(\beta)$ is continuous in $\beta \in \Theta$; (ii) $\{m_t(\beta)\}$ is a stationary ergodic process; (iii) $\hat{m}(\beta)$ converges uniformly over Θ to $m(\beta) \equiv E[m_t(\beta)]$ in probability in the sense that

$$\sup_{\beta \in \Theta} ||\hat{m}(\beta) - m(\beta)|| \to^p 0,$$

where $||\cdot||$ is an Euclidean norm; (iv) $m(\beta)$ is continuous in $\beta \in \Theta$.

Assumption 8.3 [Identification]: There exists a unique parameter β^o in Θ such that $m(\beta^o) = 0$.

Assumption 8.4 [Weighting Matrix]: $\hat{W} \to^p W$, where W is a nonstochastic $l \times l$ symmetric, finite and nonsingular matrix.

Remarks:

Assumption 8.3 is an identification condition. If the moment condition $m(\beta^o) = 0$ is implied by economic theory, β^o can be viewed as the true model parameter value. Assumptions 8.1 and 8.3 imply that the true model parameter β^o lies inside the compact parameter space Θ. Compactness is sometimes restrictive, but it greatly simplifies our asymptotic analysis and is sometime necessary (as in the case of estimating GARCH models) where some parameters must be restricted to ensure a positive conditional variance estimator.

In many applications, the moment function $m_t(\beta)$ usually has the form

$$m_t(\beta) = h_t \varepsilon_t(\beta)$$

for some weighting function h_t and some error or generalized error term $\varepsilon_t(\beta)$. Assumption 8.2 allows but does not require such a multiplicative form for $m_t(\beta)$. Also, in Assumption 8.2, we

impose a uniform WLLN for $\hat{m}(\beta)$ over Θ. Intuitively, uniform convergence implies that the largest (or worse) deviation between $\hat{m}(\beta)$ and $m(\beta)$ over Θ vanishes to 0 in probability as $n \to \infty$.

Question: How to ensure uniform convergence in probability?

This can be achieved by a suitable uniform weak law of large numbers (UWLLN). For example, when $\{Y_t, X_t'\}'$ is i.i.d., we have the following:

Lemma [Uniform Strong Law of Large Numbers for IID Processes (USLLN)]: *Let $\{Z_t, t = 1, 2, ...\}$ be an IID sequence of random $d \times 1$ vectors, with common cumulative distribution function F.*

Let Θ be a compact subset of R^K, and let $q : R^d \times \Theta \to R$ be a function such that $q(\cdot, \beta)$ is measurable for each $\beta \in \Theta$ and $q(z, \cdot)$ is continuous on Θ for each $z \in R^d$.

Suppose there exists a measurable function $D : R^d \to R^+$ such that $|q(z, \beta)| \leq D(z)$ for all $\beta \in \Theta$ and $z \in S$, where S is the support of Z_t and $E[D(Z_t)] < \infty$.

Then

(i) $Q(\beta) = E[q(Z_t, \beta)]$ is continuous on Θ;

(ii) $\sup_{\beta \in \Theta} |\hat{Q}(\beta) - Q(\beta)| \to 0$ a.s. as $n \to \infty$, where $\hat{Q}(\beta) = n^{-1} \sum_{t=1}^n q(Z_t, \beta)$.

Proof: See Jennrich (1969, Theorem 2).

A USLLN for stationary ergodic processes is following:

Lemma [Uniform Strong Law of Large Numbers for Stationary Ergodic Processes [Ranga Rao (1962)]: *Let (Ω, F, P) be a probability space, and let $T : \Omega \to \Omega$ be a one-to-one measure preserving transformation.*

Let Θ be a compact subset of R^K, and let $q : \Omega \times \Theta \to R$ be a function such that $q(\cdot, \beta)$ is measurable for each $\theta \in \Theta$ and $q(\omega, \cdot)$ is continuous on Θ for each $\omega \in \Omega$.

Suppose there exists a measurable function $D : \Omega \to R^+$ such that $|q(\omega, \beta)| \leq D(\omega)$ for all $\beta \in \Theta$ and $\omega \in \Omega$, and $E(D) = \int D dP < \infty$.

If for each $\beta \in \Theta$, $q_t(\theta) = q(T^t\omega, \beta)$ is ergodic, then

(i) $Q(\beta) = E[q_t(\beta)]$ is continuous on Θ;

(ii) $\sup_{\beta \in \Theta} |\hat{Q}(\beta) - Q(\beta)| \to 0$ a.s. as $n \to \infty$, where $\hat{Q}(\beta) = n^{-1} \sum_{t=1}^n q_t(\beta)$.

Proof: See Ranga Rao (1962).

Remark: Uniform almost sure convergence implies uniform convergence in probability.

We first state the consistency result for the GMM estimator $\hat{\beta}$.

Theorem [Consistency of the GMM Estimator]: *Suppose Assumptions 8.1–8.4 hold. Then* $\hat{\beta} \to^p \beta^o$.

To show this consistency theorem, we need the following extrema estimator lemma.

Lemma [White, 1994, Consistency of Extrema Estimators]: *Let* $\hat{Q}(\beta)$ *be a stochastic real-valued function of* $\beta \in \Theta$, *and* $Q(\beta)$ *be a nonstochastic real-valued continuous function of* β, *where* Θ *is a compact parameter space. Suppose that for each* β, $\hat{Q}(\beta)$ *is a measurable function of the random sample with sample* n, *and for each* n, $\hat{Q}(\cdot)$ *is continuous in* $\beta \in \Theta$ *with probability one. Also suppose* $\hat{Q}(\beta) - Q(\beta) \to^p 0$ *uniformly in* $\beta \in \Theta$.

Let $\hat{\beta} = \arg\max_{\beta \in \Theta} \hat{Q}(\beta)$, *and* $\beta^o = \arg\max_{\beta \in \Theta} Q(\beta)$ *is the unique maximizer. Then* $\hat{\beta} - \beta^o \to^p 0$.

Remark: This lemma continues to hold if we change all convergences in probability to almost sure convergences.

We now show the consistency of the GMM estimator $\hat{\beta}$ by applying the above lemma.

Proof: Put

$$\hat{Q}(\beta) = -\hat{m}(\beta)'\hat{W}^{-1}\hat{m}(\beta)$$

and

$$Q(\beta) = -m(\beta)'W^{-1}m(\beta).$$

Then

$$\begin{aligned}
& \left|\hat{Q}(\beta) - Q(\beta)\right| \\
= & \left|\hat{m}(\beta)'\hat{W}^{-1}\hat{m}(\beta) - m(\beta)'W^{-1}m(\beta)\right| \\
= & \left|[\hat{m}(\beta) - m(\beta) + m(\beta)]'\hat{W}^{-1}[\hat{m}(\beta) - m(\beta) + m(\beta)] - m(\beta)'W^{-1}m(\beta)\right| \\
\leq & \left|[\hat{m}(\beta) - m(\beta)]'\,\hat{W}^{-1}\,[\hat{m}(\beta) - m(\beta)]\right| \\
& +2\left|m(\beta)'\hat{W}^{-1}\,[\hat{m}(\beta) - m(\beta)]\right| \\
& +\left|m(\beta)'(\hat{W}^{-1} - W^{-1})m(\beta)\right|.
\end{aligned}$$

It follows from Assumptions 8.1, 8.2 and 8.4 that

$$\hat{Q}(\beta) \to^p Q(\beta)$$

uniformly over Θ, and $Q(\cdot) = m(\cdot)'W^{-1}m(\cdot)$ is continuous in β over Θ. Moreover, Assumption 8.3 implies that β^o is the unique minimizer of $Q(\beta)$ over Θ. It follows that $\hat{\beta} \to^p \beta^o$ by the

extrema estimator Lemma. Note that the proof of the consistency theorem does not require the existence of a FOC. This is made possible by using the extrema estimator lemma. This completes the proof of consistency.

