Lecture 6: Instrumental Variables & GMM Foundations

Econometrics 2 — *From Moment Conditions to Estimation*

Instructor: Prof. S. Arvanitis | **Digitisation of notes by:** T. Kourtalis

▷ Amber boxes = Handwritten Notes (professor's words)

◇ Teal boxes = Student's Notes

Recall from Previous Lectures

In previous lectures we developed the **extremum estimator** framework: $\hat{\beta}_n \in \arg\min_{\beta\in\Theta} M_n(\beta)$. For OLS, we had $M_n(\beta) = \frac{1}{n}\|Y_n - X_n\beta\|^2$ and the key assumption was $\mathbb{E}(X_n'\varepsilon_n) = 0$ (exogeneity). But what if some regressors are **endogenous**? This lecture introduces **instrumental variables** and the **GMM** framework to handle this situation.

1 The Endogeneity Problem & Instrumental Variables

◇ Student's Notes

Why do we need instruments?

In the linear model $Y = X\beta_0 + \varepsilon$, OLS consistency requires $\mathbb{E}(X'\varepsilon) = 0$. This fails when:

- **Omitted variable bias:** An unobserved variable affects both X and Y
- **Measurement error:** X is measured with error correlated with the true value
- **Simultaneity:** X and Y are jointly determined

The solution: Find **instruments** Z_n ($n \times q$ matrix, $q \geq p$) satisfying:

Relevance
$\operatorname{rank}(\mathbb{E}[Z'X]) = p$

Exogeneity
$\mathbb{E}(Z'\varepsilon) = 0$

Both conditions must hold for valid IV estimation

Figure 1: The two requirements for valid instruments. Relevance ensures identification; exogeneity ensures consistency.

2 System of Moment Conditions

▷ **Handwritten Notes** (what the professor said)

In the framework of the linear model with instrumental variables, we are led to the q-dimensional vector of moments ($q \geq p$):

$$f(\beta) = \mathbb{E}\left[\frac{1}{n} Z_n'(Y_n - X_n\beta)\right] = \cdots = \frac{1}{n}\mathbb{E}[Z_n'X_n](\beta_0 - \beta)$$

This utilizes the Law of Iterated Expectations, since the errors of Z_n will be orthogonal to the columns of $X_n \implies \mathbb{E}(Z_n'\varepsilon_n) = 0$.
Using the identification condition $\operatorname{rank}(X_n) = p$, the previous equation implies that we could find β_0 if we could solve the system:

$$\mathbb{E}\left(\frac{1}{n} Z_n'X_n\right)(\beta_0 - \beta) = 0_{q\times 1}$$

This is a system of **Moment Conditions**. Obviously, the left side of our system is unobservable since it depends on the true parameter β_0.

◇ **Student's Notes**

Understanding the moment conditions geometrically:
The condition $\mathbb{E}(Z'\varepsilon) = 0$ says that each column of Z is *orthogonal* to the error term in the population. This gives us q separate orthogonality conditions.

The identification requirement:
For $\mathbb{E}[Z'X](\beta_0 - \beta) = 0$ to have a *unique* solution $\beta = \beta_0$, we need:

$$\operatorname{rank}\big(\mathbb{E}[Z'X]\big) = p$$

This is a $q \times p$ matrix, so we need $q \geq p$ (at least as many instruments as parameters).

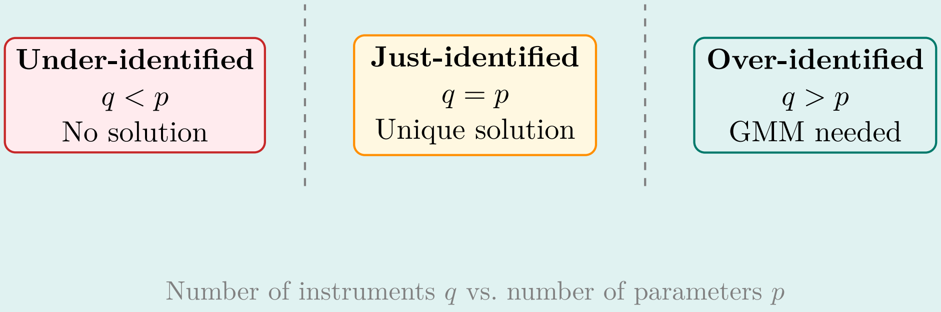


Figure 2: The three identification cases. Under-identification is fatal; just-identification gives a unique solution; over-identification requires a criterion to combine the "extra" information.

