ECON5128 MRes Econometrics 2

Week 3: GMM with Linear Moment Conditions

Santiago Montoya-Blandón

January 28, 2024

General Idea

1. Motivation
2. Quick review of instrumental variable regression
3. Moment Conditions
4. Method of Moments (MM)
5. Generalized Method of Moments (GMM)
6. GMM Estimator and a Fundamental Decomposition
7. Asymptotic Properties
8. Optimal Choice of Weighting Matrix
9. Overidentifying restrictions

Lecture outline

General Idea

- Likelihood-based analysis requires a full specification of the data-generating process (DGP) for data $y = (y_1, \ldots, y_n)$ based on parameters θ
- A *model* formally corresponds to $\mathcal{F} = \{f(y \mid \theta) : \theta \in \Theta \subset \mathbb{R}^p\}$, a family of distributions indexed by p-dimensional parameters θ
- We will assume dimension of parameters is *small* ($p \ll n$)
- Asymptotic efficiency of ML estimator $\widehat{\theta}_{\text{MLE}} := \arg\max_{\theta \in \Theta} f(y \mid \theta)$ requires *correct specification* of the likelihood
  - Intuitively: chosen model captures all features of the data, true DGP $h(y)$ should be contained in specified family $\mathcal{F}$
  - Formally: there exists "*true value*" $\theta_0 \in \Theta$ such that $f(y \mid \theta_0) = h(y)$

General Idea

- If there is much uncertainty on the distributional form, it may be preferable to apply an estimation technique that assumes less structure on the DGP
- *Generalized method of moments* (GMM) is one such alternative technique
- Estimator derived from minimal assumptions that the model should satisfy, so-called *moment conditions*
- Economic theory drives the model selection: start from economic theory $\Longrightarrow$ moment conditions for the model
- Lecture based on Ch. 13 of Hansen's "Econometrics" textbook

Lecture outline

1. Motivation
2. Quick review of instrumental variable regression
3. Moment Conditions
4. Method of Moments (MM)
5. Generalized Method of Moments (GMM)
6. GMM Estimator and a Fundamental Decomposition
7. Asymptotic Properties
8. Optimal Choice of Weighting Matrix
9. Overidentifying restrictions

Endogeneity and IV regression

- Suppose we have a simple regression

$$y_i = \alpha + \beta X_i + \epsilon_i$$

where X is endogeneous: $\mathbb{E}[X\epsilon] \neq 0 \implies \text{Cov}(X, \epsilon) \neq 0$

- $\widehat{\beta}_{OLS}$ is inconsistent under endogeneity:

$$\widehat{\beta}_{OLS} = \left(X^\top X\right)^{-1} X^\top y \xrightarrow{p} \frac{\text{Cov}(X, y)}{\text{Var}(X)}$$
$$= \frac{\text{Cov}(X, \alpha + \beta X + \epsilon)}{\text{Var}(X)} = \beta + \frac{\text{Cov}(X, \epsilon)}{\text{Var}(X)} \neq \beta$$

Endogeneity and IV regression

- An instrumental variable Z needs to satisfy:
  1. Relevance: $\text{Cov}(Z, X) \neq 0$
  2. Exclusion restriction: $\text{Cov}(Z, \epsilon) = 0$

$$\text{Cov}(y, Z) = \beta \, \text{Cov}(X, Z) + \text{Cov}(\epsilon, Z)$$
$$\implies \beta = \frac{\text{Cov}(Z, y)}{\text{Cov}(Z, X)}$$

- Consistent instrumental variable (IV) estimator $\widehat{\beta}_{IV}$ is

$$\widehat{\beta}_{IV} = \left(Z^\top X\right)^{-1} Z^\top y \xrightarrow{p} \frac{\text{Cov}(Z, y)}{\text{Cov}(Z, X)} = \beta$$

Lecture outline

1. Motivation
2. Quick review of instrumental variable regression
3. Moment Conditions
4. Method of Moments (MM)
5. Generalized Method of Moments (GMM)
6. GMM Estimator and a Fundamental Decomposition
7. Asymptotic Properties
8. Optimal Choice of Weighting Matrix
9. Overidentifying restrictions

Moments Conditions

- Typically, a *moment condition* is defined as

$$\mathbb{E}\left[f(\omega_i, z_i, \theta_0)\right] = 0 \tag{1}$$

  - θ is the $p \times 1$ vector parameters
  - $f(\cdot)$ is an ℓ-dimensional vector-valued (potentially non-linear) function, also known as *moment function*
  - ω_i is a vector of variables appearing in the model; e.g., $\omega_i = (y_i, x_i)$
  - z_i is a vector of instruments
- For simplicity in notation, we define $g(\theta) := \mathbb{E}\left[f(\omega_i, z_i, \theta)\right]$
- The moment condition holds at the true value $\theta = \theta_0$
- We say θ is *identified* if there is a *unique* θ_0 satisfying $g(\theta_0) = 0$

