# Linear Models Lecture 16: GMM Inference and Going Beyond ATE

Robert Gulotty

University of Chicago

February 26, 2026

# Recap: The GMM Framework

From Lecture 15:

**The GMM estimator** minimizes the weighted quadratic form:

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

**Key results:**

- $W = (Z'Z)^{-1}$ gives 2SLS (Thm 13.2)
- Sandwich variance: $V_\beta = (Q'WQ)^{-1}(Q'W\Omega WQ)(Q'WQ)^{-1}$ (Thm 13.3)
- Efficient GMM: $W = \Omega^{-1}$, $V_\beta = (Q'\Omega^{-1}Q)^{-1}$ (Thm 13.4)
- 2SLS efficient only under homoskedasticity (Thm 13.6)
- Two-step and iterated GMM are asymptotically efficient (Thm 13.7)

**Today:** What can we *test* with GMM? And how does GMM let us go *beyond* average treatment effects?

# The Overidentification Test: Intuition

When $\ell > k$, the model imposes **testable restrictions**.

**Under correct specification:**

$$\mathbb{E}[g_i(\beta)] = 0 \quad \Rightarrow \quad \bar{g}_n(\hat{\beta}) \approx 0$$

Even at the minimizer, $\bar{g}_n(\hat{\beta}) \neq 0$ in finite samples. But it should be *close* to zero.

**Test logic:**

- **Small** $J(\hat{\beta})$: Moment conditions are approximately satisfied $\Rightarrow$ model OK
- **Large** $J(\hat{\beta})$: Cannot simultaneously satisfy all moments $\Rightarrow$ misspecification

The criterion value $J = J(\hat{\beta}_{\text{gmm}})$ is a natural test statistic for:

$$H_0 : \mathbb{E}[Ze] = 0 \qquad \text{vs.} \qquad H_1 : \mathbb{E}[Ze] \neq 0$$

# Hansen's J-Test (Hansen Thm. 13.14)

## Theorem (13.14: Overidentification Test)

*Under* $H_0 : \mathbb{E}[Ze] = 0$ *and using an efficient weight matrix estimator,*

$$J = J(\hat{\beta}_{gmm}) = n\,\bar{g}_n(\hat{\beta})'\hat{\Omega}^{-1}\bar{g}_n(\hat{\beta}) \xrightarrow{d} \chi^2_{\ell-k}$$

*Reject* $H_0$ *if* $J > \chi^2_{\ell-k,1-\alpha}$.

**Degrees of freedom** $= \ell - k =$ number of overidentifying restrictions.

**Requirements:**

- Must use an *efficient* weight matrix ($\hat{W} = \hat{\Omega}^{-1}$)
- Generalizes the Sargan test (which assumed homoskedasticity)
- When $\ell = k$ (just identified): $J = 0$ always — no test possible

# J-Test: Strengths and Limitations

**Strengths:**

- Automatic byproduct of efficient GMM estimation
- General: works under heteroskedasticity, no distributional assumptions
- Natural diagnostic: "always report the J statistic" (Hansen, Ch. 13)

**Limitations:**

- No power if all instruments are invalid in the same direction — moments are all "wrong" together, J-test cannot detect this
- Requires $\ell > k$ (overidentification)
- Requires efficient weight matrix for $\chi^2$ distribution
- Rejection tells you *something* is wrong, but not *what*

Good practice: Use subset overidentification tests (Thm 13.15) to investigate *which* instruments may be invalid.

# J-Test for Missing Data

From the Abrevaya & Donald application (Lecture 15):

```
# Run the specification test
specTest(gmm_men)

# Hansen's J-test
# Test E(g) = 0:
# Statistics  df  p-value
# J-test       ?   2    ?
```

**Interpretation** ($\ell - k = 7 - 5 = 2$ degrees of freedom):

- The test assesses whether the linear projection restriction holds for both missing and complete data subpopulations
- **Fail to reject:** The restriction that $x = z'\gamma + \xi$ is the same for observed and missing cases is supported
- **Reject:** Missingness may depend on unobservables, violating Assumption 1

# The Wald Test (Hansen Thm. 13.8)

Test $H_0 : \theta = \theta_0$ where $\theta = r(\beta)$ for a known function $r : \mathbb{R}^k \to \mathbb{R}^q$.