★ Intuition

Why orthogonality = identification:
Think of each moment condition as a “constraint” that β must satisfy. With p unknowns:

- $q < p$: Too few constraints ⇒ infinitely many solutions
- $q = p$: Exactly enough constraints ⇒ unique solution (if rank condition holds)
- $q > p$: Over-constrained ⇒ generally no exact solution, so we find the “best approximate” solution (GMM)

3 Sample Approximation and Objective Functions

▷ **Handwritten Notes** (what the professor said)

What we could do is try to solve the sample approximation:

$$\frac{1}{n}Z_n'(Y_n - X_n\beta) = 0_{q\times 1}$$

(We try to solve for this equation when $q \geq p$).
When we talk about length and distance, we are talking about the same thing here. We will try to express this as something that results from an optimization problem. In this case, we will try to construct an objective function that will represent some notion of distance of $\frac{1}{n}\mathbb{E}(Z_n'X_n)(\beta_0 - \beta)$ from 0.
There are different ways to do this.

◇ Student’s Notes

The key insight:
When $q > p$, we have more equations than unknowns, so $\frac{1}{n}Z'(Y - X\beta) = 0$ typically has *no exact solution*. Instead, we find β that makes the sample moments “as close to zero as possible.”

The general GMM approach:
Define the **sample moment vector**:

$$g_n(\beta) = \frac{1}{n}Z_n'(Y_n - X_n\beta) \quad \in \mathbb{R}^q$$

Then minimise some “distance” of $g_n(\beta)$ from $0_{q\times 1}$. The choice of distance metric determines the estimator.

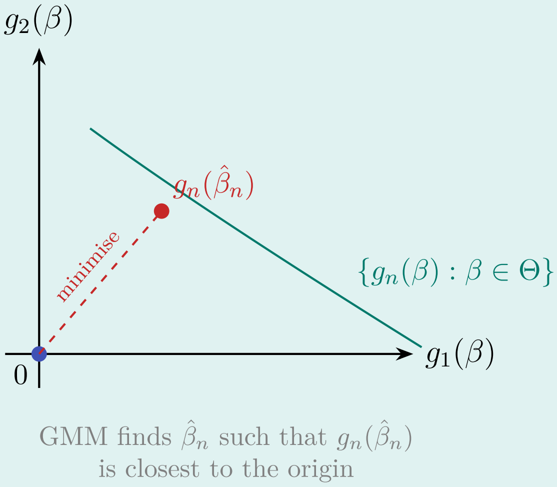


Figure 3: GMM as finding the point on the moment curve closest to zero. The "distance" depends on the choice of metric.

3.1 Example 1: Maximum Absolute Distance

▷ Handwritten Notes (what the professor said)

We could use the following distance. If $x, y \in \mathbb{R}^q$, then as a distance between them we could use $\max_{j=1,\ldots,q} |x_j - y_j|$.
So, we can define:

$$\mu_n^*(\beta) = \max_{j=1,\ldots,q} \left| \left(\frac{1}{n} \mathbb{E}(Z_n' X_n)(\beta_0 - \beta) \right)_j \right|$$

Because of the previous properties, we have that $\arg\min_{\beta\in\Theta} \mu_n^*(\beta) = \beta_0$.
Correspondingly with what we have done in our first example, under certain assumptions the unobservable μ_n^* can be approximated by the observable counterpart:

$$\mu_n(\beta) = \max_{j=1,\ldots,q} \left| \left(\frac{1}{n} Z_n'(Y_n - X_n\beta) \right)_j \right|$$

So the corresponding estimator $\hat\beta_n$ will be defined from $\hat\beta_n \in \arg\min_{\beta\in\Theta} \mu_n(\beta)$.