Moments Conditions

- The moment condition (1) is a system of ℓ equations with p unknowns
  $\implies$ We need $\ell \geq p$ for there to be a unique solution
- If $\ell = p$, we say the model is **just identified**
  - There is just enough information to identify θ
  - GMM equals the Method of Moments (MM) in this case
- If $\ell > p$, we say the model is **overidentified**
  - There is excess identifying information (which can improve estimation efficiency)
  - GMM also applies to this case
- If $\ell < p$, we say the model is **underidentified** or **unidentified**
  - there is insufficient information to identify θ

Linear vs Non-linear Moment Models

- One of the greatest advantages of moment equation framework (MM and GMM) is that it allows both linear and non-linear moment restrictions
- Linear moment equations allow for explicit solutions for the estimators, which simplifies asymptotic theory
- Therefore, in this lecture, we focus on linear moment restrictions

Lecture outline

1. Motivation
2. Quick review of instrumental variable regression
3. Moment Conditions
4. Method of Moments (MM)
5. Generalized Method of Moments (GMM)
6. GMM Estimator and a Fundamental Decomposition
7. Asymptotic Properties
8. Optimal Choice of Weighting Matrix
9. Overidentifying restrictions

Method of Moments

- In this section we consider the just-identified case $\ell = p$
- For a given set of observations (ω_i, z_i) $i = 1, \ldots, n$, we cannot compute expectations so replace these with sample averages
- That is, whenever you see $\mathbb{E}[h(\omega_i, z_i)]$, replace it with $(1/n)\sum_{i=1}^{n} h(\omega_i, z_i)$ for any function h
- Define $\bar{g}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} f(\omega_i, z_i, \theta)$, called the **estimating equations** or **sample moments**
- The method of moments estimator $\widehat{\theta}_{\mathsf{MM}}$ is the solution to $\bar{g}_n(\theta) = 0$ (sometimes written as $\widehat{\theta}_{\mathsf{MM}} := \arg\operatorname{zero} \bar{g}_n(\theta)$)
- In some contexts, there is an explicit solution for $\widehat{\theta}_{\mathsf{MM}}$ (examples below)
- In other cases, the solution must be approximated numerically

Examples of Method of Moments Estimators

- **Mean**
  - Data: $\omega_i = y_i$, $i = 1, \ldots, n$
  - Parameters: $\theta = \mu$
  - Moment function: $f(y_i, \mu) = y_i - \mu$
  - Moment condition: $\mathbb{E}\left[f(y_i, \mu)\right] = 0$ or $\mathbb{E}[y_i] = \mu$
  - Estimator: $\widehat{\mu}_{\text{MM}}$ solves $\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(y_i, \mu) = 0 \implies \widehat{\mu}_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} y_i$
- **Mean and Variance**
  - Parameters: $\theta = (\mu, \sigma^2)$
  - Moment function: $f(y_i, \mu, \sigma^2) = \begin{bmatrix} y_i - \mu \\ (y_i - \mu)^2 - \sigma^2 \end{bmatrix}$
  - Estimators: $\widehat{\mu}_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\widehat{\sigma}^2_{\text{MM}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\mu}_{\text{MM}})^2$

Examples of Method of Moments Estimators

- **OLS**
  - Data: $\omega_i = (y_i, X_i^\top)$, $i = 1, \ldots, n$, X_i is p-dimensional
  - Parameters: $\theta = (\beta_1, \ldots, \beta_K)$
  - Moment function: $f(y_i, \beta) = X_i(y_i - X_i^\top \beta)$
  - Moment condition: $\mathbb{E}[f(y_i, \mu)] = 0$ or $\mathbb{E}[y_i] = \mu$
  - Estimator: $\widehat{\beta}_{\text{MM}} = (X^\top X)^{-1} X^\top y = \widehat{\beta}_{\text{OLS}}$
- **IV**
  - Data: $\omega_i = (y_i, X_i^\top)$, $i = 1, \ldots, n$, X_i is of size $p \times 1$ and Z_i is of size $\ell \times 1$ (recall we are currently imposing $p = \ell$)
  - Parameters: $\theta = (\beta_1, \ldots, \beta_K)$
  - Moment function: $f(y_i, X_i, Z_i, \beta) = Z_i(y_i - X_i^\top \beta)$
  - Estimating equations: $\bar{g}_n(\beta) = \frac{1}{n} \sum_{i=1}^n f(y_i, X_i, Z_i, \beta) = \frac{1}{n}(Z^\top y - Z^\top X \beta)$
  - Estimator: $\widehat{\beta}_{\text{MM}} = (Z^\top X)^{-1} Z^\top y = \widehat{\beta}_{IV}$