## Theorem (13.8: Wald Test)

*Under $H_0$, as $n \to \infty$:*

$$W = n(\hat{\theta} - \theta_0)' \hat{V}_{\theta}^{-1} (\hat{\theta} - \theta_0) \xrightarrow{d} \chi^2_q$$

*where $\hat{V}_{\theta} = \hat{R}' \hat{V}_{\beta} \hat{R}$ and $\hat{R} = \frac{\partial}{\partial \beta} r(\hat{\beta}_{gmm})'$.*

**Special case:** Testing $\beta_j = 0$: $W = \hat{\beta}_j^2 / \widehat{\text{Var}}(\hat{\beta}_j) = t_j^2 \xrightarrow{d} \chi^2_1$

**Familiar:** This is exactly the $t$-test (squared) we have been using throughout the course, now justified by GMM asymptotics.

# The Distance Test (Hansen Thm. 13.12)

An alternative to the Wald test, based on comparing criterion functions.

**Idea:** Estimate unrestricted ($\hat{J}$) and restricted ($\tilde{J}$, subject to $r(\beta) = \theta_0$) models by efficient GMM.

**Theorem (13.12: Distance Test)**

*Under $H_0$ and using efficient GMM, as $n \to \infty$:*

$$D = \tilde{J} - \hat{J} \xrightarrow{d} \chi^2_q$$

**Key advantages over Wald:**

- Invariant to reparameterization—the Wald statistic is not
- Analogous to the likelihood ratio test (criterion-based)
- Thm 13.13: If $\tilde{\Omega} = \hat{\Omega}$, then $D \geq 0$; for linear $r$, $D = W$

## Three Tests Compared

| | Wald Test | Distance Test | J-Test |
|---|---|---|---|
| **Tests** | Parameter restrictions $r(\beta) = \theta_0$ | Parameter restrictions $r(\beta) = \theta_0$ | Model specification $\mathbb{E}[g_i(\beta)] = 0$ |
| **Requires** | Unrestricted estimate only | Both restricted and unrestricted | Efficient weight matrix |
| **Distribution** | $\chi^2_q$ | $\chi^2_q$ | $\chi^2_{\ell-k}$ |
| **Invariant?** | No | Yes | N/A |
| **Analogy** | $t$-test / $F$-test | Likelihood ratio | Sargan test |

**Recommendation:** Distance test for nonlinear hypotheses (invariant). Wald test for linear hypotheses (simpler, equivalent). Always report the J-test.

# Subset Overidentification Tests (Hansen Thm. 13.15)

**Question:** Are *specific* instruments valid?

Partition $Z = (Z_a, Z_b)$ where:

- $Z_a$ ($\ell_a$ instruments): believed to be valid ($\mathbb{E}[Z_a e] = 0$)
- $Z_b$ ($\ell_b$ instruments): questionable ($\mathbb{E}[Z_b e] \stackrel{?}{=} 0$)

Require $\ell_a > k$ so $Z_a$ alone identifies the model.

## Theorem (13.15: Subset Overidentification Test)

*Let $\tilde{J}$ use only $Z_a$ and $\hat{J}$ use all $Z = (Z_a, Z_b)$. Then:*

$$C = \hat{J} - \tilde{J} \xrightarrow{d} \chi^2_{\ell_b}$$

*under $H_0 : \mathbb{E}[Z_b e] = 0$.*

**Special case:** If $Z_a$ is just-identified ($\ell_a = k$), then $\tilde{J} = 0$ and $C = \hat{J}$ is the standard J-test.

# Endogeneity Test via GMM (Hansen Thm. 13.16)

**Question:** Is $Y_2$ endogenous, i.e., is $\mathbb{E}[Y_2 e] \neq 0$?

Model: $Y = Z_1'\beta_1 + Y_2'\beta_2 + e$, instruments $(Z_1, Z_2)$ with $\mathbb{E}[Z_1 e] = \mathbb{E}[Z_2 e] = 0$.