◇ Student's Notes

The ℓ^∞ (sup-norm) approach:
This uses the ℓ^∞ norm: $\|x\|_\infty = \max_j |x_j|$. The objective penalises the *worst-fitting* moment condition.

Properties:

- **Pro:** Robust to having one moment fit much worse than others
- **Con:** Non-smooth (the max function has kinks), making optimisation harder
- **Con:** Requires linear programming techniques, not standard gradient methods

Rarely used in practice due to computational difficulties. We mention it to show that GMM is flexible—many distance metrics are possible.

3.2 Example 2: Mahalanobis-Type Distance (Quadratic Form)

▷ **Handwritten Notes** (what the professor said)

We will use distances of the Mahalanobis type. Let matrix W be positive definite; the distance between x, y with respect to W is $\sqrt{(x-y)'W(x-y)}$. Here, W acts as a *weighting matrix*. For example, when $W = I_{q\times q}$, we have the usual Euclidean distance. Given such a W, we can construct:

$$\mu_n^*(\beta, W) = \left[\left(\frac{1}{n}\mathbb{E}(Z_n'X_n)(\beta_0-\beta)\right)' W \left(\frac{1}{n}\mathbb{E}(Z_n'X_n)(\beta_0-\beta)\right)\right]^{1/2}$$

Correspondingly, under assumptions, the unobservable μ_n^* can be approximated by the observable:

$$\begin{aligned}\mu_n(\beta, W) &= \left(\frac{1}{n}Z_n'(Y_n - X_n\beta)\right)' W \left(\frac{1}{n}Z_n'(Y_n - X_n\beta)\right)\\ &= \frac{1}{n^2}(Y_n - X_n\beta)'(Z_nWZ_n')(Y_n - X_n\beta)\end{aligned}$$

Definition: Mahalanobis Distance

For vectors $x, y \in \mathbb{R}^q$ and a positive definite matrix W $(q \times q)$, the **Mahalanobis distance** is:

$$d_W(x, y) = \sqrt{(x-y)'W(x-y)}$$

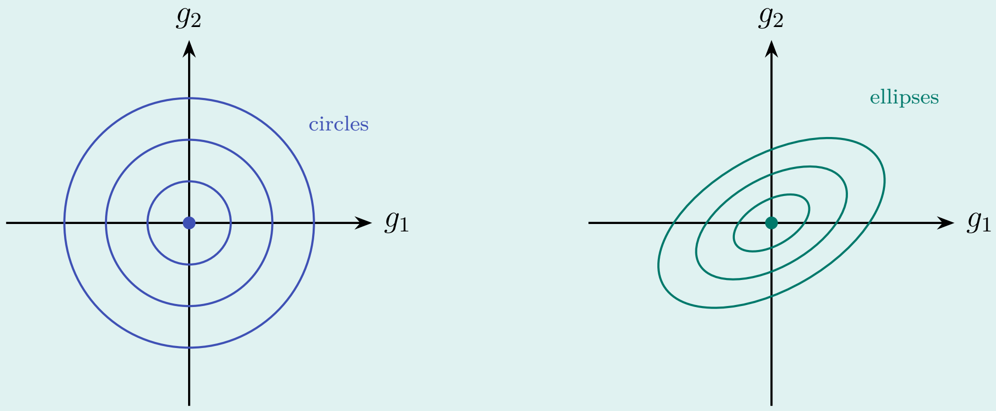
When $W = I$, this reduces to Euclidean distance.

◇ **Student's Notes**

Why use a weighting matrix W?
The weighting matrix W allows us to:

1. **Scale** moment conditions that have different units
2. **Weight** more informative moments more heavily
3. **Account for correlation** between different moments

Geometric interpretation:



$W = I$ **(Euclidean)**

General W **(Mahalanobis)**

Figure 4: Level sets of the distance function. With $W = I$, equal distance from zero forms circles. With general W, we get ellipses—some directions are "penalised" more than others.

