ECON5128 MRes Econometrics 2

Weeks 4-5: Non-Linear Moment Conditions

Santiago Montoya-Blandón

February 4 & 11, 2025

Lecture outline

Roadmap to estimator properties

When encountering or developing a new estimator, the following is a useful roadmap to analysing its properties:

- **Setup / Model:** economic model and available data
- **Estimand:** object or feature of interest from model
- **Identification:** could estimand be recovered from data? What assumptions are required?
- **Estimation:** calculate the identified estimand from data sample
- **Inference:** attach uncertainty to estimates and infer (learn) about the population from the sample and the assumed model

Properties of estimators: Example OLS with linear model

1. Model: normal linear regression model $y = X\beta + u$, $u \sim \mathcal{N}(0, \sigma^2 I_n)$
2. Estimands:
   - Coefficients: β
   - Error variance: σ^2
   - Linear prediction: $\hat{y}(X) = X\beta$
3. Identification: iid + Invertibility of $\mathbb{E}[x_i x_i^\top] \implies \beta = \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i y_i]$
4. Estimation:
   - Coefficients: $\hat{\beta} = (X^\top X)^{-1} X^\top y$
   - Error variance: $\hat{\sigma}^2 = (y - X\hat{\beta})^\top (y - X\hat{\beta})/(n - K)$
   - Linear prediction: $\hat{y}(X) = X\hat{\beta}$
   - MM, GMM and ML all yield equivalent estimators of identified quantities
5. Inference:
   - Normally distributed coefficients (asymptotically if u non-Gaussian)
   - T-test allows us to learn "causal" effects: $H_0 : \beta_j = 0$ vs $H_1 : \beta_j \neq 0$

Setup and economic model

- Description of economic phenomena of interest and economic model used to understand this phenomena
- Description can vary in its level of detail:
  - Likelihood (full generative model)
  - Moment condition (estimating equations)
  - Fully non-parametric
- Description should also touch on available data
- Initial assumptions sufficient to describe the setup of interest
- Examples:
  - Linear model for wage determination
  - Vector Autoregression (VAR) for macroeconomic time series
  - Potential outcomes of policy treatment

Estimand

- Object or quantity of interest from the analysis
- Defined only with respect to the economic model of interest
- Can be a quantity or a function
- Examples:
  - Slope coefficients: β
  - Linear prediction: $\hat{y}(X) = X\beta$
  - Average treatment effect: $\mathbb{E}[y_i(1) - y_i(0)]$
  - Average treatment on the treated effect: $\mathbb{E}[y_i(1) - y_i(0)|D_i = 1]$
  - Impulse Response Function in VARs
  - Demand function

Identification

- Hard to give a concrete definition as many exist (see Lewbel, 2019)
  - Recoverability of estimand from data given the setup
  - Observational equivalence
  - Sources of variation
- Identification asks: is there some rule that maps *observable features* to *estimands*?
- In principle, no presumption of uniqueness:
  - Identified but not uniquely: *set* or *partial identification*
  - Identified uniquely: *(point) identification*
- No presumption of constructivity or computability either:
  - Identification shown to exist but rule is unknown (non-constructive identification)
  - Rule does not need to be easy to calculate in a computer
- Identification crucially hinges on additional assumptions

Estimation

- Numerically obtaining a value for the estimand given a sample of data
- *Estimators are algorithms*: input data, output a value of the estimands
- Value of an estimator computed on a given sample are called *estimates*
- Estimators should at least be consistent: estimator converges to truth, i.e., $\hat{\theta} \xrightarrow{p} \theta_0$
- Important to distinguish from identification:
  - Identification without estimation: $\beta = \mathbb{E}[x_i x_i^\top]^{-1} \mathbb{E}[x_i y_i]$ vs $\hat{\beta}_{\text{OLS}} = (X^\top X)^{-1} X^\top y$ with multicollinearity
  - Estimation without identification: sample average $\bar{y}$ of Cauchy random variables $y_1, \ldots, y_n \sim \text{Cauchy}(0, \gamma)$
- Examples: OLS, GMM, maximum likelihood