**Key insight:** If $Y_2$ is exogenous ($H_0$), then $Y_2$ is a valid instrument for itself. So we can expand the instrument set from $(Z_1, Z_2)$ to $(Z_1, Z_2, Y_2)$.

## Theorem (13.16: Endogeneity Test)

1. *Estimate efficient GMM with instruments $(Z_1, Z_2)$. Let $\bar{J}$ be the criterion.*
2. *Estimate efficient GMM with instruments $(Z_1, Z_2, Y_2)$. Let $\hat{J}$ be the criterion.*

*Under* $H_0 : \mathbb{E}[Y_2 e] = 0, \quad C = \hat{J} - \bar{J} \xrightarrow{d} \chi^2_{k_2}.$

This is a **subset overidentification test** where the "questionable instruments" are $Y_2$ itself. It **generalizes** Durbin-Wu-Hausman: DWH requires homoskedasticity; this GMM version does not.

# GMM Subsumes All Prior Tests

| Earlier Course Test | GMM Version | Theorem |
|---|---|---|
| $t$-test for $\beta_j = 0$ | Wald test ($q = 1$) | 13.8 |
| $F$-test for $R\beta = c$ | Wald test ($q > 1$) | 13.8 |
| Hausman (OLS vs IV) | Endogeneity test | 13.16 |
| Durbin-Wu-Hausman | Endogeneity test | 13.16 |
| Sargan overid test | J-test | 13.14 |
| White's heteroskedasticity test | Special case of J-test | — |
| Likelihood ratio test | Distance test | 13.12 |

Meta-lesson #3 (General): GMM provides a *unified* framework for both estimation and inference. Every test we have seen is a special case of a GMM test—often under weaker assumptions.

# When ATE Is Not Enough

Recall from Lecture 14:

- **LATE** = effect for *compliers* only — those induced to change treatment by the instrument
- Different instruments ⇒ different complier groups ⇒ different LATEs

**Key questions that LATE cannot answer:**

1. What is the effect for *always-takers*? For *never-takers*?
2. Does the treatment effect *change sign* across subpopulations?
3. What would happen under a *different policy* (different compliance margin)?

The **Marginal Treatment Effect** $\Delta^{MTE}(u_D)$ traces out treatment effects across the *entire* resistance-to-treatment distribution. Estimating MTE requires GMM.

# West German TV in East Germany

**Kern & Hainmueller (2009):** Effect of Western media on support for the East German regime.

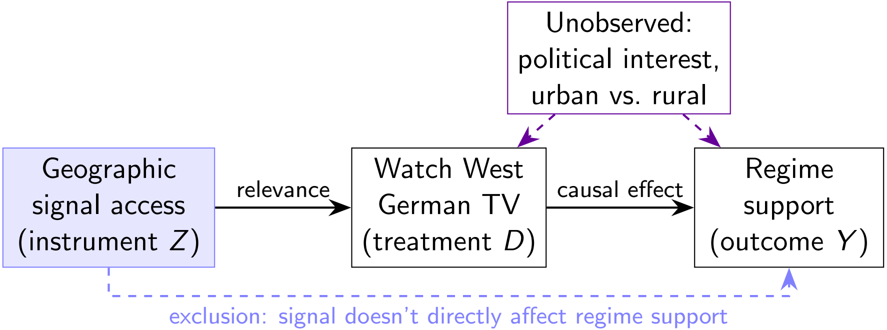
**Setting:**

- Cold War East Germany (GDR), pre-1989
- West German TV broadcast entertainment, news, and consumer culture
- The regime feared Western media as subversive
- Most East Germans could receive West German TV signals

**Research question:** Did exposure to West German TV *undermine* support for the Communist regime?

**Outcome:** Support for the regime (survey data)
**Treatment:** Regular viewing of West German TV
**Challenge:** Viewers self-select into watching — endogenous

# The Instrument: Geographic Signal Access




**Key:** Due to topography and transmitter locations, some areas of East Germany could not receive West German TV — notably the **Dresden** region ("Valley of the Clueless").

Geographic signal access is plausibly exogenous: determined by physics, not politics.

