Advanced Econometrics
07 Generalized Method of Moments (GMM)

Eduard Brüll

Fall 2025

Lecture 7

Advanced Econometrics

Literature: Hansen (1982), Greene Ch. 14, Wooldridge Ch. 14

7.1 Review: Moments of a Distribution

Reminder: What is an Expected Value?

- The **expected value** (mean) of a random variable X is its theoretical long-run average:

$$\mathbf{E}[X] = \begin{cases} \int_{-\infty}^{\infty} x f_X(x)\, dx, & \text{if } X \text{ is continuous,} \\ \sum_{x \in \mathcal{X}} x P(X = x), & \text{if } X \text{ is discrete.} \end{cases}$$

- $f_X(x)$: population density (pdf or pmf).
- $\mathbf{E}[X]$ exists if $\int |x| f_X(x)\, dx < \infty$.
- Example: fair die $\Rightarrow \mathbf{E}[X] = (1 + 2 + 3 + 4 + 5 + 6)/6 = 3.5$.

What Are Moments?

- The n^{th} moment of X:

$$\mu_n' = \mathbf{E}[X^n]$$

- Examples:

$$\mathbf{E}[X] \text{ (mean)}, \quad \mathbf{E}[X^2] \text{ (2nd moment)}, \quad \mathbf{E}[X^3], \mathbf{E}[X^4], \ldots$$

- Moments summarize the **shape** of a distribution:
  - 1st moment: location
  - 2nd: spread
  - 3rd: skewness (asymmetry)
  - 4th: kurtosis (tail thickness)

Central (or Centered) Moments

- Centered around the mean:

$$\mu_n = \mathbf{E}[(X - \mathbf{E}[X])^n]$$

- Examples:

$\mu_1 = 0$	(first central moment)
$\mu_2 = \mathbf{var}(X)$	(second moment = variance)
μ_3 measures skewness	(asymmetry)
μ_4 measures kurtosis	(tail thickness)

- For random vectors:

$$\Sigma = \mathbf{E}[(x - \mathbf{E}[x])(x - \mathbf{E}[x])']$$

The Sample Analogue of an Expectation

- The population mean $\mathbf{E}[X]$ depends on the unknown $f_X(x)$.
- Replace the population distribution by its **sample analogue**:

$$\mathbf{E}[X] \;\Rightarrow\; \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- More generally, for any function $g(X)$:

$$\mathbf{E}[g(X)] \;\Rightarrow\; \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

- This idea underlies all **moment estimators**.

Moments and Moment Conditions

- **Theoretical moment conditions:** Many economic models imply that for the true parameter θ_0 and a collection of observed data Z_i (e.g let Z_i cotanin y_i, X_i, etc.)

$$\mathbf{E}[g(Z_i, \theta_0)] = 0.$$

  Example: Exogeneity $\Rightarrow \mathbf{E}[X_i \varepsilon_i] = 0$

- **Sample analogues:** In the data, replace expectations by sample averages:

$$\bar{g}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(Z_i, \theta) \approx 0.$$

7.2 Why GMM?

Why GMM? The Big Picture

- **OLS:** Assumes exogeneity and a linear model

$$\mathbf{E}[X_i\varepsilon_i] = 0 \quad \Rightarrow \quad \hat{\beta}_{\text{OLS}} = (X'X)^{-1}X'y.$$

  Relies on one specific moment condition linking X_i and ε_i.

- **Maximum Likelihood Estimation:** Requires assumptions on the full distribution $f(y_i|X_i,\theta)$.
  - Efficient if correctly specified.
  - But sensitive to misspecification.

- **GMM:** Uses only the parts of the model we are confident about — its **moment conditions**:

$$\mathbf{E}[g(Z_i,\theta_0)] = 0.$$

- Provides a unifying framework that includes OLS, IV, 2SLS, and others as special cases.

The Method of Moments: Intuition

- Suppose we have L known **moment conditions** in the population:

$$\mathbf{E}[g^1(X,\theta)] = 0,\ \mathbf{E}[g^2(X,\theta)] = 0,\ \ldots,\ \mathbf{E}[g^L(X,\theta)] = 0.$$

- Replace population expectations by their **sample analogues**:

$$\bar{g}_n^l(\theta) = \frac{1}{n}\sum_{i=1}^{n} g^l(x_i,\theta) \approx 0.$$

- Solve $\bar{g}_n(\theta) = 0$ for θ.
  - $L = K$: exactly identified $\Rightarrow$ **Method of Moments**.
  - $L > K$: overidentified $\Rightarrow$ **Generalized Method of Moments (GMM)**.

Example: Estimating Mean and Variance

A simple, exactly identified method-of-moments (MM) example:

$$\mathbf{E}[X - \mu] = 0,$$
$$\mathbf{E}[(X - \mathbf{E}[X])^2 - \sigma^2] = 0.$$

Sample analogues:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

- $\hat{\sigma}^2$ is biased but consistent.
- **Shows the core idea:** replace expectations by averages.

What GMM Does

Core idea

Theoretical moment conditions:

$$\mathbf{E}[g(Z_i, \theta_0)] = 0.$$

GMM chooses $\hat{\theta}$ to make the corresponding sample moments as close to zero as possible:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\ \bar{g}_n(\theta)' W_n \bar{g}_n(\theta), \qquad \bar{g}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} g(Z_i, \theta).$$

- Each valid moment condition contributes information about θ.
- Exactly identified ($L = K$): solves $\bar{g}_n(\theta) = 0$.
- Overidentified ($L > K$): combines moments efficiently via W_n.

Overidentified Example: Setup

- Suppose X satisfies $\mathbf{E}(X) = \mathbf{E}(X^2) - \mathbf{E}(X)^2 = \lambda$.
- That is, both the mean and the variance of X equal λ.
- This is a property of a Poisson random variable, but we do **not** assume X is Poisson.
- We simply use these two population relationships as **moment conditions**.

$$\mathbf{E}[X - \lambda] = 0 \qquad (1)$$

$$\mathbf{E}[(X - \mathbf{E}[X])^2 - \lambda] = 0 \qquad (2)$$

Two moment conditions for one parameter $\lambda \Rightarrow$ **overidentified** system.

Sample Moment Conditions

Replace population expectations by sample averages:

$$\hat{g}_1(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \lambda) = 0,$$

$$\hat{g}_2(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left[(x_i - \bar{x})^2 - \lambda\right] = 0.$$

- These are two equations in one unknown λ.
- Generally, there is no single λ satisfying both exactly.
- Hence, the system is **overidentified**.

$$\hat{\lambda}_1 = \bar{x}, \quad \hat{\lambda}_2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

Most likely, $\hat{\lambda}_1 \neq \hat{\lambda}_2$.

Solving the Overidentified Problem (GMM)

- There is no single λ that sets both sample moments to zero.
- The idea of GMM: find $\hat{\lambda}$ that makes the sample moments **as close to zero as possible**.

