Chapter 23: Generalized Method of Moments and Extremum Estimators

23.1 Generalized Method Moments

We suppose that there is i.i.d. vector data $\{Z_i\}_{i=1}^n$ from some population. It is known that there exists a unique $\theta_0 \in \mathbb{R}^p$ such that

$$\mathbb{E}[g(Z_i, \theta_0)] = 0 \qquad (23.1)$$

for some vector of known functions $g(\cdot)$ $[q \times 1]$. For example, the first order condition from some optimization problem for the representative agent, see below. This is a **semiparametric model**, because the distribution of Z_i is unspecified apart from the q moments, but we are only interested in the parameters θ. There are several cases:

1. $p > q$: unidentified case
2. $p = q$: exactly identified case
3. $p < q$: overidentified case.

We next give some examples:

Example 23.1. Simultaneous Equations Model.

Suppose that we observe $y_i \in \mathbb{R}^L$ and $x_i \in \mathbb{R}^K$, where

$$B(\theta) y_i = C(\theta) x_i + u_i,$$

where $B(\theta)$ is an $L \times L$ matrix and $C(\theta)$ is an $L \times K$ matrix of unknown quantities depending on unknown parameters $\theta \in \mathbb{R}^p$, while the error term $u_i \in \mathbb{R}^L$ satisfies the conditional moment restriction

$$\mathbb{E}(u_i | x_i) = 0. \qquad (23.2)$$

The parameters of interest are B, C, which are not themselves identified unless the parameter vector θ encodes some restrictions (such as $B_{ij} = 0$). Notice that $\mathbb{E}(y_i | x_i) = B(\theta)^{-1} C(\theta) x_i = \Phi(\theta) x_i$ provided B is invertible, where $\Phi(\theta)$ is an $L \times K$ matrix, therefore, the parameters in Φ are identified provided $\mathbb{E}(x_i x_i^\top)$ is of full rank. However, it is the parameters in B and C that are fundamental to the economic interpretation and are therefore the quantities of interest. We will discuss the identification issue later. Let

$$g(Z_i, \theta) = (B(\theta) y_i - C(\theta) x_i) \otimes h(x_i),$$

where $h(x_i)$ is an $M \times 1$ vector of functions of x_i; here, for vectors $a \in \mathbb{R}^L$ and $b \in \mathbb{R}^M$, $a \otimes b$ means the $L \times M$ by 1 vector $(a_1 b^\top, \ldots, a_L b^\top)^\top$ that contains all the cross products. This falls into the framework (23.1) with $Z_i = (y_i^\top, x_i^\top)^\top$ and $q = L \times M$. The traditional approach here has been to assume the stronger condition that

$$u_i \sim N(0, \Sigma(\theta))$$

for some unknown covariance matrix Σ, in which case, $y_i | x_i \sim N(B(\theta)^{-1} C(\theta) x_i, B(\theta)^{-1} \Sigma(\theta) B(\theta)^{\top -1})$.

Example 23.2. (Hansen and Singleton, 1982)

One of the most influential econometric papers of the 1980s. Intertemporal consumption/Investment decision: c_i consumption $u(\cdot)$ utility $u_c > 0$, $u_{cc} < 0$, $1 + r_{j,i+1}$, $j = 1, \ldots, m$ is gross return on asset j at time $i+1$. The representative agent solves the following optimization problem

$$\max_{\{c_i,w_i\}_{i=0}^{\infty}} \sum_{t=0}^{\infty} \beta^t \mathbb{E}[u(c_{i+t})|I_i],$$

where w_i is a vector of portfolio weights and β is the discount factor. This is a dynamic programming problem. We assume that there is a unique interior solution; this is characterized by the following condition

$$u'(c_i) = \beta \mathbb{E}[(1 + r_{j,i+1})u'(c_{i+1})|I_i], \quad j = 1, \ldots, m.$$

Now suppose that

$$u(c_i) = \begin{cases} \dfrac{c_i^{1-\gamma}}{1-\gamma} & \text{if } \gamma > 0, \gamma \neq 1, \\ \log c_i & \text{if } \gamma = 1. \end{cases}$$

Here, γ is the coefficient of relative risk aversion. Then

$$c_i^{-\gamma} = \beta \mathbb{E}[(1 + r_{j,i+1})c_{i+1}^{-\gamma}|I_i].$$

Rearranging we get

$$\mathbb{E}\left[\left\{1 - \beta(1 + r_{j,i+1})\left(\frac{c_{i+1}}{c_i}\right)^{-\gamma}\right\}\middle| I_i^*\right] = 0, \ j = 1, \ldots, m$$

where $I_i^* \subset I_i$ and I_i^* is the econometrician's information set. We want to estimate the parameters and test the theory given a dataset consisting of $c_i, r_{j,i+1}, I_i^*$. Let $\theta_{px1} = (\beta, \gamma)'$ and define:

$$g(Z_i, \theta) = \begin{bmatrix} \vdots \\ \left\{1 - \beta(1 + r_{j,i+1})\left(\frac{c_{i+1}}{c_i}\right)^{-\gamma}\right\} v_i \\ \vdots \end{bmatrix}_{q \times 1}$$

where $v_i \in I_i^*$ and $Z_i = (v_i, c_i, c_{i+1}, r_{1,i+1}, \ldots, r_{m,i+1})'$.

For any $\theta \in \Theta \subseteq \mathbb{R}^p$, let

$$G_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(Z_i, \theta) \in \mathbb{R}^q.$$

Here, Θ is the parameter space or the set of allowable parameters. In the exactly identified case where $q = p$, we can hopefully solve the p equations in p-unknowns: $G_n(\theta) = 0$ exactly, possibly using some numerical methods. However, in the overidentified case when $p < q$, this will not be possible because we cannot simultaneously zero q functions with p controls. Define

$$Q_n(\theta) = G_n(\theta)' W_n(\theta) G_n(\theta) = ||G_n(\theta)||^2_{W_n}$$

where $W_n(\theta)$ is a $q \times q$ positive definite weighting matrix, and $||x||^2_A = x'Ax$. For example, $W_n(\theta) = I_{q \times q}$. Then let $\hat{\theta}_{GMM}$ minimize $Q_n(\theta)$ over $\theta \in \Theta \subseteq \mathbb{R}^p$. This defines a large class of estimators, one for each weighting matrix W_n. It is generally a nonlinear optimization problem like maximum likelihood; various techniques are available for finding the minimizer.

An alternative approach to using overidentifying information is to combine the estimating equations, that is let

$$G_n^*(\theta) = A_n(\theta) G_n(\theta) \in \mathbb{R}^p, \qquad (23.3)$$

where $A_n(\theta)$ is a full rank deterministic $p \times q$ matrix. Then find the value of $\theta \in \Theta \subseteq \mathbb{R}^p$ that solves $G_n^*(\theta) = 0$. The two approaches are broadly equivalent: for a given choice of W_n there is an equivalent choice of A_n and vice versa.

23.2 Asymptotic Properties of Extremum Estimators

We consider the asymptotic properties of a general class of estimators $\hat{\theta}$ that minimize

$$Q_n(\theta) \text{ over } \theta \in \Theta \subseteq \mathbb{R}^p \qquad (23.4)$$

for some general objective function $Q_n(\theta)$ that depends on the data. This includes GMM and Maximum Likelihood as special cases. The difficult part is consistency because in general the objective function is nonlinear in the parameters and does not have a closed form solution.

23.2.1 Consistency

There are many treatments of the asymptotic properties of extremum estimators, we give a standard version that rules out discontinuous criterion functions; there are versions that allow for discontinuity in θ, see below.

What we do is construct a sequence of functions, $\{Q_n(\theta)\}$, which for each finite n have some distribution, but which "converge to" $Q(\theta)$ in some sense as n grows large. If that sense is strong enough, then for n sufficiently large, the value of θ that minimizes Q_n will be the value that minimizes Q. What do we need for this logic to make sense?

1. **Identification Condition.** The first thing to note is that there must be only one "minimum" of $Q(\theta)$. More formally what we will require is that for every θ different from θ_0 (i.e., provided $||\theta - \theta_0|| \geq \delta > 0, Q(\theta) - Q(\theta_0) \geq \epsilon(\delta) > 0$). If this were not true then there would be two distinct θ's that would minimize the objective function, and we have no way of distinguishing which one is the true θ_0 (alternatively there is no way of knowing whether the computational algorithm stops at the right one). Consequently this is a condition we will have to impose on the problem. We will refer to it as the **identification condition** for the nonlinear model. Of course, if there is another way to choose between different θ's, then the model could be identified even if this condition is not satisfied for Q.

2. **Convergence.** The second point here is that for the logic to make sense the convergence must be uniform over $\theta \in \Theta$, i.e., $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| = O_p(1)$. If this were not the case then, even if the identification condition were met, we could go to some θ different from θ_0, say θ_* and find that $Q_n(\theta)$ hovers about $Q_n(\theta_0)$ as θ circles around that θ_* even though $Q(\theta_*) - Q(\theta_0) > \delta > 0$. Since at any fixed θ, say θ_1, G_n is just a sample mean of mean zero i.i.d. deviates, a standard LLN establishes that $||Q_n(\theta_*) - Q(\theta_*)|| = o_p(1)$. What we need for consistency is the stronger property that $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| = o_p(1)$, i.e., a "uniform law of large numbers".

These two properties, that is the ULLN and the identification condition seem to be the only properties used in the intuition underlying the consistency of the estimator. Indeed, as we show now they are more than enough to prove consistency. Note that neither of these properties have anything directly to do with smoothness of the objective function; i.e., of $Q_n(\cdot)$. So using them certainly does not rule out estimators based on objective functions that are not differentiable.

Theorem 23.1. (Consistency).

Suppose that the following conditions hold:

A. The parameter space Θ is a compact subset of Euclidean p-space.

B. $Q_n(\theta)$ is continuous in $\theta \in \Theta$ for all possible samples and is a measurable function of the data for all $\theta \in \Theta$.

C. There exists a nonstochastic function $Q(\theta)$ such that

$$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{P} 0, \qquad (23.5)$$

as $n \to \infty$.

D. The limit $Q(\theta)$ achieves a unique global minimum at $\theta = \theta_0$.

Then,

$$\hat{\theta} \xrightarrow{P} \theta_0.$$

Proof. Conditions (A) and (B) guarantee that there exists a minimizer $\hat{\theta}$. From (D), if $||\theta - \theta_0|| > \delta$, then there is an $\epsilon(\delta) > 0$ such that $Q(\theta) - Q(\theta_0) \geq \epsilon(\delta)$. Consequently,

$$\Pr(||\hat{\theta} - \theta_0|| > \delta) \leq \Pr(Q(\hat{\theta}) - Q(\theta_0) \geq \epsilon(\delta)),$$

and it is sufficient to prove that for any $\epsilon(\delta) > 0$, the latter probability goes to zero. By adding and subtracting terms we obtain

$$\begin{aligned} Q(\hat{\theta}) - Q(\theta_0) &= Q(\hat{\theta}) - Q_n(\hat{\theta}) + Q_n(\hat{\theta}) - Q_n(\theta_0) + Q_n(\theta_0) - Q(\theta_0) \\ &= I + II + III. \end{aligned}$$

By the fact that $\hat{\theta} \in \Theta$, we can bound the first and third times in absolute value by the left hand side of (23.5), i.e.,

$$|I|, |III| \leq \sup_{\theta \in \Theta} |Q(\theta) - Q_n(\theta)| \xrightarrow{P} 0,$$

where the convergence to zero is assumed in (C). Then by the definition of the estimator, $II \leq 0$. Together this implies that $\Pr(Q(\hat{\theta}) - Q(\theta_0) \geq \epsilon(\delta)) \to 0$. □

Condition (A) that the parameter space is compact (closed and bounded such as an interval) is not needed in linear regression since we automatically, under full rank condition, have the existence of an estimator, but a general criterion may not have a minimizer over the whole of $\mathbb{R}^p$.

23.2.1.1 Uniformity

The **Uniform Law of Large Numbers (ULLN)** condition in (C) is necessary for uncountable parameter spaces, as the following example illustrates.

Example 23.3.

Consider the following family of functions $\{Q_n(\cdot), n = 1, 2, \ldots\}$, where

$$Q_n(\theta) = \begin{cases} \dfrac{\theta^2}{\theta^2 + (1-\theta)^2} & 0 \leq \theta < 1 \\ 1/2 & \theta = 1. \end{cases}$$

Then, $|Q_n(\theta)| \leq 1$ and

$$\lim_{n \to \infty} |Q_n(\theta) - Q(\theta)| = 0$$

for all fixed $\theta \in [0,1]$, where $Q(\theta) = 0$ for all θ with $0 \le \theta < 1$ and $Q(1) = 1/2$. However,

$$\lim_{n\to\infty} \sup_{0\le\theta\le 1} |Q_n(\theta) - Q(\theta)| \neq 0.$$

Furthermore,

$$Q_n\left(\frac{1}{n}\right) = 1 \text{ for all } n.$$

Thus the maximizing value of Q_n [$\theta_n = 1/n$] converges to 0, while the maximizing value of Q is achieved at $\theta = 1$.

The ULLN condition is often satisfied because $Q_n(\theta)$ is typically a function of a sample average, i.e., of the form $f\left(n^{-1}\sum_{i=1}^n q_i(\theta)\right)$, where q_i depends only on the ith observation. When the data are independent across i, or at least only weakly dependent, many results can be applied to verify the required convergence. Andrews (1994) gives suitable conditions for uniform law of large numbers to hold for sample averages $n^{-1}\sum_{i=1}^n q_i(\theta)$. Bernstein's inequality or its variants is often a key tool in establishing these results. In some cases, simple arguments apply

Example 23.4. [Normal linear regression model.]