# Conventional IV Results

**Standard IV/2SLS estimate:**

$$\widehat{LATE} \approx -0.12$$

**Interpretation:** Among *compliers* (those induced to watch by having signal access), watching West German TV reduces regime support by 0.12 standard deviations.

**But who are the compliers?**

- People who watch West German TV *only when they have signal access*
- Not the most politically interested (they would find other sources — always-takers)
- Not the most regime-loyal (they would not watch even with access — never-takers)
- Compliers are a potentially narrow and unrepresentative group

**Question:** What about the effects for always-takers and never-takers?

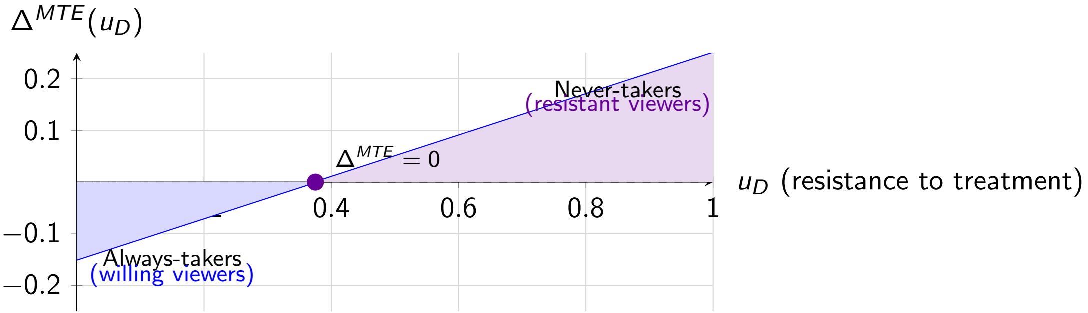
# MTE Results: Treatment Effect Heterogeneity

**Using the MTE framework** (extending Kern & Hainmueller with MTE methods):

| **Group** | **Treatment Effect** | **Interpretation** |
|---|---|---|
| Always-takers | $-0.104$ | Western TV *reduces* support (politically interested viewers) |
| Compliers (LATE) | $\approx -0.12$ | Moderate effect |
| Never-takers | $+0.189$ | Western TV *increases* support! (contrast effect / reactance) |

Sign reversal: The treatment effect is *negative* for willing viewers but *positive* for resistant viewers! Western consumer culture made regime-loyal individuals *more* supportive (reactance), while curious viewers became *less* supportive (information effect).

# Visualizing the MTE Curve




**Low $u_D$:** Always-takers (willing viewers) — effect is *negative*. **High $u_D$:** Never-takers (resistant viewers) — effect is *positive*. The **LATE** is a weighted average over compliers, masking the heterogeneity.

# Why MTE Requires GMM

**MTE estimation involves:**

1. **Nonlinear moment conditions:** $E[Y|X, Z] = X'\beta + \int_0^{P(Z)} \Delta^{MTE}(u)\,du$. Parameters enter through the integral — nonlinear in $\beta$.
2. **Overidentification:** Multiple instruments provide more moment conditions than parameters $\Rightarrow$ testable.
3. **Efficient weighting:** Different propensity score regions are estimated with different precision $\Rightarrow$ optimal weighting matters.
4. **J-test:** Tests whether the MTE specification (e.g., polynomial degree) is adequate.

Without GMM, we could not estimate or test the MTE.

# Policy Implications

The **Policy-Relevant Treatment Effect** (PRTE) depends on which part of the MTE curve the policy targets:

$$PRTE = \int_0^1 \Delta^{MTE}(u_D) \cdot \omega^{PRTE}(u_D)\, du_D$$

where $\omega^{PRTE}$ depends on how the policy shifts the propensity score.

**For the propaganda example:**

- Forcing everyone to watch (shifting never-takers) could backfire: MTE is positive for high $u_D$
- Facilitating access for willing viewers (low $u_D$) would reduce regime support
- The PRTE sign/magnitude depends on the policy's compliance margin

Lesson: Treatment effect heterogeneity means policy evaluation requires more than a single number. The MTE curve—estimated via GMM—provides the necessary detail.