Define the **criterion function**:

$$q(\lambda) = n\,\hat{g}(\lambda)^\top \mathbf{W}\,\hat{g}(\lambda),$$

where

$$\hat{g}(\lambda) = \begin{bmatrix} \hat{g}_1(\lambda) \\ \hat{g}_2(\lambda) \end{bmatrix}, \qquad \mathbf{W} \text{ is a weighting matrix.}$$

- $\mathbf{W} = \mathbf{I}$ gives equal weight to both moment conditions.
- $\mathbf{W} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$ would only consider the first moment.
- The optimal $\mathbf{W}$ minimizes the asymptotic variance of $\hat{\lambda}$.

GMM estimator:

$$\hat{\lambda}_{GMM} = \arg\min_{\lambda} J(\lambda).$$

Properties of GMM Estimators

- **Law of Large Numbers:** Sample moment conditions converge to their population counterparts:

$$\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(x_i, \theta) \xrightarrow{p} \mathbf{E}[g(X, \theta)].$$

- **Central Limit Theorem:** Sample moments are asymptotically normal:

$$\sqrt{n}(\bar{g}_n(\theta) - \mathbf{E}[g(X, \theta)]) \xrightarrow{d} \mathcal{N}(0, Q),$$

  where $Q = \mathbf{cov}(g(X, \theta))$ (adjusted under heteroskedasticity or clustering).

Implication

These properties carry over to $\hat{\theta}$, the GMM estimator solving $\bar{g}_n(\hat{\theta}) = 0$.

Outlook: Where We Are Going with GMM?

- How do we pick the best W_n for our GMM estimator?

$$\hat{\theta}_{GMM} = \arg\min_{\theta} \ \bar{g}_n(\theta)' W_n \bar{g}_n(\theta).$$

The choice of the weighting matrix W_n determines how efficiently we use the available information.

- Choosing the **optimal** W_n, and proving efficiency and inference results, will be the main task in the later part of the lecture.
- **Historical note:** This optimal weighting and efficiency result is what earned **Lars Peter Hansen (Nobel Prize, 2013)** recognition for developing GMM as a unifying estimation framework.

Next Steps

1. Build intuition from simple, exactly identified **MM** examples.
2. Then generalize to the efficient (two-step) **GMM** estimator.

Note on Notation in Greene

- In Greene's notation, the sample moment functions $\bar{g}_n(\lambda)$ are written as $m(\lambda)$.
- Each component corresponds to one sample moment condition:

$$0 = -\lambda + \frac{1}{n}\sum_{i=1}^{n} x_i = m_1(\lambda),$$

$$0 = -\lambda + \frac{1}{n}\sum_{i=1}^{n} (x_i - \lambda)^2 = m_2(\lambda).$$

- The criterion function then becomes:

$$q(\lambda, \mathbf{W}) = n\, m(\lambda)^\top \mathbf{W}\, m(\lambda), \qquad m(\lambda) = \begin{bmatrix} m_1(\lambda) \\ m_2(\lambda) \end{bmatrix}.$$

7.3 Method of Moments - Least Squares

OLS as a Method of Moments Estimator

- **Recall from the CEF decomposition (Lecture 3):** For the linear projection

$$Y_i = X_i'\beta + \varepsilon_i,$$

exogeneity implies the weaker condition of **uncorrelatedness**:

$$\mathbf{E}[\varepsilon_i \mid X_i] = 0 \;\Rightarrow\; \mathbf{E}[X_i\varepsilon_i] = 0.$$

- This yields $K+1$ **population moment conditions**:

$$\mathbf{E}[X_i(Y_i - X_i'\beta)] = 0.$$

OLS as a Method of Moments Estimator (continued)

- Expanding the expectation $\mathbf{E}[X_i(Y_i - X_i'\beta)] = 0.$ gives

$$\mathbf{E}[X_i Y_i] - \mathbf{E}[X_i X_i']\,\beta = 0,$$

  which can be rearranged as

$$\mathbf{E}[X_i X_i']\,\beta = \mathbf{E}[X_i Y_i].$$

- The sample analog replaces expectations with averages:

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i X_i'\right)\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} X_i Y_i.$$

- Multiplying both sides by n and rearranging yields the familiar OLS estimator:

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y.$$

Second Set of Moment Conditions: Variance

- After estimating β from the first set of $(K+1)$ moment conditions
$$\mathbf{E}[X_i(Y_i - X_i'\beta)] = 0,$$
we can use the residuals to form a **second set of moment conditions** for the variance.
- Under homoskedasticity,
$$\mathbf{E}[\varepsilon_i^2 - \sigma^2] = 0,$$
i.e. a single moment condition identifies σ^2.
- More generally (e.g. in feasible GLS or heteroskedasticity modeling),
$$\mathbf{E}[Z_i(\varepsilon_i^2 - \sigma^2(X_i))] = 0,$$
where Z_i is a set of instruments or functions of X_i that enter the variance equation.
- The sample analogs of these conditions yield estimators of the variance parameters after $\hat{\beta}$ is obtained.

7.4 Instrumental Variables

Motivation: Endogeneity and Bias

- So far, OLS relied on the assumption:

$$\mathbf{E}[\varepsilon_i \mid X_i] = 0 \quad \Rightarrow \quad \mathbf{E}[X_i \varepsilon_i] = 0.$$

- But if any regressor x_{ij} is correlated with the error:

$$\mathbf{E}[X_i \varepsilon_i] \neq 0,$$

  OLS becomes biased and inconsistent.

- **Example:**
  - Education $\rightarrow$ wage regression: ability is unobserved.
  - Ability affects both education and wages $\Rightarrow$ endogeneity.

Question

How can we estimate causal effects when regressors are endogenous?

Idea: Instrumental Variables (IV)

- Find variables Z_i (instruments) that satisfy:
  1. **Relevance:** correlated with the endogenous regressor

     $$\mathbf{cov}(Z_i, X_i) \neq 0$$

  2. **Exogeneity:** uncorrelated with the structural error

     $$\mathbf{cov}(Z_i, \varepsilon_i) = 0$$

- Then Z_i provides variation in X_i that is “as if exogenous.”
- The idea: use Z_i to isolate the exogenous component of X_i.

Exogenous Regressors Are Also Instruments

- Recall that OLS relied on exogeneity:

$$\mathbf{E}[X_i \varepsilon_i] = 0.$$

- In IV estimation, we replace (or augment) X_i by instruments Z_i satisfying:

$$\mathbf{E}[Z_i \varepsilon_i] = 0.$$

- **Important:** Any exogenous regressor in X_i automatically satisfies this condition. It can stay in Z_i as its **own instrument.**

$$Z_i = [X_i^{exog}, \; Z_i^{other}]$$

- Hence, we only need additional instruments for the **endogenous** regressors.

Implication

When specifying Z_i, always include all exogenous X_i; only add new instruments for the endogenous variables.