Key observation: Since we minimise $\mu_n(\beta, W)^2$ (a monotone transformation), we typically work with the **squared** distance to avoid the square root:

$$M_n(\beta, W) = g_n(\beta)' W \, g_n(\beta) = \frac{1}{n^2}(Y - X\beta)' Z W Z' (Y - X\beta)$$

This is a **quadratic form** in the residuals, which is smooth and easy to optimise.

4 The Instrumental Variables Estimator (IVE)

▷ **Handwritten Notes** (what the professor said)

The usual version of the Instrumental Variables Estimator (IVE) is:

$$\hat{\beta}_n \in \arg\min_{\beta \in \Theta} \frac{1}{n^2}(Y_n - X_n\beta)'(Z_n W Z_n')(Y_n - X_n\beta)$$

Questions arise of the form:

- Does $\hat{\beta}_n$ depend on, and what are its properties based on, the choice of the constant matrix W?
- If yes, how can we choose the optimal W depending on the sample?

If we study the optimization problem a bit more, let's initially assume that $\Theta = \mathbb{R}^p$. We have:

$$\min_{\beta\in\mathbb{R}^p} \frac{1}{n^2}(Y_n - X_n\beta)'(Z_n W Z_n')(Y_n - X_n\beta)$$

This objective function is at least twice continuously differentiable. Since $\Theta = \mathbb{R}^p$, we can use first and second-order conditions.

◇ Student's Notes

Preview of answers to the professor's questions:

Question	Answer (details in later lectures)
Does $\hat{\beta}_n$ depend on W?	In general, **yes**—different W give different estimates (unless $q = p$, then all give the same)
What properties depend on W? **Efficiency:** depends critically on W	**Consistency:** holds for any positive definite W
Optimal W?	$W_{\text{opt}} = [\text{Var}(g_n(\beta_0))]^{-1}$, the inverse of the variance of the moment conditions

Intuition for optimal weighting:
Weight moment conditions by the *inverse* of their variance. High-variance moments are less informative ⇒ give them less weight. This is analogous to GLS weighting residuals by inverse variance.

5 Parenthesis: Matrix Derivatives

▷ Handwritten Notes (what the professor said)

Let's say I have a function $G(\beta) : \mathbb{R}^p \to \mathbb{R}^n$. For example, $G(\beta) = Y_n - X_n\beta$. And an inner matrix $A_{n\times n}$ (e.g., $A = Z_n W Z_n'$).
We are interested in differentiating with respect to β the quadratic form:

$$G'(\beta)\, A\, G(\beta)$$

The first derivative with respect to β is:

$$\frac{\partial G'(\beta)\, A\, G(\beta)}{\partial \beta} = 2\frac{\partial G'(\beta)}{\partial \beta}\, A\, G(\beta)$$

Additionally, if G is linear with respect to β (which implies $\frac{\partial^2 G'(\beta)}{\partial\beta\,\partial\beta'} = 0_{p\times p}$), then the

second derivative is:
$$\frac{\partial^2 G'(\beta)\, A\, G(\beta)}{\partial\beta\,\partial\beta'} = 2\frac{\partial G'(\beta)}{\partial\beta}\, A\, \frac{\partial G(\beta)}{\partial\beta'}$$

Key Result

Matrix Calculus Rules for Quadratic Forms
Let $G(\beta) : \mathbb{R}^p \to \mathbb{R}^n$ and A be an $n \times n$ symmetric matrix. Define:

$$Q(\beta) = G(\beta)' A\, G(\beta) \quad \in \mathbb{R}$$

First derivative (gradient, $p \times 1$ vector):

$$\frac{\partial Q}{\partial\beta} = 2\left(\frac{\partial G}{\partial\beta'}\right)' A\, G(\beta) = 2\frac{\partial G'}{\partial\beta}\, A\, G(\beta)$$