Lecture outline

1. Motivation
2. Quick review of instrumental variable regression
3. Moment Conditions
4. Method of Moments (MM)
5. Generalized Method of Moments (GMM)
6. GMM Estimator and a Fundamental Decomposition
7. Asymptotic Properties
8. Optimal Choice of Weighting Matrix
9. Overidentifying restrictions

Setup

Consider the linear regression model

$$y_i = x_i^\top \theta_0 + u_i, \quad i = 1, \ldots, n$$

- y_i is a scalar observed outcome
- x_i is a $p \times 1$ vector of observed explanatory variables
- θ_0 is a $p \times 1$ vector of parameters belonging to Θ (parameter space)
- z_i is a $\ell \times 1$ vector of instrumental variables
- Define $u_i(\theta) \coloneqq y_i - x_i^\top \theta$ as the disturbances. Note $u_i(\theta_0) = u_i$

Setup

Assumption 1 (Strict Stationarity)

$v_i = (x_i^\top, z_i^\top, u_i)^\top$ *is a strictly stationary process*

Assumption 2 (Population Moment Condition)

The $\ell \times 1$ *vector* z_i *satisfies* $\mathbb{E}[z_i u_i(\theta_0)] = 0$

- In a time series context, the first assumption implies that any population moments of v_t do not depend on the time index t
- Second assumption is the instrument's exclusion restriction
- This is sometimes referred to as an "orthogonality condition", as it comes from instruments being orthogonal to disturbances

Identification

- Assumption 2 specifies the information upon which estimation is based
- The resulting estimation is only going to be successful if this population moment condition provides enough information to determine θ_0 uniquely
- θ_0 is only uniquely determined by the moment condition if

$$\mathbb{E}[z_i u_i(\theta)] \neq 0$$

  at all other values of $\theta \neq \theta_0$
- In this case, θ_0 is said to be *identified* by the population moment condition
- Hence, θ_0 is identified if $\mathbb{E}[z_i x_i^\top](\theta_0 - \theta) \neq 0$ at all $\theta \neq \theta_0$, since using the moment condition we have

$$\begin{aligned}\mathbb{E}[z_i u_i(\theta)] &= \mathbb{E}[z_i u_i(\theta_0)] + \mathbb{E}[z_i x_i^\top](\theta_0 - \theta) \\ &= \mathbb{E}[z_i x_i^\top](\theta_0 - \theta)\end{aligned}$$

Identification

Assumption 3 (Identification Condition)

$\text{rank}\left(\mathbb{E}[z_i x_i^\top]\right) = p$ *with* $p \leq \ell$

- This full rank condition is necessary to guarantee identification
- Recall we need to show that for any $\theta \neq \theta_0$

$$\mathbb{E}[z_i u_i(\theta)] = \mathbb{E}[z_i x_i^\top](\theta_0 - \theta) \neq 0$$

- Set $v := \theta - \theta_0$, such that $\theta \neq \theta_0 \implies v \neq 0$
- $\mathbb{E}[z_i x_i^\top] v \neq 0$ for all $v \neq 0$ *if and only if* $\mathbb{E}[z_i x_i^\top]$ is of full rank as in Assumption 3

How can identification fail?

There are two basic scenarios

1. A failure can occur because there are fewer moment conditions than parameters
   - For any $(\ell \times p)$ matrix A, $rank(A) \leq min(p, \ell)$
   - Hence, $rank\left(\mathbb{E}[z_i x_i^\top]\right) \leq min(p, \ell)$. Suppose $\ell < p$. Then

$$rank\left(\mathbb{E}[z_i x_i^\top]\right) \leq \ell < p$$

   - Intuition: it is impossible to uniquely extract the p pieces of information needed to determine θ_0 with only ℓ moment conditions

How can identification fail?

2. A failure can occur even when $\ell \geq p$ because collectively the population moment conditions still do not provide enough information to uniquely determine θ_0

- Suppose $p = \ell = 2$. Let $x_i = (x_{1t}, x_{2t})^\top$, $z_i = (z_{1t}, z_{2t})^\top$
- We can express

$$\mathbb{E}[z_i u_i(\theta)] = \begin{pmatrix} \mathbb{E}[z_{1t}x_{1t}] & \mathbb{E}[z_{1t}x_{2t}] \\ \mathbb{E}[z_{2t}x_{1t}] & \mathbb{E}[z_{2t}x_{2t}] \end{pmatrix} \begin{pmatrix} \theta_{01} - \theta_1 \\ \theta_{02} - \theta_2 \end{pmatrix}$$

- Identification requires $rank\left(\mathbb{E}[z_i x_i^\top]\right) = 2$
- A failure can occur either because (1) it contains a row of zeros or (2) the first row is a linear transformation of the second