Philip G. Wright (1928) *"The Tariff on Animal and Vegetable Oils"*

- First known **empirical use of instrumental variables (IV)** in economics.
- Used **exogenous supply shifters** (tariffs, transport costs) as instruments to estimate demand elasticities for oil products.
- Appendix B develops the IV method with help from his son, **Sewall Wright**, a biostatistician, who introduced the same algebraic logic in genetics through **path analysis** (causal diagrams).

Key idea

Identify demand by exploiting **supply-side variation** that is uncorrelated with demand shocks: the fundamental IV logic we still use today.

References:
Wright, P. G. (1928), *The Tariff on Animal and Vegetable Oils*.
Stock, J. H. (2003), “Who Invented Instrumental Variable Regression?”
Cunningham, S. (2021), *Causal Inference: The Mixtape*, Ch. 7.

Modern Example: Customer Satisfaction and Loyalty

Huang, G. & Sudhir, K. (2021). *The Causal Effect of Service Satisfaction on Customer Loyalty*. Management Science, 67(1), 317–341.

Research question: What is the causal effect of service satisfaction on customer loyalty?

Challenge: Satisfaction may be endogenous (e.g. unobserved traits of customers, reverse causality).

Instrument: Use variation in exogenous service shocks (e.g. unexpected disruptions or external factors) that affect satisfaction but are plausibly unrelated to demand or loyalty directly.

Rationale:

- Exogenous shocks influence satisfaction not via customers' latent types.
- They shift satisfaction but (arguably) don't directly shift loyalty except through satisfaction.

Sources of Endogeneity

Endogeneity arises whenever regressors are correlated with the error term:

$$\mathbf{E}[X_i \varepsilon_i] \neq 0.$$

Main sources:

1. **Omitted Variable Bias (OVB):** Unobserved factor affects both X and Y.
   Example: Ability affects both education and earnings.
2. **Simultaneity:** X and Y determined together.
   Example: Price and quantity in supply–demand models.
3. **Measurement Error:** Mismeasured regressors create correlation with ε.
   $\rightarrow$ **Special case - Attenuation Bias:** Bias toward zero.
4. **Lagged Dependent Variable:** y_{t-1} correlated with error term u_t in dynamic panels.

Omitted Variable Bias (OVB)

Setup: Partition the full regressor matrix as

$$X = [\,X_1\ X_2\,],$$

where X_1 are the included regressors and X_2 are the omitted ones. The true model is

$$y = X_1\beta_1 + X_2\beta_2 + \varepsilon.$$

If we estimate the short regression omitting X_2,

$$y = X_1\tilde{\beta}_1 + \tilde{\varepsilon},$$

the OLS estimator $\tilde{\beta}_1$ will generally be biased because the omitted block X_2 can be correlated with the included block X_1.

Intuition:

- Omitted variables that affect y and are correlated with X_1 violate $\mathbf{E}[X_1'\varepsilon] = 0$.
- Their effect is partially attributed to X_1, distorting $\tilde{\beta}_1$.
- Example: Unobserved ability affects both education (X_1) and earnings (y).

OVB via Frisch-Waugh-Lovell Decomposition

Using the FWL theorem (see Lecture 4), the coefficient from the short regression is

$$\tilde{\beta}_1 = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2.$$

Interpretation:

- The bias term $(X_1'X_1)^{-1}X_1'X_2\beta_2$ shows how the omitted regressors X_2 project onto the included regressors X_1.
- Bias arises only if both:

$$\begin{aligned} X_1'X_2 &\neq 0 \quad \text{(correlation between regressors)} \\ \beta_2 &\neq 0 \quad \text{(omitted variables matter for } y\text{)}. \end{aligned}$$

- In the scalar one-omitted-variable case:

$$\text{Bias} = \beta_2 \frac{\mathbf{cov}(x_1, x_2)}{\mathbf{var}(x_1)}.$$

Simultaneity: The Supply and Demand Example

Market equilibrium:

$$Q^d = \alpha_1 - \alpha_2 P + u_d \qquad \text{(demand)}$$
$$Q^s = \beta_1 + \beta_2 P + u_s \qquad \text{(supply)}$$
$$Q^d = Q^s = Q \qquad \text{(equilibrium condition)}$$

Solve for the equilibrium price and quantity:

$$\alpha_1 - \alpha_2 P + u_d = \beta_1 + \beta_2 P + u_s$$

$$\Rightarrow \quad P = \frac{\alpha_1 - \beta_1 + u_d - u_s}{\alpha_2 + \beta_2}.$$

Substitute this price into the demand equation:

$$Q = \alpha_1 - \alpha_2 P + u_d = \alpha_1 - \alpha_2 \frac{\alpha_1 - \beta_1 + u_d - u_s}{\alpha_2 + \beta_2} + u_d.$$

- Q depends on both u_d and u_s.
- Hence P and Q are jointly determined: P correlated with the demand shock u_d.
- $\Rightarrow$ OLS of Q on P gives a biased estimate of the demand slope.

Solution to Simultaneity: Instrumental Variables

We look for an instrument Z that satisfies:

$$\underbrace{\mathbf{cov}(Z, P) \neq 0}_{\text{Relevance: } Z \text{ shifts supply}}, \qquad \underbrace{\mathbf{cov}(Z, u_d) = 0}_{\text{Exogeneity: } Z \text{ does not affect demand directly}}.$$

- Z affects equilibrium price P only through its effect on supply.
- Z is unrelated to unobserved demand shocks u_d.
- **Intuitively:** Z provides variation in P that is "as if random" from the perspective of demand.

Economic Interpretation

- **Valid instrument:** a variable that moves the equilibrium point along the demand curve by shifting the supply curve.
- **Example:** Weather, input costs, or policy shocks changing producers' behavior but not consumers' preferences.

Dynamic Models and the Lagged Dependent Variables

Model:

$$y_{it} = \rho y_{i,t-1} + x_{it}'\beta + u_{it}, \quad u_{it} = \mu_i + \nu_{it}$$

- $y_{i,t-1}$ is correlated with μ_i, the individual fixed effect μ_i.
- Even if $\mathbf{E}[\nu_{it}] = 0$, we get:

$$\mathbf{E}[y_{i,t-1}u_{it}] \neq 0.$$

- This violates the exogeneity condition.
- Common solution: **First-difference** the equation:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta x_{it}'\beta + \Delta \nu_{it},$$

and use instruments like $y_{i,t-2}$ (Arellano–Bond).

Measurement Error and Attenuation Bias

True model:

$$y_i = \beta x_i^* + u_i, \quad \text{but we observe } x_i = x_i^* + v_i,$$

where v_i is classical measurement error:

$$\mathbf{E}[v_i] = 0, \quad \mathbf{cov}(x_i^*, v_i) = 0, \quad \mathbf{cov}(u_i, v_i) = 0.$$

OLS with observed x_i:

$$\hat{\beta} = \frac{\mathbf{cov}(y_i, x_i)}{\mathbf{var}(x_i)}.$$

Substitute and expand:

$$\mathbf{cov}(y_i, x_i) = \mathbf{cov}(\beta x_i^* + u_i,\ x_i^* + v_i) = \beta\ \mathbf{var}(x_i^*).$$

But

$$\mathbf{var}(x_i) = \mathbf{var}(x_i^* + v_i) = \mathbf{var}(x_i^*) + \mathbf{var}(v_i).$$

Measurement Error and Attenuation Bias (contd.)