Second derivative (Hessian, $p \times p$ matrix), when G is *linear* in β:

$$\frac{\partial^2 Q}{\partial\beta\,\partial\beta'} = 2\frac{\partial G'}{\partial\beta}\, A\, \frac{\partial G}{\partial\beta'}$$

◇ Student's Notes

Application to our problem:
With $G(\beta) = Y - X\beta$ and $A = ZWZ'$:

$$\frac{\partial G}{\partial\beta'} = -X \quad \text{(an } n \times p \text{ matrix)}$$
$$\frac{\partial G'}{\partial\beta} = -X' \quad \text{(a } p \times n \text{ matrix)}$$

Therefore:

$$\frac{\partial Q}{\partial\beta} = 2(-X')\,(ZWZ')\,(Y - X\beta) = -2X'ZWZ'(Y - X\beta)$$
$$\frac{\partial^2 Q}{\partial\beta\,\partial\beta'} = 2(-X')\,(ZWZ')\,(-X) = 2X'ZWZ'X$$

! Watch Out

Notation pitfall: Be careful with the dimensions!

Object	Dimension	Notes
Y	$n \times 1$	Response vector
X	$n \times p$	Regressor matrix
Z	$n \times q$	Instrument matrix
W	$q \times q$	Weighting matrix
β	$p \times 1$	Parameter vector
$Z'X$	$q \times p$	Relevance matrix
$X'ZWZ'X$	$p \times p$	Appears in FOC

6 First-Order Conditions and Optimisation

▷ **Handwritten Notes** (what the professor said)

Building on our matrix derivative rules, we want to find the parameter that minimizes our objective function.

1st Order Conditions:

We set the first derivative of the objective function $\mu_n(\beta, W)$ with respect to β equal to zero:

$$\frac{\partial \mu_n(\beta, W)}{\partial \beta} = 0_{p\times 1}$$

Applying our chain rule:

$$\Leftrightarrow \quad \frac{1}{n^2} \cdot 2\, \frac{\partial (Y_n - X_n\beta)'}{\partial \beta}(Z_n W Z_n')(Y_n - X_n\beta) = 0_{p\times 1}$$

Let's evaluate the inner derivative:

$$\frac{\partial (Y_n - X_n\beta)'}{\partial \beta} = \frac{\partial (Y_n' - \beta' X_n')}{\partial \beta} = -\frac{\partial (\beta' X_n')}{\partial \beta} = -X_n'$$

Substituting this back into our first-order condition:

$$\frac{2}{n^2}(-X_n')(Z_n W Z_n')(Y_n - X_n\beta) = 0_{p\times 1}$$

By simplifying and rearranging the terms, we get:

$$X_n'(Z_n W Z_n')X_n\beta = X_n'(Z_n W Z_n')Y_n$$

(Note: This structure corresponds to the normal equations).

◇ Student's Notes

Comparing to OLS normal equations:

Estimator	Normal Equations
OLS	$X'X\,\hat{\beta} = X'Y$
IV (with weighting W)	$X'ZWZ'X\,\hat{\beta} = X'ZWZ'Y$

Pattern: Replace $X'(\cdot)X$ with $X'ZWZ'(\cdot)X$ and $X'(\cdot)Y$ with $X'ZWZ'(\cdot)Y$.
The instruments Z "mediate" the relationship between X and Y, filtering out the endogenous variation.

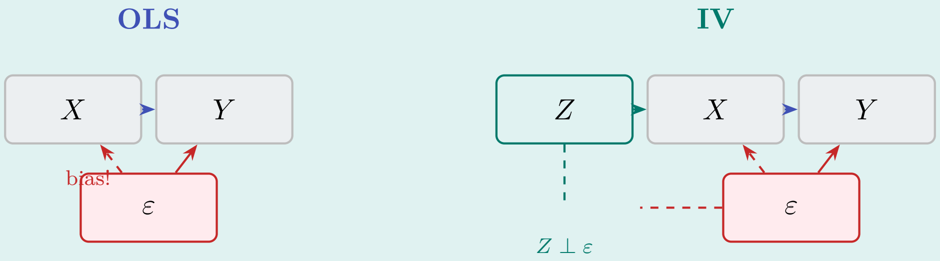


Figure 5: OLS uses variation in X directly, which is contaminated by ε. IV uses only the variation in X that comes from Z, which is "clean" (exogenous).