Inference

- Estimates for any given sample carry uncertainty (e.g., standard errors)
- Uncertainty impacts how easily we can learn from the data
- In broader context, inference means to infer (learn) from the sample to draw conclusions about the population
- Inference for econometricians usually means to derive and estimate the uncertainty for an estimator
- Our estimators are generally asymptotically normally distributed, so inference sometimes narrowly refers to obtaining the mean and variance of this normal distribution
- Confidence intervals and hypothesis testing are basic forms of inference
- "Feasible" (i.e., computable) inference sometimes requires additional estimation steps as uncertainty might depend on other unknown quantities in the model

Proofs for linear GMM

- Setup: linear model $y = X\theta + u$ with endogeneity $\mathbb{E}[X^\top u] \neq 0$
- Estimand: model parameters θ and functions of them $\psi(\theta)$
- Identification:
  - Assumptions on iid sampling and moment condition $\mathbb{E}[Z^\top(y - X\theta_0)] = 0$
  - Identification holds when moments are exactly zero only at θ_0
  - Condition implied by order $(\ell \geq p)$ and rank assumptions $\text{rank}(\mathbb{E}[z_i x_i^\top]) = p$
- Estimation: Minimize quadratic on sample moment conditions
- Consistency:
  - Plug-in linear model into estimator, take probability limits (using a WLLN)
  - Need rank condition and population moment assumption as well
- Inference:
  - Moment conditions are asymptotically normal themselves (using a CLT)
  - Estimator can be written as linear function of moment conditions
  - Linear combination of normals is normal, finalizing proof

Lecture outline

Setup

- We now turn to providing a similar roadmap to understanding the GMM estimator when moment conditions can be *non-linear*
- Special emphasis needs to be placed on identification
- Estimation proceeds just as before, though properties are more difficult to analyse due to non-linearities
- Will deal with dependent data (e.g., time series) for generality
- Will carefully introduce all relevant assumptions as we need them

Setup

- The population moment condition involves a function $f(v_t, \theta)$ of observed random variable v_t and unknown parameters $\theta \in \Theta \subset \mathbb{R}^p$

Assumption 1 (Strict Stationarity)

The $r \times 1$ random vectors $\{v_t : -\infty < t < \infty\}$ form a strictly stationary process with sample space $\boldsymbol{V} \subseteq \mathbb{R}^r$.

- All expectations of functions of v_t are independent of time
- v_t includes all available observable data, e.g, $v_t = (y_t, x_t^\top, z_t^\top)^\top$
- As before, we denote Θ as the parameter space

Setup

Assumption 2 (Regularity Conditions for Moment Function)

The function $f : \boldsymbol{V} \times \Theta \to \mathbb{R}^{\ell}$, *where* $\ell < \infty$, *satisfies:*

(i) *it is continuous on* Θ *for each* v_t;

(ii) $\mathbb{E}[f(v_t, \theta)]$ *exists and is finite for every* $\theta \in \Theta$;

(iii) $\mathbb{E}[f(v_t, \theta)]$ *is continuous on* Θ.

Assumption 3 (Population Moment Condition)

There exists some parameter vector θ_0 *that satisfies the* $(\ell \times 1)$ *population moment condition:* $\mathbb{E}[f(v_t, \theta_0)] = 0$ *for almost all data vectors* v_t.

Setup

- The population moment condition can only be used as a basis for estimation if it provides enough information to uniquely identify the parameter vector θ_0
- In the linear model, we could relate parameter identification to a simple condition only depending on data
- In nonlinear models, the situation is more complicated
- Identification can fail due to the properties of the data, v_t, or due to the properties of f as function of θ or an interaction of the two
- To characterize how these types of failure can occur in nonlinear models, it is necessary to distinguish between global and local identification

Setup

Assumption 4 (Global Identification)

$\mathbb{E}[f(v_t, \theta_0)] \neq 0$ *for all* $\theta \in \Theta$ *such that* $\theta \neq \theta_0$.