8.4 Asymptotic Normality of GMM

To derive the asymptotic distribution of the GMM estimator, we impose two additional regularity conditions.

Assumption 8.5 [Interiorness]: $\beta^o \in int(\Theta)$.

Assumption 8.6 [CLT]:

(i) For each t, $m_t(\beta)$ is continuously differentiable with respect to $\beta \in \Theta$ with probability one.

(ii) As $n \to \infty$,

$$\sqrt{n}\hat{m}(\beta^o) \equiv n^{-1/2} \sum_{t=1}^{n} m_t(\beta^o) \xrightarrow{d} N(0, V_o),$$

where $V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$ is finite and p.d.

(iii) $\{\frac{\partial m_t(\beta)}{\partial \beta}\}$ obeys the uniform weak law of large numbers (UWLLN), i.e.,

$$\sup_{\beta \in \Theta} \left\| n^{-1} \sum_{t=1}^{n} \frac{\partial m_t(\beta)}{\partial \beta} - D(\beta) \right\| \to^p 0,$$

where the $l \times K$ matrix

$$\begin{aligned} D(\beta) &\equiv E\left[\frac{\partial m_t(\beta)}{\partial \beta}\right] \\ &= \frac{dm(\beta)}{d\beta} \end{aligned}$$

is continuous in $\beta \in \Theta$ and is of full rank K.

Remarks:

Question: Why do we need to assume that β^o is an interior point in Θ?

This is because we will have to use a Taylor series expansion. We need to make use of the FOC for GMM in order to derive the asymptotic distribution of $\hat{\beta}$.

In Assumption 8.6, we assume both CLT and UWLLN directly. These are called "high-level assumptions." They can be ensured by imposing more primitive conditions on the data generating processes (e.g., i.i.d. random samples or MDS random samples), and the moment and smoothness conditions of $m_t(\beta)$. Fore more discussion, see White (1994).

We now establish the asymptotic normality of the GMM estimator $\hat{\beta}$.

Theorem [Asymptotic Normality]: *Suppose Assumptions 8.1–8.6 hold. Then as* $n \to \infty$,

$$\sqrt{n}\left(\hat{\beta} - \beta^o\right) \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = (D_o' W^{-1} D_o)^{-1} D_o' W^{-1} V_o W^{-1} D_o (D_o' W^{-1} D_o)^{-1},$$

and $D_o \equiv D(\beta^o) = \frac{\partial m(\beta^o)}{\partial \beta}$.

Proof: Because β^o is an interior element in Θ, and $\hat{\beta} \to^p \beta^o$ as $n \to \infty$, we have that $\hat{\beta}$ is an interior element of Θ with probability approaching one as $n \to \infty$.

For n sufficiently large, the first order conditions for the maximization of $\hat{Q}(\beta) = -\hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta)$ are

$$\begin{aligned} 0 &= \frac{d\hat{Q}(\beta)}{d\beta}|_{\beta=\hat{\beta}} \\ &= -2\frac{d\hat{m}(\hat{\beta})}{d\beta'} \hat{W}^{-1} \hat{m}(\hat{\beta}). \\ 0 &= \frac{d\hat{m}(\hat{\beta})}{d\beta'} \hat{W}^{-1} \sqrt{n}\hat{m}(\hat{\beta}). \\ K \times 1 &= (K \times l) \times (l \times l) \times (l \times 1) \end{aligned}$$

Note that $\hat{W}$ is not a function of β. Also, this FOC does not necessarily imply $\hat{m}(\hat{\beta}) = 0$. Instead, it only says that a set (with dimension $K \leq l$) of linear combinations of the l components in $\hat{m}(\hat{\beta})$ is equal to zero. Here, the $l \times K$ matrix $\frac{d\hat{m}(\beta)}{d\beta}$ is the gradient of the $l \times 1$ vector $\hat{m}(\hat{\beta})$ with respect to the $K \times 1$ vector β.

Using the Taylor series expansion around the true parameter value β^o, we have

$$\sqrt{n}\hat{m}(\hat{\beta}) = \sqrt{n}\hat{m}(\beta^o) + \frac{d\hat{m}(\bar{\beta})}{d\beta}\sqrt{n}(\hat{\beta} - \beta^o),$$

where $\bar{\beta} = \lambda\hat{\beta} + (1-\lambda)\beta^o$ lies between $\hat{\beta}$ and β^o, with $\lambda \in [0,1]$. Here, for notational simplicity, we have amused the notation in the expression of$\frac{d\hat{m}(\bar{\beta})}{d\beta}$. Precisely speaking, a different $\bar{\beta}$ is needed for each partial derivative of $\hat{m}(\cdot)$ with respect to each parameter β_i, $i = 1, ..., K$.

The first term in the above Taylor series expansion is contributed by the sampling randomness of the sample average of the moment functions evaluated at the true parameter β^o, and the second term is contributed by the randomness of parameter estimator $\hat{\beta} - \beta^o$.

It follows from FOC that

$$\begin{aligned} 0 &= \frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\sqrt{n}\hat{m}(\hat{\beta}) \\ &= \frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\sqrt{n}\hat{m}(\beta^o) \\ &\quad + \frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\sqrt{n}(\hat{\beta}-\beta^o). \end{aligned}$$

Now let us show that $\frac{d\hat{m}(\hat{\beta})}{d\beta} \to^p D_o \equiv D(\beta^o)$. To show this, consider

$$\begin{aligned} & \left\| \frac{d\hat{m}(\hat{\beta})}{d\beta} - D_0 \right\| \\ = & \left\| \frac{d\hat{m}(\hat{\beta})}{d\beta} - D(\hat{\beta}) + D(\hat{\beta}) - D(\beta^o) \right\| \\ \leq & \left\| \frac{d\hat{m}(\hat{\beta})}{d\beta} - D(\hat{\beta}) \right\| + \left\| D(\hat{\beta}) - D(\beta^o) \right\| \\ \leq & \sup_{\beta\in\Theta} \left\| \frac{d\hat{m}(\beta)}{d\beta} - D(\beta) \right\| + \left\| D(\hat{\beta}) - D(\beta^o) \right\| \\ \to & {}^p 0 \end{aligned}$$

by the triangle inequality and Assumption 8.6 (the UWLLN, the continuity of $D(\beta)$, and $\hat{\beta} - \beta^o \to^p 0$).