Lecture outline

1. Motivation
2. Quick review of instrumental variable regression
3. Moment Conditions
4. Method of Moments (MM)
5. Generalized Method of Moments (GMM)
6. GMM Estimator and a Fundamental Decomposition
7. Asymptotic Properties
8. Optimal Choice of Weighting Matrix
9. Overidentifying restrictions

Motivation

Vector/matrix notations:

- Let y be the $n \times 1$ vector with the ith element y_i
- X be the $n \times p$ matrix with the ith row being $x_i^\top$
- Z be the $n \times \ell$ matrix with the ith row being $z_i^\top$
- $u(\theta) \coloneqq y - X\theta$ produces an $n \times 1$ vector of disturbances
- True value of $p \times 1$ vector of parameters $\theta_0 \in \Theta \subset \mathbb{R}^p$
- $u \coloneqq u(\theta_0)$ is the $n \times 1$ vector of true disturbances

Motivation

- Model: $y = X\theta_0 + u$ with residual $u(\theta) := y - X\theta$
- OLS Idea: residual should be "smallest" at $\theta = \theta_0$
- Naively, you would minimize $u(\theta)$. But this is a vector, what is "minimum"?
- Idea: minimize sum of squared residuals instead, $\sum_{i=1}^{n} u_i(\theta)^2 = u(\theta)^\top u(\theta)$

Motivation

- Model: $y = X\theta_0 + u$ with residual $u(\theta) := y - X\theta$
- OLS Idea: residual should be "smallest" at $\theta = \theta_0$
- Naively, you would minimize $u(\theta)$. But this is a vector, what is "minimum"?
- Idea: minimize sum of squared residuals instead, $\sum_{i=1}^{n} u_i(\theta)^2 = u(\theta)^\top u(\theta)$
- Additionally, OLS is efficient when $\text{Var}[u(\theta)] = \sigma^2 I_n$
- Inefficiency arises with more general dependence structures in the error term: $\text{Var}[u(\theta)] = \Omega$, a general $n \times n$ matrix
- Generalized least squares (GLS) is more efficient by weighing residuals using matrix W

$$\widehat{\theta} := \arg\min_{\theta \in \Theta} u(\theta)^\top W u(\theta) = (X^\top W X)^{-1} X^\top W y$$

- Under these conditions, $W = \Omega^{-1}$ is the most efficient weighting scheme

Motivation

- Model: $g(\theta_0) := \mathbb{E}[f(\omega_i, z_i, \theta_0)] = 0$
- GMM idea: since population moments $g(\theta)$ are 0 at $\theta = \theta_0$, sample moments $\bar{g}_n(\theta)$ should also be small at this value
- Naively, you would then minimize $\bar{g}_n(\theta) := \frac{1}{n} \sum_{i=1} f(\omega_i, z_i, \theta)$
- In linear just-identified models ($\ell = p$), $\bar{g}_n(\theta_0) = 0$ is achievable
- In overidentified ($\ell > p$) or non-linear models, generally $\bar{g}_n(\theta_0) \neq 0$
- Idea: minimize sum of squares instead, $\sum_{j=1}^{\ell} \bar{g}_{n,j}(\theta)^2 = \bar{g}_n(\theta)^\top \bar{g}_n(\theta)$

Motivation

- Model: $g(\theta_0) := \mathbb{E}[f(\omega_i, z_i, \theta_0)] = 0$
- GMM idea: since population moments $g(\theta)$ are 0 at $\theta = \theta_0$, sample moments $\bar{g}_n(\theta)$ should also be small at this value
- Naively, you would then minimize $\bar{g}_n(\theta) := \frac{1}{n}\sum_{i=1} f(\omega_i, z_i, \theta)$
- In linear just-identified models ($\ell = p$), $\bar{g}_n(\theta_0) = 0$ is achievable
- In overidentified ($\ell > p$) or non-linear models, generally $\bar{g}_n(\theta_0) \neq 0$
- Idea: minimize sum of squares instead, $\sum_{j=1}^{\ell} \bar{g}_{n,j}(\theta)^2 = \bar{g}_n(\theta)^\top \bar{g}_n(\theta)$
- More efficient to weight moment conditions if $\text{Var}[\bar{g}_n(\theta)] \neq \sigma^2 I_\ell$
- Correlated moments are the standard in empirical applications
- Weighting affects parameter estimates somewhat, but largest effect is through the asymptotic variance

Estimation

- The *GMM estimator* is defined as

$$\widehat{\theta}_n = \arg\min_{\theta \in \Theta} Q(\theta, \widehat{W}),$$

where

$$Q(\theta, \widehat{W}) := \bar{g}_n(\theta)^\top \widehat{W} \bar{g}_n(\theta)$$