- The adjective "global" emphasizes that the population moment condition holds at only one value in the entire parameter space Θ
- This is the same identification concept as for linear GMM
- Within that context, it was possible to derive a convenient condition for global identification
- Unfortunately, this is rarely possible in nonlinear models
- General global identification can thus only be assumed

Local Identification

- It is typically impossible to find a useful alternative representation for $f(v_t, \theta)$ which holds (globally) over all $\theta \in \Theta$
- However, such a representation might be found if we restrict our attention to some suitably defined *neighbourhood* of θ_0
- *Local identification* means that we can only guarantee identification within this neighbourhood
- This is the price we pay due to nonlinearity, can only provide local identification conditions
- Global identification $\implies$ Local identification, but the converse is not generally true

Local Identification

- An ε-neighbourhood of θ_0 is defined to be the set $N_\varepsilon = \{\theta : \|\theta - \theta_0\| < \varepsilon\}$
- Under some conditions, f should behave nicely within the neighbourhood N_ε
- Then, we will first-order Taylor expand f over N_ε

Assumption 5 (Regularity Conditions on $\partial f(v_t, \theta)/\partial\theta^\top$)

(i) *The derivative matrix* $\partial f(v_t, \theta)/\partial\theta^\top$ *exists and is continuous on* Θ *for each* $v_t \in \boldsymbol{V}$;

(ii) θ_0 *is an interior point of* Θ;

(iii) $\mathbb{E}[\partial f(v_t, \theta_0)/\partial\theta^\top]$ *exists and is finite.*

Local Identification

- Condition for local identification derived by restricting attention to sufficiently small ε so that f equals the following first order Taylor series expansion around θ_0 in N_ε:

$$f(v_t, \theta) = f(v_t, \theta_0) + \{\partial f(v_t, \theta_0)/\partial \theta^\top\}(\theta - \theta_0)$$

- The advantage of this approach is that the above implies $f(v_t, \theta)$ is a linear function of $\theta - \theta_0$ in N_ε
- This local approximation provides insight into the necessary local identification condition
- Taking expectation on both sides, Assumption 3 (Population Moment Condition) and Assumption 5 (Regularity Conditions on derivative) imply

$$\mathbb{E}[f(v_t, \theta)] = \{\mathbb{E}[\partial f(v_t, \theta_0)/\partial \theta^\top]\}(\theta - \theta_0)$$

Local Identification

- Note form of local condition:

$$\mathbb{E}[f(v_t, \theta)] = \{\mathbb{E}[\partial f(v_t, \theta_0)/\partial \theta^\top]\}(\theta - \theta_0)$$

- Compare to global condition for linear IV regression:

$$\mathbb{E}[z_t u_t(\theta)] = \mathbb{E}[z_t x_t^\top](\theta_0 - \theta),$$

- In the linear world, we required $rank\, \mathbb{E}[z_t x_t^\top] = p$ for the identification
- Condition for local identification under non-linearity:

Assumption 6 (Local Identification)

$rank\{\mathbb{E}[\partial f(v_t, \theta_0)/\partial \theta^\top]\} = p$

Lecture outline

Estimation

As in the linear IV regression case, the GMM estimator is defined as

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} Q_T(\theta),$$

where

$$Q_T(\theta) = \left\{ T^{-1} \sum_{t=1}^{T} f(v_t, \theta) \right\}^{\top} W_T \left\{ T^{-1} \sum_{t=1}^{T} f(v_t, \theta) \right\}.$$

Assumption 7 (Properties of the Weighting Matrix)

W_T is a positive semi-definite matrix which converges in probability to the positive definite matrix of constants W.