7 Matrix Invertibility and the IV Estimator

▷ Handwritten Notes (what the professor said)

To solve for β, we must ensure the matrix on the left-hand side is invertible.
Since we assume:

- $\text{rank}(Z_n) = q$
- $\text{rank}(W) = q$
- $\text{rank}(X_n) = p$

The matrix $(X_n'Z_n)W(Z_n'X_n)$ has rank p and is therefore invertible. (It is also positive definite).
Critical Point:
Solving for the estimator gives us:

$$\hat{\beta}_n = (X_n'Z_nWZ_n'X_n)^{-1}X_n'Z_nWZ_n'Y_n$$

2nd Order Conditions:
Checking the second derivative confirms we have a minimum:

$$\frac{\partial^2 \mu_n(\beta)}{\partial \beta \, \partial \beta'} = \frac{2}{n^2} X_n' Z_n W Z_n' X_n$$

Since the resulting matrix is positive definite, our critical point is indeed a minimum.

Key Result

The General IV Estimator
With weighting matrix W (positive definite, $q \times q$):

$$\boxed{\hat{\beta}_n^{IV} = (X'ZWZ'X)^{-1}X'ZWZ'Y}$$

Existence requires: $\operatorname{rank}(Z'X) = p$ (relevance condition)
The estimator is unique because: $X'ZWZ'X$ is positive definite (hence invertible)

◇ Student's Notes

Special cases:

1. **Just-identified case** ($q = p$):

   When Z is square and invertible, a remarkable simplification occurs. Taking $W = (Z'Z)^{-1}$:
   $$\hat{\beta}_n^{IV} = (Z'X)^{-1}Z'Y$$
   This is the classic IV/2SLS estimator. Note: **the choice of W doesn't matter** when $q = p$ (all give the same $\hat{\beta}$)!

2. **Over-identified case** ($q > p$):

   Different W give different estimators. The **efficient GMM** chooses W optimally (inverse of moment variance).

3. **Two-Stage Least Squares (2SLS):**

   A popular choice is $W = (Z'Z)^{-1}$, giving:
   $$\hat{\beta}_n^{2SLS} = (X'P_Z X)^{-1}X'P_Z Y$$
   where $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix onto the column space of Z.

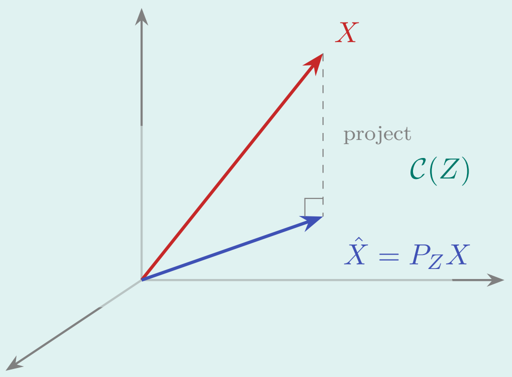


Figure 6: 2SLS first projects X onto the column space of Z, obtaining $\hat{X} = P_Z X$. Then it runs OLS of Y on $\hat{X}$. This removes the "bad" variation in X that is correlated with ε.