- Intuition: $Q(\theta, \widehat{W})$ is akin to the GLS weighted objective where $\bar{g}_n(\theta)$ plays the role of the residual
- With linear moment conditions: $\bar{g}_n(\theta) = n^{-1} \sum_{i=1}^n z_i u_i(\theta) = n^{-1} Z^\top (y - X\theta)$
- $\widehat{W}$ is the weight matrix, which can be data-driven
- Only requirement: $\text{plim}(\widehat{W}) = W$, positive-definite and symmetric
- We discuss the (optimal) choice of weight matrix $\widehat{W}$ later

GMM Estimator

- Note that

$$Q(\theta, \widehat{W}) = n^{-2} \left\{ y^\top Z \widehat{W} Z^\top y + \theta^\top X^\top Z \widehat{W} Z^\top X \theta - 2 y^\top Z \widehat{W} Z^\top X \theta \right\}$$

- First-order condition (FOC):

$$\left(n^{-1} X^\top Z\right) \widehat{W} \left(n^{-1} Z^\top y\right) = \left(n^{-1} X^\top Z\right) \widehat{W} \left(n^{-1} Z^\top X\right) \widehat{\theta}_n$$

- Provided that $\left(n^{-1} X^\top Z\right) \widehat{W} \left(n^{-1} Z^\top X\right)$ is non-singular, the GMM estimator is given in closed form by

$$\widehat{\theta}_n = \left[\left(n^{-1} X^\top Z\right) \widehat{W} \left(n^{-1} Z^\top X\right) \right]^{-1} \left(n^{-1} X^\top Z\right) \widehat{W} \left(n^{-1} Z^\top y\right) \qquad (2)$$

Interpretation

- The FOC can be written as

$$\left[(n^{-1}X^\top Z)\right]\widehat{W}\left[n^{-1}Z^\top u(\widehat{\theta}_n)\right] = 0$$

- The population analogue of the above is

$$\mathbb{E}[x_i z_i^\top] W \mathbb{E}[z_i u_i(\theta_0)] = 0 \tag{3}$$

- The above is called is the Method of Moment interpretation of the GMM estimator
- This interpretation clarifies the importance of the orthogonality condition

Interpretation

- Note that in the special case of just identification $p = \ell$, the inverses in the estimator are well-defined, such that

$$\left[\left(n^{-1} X^\top Z \right) \widehat{W} \left(n^{-1} Z^\top X \right) \right]^{-1} = \left[\left(n^{-1} Z^\top X \right) \right]^{-1} \widehat{W}^{-1} \left[\left(n^{-1} X^\top Z \right) \right]^{-1}$$

- This result leads to an equivalence between GMM and IV:

$$\widehat{\theta}_{GMM} = \left[\left(n^{-1} Z^\top X \right) \right]^{-1} \left(n^{-1} Z^\top y \right) = \widehat{\theta}_{IV}$$

- In the just-identified case the weight matrix $\widehat{W}$ plays no role!
- If $\ell > p$, no such reduction is possible

Interpretation

- In the overidentified case $\ell > p$ there is a difference between the orthogonality condition and the method of moments representation
- The information with which we began, Assumption 2, states

$$\mathbb{E}[z_i u_i(\theta_0)] = 0$$

- The information used in estimation is instead given by Equation (3):

$$\mathbb{E}[x_i z_i^\top] W \, \mathbb{E}[z_i u_i(\theta_0)] = 0$$

- The choice of weighting matrix is important as it determines the exact nature of the linear combinations of $\mathbb{E}[z_i u_i(\theta_0)]$ set to zero in the FOC

Lecture outline

Asymptotic Properties

- To illustrate ideas, we strengthen Assumption 1 to incorporate a cross-sectional random sample assumption:

Assumption 4 (Independence)

The vector $v_i := (x_i^\top, z_i^\top, u_i)^\top$ *is independent of* v_j *for all* $i \neq j$.

- Assumptions 1 and 4 jointly imply v_i is an independent and identically distributed (iid) process.
- Plug $y = X\theta_0 + U$ into Equation (2):

$$\widehat{\theta}_n = \theta_0 + \left[\left(n^{-1} X^\top Z\right) \widehat{W} \left(n^{-1} Z^\top X\right) \right]^{-1} \left(n^{-1} X^\top Z\right) \widehat{W} \left(n^{-1} Z^\top u\right) \quad (4)$$

- Consistency and asymptotic normality of $\widehat{\theta}_n$ follow directly from Equation (4)

Consistency

- Remember consistency means: $\widehat{\theta}_n \xrightarrow{p} \theta_0$ or $\text{plim}(\widehat{\theta}_n) = \theta_0$
- From Equation (4),

$$\text{plim}\,\widehat{\theta}_n = \theta_0 + \text{plim}\left(\left[\left(n^{-1}X^\top Z\right) \widehat{W} \left(n^{-1}Z^\top X\right) \right]^{-1} \left(n^{-1}X^\top Z\right) \widehat{W} \left(n^{-1}Z^\top u\right) \right)$$