Estimation

If Assumption 5 (Regularity Conditions on $\partial f(v_t, \theta)/\partial \theta^\top$) holds, then the first order conditions for this minimization imply $\partial Q_T(\hat{\theta}_T)/\partial \theta = 0$, which implies

$$0 = \left\{ T^{-1} \sum_{t=1}^{T} \frac{\partial f(v_t, \hat{\theta}_T)}{\partial \theta^\top} \right\}^\top W_T \left\{ T^{-1} \sum_{t=1}^{T} f(v_t, \hat{\theta}_T) \right\}.$$

- In the linear IV model, these conditions could be solved to obtain a closed form solution for $\hat{\theta}_T$ as a function of the data
- In nonlinear models this is typically impossible
- Fortunately, the development of numerical techniques and advance in computers allow us to numerically tackle the optimization

Lecture outline

Asymptotic Properties

- In the linear model, asymptotic analysis rested crucially on a closed form expression for $\hat{\theta}_T$
- However, as discussed, such a representation typically does not exist in nonlinear models
- Necessary to develop a different proof strategy
- As it turns out, the difference is most marked in the proof of consistency
- Once consistency is established, then it is possible to invoke the Mean Value Theorem to obtain a representation for $\hat{\theta}_T - \theta_0$
- Such representation facilitates the derivation of asymptotic normality using similar arguments as in the linear case

Asymptotic Properties

- Before developing the asymptotic analysis it is necessary to place a further restriction on v_t

Assumption 8 (Ergodicity)

The random process $\{v_t : -\infty < t < \infty\}$ *is ergodic.*

- A formal definition of Ergodicity is beyond this course
- See Davidson (1994) for formal treatment
- Necessary to control amount of dependence in data
- A stronger condition is called Mixing: the dependence between v_t and v_{t-m} decreases as $m \to \infty$

Consistency

The GMM estimator does not admit a closed form, but is defined as

$$\hat{\theta}_T = \arg\min_{\theta \in \Theta} Q_T(\theta),$$

where

$$Q_T(\theta) = \left\{ T^{-1} \sum_{t=1}^{T} f(v_t, \theta) \right\}^\top W_T \left\{ T^{-1} \sum_{t=1}^{T} f(v_t, \theta) \right\}.$$

A key is to consider the population analogue of the objective function

$$Q_0(\theta) = \left\{ E[f(v_t, \theta)] \right\}^\top W \left\{ E[f(v_t, \theta)] \right\}$$

Consistency

- The population moment condition (Assumption 3) implies $Q_0(\theta) = 0$
- The global identification condition (Assumption 4) and the positive definiteness of W (Assumption 7) imply $Q_0(\theta) > 0$ for all $\theta \neq \theta_0$
- Taken together these two properties imply $Q_0(\theta)$ has a unique minimum at $\theta = \theta_0$
- $\hat{\theta}_T$ should converge in probability to θ_0:
  - (i) $\hat{\theta}_T$ minimizes $Q_T(\theta)$ and θ_0 minimizes $Q_0(\theta)$
  - (ii) $Q_T(\theta)$ converges in probability to function $Q_0(\theta)$
  - (iii) Require conditions so that the sequence of minimizers convergences to the minimizer of the limit

Consistency

- Mathematically, it is not necessarily the case that the minimum of a sequence of functions converges to the minimum of the limit of the sequence of functions
- A sufficient condition is for $Q_T(\theta)$ to converge *uniformly* to $Q_0(\theta)$
- We require two additional technical conditions

Assumption 9 (Compactness of Θ)

Θ *is a compact set.*

Assumption 10 (Domination of $f(v_t, \theta)$)