Similarly, because $\bar{\beta} = \lambda\hat{\beta} + (1-\lambda)\beta^o$ for $\lambda \in [0,1]$, we have

$$||\bar{\beta} - \beta^o|| = ||\lambda(\hat{\beta} - \beta^o)|| \leq ||\hat{\beta} - \beta^o|| \to^p 0.$$

It follows that

$$\frac{d\hat{m}(\bar{\beta})}{d\beta} \to^p D_o.$$

Then the $K \times K$ matrix

$$D_o' W^{-1} D_o$$

is nonsingular by Assumptions 8.4 and 8.6. Therefore, for n sufficiently large, the inverse

$$\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta} \right]^{-1}$$

exists and it converges in probability to $(D_o' W^{-1} D_o)^{-1}$. Therefore, when n is sufficiently large,

we have

$$\begin{aligned}\sqrt{n}(\hat{\beta}-\beta^o) &= -\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\sqrt{n}\hat{m}(\beta^o)\\ &= \hat{A}\sqrt{n}\hat{m}(\beta^o),\end{aligned}$$

where

$$\hat{A} = -\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}.$$

By Assumption 8.6(ii), the CLT for $\{m_t(\beta^o)\}$, we have

$$\sqrt{n}\hat{m}(\beta^o) \xrightarrow{d} N(0, V_o),$$

where $V_o \equiv \text{avar}[n^{-1/2}\Sigma_{t=1}^n m_t(\beta^o)]$. Moreover,

$$\begin{aligned}\hat{A} &= \left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\hat{W}^{-1}\\ &\to {}^p \left(D_o'W^{-1}D_o\right)^{-1}D_o'W^{-1} \equiv A.\end{aligned}$$

It follows from the Slutsky theorem that

$$\sqrt{n}(\hat{\beta}-\beta^o) \xrightarrow{d} A\cdot N(0,V_o) \sim N(0,\Omega),$$

where

$$\begin{aligned}\Omega &= AV_oA'\\ &= (D_o'W^{-1}D_o)^{-1}D_o'W^{-1}V_oW^{-1}D_o(D_o'W^{-1}D_o)^{-1}.\end{aligned}$$

This completes the proof.

Remarks:

The structure of $\text{avar}(\sqrt{n}\hat{\beta})$ is very similar to that of $\text{avar}(\sqrt{n}\hat{\beta}_{2sls})$. In fact, as pointed out earlier, 2SLS is a special case of the GMM estimator with the choice of

$$\begin{aligned}m_t(\beta) &= Z_t(Y_t - X_t'\beta)\\ W &= E(Z_tZ_t') = Q_{zz}.\end{aligned}$$

Similarly, the OLS estimator is a special case of GMM with the choice of

$$\begin{aligned} m_t(\beta) &= X_t(Y_t - X_t'\beta), \\ W &= E(X_t X_t') = Q_{xx}. \end{aligned}$$

Most econometric estimators can be viewed as a special case of GMM, at least asymptotically. In other words, GMM provides a convenient unified framework to view most econometric estimators. See White (1994) for more discussion.

8.5 Asymptotic Efficiency

Question: There are many possible choices of $\hat{W}$. Is there any optimal choice for $\hat{W}$? If so, what is the optimal choice of $\hat{W}$?

The following theorem shows that the optimal choice of W is given by $W = V_o \equiv \text{var}[\sqrt{n}\hat{m}(\beta^o)]$.

Theorem [Asymptotic Efficiency]: *Suppose Assumptions 8.4 and 8.6 hold. Define* $\Omega_o = (D_o' V_o^{-1} D_o)^{-1}$, *which is obtained from* Ω *by choosing* $W = V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$. *Then*

$$\Omega - \Omega_o \text{ is p.s.d.}$$

for any finite, symmetric and nonsingular matrix W.

Proof: Observe that $\Omega - \Omega_o$ is p.s.d. if and only if $\Omega_o^{-1} - \Omega^{-1}$ is p.s.d. We therefore consider

$$\begin{aligned} & \Omega_o^{-1} - \Omega^{-1} \\ = \; & D_o' V_o^{-1} D_o - D_o' W^{-1} D_o (D_o' W^{-1} V_o W^{-1} D_o)^{-1} D_o' W^{-1} D_o \\ = \; & D_o' V_o^{-1/2} [I - V_o^{1/2} W^{-1} D_o (D_o' W^{-1} V_o W^{-1} D_o)^{-1} D_o' W^{-1} V_o^{1/2}] V_o^{-1/2} D_o \\ = \; & D_o' V_o^{-1/2} G V_o^{-1/2} D_o, \end{aligned}$$

where $V_o = V_o^{1/2} V_o^{1/2}$ for some symmetric and nonsingular matrix $V_o^{1/2}$, and

$$G \equiv I - V_o^{1/2} W^{-1} D_o (D_o' W^{-1} V_o W^{-1} D_o)^{-1} D_o' W^{-1} V_o^{1/2}$$

is a symmetric idempotent matrix (i.e., $G = G'$ and $G^2 = G$). It follows that we have

$$\begin{aligned} \Omega_o^{-1} - \Omega^{-1} &= (D_o' V_o^{-1/2} G)(G V_o^{-1/2} D_o) \\ &= (G V_o^{-1/2} D_o)'(G V_o^{-1/2} D_o) \\ &= B'B \\ &\sim \text{ p.s.d. (why?)}, \end{aligned}$$

where $B = GV_o^{-1/2}D_o$ is a $l \times K$ matrix. This completes the proof.

Remark:

The optimal choice of $W = V_o$ is not unique. The choice of $W = cV_o$ for any nonzero constant c is also optimal.

In practice, the matrix V_o is unavailable. However, we can use a feasible asymptotically optimal choice $\hat{W} = \tilde{V}$, a consistent estimator for $V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$.

Question: What is the intuition that $\hat{W} = \tilde{V}$ is an optimal weighting matrix?

Answer: $\hat{W} \to^p V_o$, and V_o is the variance-covariance matrix of the sample moments $\sqrt{n}\hat{m}(\beta^o)$. The use of $\hat{W}^{-1} \to^p V_o^{-1}$, therefore, downweighs the sample moments which have large sampling variations and differences out correlations between different components$\sqrt{n}\hat{m}_i(\beta^o)$ and $\sqrt{n}\hat{m}_j(\beta^o)$ for $i \neq j$, where $i, j = 1, ..., K$. This is similar in spirit to the GLS estimator in the linear regression model. It also corrects serial correlations between different sample moments when they exist.

Optimality of the 2SLS Estimator $\hat{\beta}_{2sls}$

As pointed out earlier, the 2SLS estimator $\hat{\beta}_{2sls}$ is a special case of the GMM estimator with $m_t(\beta) = Z_t(Y_t - X_t'\beta)$ and the the choice of weighting matrix $W = E(Z_tZ_t') = Q_{zz}$. Suppose $\{m_t(\beta^o)\}$ is an MDS and $E(\varepsilon_t^2|Z_t) = \sigma^2$, where $\varepsilon_t = Y_t - X_t'\beta^o$. Then

$$\begin{aligned} V_o &= \text{avar}[\sqrt{n}\hat{m}(\beta^o)] \\ &= E\left[m_t(\beta^o)m_t(\beta^o)'\right] \\ &= \sigma^2 Q_{zz} \end{aligned}$$

where the last equality follows by the law of iterated expectations and conditional homoskedasticity. Because $W = Q_{zz}$ is proportional to V_o, the 2SLS estimator $\hat{\beta}$ is asymptotically optimal in this case. In contrast, when $\{m_t(\beta^o)\}$ is an MDS with conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|Z_t) \neq \sigma^2$) or $\{m_t(\beta^o)\}$ is not an MDS, then the choice of $W = Q_{zz}$ does not deliver an asymptotically optimal 2SLS estimator. Instead, the GMM estimator with the choice of $W = V_o = E(Z_tZ_t'\varepsilon_t^2)$ is asymptotically optimal.

Two Stage GMM Estimator

The previous theorem suggests that the following two-stage GMM estimator will be asymptotically optimal.

Step 1: Find a consistent preliminary estimator $\tilde{\beta}$:

$$\tilde{\beta} = \arg\min_{\beta\in\Theta} \hat{m}(\beta)'\tilde{W}^{-1}\hat{m}(\beta),$$

for some prespecified $\tilde{W}$ which converges in probability to some finite and p.d. matrix. For convenience, we can set $\tilde{W} = I$, an $l \times l$ identity matrix. This is not an optimal estimator, but it is a consistent estimator for β^o.

Step 2: Find a preliminary consistent estimator $\tilde{V}$ for $V_o \equiv$avar$[\sqrt{n}\hat{m}(\beta^o)]$, and choose $\hat{W} = \tilde{V}$.

