Linear Models Lecture 15: GMM Estimation and Efficiency

Robert Gulotty

University of Chicago

February 26, 2026

Where We Left Off

In Lecture 14, we introduced IV estimation as a method of moments:

$$\frac{1}{n}\sum_{i=1}^{n} Z_i(Y_i - X_i'\beta) = 0$$

This works when the number of moment conditions ℓ equals the number of parameters k.

But what if we have more instruments than parameters?

- We cannot simultaneously set all moment conditions to zero
- We need a principled way to *get as close as possible*
- This is the **overidentification problem**

Three Reasons to Learn GMM

Three Meta-Lessons from GMM

1. **Semiparametric:** GMM requires only moment conditions — no distributional assumptions. It achieves the *semiparametric efficiency bound* (Chamberlain, 1987).
2. **Efficient:** The optimal weight matrix minimizes asymptotic variance across all GMM estimators. Two-step and iterated GMM achieve this bound in practice.
3. **General:** GMM provides a unified framework for estimation *and* testing. It nests OLS, GLS, IV, and 2SLS as special cases — and extends to nonlinear models and treatment effect heterogeneity.

Moment Equation Models (Hansen §13.2)

Let $g_i(\beta)$ be a known $\ell \times 1$ function of the i^{th} observation and a $k \times 1$ parameter β.

Definition

A **moment equation model** *is defined by*

$$\mathbb{E}[g_i(\beta)] = 0$$

and a parameter space $\beta \in B$.

Identification requires $\ell \geq k$**:**

- $\ell = k$: **Just identified** — exactly enough information
- $\ell > k$: **Overidentified** — excess information (testable restrictions)
- $\ell < k$: **Underidentified** — insufficient information

Example: IV model: $g_i(\beta) = Z_i(Y_i - X_i'\beta)$

Everything Is a Moment Condition

Estimator	**Moment condition $g_i(\beta)$**	**MME**
Sample mean	$Y_i - \mu$	$\hat{\mu} = \bar{Y}$
OLS	$X_i(Y_i - X_i'\beta)$	$\hat{\beta} = (X'X)^{-1}X'Y$
IV	$Z_i(Y_i - X_i'\beta)$	$\hat{\beta} = (Z'X)^{-1}Z'Y$
2SLS	$Z_i(Y_i - X_i'\beta)$ with $W = (Z'Z)^{-1}$	$\hat{\beta}_{2SLS}$
GLS	$\bar{X}_i\Sigma^{-1}(Y_i - \bar{X}_i'\beta)$	$\hat{\beta}_{GLS}$

Key insight: All the estimators we have studied are *method of moments estimators* — they solve $\bar{g}_n(\hat{\beta}) = 0$. GMM generalizes to the overidentified case.

The Overidentification Problem (Hansen §13.4)

Define the sample moment:

$$\bar{g}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} g_i(\beta)$$

When $\ell = k$: We can solve $\bar{g}_n(\hat{\beta}) = 0$ exactly.

When $\ell > k$: There are more equations than unknowns. In general, no β sets $\bar{g}_n(\beta) = 0$.

Intuition: Think of $\mu = Z'Y$, $G = Z'X$, and $\eta = \mu - G\beta$. We want η small. Regressing μ on G:

$$\tilde{\beta} = (G'G)^{-1}G'\mu$$

minimizes $\eta'\eta$. But we can do *better* with weighted least squares...