$\mathbb{E}[\sup_{\theta \in \Theta} \|f(v_t, \theta)\|] < \infty$

Consistency

- In Euclidean spaces, a set is compact when it is closed and bounded
- Compactness would then require knowledge of bounds on θ_0, which are typically unavailable
- In practice, one can reasonably assume the parameter space to be as large as necessary while still being closed and bounded
- Compactness is thus often ignored in practice due to being a mild assumption, though other proof strategies try to proceed without imposing it
- Domination requires $f(v_t, \theta)$ to be bounded by a function with finite expectation for all θ
- We mainly require it for applying a dominated convergence theorem (DCT) to obtain the required uniform convergence

Consistency

Lemma 1 (Uniform Convergence in Probability of $Q_T(\theta)$)

Under Assumptions 1 (Strict Stationarity of v_t), 2 (Regularity conditions of f), 7 (Weight Matrix), 8 (Ergodicity of v_t), 9 (Compactness of Θ), and 10 (Domination of f), then $Q_T(\theta)$ converges uniformly in probability to $Q_0(\theta)$, written as

$$\sup_{\theta \in \Theta} |Q_T(\theta) - Q_0(\theta)| \xrightarrow{p} 0$$

Theorem 2 (Consistency of the Parameter Estimator)

Under the assumptions for Lemma 1 and Assumptions 3 and 4,

$$\hat{\theta}_T \xrightarrow{p} \theta_0$$

Consistency (proof)

- We break out the proof into two parts
- First we will show that

$$\lim_{T\to\infty} P\left[0 \leq Q_0(\hat{\theta}_T) < \varepsilon\right] = 1 \text{ for any } \varepsilon > 0 \tag{1}$$

- The above says that $\hat{\theta}_T$ minimizes $Q_0(\theta)$ with probability one as $T \to \infty$
- The second part of the proof shows that (1) implies consistency
- As a consequence of Lemma 1, for any $\theta \in \Theta$ and any constant $\varepsilon > 0$

$$\lim_{T\to\infty} P[|Q_0(\hat{\theta}_T) - Q_T(\hat{\theta}_T)| < \varepsilon] = 1$$

Consistency (proof: part 1)

- Fix $\varepsilon > 0$. Applying Lemma 1 with $\hat{\theta}_T \in \Theta$:

$$\lim_{T\to\infty} P[Q_0(\hat{\theta}_T) < Q_T(\hat{\theta}_T) + \varepsilon/3] = 1 \tag{2}$$

- Applying Lemma 1 with $\theta_0 \in \Theta$:

$$\lim_{T\to\infty} P[Q_T(\theta_0) < Q_0(\theta_0) + \varepsilon/3] = 1 \tag{3}$$

- Since $\hat{\theta}_T$ minimizes $Q_T(\theta)$, we also have

$$\lim_{T\to\infty} P[Q_T(\hat{\theta}_T) < Q_T(\theta_0) + \varepsilon/3] = 1 \tag{4}$$

Consistency (proof: part 1)

- Equations (2) and (4) imply

$$\lim_{T\to\infty} P[Q_0(\hat{\theta}_T) < Q_T(\theta_0) + 2\varepsilon/3] = 1,$$

- Together with (3), we get

$$\lim_{T\to\infty} P[Q_0(\hat{\theta}_T) < Q_0(\theta_0) + \varepsilon] = 1.$$

- The positive definiteness of W implies $Q_0(\theta) \geq 0$ and the population moment condition gives $Q_0(\theta_0) = 0$
- Equation (1) then follows

Consistency (proof: part 2)

- We now show the previous condition is equivalent to consistency
- Let N_δ be a neighbourhood of θ_0, an open subset of Θ which contains θ_0
- Let $\bar{N}_\delta$ be the complement of N_δ relative to Θ
- $\bar{N}_\delta$ is a closed subset of a compact set Θ and so it is also compact
- Since $\bar{N}_\delta$ is compact and $Q_0(\theta)$ is a continuous function, it has an infimum on $\bar{N}_\delta$, denoted as $\underline{Q}_0 := \inf_{\theta \in \bar{N}_\delta} Q_0(\theta)$.
- By Assumption 4, as $\theta_0 \notin \bar{N}_\delta$, $Q_0(\theta) > 0$ for all $\theta \in \bar{N}_\delta$ so that this infimum is strictly positive: $\underline{Q}_0 > 0$