Suppose that $Q_n(\beta) = -n^{-1}\sum_{i=1}^n (y_i - \beta^\top x_i)^2$. Then

$$\begin{aligned} -Q_n(\beta) &= \frac{1}{n}\sum_{i=1}^n \{\epsilon_i - (\beta-\beta_0)^\top x_i\}^2 \\ &= \frac{1}{n}\sum_{i=1}^n \epsilon_i^2 + (\beta-\beta_0)^\top \frac{1}{n}\sum_{i=1}^n x_i x_i^\top (\beta-\beta_0) + 2(\beta-\beta_0)^\top \frac{1}{n}\sum_{i=1}^n \epsilon_i x_i. \end{aligned}$$

But $n^{-1}\sum_{i=1}^n \epsilon_i x_i \xrightarrow{P} 0$, $n^{-1}\sum_{i=1}^n x_i x_i^\top \xrightarrow{P} M = \mathbb{E} x_i x_i^\top$, and $n^{-1}\sum_{i=1}^n \epsilon_i^2 \xrightarrow{P} \sigma^2$. Therefore,

$$Q_n(\beta) \xrightarrow{P} \sigma^2 + (\beta-\beta_0)^\top M (\beta-\beta_0) = Q(\beta).$$

The convergence is uniform over $\beta \in B$, where B is a compact set because for example (take $K = 1$ and $B = [-b, b]$)

$$\sup_{\beta\in B} (\beta-\beta_0)^\top \frac{1}{n}\sum_{i=1}^n \epsilon_i x_i \le \sup_{\beta\in B} |\beta-\beta_0| \times \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\right| \le 2b \times \left|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\right| \xrightarrow{P} 0.$$

We next give a classic ULLN result.

Theorem 23.2. (Glivenko, 1933; Cantelli, 1933)

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow{P} 0.$$

Proof. We assume for simplicity that F is continuous and strictly monotonic. Let x_{jk} be the value of x that satisfies $F(x_{jk}) = j/k$ for integer j, k with $j \le k$. For any x between x_{jk} and $x_{j+1,k}$,

$$F(x_{jk}) \le F(x) \le F(x_{j+1,k}); \quad F_n(x_{jk}) \le F_n(x) \le F_n(x_{j+1,k}),$$

while $0 \le F(x_{j+1,k}) - F(x_{jk}) \le 1/k$, so that

$$F_n(x) - F(x) \le F_n(x_{j+1,k}) - F(x_{jk}) \le F_n(x_{j+1,k}) - F(x_{j+1,k}) + \frac{1}{k}$$
$$F_n(x) - F(x) \ge F_n(x_{j,k}) - F(x_{j+1,k}) \ge F_n(x_{j,k}) - F(x_{j,k}) - \frac{1}{k}.$$

Therefore, for any x and k,

$$|F_n(x) - F(x)| \le \max_{1\le j\le k} |F_n(x_{jk}) - F(x_{jk})| + \frac{1}{k} \qquad (23.6)$$

Since the right hand side of (23.6) does not depend on x, we can replace the left hand side by $\sup_{-\infty<x<\infty} |F_n(x) - F(x)|$.

Let $\epsilon > 0$ be fixed and take $k = \lceil 1/2\epsilon \rceil$. It suffices to show that $\max_{1\le j\le k} |F_n(x_{jk}) - F(x_{jk})| \xrightarrow{P} 0$. Now let $A_{jk}(\epsilon) = \{\omega : |F_n(x_{jk}) - F(x_{jk})| \ge \epsilon\}$ and

$$A_k(\epsilon) = \bigcup_{j=1}^{k} A_{jk}(\epsilon) = \left\{\omega : \max_{1\le j\le k} |F_n(x_{jk}) - F(x_{jk})| \ge \epsilon\right\}.$$

We have for any $\delta > 0$ there exists n_j such that for all $n \ge n_j$, $\Pr(A_{jk}(\epsilon)) \le \delta$. Therefore,

$$\Pr\left(\max_{1\le j\le k} |F_n(x_{jk}) - F(x_{jk})| \ge \epsilon\right) = \Pr(A_k(\epsilon)) \le k\delta$$

for all $n \ge \max_{1\le j\le k} n_j$. It follows that for any δ' we can find δ such that $\delta' = k\delta$ and n' such that for all $n > n'$

$$\Pr\left(\max_{1\le j\le k} |F_n(x_{jk}) - F(x_{jk})| \ge \epsilon\right) \le \delta.$$

The result follows.

An alternative proof. We take $k = \log n$, which ensures that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \le \max_{1\le j\le k} |F_n(x_{jk}) - F(x_{jk})| + o(1)$$

as $n \to \infty$. For any $\epsilon > 0$, let $A_j = \{|F_n(x_{jk}) - F(x_{jk})| > \epsilon - 1/k\}$ and

$$A = \bigcup_{j=1}^{k} A_j = \left\{\max_{1\le j\le k} |F_n(x_{jk}) - F(x_{jk})| > \epsilon - \frac{1}{k}\right\}.$$

Then for large enough n, $\epsilon - 1/k > \epsilon/2$ and

$$\begin{aligned}
\Pr(A) &\le \sum_{j=1}^{k} \Pr(A_j) \\
&\le \sum_{j=1}^{k} \frac{4\mathbb{E}(F_n(x_{jk}) - F(x_{jk}))^2}{\epsilon^2} \\
&= \sum_{j=1}^{k} \frac{4F(x_{jk})(1 - F(x_{jk}))}{\epsilon^2 n} \\
&\le \frac{\log n}{\epsilon^2 n},
\end{aligned}$$

because $F(x)(1-F(x)) \leq 1/4$ for all x. The first inequality follows by the Bonferroni inequality, while the second one uses the Chebychev or Markov inequality. Now $\log n/n \to 0$ so that $\Pr(A) \to 0$ as $n \to \infty$, which implies the result. □

The asymptotic properties of the empirical c.d.f. have been established in 1933 in two separate papers by Glivenko and Cantelli; in fact they showed the strong law version (convergence with probability one). The only 'condition' in their result is that X_i are i.i.d., although note that since F is a distribution function it has at most a countable number of discontinuities, is bounded between zero and one and is right continuous. Note also that the supremum is over a non-compact set – in much subsequent work generalizing this theorem it has been necessary to restrict attention to compact sets. The proof of this theorem exploits some special structure: specifically that for each x, $\mathbb{1}(X_i \leq x)$ is Bernoulli with probability $F(x)$. This proof is very special and uses the structure of the empirical c.d.f. quite a lot. Much work has gone into establish ULLN's for more general settings.

23.2.1.2 Identification

Often more difficult to establish is condition (D). For the MLE the following lemma is available.

Lemma 23.1.

Suppose that θ_0 is identified, i.e., for any $\theta \in \Theta$ with $\theta \neq \theta_0$ we have with positive probability $f(X|\theta) \neq f(X|\theta_0)$. Suppose also that for all $\theta \in \Theta$, $\mathbb{E}\{|\ln f(X|\theta)|\} < \infty$. Then, $Q(\theta) = \mathbb{E}\{\ln f(X|\theta)\}$ has a unique maximum at θ_0.

Proof. For all $\theta \neq \theta_0$,

$$Q(\theta_0) - Q(\theta) = \mathbb{E}\left[-\ln \frac{f(X|\theta)}{f(X|\theta_0)}\right] > -\ln\left[\mathbb{E}\frac{f(X|\theta)}{f(X|\theta_0)}\right] = 0,$$

by Jensen's inequality. □

Example 23.5. [Normal linear regression model.]

Suppose that $Q_n(\beta) = -n^{-1}\sum_{i=1}^n (y_i - \beta^\top x_i)^2$. Then we have

$$Q(\beta) = \sigma^2 + (\beta - \beta_0)^\top M (\beta - \beta_0).$$

Provided $M > 0$, the function $Q(\beta)$ is uniquely minimized at $\beta = \beta_0$.

Example 23.6. The simultaneous Equation system.

The normalized negative likelihood function is

$$Q_n(\theta) = \log\det(B(\theta)) + \frac{1}{2n}\sum_{i=1}^n (B(\theta)y_i - C(\theta)x_i)^\top \Sigma(\theta)^{-1}(B(\theta)y_i - C(\theta)x_i).$$

Under the assumption that $\mathbb{E}(y_i|x_i) = \Phi(\theta_0)x_i$ and $\operatorname{var}(y_i|x_i) = B(\theta_0)^{-1}\Sigma(\theta_0)B(\theta_0)^{\top -1}$ for some $\theta_0 \in \Theta$, $Q_n(\theta)$ has the probability limit

$$\begin{aligned} Q(\theta) = \log\det(B(\theta)) &+ \frac{1}{2}\operatorname{tr}\left(M(C(\theta_0) - C(\theta))^\top \Sigma(\theta)^{-1}(C(\theta_0) - C(\theta))\right) \\ &+ \frac{1}{2}\operatorname{tr}\left(B(\theta_0)^{-1}\Sigma(\theta_0)B(\theta_0)^{\top -1}B(\theta)^\top \Sigma(\theta)^{-1}B(\theta)\right) \end{aligned}$$

for each $\theta \in \Theta$. We have

$$Q(\theta_0) = \log\det(B(\theta_0)) + \frac{L}{2}.$$

The question is, whether there exists any other $\theta \in \Theta$ such that $Q(\theta) = Q(\theta_0)$. In the case where $\theta \in \mathbb{R}^{L^2+KL+L(L+1)/2}$ is unrestricted, then the answer is positive, meaning the model is not identified. This is because the triple $B^* = FB$, $C^* = FC$, and $\Sigma^* = F\Sigma F^\top$, where F is any nonsingular $L \times L$ matrix, will yield exactly the same $Q(\theta)$ because

$$\begin{aligned} B^{*-1}C^* &= B^{-1}F^{-1}FC = B^{-1}C \\ B^{*-1}\Sigma^* B^{*\top-1} &= B^{-1}F^{-1}F\Sigma F^\top F^{\top-1}B^{\top-1} = B^{-1}\Sigma B^{\top-1}. \end{aligned}$$

23.2.2 Asymptotic Normality

Once we have consistency we can confine ourselves to "local conditions" to prove subsequent limit properties of the estimator. That is if we now can prove that provided that any $\hat{\theta}$ that is eventually within a δ neighbourhood of θ_0 will have a particular property, then our estimator will have that property with probability tending to one (since our estimator will be in that neighbourhood with probability tending to one). This allows us to focus in on conditions on Q_n and Q in a neighbourhood of θ_0, and ignore entirely the behaviour of these functions outside of this neighbourhood. This is in distinct contrast to consistency, which is generally thought of as a "global" property; it depends on the properties of Q_n and Q over all of Θ. That is the conditions we need for consistency are global, but once we have consistency, the additional conditions we need for asymptotic normality are local. The literature often goes one step further than this in its discussion of local properties.

Theorem 23.3.

Suppose that $\hat{\theta} \xrightarrow{P} \theta_0$. Then there exists a sequence $\{\delta_n\}$ with $\delta_n \to 0$, such that

$$\lim_{n\to\infty} \Pr(||\hat{\theta} - \theta_0|| > \delta_n) = 0.$$

Proof. Consistency implies that for all $\epsilon > 0$ and for all positive integers J, there exists a positive integer $n_0(J)$ such that for all $n \geq n_0(J)$

$$\Pr(||\hat{\theta} - \theta_0|| > 1/J) \leq \epsilon.$$

For every J, let $n^*(J)$ be the smallest value of n that satisfies this condition, so that $n^*(J)$ is an increasing sequence. Then set

$$\delta_n = 1/J \text{ for } n^*(J) < n \leq n^*(J+1).$$

Clearly, $\lim_{n\to\infty} \delta_n = 0$. Therefore, by construction, for all $n \geq n^*(1/\delta_n)$

$$\Pr(||\hat{\theta} - \theta_0|| > \delta_n) \leq \epsilon. \qquad \square$$

The sequence δ_n can go to zero arbitrarily slowly depending on the case. The discussion of subsequent properties can confine itself to conditions that need only hold in "shrinking neighbourhoods" of θ_0; i.e., neighbourhoods of θ_0 that can get arbitrarily small as n grows large, and still we know that our estimator will have that property with probability tending to one. Define the event $\Theta_n = \{||\hat{\theta} - \theta_0|| \leq \delta_n\}$. Then, for any event A, we have

$$\Pr(A) = \Pr(A \cap \Theta_n) + \Pr(A \cap \Theta_n^c) \leq \Pr(A \cap \Theta_n) + \Pr(\Theta_n^c) \leq \Pr(A \cap \Theta_n) + \epsilon.$$

Since ϵ is arbitrary, we can effectively assume that Θ_n is true.

We next consider the asymptotic distribution of the optimization estimator.

Theorem 23.4. (Asymptotic Normality).

Suppose that the following conditions hold:

A. $\hat{\theta} \xrightarrow{P} \theta_0$;

B. θ_0 is an interior point of Θ

C. $\dfrac{\partial^2 Q_n}{\partial\theta\partial\theta^\top}(\theta)$ exists and is continuous in an open convex neighbourhood of θ_0.

D. There exists a finite nonsingular matrix A, such that

$$\sup_{|\theta-\theta_0|<\delta_n} \left|\left| \frac{\partial^2 Q_n}{\partial\theta\partial\theta^\top}(\theta) - A \right|\right| \xrightarrow{P} 0$$

for any sequence $\delta_n \to 0$.

E. $n^{1/2}\dfrac{\partial Q_n}{\partial\theta}(\theta_0) \xrightarrow{D} N(0, B)$ for some positive definite matrix B.

Then with $V(\theta) = A(\theta)^{-1}B(\theta)A(\theta)^{-1}$ and $V = V(\theta_0)$ we have

$$n^{1/2}(\hat{\theta} - \theta_0) \xrightarrow{D} N(0, V). \qquad (23.7)$$

Proof. Conditions A and B ensure that with probability tending to one, $\hat{\theta}$ satisfies the first order condition. By the Mean Value Theorem

$$0 = \frac{\partial Q_n}{\partial\theta}(\hat{\theta}) = n^{1/2}\frac{\partial Q_n}{\partial\theta}(\theta_0) + \frac{\partial^2 Q_n}{\partial\theta\partial\theta^\top}(\theta^*)n^{1/2}(\hat{\theta} - \theta_0),$$

where θ^* is intermediate between $\hat{\theta}$ and θ_0, i.e., $||\theta^* - \theta_0|| \leq ||\hat{\theta} - \theta_0||$. Actually, we need a different such θ^* for each row, but each will satisfy the contraction condition. By assumption D we can replace $\dfrac{\partial^2 Q_n}{\partial\theta\partial\theta^\top}(\theta^*)$ by the limiting matrix A with probability tending to one, and obtain that

$$n^{1/2}(\hat{\theta} - \theta_0) = -A^{-1}n^{1/2}\frac{\partial Q_n}{\partial\theta}(\theta_0) + R_n,$$

where the remainder term $R_n \xrightarrow{P} 0$. Then apply Assumption E and Slutsky's theorem to conclude the result (23.7). □

The uniformity condition (B) is needed, i.e. it is not generally sufficient that

$$\frac{\partial^2 Q_n}{\partial\theta\partial\theta^\top}(\theta_0) \xrightarrow{P} A.$$

In the linear regression case, the usual least squares objective function satisfies

$$\frac{\partial^2 Q_n}{\partial\beta\partial\beta^\top}(\beta) = -X^\top X$$

for all parameter values β and so this condition is automatically satisfied.