The GMM Criterion Function

For a positive definite $\ell \times \ell$ weight matrix W, the **GMM criterion function** is:

$$\boxed{J(\beta) = n\,\bar{g}_n(\beta)' W\,\bar{g}_n(\beta)}$$

- $J(\beta) \geq 0$ for all β (since $W > 0$)
- When $W = I_\ell$: $J(\beta) = n\|\bar{g}_n(\beta)\|^2$ (Euclidean distance)
- The factor n is for distributional convenience
- Different choices of W yield different estimators — and different efficiency

For the linear IV model:

$$J(\beta) = n(Z'Y - Z'X\beta)'W(Z'Y - Z'X\beta)$$

The GMM Estimator (Hansen Def. 13.1, Thm. 13.1)

Definition (GMM Estimator)

$$\hat{\beta}_{gmm} = \arg\min_{\beta} J(\beta) = \arg\min_{\beta} \; n\,\bar{g}_n(\beta)' W\, \bar{g}_n(\beta)$$

Theorem (13.1)

For the overidentified linear IV model, the GMM estimator is:

$$\hat{\beta}_{gmm} = (X'ZWZ'X)^{-1}(X'ZWZ'Y)$$

- When $\ell = k$ (just identified): $\hat{\beta}_{\text{gmm}} = (Z'X)^{-1}(Z'Y) = \hat{\beta}_{\text{iv}}$, regardless of W
- When $\ell > k$: the estimator depends on W

2SLS as GMM (Hansen Thm. 13.2)

Theorem (13.2)

If $W = (Z'Z)^{-1}$, *then* $\hat{\beta}_{gmm} = \hat{\beta}_{2sls}$.
Furthermore, if $k = \ell$ *then* $\hat{\beta}_{gmm} = \hat{\beta}_{iv}$.

Proof sketch: Substitute $W = (Z'Z)^{-1}$ into Theorem 13.1:

$$\hat{\beta} = (X'Z(Z'Z)^{-1}Z'X)^{-1}(X'Z(Z'Z)^{-1}Z'Y) = (X'P_ZX)^{-1}(X'P_ZY)$$

where $P_Z = Z(Z'Z)^{-1}Z'$ is the projection matrix — exactly the 2SLS formula.

Takeaway: 2SLS is a *one-step* GMM estimator. But $(Z'Z)^{-1}$ is not always the best weight matrix.

Asymptotic Distribution (Hansen Thm. 13.3)

Let $Q = \mathbb{E}[ZX']$ and $\Omega = \mathbb{E}[ZZ'e^2]$.

Theorem (13.3: Asymptotic Distribution of GMM)

As $n \to \infty$,

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

where

$$\boxed{V_\beta = (Q'WQ)^{-1}(Q'W\Omega WQ)(Q'WQ)^{-1}}$$

- This is a **sandwich form**: "bread" $(Q'WQ)^{-1}$ wraps "meat" $Q'W\Omega WQ$
- Applies for *any* positive definite weight matrix W
- Different $W \Rightarrow$ different $V_\beta \Rightarrow$ different efficiency

What Is the Best Weighting?

Question: Which W minimizes V_β?

Intuition: We want to weight moments that are

- *more informative* (high signal) $\Rightarrow$ weight up
- *less noisy* (low variance) $\Rightarrow$ weight up

The variance of the sample moments is $\Omega = \mathbb{E}[g_i g_i']$. So weight inversely to their variance: $W = \Omega^{-1}$.

Analogy: Just like GLS weights observations by the inverse of $\text{Var}(e_i)$, efficient GMM weights *moment conditions* by the inverse of their covariance.

Efficient GMM (Hansen Thms. 13.4–13.5)

Theorem (13.4: Efficient GMM)

Setting $W = \Omega^{-1}$, *as* $n \to \infty$: $\sqrt{n}(\hat{\beta}_{gmm} - \beta) \xrightarrow{d} N(0, V_\beta)$ *where* $V_\beta = (Q'\Omega^{-1}Q)^{-1}$.

The sandwich collapses: $(Q'\Omega^{-1}Q)^{-1}(\underbrace{Q'\Omega^{-1}\Omega\Omega^{-1}Q}_{Q'\Omega^{-1}Q})(Q'\Omega^{-1}Q)^{-1}$

Theorem (13.5: Efficiency)

For any $W > 0$: $(Q'WQ)^{-1}(Q'W\Omega WQ)(Q'WQ)^{-1} - (Q'\Omega^{-1}Q)^{-1} \geq 0$.