Expected OLS coefficient:

$$\mathbf{E}[\hat{\beta}] = \beta \cdot \frac{\mathbf{var}(x_i^*)}{\mathbf{var}(x_i^*) + \mathbf{var}(v_i)} = \beta \cdot \lambda, \quad 0 < \lambda < 1.$$

Interpretation:

- $\lambda = \frac{\text{signal}}{\text{signal+noise}}$
- Measurement error inflates $\mathbf{var}(x_i)$ but not $\mathbf{cov}(y_i, x_i)$
- $\Rightarrow$ Estimated slope shrinks toward zero

Intuition

Noisy regressors mix signal and noise $\Rightarrow$ weaker correlation with y_i $\Rightarrow$ slope estimate pulled toward zero.

Method of Moments Perspective on IV

- Recall the structural model:

$$y_i = x_i'\beta + \varepsilon_i$$

- OLS moment condition (fails with endogeneity):

$$\mathbf{E}[x_i\varepsilon_i] = 0.$$

- IV replaces this by valid instruments:

$$\mathbf{E}[z_i\varepsilon_i] = 0.$$

- Substitute $\varepsilon_i = y_i - x_i'\beta$:

$$\mathbf{E}[z_i(y_i - x_i'\beta)] = 0.$$

- These are L moment conditions for K parameters.

Solving the IV Moment Conditions

- Expand:

$$\mathbf{E}[z_i y_i] - \mathbf{E}[z_i x_i']\beta = 0 \quad \Rightarrow \quad \mathbf{E}[z_i x_i']\beta = \mathbf{E}[z_i y_i].$$

- Under full rank of $\mathbf{E}[z_i x_i']$, the population solution is:

$$\beta = \left(\mathbf{E}[z_i x_i']\right)^{-1} \mathbf{E}[z_i y_i].$$

- Replace expectations by sample averages:

$$\hat{\beta}_{IV} = \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right).$$

Matrix Notation and the IV Estimator

Let

$$Z = \begin{bmatrix} z_1' \\ \vdots \\ z_n' \end{bmatrix}, \quad X = \begin{bmatrix} x_1' \\ \vdots \\ x_n' \end{bmatrix}, \quad y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Then the sample IV estimator is:

$$\boxed{\hat{\beta}_{IV} = (Z'X)^{-1} Z'y.}$$

- If $L = K$ (exactly identified): this is the simple IV estimator.
- If $L > K$ (overidentified): **2SLS** covers this case!

Two-Stage Least Squares (2SLS)

Stage 1: Regress endogenous regressor(s) on instruments:

$$X = Z\Pi + v \quad \Rightarrow \quad \hat{X} = Z\hat{\Pi}.$$

Stage 2: Regress y on predicted values $\hat{X}$:

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'y = (X'P_ZX)^{-1}X'P_Zy,$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is the projection onto the instrument space.

Interpretation

2SLS isolates the exogenous variation in X explained by Z and uses it to estimate the causal effect of X on y.

IV as GMM problem

- Recall the 2SLS estimator:

$$\hat{\beta}_{2SLS} = (X'P_Z X)^{-1} X'P_Z y, \qquad P_Z = Z(Z'Z)^{-1}Z'$$

- 2SLS is a special case of GMM with moment conditions

$$\mathbf{E}[Z_i(y_i - X_i'\beta)] = 0$$

and weighting matrix

$$W = (Z'Z/n)^{-1}.$$

- Then the GMM estimator becomes:

$$\hat{\beta}_{GMM} = (X'ZWZ'X)^{-1} X'ZWZ'y = (X'P_Z X)^{-1} X'P_Z y.$$

Key Insight

The projection matrix P_Z in IV is just the GMM weighting matrix that projects residuals onto the instrument space.

Overidentification in IV

Setup:

$$\mathbf{E}[Z_i(y_i - X_i'\beta_0)] = 0, \qquad L > K.$$

- More valid instruments (L) than endogenous regressors (K) $\Rightarrow$ system is **overidentified.**
- GMM (and 2SLS) combine all available instruments efficiently.
- The extra moment conditions can be used to test instrument validity.

Geometric intuition:

- Each instrument defines a "moment hyperplane" in parameter space.
- With overidentification, these hyperplanes may not intersect perfectly.
- GMM chooses $\hat{\beta}$ minimizing the weighted distance to all hyperplanes.

Interpretation

Overidentification is both a blessing (efficiency) and a curse (risk of invalid instruments).

Testing Instrument Validity: Hansen's (1982) J-Test

Purpose: Tests joint validity of instruments when the model is **overidentified** ($L > K$).

$$J = n\,\bar{g}(\hat{\beta})'\hat{W}\,\bar{g}(\hat{\beta}), \quad J \sim \chi^2_{L-K}.$$

Interpretation:

- Checks whether all moment conditions (instruments) are consistent with exogeneity.
- **High J-statistic:** at least one instrument likely invalid (correlated with u_i).
- **Low J-statistic:** cannot reject joint validity.
- Only applies when $L > K$, i.e., there are more instruments than endogenous regressors.

The Problem of Weak Instruments

First stage of 2SLS:

$$X_i = Z_i\pi + v_i$$

where Z_i are instruments and π measures their strength.
If instruments are weak:

- $\text{Cov}(Z, X)$ is small $\Rightarrow$ fitted values $\hat{X}_i = Z_i\hat{\pi}$ barely differ from X_i.
- The 2SLS estimator

  $$\hat{\beta}_{2SLS} = (X'P_ZX)^{-1}X'P_ZY$$

  becomes noisy and biased toward OLS.
- Even if instruments are exogenous ($\text{Cov}(Z, \varepsilon) = 0$), weak relevance ($\text{Cov}(Z, X) \approx 0$) causes:
  - finite-sample bias $\approx$ OLS bias,
  - large standard errors and size distortions in t-tests.

Testing for Weak Instruments

Diagnostics:

- Check **first-stage F-statistic**
- Old rule: $F > 10$ (Staiger & Stock, 1997).
- **Recent work:** higher thresholds needed.
  - Montiel Olea & Pflueger (2013): use **effective** F_{eff}.
  - Lee et al. (2022): reliable 5% t-tests require $F \approx 100$; propose t_F adjustment.
- Multiple endogenous regressors: use **Sanderson–Windmeijer (SW)** partial F or **Kleibergen–Paap rk** statistic.

Refs: Staiger & Stock (1997); Montiel Olea & Pflueger (2013, JBES); Lee et al. (2022, Econometrica); Sanderson & Windmeijer (2016, J. Econometrics).