Condition B is needed because otherwise $\hat{\theta}$ may not satisfy the first order condition.

Example 23.7.

Suppose that $X \sim N(\mu, 1)$, where $\mu \in \Theta = [\underline{\mu}, \overline{\mu}]$. The maximum likelihood estimator in this case can be shown to be

$$\hat{\mu} = \begin{cases} \overline{\mu} & \text{if } \overline{X} > \overline{\mu} \\ \overline{X} & \text{if } \overline{X} \in [\underline{\mu}, \overline{\mu}] \\ \underline{\mu} & \text{if } \overline{X} \leq \underline{\mu}. \end{cases}$$

This will satisfy

$$\sum_{i=1}^{n} (X_i - \hat{\mu}) = 0 \qquad (23.8)$$

only when $\overline{X} \in (\underline{\mu}, \overline{\mu})$, i.e., the constraint is not binding. If the true parameter $\mu_0 = \underline{\mu}$, then we can see that only 50% of the time will (23.8) occur. In fact, $\hat{\mu}$ is not asymptotically normal in this case. See the exercise.

The precise form of the limiting distribution depends on the details of Q_n. We consider the main leading cases separately.

Correctly Specified Likelihood. The score function and Hessian are sample averages of independent variables (under the i.i.d. sampling) and thus satisfy the appropriate CLTs and ULLNs under some regularity conditions. By virtue of the information matrix equality, $A = B = \mathcal{I}$, the asymptotic variance has the simpler form $\mathcal{I}^{-1}$.

A number of different methods exist for estimating the asymptotic covariance matrix using either the Hessian or outer product:

$$\frac{1}{n} \frac{\partial^2 l}{\partial \theta \partial \theta^\top}(\hat{\theta}) \quad \text{or} \quad \frac{1}{n} \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta} \frac{\partial l_i}{\partial \theta^\top}(\hat{\theta}).$$

Misspecified Likelihood. Clearly, the meaning of the limit θ_0 is not obvious when the model is not correct, since these parameters may have nothing to do with the true distribution of the data. However, under regularity conditions, θ_0 is the parameter value that minimizes the Kullback-Liebler distance between the distribution of the data and the specified model. Under partial misspecification, one may retain consistency. For example, in a regression model if the mean is correctly specified but the distribution of the errors is not normal or the heteroskedasticity is ignored. In these cases the (pseudo or quasi) ML estimates of the mean parameters can be consistent. In the general case, the limiting variance is of the sandwich form $A^{-1}BA^{-1}$. One can carry out robust inference by allowing for this more general structure when estimating the asymptotic covariance matrix. The leading example here is when the parametric model is linear regression with normal homoskedastic errors but the distribution of the data has heteroskedasticity. In this case, $A = X^\top X$ and $B = X^\top \Sigma X$, which can be estimated by replacing the diagonal matrix Σ by the diagonal matrix whose elements are the squared least squares residuals. For a general (pseudo) likelihood criterion, we estimate the covariance matrix robustly by

$$\hat{V} = \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \theta \partial \theta^\top}(\hat{\theta}) \right\}^{-1} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial l_i}{\partial \theta} \frac{\partial l_i}{\partial \theta^\top}(\hat{\theta}) \right\} \left\{ \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 l_i}{\partial \theta \partial \theta^\top}(\hat{\theta}) \right\}^{-1}.$$

See White (1982) for a comprehensive discussion of this theory.

Generalized Method of Moments. Suppose that $W_n \xrightarrow{P} W > 0$ and that the moments are correctly specified. Then

$$\sqrt{n}(\hat{\theta}_{GMM} - \theta_0) \xrightarrow{D} N(0, (\Gamma^\top W \Gamma)^{-1} \Gamma^\top W \Omega W \Gamma (\Gamma^\top W \Gamma)^{-1}),$$

$$\Omega = \text{var}\sqrt{n}G_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(g(Z_i,\theta_0)g(Z_i,\theta_0)^\top)$$

$$\Gamma = \mathbb{E}\left(\frac{\partial g(Z_i,\theta_0)}{\partial\theta}\right).$$

If $p = q$, then $(\Gamma^\top W\Gamma)^{-1}\Gamma^\top W\Omega W\Gamma(\Gamma^\top W\Gamma)^{-1} = \Gamma^{-1}\Omega\Gamma^{\top-1}$, and the asymptotic variance simplifies. We estimate Γ and Ω by

$$\hat{\Gamma} = \frac{\partial G_n(\hat{\theta})}{\partial\theta}; \quad \hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} g(Z_i,\hat{\theta})g(Z_i,\hat{\theta})^\top$$

and hence $\hat{V} = (\hat{\Gamma}^\top W_n\hat{\Gamma})^{-1}\hat{\Gamma}^\top W_n\hat{\Omega}W_n\hat{\Gamma}(\hat{\Gamma}^\top W_n\hat{\Gamma})^{-1}$ is used to estimate $V = (\Gamma^\top W\Gamma)^{-1}\Gamma^\top W\Omega W\Gamma(\Gamma^\top W\Gamma)^{-1}$.

Example 23.8. A simple example, linear regression $y = X\beta + \epsilon$.

In this case, the moment conditions are

$$\mathbb{E}[X_i\epsilon_i(\beta_0)] = 0.$$

Here there are K conditions and K parameters, it corresponds to the exactly identified case. In this case, there exists a unique β that satisfies the empirical conditions provided X is of full rank.

Example 23.9. Instrumental variables.

Suppose now $\mathbb{E}[x_i\epsilon_i(\beta_0)] \neq 0$ because of omitted variables/endogeneity. However, suppose that

$$\mathbb{E}[z_i\epsilon_i(\beta_0)] = 0$$

for instruments $z_i \in \mathbb{R}^J$, with $J > K$. In this case, we can't solve uniquely for $\hat{\beta}_{IV}$ because there are too many equations which can't all be satisfied simultaneously. Take

$$G_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\theta)g_i(\theta)^\top; \quad g_i(\theta) = z_i\epsilon_i(\beta_0) = z_i(y_i - x_i^\top\beta_0).$$

If we take weighting matrix $W_n = (Z^\top Z)^{-1}$, where Z is the $n \times J$ matrix of instruments, then the objective function becomes

$$Q_n(\beta) = (y - X\beta)^\top Z(Z^\top Z)^{-1}Z^\top(y - X\beta) = ||P_Z(y - X\beta)||^2$$

where $P_Z = Z(Z^\top Z)^{-1}Z^\top$. This has a closed form solution

$$\hat{\beta}_{GMM} = ((P_Z X)^\top P_Z X)^{-1}(P_Z X)^\top(P_Z y) = (X^\top P_Z X)^{-1}X^\top P_Z y,$$

i.e., it is an instrumental variables estimator with instruments $X^* = P_Z X$. We require Z to be full rank.

23.3 Quantile Regression

The quantile regression model is of the form

$$y_i = x_i^\top\beta + u_i, \qquad (23.9)$$

where the conditional α-quantile of u_i given x_i is zero. Therefore, the conditional quantile function satisfies $Q_{y_i|x_i}(y_i|x_i;\alpha) = x_i^\top \beta$. The quantile restriction can be expressed as

$$\mathbb{E}(\psi_\alpha(u_i)|x_i) = 0,$$

where $\psi_\alpha(x) = \text{sgn}(x) - (1-2\alpha)$, which is a conditional moment restriction. To estimate β we may consider the objective function

$$Q_n(\beta) = \frac{1}{n}\sum_{i=1}^n \rho_\alpha(y_i - x_i^\top \beta), \text{ where } \rho_\alpha(x) = x(\alpha - \mathbb{1}(x<0),$$

or its first order condition

$$G_n(\beta) = \frac{1}{n}\sum_{i=1}^n x_i \psi_\alpha(y_i - x_i^\top \beta), \text{ where}$$

is the check function. The median corresponds to the case $\rho_{1/2}(x) = |x|$ and $\psi_{1/2}(x) = \text{sgn}(x)$.

The solution is not generally unique. Note that the objective function is not differentiable (at one point) and the first order condition is not even continuous at the same point, so that typically different methods have to be employed to study the large sample properties. Nevertheless, the consistency theorem can be applied directly. In fact, conditions A and B are unnecessary, because one can show that the objective function is globally convex and so there exists a minimizing value $\beta \in \mathbb{R}^K$ although it won't generally be unique.

We have the following result (an adaptation of Koenker (2005)).

Assumptions Q.

Q1. Suppose that u_i, x_i are i.i.d. and $u_i|x_i$ is continuous with density $f_{u|x}$ and α-quantile 0 such that $0 < \inf_x f_{u|x}(0|x) \leq \sup_x f_{u|x}(0|x) < \infty$.

Q2. The matrices $M = \mathbb{E}[x_i x_i^\top]$ and $M_\alpha = \mathbb{E}[x_i x_i^\top f_{u|x}(0|x_i)]$ exist and are positive definite.

Theorem 23.5.

Suppose that assumptions Q1 and Q2 hold. Then

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{D} N(0, \alpha(1-\alpha)M_\alpha^{-1} M M_\alpha^{-1}).$$

Proof. We just give some heuristics. We may show that for any $\beta \in \mathbb{R}^K$ that

$$G_n(\beta) = \frac{1}{n}\sum_{i=1}^n x_i \psi_\alpha(y_i - x_i^\top \beta) \xrightarrow{P} G(\beta) = \mathbb{E}[x_i \psi_\alpha(u_i - x_i^\top(\beta - \beta_0))].$$

There are special arguments that make use of the convexity of the objective function that guarantees that this convergence is uniform. We suppose for convenience of notation that x is continuously distributed. We can write

$$
\begin{aligned}
G(\beta) &= \int [x\psi_\alpha(u - x^\top(\beta - \beta_0))] f_{u|x}(u|x) f_x(x) du dx \\
&= \int [x \text{sgn}(u - x^\top(\beta - \beta_0))] f_{u|x}(u|x) f_x(x) du dx - (1 - 2\alpha) \\
&= \int [x \text{sgn}(v)] f_{u|x}(v + x^\top(\beta - \beta_0)|x) f_x(x) dv dx - (1 - 2\alpha) \\
&= \int_0^\infty x f_{u|x}(v + x^\top(\beta - \beta_0)|x) f_x(x) dv dx - \int_{-\infty}^0 x f_{u|x}(v + x^\top(\beta - \beta_0)|x) f_x(x) dv dx - (1 - 2\alpha) \\
&= \int x[1 - F_{u|x}(x^\top(\beta - \beta_0)|x)] f_x(x) dx - \int x F_{u|x}(x^\top(\beta - \beta_0)|x) f_x(x) dx - (1 - 2\alpha) \\
&= \int x[1 - 2F_{u|x}(x^\top(\beta - \beta_0)|x)] f_x(x) dx - (1 - 2\alpha),
\end{aligned}
$$

where we used the change of variable $u \to v = u - x^\top(\beta - \beta_0)$. The function $G(\beta)$ satisfies $G(\beta_0) = 0$ and it is differentiable with

$$
\frac{\partial G}{\partial \beta}(\beta_0) = -2 \int x x^\top f_{u|x}(0|x) f_x(x) dx = -2\mathbb{E}[x x^\top f_{u|x}(0|x)],
$$

which we will assume below is non-zero. This implies that at least in a neighbourhood of β_0 there can be no other zero of $G(\beta)$ and indeed $Q(\beta_0) < Q(\beta)$ for all such β, where $Q(\beta) = \mathbb{E}Q_n(\beta)$. Consistency follows. Regarding the asymptotic normality, the tricky part is to show that the estimator satisfies the condition

$$
n^{1/2}(\hat{\beta} - \beta_0) = \left[-\frac{\partial G}{\partial \beta}(\beta_0) \right]^{-1} n^{1/2} G_n(\beta_0) + R_n, \qquad (23.10)
$$

where $R_n \xrightarrow{P} 0$. Note that we have $\partial G / \partial \beta$ not $\mathbb{E}\partial G_n / \partial \beta$, which is not well defined due to the discontinuity of G. Once this is established, the result follows by the Slutsky theorem etc. since

$$
n^{1/2} G_n(\beta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^n x_i \psi_\alpha(u_i)
$$

satisfies a CLT because by assumption $\mathbb{E}(\psi_\alpha(u_i)|x_i) = 0$. $\square$

Quantile regression is robust in the sense that nowhere in the above theory have we required $\mathbb{E}(u_i^2) < \infty$ or even $\mathbb{E}(|u_i|) < \infty$, which contrasts with the theory for OLS. The monotone equivariance property implies that

$$
Q_{\Lambda(y_i)|x_i}(\Lambda(y_i)|x_i; \alpha) = \Lambda(x_i^\top \beta)
$$

for any strictly increasing transformation Λ such as the logarithm. This property says for example that we can infer the effect of x_i on say $\log(y_i)$ from the quantile regression of y_i on x_i.

Exercises

Exercise 1

Solution 1

Consider the simultaneous equations model in Example 23.1. Suppose $L = 2$, $K = 1$, and we have the following structural equations:

$$
\begin{aligned}
y_{1i} &= \gamma y_{2i} + \beta_1 x_i + u_{1i} \\
y_{2i} &= \delta y_{1i} + \beta_2 x_i + u_{2i},
\end{aligned}
$$

where $\mathbb{E}[u_{1i}|x_i] = \mathbb{E}[u_{2i}|x_i] = 0$. Express this model in the matrix form $B(\theta)y_i = C(\theta)x_i + u_i$, and identify the matrices $B(\theta)$, $C(\theta)$, and the parameter vector θ.

