EC708 Discussion 3
Hausman Pretest and LATE

Yan Liu

Department of Economics
Boston University

February 3, 2023

Outline

Table of Contents

Proving Positive Semidefiniteness

Best Linear Combination of Instruments

Consider the linear IV model:

$$Y_t = X_t'\beta + U_t,$$
$$X_t' = Z_t'\pi + V_t,$$

where X_t is $k \times 1$ and Z_t is $\ell \times 1$ with $\ell > k$. Assume $E[U_t^2|Z_t] = \sigma^2$.

- Create a new instrument $\tilde{Z}_t = B'Z_t$ for some ℓ-by-k matrix B. Let $\tilde{\beta} = (\tilde{\mathbf{Z}}'\mathbf{X})^{-1}\tilde{\mathbf{Z}}Y$. The asymptotic variance of $\tilde{\beta}$ is

$$V_{\tilde{\beta}} = \sigma^2(b'Q_{ZX})^{-1}b'Q_{ZZ}b(Q_{ZX}'b)^{-1}, \text{ where } b = \operatorname{plim}_{T\to\infty} B.$$

- Setting $B = (\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}$ gives the 2SLS estimator $\hat{\beta}_{2SLS}$, whose asymptotic variance is

$$V_{\hat{\beta}_{2SLS}} = \sigma^2(Q_{ZX}'Q_{ZZ}^{-1}Q_{ZX})^{-1}.$$

Proving Positive Semidefiniteness

Best Linear Combination of Instruments

We want to show that $V_{\tilde{\beta}} - V_{\hat{\beta}_{2SLS}}$ is positive semi-definite. Exercise 13.4 of Hansen (2022) provides a proof strategy:

1. Find matrices A and A^* such that $V_{\tilde{\beta}} = \sigma^2 A' Q_{ZZ} A$ and $V_{\hat{\beta}_{2SLS}} = \sigma^2 A^{*\prime} Q_{ZZ} A^*$.
2. Show that $A^{*\prime} Q_{ZZ} A = A^{*\prime} Q_{ZZ} A^*$ and therefore that $A^{*\prime} Q_{ZZ}(A - A^*) = 0$.
3. Use the expressions $V_{\tilde{\beta}} = \sigma^2 A' Q_{ZZ} A$, $A = A^* + (A - A^*)$, and $A^{*\prime} Q_{ZZ}(A - A^*) = 0$ to show that $V_{\tilde{\beta}} \geq V_{\hat{\beta}_{2SLS}}$.

Intuitively, we are showing $V_{\tilde{\beta}} - V_{\hat{\beta}_{2SLS}} = \text{Asy.}\,\text{Var}(\tilde{\beta} - \hat{\beta}_{2SLS})$.

Proving Positive Semidefiniteness

Best Linear Combination of Instruments

Step 1:

$$V_{\tilde{\beta}} = \sigma^2 (b' Q_{ZX})^{-1} b' Q_{ZZ} \underbrace{b (Q'_{ZX} b)^{-1}}_{=:A},$$

$$\begin{aligned} V_{\hat{\beta}_{2SLS}} &= \sigma^2 (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} \\ &= \sigma^2 (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX} (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} \\ &= \sigma^2 (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} Q'_{ZX} Q_{ZZ}^{-1} Q_{ZZ} \underbrace{Q_{ZZ}^{-1} Q_{ZX} (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1}}_{=:A^*} \end{aligned}$$

Proving Positive Semidefiniteness

Best Linear Combination of Instruments

Step 2:

$$
\begin{aligned}
A^{*\prime} Q_{ZZ} A &= [(Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} Q'_{ZX} Q_{ZZ}^{-1}] Q_{ZZ} [b(Q'_{ZX} b)^{-1}] \\
&= (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} Q'_{ZX} b (Q'_{ZX} b)^{-1} \\
&= (Q'_{ZX} Q_{ZZ}^{-1} Q_{ZX})^{-1} \\
&= A^{*\prime} Q_{ZZ} A^{*}.
\end{aligned}
$$

Therefore, $A^{*\prime} Q_{ZZ} (A - A^{*}) = 0$.