7.5 GMM, Optimal Weighting and Efficiency

GMM with Overidentification: Criterion & Weights

Setup: For the true parameter β_0,

$$\mathbf{E}[m(y_i, x_i, z_i, \beta_0)] = 0, \qquad \bar{m}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, z_i, \beta).$$

Criterion Function:

$$q_n(\beta) = \bar{m}_n(\beta)' \, W_n \, \bar{m}_n(\beta), \qquad \hat{\beta}_{GMM} = \operatorname*{argmin}_{\beta} q_n(\beta).$$

- Any symmetric positive definite W_n yields a consistent estimator.
- Simple start: $W_n = I$ (equal weight on each moment).
- Why consider other W_n? **Efficiency.** The optimal choice is $W_n \xrightarrow{p} S^{-1}$, where

$$S \xrightarrow{p} \mathbf{var}\left(\sqrt{n}\, \bar{m}_n(\beta_0)\right).$$

Assumptions for Consistency (GMM1–GMM4)

GMM1 **Valid moments & LLN:** $\mathbf{E}[m(y_i, x_i, z_i, \beta_0)] = 0$ and $\bar{m}_n(\beta_0) \xrightarrow{p} 0$ (i.i.d. or weak dependence; finite second moments).

GMM2 **Continuity/Compactness:** $q_n(\beta)$ is continuous in β and the parameter space is compact (or standard conditions ensuring existence of a minimizer).

GMM3 **Identification:**

$$Q(\beta) = \bar{m}(\beta)' W \bar{m}(\beta) \text{ has a unique global minimum at } \beta_0,$$

where $\bar{m}(\beta) = \mathbf{E}[m(y_i, x_i, z_i, \beta)]$ and W is positive definite.

GMM4 **Weight stability:** $W_n \xrightarrow{p} W$ with W positive definite.

Implication

Under **GMM1–GMM4**, $\hat{\beta}_{GMM} \xrightarrow{p} \beta_0$.

Consistency: Step 1–2 — Getting $q_n(\hat{\beta}) \to 0$

Setup:

$$q_n(\beta) = \bar{m}_n(\beta)' W_n \bar{m}_n(\beta), \qquad \hat{\beta} = \arg\min_{\beta} q_n(\beta).$$

Step 1: Moment convergence (GMM1, GMM4) At the true parameter, the sample moments approach zero:

$$\bar{m}_n(\beta_0) \xrightarrow{p} 0.$$

Then, since $q_n(\beta)$ is just a quadratic form,

$$q_n(\beta_0) = \bar{m}_n(\beta_0)' W_n \bar{m}_n(\beta_0) \xrightarrow{p} 0.$$

Interpretation: The objective is small when evaluated at the truth.

Step 2: By minimization,

$$0 \leq q_n(\hat{\beta}) \leq q_n(\beta_0) \xrightarrow{p} 0 \;\Rightarrow\; q_n(\hat{\beta}) \xrightarrow{p} 0.$$

Interpretation: The estimator fits the moment conditions at least as well as the true parameter. Therefore, the minimized criterion also goes to zero.

Consistency: Step 3–4 — From $q_n(\hat{\beta}) \to 0$ to $\hat{\beta} \to \beta_0$

Step 3: Positive definiteness (GMM4) Because $W_n \succ 0$, the quadratic form is zero only when the moments are zero:

$$q_n(\hat{\beta}) \to 0 \;\Rightarrow\; \bar{m}_n(\hat{\beta}) \xrightarrow{p} 0.$$

Interpretation: The only way to make the criterion small is to make the sample moments small.

Step 4: Identification (GMM3) If the population moments equal zero only at the true parameter,

$$\bar{m}(\beta) = 0 \text{ only at } \beta_0,$$

then

$$\bar{m}_n(\hat{\beta}) \xrightarrow{p} 0 \;\Rightarrow\; \boxed{\hat{\beta} \xrightarrow{p} \beta_0.}$$

Summary of Logic

(GMM1) Valid moments $\Rightarrow q_n(\beta_0) \to 0 \Rightarrow q_n(\hat{\beta}) \to 0 \Rightarrow \bar{m}_n(\hat{\beta}) \to 0 \Rightarrow \hat{\beta} \to \beta_0$.

Variance of the Moment Conditions

- Recall: For the true parameter β_0,

$$\mathbf{E}[m(y_i, x_i, z_i, \beta_0)] = 0.$$

- But each $m(y_i, x_i, z_i, \beta_0)$ is a **random vector** — its components vary across observations.
- The sample average

$$\bar{m}_n(\beta_0) = \frac{1}{n}\sum_{i=1}^{n} m(y_i, x_i, z_i, \beta_0)$$

has variance

$$\mathbf{var}(\sqrt{n}\,\bar{m}_n(\beta_0)) = \Phi, \quad \text{where } \Phi = \mathbf{E}[m_i(\beta_0)m_i(\beta_0)'].$$

- Φ summarizes how **noisy** and **correlated** the moment conditions are.

Why Variance Matters for GMM

Intuition

Each moment condition contributes information about β_0, but some are more precise or correlated than others.

- If some $m^l(\cdot)$ have high variance $\Rightarrow$ less reliable.
- If some are correlated $\Rightarrow$ contain overlapping information.
- Therefore, when we form the quadratic form

$$q_n(\beta) = \bar{m}_n(\beta)' W_n \bar{m}_n(\beta),$$

  the weighting matrix W_n should give:
  - more weight to precise (low variance) moments,
  - less weight to noisy or redundant ones.

Covariance Structure of the Moments

$$\Phi = \mathbf{E}[m_i(\beta_0)m_i(\beta_0)'] = \begin{bmatrix} \mathbf{var}(m_1) & \mathbf{cov}(m_1, m_2) & \cdots \\ \mathbf{cov}(m_2, m_1) & \mathbf{var}(m_2) & \cdots \\ \vdots & & \ddots \end{bmatrix}$$

- If moment conditions are independent $\Rightarrow \Phi$ is diagonal.
- If correlated $\Rightarrow$ off-diagonal elements nonzero.
- Estimation efficiency depends on how we incorporate this covariance.

Goal

Choose W that accounts for this covariance to make $\hat{\beta}_{GMM}$ efficient.

The Matrix $\Phi = E[m_i m_i']$

Suppose we have two centered moment conditions m_1 and m_2 with

$$E[m_1] = 0, \qquad E[m_2] = 0.$$

Then their covariance matrix is

$$\Phi = E\left[\begin{pmatrix} m_1 \\ m_2 \end{pmatrix} \begin{pmatrix} m_1 & m_2 \end{pmatrix}\right] = \begin{pmatrix} E[m_1^2] & E[m_1 m_2] \\ E[m_2 m_1] & E[m_2^2] \end{pmatrix}.$$

Why this simplification holds:

$$\mathbf{var}(m_1) = E[(m_1 - E[m_1])^2] = E[m_1^2]$$
$$\mathbf{cov}(m_1, m_2) = E[(m_1 - E[m_1])(m_2 - E[m_2])] = E[m_1 m_2]$$

Intuition

Because each moment condition is defined to have mean zero at the true parameter ($E[m_i(\beta_0)] = 0$), their variance and covariance reduce to simple expectations of products. This is what makes the matrix $\Phi = E[m_i m_i']$ appear throughout the GMM variance formulas.