Exercise 2

Solution 2

Referring to Exercise 1, write down the reduced form equations for y_{1i} and y_{2i} in terms of x_i and the parameters. Explain how the reduced form relates to the matrix $\Phi(\theta)$ in Example 23.1. What condition on γ and δ is required for $B(\theta)$ to be invertible?

Exercise 3

Solution 3

For the model in Exercise 1, derive an expression for $g(Z_i, \theta)$ using $h(x_i) = x_i$. What is the dimension of $g(Z_i, \theta)$ in this case?

Exercise 4

Solution 4

In Example 23.2 (Hansen and Singleton, 1982), assume there is only one asset, $m = 1$, and $r_{1,i+1} = r_{i+1}$. Let $v_i = 1$. Derive the explicit form of $g(Z_i, \theta)$ and specify Z_i and θ. What is the dimension q of the moment condition?

Exercise 5

Solution 5

Explain the difference between the **exactly identified**, **overidentified**, and **unidentified** cases in the context of GMM. Provide an example of each case using a simple linear model with moment conditions.

Exercise 6

Solution 6

What is the role of the **weighting matrix** $W_n(\theta)$ in GMM? Explain how different choices of $W_n(\theta)$ lead to different estimators.

Exercise 7

Solution 7

Explain the concept of an **extremum estimator**. Give three examples of extremum estimators.

Exercise 8

Solution 8(#sec-ch23solution8}

State the **identification condition** for consistency of an extremum estimator. Explain why this condition is necessary.

Exercise 9

Solution 9(#sec-ch23solution9}

Explain the concept of **uniform convergence** in the context of extremum estimators and why it is important for consistency. Relate it to the Uniform Law of Large Numbers (ULLN).

Exercise 10

Solution 10(#sec-ch23solution10}

In the proof of Theorem 23.1 (Consistency), explain the roles of terms I, II, and III in establishing that $\Pr(||\hat{\theta} - \theta_0|| > \delta)$ goes to zero.

Exercise 11

Solution 11(#sec-ch23solution11}

Give an example where the parameter space Θ is *not* compact, but consistency of an estimator can still be established. Explain how the proof of Theorem 23.1 might be adapted.

Exercise 12

Solution 12(#sec-ch23solution12}

In Example 23.4 (Normal linear regression model), verify that $n^{-1}\sum_{i=1}^{n}\epsilon_i x_i \xrightarrow{P} 0$ under the assumption that $\mathbb{E}[\epsilon_i|x_i] = 0$ and $\mathbb{E}[||x_i||^2] < \infty$.

Exercise 13

Solution 13(#sec-ch23solution13}

In Example 23.7, show that if the true parameter $\mu_0 = \overline{\mu}$, then the MLE $\hat{\mu}$ is not asymptotically normal. Hint: Consider the probability $\Pr(\overline{X} > \overline{\mu})$.

Exercise 14

Solution 14(#sec-ch23solution14}

Explain the difference between **local** and **global conditions** in the context of establishing the asymptotic properties of extremum estimators.

Exercise 15

Solution 15(#sec-ch23solution15}

Explain why, under correct specification, the asymptotic variance of the MLE takes the simpler form $\mathcal{I}^{-1}$, where $\mathcal{I}$ is the information matrix.

Exercise 16

Solution 16(#sec-ch23solution16}

What is meant by a **misspecified likelihood**? Give an example where the MLE can still be consistent despite misspecification.

Exercise 17

Solution 17(#sec-ch23solution17}

In the context of GMM, explain how the asymptotic variance simplifies when $p = q$ (the exactly identified case).

Exercise 18

Solution 18(#sec-ch23solution18}

In Example 23.9 (Instrumental variables), explain why we require $J > K$ (the number of instruments to be greater than the number of endogenous regressors). What happens if $J = K$?

Exercise 19

Solution 19(#sec-ch23exercise19}

What is the **check function** in quantile regression? Write down the check function for the median regression ($\alpha = 0.5$).

Exercise 20

Solution 20(#sec-ch23solution20}

Explain the **monotone equivariance property** of quantile regression. Provide an example of how this property can be useful.

Solutions

Solution 1

Exercise 1

The given system of equations is:

$$
\begin{aligned}
y_{1i} &= \gamma y_{2i} + \beta_1 x_i + u_{1i} \\
y_{2i} &= \delta y_{1i} + \beta_2 x_i + u_{2i}
\end{aligned}
$$

We can rewrite this system in matrix form as follows:

$$
\begin{bmatrix} 1 & -\gamma \\ -\delta & 1 \end{bmatrix} \begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} x_i + \begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix}
$$

Comparing this to the general form $B(\theta) y_i = C(\theta) x_i + u_i$, we can identify:

$$
B(\theta) = \begin{bmatrix} 1 & -\gamma \\ -\delta & 1 \end{bmatrix}, \quad y_i = \begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix}, \quad C(\theta) = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad x_i = x_i, \quad u_i = \begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix}.
$$

The parameter vector θ consists of the unknown coefficients in $B(\theta)$ and $C(\theta)$, so $\theta = (\gamma, \delta, \beta_1, \beta_2)^\top$.

Solution 2

Exercise 2

To find the reduced form, we need to solve for y_{1i} and y_{2i} in terms of x_i and the parameters. We can do this by inverting $B(\theta)$:

$$B(\theta)^{-1} = \frac{1}{1-\gamma\delta}\begin{bmatrix} 1 & \gamma \\ \delta & 1 \end{bmatrix},$$

provided that $1-\gamma\delta \neq 0$. Multiplying both sides of the matrix equation by $B(\theta)^{-1}$, we get:

$$\begin{bmatrix} y_{1i} \\ y_{2i} \end{bmatrix} = \frac{1}{1-\gamma\delta}\begin{bmatrix} 1 & \gamma \\ \delta & 1 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} x_i + \frac{1}{1-\gamma\delta}\begin{bmatrix} 1 & \gamma \\ \delta & 1 \end{bmatrix}\begin{bmatrix} u_{1i} \\ u_{2i} \end{bmatrix}.$$

This simplifies to:

$$\begin{aligned} y_{1i} &= \frac{\beta_1 + \gamma\beta_2}{1-\gamma\delta} x_i + \frac{u_{1i} + \gamma u_{2i}}{1-\gamma\delta} \\ y_{2i} &= \frac{\delta\beta_1 + \beta_2}{1-\gamma\delta} x_i + \frac{\delta u_{1i} + u_{2i}}{1-\gamma\delta}. \end{aligned}$$

These are the reduced form equations. The matrix $\Phi(\theta)$ in Example 23.1 is the matrix that multiplies x_i in the reduced form, i.e.,

$$\Phi(\theta) = B(\theta)^{-1}C(\theta) = \frac{1}{1-\gamma\delta}\begin{bmatrix} 1 & \gamma \\ \delta & 1 \end{bmatrix}\begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} = \frac{1}{1-\gamma\delta}\begin{bmatrix} \beta_1 + \gamma\beta_2 \\ \delta\beta_1 + \beta_2 \end{bmatrix}.$$

The condition required for $B(\theta)$ to be invertible is that its determinant is non-zero, which is $1-\gamma\delta \neq 0$.

Solution 3

Exercise 3

From Exercise 1, we have:

$$B(\theta) = \begin{bmatrix} 1 & -\gamma \\ -\delta & 1 \end{bmatrix}, \quad C(\theta) = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \theta = (\gamma, \delta, \beta_1, \beta_2)^\top.$$

With $h(x_i) = x_i$, we have:

$$g(Z_i, \theta) = (B(\theta)y_i - C(\theta)x_i) \otimes h(x_i) = \begin{bmatrix} y_{1i} - \gamma y_{2i} - \beta_1 x_i \\ y_{2i} - \delta y_{1i} - \beta_2 x_i \end{bmatrix} \otimes x_i = \begin{bmatrix} (y_{1i} - \gamma y_{2i} - \beta_1 x_i)x_i \\ (y_{2i} - \delta y_{1i} - \beta_2 x_i)x_i \end{bmatrix}.$$

Since $B(\theta)$ is 2×2 and $h(x_i)$ is 1×1, the dimension of $g(Z_i, \theta)$ is $L \times M = 2 \times 1 = 2$, which is q. So, $g(Z_i, \theta)$ is a 2×1 vector.

Solution 4

Exercise 4

In Example 23.2, with $m = 1$, $r_{1,i+1} = r_{i+1}$, and $v_i = 1$, the moment condition is:

$$\mathbb{E}\left[\left\{1 - \beta(1 + r_{i+1})\left(\frac{c_{i+1}}{c_i}\right)^{-\gamma}\right\} \middle| I_i^*\right] = 0.$$

The function $g(Z_i, \theta)$ is given by:

$$g(Z_i, \theta) = \left\{ 1 - \beta(1 + r_{i+1}) \left(\frac{c_{i+1}}{c_i} \right)^{-\gamma} \right\} v_i.$$

With $v_i = 1$, this becomes:

$$g(Z_i, \theta) = 1 - \beta(1 + r_{i+1}) \left(\frac{c_{i+1}}{c_i} \right)^{-\gamma}.$$

Here, $Z_i = (c_i, c_{i+1}, r_{i+1})^\top$, and $\theta = (\beta, \gamma)^\top$. Since there is only one moment condition, the dimension q is 1.

Solution 5

Exercise 5

- **Exactly identified (p=q):** The number of moment conditions (q) equals the number of parameters (p). In this case, we can typically find a unique solution to the sample moment conditions $G_n(\theta) = 0$.

  Example: $y_i = \beta x_i + u_i$, with $\mathbb{E}[u_i] = 0$. Here, $p = q = 1$, and the sample moment condition is $\frac{1}{n} \sum_{i=1}^{n} (y_i - \beta x_i) = 0$.

- **Overidentified (p<q):** The number of moment conditions (q) is greater than the number of parameters (p). In this case, we generally cannot find a solution that satisfies all sample moment conditions simultaneously. We use a weighting matrix to find the "best" solution.

  Example: $y_i = \beta x_i + u_i$, with $\mathbb{E}[u_i] = 0$ and $\mathbb{E}[x_i u_i] = 0$. Here, $p = 1$ and $q = 2$. We have two moment conditions, but only one parameter.

- **Unidentified (p>q):** The number of moment conditions (q) is less than the number of parameters (p). In this case, there are infinitely many solutions that satisfy the moment conditions, and we cannot uniquely estimate the parameters.

  Example: $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i$, with only one moment condition $\mathbb{E}[u_i] = 0$. Here p=2 and q=1. We have one moment condition and two parameters.

Solution 6

Exercise 6

The **weighting matrix** $W_n(\theta)$ in GMM is a $q \times q$ positive definite matrix that determines how the sample moment conditions $G_n(\theta)$ are combined to estimate the parameters. In the overidentified case ($p < q$), we cannot satisfy all moment conditions exactly, so $W_n(\theta)$ determines the relative importance given to each moment condition.

Different choices of $W_n(\theta)$ lead to different GMM estimators. For example:

- $W_n(\theta) = I$: This gives equal weight to all moment conditions.
- $W_n(\theta) = \hat{\Omega}^{-1}$: This is the optimal weighting matrix, where $\hat{\Omega}$ is a consistent estimator of the variance-covariance matrix of the sample moment conditions. This choice leads to the most efficient GMM estimator.

In general, any positive definite weighting matrix will lead to a consistent estimator, but the optimal weighting matrix leads to the most efficient estimator (smallest asymptotic variance).

Solution 7

Exercise 7

An **extremum estimator** is an estimator that is obtained by minimizing (or maximizing) some objective function $Q_n(\theta)$ over the parameter space Θ.

Examples of extremum estimators:

1. **Ordinary Least Squares (OLS):** Minimizes the sum of squared residuals: $Q_n(\beta) = \sum_{i=1}^{n}(y_i - x_i^\top \beta)^2$.
2. **Maximum Likelihood Estimation (MLE):** Maximizes the likelihood function (or equivalently, minimizes the negative log-likelihood function).
3. **Generalized Method of Moments (GMM):** Minimizes a quadratic form of the sample moment conditions: $Q_n(\theta) = G_n(\theta)^\top W_n(\theta) G_n(\theta)$.

Solution 8

Exercise 8

The **identification condition** for consistency of an extremum estimator states that the limit of the objective function, $Q(\theta)$, must have a unique global minimum at the true parameter value θ_0. Formally, for every $\theta \neq \theta_0$, we require $Q(\theta) > Q(\theta_0)$.

This condition is necessary because if there were another value $\theta_1 \neq \theta_0$ such that $Q(\theta_1) = Q(\theta_0)$, then the objective function would have two global minima. In this case, we would have no way of distinguishing between θ_0 and θ_1 based on the objective function, and the estimator would not be consistent.

Solution 9

Exercise 9

Uniform convergence in the context of extremum estimators means that the objective function $Q_n(\theta)$ converges to its limit $Q(\theta)$ uniformly over the parameter space Θ. Formally, $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{P} 0$.

This is important for consistency because it ensures that the minimizer of $Q_n(\theta)$ converges to the minimizer of $Q(\theta)$. If the convergence were not uniform, then $Q_n(\theta)$ could be very different from $Q(\theta)$ for some values of θ, even for large n, and the minimizer of $Q_n(\theta)$ might not be close to the true parameter value θ_0.

The Uniform Law of Large Numbers (ULLN) provides conditions under which sample averages converge uniformly to their expectations. Since many objective functions in extremum estimation involve sample averages (e.g., GMM, MLE), the ULLN is often used to establish uniform convergence.

Solution 10

Exercise 10

In the proof of Theorem 23.1, we have:

$$Q(\hat{\theta}) - Q(\theta_0) = \underbrace{Q(\hat{\theta}) - Q_n(\hat{\theta})}_{I} + \underbrace{Q_n(\hat{\theta}) - Q_n(\theta_0)}_{II} + \underbrace{Q_n(\theta_0) - Q(\theta_0)}_{III}.$$

- **Term I:** By uniform convergence (condition C), $|I| = |Q(\hat{\theta}) - Q_n(\hat{\theta})| \leq \sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{P} 0$. This means that the difference between the limit objective function and the sample objective function evaluated at the estimator goes to zero in probability.