The construction of $\tilde{V}$ differs in the following two cases, depending on whether $\{m_t(\beta^o)\}$ is an MDS:

Case (i): $\{m_t(\beta^o)\}$ is an ergodic stationary MDS process. In this case,

$$V_o \equiv \text{ var}[\sqrt{n}\hat{m}(\beta^o)] = E[m_t(\beta^o)m_t(\beta^o)'].$$

The asymptotic variance estimator

$$\tilde{V} = n^{-1}\sum_{t=1}^{n} m_t(\tilde{\beta})m_t(\tilde{\beta})'$$

will be consistent for

$$V_o = E[m_t(\beta^o)m_t(\beta^o)'].$$

Question: How to show this?

Answer: We need to assume that $\{m_t(\beta)m_t(\beta)' - E[m_t(\beta)m_t(\beta)']\}$ satisfies the uniform convergence:

$$\sup_{\beta\in\Theta}\left\| n^{-1}\sum_{t=1}^{n} m_t(\beta)m_t(\beta)' - E[m_t(\beta)m_t(\beta)']\right\| \to^p 0.$$

Also, we need to assume that $E[m_t(\beta)m_t(\beta)']$ is continuous in $\beta \in \Theta$.

Case (ii): $\{m_t(\beta^o)\}$ is not MDS. In this case, a long-run variance estimator for $V_o \equiv \text{var}[\sqrt{n}\hat{m}(\beta^o)]$ is needed:

$$\tilde{V} = \sum_{j=1-n}^{n-1} k(j/p)\tilde{\Gamma}(j),$$

where $k(\cdot)$ is a kernel function, $p = p(n)$ is a smoothing parameter,

$$\tilde{\Gamma}(j) = n^{-1}\sum_{t=j+1}^{n} m_t(\tilde{\beta})m_{t-j}(\tilde{\beta})' \text{ for } j \geq 0,$$

and $\tilde{\Gamma}(j) = \tilde{\Gamma}(-j)'$ if $j < 0$. Under regularity conditions, it can be shown that $\tilde{V}$ is consistent for the long-run variance

$$V_o = \sum_{j=-\infty}^{\infty} \Gamma(j),$$

where $\Gamma(j) = \text{cov}[m_t(\beta^o), m_{t-j}(\beta^o)] = E[m_t(\beta^o)m_{t-j}(\beta^o)']$. See more discussion in Chapter 6.

Question: Why do not we need demean when defining $\tilde{\Gamma}(j)$?

Step 3: Find an asymptotically optimal estimator $\hat{\beta}$:

$$\hat{\beta} = \arg\min_{\beta\in\Theta} \hat{m}(\beta)'\tilde{V}^{-1}\hat{m}(\beta).$$

Remark: The weighting matrix $\tilde{V}$ does not involve the unknown parameter β. It is a given (stochastic) weighting matrix. This two-stage GMM estimator $\hat{\beta}$ is asymptotically optimal because $\tilde{V} \to^p V_o = \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$.

Theorem [Two-Stage Asymptotically Most Efficient GMM]: *Suppose Assumptions 8.1–8.3, 8.5 and 8.6 hold, $\tilde{V} \to^p V$, and $\tilde{W} \to^p W$ for some symmetric finite and positive definite matrix W. Then*

$$\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \Omega_o) \text{ as } n \to \infty,$$

where $\Omega_o = (D_o'V_o^{-1}D_o)^{-1}$.

Question: Why do we need the asymptotically two-stage GMM estimator?

First, most macroeconomic time series data sets are usually short, and second, the use of instruments Z_t is usually inefficient. These factors lead to a large estimation error so it is desirable to have an asymptotically efficient estimator.

Although the two-stage GMM procedure is asymptotically efficient, one may like to iterate the procedure further until the GMM parameter estimates and the values of the minimized objective function converge. This will eliminate any dependence of the GMM estimator on the choice of the initial weighting matrix $\tilde{W}$, and it may improve the finite sample performance of the GMM estimator when the number of parameters is large (e.g., Ferson and Foerster 1994).

8.6 Asymptotic Variance Estimator

To construct confidence interval estimators and conduct hypothesis tests, we need to estimate the asymptotic variance Ω_o of the optimal GMM estimator.

Question: How to estimate $\Omega_o \equiv (D_o'V_o^{-1}D_o)^{-1}$?

We need to estimate both D_o and V_o.

(i) To estimate $D_o = E[\frac{\partial m_t(\beta^o)}{\partial \beta}]$, we can use

$$\hat{D} = \frac{d\hat{m}(\hat{\beta})}{d\beta}.$$

We have shown earlier that

$$\hat{D} \to^p D_o.$$

(ii) To estimate V_o, we need to consider two cases—MDS and non-MDS separately:

Case 1: $\{m_t(\beta^o)\}$ is ergodic stationary MDS. In this case,

$$V_o = E[m_t(\beta^o)m_t(\beta^o)'].$$

A consistent variance estimator is

$$\hat{V} = n^{-1}\sum_{t=1}^{n} m_t(\hat{\beta})m_t(\hat{\beta})'.$$

Assuming the UWLLN for $\{m_t(\beta)m_t(\beta)'\}$, we can show that $\hat{V}$ is consistent for

$$V_o = E[m_t(\beta^o)m_t(\beta^o)'].$$

Case 2: $\{m_t(\beta^o)\}$ is not MDS. In this case,

$$V_0 = \sum_{j=-\infty}^{\infty} \Gamma(j),$$

where $\Gamma(j) = E[m_t(\beta^o)m_{t-j}(\beta^o)']$. A consistent variance estimator is

$$\hat{V} = \sum_{j=1-n}^{n-1} k(j/p)\hat{\Gamma}(j),$$

where $k(\cdot)$ is a kernel function, and

$$\hat{\Gamma}(j) = n^{-1}\sum_{t=j+1}^{n} m_t(\hat{\beta})m_{t-j}(\hat{\beta})' \text{ for } j \geq 0,$$

Under suitable conditions (e.g., Newey and West 1994, Andrews 1991), we can show

$$\hat{V} \to^p V_o$$

but the proof of this is beyond the scope of this course.

To cover both cases, we directly impose the following "high-level assumption":

Assumption 8.7: $\hat{V} - V_o \to^p 0$, where $V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$.

Theorem [Asymptotic Variance Estimator for the Optimal GMM Estimator]: *Suppose Assumptions 8.1–8.7 hold. Then*

$$\hat{\Omega}_o \equiv (\hat{D}'\hat{V}^{-1}\hat{D})^{-1} \to^p \Omega_o \text{ as } n \to \infty.$$

8.7 Hypothesis Testing

We now consider testing the hypothesis of interest

$$\mathbf{H}_0 : R(\beta^o) = r,$$

where $R(\cdot)$ is a $J \times 1$ continuously differentiable vector-valued function, $J \leq K$, and the $J \times K$ matrix $\frac{dR(\beta^o)}{d\beta} = R'(\beta^o)$ is of full rank J. Note that $R(\beta^o) = r$ covers both linear and nonlinear restrictions on model parameters. An example of nonlinear restriction on β^o is $\beta_1^o\beta_2^o = 1$.

Remark: We need $J \leq K$. The number of restrictions is less than the number of parameters. We now allow hypotheses of both linear and nonlinear restrictions on β^o.

Question: How to construct a test statistic for $\mathbf{H}_0$?