Properties of the Quadratic Form

- Recall the GMM criterion:

$$q_n(\beta) = \bar{m}_n(\beta)' W_n \bar{m}_n(\beta),$$

  where $\bar{m}_n(\beta) : \mathbb{R}^K \to \mathbb{R}^L$ collects the sample moments.

- Dimensions:

$$q_n(\beta) = \underbrace{\bar{m}_n(\beta)'}_{1\times L} \underbrace{W_n}_{L\times L} \underbrace{\bar{m}_n(\beta)}_{L\times 1} \Rightarrow q_n(\beta) \in \mathbb{R}.$$

- W_n symmetric and positive definite:

$$x' W_n x > 0 \quad \text{for all } x \neq 0.$$

Interpretation

$q_n(\beta)$ is a weighted squared distance between the sample moments and 0.

7.5.1 Asymptotic Distribution of GMM

Goal and Key Objects

Goal: Derive the asymptotic distribution (sampling variability) of the GMM estimator.

$$\hat{\beta}_{GMM} = \arg\min_{\beta} \bar{m}_n(\beta)' W_n \bar{m}_n(\beta), \quad \bar{m}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} m_i(\beta).$$

At the true parameter β_0:

$$\mathbf{E}[m_i(\beta_0)] = 0, \qquad \Gamma = \mathbf{E}\left[\frac{\partial m_i(\beta_0)}{\partial \beta'}\right], \qquad \Phi = \mathbf{E}[m_i(\beta_0) m_i(\beta_0)'].$$

Dimensions:

$$\bar{m}_n(\beta) : L \times 1, \quad \Gamma : L \times K, \quad W_n : L \times L.$$

Intuition

GMM combines L noisy moment conditions to estimate K parameters. We want to understand how $\hat{\beta}_{GMM}$ fluctuates around β_0 as n grows.

Step 1: Linearize the Sample Moments

Use a first-order (mean value) expansion of the sample moments around β_0:

$$\bar{m}_n(\hat{\beta}_{GMM}) \approx \bar{m}_n(\beta_0) + \Gamma_n(\tilde{\beta})(\hat{\beta}_{GMM} - \beta_0),$$

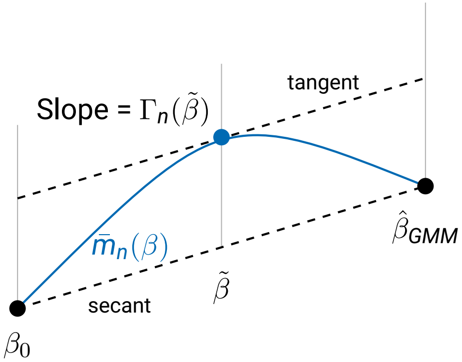
where

$$\Gamma_n(\tilde{\beta}) = \frac{\partial \bar{m}_n(\tilde{\beta})}{\partial \beta'}, \quad \tilde{\beta} \text{ lies between } \hat{\beta}_{GMM} \text{ and } \beta_0.$$

Intuition

We approximate how the sample moments react to small changes in β. The Jacobian Γ plays the same role as the "design matrix" in regression. It captures how informative the moments are.

Step 1: Linearize via the Mean Value Theorem



Intuition

By the Mean Value Theorem, there exists $\tilde{\beta} \in (\beta_0, \hat{\beta}_{\text{GMM}})$ such that the derivative $\bar{G}_n(\tilde{\beta})$ equals the average slope between the endpoints. This $\tilde{\beta}$ is the linearization point used to approximate $\tilde{m}_n(\hat{\beta})$ around β_0 in the GMM derivation.

Mean Value vs. Taylor Approximation

Why we say "Mean Value Approximation" rather than "Taylor Expansion":

- The true parameter β_0 is **unknown**, so we cannot directly evaluate $\Gamma_n(\beta_0) = \frac{\partial \bar{m}_n(\beta)}{\partial \beta'}\Big|_{\beta_0}$.
- The **Mean Value Theorem** ensures there exists some point $\tilde{\beta}$ between β_0 and $\hat{\beta}_{GMM}$ such that

$$\bar{m}_n(\hat{\beta}_{GMM}) = \bar{m}_n(\beta_0) + \Gamma_n(\tilde{\beta})\,(\hat{\beta}_{GMM} - \beta_0).$$

- As $n \to \infty$, $\hat{\beta}_{GMM} \xrightarrow{p} \beta_0$, so $\Gamma_n(\tilde{\beta}) \xrightarrow{p} \Gamma$. This lets us **treat it like a first-order Taylor expansion asymptotically**.

Key takeaway

The "mean value approximation" is the mathematically valid form of the linearization when the true parameter is unknown.

Step 2: First-Order Condition (FOC)

The estimator minimizes the quadratic form

$$q_n(\beta) = \bar{m}_n(\beta)' W_n \bar{m}_n(\beta).$$

Differentiate with respect to β and set to zero:

$$\frac{\partial q_n(\hat{\beta}_{GMM})}{\partial \beta} = 2\,\Gamma_n(\hat{\beta}_{GMM})' W_n \bar{m}_n(\hat{\beta}_{GMM}) = 0.$$

Intuition

At the minimum, the weighted average of sample moments (the "residual moments") must be orthogonal to the gradient direction $\Gamma_n' W_n$. This ensures that we are at the point where the sample moments are as close to zero as possible under W_n.

Step 3: Substitute the Linear Approximation

Plug the linearized form of $\bar{m}_n(\hat{\beta}_{GMM})$ into the FOC:

$$\Gamma_n(\hat{\beta}_{GMM})' W_n \left[\bar{m}_n(\beta_0) + \Gamma_n(\tilde{\beta})(\hat{\beta}_{GMM} - \beta_0) \right] \approx 0.$$

Rearranging gives:

$$\hat{\beta}_{GMM} - \beta_0 \approx -\left(\Gamma_n' W_n \Gamma_n\right)^{-1} \Gamma_n' W_n \, \bar{m}_n(\beta_0).$$

Step 4: Replace Sample Terms by Population Limits

Under standard regularity conditions:

$$\Gamma_n(\tilde{\beta}) \xrightarrow{p} \Gamma, \qquad W_n \xrightarrow{p} W.$$

Hence,

$$\boxed{\hat{\beta}_{GMM} - \beta_0 \;\approx\; -(\Gamma' W \Gamma)^{-1} \Gamma' W \bar{m}_n(\beta_0).}$$

Intuition

This linearization says: the GMM estimator is just a weighted linear transformation of the sample moments. Errors in $\bar{m}_n(\beta_0)$ propagate to $\hat{\beta}_{GMM}$ through Γ and W.