- **Term II:** Since $\hat{\theta}$ minimizes $Q_n(\theta)$, we have $Q_n(\hat{\theta}) \leq Q_n(\theta_0)$. Therefore, $II = Q_n(\hat{\theta}) - Q_n(\theta_0) \leq 0$.

- **Term III:** Again, by uniform convergence, $|III| = |Q_n(\theta_0) - Q(\theta_0)| \leq \sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{P} 0$. This means the difference between the sample objective function and the limit objective function evaluated at the true parameter goes to zero in probability.

Combining these results, we have that $Q(\hat{\theta}) - Q(\theta_0)$ is bounded by terms that go to zero in probability, plus a non-positive term. This implies that $Q(\hat{\theta}) - Q(\theta_0)$ converges to a non-positive number. But from the identification condition $Q(\hat{\theta}) - Q(\theta_0) \geq \epsilon(\delta)$ when $||\hat{\theta} - \theta_0|| > \delta$. This implies that
$\Pr(||\hat{\theta} - \theta_0|| > \delta) \leq \Pr(Q(\hat{\theta}) - Q(\theta_0) \geq \epsilon(\delta)) \to 0$

Solution 11

Exercise 11

Consider the linear regression model $y_i = \beta x_i + u_i$, where x_i and u_i are i.i.d. with $\mathbb{E}[u_i] = 0$ and $\mathbb{E}[x_i u_i] = 0$. The parameter space for β is often taken to be $\Theta = \mathbb{R}$, which is *not* compact (it is not bounded).

Consistency of the OLS estimator $\hat{\beta} = (\sum_{i=1}^n x_i^2)^{-1} \sum_{i=1}^n x_i y_i$ can still be established under suitable assumptions (e.g., $\mathbb{E}[x_i^2] > 0$).

The proof of Theorem 23.1 relies on the existence of a minimizer $\hat{\theta}$, which is guaranteed by the compactness of Θ and the continuity of $Q_n(\theta)$. In the linear regression case, even though Θ is not compact, the OLS estimator exists and is unique (with probability 1) as long as $\sum_{i=1}^n x_i^2 > 0$. We can work directly with probability limits. The objective function $Q_n(\beta) = \frac{1}{n} \sum (y_i - \beta x_i)^2$. We have already shown in Example 23.4 that
$Q_n(\beta) \xrightarrow{P} Q(\beta) = \sigma^2 + (\beta - \beta_0)^2 \mathbb{E}(x_i^2)$ which has a unique minimum at $\beta = \beta_0$. We can modify the proof by restricting attention to a compact subset of Θ, say $B_M = [-M, M]$, where M is a large positive number. We can show that, with probability approaching 1, the minimizer of $Q_n(\beta)$ lies within B_M. Then, we can apply Theorem 23.1 on B_M to show consistency.

Solution 12

Exercise 12

Let $w_i = \epsilon_i x_i$. We are given that $\mathbb{E}[\epsilon_i | x_i] = 0$ and $\mathbb{E}[||x_i||^2] < \infty$. By the law of iterated expectations,
$\mathbb{E}[w_i] = \mathbb{E}[\mathbb{E}[\epsilon_i x_i | x_i]] = \mathbb{E}[x_i \mathbb{E}[\epsilon_i | x_i]] = \mathbb{E}[x_i \cdot 0] = 0$. Since x_i and ϵ_i are i.i.d., w_i are also i.i.d.

We need to show that $n^{-1} \sum_{i=1}^n w_i \xrightarrow{P} 0$. By the Weak Law of Large Numbers (WLLN), if $\mathbb{E}[w_i] = 0$ and $\text{Var}(w_i) < \infty$, then $n^{-1} \sum_{i=1}^n w_i \xrightarrow{P} 0$.

We have $\mathbb{E}[w_i] = 0$. Now we need to check if $\text{Var}(w_i) < \infty$. Assuming that ϵ_i and x_i have finite second moments, we consider:

$$\text{Var}(w_i) = \mathbb{E}[w_i^2] - (\mathbb{E}[w_i])^2 = \mathbb{E}[w_i^2] = \mathbb{E}[\epsilon_i^2 x_i^2]$$

If we also assume that $\mathbb{E}[\epsilon_i^2 | x_i] = \sigma^2$, then

$$\mathbb{E}[\epsilon_i^2 x_i^2] = \mathbb{E}[\mathbb{E}[\epsilon_i^2 x_i^2 | x_i]] = \mathbb{E}[x_i^2 \mathbb{E}[\epsilon_i^2 | x_i]] = \sigma^2 \mathbb{E}[x_i^2]$$

Since we are assuming $\mathbb{E}[x_i^2] < \infty$ we conclude that $\text{Var}(w_i) < \infty$. By the WLLN, $n^{-1} \sum_{i=1}^n \epsilon_i x_i \xrightarrow{P} 0$.

Solution 13

Exercise 13

If $\mu_0 = \overline{\mu}$, the MLE is

$$\hat{\mu} = \begin{cases} \overline{\mu} & \text{if } \overline{X} > \overline{\mu} \\ \overline{X} & \text{if } \overline{X} \leq \overline{\mu}. \end{cases}$$

Let $Z = \sqrt{n}(\overline{X} - \mu_0) = \sqrt{n}(\overline{X} - \overline{\mu})$. Since $X_i \sim N(\mu_0, 1)$, we have $\overline{X} \sim N(\mu_0, 1/n)$, and thus $Z \sim N(0, 1)$.

Now, consider $\sqrt{n}(\hat{\mu} - \overline{\mu})$. If $\overline{X} > \overline{\mu}$, then $\hat{\mu} = \overline{\mu}$, so $\sqrt{n}(\hat{\mu} - \overline{\mu}) = 0$. If $\overline{X} \leq \overline{\mu}$, then $\hat{\mu} = \overline{X}$, so $\sqrt{n}(\hat{\mu} - \overline{\mu}) = \sqrt{n}(\overline{X} - \overline{\mu}) = Z$.

Therefore,

$$\sqrt{n}(\hat{\mu} - \overline{\mu}) = \begin{cases} 0 & \text{if } Z > 0 \\ Z & \text{if } Z \leq 0. \end{cases}$$

This means that $\sqrt{n}(\hat{\mu} - \overline{\mu})$ converges in distribution to a random variable that is 0 with probability 1/2 (since $Z \sim N(0, 1)$) and equal to Z with probability 1/2. This is a mixed distribution, and it is not a normal distribution. Therefore, $\hat{\mu}$ is not asymptotically normal when $\mu_0 = \overline{\mu}$.

Solution 14

Exercise 14

- **Global conditions:** These conditions relate to the behavior of the objective function $Q_n(\theta)$ and its limit $Q(\theta)$ over the *entire* parameter space Θ. Consistency typically requires global conditions (e.g., the identification condition and uniform convergence).
- **Local conditions:** These conditions relate to the behavior of $Q_n(\theta)$ and $Q(\theta)$ in a *neighborhood* of the true parameter value θ_0. Asymptotic normality typically requires local conditions (e.g., smoothness of $Q_n(\theta)$ around θ_0, and convergence of the second derivative).

The idea is that once we have established consistency ($\hat{\theta} \xrightarrow{P} \theta_0$), we know that the estimator will be close to the true parameter value with high probability for large n. Therefore, we only need to examine the behavior of the objective function *locally* around θ_0 to derive the asymptotic distribution.

Solution 15

Exercise 15

Under correct specification, the score function $s(Z_i, \theta) = \frac{\partial l_i}{\partial \theta}$ has mean zero at the true parameter value: $\mathbb{E}[s(Z_i, \theta_0)] = 0$. The information matrix is defined as

$$\mathcal{I} = \mathbb{E}\left[s(Z_i, \theta_0) s(Z_i, \theta_0)^\top\right] = -\mathbb{E}\left[\frac{\partial^2 l_i}{\partial \theta \partial \theta^\top}(\theta_0)\right].$$

This is the information matrix equality. In the context of Theorem 23.4, this means that $A = -\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \frac{\partial^2 l_i}{\partial \theta \partial \theta^\top}(\theta_0)\right]$ and $B = \text{Var}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^n \frac{\partial l_i}{\partial \theta}(\theta_0)\right) = \mathbb{E}\left[s(Z_i, \theta_0) s(Z_i, \theta_0)^\top\right]$. Thus, under the information matrix equality, $A = B = \mathcal{I}$.

The asymptotic variance in Theorem 23.4 is given by $V = A^{-1} B A^{-1}$. Under correct specification and the information matrix equality, we have $A = B = \mathcal{I}$, so $V = \mathcal{I}^{-1} \mathcal{I} \mathcal{I}^{-1} = \mathcal{I}^{-1}$.

Solution 16

Exercise 16

A **misspecified likelihood** means that the assumed parametric model for the data is incorrect. That is, the true data generating process does not belong to the family of distributions assumed by the likelihood function.

Example: Suppose we observe data y_i and x_i, and we assume a linear model with normal errors: $y_i = \beta x_i + u_i$, where $u_i \sim N(0, \sigma^2)$. We use maximum likelihood to estimate β and σ^2 under this assumption.

However, suppose the true model is $y_i = \beta x_i + u_i$, where u_i follows a t-distribution with 3 degrees of freedom (which has heavier tails than the normal distribution). In this case, the likelihood is misspecified.

Despite the misspecification, the MLE for β (which is equivalent to the OLS estimator in this case) can still be consistent for the true β if $\mathbb{E}[u_i|x_i] = 0$. This is because the OLS estimator only relies on the conditional mean being correctly specified. However, the MLE for σ^2 will generally be inconsistent, and the standard errors for $\hat{\beta}$ based on the normality assumption will be incorrect.

Solution 17

Solution 17

In the exactly identified case ($p = q$), the number of moment conditions equals the number of parameters. The asymptotic variance of the GMM estimator is given by:

$$V = (\Gamma^\top W \Gamma)^{-1} \Gamma^\top W \Omega W \Gamma (\Gamma^\top W \Gamma)^{-1},$$

where $\Gamma = \mathbb{E}\left[\frac{\partial g(Z_i, \theta_0)}{\partial \theta}\right]$ and $\Omega = \text{Var}\sqrt{n} G_n(\theta_0) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(g(Z_i, \theta_0) g(Z_i, \theta_0)^\top)$.

Since $p = q$, Γ is a square matrix (assuming it's full rank). Therefore, we can simplify:

$$\begin{aligned} V &= (\Gamma^\top W \Gamma)^{-1} \Gamma^\top W \Omega W \Gamma (\Gamma^\top W \Gamma)^{-1} \\ &= \Gamma^{-1} W^{-1} (\Gamma^\top)^{-1} \Gamma^\top W \Omega W \Gamma \Gamma^{-1} W^{-1} (\Gamma^\top)^{-1} \\ &= \Gamma^{-1} W^{-1} \Omega W \Gamma^{-1} \\ &= \Gamma^{-1} \Omega \Gamma^{\top -1}. \end{aligned}$$

So, the asymptotic variance simplifies to $\Gamma^{-1} \Omega \Gamma^{\top -1}$.

Solution 18

Exercise 18

In the instrumental variables (IV) setting, we have endogenous regressors (correlated with the error term), and we use instruments to address this endogeneity.

We require $J > K$ (overidentification) because:

1. **Identification:** If $J < K$ (underidentification), we have fewer instruments than endogenous regressors. In this case, we cannot uniquely identify the parameters of interest. We have more parameters to estimate than moment conditions to use. The parameters are not identified.

2. **Estimation:** If $J = K$ (exact identification), we can solve the sample moment conditions exactly, and the IV estimator is equivalent to setting the sample moments to zero. This is analogous to the method of moments.

If $J > K$, we have more moment conditions than parameters. This allows us to use the extra moment conditions to improve the efficiency of the estimator and to test the validity of the instruments (overidentifying restrictions test). With more information (instruments), we can get a "better" estimate by minimizing the weighted distance of sample moments to zero.

Solution 19

Exercise 19

The **check function** in quantile regression, denoted by $\rho_\alpha(u)$, is defined as:

$$\rho_\alpha(u) = u(\alpha - \mathbb{1}(u < 0)),$$

where α is the quantile of interest ($0 < \alpha < 1$), and $\mathbb{1}(u < 0)$ is the indicator function, which equals 1 if $u < 0$ and 0 otherwise.

For the median regression, $\alpha = 0.5$. Substituting this into the check function, we get:

$$\rho_{0.5}(u) = u(0.5 - \mathbb{1}(u < 0)).$$

This can be simplified to:

$$\rho_{0.5}(u) = \begin{cases} 0.5u & \text{if } u \geq 0 \\ 0.5u - u = -0.5u & \text{if } u < 0. \end{cases}$$

This is equivalent to $\rho_{0.5}(u) = 0.5|u|$. Thus the check function that we minimize is equivalent to minimizing the sum of absolute values of the residuals.

Solution 20

Exercise 20

The **monotone equivariance property** of quantile regression states that if we apply a strictly increasing transformation $\Lambda(\cdot)$ to the response variable y, the conditional quantiles of the transformed variable $\Lambda(y)$ are equal to the transformed conditional quantiles of y. Formally:

$$Q_{\Lambda(y_i)|x_i}(\Lambda(y_i)|x_i; \alpha) = \Lambda(Q_{y_i|x_i}(y_i|x_i; \alpha)) = \Lambda(x_i^\top \beta).$$

Example: Suppose we are interested in the effect of education (x) on wages (y). Instead of modeling the conditional quantiles of wages directly, we might be interested in the conditional quantiles of *log* wages, $\log(y)$.

The monotone equivariance property tells us that if we estimate the α-quantile regression of y on x:

$$Q_{y_i|x_i}(y_i|x_i; \alpha) = x_i^\top \beta,$$

then the α-quantile regression of $\log(y)$ on x is simply:

$$Q_{\log(y_i)|x_i}(\log(y_i)|x_i; \alpha) = \log(x_i^\top \beta).$$

That is the coefficient of the log quantile regression can be interpreted as the effect of the covariates on the log of the quantiles of y.