The basic idea is to check whether $R(\hat{\beta}) - r$ is close to 0. By the Taylor series expansion and $R(\beta^o) = r$ under $\mathbf{H}_0$, we have

$$\begin{aligned}
\sqrt{n}[R(\hat{\beta}) - r] &= \sqrt{n}[R(\beta^o) - r] \\
&\quad + R'(\bar{\beta})\sqrt{n}(\hat{\beta} - \beta^o) \\
&= R'(\bar{\beta})\sqrt{n}(\hat{\beta} - \beta^o) \\
&\to {}^d R'(\beta^o) \cdot N(0, \Omega_o) \\
&\sim N[0, R'(\beta^o)\Omega_o R'(\beta^o)'].
\end{aligned}$$

where $\bar{\beta}$ lies between $\hat{\beta}$ and β^o, i.e., $\bar{\beta} = \lambda\hat{\beta} + (1-\lambda)\beta^o$ for some $\lambda \in [0, 1]$.

Because $R'(\bar{\beta}) \to^p R'(\beta^o)$ given continuity of $R'(\cdot)$ and $\bar{\beta} - \beta^o \to^p 0$, and

$$\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \Omega_o),$$

we have

$$\sqrt{n}[R(\hat{\beta}) - r] \xrightarrow{d} N[0, R'(\beta^o)\Omega_o R'(\beta^o)'].$$

by the Slutsky theorem. It follows that the quadratic form

$$\sqrt{n}[R(\hat{\beta}) - r]'[R'(\beta^o)\Omega_o R'(\beta^o)']^{-1}\sqrt{n}[R(\hat{\beta}) - r] \xrightarrow{d} \chi^2_J.$$

The Wald test statistic is then

$$W = n[R(\hat{\beta}) - r]'[R'(\hat{\beta})\hat{\Omega}_o R'(\hat{\beta})']^{-1}[R(\hat{\beta}) - r] \xrightarrow{d} \chi^2_J$$

where the convergence in distribution to χ^2_J follows by the Slutsky theorem.

When $J = 1$, we can have an asymptotically N(0,1) test statistic

$$T = \frac{\sqrt{n}[R(\hat{\beta}) - r)]}{\sqrt{R'(\hat{\beta})\hat{\Omega}_o R'(\hat{\beta})'}} \to^d N(0,1) \text{ as } n \to \infty.$$

Theorem [Wald Test Statistic]: *Suppose Assumptions 8.1–8.7 hold. Then under* $\mathbf{H}_0$: $R(\beta^o) = r$, *we have*

$$W = n[R(\hat{\beta}) - r]'[R'(\hat{\beta})\hat{\Omega}_o R'(\hat{\beta})']^{-1}[R(\hat{\beta}) - r] \xrightarrow{d} \chi^2_J.$$

Remark: This can be used for hypothesis testing. This Wald test is built upon an asymptotically optimal GMM estimator. One could also construct a Wald test using a consistent but suboptimal GMM estimator (how?).

8.8 Model Specification Testing

As pointed out earlier, many dynamic economic theories can be formulated as a moment condition or a set of moment conditions. Thus, to test validity of an economic theory, one can check whether the related moment condition holds.

Question: How to test whether the econometric model as characterized by

$$E\left[m_t(\beta^o)\right] = 0 \text{ for some } \beta^o$$

is correctly specified?

Answer: We can check correct model specification by testing whether the above moment condition holds.

Question: How to check if the moment condition

$$E[m_t(\beta^o)] = 0$$

holds?

Answer: Use the sample moment

$$\hat{m}(\hat{\beta}) = n^{-1} \sum_{t=1}^{n} m_t(\hat{\beta})$$

and see if it is significantly different from zero (the value of the population moment evaluated at the true parameter value β^o). For this purpose, we need to know the asymptotic distribution of $\sqrt{n}\hat{m}(\hat{\beta})$.

Consider the test statistic

$$\begin{aligned} \sqrt{n}\hat{m}(\hat{\beta}) &= \sqrt{n}\hat{m}(\beta^o) \\ &\quad + \frac{d\hat{m}(\bar{\beta})}{d\beta}\sqrt{n}(\hat{\beta} - \beta^o) \end{aligned}$$

which follows by a first order Taylor series expansion, and $\bar{\beta}$ lies between $\hat{\beta}$ and β^o. The asymptotic distribution of $\sqrt{n}\hat{m}(\hat{\beta})$ is contributed from two sources.

Recall that the two-stage GMM

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \tilde{V}^{-1} \hat{m}(\beta).$$

The FOC of the two-stage GMM estimation is given by

$$0 = \frac{d}{d\beta}\left[\hat{m}(\hat{\beta})' \tilde{V}^{-1} \hat{m}(\hat{\beta})\right].$$

It is very important to note that $\tilde{V}$ is not a function of β, so it has nothing to do with the differentiation with respect to β. We then have

$$\begin{aligned} 0 &= \frac{d\hat{m}(\hat{\beta})}{d\beta'} \tilde{V}^{-1} \sqrt{n}\hat{m}(\beta^o) \\ &\quad + \frac{d\hat{m}(\hat{\beta})}{d\beta'} \tilde{V}^{-1} \frac{d\hat{m}(\bar{\beta})}{d\beta} \sqrt{n}(\hat{\beta} - \beta^o). \end{aligned}$$

It follows that for n sufficiently large, we have

$$\begin{aligned}
& \sqrt{n}(\hat{\beta} - \beta^o) \\
= & -\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1} \\
& \times \frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1}\sqrt{n}\hat{m}(\beta^o).
\end{aligned}$$

Hence,

$$\begin{aligned}
& \tilde{V}^{-1/2}\sqrt{n}\hat{m}(\hat{\beta}) \\
= & \tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o) \\
& + \tilde{V}^{-1/2}\frac{d\hat{m}(\bar{\beta})}{d\beta}\sqrt{n}(\hat{\beta} - \beta^o) \\
= & \left[I - \tilde{V}^{-1/2}\frac{d\hat{m}(\bar{\beta})}{d\beta}\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1/2}\right]\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o) \\
= & \hat{\Pi}[\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o)].
\end{aligned}$$

By the CLT for $\{m_t(\beta^o)\}$ and the Slutsky theorem, we have

$$\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o) \xrightarrow{d} N(0, I).$$

where I is a $l \times l$ identity matrix. Also, we have

$$\begin{aligned}
\hat{\Pi} &= I - \tilde{V}^{-1/2}\frac{d\hat{m}(\bar{\beta})}{d\beta}\left[\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1}\frac{d\hat{m}(\bar{\beta})}{d\beta}\right]^{-1}\frac{d\hat{m}(\hat{\beta})}{d\beta'}\tilde{V}^{-1/2} \\
&\to {}^{p} I - V_o^{-1/2}D_o(D_o'V_o^{-1}D_o)^{-1}D_o'V_o^{-1/2} \\
&= \Pi,
\end{aligned}$$

where

$$\Pi = I - V_o^{-1/2}D_o(D_o'V_o^{-1}D_o)^{-1}D_o'V_o^{-1/2}$$

is a $l \times l$ symmetric matrix which is also idempotent (i.e., $\Pi^2 = \Pi$) with $\text{tr}(\Pi) = l - K$ (why? Use $\text{tr}(AB) = \text{tr}(BA)$!).

It follows that under correct model specification, we have

$$\begin{aligned} n[\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})] &= [\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o)]'\hat{\Pi}^2[\tilde{V}^{-1/2}\sqrt{n}\hat{m}(\beta^o)] + o_P(1) \\ &\to {}^dG'\Pi G \\ &\sim \chi^2_{l-K} \end{aligned}$$

by the following lemma, where $G \sim N(0, I)$:

Lemma [Quadratic Form in Normal Random variables]: *If $v \sim N(0, I)$ and Π is an $l \times l$ symmetric and idempotent with rank $q \leq l$, then the quadratic form*

$$v'\Pi v \sim \chi^2_q.$$

Remark: The adjustment of degrees of freedom from l to $l - K$ is due to the impact of the asymptotically optimal parameter estimator $\hat{\beta}$.