8 Properties of the IV Estimator

▷ **Handwritten Notes** (what the professor said)

Therefore, in the parameter space $\Theta = \mathbb{R}^p$, with a specific choice of weighting matrix W, the Instrumental Variables Estimator (IVE) is:

$$\hat{\beta}_n = (X_n' Z_n W Z_n' X_n)^{-1} X_n' Z_n W Z_n' Y_n$$

We know the true data-generating process is $Y_n = X_n\beta_0 + \varepsilon_n$. Substituting Y_n into our estimator:

$$\hat{\beta}_n = (X_n' Z_n W Z_n' X_n)^{-1} X_n' Z_n W Z_n' (X_n\beta_0 + \varepsilon_n)$$
$$\Rightarrow \quad \hat{\beta}_n = \beta_0 + (X_n' Z_n W Z_n' X_n)^{-1} X_n' Z_n W Z_n' \varepsilon_n$$

Is it Unbiased?
We don't know yet, we shall see. Taking the expectation:

$$\mathbb{E}(\hat{\beta}_n) = \beta_0 + \mathbb{E}\left[(X_n' Z_n W Z_n' X_n)^{-1} X_n' Z_n W Z_n' \varepsilon_n\right]$$

(Note: Evaluating this expectation to $0_{p\times 1}$ is not straightforward due to the non-linear way the random variables interact inside the inverse matrix).

◇ **Student's Notes**

The decomposition:

$$\hat{\beta}_n - \beta_0 = \underbrace{(X'ZWZ'X)^{-1}X'ZWZ'}_{\text{"IV projection matrix"}} \underbrace{\varepsilon}_{\text{error}}$$

This is analogous to the OLS decomposition $\hat{\beta}^{OLS} - \beta_0 = (X'X)^{-1}X'\varepsilon$, but with the instruments filtering the projection.

Why unbiasedness is not straightforward:
For OLS with fixed regressors, $\mathbb{E}[(X'X)^{-1}X'\varepsilon] = (X'X)^{-1}X'\mathbb{E}[\varepsilon] = 0$ because X is non-random.
For IV (and OLS with random regressors):

- X, Z, and ε are *all random*
- The matrix inverse $(X'ZWZ'X)^{-1}$ is a highly nonlinear function of the random data
- $\mathbb{E}[f(X,Z)\cdot\varepsilon] \neq \mathbb{E}[f(X,Z)]\cdot\mathbb{E}[\varepsilon]$ in general (even if ε is mean-zero)

Conclusion: IV estimators are generally **biased in finite samples**, but **consistent** (unbiased asymptotically).

! Watch Out

Finite-sample bias of IV can be severe!
The bias of IV is approximately:

$$\text{Bias}(\hat{\beta}^{IV}) \approx \frac{\sigma^2}{n} \cdot \frac{q-p+2}{\text{concentration parameter}}$$

Key points:

- Bias increases with the *number of instruments* (q)
- Bias increases when instruments are *weak* (low correlation with X)
- With many weak instruments, IV can be *more biased than OLS*!

This is the "many instruments" / "weak instruments" problem—a major practical concern in applied econometrics.

9 Consistency of the IV Estimator

◇ Student's Notes

Proving consistency (preview):
From $\hat{\beta}_n - \beta_0 = (X'ZWZ'X)^{-1}X'ZWZ'\varepsilon$:

$$\hat{\beta}_n - \beta_0 = \left(\frac{X'Z}{n}W\frac{Z'X}{n}\right)^{-1}\frac{X'Z}{n}W\frac{Z'\varepsilon}{n}$$

By LLN (under appropriate conditions):

$$\frac{X'Z}{n} \xrightarrow{p} \mathbb{E}[X'Z] =: Q_{XZ} \quad \text{(relevance: full rank)}$$

$$\frac{Z'\varepsilon}{n} \xrightarrow{p} \mathbb{E}[Z'\varepsilon] = 0 \quad \text{(exogeneity)}$$

Therefore:

$$\hat{\beta}_n - \beta_0 \xrightarrow{p} (Q'_{XZ} W Q_{XZ})^{-1} Q'_{XZ} W \cdot 0 = 0$$

So $\hat{\beta}_n \xrightarrow{p} \beta_0$. **Consistency holds for any positive definite W!**