This is useful because it allows us to model the conditional quantiles of a transformed variable without having to re-estimate the model. It also implies that the quantile regression model is robust to monotonic transformations of the response variable.

R Scripts

R Script 1: Simultaneous Equations Model and Identification

```
# Load necessary packages
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✓ dplyr     1.1.4     ✓ readr     2.1.5
✓ forcats   1.0.0     ✓ stringr   1.5.1
✓ ggplot2   3.5.1     ✓ tibble    3.2.1
✓ lubridate 1.9.4     ✓ tidyr     1.3.1
✓ purrr     1.0.2
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ dplyr::filter() masks stats::filter()
✖ dplyr::lag()    masks stats::lag()
ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(matlib) # For matrix operations

# Set seed for reproducibility
set.seed(123)

# Simulate data for a simultaneous equations model (Example 23.1)
n <- 1000  # Sample size
gamma <- 0.6 # Parameter
delta <- 0.2 # Parameter
beta1 <- 1  # Parameter
beta2 <- -0.5 # Parameter
x <- rnorm(n) # Exogenous variable
u1 <- rnorm(n) # Error term 1
u2 <- rnorm(n) # Error term 2

# Structural equations
# y1 = gamma * y2 + beta1 * x + u1
# y2 = delta * y1 + beta2 * x + u2

# Reduced form equations (see Exercise 2 Solution)
# y1 = (beta1 + gamma*beta2)/(1 - gamma*delta) * x + (u1 + gamma*u2)/(1 - gamma*delta)
# y2 = (delta*beta1 + beta2)/(1 - gamma*delta) * x + (delta*u1 + u2)/(1 - gamma*delta)
y1 <- (beta1 + gamma*beta2)/(1 - gamma*delta) * x + (u1 + gamma*u2)/(1 - gamma*delta)
y2 <- (delta*beta1 + beta2)/(1 - gamma*delta) * x + (delta*u1 + u2)/(1 - gamma*delta)

# Create a data frame
df <- data.frame(y1, y2, x)

# Define B(theta) and C(theta) (see Exercise 1 Solution)
B_theta <- function(gamma, delta) {
  matrix(c(1, -gamma, -delta, 1), nrow = 2, byrow = TRUE)
}

C_theta <- function(beta1, beta2) {
  matrix(c(beta1, beta2), nrow = 2)
}

# Check invertibility of B(theta) (determinant should be non-zero)
det(B_theta(gamma, delta))  # 1 - gamma*delta
```

```
[1] 0.88
```

```
# Calculate Phi(theta) = B(theta)^(-1) * C(theta) (Reduced form coefficients)
Phi_theta <- solve(B_theta(gamma, delta)) %*% C_theta(beta1, beta2)
Phi_theta
```

```
           [,1]
[1,]  0.7954545
[2,] -0.3409091
```

```
# Estimate reduced form by OLS
lm_y1 <- lm(y1 ~ x, data = df)
lm_y2 <- lm(y2 ~ x, data = df)

# Compare estimated reduced form coefficients with true Phi(theta)
coef(lm_y1)
```

```
(Intercept)           x
 0.03313892  0.88250631
```

```
coef(lm_y2)
```

```
(Intercept)           x
-0.01317719 -0.34256831
```

```
print("True Phi(theta):")
```

```
[1] "True Phi(theta):"
```

```
Phi_theta
```

```
           [,1]
[1,]  0.7954545
[2,] -0.3409091
```

```
# Visualize the relationship between y1 and x, and y2 and x
ggplot(df, aes(x = x)) +
  geom_point(aes(y = y1, color = "y1"), alpha = 0.6) +
  geom_point(aes(y = y2, color = "y2"), alpha = 0.6) +
  geom_smooth(aes(y = y1, color = "y1"), method = "lm", se = FALSE) +
  geom_smooth(aes(y = y2, color = "y2"), method = "lm", se = FALSE) +
  labs(title = "Simultaneous Equations Model: Reduced Form",
       x = "x",
       y = "y1 and y2",
       color = "Variable") +
  theme_bw()
```

```
`geom_smooth()` using formula = 'y ~ x'
`geom_smooth()` using formula = 'y ~ x'
```

Explanation:

1. **Load Packages:** We load `tidyverse` for data manipulation and visualization, and `matlib` for matrix operations like `solve()` (matrix inverse).
2. **Simulate Data:** We simulate data from a simultaneous equations model with two endogenous variables (y_1, y_2) and one exogenous variable (x). The parameters γ, δ, β_1, and β_2 determine the relationships between the variables. We generate the data using the *reduced form* equations, derived from the structural equations.
3. **Create Data Frame:** We combine the simulated variables into a data frame `df`.
4. **Define `B_theta` and `C_theta`:** These functions represent the matrices $B(\theta)$ and $C(\theta)$ from the simultaneous equations model, as described in Example 23.1 and Exercise 1.
5. **Check Invertibility:** We calculate the determinant of $B(\theta)$. For the reduced form to exist (and for the model to be identified in a certain sense), $B(\theta)$ must be invertible, meaning its determinant must be non-zero.

6. **Calculate `Phi_theta`:** We calculate the matrix $\Phi(\theta) = B(\theta)^{-1}C(\theta)$, which represents the coefficients of the reduced form equations.
7. **Estimate Reduced Form (OLS):** We estimate the reduced form equations using ordinary least squares (OLS) regression. This gives us estimates of the reduced form coefficients.
8. **Compare Coefficients:** We compare the OLS estimates of the reduced form coefficients with the true $\Phi(\theta)$ calculated from the known parameters. They should be close, demonstrating that OLS can consistently estimate the reduced form parameters.
9. **Visualize:** The `ggplot2` code creates a scatter plot of y_1 and y_2 against x, along with the OLS regression lines. This visualizes the reduced form relationships. The plot illustrates how the exogenous variable x affects both endogenous variables y_1 and y_2.

Connection to Text:

- **Example 23.1:** This script directly implements the simultaneous equations model described in the example. It shows how the structural equations are related to the reduced form.
- **Identification:** The condition `det(B_theta(gamma, delta)) != 0` relates to the identification of the structural parameters. If this determinant were zero, we couldn't uniquely recover the structural parameters from the reduced form estimates.
- **Reduced form:** The estimation of reduced form using OLS is discussed in the text.

R Script 2: Generalized Method of Moments (GMM) Estimation

```
# Load necessary packages (if not already loaded)
# library(tidyverse)
library(gmm) # For GMM estimation
```

```
Loading required package: sandwich
```

```
# Simulate data for a simple linear model with endogeneity (Example 23.9)
n <- 1000
beta_true <- 2
x <- rnorm(n) # Endogenous regressor
z <- rnorm(n) # Instrument
u <- 0.5 * z + rnorm(n) # Error term correlated with x (endogeneity)
y <- beta_true * x + u

# Create a data frame
df <- data.frame(y, x, z)

# Define the moment function g(Z_i, theta) (see Example 23.9)
g_func <- function(theta, data) {
  y <- data[, "y"]
  x <- data[, "x"]
  z <- data[, "z"]
  return(z * (y - theta * x)) # Moment condition: E[z * (y - theta * x)] = 0
}

# GMM estimation (exactly identified case: 1 parameter, 1 instrument)
gmm_result <- gmm(g = g_func, x = df, t0 = 0) # t0 is the initial value for theta
```

```
Warning in (function (par, fn, gr = NULL, ..., method = c("Nelder-Mead", : one-dimensional optimization
by Nelder-Mead is unreliable:
use "Brent" or optimize() directly
```

```
# Summary of GMM results
summary(gmm_result)
```

```

Call:
gmm(g = g_func, x = df, t0 = 0)
```

```
Method:  twoStep

Kernel:  Quadratic Spectral

Coefficients:
          Estimate     Std. Error   t value      Pr(>|t|)
Theta[1]  -4.9974e+03   1.5777e+06  -3.1675e-03   9.9747e-01

J-Test: degrees of freedom is 0
                J-test               P-value
Test E(g)=0:    1.0794189986549e-06  *******

#############
Information related to the numerical optimization
Convergence code =  0
Function eval. =  54
Gradian eval. =  NA

# Extract the estimated coefficient
beta_hat_gmm <- coef(gmm_result)
beta_hat_gmm

Theta[1]
 -4997.4

# Compare with true beta
beta_true

[1] 2

# Compare with OLS (which is biased due to endogeneity)
lm_result <- lm(y ~ x, data = df)
coef(lm_result)

(Intercept)           x
 0.01879416  2.04357828

# Overidentified case (add another instrument)
z2 <- rnorm(n)
df$z2 <- z2

g_func_overid <- function(theta, data) {
  y <- data[, "y"]
  x <- data[, "x"]
  z <- data[, c("z", "z2")] # Now using two instruments
  return(cbind(z[,1] * (y - theta * x), z[,2] * (y - theta * x)))
}

gmm_result_overid <- gmm(g = g_func_overid, x = df, t0 = 0)

Warning in (function (par, fn, gr = NULL, ..., method = c("Nelder-Mead", : one-dimensional optimization
by Nelder-Mead is unreliable:
use "Brent" or optimize() directly

Warning in (function (par, fn, gr = NULL, ..., method = c("Nelder-Mead", : one-dimensional optimization
by Nelder-Mead is unreliable:
use "Brent" or optimize() directly

summary(gmm_result_overid)


Call:
gmm(g = g_func_overid, x = df, t0 = 0)
```

```
Method:  twoStep

Kernel:  Quadratic Spectral(with bw =  0.46692 )

Coefficients:
          Estimate  Std. Error  t value  Pr(>|t|)
Theta[1]  0.45625   3.08024     0.14812  0.88225

J-Test: degrees of freedom is 1
                J-test     P-value
Test E(g)=0:    9.110e+01  1.366e-21

Initial values of the coefficients
Theta[1]
    2.95

#############
Information related to the numerical optimization
Convergence code =  0
Function eval. =  16
Gradian eval. =  NA

# J-test for overidentifying restrictions (test of instrument validity)
# The null hypothesis is that all instruments are valid
j_test <- summary(gmm_result_overid)$stest
print(j_test)


 ##  J-Test: degrees of freedom is 1  ##

                J-test     P-value
Test E(g)=0:    9.110e+01  1.366e-21
```

Explanation:

1. **Load Packages:** Load `gmm` for GMM estimation.
2. **Simulate Data:** Simulate data from a simple linear model $y = \beta x + u$, where x is endogenous (correlated with u). We create an instrument z that is correlated with x but uncorrelated with u.
3. **Create Data Frame:** Combine the variables into a data frame.
4. **Define Moment Function:** Define the function `g_func` that calculates the sample moment conditions. This function corresponds to $g(Z_i, \theta)$ in the text. The moment condition is based on the assumption that the instrument z is uncorrelated with the error term u: $\mathbb{E}[z(y - \theta x)] = 0$.
5. **GMM Estimation (Exactly Identified):** Use the `gmm()` function to estimate the parameter β using GMM. Since we have one parameter and one instrument, this is the exactly identified case. `t0` is the starting value for the optimization.
6. **Summary and Coefficient:** We print a summary of the GMM results and extract the estimated coefficient $\hat{\beta}_{GMM}$.
7. **Comparison:** We compare $\hat{\beta}_{GMM}$ with the true β and with the OLS estimate (which is biased due to endogeneity).
8. **Overidentified Case:** We add a second instrument (`z2`) and repeat the GMM estimation. Now we have one parameter and two instruments (overidentified).
9. **J-test:** We perform the J-test (test for overidentifying restrictions). This tests the null hypothesis that *all* instruments are valid (uncorrelated with the error term). A small p-value suggests that at least one instrument is invalid.

Connection to Text:

- **Example 23.9:** This script directly implements the instrumental variables example.
- **Moment Function:** The `g_func` corresponds to the moment function $g(Z_i, \theta)$ discussed in the text.

- **Exactly Identified and Overidentified:** The script demonstrates both the exactly identified and overidentified cases.
- **GMM Estimation:** The `gmm()` function performs the GMM estimation, minimizing the quadratic form of the sample moment conditions.
- **J-test:** The J-test is mentioned in the text in the context of overidentified models (page 326).

R Script 3: Maximum Likelihood Estimation (MLE) and Misspecification

```
# Load necessary packages (if not already loaded)
# library(tidyverse)

# Simulate data from a normal distribution
n <- 1000
mu_true <- 5
sigma_true <- 2
y <- rnorm(n, mean = mu_true, sd = sigma_true)

# Define the negative log-likelihood function for the normal distribution
neg_loglik_normal <- function(theta, data) {
  mu <- theta[1]
  sigma <- theta[2]
  if (sigma <= 0) {  # Ensure sigma is positive
    return(Inf)
  }
  -sum(dnorm(data, mean = mu, sd = sigma, log = TRUE))
}

# MLE estimation (correctly specified)
mle_result_correct <- optim(par = c(0, 1), fn = neg_loglik_normal, data = y,
                            method = "L-BFGS-B", lower = c(-Inf, 0.001)) # Constrain sigma > 0

# Estimated parameters
mle_result_correct$par

[1] 5.010923 1.998936

# True parameters
c(mu_true, sigma_true)

[1] 5 2

# Simulate data from a t-distribution (misspecified case)
y_t <- rt(n, df = 3) * sigma_true + mu_true # t-distribution with 3 df

# MLE estimation (misspecified) assuming normal distribution
mle_result_misspec <- optim(par = c(0, 1), fn = neg_loglik_normal, data = y_t,
                            method = "L-BFGS-B", lower = c(-Inf, 0.001))

# Estimated parameters under misspecification
mle_result_misspec$par

[1] 4.869772 3.157655

# True parameters (for comparison, although the "true" sigma doesn't exist in the same way)
c(mu_true, sigma_true)

[1] 5 2

# Illustrative Plot:
hist(y_t, breaks=30, probability=TRUE, main="Histogram of t-distributed Data and Fitted Normal")
curve(dnorm(x, mean=mle_result_misspec$par[1], sd=mle_result_misspec$par[2]),
      col="red", add=TRUE, lwd=2)
```

```
curve(dnorm(x, mean=mu_true, sd=sigma_true*sqrt(3)/sqrt(1) ), col="blue", add=TRUE, lty=2, lwd=2)
legend("topright", legend=c("Fitted Normal", "True Normal(same variance)"), col=c("red", "blue"),
         lty=1:2, lwd=2)
```

Explanation:

1. **Load Packages:** We might not need additional packages beyond what's already loaded for basic MLE.
2. **Simulate Data (Normal):** Simulate data from a normal distribution with known mean (`mu_true`) and standard deviation (`sigma_true`).
3. **Define Negative Log-Likelihood:** Define the negative log-likelihood function for the normal distribution. We minimize the *negative* log-likelihood because `optim()` is a minimization function. The `if (sigma <= 0)` part is important: it ensures that the optimization doesn't try invalid parameter values (standard deviation must be positive).
4. **MLE (Correctly Specified):** Use `optim()` to find the MLE estimates of μ and σ assuming a normal distribution. We use the "L-BFGS-B" method, which allows for box constraints (we constrain $\sigma > 0$).
5. **Estimated Parameters:** Print the estimated parameters from the correctly specified model.
6. **Simulate Data (t-distribution):** Simulate data from a t-distribution with 3 degrees of freedom. This represents the case where the true data generating process is *not* normal. We scale and shift the t-distribution so it has, approximately, mean of `mu_true` and standard deviation of `sigma_true`.
7. **MLE (Misspecified):** Use `optim()` again, but this time apply the *normal* log-likelihood function to the *t-distributed* data. This is the misspecified case.
8. **Estimated Parameters (Misspecified):** Print the estimated parameters under the misspecified model. The estimated mean will likely still be close to `mu_true`, demonstrating consistency of the mean estimator, but the estimated standard deviation will be biased.
9. **Illustrative plot**: This plots the histogram of the t-distributed data, showing the data's heavier tails. It also shows the fitted normal density from the misspecified MLE, along with a normal curve that has the same variance as the true t-distribution (which is $\frac{df}{df-2}\sigma^2$ for $df > 2$).