Proving Positive Semidefiniteness

Best Linear Combination of Instruments

Step 3:

$$
\begin{aligned}
V_{\tilde{\beta}} &= \sigma^2 A' Q_{ZZ} A \\
&= \sigma^2 [A^* + (A - A^*)]' Q_{ZZ} [A^* + (A - A^*)] \\
&= \sigma^2 A^{*\prime} Q_{ZZ} A^* + 2 \underbrace{A^{*\prime} Q_{ZZ} (A - A^*)}_{=0 \text{ by Step 2}} + (A - A^*)' Q_{ZZ} (A - A^*) \\
&= V_{\hat{\beta}_{2SLS}} + \underbrace{(A - A^*)' Q_{ZZ} (A - A^*)}_{\text{Asy.Var}(\tilde{\beta} - \hat{\beta}_{2SLS})}.
\end{aligned}
$$

Table of Contents

Consider two estimators $\hat{\beta}_{\mathrm{I}}$ and $\hat{\beta}_{\mathrm{II}}$ and a general hypothesis H_0 against alternative H_1. Suppose

- $\hat{\beta}_{\mathrm{I}}$: consistent & efficient (e.g. OLS) under H_0;
- $\hat{\beta}_{\mathrm{II}}$: consistent (e.g. IV) under H_0 and H_1.

Hausman-Wu test statistic:

$$H_T = T(\hat{\beta}_{\mathrm{II}} - \hat{\beta}_{\mathrm{I}})'[\hat{V}_{\mathrm{II}} - \hat{V}_{\mathrm{I}}]^{-}(\hat{\beta}_{\mathrm{II}} - \hat{\beta}_{\mathrm{I}})$$

- $\hat{V}_{\mathrm{I}}, \hat{V}_{\mathrm{II}}$: consistent estimators of the asymptotic variance of $\hat{\beta}_{\mathrm{I}}, \hat{\beta}_{\mathrm{II}}$
- A^{-}: generalized inverse of A

Under H_0, $H_T \overset{d}{\to} \chi^2_m$, where $m = \mathrm{rank}(\hat{V}_{\mathrm{II}} - \hat{V}_{\mathrm{I}})$.

Hausman Pretest

Two-Stage Test

Consider the linear IV model:

$$Y_t = X_t\beta + U_t,$$
$$X_t = Z_t'\pi + V_t,$$

where X_t is scalar and Z_t is $k \times 1$. We want to test

$$H_0 : \beta = \beta_0 \text{ v.s. } H_1 : \beta \neq \beta_0.$$

Under conditional homoskedasticity,

- If X_t is exogenous, $\hat{\beta}_{OLS}$ and $\hat{\beta}_{2SLS}$ are both consistent while $\hat{\beta}_{OLS}$ is efficient;
- If X_t is endogenous, only $\hat{\beta}_{2SLS}$ is consistent.

Hausman Pretest

Two-Stage Test

A natural thought is conduct a two-stage test:

1. Test endogeneity of X_t using the Hausman-Wu test:

$$H_0^{\text{Hausman}} : X_t \text{ is exogenous}, \quad H_T = T\frac{(\hat{\beta}_{2SLS} - \hat{\beta}_{OLS})^2}{\hat{V}_{2SLS} - \hat{V}_{OLS}}.$$

2. Run a 2SLS-based t-test if H_0^{Hausman} is rejected at level α^{H}: $t_{2SLS}(\beta_0) = \sqrt{T}(\hat{\beta}_{2SLS} - \beta_0)/\sqrt{\hat{V}_{2SLS}}$.
   Run an OLS-based t-test if H_0^{Hausman} is not rejected at level α^{H}: $t_{OLS}(\beta_0) = \sqrt{T}(\hat{\beta}_{OLS} - \beta_0)/\sqrt{\hat{V}_{OLS}}$.

The two-stage test statistic is

$$t_T(\beta_0) = t_{OLS}(\beta_0)1(H_T \leq \chi^2_{1,1-\alpha^{\text{H}}}) + t_{2SLS}(\beta_0)1(H_T > \chi^2_{1,1-\alpha^{\text{H}}}).$$

This test is problematic: severe size distortion in the second stage.