Step 5: Scale by $\sqrt{n}$

Multiply both sides by $\sqrt{n}$:

$$\boxed{\sqrt{n}(\hat{\beta}_{GMM} - \beta_0) \approx -(\Gamma' W \Gamma)^{-1} \Gamma' W \sqrt{n}\, \bar{m}_n(\beta_0).}$$

Interpretation

Sampling error in $\bar{m}_n(\beta_0)$ drives the sampling error in $\hat{\beta}_{GMM}$. The term $\Gamma' W$ transforms the moment noise into parameter noise.

Step 6: Apply the Central Limit Theorem

By the multivariate CLT:

$$\sqrt{n}\,\bar{m}_n(\beta_0) \xrightarrow{d} \mathcal{N}(0, \Phi), \qquad \Phi = \mathbf{E}[m_i(\beta_0)m_i(\beta_0)'].$$

Combining with the previous step:

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta_0) \xrightarrow{d} \mathcal{N}\Big(0,\ (\Gamma' W \Gamma)^{-1}\Gamma' W \Phi W \Gamma (\Gamma' W \Gamma)^{-1}\Big).$$

Intuition

Moment fluctuations are asymptotically normal, and the estimator inherits that normality—scaled and rotated by Γ and W.

Step 7: Asymptotic Variance and Efficiency

$$\boxed{\mathrm{Avar}(\hat{\beta}_{GMM}) = (\Gamma' W \Gamma)^{-1} \Gamma' W \Phi W \Gamma (\Gamma' W \Gamma)^{-1}.}$$

Interpretation

- Φ – covariance of moment conditions (noise in the data).
- Γ – sensitivity of moments to parameters (identification strength).
- W – weighting scheme that determines efficiency.

The efficient GMM estimator uses $W = \Phi^{-1}$, minimizing this variance.

Summary of the Derivation

1. **Linearize** sample moments around β_0.
2. **Use FOC** to link $\hat{\beta}$ and $\bar{m}_n(\beta_0)$.
3. **Replace** sample Jacobians by their probability limits.
4. **Scale** by $\sqrt{n}$ to study sampling variation.
5. **Apply CLT** to the sample moments.
6. **Derive** asymptotic normality:

$$\sqrt{n}(\hat{\beta}_{GMM} - \beta_0) \xrightarrow{d} \mathcal{N}(0, V_{GMM}),$$

with V_{GMM} as above.

Applying the General GMM Variance Formula to OLS

General GMM asymptotic variance:

$$V_{GMM} = (\Gamma' W \Gamma)^{-1} \Gamma' W \Phi W \Gamma (\Gamma' W \Gamma)^{-1}.$$

For OLS:

$$m_i(\beta) = x_i(y_i - x_i'\beta) \quad \Rightarrow \quad \Gamma = -\mathbf{E}[x_i x_i'],\ W = I,\ \Phi = \mathbf{E}[x_i x_i' \varepsilon_i^2].$$

Under homoskedasticity:

$$\mathbf{E}[\varepsilon_i^2 \mid X_i] = \sigma^2 \quad \Rightarrow \quad \Phi = \sigma^2\,\mathbf{E}[x_i x_i'].$$

Plug in:

$$V_{OLS} = \sigma^2 (\mathbf{E}[x_i x_i'])^{-1}.$$

Interpretation

The general GMM variance collapses to the textbook OLS variance once we substitute the OLS moment conditions and homoskedasticity.

Sample Analogues: From Population to Data

Population matrices:

$$Q_{xx} = \mathbf{E}[x_i x_i'], \qquad \Phi = \sigma^2 Q_{xx}.$$

Sample analogues:

$$\frac{1}{n} X'X \xrightarrow{p} Q_{xx}, \qquad \hat{\sigma}^2 = \frac{\hat{\varepsilon}'\hat{\varepsilon}}{n-k} \xrightarrow{p} \sigma^2.$$

Hence:

$$\widehat{V}_{OLS} = \hat{\sigma}^2 (X'X/n)^{-1} \xrightarrow{p} V_{OLS}.$$

Why TSLS is a GMM Estimator

Moment conditions:

$$E[z_i(y_i - x_i'\beta)] = 0$$

GMM criterion:

$$Q(\beta) = g_n(\beta)' W g_n(\beta) \quad \text{where} \quad g_n(\beta) = \frac{1}{n} Z'(y - X\beta)$$

Minimization problem:

$$\hat{\beta}_{GMM} = \arg\min_{\beta} (y - X\beta)' Z W Z' (y - X\beta)$$

First-order condition:

$$X'ZWZ'(y - X\hat{\beta}_{GMM}) = 0 \quad \Rightarrow \quad \hat{\beta}_{GMM} = (X'ZWZ'X)^{-1} X'ZWZ'y$$

Special case: If $W = (Z'Z)^{-1}$, then

$$\hat{\beta}_{GMM} = (X'P_Z X)^{-1} X' P_Z y \quad \text{where} \quad P_Z = Z(Z'Z)^{-1} Z'$$

7.5.2 Optimal Weighting and Efficiency

Goal: Find W that minimizes the asymptotic variance V_{GMM}.

$$V_{GMM} = (\Gamma' W \Gamma)^{-1} \Gamma' W \Phi W \Gamma (\Gamma' W \Gamma)^{-1}.$$

The minimizing (optimal) weighting matrix is

$$\boxed{W_{\text{opt}} = \Phi^{-1}}.$$

Substituting W_{opt} yields

$$\boxed{V_{GMM,opt} = (\Gamma' \Phi^{-1} \Gamma)^{-1}.}$$

- This is the smallest possible asymptotic variance among all GMM estimators.
- The corresponding estimator is the **efficient GMM** (or two-step GMM).

Intuition for the Optimal $W_{\text{opt}} = \Phi^{-1}$

- Think of W as telling us how much to “trust” each moment.
- If a moment condition has:
  - high variance $\Rightarrow$ down-weight it,
  - low variance $\Rightarrow$ give it more influence.
- Correlated moments: Φ^{-1} also de-correlates them.

Practical Implementation

1. **Step 1:** Estimate with $W = I$ to get preliminary $\hat{\beta}$.
2. **Step 2:** Estimate $\hat{\Phi}$ using residuals at $\hat{\beta}$.
3. **Step 3:** Re-estimate with $W = \hat{\Phi}^{-1}$ (efficient 2-step GMM).

Review: GMM and the Wald Test Analogy

An Analogy: Both the **GMM criterion function** and the **Wald test** measure how far some sample quantities are from zero, using an appropriate weighting matrix.

$$\underbrace{J_n(\theta)}_{\text{GMM criterion}} = n\,\bar{g}_n(\theta)'\,W_n\,\bar{g}_n(\theta)$$

$$\underbrace{W}_{\text{Wald statistic}} = (R\hat{\beta} - r)'\,[R\,\widehat{\mathrm{Var}}(\hat{\beta})\,R']^{-1}\,(R\hat{\beta} - r)$$

- **GMM:** minimizes the weighted distance of sample moments $\bar{g}_n(\theta)$ from zero.
- **Wald:** measures the weighted distance of estimated restrictions $(R\hat{\beta} - r)$ from zero.
- In both: the weighting matrix gives more weight to *precise* and *less correlated* components.