Consistency (proof: part 2)

- Therefore, plug in $\varepsilon = \underline{Q}_0 > 0$ in the above equation (1):

$$\lim_{T\to\infty} P\left[Q_0(\hat{\theta}_T) < \inf_{\theta\in\bar{N}_\delta} Q_0(\theta)\right] = 1$$

- As θ_0 minimizes $Q_0(\theta)$, this means $\lim_{T\to\infty} P[\hat{\theta}_T \notin \bar{N}_\delta] = 1$
- Hence, $\lim_{T\to\infty} P[\hat{\theta}_T \in N_\delta] = 1$
- Finally, the above argument holds for any choice of $\delta > 0$ no matter how small, so it must follow that $\lim_{T\to\infty} P[\hat{\theta}_T = \theta_0] = 1$
- In other words, $\hat{\theta}_T \xrightarrow{p} \theta_0$ as required

Lecture outline

Asymptotic normality

- To develop the asymptotic distribution of the estimator, we require an asymptotically valid closed form representation for $\sqrt{T}(\hat{\theta}_T - \theta_0)$.
- This representation comes from applying the Mean Value Theorem (MVT)
- MVT relates f to its first derivatives $\partial f(v_t, \theta)/\partial \theta^\top$, so it is necessary to impose Assumption 5
- Define $g_T(\theta) := T^{-1} \sum_{t=1}^{T} f(v_t, \theta)$ and $G_T(\theta) := T^{-1} \sum_{t=1}^{T} \partial f(v_t, \theta)/\partial \theta^\top$
- GMM objective can be written as: $Q_T(\theta) = g_T(\theta) W_T g_T(\theta)$
- First-order condition: $0 = \partial Q_T(\hat{\theta}_T)/\partial \theta = G_T(\hat{\theta}_T) W_T g_T(\hat{\theta}_T)$

Asymptotic normality

- MVT implies that

$$g_T(\hat{\theta}_T) = g_T(\theta_0) + G_T(\hat{\theta}_T, \theta_0, \lambda_T)(\hat{\theta}_T - \theta_0),$$

  where $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$ is the $(\ell \times p)$ matrix whose jth row is the corresponding row of $G_T(\bar{\theta}_T^{(j)})$ where $\bar{\theta}_T^{(j)} = \lambda_{T,j}\theta_0 + (1 - \lambda_{T,j})\hat{\theta}_T$ for some $\lambda_{T,j} \in [0, 1)$.
- Pre-multiply this by $G_T(\hat{\theta}_T)^\top W_T$ gives

$$G_T(\hat{\theta}_T)^\top W_T g_T(\hat{\theta}_T) =$$
$$G_T(\hat{\theta}_T)^\top W_T g_T(\theta_0) + G_T(\hat{\theta}_T)^\top W_T G_T(\hat{\theta}_T, \theta_0, \lambda_T)(\hat{\theta}_T - \theta_0)$$

Asymptotic normality

- The FOC implies that the left hand side is zero, which implies

$$\begin{aligned}
&\sqrt{T}(\hat{\theta}_T - \theta_0) \\
&= -\left[G_T(\hat{\theta}_T)^\top W_T G_T(\hat{\theta}_T, \theta_0, \lambda_T)\right]^{-1} G_T(\hat{\theta}_T)^\top W_T \sqrt{T} g_T(\theta_0) \\
&:= -\bar{M}_T \sqrt{T} g_T(\theta_0)
\end{aligned}$$

- Notice that this equation has the same basic structure as arose in the linear model at this stage: it is a product of
  - A random matrix $-\bar{M}_T$ converging to a constant matrix, with
  - a random vector $\sqrt{T} g_T(\theta_0)$, to which we can apply a CLT
- Just like in the linear case, we analyze the two terms separately