No GMM estimator with these moment conditions can have a smaller asymptotic variance. Chamberlain (1987): this is the **semiparametric efficiency bound**.

When Is 2SLS Efficient? (Hansen Thm. 13.6)

Recall 2SLS uses $W = (Z'Z)^{-1}$, which converges to $(\mathbb{E}[ZZ'])^{-1}$.

The efficient weight is $\Omega^{-1} = (\mathbb{E}[ZZ'e^2])^{-1}$.

These are equal when $\mathbb{E}[e^2|Z] = \sigma^2$ (conditional homoskedasticity):

$$\mathbb{E}[ZZ'e^2] = \sigma^2\mathbb{E}[ZZ']$$

Theorem (13.6)

Under conditional homoskedasticity $\mathbb{E}[e^2|Z] = \sigma^2$, the 2SLS estimator $\hat{\beta}_{2sls}$ is efficient GMM.

Implications:

- Homoskedastic errors $\Rightarrow$ 2SLS is fine, no need for GMM
- Heteroskedastic errors $\Rightarrow$ 2SLS is *inefficient*; use efficient GMM

Two-Step GMM (Hansen §13.10, Thm. 13.7)

Problem: $\Omega = \mathbb{E}[ZZ'e^2]$ is unknown. We need to estimate it.

Two-Step GMM:

1. **Step 1:** Estimate β by 2SLS (using $W = (Z'Z)^{-1}$). Compute residuals $\tilde{e}_i = Y_i - X_i'\hat{\beta}_{2sls}$.
2. **Step 2:** Estimate $\hat{\Omega} = \frac{1}{n}\sum_{i=1}^{n} Z_i Z_i' \tilde{e}_i^2$, set $\hat{W} = \hat{\Omega}^{-1}$, and re-estimate β.

Theorem (13.7)

The two-step GMM estimator is asymptotically efficient:

$$\sqrt{n}(\hat{\beta}_{gmm} - \beta) \xrightarrow{d} N(0, V_\beta) \quad \textit{where} \quad V_\beta = (Q'\Omega^{-1}Q)^{-1}$$

Key point: The initial estimator does not affect the asymptotic distribution.

Iterated GMM (Hansen §13.11)

The two-step estimator's *finite-sample* performance can depend on the initial estimator. To remove this dependence:

Iterated GMM:

1. Start with $\hat{\beta}^{(0)}$ (e.g., 2SLS)
2. Compute $\hat{\Omega}^{(s)} = \frac{1}{n}\sum Z_i Z_i' (\hat{e}_i^{(s)})^2$
3. Re-estimate: $\hat{\beta}^{(s+1)} = (X'Z\hat{\Omega}^{(s)-1}Z'X)^{-1}(X'Z\hat{\Omega}^{(s)-1}Z'Y)$
4. Repeat until convergence

In R's `gmm` package: `type = "twoStep"` (two-step) or `type = "iterative"` (iterated).

Note: Hansen and Lee (2021) show the iterated GMM estimator is unaffected by whether the weight matrix is computed with or without centering.

Covariance Matrix Estimation (Hansen §13.12)

One-step or two-step GMM (general W):

$$\hat{V}_\beta = (\hat{Q}'\hat{W}\hat{Q})^{-1}(\hat{Q}'\hat{W}\hat{\Omega}\hat{W}\hat{Q})(\hat{Q}'\hat{W}\hat{Q})^{-1}$$

where $\hat{Q} = \frac{1}{n}\sum Z_i X_i'$ and $\hat{\Omega}$ uses residuals from the final step.

Efficient iterated GMM ($\hat{W} = \hat{\Omega}^{-1}$):

$$\hat{V}_\beta = \left(\hat{Q}'\hat{\Omega}^{-1}\hat{Q}\right)^{-1}$$

The sandwich simplifies — just like GLS vs. OLS variance formulas.