Connection to Text:

- **Maximum Likelihood Estimation:** This script demonstrates MLE, which is a special case of an extremum estimator.
- **Correctly Specified Likelihood:** The first part of the script shows MLE when the model is correctly specified.
- **Misspecified Likelihood:** The second part shows what happens when the likelihood is misspecified (assuming normality when the data is t-distributed). The example demonstrates that the MLE for the mean can still be consistent even with misspecification of the distributional form, as discussed on page 325.
- **Optimization:** The script uses the R function `optim()`. This is a numerical method for finding the values that minimizes the objective function (the negative log-likelihood in this case)

R Script 4: Quantile Regression

```
# Load necessary packages
library(quantreg) # For quantile regression
```

```
Loading required package: SparseM
```

```
# library(tidyverse) # If not already loaded

# Simulate data for quantile regression (Example 23.3)
n <- 1000
beta0 <- 1
beta1 <- 2
x <- rnorm(n)
```

```
u <- rnorm(n) # Homoscedastic errors
y <- beta0 + beta1 * x + u

# Quantile regression for different quantiles
tau_values <- c(0.1, 0.25, 0.5, 0.75, 0.9) # Quantiles of interest

# Store results
results <- list()

for (tau in tau_values) {
  # rq() performs quantile regression
  rq_fit <- rq(y ~ x, tau = tau)
  results[[as.character(tau)]] <- coef(rq_fit)
}

# Print the estimated coefficients for each quantile
print(results)
```

```
$`0.1`
(Intercept)           x
 -0.2879848   2.0140112

$`0.25`
(Intercept)           x
  0.3539796   1.9972320

$`0.5`
(Intercept)           x
  0.9712894   2.0134440

$`0.75`
(Intercept)           x
   1.664000    2.028558

$`0.9`
(Intercept)           x
   2.299093    2.081870
```

```
# True coefficients (for comparison - in this case, they are the same for all quantiles)
c(beta0, beta1)
```

```
[1] 1 2
```

```
# Heteroscedastic errors
u_het <- (1 + 0.5 * abs(x)) * rnorm(n)
y_het <- beta0 + beta1 * x + u_het

# Quantile regression with heteroscedastic errors
results_het <- list()

for (tau in tau_values) {
  rq_fit_het <- rq(y_het ~ x, tau = tau)
  results_het[[as.character(tau)]] <- coef(rq_fit_het)
}

print(results_het)
```

```
$`0.1`
(Intercept)           x
 -0.9632439   1.9441107

$`0.25`
(Intercept)           x
 0.05275925  1.96584052
```

```
$`0.5`
(Intercept)           x
  0.9845508   1.9546655

$`0.75`
(Intercept)           x
   2.030928    1.876352

$`0.9`
(Intercept)           x
   2.925622    1.934705
```

```
# Create a plot to visualize quantile regression
df_plot <- data.frame(x=x, y=y_het)

ggplot(df_plot, aes(x = x, y = y_het)) +
    geom_point() +
    geom_quantile(quantiles = tau_values, color = "red", size=1) +
    labs(title = "Quantile Regression with Heteroscedastic Errors",
         x = "x", y = "y") +
    theme_bw()
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
ℹ Please use `linewidth` instead.

Smoothing formula not specified. Using: y ~ x
```

Explanation:

1. **Load Packages:** Load `quantreg` for quantile regression.
2. **Simulate Data (Homoscedastic):** Simulate data from a linear model $y = \beta_0 + \beta_1 x + u$, where the errors u are homoscedastic (constant variance).
3. **Quantile Regression:** Perform quantile regression for different quantiles (`tau_values`) using the `rq()` function. We store the estimated coefficients for each quantile.
4. **Print Results (Homoscedastic):** Print the estimated coefficients for each quantile. With homoscedastic errors, the slope coefficients should be similar across quantiles.
5. **Simulate Data (Heteroscedastic):** Simulate data with heteroscedastic errors, where the variance of u depends on x.
6. **Quantile Regression (Heteroscedastic):** Perform quantile regression again with the heteroscedastic data.
7. **Print Results (Heteroscedastic):** Print the estimated coefficients. With heteroscedastic errors, the slope coefficients will generally be different across quantiles.
8. **Visualize**: The `ggplot2` code creates a scatter plot and overlays the estimated quantile regression lines. With heteroscedasticity, you'll see that the lines are not parallel, reflecting the changing conditional distribution of y given x.

Connection to Text:

- **Quantile Regression:** This script directly implements quantile regression as described in Section 23.3.
- **`rq()` Function:** The `rq()` function from the `quantreg` package performs the quantile regression estimation.
- **Heteroscedasticity:** The script demonstrates how quantile regression can be used to model the conditional distribution of y given x even when the errors are heteroscedastic. The coefficients obtained for the different quantiles will be different. This is discussed in the context of mispecification. The conditional mean model may be correct but not the assumptions about the error term.
- **Visualization:** The visualization helps to see the difference between OLS regression and quantile regression. Quantile regression captures how the *entire conditional distribution* changes with x, not just the conditional mean.

R Script 5: Consistency of Extremum Estimators

```r
# Load necessary packages
# library(tidyverse)

# Simulate data from a simple linear model
n_values <- c(10, 50, 100, 500, 1000, 5000)  # Different sample sizes
beta_true <- 2
results_ols <- numeric(length(n_values))

for (i in seq_along(n_values)) {
  n <- n_values[i]
  x <- rnorm(n)
  u <- rnorm(n)
  y <- beta_true * x + u

  # OLS estimation (an extremum estimator)
  lm_fit <- lm(y ~ x)
  results_ols[i] <- coef(lm_fit)[2] # Store the estimated slope coefficient
}

# Create a data frame for plotting
df_results <- data.frame(n = n_values, beta_hat = results_ols)

# Plot the estimated coefficients against sample size
ggplot(df_results, aes(x = n, y = beta_hat)) +
  geom_point() +
  geom_hline(yintercept = beta_true, linetype = "dashed", color = "red") +
  labs(title = "Consistency of OLS Estimator",
       x = "Sample Size (n)",
       y = "Estimated Slope Coefficient (beta_hat)") +
    scale_x_continuous(breaks=n_values) +
  theme_bw()
```

```r
# Example with a non-linear objective function.
# We will use an estimator that minimizes the sum of 4th powers of errors.
results_m <- numeric(length(n_values))
obj_fun <- function(beta, x, y) {
    sum((y-beta*x)^4)
}
for (i in seq_along(n_values)) {
  n <- n_values[i]
  x <- rnorm(n)
  u <- rnorm(n)
  y <- beta_true * x + u

  # M-estimation (minimizing sum of 4th powers of errors)
  m_fit <- optim(par=0, fn=obj_fun, x=x, y=y, method="Brent", lower=-10, upper=10)
  results_m[i] <- m_fit$par # Store the estimated slope coefficient
}

df_results_m <- data.frame(n = n_values, beta_hat = results_m)
ggplot(df_results_m, aes(x = n, y = beta_hat)) +
  geom_point() +
  geom_hline(yintercept = beta_true, linetype = "dashed", color = "red") +
  labs(title = "Consistency of M-Estimator",
       x = "Sample Size (n)",
       y = "Estimated Slope Coefficient (beta_hat)") +
    scale_x_continuous(breaks=n_values) +
  theme_bw()
```

Explanation:

1. **Load Packages:** We might not need additional packages beyond what's been used before.
2. **Simulate Data:** Simulate data from a simple linear model $y = \beta x + u$ for different sample sizes (`n_values`).
3. **OLS Estimation:** For each sample size, estimate the slope coefficient β using OLS (which is an extremum estimator, minimizing the sum of squared residuals).
4. **Store Results:** Store the estimated slope coefficients in `results_ols`.
5. **Create Data Frame:** Create a data frame to hold the sample sizes and corresponding estimated coefficients.
6. **Plot Results:** Plot the estimated coefficients against the sample size. You should see that as the sample size increases, the estimated coefficients converge to the true value of β (which is 2 in this simulation). This illustrates **consistency**.
7. **M-estimation:** Repeat the simulation, but now instead of OLS, estimate the parameter by minimizing the sum of the 4th powers of the error. This shows consistency of a different extremum estimator.

Connection to Text:

- **Extremum Estimators:** This script uses OLS as an example of an extremum estimator.
- **Consistency:** The script demonstrates the concept of consistency (Theorem 23.1), which is a key property of extremum estimators. As the sample size grows, the estimator converges in probability to the true parameter value.
- **Simulation:** Simulation is a powerful tool for understanding asymptotic properties like consistency. We can visually see how the estimator behaves as the sample size increases.
- **Non-linear objective function:** The second example demonstrates the consistency when using a different objective function (sum of 4th power of errors).

YouTube Videos on GMM and Extremum Estimators

Here are some YouTube videos that explain concepts related to the attached text, along with verification of their availability and explanations of their relevance:

1. Generalized Method of Moments (GMM) - Introduction

- **Title:** "11.1 Generalized Method of Moments (GMM): Introduction"
- **Channel:** Ben Lambert
- **Link:** https://www.youtube.com/watch?v=GMVh02P-Oug
- **Availability:** Verified (October 26, 2023)
- **Relevance:** This video provides a good introduction to GMM, explaining the basic idea of using moment conditions to estimate parameters. It connects directly to **Section 23.1** of the text. It introduces the concept in a clear, intuitive manner, starting with simpler examples before building up to the general framework. Covers the unidentified, exactly identified and overidentified cases.

2. GMM - Intuition

- **Title:** "11.2 Generalized Method of Moments (GMM): Intuition"
- **Channel:** Ben Lambert
- **Link:** https://www.youtube.com/watch?v=3FzKZv4s8O8
- **Availability:** Verified (October 26, 2023)
- **Relevance:** This video builds intuition for GMM, explaining why it works and how the weighting matrix is involved. It helps understand the underlying principles behind minimizing the sample moment conditions, as described in **Section 23.1**. The intuition behind identification is presented.

3. GMM - Example

- **Title:** "11.3 Generalized Method of Moments (GMM): Example"
- **Channel:** Ben Lambert

- **Link:** https://www.youtube.com/watch?v=YS-ojO-t8aM
- **Availability:** Verified (October 26, 2023)
- **Relevance**: This video goes through a specific example using GMM. It closely relates to **Example 23.8** and **Example 23.9**, showing a practical application of the GMM estimator in context that are very similar to linear regression and instrumental variables setting.

4. GMM Derivation of Asymptotic Distribution

- **Title:** "11.6 Generalized Method of Moments (GMM): Asymptotic properties"
- **Channel:** Ben Lambert
- **Link:** https://www.youtube.com/watch?v=jPAwA-qiL4Y
- **Availability:** Verified (October 26, 2023)
- **Relevance:** This video covers the asymptotic properties of the GMM estimator, including consistency and asymptotic normality. It aligns with **Section 23.2**, particularly **Theorem 23.4**, although it focuses specifically on GMM rather than general extremum estimators.

5. Introduction to Maximum Likelihood Estimation (MLE)

- **Title:** "Maximum Likelihood, clearly explained!!!"
- **Channel:** StatQuest with Josh Starmer
- **Link:** https://www.youtube.com/watch?v=XepXtl9YKwc
- **Availability:** Verified (October 26, 2023)
- **Relevance:** While the text focuses more on GMM, it also mentions Maximum Likelihood Estimation (MLE) as a special case of an extremum estimator. This video provides an excellent, intuitive introduction to MLE, explaining the core concepts in a visually clear way. This relates to the discussion of M-estimators and to the section on **Correctly Specified Likelihood** and **Misspecified Likelihood**, page 325.

6. Consistency, Bias and Efficiency of Estimators

- **Title:** "Estimators - Unbiasedness, Consistency, Efficiency (Intuition + Examples)"
- **Channel:** Shokoufeh Mirzaei
- **Link:** https://www.youtube.com/watch?v=_SpL2HDA_Cc&t=607s
- **Availability:** Verified (October 26, 2023)
- **Relevance:** This video helps build a solid foundation for **Section 23.2** by clearly explaining the concepts of **consistency**, bias, and efficiency of estimators. It is not specific to extremum estimators but introduces essential ideas for understanding their properties. It is a good primer before tackling Theorems 23.1, 23.3 and 23.4.