# Connection to ivmte

Recall from Lecture 14: the `ivmte` package (Shea & Torgovitsky) estimates MTE:

```
library(ivmte)
result <- ivmte(
  data = df,
  outcome = "regime_support",
  treatment = "watch_western_tv",
  instrument = "signal_access",
  target = "ate",         # or "att", "prte"
  m0 = ~ u + I(u^2),    # MTE polynomial for control
  m1 = ~ u + I(u^2),    # MTE polynomial for treated
  propensity = D ~ signal_access
)
```

**Under the hood,** `ivmte` **performs:** nonlinear GMM estimation of MTE parameters, efficient weighting across moment conditions, and overidentification testing (J-test).

# Restricted GMM (Hansen §13.15–13.16)

**Linear constraints:** $R'\beta = c$

$\hat{\beta}_{\text{cgmm}} = \hat{\beta}_{\text{gmm}} - (X'ZWZ'X)^{-1}R\left(R'(X'ZWZ'X)^{-1}R\right)^{-1}(R'\hat{\beta}_{\text{gmm}} - c)$

This is the GMM analog of restricted OLS / minimum distance.

**Nonlinear constraints:** $r(\beta) = 0$ — minimize $J(\beta)$ subject to constraint (numerical optimization).

## Theorem (13.10–13.11)

*The constrained efficient GMM estimator has asymptotic variance $V_{cgmm} = V_\beta - V_\beta R(R'V_\beta R)^{-1}R'V_\beta$. Constrained estimation is (weakly) more efficient than unconstrained — if the constraints are true.*

# Bootstrap for GMM (Hansen §13.26)

Standard bootstrap for GMM: resample $(Y_i^*, X_i^*, Z_i^*)$ and re-estimate.

Problem: When overidentified, the bootstrap estimator does not satisfy the orthogonality condition $\Rightarrow$ no asymptotic refinement.

**Solution: Recentered bootstrap** (Hall & Horowitz, 1996)

$$\hat{\beta}^{**}_{\text{gmm}} = (X^{*\prime} Z^* W^* Z^{*\prime} X^*)^{-1} (X^{*\prime} Z^* W^* (Z^{*\prime} Y^* - Z' \hat{e}))$$

where $\hat{e} = Y - X\hat{\beta}_{\text{gmm}}$ from the *original* sample. Subtracts original residuals' moment contribution to recenter around zero.

**Practical advice:** Use bootstrap for confidence intervals (not SEs). Percentile-$t$ bootstrap provides best finite-sample coverage.

# The Full GMM Testing Toolkit

| Test | Distribution | Purpose |
|---|---|---|
| J-test (Thm 13.14) | $\chi^2_{\ell-k}$ | Model specification |
| Wald (Thm 13.8) | $\chi^2_q$ | Parameter restrictions |
| Distance (Thm 13.12) | $\chi^2_q$ | Parameter restrictions (invariant) |
| Subset overid (Thm 13.15) | $\chi^2_{\ell_b}$ | Validity of specific instruments |
| Endogeneity (Thm 13.16) | $\chi^2_{k_2}$ | Exogeneity of regressors |
| Constrained (Thm 13.10) | — | Efficient estimation under $H_0$ |
| Bootstrap | — | Finite-sample inference |

**All tests are:** valid under heteroskedasticity, derived from the GMM criterion function, and applicable to both linear and nonlinear models.

# Semiparametric Efficiency Bound

**Chamberlain (1987):** If all that is known is $\mathbb{E}[g_i(\beta)] = 0$, this is a **semiparametric** problem (the distribution of the data is unknown).

Chamberlain showed that no semiparametric estimator can have asymptotic variance smaller than $(G'\Omega^{-1}G)^{-1}$ where $G = \mathbb{E}\left[\frac{\partial}{\partial\beta'} g_i(\beta)\right]$.