Practical advice: Always report robust (sandwich) standard errors for one-step GMM. For efficient GMM, the simplified formula is already robust.

Why Missing Data Needs GMM

Meta-lesson #1: GMM is semiparametric.

Missing data on regressors is pervasive (Abrevaya & Donald 2017, surveying top journals 2006–2008): ∼50% of empirical papers in JLE/QJE have missing data; ∼70% use the *complete case method* (drop observations).

Standard approaches:

1. **Complete case:** Drop missing obs. — loses efficiency
2. **Dummy variable:** Fill in 0, add indicator — can be inconsistent
3. **Linear imputation:** Predict missing values — requires extra assumptions

GMM offers a better way: Exploit moment conditions from *both* complete and incomplete observations simultaneously.

The Missing Data Model (Abrevaya & Donald, 2017)

Structural regression:

$$y_i = \alpha_0 x_i + z_i'\beta_0 + \varepsilon_i \qquad \text{where } E(x_i\varepsilon_i) = 0,\ E(z_i\varepsilon_i) = 0$$

x_i is a (possibly missing) scalar regressor; z_i is always observed.

Linear projection of x_i on z_i: $x_i = z_i'\gamma_0 + \xi_i$ where $E(z_i\xi_i) = 0$. **Missingness indicator:** $m_i = 1$ if x_i missing, $m_i = 0$ if observed.

Key substitution: For observations with $m_i = 1$:

$$y_i = z_i'(\gamma_0\alpha_0 + \beta_0) + (\varepsilon_i + \xi_i\alpha_0) \equiv z_i'\delta_0 + \eta_i$$

Assumption 1: (a) $E(m_i z_i \varepsilon_i) = 0$; (b) $E(m_i z_i \xi_i) = 0$; (c) $E(m_i x_i \varepsilon_i) = 0$.
$\Rightarrow$ Missingness may depend on z_i but not on ε_i or ξ_i.

Three Blocks of Moment Conditions

The parameters are $\theta = (\alpha_0, \beta_0, \gamma_0)'$ with $k = 2K + 1$ unknowns.

Block 1 (observed cases, structural equation): $(1 - m_i) \cdot w_i(y_i - \alpha_0 x_i - z_i'\beta_0) = 0$ [$K + 1$ conditions]
Block 2 (observed cases, projection equation): $(1 - m_i) \cdot z_i(x_i - z_i'\gamma_0) = 0$ [K conditions]
Block 3 (missing cases, substituted equation): $m_i \cdot z_i(y_i - z_i'(\gamma_0\alpha_0 + \beta_0)) = 0$ [K conditions]

Total: $\ell = 3K + 1$ moments for $k = 2K + 1$ parameters $\Rightarrow$ K overidentifying restrictions (testable via J-test!)

With $K = 2$ (intercept + IQ): 7 moments, 5 parameters, **2 overidentifying restrictions**

Why GMM Beats the Alternatives

Method	Consistent?	Efficient?	Testable?
Complete case	Yes (drops $m_i = 1$)	No (fewer obs)	No
Dummy variable	Sometimes no	No	No
Linear imputation	Yes (under Assn 1)	Somewhat	No
GMM	**Yes** (under Assn 1)	**Yes** (optimal W)	**Yes** (J-test)

Abrevaya & Donald Monte Carlo ($n = 400$, 1000 replications):

Method	Bias (α_0)	$n \times$Var (α_0)	MSE (α_0)
Complete case	0.011	13.93	0.035
Dummy variable	-0.194	13.94	0.073
FGLS	0.011	13.93	0.035
GMM	**0.006**	**12.53**	**0.031**

WLS Data: Education and BMI

```
library(gmm); library(haven); library(estimatr)

# Wisconsin Longitudinal Study data
data1 <- read_dta("wls-data.dta")
data_men <- subset(data1, male == 1)

# How much data is missing?
table(data_men$bmimissing)
#   0    1
# 4231 1095    (~21% missing)

# Complete-case regression
data_complete <- subset(data_men, bmimissing == 0)
cc_reg <- lm(educ ~ bmirating + iq, data = data_complete)
```

Question: Education (y) regressed on BMI rating (x, sometimes missing) and IQ (z, always observed).