Theorem [Overidentification Test] *Suppose Assumptions 8.1–8.6 hold, and $\tilde{V} \to^p V_o$ as $n \to \infty$. Then under the null hypothesis that $E[m_t(\beta^o)] = 0$ for some unknown β^o, the test statistic*

$$n \cdot \hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta}) \to^d \chi^2_{l-K}.$$

Remark: This is often called the J-test or the test for overidentification in the GMM literature, because it requires $l > K$. This test can be used to check if the model characterized as $E[m_t(\beta^o)] = 0$ is correctly specified.

It is important to note that the fact that

$$n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta}) \to G'\Pi G$$

where Π is an idempotent matrix is due to the fact that $\hat{\beta}$ is an asymptotically optimal GMM estimator that minimizes the objective function $n\hat{m}(\beta)'\tilde{V}^{-1}\hat{m}(\beta)$. If a suboptimal GMM estimator is used, we would have no above result. Instead, we need to use a different asymptotic variance estimator to replace $\tilde{V}$ and obtain an asymptotically χ^2_l distribution under correct model specification. Because the critical value of χ^2_{l-K} is smaller than that of χ^2_l when $K > 0$, the use of the asymptotically optimal estimator $\hat{\beta}$ leads to an asymptotically more efficient test.

Remark: When $l = K$, the exactly identified case, the moment conditions cannot be tested by the asymptotically optimal GMM $\hat{\beta}$, because $\hat{m}(\hat{\beta})$ will be identically zero, no matter whether $E[m(\beta^o)] = 0$.

Question: Why is the degree of freedom equal to $l - K$?

Answer: The adjustment of degrees of freedom (minus K) is due to the impact of the sampling variation of the asymptotically optimal GMM estimator. In other words, the use of an asymptotically optimal GMM estimator $\hat{\beta}$ instead of β^o renders the degrees of freedom to change from l to $l-K$. Note that if $\hat{\beta}$ is not an asymptotically optimal GMM estimator, the asymptotic distribution of $n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})$ will be changed.

Question: In the J test, why do we use the preliminary weighting matrix $\tilde{V}$, which is evaluated at a preliminary parameter estimator $\tilde{\beta}$? Why not use $\hat{V}$, a consistent estimator for V that is evaluated at the asymptotically optimal estimator $\hat{\beta}$?

Answer: With the preliminary matrix $\tilde{V}$, the J-test statistic is n times the minimum value of the objective function—the quadratic form in the second stage of GMM estimation. Thus, the value of the test statistic $n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})$ is directly available as a by-product of the second stage GMM estimation. For this reason and for its asymptotic χ^2 distribution, the J-test is also called the minimum chi-square test.

Question: Can we use $\hat{V}$ to replace $\tilde{V}$ in the J-test statistic?

Answer: Yes. The test statistic $n\hat{m}(\hat{\beta})'\hat{V}^{-1}\hat{m}(\hat{\beta})$ is also asymptotically χ^2_{l-K} under correct model specification (please verify!), but this statistic is less convenient to compute than $n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})$, because the latter is the objective function of the second stage GMM estimation. This is analogous to the F-test statistic, which is based on the sums of squared residuals of linear regression models.

Question: Can we replace $\hat{\beta}$ by some suboptimal but consistent GMM estimator $\tilde{\beta}$, say?

Answer: No. We cannot obtain the asymptotically χ^2_{l-K} distribution. We need to replace $\tilde{V}$ in the $n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta})$ with a suitable asymptotic variance estimator and will obtain an asymptotic χ^2_l distribution. Note that $\text{avar}(\sqrt{n}\hat{\beta}) \neq \text{avar}(\sqrt{n}\tilde{\beta})$ if $\tilde{\beta}$ is a consistent but suboptimal estimator for β^o.

Testing for Validity of Instruments

In the linear IV estimation context, where

$$m_t(\beta) = Z_t(Y_t - X_t'\beta),$$

the overidentification test can be used to check the validity of the moment condition

$$\begin{aligned} E[m_t(\beta^o)] &= E[Z_t(Y_t - X_t'\beta^o)] \\ &= 0 \text{ for some } \beta^o. \end{aligned}$$

This is essentially to check whether Z_t is a valid instrument vector, that is, whether Z_t is orthogonal to $\varepsilon_t = Y_t - X_t'\beta^o$. Put $\hat{e}_t = Y_t - X_t'\hat{\beta}_{2sls}$. We can use the following test statistic

$$\frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{e}'\hat{e}/n}$$

Note that the numerator

$$\hat{e}'Z(Z'Z)^{-1}Z'\hat{e} = n \cdot \hat{m}(\hat{\beta}_{2sls})'\hat{W}^{-1}\hat{m}(\hat{\beta}_{2sls})$$

is n times the value of the objective function of the GMM minimization with the choice of $\hat{W} = (Z'Z/n)$, which is an optimal choice when $\{m_t(\beta^o)\}$ is an MDS with conditional homoskedasticity (i.e., $E(\varepsilon_t^2|Z_t) = \sigma^2$). In this case,

$$\frac{\hat{e}'\hat{e}}{n}\frac{Z'Z}{n} \to^p \sigma^2 Q_{zz} = V_o.$$

It follows that the test statistic

$$\frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{e}'\hat{e}/n} \to^d \chi^2_{l-K}$$

under the null hypothesis that $E(\varepsilon_t|Z_t) = 0$ for some β^o.

Corollary: *Suppose Assumptions 7.1–7.4, 7.6 and 7.7 hold, and $l > K$. Then under the null hypothesis that $E(\varepsilon_t|Z_t) = 0$, the test statistic*

$$\frac{\hat{e}'Z(Z'Z)^{-1}Z'\hat{e}}{\hat{e}'\hat{e}/n} \to^d \chi^2_{l-K},$$

where $\hat{e} = Y - X\hat{\beta}_{2sls}$.

In fact, the overidentification test statistic is equal to nR^2_{uc}, where R^2_{uc} is the uncentered R^2 from the auxiliary regression

$$\hat{e}_t = \alpha' Z_t + w_t.$$

In fact, it can be shown that under the null hypothesis of $E(\varepsilon_t|Z_t) = 0$, nR^2_{uc} is asymptotically equivalent to nR^2 in the sense that $nR^2_{uc} = nR^2 + o_P(1)$, where R^2 is the uncentered R^2 of regressing $\hat{e}_t$ on Z_t.This provides a convenient way to calculate the test statistic. However, it is important to emphasize that this convenient procedure is asymptotically valid only when $E(\varepsilon_t^2|Z_t) = \sigma^2$.