**Efficient GMM achieves this bound:** $V_\beta = (Q'\Omega^{-1}Q)^{-1} = (G'\Omega^{-1}G)^{-1}$

**This means:**

- No other estimator using only these moment conditions can do better
- MLE *can* do better — but requires specifying the full distribution
- GMM is the best you can do without distributional assumptions

# When to Use GMM vs. Other Methods

| | When to use | Advantages | Cost |
|---|---|---|---|
| **OLS/GLS** | $\mathbb{E}[Xe] = 0$ (no endogeneity) | Simple, efficient under assumptions | Biased if endogenous |
| **2SLS** | Endogeneity, homoskedastic | Simple, familiar | Inefficient under heterosked. |
| **GMM** | Endogeneity, heteroskedastic | Efficient, flexible, testable | More complex |
| **MLE** | Full distribution known | Most efficient (Cramér–Rao) | Misspecification bias |

**Decision rule:**

- No endogeneity? $\Rightarrow$ OLS/GLS
- Endogeneity + homoskedasticity + few instruments? $\Rightarrow$ 2SLS
- Endogeneity + heteroskedasticity or many instruments? $\Rightarrow$ **GMM**
- Full distributional knowledge? $\Rightarrow$ MLE

# The Three Meta-Lessons Revisited

1. **Semiparametric**
   - Requires only moment conditions — achieves Chamberlain bound
   - Missing data: no distributional assumptions on missingness
   - MTE: no parametric model for selection — only moments from the propensity score
2. **Efficient**
   - Optimal weighting $W = \Omega^{-1}$ minimizes variance
   - Missing data: GMM has lowest MSE across methods
   - Propaganda: efficient weighting across propensity score values
3. **General**
   - Unified estimation: OLS, GLS, IV, 2SLS all as special cases
   - Unified testing: $t$, $F$, Hausman, Sargan all as special cases
   - Extends to nonlinear models, treatment effect heterogeneity, policy evaluation

# The Course in One Slide

OLS $E[Xe] = 0$ $\xrightarrow{\text{heterosked.}}$ GLS $E[\bar{X}e] = 0$ $\xrightarrow{\text{endogeneity}}$ IV $E[Ze] = 0$, $\ell = k$ $\xrightarrow{\text{overid.}}$ 2SLS $E[Ze] = 0$, $\ell > k$ $\xrightarrow{\text{efficiency}}$ **GMM** $E[g_i(\beta)] = 0$, **optimal** $W$

All are moment condition estimators: $\mathbb{E}[g_i(\beta)] = 0$

**The progression:** Each step addresses a *new problem* (heteroskedasticity, endogeneity, overidentification, efficiency). Each estimator is a *special case* of the next. GMM is the *most general*: it nests everything and achieves the semiparametric bound.

# Looking Ahead: Panel Data

**Next week: Panel data and fixed effects**

Panel data (repeated observations on the same units) introduces new moment conditions:

**Fixed effects as moments:**

$$\mathbb{E}[\tilde{X}_i \tilde{e}_i] = 0 \qquad \text{(within-group orthogonality)}$$

where $\tilde{X}_i = X_{it} - \bar{X}_i$ is the demeaned regressor.

**Arellano-Bond (1991) GMM:**

- Dynamic panels: $Y_{it} = \rho Y_{it-1} + X_{it}'\beta + \alpha_i + e_{it}$
- Lagged levels as instruments for first-differenced equation
- Growing set of moment conditions: $\mathbb{E}[\Delta e_{it} \cdot Y_{is}] = 0$ for $s \leq t - 2$
- Naturally overidentified $\Rightarrow$ GMM with efficient weighting + J-test

GMM is the workhorse estimator for dynamic panel data.

# Summary

**Five takeaways from Lectures 15–16:**

1. **GMM is a unified estimation framework** that nests OLS, GLS, IV, and 2SLS as special cases, requiring only moment conditions $\mathbb{E}[g_i(\beta)] = 0$.
2. **Efficient GMM** uses $W = \Omega^{-1}$ to achieve the semiparametric efficiency bound. Two-step and iterated GMM achieve this in practice.
3. **The J-test** provides a natural diagnostic for overidentified models. Always report it.
4. **GMM subsumes all prior tests:** Wald, Distance, subset overidentification, and endogeneity tests are all GMM tests—valid under heteroskedasticity.
5. **Going beyond ATE:** GMM enables estimation of the MTE curve, revealing treatment effect heterogeneity that LATE conceals.