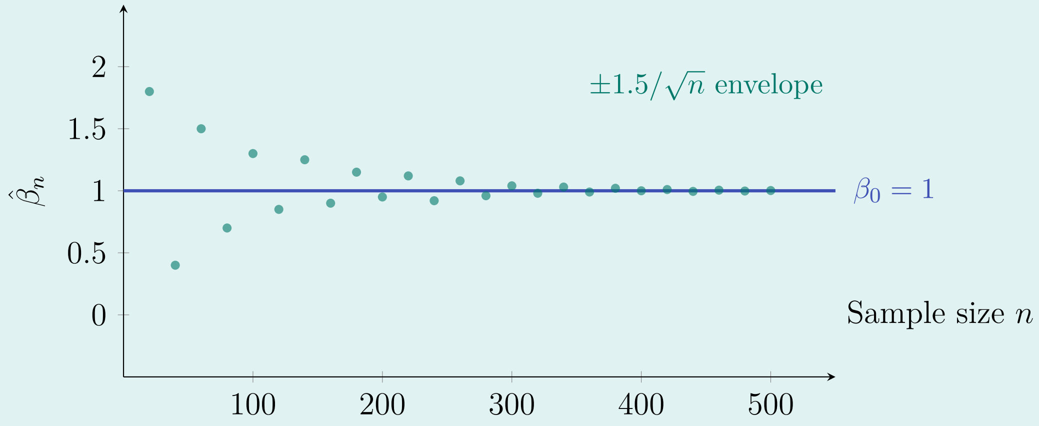


Figure 7: IV estimates converge to the true value β_0 as sample size increases. The "funnel" shape reflects the $1/\sqrt{n}$ rate of convergence.

Connection to the GMM Framework

★ Intuition

IV estimation is a special case of GMM!
The **Generalised Method of Moments** framework:

1. Define population moment conditions: $\mathbb{E}[g(Z_i, \beta_0)] = 0$
2. Form sample moments: $g_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g(Z_i, \beta)$
3. Minimise: $\hat{\beta} = \arg\min_{\beta} \; g_n(\beta)' W \, g_n(\beta)$

For IV in the linear model:

$$g(Z_i, \beta) = z_i(y_i - x_i'\beta) \quad \Rightarrow \quad g_n(\beta) = \frac{1}{n} Z'(Y - X\beta)$$

The moment condition $\mathbb{E}[z_i \varepsilon_i] = 0$ is exactly the IV exogeneity assumption!

GMM is much more general:

- Works for nonlinear models
- Can combine multiple types of moment conditions
- Optimal weighting gives efficient estimation
- Provides a framework for testing (over-identifying restrictions)

Quick-Reference Summary

◇ Student's Notes

Lecture 6 narrative arc:

Topic	What was accomplished
Endogeneity	OLS fails when $\mathbb{E}(X'\varepsilon) \neq 0$; need instruments
Moment conditions	$\mathbb{E}[Z'(Y - X\beta_0)] = 0$ gives q equations in p unknowns
Identification	Need $q \geq p$ and $\text{rank}(\mathbb{E}[Z'X]) = p$
GMM objective	Minimise $g_n(\beta)'Wg_n(\beta)$ where $g_n = \frac{1}{n}Z'(Y - X\beta)$
IV estimator	$\hat{\beta}^{IV} = (X'ZWZ'X)^{-1}X'ZWZ'Y$
Properties	Biased in finite samples, but consistent for any $W > 0$

Key formulas:

Estimator	Formula
General IV	$(X'ZWZ'X)^{-1}X'ZWZ'Y$
Just-identified IV	$(Z'X)^{-1}Z'Y$
2SLS	$(X'P_ZX)^{-1}X'P_ZY$ where $P_Z = Z(Z'Z)^{-1}Z'$

Figures in this lecture:

Fig.	**Content**
1	Two requirements for valid instruments
2	Under/just/over-identification cases
3	GMM as minimising distance to zero
4	Euclidean vs. Mahalanobis distance geometry
5	OLS vs. IV: how instruments filter endogeneity
6	2SLS geometric interpretation (projection)
7	Consistency: estimates converging to β_0