8.9 Empirical Applications

8.10 Summary of GMM

Most economic and financial theories have implications on and only on a moment restriction

$$E[m_t(\beta^o)] = 0,$$

where $m_t(\beta)$ is a $l \times 1$ moment function. This moment condition can be used to estimate model parameter β^o via the so-called GMM estimation method. The GMM estimator is defined as:

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta),$$

where

$$\hat{m}(\beta) = n^{-1} \sum_{t=1}^{n} m_t(\beta).$$

Under a set of regularity conditions, it can be shown that

$$\hat{\beta} \ \rightarrow^p \ \beta^o$$

and

$$\sqrt{n}(\hat{\beta} - \beta^o) \xrightarrow{d} N(0, \Omega),$$

where

$$\Omega = (D_o' W^{-1} D_o)^{-1} D_o' W^{-1} V_o W^{-1} D_o (D_o' W^{-1} D_o)^{-1}.$$

The asymptotic variance Ω of the GMM estimator $\hat{\beta}$ depends on the choice of weighting matrix W. An asymptotically most efficient GMM estimator is to choose $W = V_o \equiv \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$. In this case, the asymptotic variance of the GMM estimator is given by

$$\Omega_o = (D_o' V_o^{-1} D_o)^{-1}$$

which is a minimum variance. This is similar in spirit to the GLS estimator in a linear regression model. This suggests a two-stage asymptotically optimal GMM estimator $\hat{\beta}$: First, one can obtain a consistent but suboptimal GMM estimator $\tilde{\beta}$ by choosing some convenient weighting matrix $\tilde{W}$. Then one can use $\tilde{\beta}$ to construct a consistent estimator $\tilde{V}$ for V_o, and use it as a weighting matrix to obtain the second stage GMM estimator $\hat{\beta}$.

To construct confidence interval estimators and hypothesis tests, one has to obtain consistent asymptotic variance estimators for GMM estimators. A consistent asymptotic variance estimator for an asymptotically optimal GMM estimator is

$$\hat{\Omega}_o = (\hat{D}' \hat{V}^{-1} \hat{D})^{-1},$$

where

$$\hat{D} = n^{-1} \sum_{t=1}^{n} \frac{dm_t(\hat{\beta})}{d\beta},$$

and the construction of $\hat{V}$ depends on the properties of $\{m_t(\beta^o)\}$, particularly on whether $\{m_t(\beta^o)\}$ is an ergodic stationary MDS process.

Suppose a two-stage asymptotically optimal GMM estimator is used. Then the associated Wald test statistic for the null hypothesis

$$H_0 : R(\beta^o) = r.$$

is given by

$$\hat{W} = n[R(\hat{\beta}) - r]'[R'(\hat{\beta})(\hat{D}'\hat{V}^{-1}\hat{D})^{-1}R'(\hat{\beta})']^{-1}[R(\hat{\beta}) - r] \xrightarrow{d} \chi^2_J \quad .$$

The moment condition $E[m_t(\beta^o)] = 0$ also provides a basis to check whether an economic theory or economic model is correctly specified. This can be done by checking whether the sample moment $\hat{m}(\hat{\beta})$ is close to zero. A popular model specification test in the GMM framework is the J-test statistic

$$n\hat{m}(\hat{\beta})'\tilde{V}^{-1}\hat{m}(\hat{\beta}) \xrightarrow{d} \chi^2_{l-K}$$

under correct model specification, where $\hat{\beta}$ is an asymptotically optimal GMM estimator (question: what will happen if a consistent but suboptimal GMM estimator is used). This is also called the overidentification test. The J-test statistic $n\hat{m}(\hat{\beta})\tilde{V}^{-1}\hat{m}(\hat{\beta})$ is rather convenient to compute, because it is the objective function of the GMM estimator.

GMM provides a convenient unified framework to view most econometric estimators. In other words, most econometric estimators can be viewed as a special case of the GMM framework with suitable choice of moment function and weighting matrix. In particular, the OLS and 2SLS estimators are special cases of the class of GMM estimators.

References

White, H. (1994)

EXERCISES

8.1. A generalized method of moment (GMM) estimator is defined as

$$\hat{\beta} = \arg\min_{\beta \in \Theta} \hat{m}(\beta)' \hat{W}^{-1} \hat{m}(\beta),$$

where β is a $K \times 1$ vector, $\hat{W}$ is a possibly stochastic $l \times l$ symmetric and nonsingular matrix,

$$\hat{m}(\beta) = n^{-1} \sum_{t=1}^{n} m_t(\beta),$$

and $m_t(\beta)$ is a $l \times 1$ moment function of random vector Z_t, and $l \geq K$. We make the following assumptions:

Assumption 1.1: β^o is the unique solution to $E[m(Z_t, \beta^o)] = 0$, and β^o is an interior point in Θ.

Assumption 1.2: $\{Z_t\}$ is a stationary time series process and $m(Z_t, \beta^o)$ is a martingale difference sequence in the sense that

$$E\left[m(Z_t, \beta^o) | \, Z^{t-1} \right] = 0,$$

where $Z^{t-1} = \{Z_{t-1}, Z_{t-2}, ..., Z_1\}$ is the information available at time $t-1$..

Assumption 1.3: $m(Z_t, \beta)$ is continuously differentiable with respect to $\beta \in \Theta$ such that

$$\sup_{\beta \in \Theta} \|\hat{m}'(\beta) - m'(\beta)\| \to^p 0,$$

where $\hat{m}'(\beta) = \frac{d}{d\beta} \hat{m}(\beta)$ and $m'(\beta) = \frac{d}{d\beta} E[m(Z_t, \beta)] = E[\frac{\partial}{\partial \beta} m(Z_t, \beta)]$.

Assumption 1.4: $\sqrt{n}\hat{m}(\beta^o) \to^d N(0, V_o)$ for some finite and positive definite matrix V_o.

Assumption 1.5: $\hat{W} \to^p W$, where W is a finite and positive definite matrix.

From these assumptions, one can show that $\hat{\beta} \to^p \beta^o$, and this result can be used in answering the following questions in parts (a)–(d). Moreover, you can make additional assumptions if you feel appropriate and necessary.

(a) Find the expression of V_o in terms of $m(Z_t, \beta^o)$.
(b) Find the first order condition of the above GMM minimization problem.
(c) Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta^o)$.

(d) Find the optimal choice of $\hat{W}$. Explain why your choice of $\hat{W}$ is optimal.

8.2. (a) Show that the 2SLS $\hat{\beta}_{2sls}$ for the parameter β^o in the regression model $Y_t = X_t'\beta^o + \varepsilon_t$ is a special case of the GMM estimator with suitable choices of moment function $m_t(\beta)$ and weighting matrix $\hat{W}$;

(b) Assume that $\{Z_t\varepsilon_t\}$ is a stationary ergodic process and other regularity conditions hold. Compare the relative efficiency between an asymptotically optimal GMM estimator (with the optimal choice of the weighting matrix) and $\hat{\beta}_{2sls}$ under conditional homoskedasticity and conditional heteroskedasticity respectively.

8.3. Use a suboptimal GMM estimator $\hat{\beta}$ with a given weighting function $\hat{W} \to^p W$ to construct a Wald test statistic for the null hypothesis $\mathbf{H}_0 : R\beta^o = r$, and justify your reasoning. Assume all necessary regularity conditions hold.

8.4. Suppose that $\{m_t(\beta)\}$ is an ergodic stationary MDS process, where $m_t(\cdot)$ is continuous on a compact parameter set Θ, and $\{m_t(\beta)m_t(\beta)'\}$ follows a uniform weak law of large numbers, and $V_o = E[m_t(\beta^o)m_t(\beta^o)']$ is finite and nonsingular. Let $\hat{V} = n^{-1}\sum_{t=1}^n m_t(\hat{\beta})m_t(\hat{\beta})'$, where $\hat{\beta}$ is a consistent estimator of β^o. Show $\hat{V} \to^p V_o$.

8.5. Suppose $\hat{V}$ is a consistent estimator for $V_o = \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$. Show that replacing $\tilde{V}$ by $\hat{V}$ has no impact on the asymptotic distribution of the overidentification test statistic, that is, show

$$n\hat{m}(\hat{\beta})\tilde{V}^{-1}\hat{m}(\hat{\beta}) - n\hat{m}(\hat{\beta})\hat{V}^{-1}\hat{m}(\hat{\beta}) \to^p 0.$$

Assume all necessary regularity conditions hold.