Asymptotic normality

- The asymptotic behavior of $\sqrt{T}g_T(\theta_0)$ will be given by the CLT
- To apply CLT, we need to assume the second order moment matrices of the sample moments satisfy certain restrictions

Assumption 11 (Properties of the Variance of the Sample Moment)

(i) $\mathbb{E}[f(v_t, \theta_0)f(v_t, \theta_0)^\top]$ *exists and is finite;*

(ii) $\lim_{T\to\infty} Var[\sqrt{T}g_T(\theta_0)] = S$ *exists and is a finite valued positive definite matrix.*

Theorem 3 (Central Limit Theorem for $\sqrt{T}g_T(\theta_0)$)

If Assumptions 1, 3, 8, and 11 hold, then $\sqrt{T}g_T(\theta_0) \xrightarrow{d} N(0, S)$.

Asymptotic normality

- The analysis of $\bar{M}_T$ is more complicated than in the linear model because it depends on $G_T(\hat{\theta}_T)$ and $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$.
- Since $\hat{\theta}_T \xrightarrow{p} \theta_0$ and $\bar{\theta}_T^{(j)}$ lies on the line between $\hat{\theta}_T$ and θ_0 for $j = 1, \ldots, p$.
- Intuition suggests that this should imply both $G_T(\hat{\theta}_T)$ and $G_T(\hat{\theta}_T, \theta_0, \lambda_T)$ converge in probability to $G_0 := \mathbb{E}[\partial f(v_t, \theta_0)/\partial \theta^\top]$.
- To formalize this intuition, we need the following two additional assumptions

Assumption 12

$\mathbb{E}[\partial f(v_t, \theta)/\partial \theta^\top]$ *is continuous on some neighbourhood* N_ε *of* θ_0.

Assumption 13 (Uniform convergence of $G_T(\theta)$)

$\sup_{\theta \in N_\varepsilon} ||G_T(\theta) - \mathbb{E}[\partial f(v_t, \theta)/\partial \theta^\top]|| \xrightarrow{p} 0$

Asymptotic normality

Lemma 4

Under Assumptions 1-5, 7-10, 12, and 13, we have $G_T(\hat{\theta}_T) \xrightarrow{p} G_0$ *and* $G_T(\hat{\theta}_T, \theta_0, \lambda_T) \xrightarrow{p} G_0$.

- Recall $\bar{M}_T = \left[G_T(\hat{\theta}_T)^\top W_T G_T(\hat{\theta}_T, \theta_0, \lambda_T) \right]^{-1} G_T(\hat{\theta}_T)^\top W_T$
- With the previous Lemma, combined with Assumption 7 and Slutsky's theorem, we can deduce that

$$\bar{M}_T \xrightarrow{p} (G_0^\top W G_0)^{-1} G_0^\top W$$

Asymptotic normality

Therefore, as in the linear model, $\sqrt{T}(\hat{\theta}_T - \theta_0)$ is asymptotically the product of

- a random matrix which converges in probability to a constant, and
- a random vector which converges to a normal distribution

Theorem 5 (Asymptotic Normality)

If Assumptions 1-5 and 7-13 hold, then

$$\sqrt{T}(\hat{\theta}_T - \theta_0) \xrightarrow{d} N(0, MSM^\top),$$

where $M = (G_0^\top W G_0)^{-1} G_0^\top W$.

Asymptotic normality

Theorem implies that an approximate $100(1-\alpha)\%$ confidence interval for $\theta_{0,j}$ in large samples is given by

$$\hat{\theta}_{j,T} \pm z_{\alpha/2}\sqrt{\hat{V}_{T,ii}/T},$$

where $\hat{V}_{T,ii}/T$ is the (j,j)th element of a consistent estimator of $MSM^\top$.