Guggenberger (2010) classifies nuisance parameters γ into three categories:

1. $\gamma_1 = \rho = \mathrm{Corr}(U_t, V_t)$: degree of endogeneity
2. $\gamma_2 = \|Q_{ZZ}^{1/2}\pi/\sigma_v\|$: strength of instruments
3. γ_3: everything else

He considers a sequence $\gamma_T = (\gamma_{1T}, \gamma_{2T}, \gamma_{3T})$ such that

$$\sqrt{T}\gamma_{1T} \to h_1, \quad \gamma_{2T} \to h_2 > \kappa > 0.$$

When X_t is weakly endogenous ($|h_1| < \infty$), he shows that

$$t_T(\beta_0) \overset{\gamma_T}{\rightsquigarrow} \underbrace{\eta^{\mathsf{OLS}}_{h_1,h_2} 1(\eta^{H}_{h_1,h_2} \leq \chi^2_{1,1-\alpha^{\mathsf{H}}}) + \eta^{\mathsf{2SLS}}(\beta_0) 1(\eta^{\mathsf{H}}_{h_1,h_2} > \chi^2_{1,1-\alpha^{\mathsf{H}}})}_{=:\eta_{h_1,h_2}}.$$

Hausman Pretest

Two-Stage Test: Size Distortion

In particular,

- $\eta^{\mathsf{OLS}}_{h_1,h_2} \sim N(h_1(1+h_2^2)^{-1/2}, 1)$
- $\eta^{\mathsf{2SLS}} \sim N(0,1)$
- $\eta^{\mathsf{H}}_{h_1,h_2} \sim \chi^2_1(h_1^2 h_2^2 (h_2^2+1)^{-1})$ (noncentral chi-square distribution)

As a result,

- Hausman pretest does not have enough power:
  $P(\eta^{\mathsf{H}}_{h_1,h_2} > \chi^2_{1,1-\alpha^{\mathsf{H}}}) < 1$
- Whenever OLS-based t-test is used, the maximal asymptotic probability of rejecting $H_0 : \beta = \beta_0$ equals 1 for h_1 large enough

Hausman Pretest

Partial Solution

- Ideally, we want to use the correct critical value $c_{h_1,h_2}(1-\alpha)$, which is the $1-\alpha$ quantile of the distribution of η_{h_1,h_2}
- However, it is impossible to consistently estimate h_1
- Andrews and Guggenberger (2009) propose to use the plug-in size-corrected fixed critical value:

$$cv_T^* = \sup_{h_1} c_{h_1,\hat{\gamma}_{2T}}(1-\alpha).$$

This ensures uniform size control but may prescribe an overly pessimistic critical value.

Table of Contents

2SLS and LATE

Without Covariates

In class, we learned that under

- **IV independence:** $(Y_t(1), Y_t(0), D_t(1), D_t(0)) \perp Z_t$,
- **LATE monotonicity:** $D_t(1) \geq D_t(0)$,

the LATE is identified as

$$E[Y_t(1) - Y_t(0)|T_t = co] = \frac{\text{ITT}_Y}{\text{ITT}_D}.$$

Further, we can show that

$$\frac{\text{ITT}_Y}{\text{ITT}_D} = \frac{\text{Cov}(Y_t, Z_t)}{\text{Cov}(D_t, Z_t)} = \text{plim}_{T\to\infty}\, \hat{\beta}_{2SLS}. \tag{1}$$

Hence, the linear IV estimator is consistent for the LATE.

Proof of (1): By IV independence,

$$
\begin{aligned}
\mathrm{ITT}_Y &= E[Y_t(D_t(1)) - Y_t(D_t(0))] \\
&= E[Y_t(D_t(1))|Z_t = 1] - E[Y_t(D_t(0))|Z_t = 0] \\
&= E[Y_t|Z_t = 1] - E[Y_t|Z_t = 0] \\
&= \frac{E[Y_t Z_t]}{P(Z_t = 1)} - \frac{E[Y_t(1 - Z_t)]}{P(Z_t = 0)} \\
&= \frac{E[Y_t Z_t] - E[Y_t]E[Z_t]}{P(Z_t = 1)P(Z_t = 0)}.
\end{aligned}
$$

Similarly, $\mathrm{ITT}_D = \frac{E[D_t Z_t] - E[D_t]E[Z_t]}{P(Z_t=1)P(Z_t=0)}$.