- Using Slutsky's Theorem (distributing probability limits),

$$\begin{aligned} \text{plim}\,\widehat{\theta}_n = \theta_0 + & \left[\text{plim}\left(n^{-1}X^\top Z\right) \text{plim}\,\widehat{W}\, \text{plim}\left(n^{-1}Z^\top X\right) \right]^{-1} \\ & \times \text{plim}\left(n^{-1}X^\top Z\right) \text{plim}\,\widehat{W}\, \text{plim}\left(n^{-1}Z^\top u\right) \end{aligned} \tag{5}$$

Consistency

- A weak law of large number (WLLN) gives condition for $\frac{1}{n}h(\omega_i) \xrightarrow{p} \mathbb{E}[h(\omega_i)]$
- The limiting behaviour of the other matrices above can be deduced from the WLLN
- Since $z_i x_i^\top$ and $z_i u_i$ are functions of independent processes, they are themselves independent processes
- Therefore, WLLN implies

$$n^{-1}Z^\top X = n^{-1}\sum_{i=1}^{n} z_i x_i^\top \xrightarrow{p} \mathbb{E}[z_i x_i^\top]$$

$$n^{-1}Z^\top u = n^{-1}\sum_{i=1}^{n} z_i u_i^\top \xrightarrow{p} \mathbb{E}[z_i u_i]$$

Consistency

- It is at this point that the population moment and identification conditions become important
- The identification condition (Assumption 3) states that $\mathbb{E}[z_i x_i^\top]$ is of rank p and so the inverse of $\mathbb{E}[x_i z_i^\top] W \mathbb{E}[z_i x_i^\top]$ exists
- The population moment condition (Assumption 2) states that $\mathbb{E}[z_i u_i] = 0$
- Using the two conditions in Equation (5),

$$\operatorname{plim} \widehat{\theta}_n = \theta_0 + M \mathbb{E}[z_i u_i] = \theta_0$$

  where $M \coloneqq \left\{ \mathbb{E}[x_i z_i^\top] W \mathbb{E}[z_i x_i^\top] \right\}^{-1} \mathbb{E}[x_i z_i^\top] W$.
- Therefore, $\widehat{\theta}_n$ is consistent for θ_0!

Consistency

- Why does the inverse exist? Facts from linear algebra:
- Let A be a $(\ell \times p)$ matrix. If W is a $(p \times p)$ matrix of rank p, then

$$rank(AW) = rank(A)$$

- If a $(p \times p)$ matrix is positive definite, then $rank(W) = p$
- A $(\ell \times \ell)$ matrix is invertible if and only if $rank(B) = \ell$

Asymptotic normality

- The asymptotic distribution of the GMM estimator is derived by rewriting Equation (4) as

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) = \left[\left(n^{-1}X^\top Z\right)\widehat{W}\left(n^{-1}Z^\top X\right)\right]^{-1}\left(n^{-1}X^\top Z\right)\widehat{W}\left(\frac{1}{\sqrt{n}}Z^\top u\right)$$

- Take a look at the right-most component: since $z_i u_i$ is an independent process, a Central Limit Theorem (CLT) can invoked to deduce that

$$\frac{1}{\sqrt{n}}Z^\top u = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} z_i u_i - \mathbb{E}[z_i u_i]\right) \xrightarrow{d} \mathcal{N}(0, S),$$

- $S := \lim_{n\to\infty} Var[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i]$ and the zero-mean follows from the population moment condition $\mathbb{E}[z_i u_i] = 0$

Asymptotic normality

- Putting it all together:

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} \mathcal{N}(0, MSM^\top), \tag{6}$$

- Recall that $M = \left\{\mathbb{E}[x_i z_i^\top] W \mathbb{E}[z_i x_i^\top]\right\}^{-1} \mathbb{E}[x_i z_i^\top] W$
- In the special just-identified case with $p = \ell$, $M = \mathbb{E}[z_i x_i^\top]$, so the asymptotic variance reduces to

$$MSM^\top = \{\mathbb{E}[z_i x_i^\top]\}^{-1} S \{\mathbb{E}[x_i z_i^\top]\}^{-1}$$

Asymptotic normality

- The asymptotic normality (6) implies that an approximate large sample $100(1-\alpha)$ % confidence interval for $\theta_{0,j}$ is

$$\widehat{\theta}_{n,j} \pm z_{\alpha/2}\sqrt{\widehat{V}_{n,jj}/n},$$

- $\widehat{V}$ is a consistent estimator of $MSM^{\top}$ and $\widehat{V}_{n,ii}$ is its i-th diagonal element
- $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution

Covariance Estimation

- A consistent estimator of $MSM^\top$ can be obtained from consistent estimators of its components M and S
- By Slutsky's Theorem, if $\widehat{M}_n \xrightarrow{p} M$ and $\widehat{S}_n \xrightarrow{p} S$ then

$$\widehat{M}_n \widehat{S}_n \widehat{M}_n^\top \xrightarrow{p} MSM^\top$$

- The obvious choice for $\widehat{M}_n$ is

$$\widehat{M}_n = \left[\left(n^{-1} X^\top Z \right) \widehat{W} \left(n^{-1} Z^\top X \right) \right]^{-1} \left(n^{-1} X^\top Z \right) \widehat{W}$$

- To construct $\widehat{S}_n$, we need to be more specific about the form of the long run covariance matrix S
- Our maintained assumptions pin down the form of S

Covariance Estimation

- Under our assumptions $z_i u_i$ is an independently and identically distributed process with a mean of zero
- Together, these restrictions imply

$$S = \lim_{n\to\infty} \text{Var}\left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n} z_i u_i\right] = \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[u_i u_j z_i z_j^\top] = \mathbb{E}[u_i^2 z_i z_i^\top]$$

- Let $\widehat{u}_i := y_i - x_i^\top \widehat{\theta}_n$ be the GMM residuals. White (1984) demonstrates that S can be consistently estimated by

$$\widehat{S}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{u}_i^2 z_i z_i^\top,$$

- In certain circumstances, more structure can be placed on $\mathbb{E}[u^2 zz^\top]$ that can be exploited in the construction of $\widehat{S}_n$

Lecture outline

1. Motivation
2. Quick review of instrumental variable regression
3. Moment Conditions
4. Method of Moments (MM)
5. Generalized Method of Moments (GMM)
6. GMM Estimator and a Fundamental Decomposition
7. Asymptotic Properties
8. Optimal Choice of Weighting Matrix
9. Overidentifying restrictions

Optimal Weighting Matrix

- So far, the analysis has taken the weighting matrix as given
- We noted that this matrix plays a crucial role in the analysis because it determines the exact nature of the minimand
- We can characterize the optimal choice of weighting matrix, leading us to a discussion of the two step and iterated GMM estimators
- We must consider what is meant by "optimality" in this context
- Previous analysis indicates that the weighting matrix only affects the asymptotic properties of the estimator via the asymptotic covariance matrix

Optimal Weighting Matrix

- The estimator will converge in probability to the true value as long as the population moment and identification conditions hold
- Our assumptions ensure there is sufficient information from which to estimate θ_0 and that this information is correct
- The choice of weighting matrix determines how this information is used and so impacts directly upon the precision of the estimation
- It is this feature which is captured by the variance of the asymptotic distribution
- Therefore the optimal weighting matrix is defined to be the value which minimizes the asymptotic variance

Optimal Weighting Matrix

- Again, $W = \text{plim}(\widehat{W})$ affects the asymptotic variance of $\widehat{\theta}_n$:

$$V(W) = \left\{\mathbb{E}[x_i z_i^\top] W \, \mathbb{E}[z_i x_i^\top]\right\}^{-1} \mathbb{E}[x_i z_i^\top] WSW \, \mathbb{E}[z_i x_i^\top] \left\{\mathbb{E}[z_i x_i^\top] W \, \mathbb{E}[z_i x_i^\top]\right\}^{-1}$$

- The optimal value of W is the value that minimizes $V(W)$, denoted as W^0
- Minimizing in a matrix sense means that for any valid choice of weighting matrix W

$$V(W) - V(W^0) = \text{ a positive semi-definite matrix}$$

- Hansen (1982) shows that $W^0 = S^{-1}$, the inverse of the moment covariance matrix (long-run covariance)

Optimal Weighting Matrix: Proof

Let W be any weight matrix. Let $Q \coloneqq \mathbb{E}[z_i x_i^\top]$.

$$
\begin{aligned}
&V(W) - V(S^{-1}) \\
&= \left[Q^\top W Q\right]^{-1} Q^\top W S W Q^\top \left[Q^\top W Q\right]^{-1} - \left[Q^\top S^{-1} Q\right]^{-1} \\
&= \left[Q^\top W Q\right]^{-1} \left(Q^\top W S W Q^\top - \left[Q^\top W Q\right] \left[Q^\top S^{-1} Q\right]^{-1} \left[Q^\top W Q\right] \right) \left[Q^\top W Q\right]^{-1} \\
&= \underbrace{\left[Q^\top W Q\right]^{-1} Q^\top W S^{1/2}}_{A} \left(I - \underbrace{S^{-1/2} Q \left[Q^\top S^{-1} Q\right]^{-1} Q^\top S^{-1/2}}_{B} \right) \underbrace{S^{1/2} W Q \left[Q^\top W Q\right]^{-1}}_{A^\top} \\
&\coloneqq A(I - B)A^\top
\end{aligned}
$$