7. Quantile Regression

- **Title:** "Quantile Regression"
- **Channel:** Ben Jann
- **Link:** https://www.youtube.com/watch?v=gGz-T3qe5vg
- **Availability:** Verified (October 26, 2023)
- **Relevance:** This video gives a comprehensive introduction to **quantile regression**, covering the basic idea, estimation, and interpretation. This is a good source for better understanding **Section 23.3** of the provided material. It complements the text by providing visual examples.

These videos provide a combination of theoretical explanations, intuitive examples, and practical applications related to the core concepts in the provided text. They cover GMM, MLE, consistency, and quantile regression, and they use a mix of visual aids, derivations, and examples to enhance understanding.

Multiple Choice Exercises

MC Exercise 1

MC Solution 1

In the Generalized Method of Moments (GMM), what does the equation $\mathbb{E}[g(Z_i, \theta_0)] = 0$ represent?

a. The objective function to be minimized.
b. The asymptotic variance of the estimator.
c. The moment conditions that identify the parameters.
d. The weighting matrix used in estimation.

MC Exercise 2

MC Solution 2

Which of the following cases describes an **overidentified** GMM model?

a. The number of parameters (p) is greater than the number of moment conditions (q).
b. The number of parameters (p) is equal to the number of moment conditions (q).
c. The number of parameters (p) is less than the number of moment conditions (q).
d. The number of parameters (p) is unrelated to the number of moment conditions (q).

MC Exercise 3

MC Solution 3

In the simultaneous equations model $B(\theta)y_i = C(\theta)x_i + u_i$, what is required for the reduced form to exist?

a. $C(\theta)$ must be invertible.
b. $B(\theta)$ must be invertible.
c. u_i must be normally distributed.
d. x_i and u_i must be uncorrelated.

MC Exercise 4

MC Solution 4

In the Hansen and Singleton (1982) model (Example 23.2), what does γ represent?

a. The discount factor.
b. The coefficient of relative risk aversion.
c. The gross return on an asset.
d. The utility function.

MC Exercise 5

MC Solution 5

What is the role of the weighting matrix $W_n(\theta)$ in GMM?

a. It determines the asymptotic variance of the estimator, regardless of the moment conditions.
b. It determines the relative importance given to each moment condition in estimation.
c. It ensures that the estimator is unbiased.
d. It is only relevant in the exactly identified case.

MC Exercise 6

MC Solution 6

Which of the following is NOT an example of an extremum estimator?

a. Ordinary Least Squares (OLS)
b. Maximum Likelihood Estimation (MLE)
c. Generalized Method of Moments (GMM)
d. The sample median

MC Exercise 7

MC Solution 7

What does the **identification condition** for consistency of an extremum estimator require?

a. The objective function must be differentiable.
b. The limit of the objective function must have a unique global minimum at the true parameter value.
c. The sample size must be sufficiently large.
d. The data must be normally distributed.

MC Exercise 8

MC Solution 8

What does **uniform convergence** of $Q_n(\theta)$ to $Q(\theta)$ mean in the context of extremum estimators?

a. $Q_n(\theta)$ converges to $Q(\theta)$ pointwise for each θ.
b. $Q_n(\theta)$ converges to $Q(\theta)$ at the same rate for all θ.
c. $\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)|$ converges to zero.
d. $Q_n(\theta)$ and $Q(\theta)$ are both continuous functions.

MC Exercise 9

MC Solution 9

In the proof of Theorem 23.1 (Consistency), what is the role of term II, $Q_n(\hat{\theta}) - Q_n(\theta_0)$?

a. It converges to zero in probability due to uniform convergence.
b. It is always non-positive because $\hat{\theta}$ minimizes $Q_n(\theta)$.
c. It is equal to the asymptotic variance of the estimator.
d. It is used to establish the identification condition.

MC Exercise 10

MC Solution 10(#sec-ch23mcsolution10}

Which condition is NOT required for the consistency of an extremum estimator (Theorem 23.1)?

a. The parameter space Θ is compact.
b. $Q_n(\theta)$ is continuous in θ.
c. $Q_n(\theta)$ converges uniformly to $Q(\theta)$.
d. $Q_n(\theta)$ is differentiable.

MC Exercise 11

MC Solution 11(#sec-ch23mcsolution11}

In the context of Theorem 23.4 (Asymptotic Normality), what does the matrix A represent?

a. The weighting matrix.
b. The asymptotic variance of the estimator.
c. The limit of the second derivative of the objective function.
d. The score function.

MC Exercise 12

MC Solution 12(#sec-ch23mcsolution12}

Under correct specification of the likelihood, what is the relationship between the matrices A and B in Theorem 23.4?

a. $A = B^{-1}$
b. $A = B$
c. $A = 2B$
d. A and B are unrelated.

MC Exercise 13

MC Solution 13(#sec-ch23mcsolution13}

What is the typical form of the asymptotic variance of the GMM estimator in the overidentified case?

a. $\Gamma^{-1}\Omega\Gamma^{\top -1}$
b. $(\Gamma^\top W \Gamma)^{-1}$
c. $(\Gamma^\top W \Gamma)^{-1}\Gamma^\top W \Omega W \Gamma (\Gamma^\top W \Gamma)^{-1}$
d. W^{-1}

MC Exercise 14

MC Solution 14(#sec-ch23mcsolution14}

In Example 23.9 (Instrumental Variables), what condition is necessary for the IV estimator to be well-defined?

a. The number of instruments (J) must be less than the number of parameters (K).
b. The number of instruments (J) must be equal to the number of parameters (K).
c. The number of instruments (J) must be greater than or equal to the number of parameters (K).
d. The instruments must be perfectly correlated with the endogenous regressors.

MC Exercise 15

MC Solution 15(#sec-ch23mcsolution15}

What is the **check function**, $\rho_\alpha(u)$, used for in quantile regression?

a. To define the weighting matrix.
b. To define the objective function that is minimized.
c. To define the asymptotic variance of the estimator.
d. To define the moment conditions.

MC Exercise 16

MC Solution 16(#sec-ch23mcsolution16}

What is the **monotone equivariance property** of quantile regression?

a. Quantile regression is robust to outliers.
b. Quantile regression coefficients are always positive.
c. Applying a monotonic transformation to the response variable transforms the conditional quantiles in the same way.
d. The sum of the residuals in quantile regression is always zero.

MC Exercise 17

MC Solution 17(#sec-ch23mcsolution17}

In the Glivenko-Cantelli theorem (Theorem 23.2), what does $F_n(x)$ represent? a) The probability density function. b) The true cumulative distribution function. c) The empirical cumulative distribution function. d) The quantile function

MC Exercise 18

MC Solution 18(#sec-ch23mcsolution18}

In a **misspecified likelihood** setting, what can be said about the MLE of the parameters of interest?

a. It is always inconsistent.
b. It is always unbiased.
c. It may be consistent under certain conditions, such as correct specification of the conditional mean.
d. It is asymptotically normally distributed.

MC Exercise 19

MC Solution 19 $Q(\theta_0) - Q(\theta) < 0$ for all $\theta \neq \theta_0$. b) $Q(\theta_0) - Q(\theta) = 0$ for all $\theta \neq \theta_0$. c) $Q(\theta_0) - Q(\theta) > 0$ for all $\theta \neq \theta_0$. d) $Q(\theta_0) - Q(\theta)$ can be positive or negative.

MC Exercise 20

MC Solution 20 Neighborhoods of θ_0 of fixed size. b) Neighborhoods of θ_0 whose size goes to infinity with n. c) Neighborhoods of θ_0 whose size decreases to zero as n goes to infinity. d) The entire parameter space Θ.

Multiple Choice Solutions

MC Solution 1

MC Exercise 1

The correct answer is (c).

The equation $\mathbb{E}[g(Z_i, \theta_0)] = 0$ represents the **moment conditions** that identify the parameters in GMM. These are the population moment conditions that the GMM estimator seeks to match using sample moments. This is the foundation of GMM, as explained in **Section 23.1**.

MC Solution 2

MC Exercise 2

The correct answer is (c).

An **overidentified** GMM model has more moment conditions (q) than parameters (p). This means we have more information than is strictly necessary to identify the parameters, which allows for testing the validity of the moment conditions. The different cases are described on page 313.

MC Solution 3

MC Exercise 3

The correct answer is (b).

The reduced form is obtained by solving for y_i in terms of x_i and u_i. This requires multiplying both sides of the equation by $B(\theta)^{-1}$, which is only possible if $B(\theta)$ is invertible. See **Example 23.1**.

MC Solution 4

MC Exercise 4

The correct answer is (b).

In the Hansen and Singleton (1982) model, γ represents the **coefficient of relative risk aversion**. This parameter determines the curvature of the utility function and reflects the agent's aversion to risk. See **Example 23.2**.

MC Solution 5

MC Exercise 5

The correct answer is (b).

The weighting matrix $W_n(\theta)$ determines the **relative importance** given to each moment condition in forming the GMM estimator. Different choices of $W_n(\theta)$ lead to different GMM estimators, with the optimal weighting matrix (inverse of the asymptotic variance of the sample moments) leading to the most efficient estimator. This is discussed on page 315.

MC Solution 6

MC Exercise 6

The correct answer is (d).

The sample median is an estimator, but not necessarily an *extremum* estimator in the general sense described in the text. While the sample median *does* minimize the sum of absolute deviations, it is not generally presented within the framework of minimizing a smooth objective function. OLS, MLE, and GMM are all found by minimizing or maximizing a specific objective function.

MC Solution 7

MC Exercise 7

The correct answer is (b).

The **identification condition** requires that the limit of the objective function, $Q(\theta)$, has a **unique global minimum** at the true parameter value θ_0. This ensures that we can distinguish the true parameter value from other values based on the objective function. This concept is defined on page 316.

MC Solution 8

MC Exercise 8

The correct answer is (c).

Uniform convergence means that the *maximum* difference between $Q_n(\theta)$ and $Q(\theta)$ over the entire parameter space Θ converges to zero. This is a stronger condition than pointwise convergence (which only requires convergence for each fixed θ). This is defined on page 317.

MC Solution 9

MC Exercise 9

The correct answer is (b).

Term II, $Q_n(\hat{\theta}) - Q_n(\theta_0)$, is **always non-positive** because $\hat{\theta}$ is defined as the value that *minimizes* $Q_n(\theta)$. Therefore, $Q_n(\theta)$ evaluated at $\hat{\theta}$ cannot be greater than $Q_n(\theta)$ evaluated at any other point, including θ_0. This is explained in the proof of Theorem 23.1 on page 317.

MC Solution 10

MC Exercise 10

The correct answer is (d).

While differentiability is often assumed for convenience and for deriving asymptotic normality, it is *not* strictly required for consistency. The consistency theorem (Theorem 23.1) only requires continuity of $Q_n(\theta)$, not differentiability.

MC Solution 11

MC Exercise 11

The correct answer is (c).

The matrix A represents the **limit of the second derivative** (Hessian) of the objective function $Q_n(\theta)$. This is used in deriving the asymptotic normality of the estimator, as it relates to the curvature of the objective function around the true parameter value. This is defined in condition (D) of **Theorem 23.4**.

MC Solution 12

MC Exercise 12

The correct answer is (b).

Under correct specification of the likelihood, the **information matrix equality** holds. This states that the negative expected value of the Hessian matrix (which corresponds to A) is equal to the variance of the score function (which corresponds to B). Therefore, $A = B$. This is discussed on page 325.

MC Solution 13

MC Exercise 13

The correct answer is (c).

The typical form of the asymptotic variance of the GMM estimator in the overidentified case is $(\Gamma^\top W \Gamma)^{-1}\Gamma^\top W \Omega W \Gamma (\Gamma^\top W \Gamma)^{-1}$, where Γ is the expected derivative of the moment conditions, W is the weighting matrix, and Ω is the variance of the moment conditions. This is the "sandwich" form, and is shown on page 326.

MC Solution 14

MC Exercise 14

The correct answer is (c).

For the IV estimator to be well-defined, the number of instruments (J) must be **greater than or equal to** the number of parameters (K). If $J < K$, the model is underidentified. If $J = K$ (exact identification) we have a special case. If $J > K$ we have overidentification, as discussed on page 326.

MC Solution 15

MC Exercise 15

The correct answer is (b).

The **check function**, $\rho_\alpha(u)$, is used to define the **objective function** that is minimized in quantile regression. The objective function is the sum of the check function applied to the residuals. This is defined on page 327.

MC Solution 16

MC Exercise 16

The correct answer is (c).

The **monotone equivariance property** states that if we apply a strictly increasing (monotonic) transformation to the response variable, the conditional quantiles of the transformed variable are equal to the transformed conditional quantiles of the original variable. This is a key property of quantile regression and it is defined on page 329.

MC Solution 17

MC Exercise 17

The correct answer is (c). In the Glivenko-Cantelli theorem, $F_n(x)$ represents the **empirical cumulative distribution function**. This theorem states the uniform convergence of $F_n(x)$ to the true cumulative distribution function $F(x)$.

MC Solution 18

MC Exercise 18

The correct answer is (c).

In a **misspecified likelihood** setting, the MLE may still be consistent under certain conditions. For example, in a regression model, if the conditional mean is correctly specified, the MLE (or OLS) estimator for the coefficients may still be consistent, even if the distributional assumption about the errors is incorrect. This is mentioned on page 325.

MC Solution 19

MC Exercise 19

The correct answer is (c).

Lemma 23.1 states that $Q(\theta_0) - Q(\theta) = \mathbb{E}\left[-\ln \frac{f(X|\theta)}{f(X|\theta_0)}\right]$. By Jensen's inequality, $\mathbb{E}\left[-\ln \frac{f(X|\theta)}{f(X|\theta_0)}\right] > -\ln \mathbb{E}\left[\frac{f(X|\theta)}{f(X|\theta_0)}\right] = -\ln(1) = 0$, for all $\theta \neq \theta_0$.

MC Solution 20

MC Exercise 20

The correct answer is (c).

Shrinking neighborhoods refer to neighborhoods of the true parameter value θ_0 whose size **decreases to zero** as the sample size n goes to infinity. This concept is used in the context of local analysis of extremum estimators, where we focus on the behavior of the objective function close to θ_0 as n increases. The discussion appears on page 323.

Author: Peter Fuleky