2SLS and LATE

With Covariates

Two reasons for incorporating covariates:

- IV independence may only be valid after conditioning on covariates
  - There are year-of-birth differences in both draft eligibility and earnings
  - College proximity is an active choice by parents, which might be related to characteristics that affect children's subsequence wages
  - Parental education is correlated with parents' profession, family income & wealth, which may directly affect their offspring's wage prospects
- Conditioning on covariates may lead to more precise 2SLS estimates (under constant conditional treatment effects)

2SLS and LATE

With Covariates

Conditional IV independence:

$$(Y_t(1), Y_t(0), D_t(1), D_t(0)) \perp Z_t | X_t$$

For each value of X_t, define the **covariate-specific LATE** as

$$\Delta(X_t) = E[Y_t(1) - Y_t(0) | D_t(1) > D_t(0), X_t].$$

Consider a linear IV regression based on

$$Y_t = D_t \beta + X_t' \gamma + U_t. \tag{2}$$

Is the 2SLS estimand a weighted average of covariate-specific LATE?

2SLS and LATE

With Covariates: Saturate and Weight

When covariates are discrete, Angrist and Imbens (1995) consider a fully saturated first stage

$$D_t = \pi_X + Z_t \pi_{1X} + V_t$$

and a saturated model for covariates in the second stage

$$Y_t = \alpha_X + D_t \beta + U_t.$$

Then, $\hat{\beta}_{2SLS} \xrightarrow{p} E[\omega_s(X_t)\Delta(X_t)]$, where

$$\omega_s(X_t) = \frac{P(T_t = co|X_t)^2 \cdot \mathrm{Var}[Z_t|X_t]}{E[P(T_t = co|X_t)^2 \cdot \mathrm{Var}[Z_t|X_t]]}.$$

In practice, we may not want to work with a model with a first-stage parameter for each value of the covariates ...

2SLS and LATE

With Covariates: Nonsaturated Specifications

Using the nonsaturated specification (2),

$$\hat{\beta}_{2SLS} \xrightarrow{p} \frac{E[Y_t \tilde{Z}_t]}{E[D_t \tilde{Z}_t]}, \quad \text{where } \tilde{Z}_t = Z_t - L[Z_t | X_t]$$

is the residual from linearly projecting Z_t onto X_t:

$$L[Z_t | X_t] = X_t' E[X_t X_t']^{-1} E[X_t Z_t].$$

Let us take a closer look at the numerator $E[Y_t \tilde{Z}_t]$.

2SLS and LATE

With Covariates: Nonsaturated Specifications

$$
\begin{aligned}
E[Y_t \tilde{Z}_t] &= E[E[Y_t \tilde{Z}_t | X_t]] \\
&= E[\mathrm{Cov}[Y_t, \tilde{Z}_t | X_t]] + E[E[Y_t | X_t] E[\tilde{Z}_t | X_t]] \\
&= E[\mathrm{Cov}[Y_t, Z_t | X_t]] + E[E[Y_t | X_t] E[\tilde{Z}_t | X_t]]
\end{aligned}
$$

- $E[\mathrm{Cov}[Y_t, Z_t | X_t]]$ only contains complier treatment effects
- $E[E[Y_t | X_t] E[\tilde{Z}_t | X_t]]$ contains all three groups
  - reflects the difference between nonparametric conditioning and linear projection (recall $E[\tilde{Z}_t | X_t] = E[Z_t | X_t] - L[Z_t | X_t]$)
  - depends on levels of always-taker and never-taker potential outcomes, rather than the difference $Y_t(1) - Y_t(0)$

The second term in the numerator $E[Y_t \tilde{Z}_t]$ goes away if we assume

- **Rich covariates:** $E[Z_t|X_t] = L[Z_t|X_t]$
  - automatically holds in saturated specifications and when $Z_t \perp X_t$
  - may hold when X_t included powers and cross-products of original covariates

Then, $\hat{\beta}_{2SLS} \xrightarrow{p} E[\omega_l(X_t)\Delta(X_t)]$, where

$$\omega_l(X_t) = \frac{P(T_t = co|X_t) \cdot \mathrm{Var}[Z_t|X_t]}{E[P(T_t = co|X_t) \cdot \mathrm{Var}[Z_t|X_t]]}.$$

See Blandhol et al. (2022) and Słoczyński (2020) for more details.