Optimal Weighting Matrix: Proof

- You can check that B is symmetric and idempotent: $B = B^\top$ and $B \cdot B = B$
- This means that $I - B$ is symmetric and idempotent as well, which means $I - B = (I - B)(I - B)^\top$
- Finally,

$$V(W) - V(S^{-1}) = A(I - B)(I - B)^\top A = \Big[A(I - B)\Big]\Big[A(I - B)\Big]^\top$$

- Result matrix is an outer product of a matrix with itself, which is always positive semi-definite: $v^\top CC^\top v = (C^\top v)(C^\top v)^\top = (C^\top v)^2 \geq 0$ if $v \neq 0$
- Conclusion: $V(W) - V(S^{-1})$ is positive semi-definite proving the claim

Optimal Weighting Matrix

- The optimal choice $W^0 = S^{-1}$ yields

$$V(S^{-1}) = \left(\mathbb{E}[x_i z_i^\top] S^{-1} \mathbb{E}[z_i x_i^\top] \right)^{-1}$$

- This is the covariance matrix achieved by the efficient GMM estimator
- Chamberlain has shown that there is no semiparametric estimator with a smaller asymptotic variance than this
- Because the efficient GMM estimator attains this asymptotic variance, it is *semiparametrically efficient*
- Caveat: this is the unfeasible GMM estimator that assumes S as known

Optimal Weighting Matrix

- In practice $W^0 = S^{-1}$ is unknown, meaning it needs to be estimated
- It suffices to set $\widehat{W} = \widehat{S}_n^{-1}$, where $\widehat{S}_n$ is a consistent estimator of S
- Note that consistency of $\widehat{S}_n$ only requires a consistent estimator of θ_0 (i.e., it does not need to be the optimal one)
- This leads to Hansen (1982)'s *two-step GMM estimator*:
  1. A consistent estimator $\widetilde{\theta}$ is obtained using GMM with a sub-optimal weighting matrix, such as $\widehat{W} = I_\ell$ or $\widehat{W} = n^{-1} Z^\top Z$
  2. $\widehat{S}_n$ is computed from the GMM residuals $\widetilde{u} = y - X\widetilde{\theta}$ and the model is re-estimated using $\widehat{W} = \widehat{S}_n^{-1}$

Optimal Weighting Matrix

- These two steps are sufficient to obtain an asymptotically efficient estimator with asymptotic covariance matrix equal to $V(S^{-1})$
- However, the estimator of S used in the second step estimation is based on a suboptimal estimator of θ_0
- Finite sample performance might be improved by iterating the two-step procedure until convergence
- This leads to the *Iterated GMM estimator*
- Directly maximizing the objective replacing the optimal weighting matrix formula leads to the *Continuously Updated GMM*:

$$\widehat{\theta}_{\text{CU-GMM}} \coloneqq \arg\min_{\theta\in\Theta} \bar{g}_n(\theta)^\top \left[\frac{1}{n}u(\theta)^\top Z^\top Z u(\theta)\right]^{-1} \bar{g}_n(\theta)$$

Lecture outline

1. Motivation
2. Quick review of instrumental variable regression
3. Moment Conditions
4. Method of Moments (MM)
5. Generalized Method of Moments (GMM)
6. GMM Estimator and a Fundamental Decomposition
7. Asymptotic Properties
8. Optimal Choice of Weighting Matrix
9. Overidentifying restrictions

Test for overidentifying restrictions

- Sargan (1958) was the first to introduce testing the validity of instrumental variances using overidentifying restrictions in a linear model estimated by IV
- Hansen (1982) extended the statistic to the GMM framework
- Null hypothesis of the test: $H_0 : \mathbb{E}[z_i u_i(\theta_0)] = 0$
- Test statistic is simply the GMM objective replacing the optimal weighting matrix:

$$J_n = nQ(\widehat{\theta}_n, \widehat{S}_n^{-1}) = \left[\frac{1}{\sqrt{n}} u(\widehat{\theta}_n)^\top Z\right] \widehat{S}_i^{-1} \left[\frac{1}{\sqrt{n}} Z^\top u(\widehat{\theta}_n)\right]$$

- Rejects when there is evidence of $\mathbb{E}[z_i u_i(\theta_0)] \neq 0$, meaning when instruments are not valid

Test for overidentifying restrictions

- Under the null, J_n converges in distribution to a $\chi^2_{\ell-p}$
- Degrees of freedom given by the number of overidentifying restrictions $\ell - p$
- Test statistic provides evidence the moment conditions do not hold in the sample
- Intuition: J_n will be large when $\bar{g}_n(\widehat{\theta}_n)$ is not close to 0, meaning when the moment conditions are not close to 0 at the current estimate
- Useful estimation by-production as it is easily evaluated
- In Stata, *estat overid* after *ivregress gmm*