Efficient (Two-Step) GMM in Practice

Step 1: Use a simple weight (e.g., $W_n = I$) to obtain a preliminary estimate:

$$\hat{\beta}^{(1)} = \arg\min_{\beta} \bar{m}_n(\beta)'\bar{m}_n(\beta).$$

Step 2: Estimate the covariance of the moments:

$$\hat{\Phi}_n = \frac{1}{n}\sum_{i=1}^{n} \hat{m}_i(\hat{\beta}^{(1)})\hat{m}_i(\hat{\beta}^{(1)})', \quad \hat{m}_i(\hat{\beta}^{(1)}) = m(y_i, x_i, z_i, \hat{\beta}^{(1)}).$$

Step 3: Re-estimate using the optimal weight:

$$W_n = \hat{\Phi}_n^{-1}, \quad \hat{\beta}^{(2)} = \arg\min_{\beta} \bar{m}_n(\beta)'W_n\bar{m}_n(\beta).$$

Result:

$$\sqrt{n}(\hat{\beta}^{(2)} - \beta_0) \xrightarrow{d} \mathcal{N}(0,\ (\Gamma'\Phi^{-1}\Gamma)^{-1}).$$

7.6 GMM Applications

Why Economists Like GMM

- **Flexible:** needs only moment conditions — no full likelihood.
- **Unifying:** OLS, IV, 2SLS, dynamic panels all fit in one framework.
- **Theory-based:** estimates parameters implied by equilibrium or optimality.
- **Robust:** valid under heteroskedasticity or mild misspecification.
- **Widely used:**
  - **Macroeconomics:** Structural Models
  - **Finance:** Asset pricing and risk premia
  - **IO:** Demand and cost estimation

Bottom Line

GMM connects **economic theory** to **data** with minimal assumptions.

Structural Models and Moment Conditions

- **Idea:** GMM allows estimation of parameters in theoretical systems of equations where equilibrium conditions imply specific moments.
- Structural models:

$$f(y_i, x_i, \varepsilon_i; \theta_0) = 0 \quad \Rightarrow \quad \mathbf{E}[g(Z_i, \theta_0)] = 0$$

  with $g(\cdot)$ derived from the model's behavioral or equilibrium relations.
- **Examples:**
  - Demand and supply systems
  - Consumption Euler equations
  - Investment models with adjustment costs
- GMM estimates $\hat{\theta}$ such that these model-implied moments match the data.

Example: Consumption Smoothing Intuition

Idea: Consumers prefer smooth consumption over time — spending and saving so that the value of a euro today equals the value of a euro tomorrow.

Basic trade-off:

$$u'(c_t) = \beta(1 + r_{t+1})\ \mathbf{E}_t[u'(c_{t+1})]$$

- $u'(c_t)$ = value of an extra unit of consumption today
- β = how patient the consumer is
- $(1 + r_{t+1})$ = return from saving

Economic meaning:

- If today's marginal utility $>$ expected future value $\rightarrow$ consume less today (save more).
- If it's lower $\rightarrow$ consume more today.

When consumers make these adjustments optimally, the equation holds *on average* in the data.

From Economic Rule to GMM Estimation

Model-implied moment condition:

$$\mathbf{E}_t \left[u'(c_t) \left(\beta(1 + r_{t+1}) u'(c_{t+1}) - u'(c_t) \right) \right] = 0.$$

Step 1: Use data on consumption growth and interest rates to construct the sample analogue of this moment.
Step 2: Find $\hat{\beta}$ (and possibly risk aversion γ) that makes the sample moment as close to zero as possible:

$$\hat{\beta}_{GMM} = \arg\min_{\beta} \bar{g}_n(\beta)' W \bar{g}_n(\beta)$$

Interpretation:

- GMM checks whether consumers' observed choices are consistent with the theory.
- If the model's optimality condition fits the data well, our estimated $\hat{\beta}$ measures how patient consumers are.

Structural Systems and Moment Restrictions

- Consider a simultaneous system:

$$\begin{aligned} y_{1i} &= \alpha_1 y_{2i} + x_{1i}'\beta_1 + u_{1i}, \\ y_{2i} &= \alpha_2 y_{1i} + x_{2i}'\beta_2 + u_{2i}. \end{aligned}$$

- Theoretical model implies cross-equation restrictions such as:

$$\mathbf{E}[z_{1i}u_{1i}] = 0, \quad \mathbf{E}[z_{2i}u_{2i}] = 0.$$

- Stack all equations into a single GMM system:

$$\mathbf{E}[g(Z_i, \theta_0)] = 0, \quad g(Z_i, \theta) = \begin{bmatrix} z_{1i}(y_{1i} - \alpha_1 y_{2i} - x_{1i}'\beta_1) \\ z_{2i}(y_{2i} - \alpha_2 y_{1i} - x_{2i}'\beta_2) \end{bmatrix}.$$

- Allows joint estimation and testing of cross-equation restrictions.

Arellano–Bond (1991): Dynamic Panel GMM

Dynamic panel model:

$$y_{it} = \rho y_{i,t-1} + x_{it}'\beta + \mu_i + \nu_{it}.$$

Problem: $y_{i,t-1}$ correlated with μ_i.

- Difference to remove μ_i:

$$\Delta y_{it} = \rho \Delta y_{i,t-1} + \Delta x_{it}'\beta + \Delta \nu_{it}.$$

- Instruments: earlier lags of y_{it} that remain correlated with $\Delta y_{i,t-1}$ but uncorrelated with $\Delta \nu_{it}$.

$$\mathbf{E}[y_{i,t-s}\, \Delta \nu_{it}] = 0 \quad \text{for } s \geq 2.$$

- GMM stacks these as valid moment conditions:

$$g_i(\theta) = \sum_{t=3}^{T} y_{i,t-2}\,(\Delta y_{it} - \rho \Delta y_{i,t-1} - \Delta x_{it}'\beta).$$

- Efficient estimation uses all available lags and instruments.

Instruments in Arellano–Bond

Example: $T = 5$ periods.

$$\underbrace{\begin{bmatrix} y_{i1} & 0 & 0 \\ y_{i1} & y_{i2} & 0 \\ y_{i1} & y_{i2} & y_{i3} \end{bmatrix}}_{Z_i} \quad \text{instruments for} \quad \begin{bmatrix} \Delta y_{i3} \\ \Delta y_{i4} \\ \Delta y_{i5} \end{bmatrix}$$

- Each row: valid instruments for Δy_{it} using all available lags $y_{i,t-2}, y_{i,t-3}, \ldots$.
- Lower-triangular structure $\Rightarrow$ expanding set of moment conditions.
- GMM combines them efficiently