- As in the linear model, a natural candidate is based on consistent estimators of the component matrices M and S
- Notice that this time the matrix $\bar{M}_T$ cannot be used (even when it is consistent) as it depends on the unknown θ_0 values through $\bar{\theta}_T^{(j)}, j = 1, \ldots, p$
- Instead, replace $\bar{\theta}_T^{(j)}$ with $\hat{\theta}_{j,T}$ and using $\hat{M}_T = [G_T(\hat{\theta}_T)^\top W_T G_T(\hat{\theta}_T)]^{-1} G_T(\hat{\theta}_T)^\top W_T$ to estimate M
- However, the consistent estimation of S is more complicated

Lecture outline

1. General Concepts
2. Setup
3. Estimation
4. Consistency
5. Asymptotic Normality
6. Long-Run Covariance Matrix Estimation

Estimation of S

- So far, very little has been said about S except that it exists and is positive definite (Assumption 11)
- The latter is the matrix generalization of the requirement that a scalar variance be positive
- The estimator should also be at least positive semi-definite, otherwise the estimated variances of individual coefficients could be negative
- This is not always a trivial property to impose and is one aspect of the various estimators upon which we focus below

Estimation of S

- To understand more about the structure of S, it is useful to re-express it
- Define $f_t := f(v_t, \theta)$

$$
\begin{aligned}
S &= \lim_{T\to\infty} \operatorname{Var}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t\right] \\
&= \mathbb{E}\left[\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t - \mathbb{E}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t\right]\right)\left(\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t - \mathbb{E}\left[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t\right]\right)^{\top}\right]
\end{aligned}
$$

- Stationarity implies $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t - \mathbb{E}[\frac{1}{\sqrt{T}}\sum_{t=1}^{T} f_t] = \frac{1}{\sqrt{T}}\sum_{t=1}^{T}(f_t - \mathbb{E}[f_t])$

Estimation of S

- It follows that

$$
\begin{aligned}
S &= \lim_{T\to\infty} \mathbb{E}\left[\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(f_t - \mathbb{E}[f_t])\right\}\left\{\frac{1}{\sqrt{T}}\sum_{t=1}^{T}(f_t - \mathbb{E}[f_t])\right\}^\top\right] \\
&= \lim_{T\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T}\sum_{s=1}^{T}(f_t - \mathbb{E}[f_t])(f_s - \mathbb{E}[f_s])^\top\right]
\end{aligned}
$$

- The stationary assumption implies that for all t,

$$
\Gamma_j := \mathbb{E}(f_t - \mathbb{E}[f_t])(f_{t-j} - \mathbb{E}[f_{t-j}])^\top
$$

Estimation of S

- Therefore, we can finally write

$$S = \Gamma_0 + \lim_{T\to\infty}\left\{\sum_{t=1}^{T}\left(\frac{T-j}{T}\right)(\Gamma_j + \Gamma_j^\top)\right\}$$
$$= \Gamma_0 + \sum_{j=1}^{\infty}(\Gamma_j + \Gamma_j^\top)$$

- The matrix Γ_j is known as the jth autocovariance matrix of f_t
- Estimation of S is going to require assumptions about these autocovariance matrices

Serially uncorrelated sequences

- If f_t is a serially uncorrelated sequence then $\Gamma_j = 0$ for $j \neq 0$
- In this case, it follows that $S = S_{SU} = E[f_t f_t^\top]$
- The form of S_{SU} is essentially the same as the S matrix in the linear IV model under iid sampling:

$$\lim_{T\to\infty} T^{-1} \sum_{t=1}^{T} \sum_{s=1}^{T} \mathbb{E}[u_t u_s z_t z_s^\top].$$

- White (1994) showed that the following is consistent for S_{SU}

$$\hat{S}_{SU} = T^{-1} \sum_{t=1}^{T} f(v_t, \hat{\theta}_T) f(v_t, \hat{\theta}_T)^\top$$

- Under less restrictive dependence assumptions, can use Newey-West or other Heteroskedasticity and Autocorrelation Consistent (HAC) estimators