8.6. Suppose $\tilde{\beta}$ is a suboptimal but consistent GMM estimator. Could we simply replace $\hat{\beta}$ by $\tilde{\beta}$ and still obtain the asymptotic χ^2_{l-K} distribution for the overidentification test statistic? Give your reasoning. Assume all necessary regularity conditions hold.

8.7. Suppose Assumptions 7.1–7.4, 7.6 and 7.7 hold. To test the null hypothesis that $E(\varepsilon_t|Z_t) = 0$, where Z_t is a $l \times 1$ instrumental vector, one can consider the auxiliary regression

$$\hat{e}_t = \alpha' Z_t + w_t,$$

where $\hat{e}_t = Y_t - X_t'\hat{\beta}_{2sls}$. Show $nR^2_{uc} = nR^2 + o_P(1)$ as $n \to \infty$ under the null hypothesis. [*Hint: Recall the definitions of* R^2_{uc} *and* R^2 *in Chapter 3.*]

8.8 [Nonlinear Least Squares Estimation]. Consider a nonlinear regression model

$$Y_t = g(X_t, \beta^o) + \varepsilon_t,$$

where β^o is an unknown $K \times 1$ parameter vector and $E(\varepsilon_t|X_t) = 0$ a.s. Assume that $g(X_t, \cdot)$ is twice continuously differentiable with respect to β with the $K \times K$ matrices $E[\frac{\partial g(X_t,\beta)}{\partial \beta}\frac{\partial g(X_t,\beta)}{\partial \beta'}]$ and $E[\frac{\partial^2 g(X_t,\beta)}{\partial \beta \partial \beta'}]$ finite and nonsingular for all $\theta \in \Theta$.

The nonlinear least squares (NLS) estimator solves the minimization of the sum of squared residual problem

$$\hat{\beta} = \arg\min_{\beta} \sum_{t=1}^{n} [Y_t - g(X_t, \beta)]^2 .$$

The first order condition is

$$D(\hat{\beta})'e = \sum_{t=1}^{n} \frac{\partial g(X_t, \hat{\beta})}{\partial \beta}[Y_t - g(X_t, \hat{\beta})] = 0,$$

where $D(\beta)$ is a $n \times K$ matrix, with the t-th row being $\frac{\partial}{\partial \beta} g(X_t, \beta)$. This FOC can be viewed as the FOC

$$\hat{m}(\hat{\beta}) = 0$$

for an GMM estimation with

$$m_t(\beta) = \frac{\partial g(X_t, \beta)}{\partial \beta}[Y_t - g(X_t, \beta)]$$

in an exact identification case ($l = K$). Generally, there exists no closed form expression for $\hat{\beta}$. Assume all necessary regularities conditions hold.

(a) Show that $\hat{\beta} \to^p \beta^o$ as $n \to \infty$.

(b) Derive the asymptotic distribution of $\sqrt{n}(\hat{\beta} - \beta^o)$.

(c) What is the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ if $\{\frac{\partial g(X_t,\beta)}{\partial \beta}\varepsilon_t\}$ is an MDS with conditional homoskedasticity (i.e., $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s.)? Give your reasoning.

(d) What is the asymptotic variance of $\sqrt{n}(\hat{\beta} - \beta^o)$ if $\{\frac{\partial g(X_t,\beta)}{\partial \beta}\varepsilon_t\}$ is an MDS with conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|X_t) \neq \sigma^2$ a.s.)? Give your reasoning.

(e) Suppose $\{\frac{\partial g(X_t,\beta)}{\partial \beta}\varepsilon_t\}$ is an MDS with conditional homoskedasticity (i.e., $E(\varepsilon_t^2|X_t) = \sigma^2$ a.s.). Construct a test for the null hypothesis $\mathbf{H}_0 : R(\beta^o) = r$, where $R(\beta)$ is a $J \times K$ nonstochastic matrix such that $R'(\beta^o) = \frac{\partial}{\partial \beta} R(\beta^o)$ is a $J \times L$ matrix with full rank $J \leq L$, and r is a $J \times 1$ nonstochastic vector.

8.9. [Nonlinear IV Estimation] Consider a nonlinear regression model

$$Y_t = g(X_t, \beta^o) + \varepsilon_t,$$

where $g(X_t, \cdot)$ is twice continuously differentiable with respect to β, $E(\varepsilon_t|X_t) \neq 0$ but $E(\varepsilon_t|Z_t) = 0$, where Y_t is a scalar, X_t is a $K \times 1$ vector and Z_t is a $l \times 1$ vector with $l \geq K$.

Suppose $\{Y_t, X_t, Z_t\}$ is a stationary ergodic process, and $\{Z_t\varepsilon_t\}$ is an MDS.

The unknown parameter β^o can be consistently estimated based on the moment condition

$$E[m_t(\beta^o)] = 0,$$

where $m_t(\beta) = Z_t[Y_t - g(X_t, \beta^o)]$. Suppose a nonlinear IV estimator solves the minimization problem

$$\hat{\beta} = \arg\min_{\beta} \hat{m}(\beta)'\hat{W}^{-1}\hat{m}(\beta),$$

where $\hat{m}(\beta) = n^{-1}\sum_{t=1}^{n} Z_t[Y_t - g(X_t, \beta)]$, and $\hat{W} \to^p W$, a finite and positive definite matrix.

(a) Show $\hat{\beta} \to^p \beta^o$.

(b) Derive FOC.

(c) Derive the asymptotic distribution of $\hat{\beta}$. Discuss the cases of conditional homoskedasticity and conditional heteroskedasticity respectively.

(d) What is the optimal choice of W so that $\hat{\beta}$ is asymptotically most efficient?

(e) Construct a test for the null hypothesis that $\mathbf{H}_0 : R(\beta^o) = r$, where $R(\beta)$ is a $J \times K$ nonstochastic matrix with $R'(\beta^o)$ of full rank, r is a $J \times 1$ nonstochastic vector, and $J \leq K$.

(f) Suppose $\{\frac{\partial g(X_t,\beta)}{\partial \beta}\varepsilon_t\}$ is an MDS with conditional heteroskedasticity (i.e., $E(\varepsilon_t^2|X_t) \neq \sigma^2$ a.s.). Construct a test for the null hypothesis $\mathbf{H}_0 : R(\beta^o) = r$, where $R(\beta)$ is a $J \times K$ nonstochastic matrix such that $R'(\beta^o) = \frac{\partial}{\partial \beta}R(\beta^o)$ is a $J \times L$ matrix with full rank $J \leq L$, and r is a $J \times 1$ nonstochastic vector.

8.10. Consider testing the hypothesis of interest $H_0 : R(\beta^o) = r$ under the GMM framework, where $R(\beta^o)$ is a $J \times K$ nonstochastic matrix, r is a $J \times 1$ nonstochastic vector, and $R'(\beta^o)$ is a $J \times K$ matrix with full rank J, where $J \leq K$. We can construct a Lagrangian multiplier test based on the Lagrangian multiplier $\hat{\lambda}^*$, where $\hat{\lambda}^*$ is the optimal solution of the following constrained GMM minimization problem:

$$(\hat{\beta}^*, \hat{\lambda}^*) = \arg\min_{\beta\in\Theta, \lambda\in R} \left[\hat{m}(\beta)'\tilde{V}^{-1}\hat{m}(\beta) + \lambda'[r - R(\beta)\right],$$

where $\tilde{V}$ is a preliminary consistent estimator for $V_o = \text{avar}[\sqrt{n}\hat{m}(\beta^o)]$ that does not depend β. Construct the LM test statistic and derive its asymptotic distribution. Assume all regularity conditions hold.