Defining GMM Moment Functions in R

```
g_men <- function(theta, x) {
  beta_0 <- theta[1]; alpha <- theta[2]; beta_iq <- theta[3]
  gamma_0 <- theta[4]; gamma_iq <- theta[5]
  educ <- x[,1]; bmirating <- x[,2]; iq <- x[,3]; bmimissing <- x[,4]

  # Block 1: structural eq (observed cases)
  r1 <- (1 - bmimissing) * (educ - beta_0 - alpha*bmirating - beta_iq*iq)
  # Block 2: projection eq (observed cases)
  r2 <- (1 - bmimissing) * (bmirating - gamma_0 - gamma_iq*iq)
  # Block 3: substituted eq (missing cases)
  r3 <- bmimissing * (educ - (gamma_0*alpha+beta_0) - (gamma_iq*alpha+beta_iq)*iq)

  cbind(r1, r1*bmirating, r1*iq,  # 3 moments
        r2, r2*iq,                # 2 moments
        r3, r3*iq)                # 2 moments = 7 total, 5 params
}
```

Running GMM in R

```
# Starting values from complete-case regressions
start_men <- c(beta_0 = coef(cc_reg)[1],
               alpha  = coef(cc_reg)[2],
               beta_iq = coef(cc_reg)[3],
               gamma_0 = coef(proj_reg)[1],
               gamma_iq = coef(proj_reg)[2])

# Two-step GMM with HAC variance
gmm_men <- gmm(g_men, x = x_men, t0 = start_men,
               type = "twoStep",
               wmatrix = "ident", vcov = "HAC")
summary(gmm_men)
```

Key arguments:

- `type = "twoStep"`: Initial $W = I$, then update with $\hat{\Omega}^{-1}$
- `wmatrix = "ident"`: First-step weight matrix is identity

Comparing Methods: Results

	Complete Case	Dummy Variable	GMM
$\hat{\alpha}$ (BMI)	-0.0343	-0.0343	-0.0340
SE	(0.0077)	(0.0073)	(0.0072)
$\hat{\beta}_{IQ}$	0.0563	0.0567	0.0564
SE	(0.0017)	(0.0015)	(0.0015)
n used	4231	5326	5326

J-test for overidentifying restrictions:

```
specTest(gmm_men)  # J-test = ?, df = 2, p-value = ?
```

- Large p-value: fail to reject $\Rightarrow$ Assumption 1 is plausible
- Small p-value: evidence against the linear projection restriction

Nonlinear GMM (Hansen §13.25)

So far: $g_i(\beta)$ is linear in β. In general, $g_i(\beta)$ can be **nonlinear**.

The GMM estimator still minimizes $J(\beta) = n\,\bar{g}_n(\beta)'\hat{W}\,\bar{g}_n(\beta)$ but now requires numerical optimization (no closed form).

Proposition 13.1: Distribution of Nonlinear GMM

$\sqrt{n}(\hat{\beta}_{\text{gmm}} - \beta) \xrightarrow{d} N(0, V_\beta)$ where $V_\beta = (Q'WQ)^{-1}(Q'W\Omega WQ)(Q'WQ)^{-1}$ with the **Jacobian** $Q = \mathbb{E}\left[\frac{\partial}{\partial \beta'} g_i(\beta)\right]$ replacing the linear $\mathbb{E}[ZX']$.

Example: The Abrevaya & Donald missing data model is nonlinear — Block 3 contains $\gamma_0\alpha_0$ (product of parameters).

The MTE Connection

Recall from Lecture 14: the **Marginal Treatment Effect** is

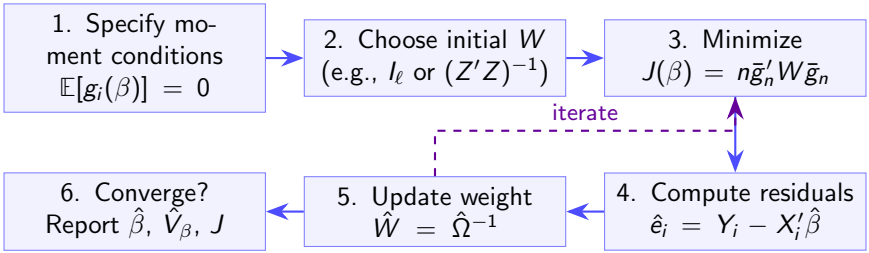
$$\Delta^{MTE}(x, u_D) = \mathbb{E}[Y_1 - Y_0 \mid X = x, U_D = u_D]$$

MTE estimation is nonlinear GMM:

- Moment conditions from the relationship between $E[Y \mid X, Z]$ and propensity score $P(Z)$
- MTE function modeled as a polynomial in u_D
- Parameters enter nonlinearly through $\int_0^{P(z)} \Delta^{MTE}(x, u) du$
- Overidentification arises from multiple instruments

The `ivmte` package (Shea & Torgovitsky) implements GMM estimation of MTE under the hood — specifying moment conditions, choosing weight matrices, and testing overidentifying restrictions.

Summary of GMM Estimation



- **One-step:** Stop after step 3 (e.g., 2SLS with $W = (Z'Z)^{-1}$)
- **Two-step:** One pass through steps 3–5
- **Iterated:** Repeat steps 3–5 until convergence

Key Theorems Summary

Theorem	Result
13.1	GMM estimator (linear): $\hat{\beta} = (X'ZWZ'X)^{-1}(X'ZWZ'Y)$
13.2	$W = (Z'Z)^{-1}$ gives 2SLS; just-identified gives IV
13.3	Asymptotic normality with sandwich variance
13.4	Efficient GMM: $W = \Omega^{-1}$, variance $(Q'\Omega^{-1}Q)^{-1}$
13.5	Efficient GMM has smallest variance among GMM estimators
13.6	2SLS is efficient GMM under homoskedasticity
13.7	Two-step GMM is asymptotically efficient

Running theme: These parallel the OLS $\rightarrow$ GLS progression. GMM is to IV what GLS is to OLS.

The Three Meta-Lessons (Part 1)

1. **Semiparametric**
   - GMM requires only $\mathbb{E}[g_i(\beta)] = 0$ — no distributional assumptions
   - The missing data application: no assumption on the distribution of missingness, errors, or regressors — only moment conditions
   - Chamberlain (1987): efficient GMM achieves the semiparametric bound
2. **Efficient**
   - Optimal weighting $W = \Omega^{-1}$ minimizes asymptotic variance
   - Missing data: GMM achieves lower MSE than complete case, dummy variable, or imputation
3. **General**
   - Nests OLS, IV, 2SLS, GLS as special cases
   - Extends to nonlinear models (missing data, MTE)
   - *Next lecture:* also provides a unified testing framework

Preview of Next Lecture

Lecture 16: GMM Inference and Going Beyond ATE

- **Testing:** The J-test for overidentification (Thm 13.14)
- **Hypothesis tests:** Wald test and Distance test in the GMM framework
- **Subset tests:** Testing specific instruments, endogeneity tests
- **Application: Propaganda and Foreign Media**
  - Kern & Hainmueller: West German TV in East Germany
  - IV gives LATE ≈ -0.12 (media reduces support for regime)
  - MTE reveals treatment effect heterogeneity: always-takers vs. never-takers have *opposite signs*
- **The course in one slide:** OLS $\rightarrow$ GLS $\rightarrow$ IV $\rightarrow$ 2SLS $\rightarrow$ GMM