ECON 510 Final Exam Speedrun

A Self-Contained Textbook for B-

Patrik Guggenberger
Course Instructor
Penn State University
LaTeX compiled by Rui Zhou

Spring 2026
(April 27 – May 5, 9-Day Sprint Edition)

How to Use This Textbook

This textbook is written for someone who has **not been to class** since the midterm and has **not opened the lecture notes**. You have 8 days, you need a B−, and you do not have time to learn the material the long way.

This book is the only resource you need. I have boiled down five lectures, four problem sets, and two midterms into a story you can read straight through. The chapters are ordered by exam priority. The first two chapters are the ones you must master. Everything else is gravy.

What This Book Is, and What It Is Not

This is a teaching book, not a cheat sheet. Each chapter starts with a *story* (in plain English, no symbols) explaining what problem we are solving and why we are about to introduce a particular tool. Then we formalize it. Then we work through a representative exam example. Then we list the answer templates you should copy verbatim under exam pressure.

Read each chapter sequentially. Do not skim. The goal is not memorization but understanding the structure: *what does this technique do, when do we deploy it, and what does Patrik want to see in your answer.*

Reading Strategy: 8 Days to the Exam

The chapters and time budget assume you are starting from zero on the post-midterm material:

- **Days 1–2 (Apr 28–29)**: Chapter 1 *Q1 Template*. This material was on the midterm. You should already know it; this chapter is your refresher. Aim for one full pass and one self-test pass.
- **Days 3–4 (Apr 30–May 1)**: Chapter 2 *Bootstrap*. This is new material. Allocate two full days. The four-step consistency proof is the single most important page of the book.
- **Day 5 (May 2)**: Chapter 3 *Identification* and Chapter 4 *Hypothesis Tests*. Read carefully; do the worked examples.
- **Day 6 (May 3)**: Chapter 5 *Weak IV* and Chapter 8 *AsyCS*. These are the hardest conceptually. The goal is to write *some* of the answer correctly, not all of it.
- **Day 7 (May 4)**: Chapter 9 *Ridge / Lasso / Thresholding*. Three clean closed-form problems from HW7. Get them all.
- **Day 8 (May 5)**: Mock exam (midterm 2025) + cheat sheet + early sleep. **Do not study new material this day.**

Tier System

- **Tier 1** (must master): Chapters 1, 2. *Goal: 9/10 on Q1, 7/10 on Q3.*
- **Tier 2** (should master): Chapters 3, 4, 5. *Goal: 6/10 on Q2.*

- **Tier 3** (partial credit): Chapters 7, 8, 9. *Goal: 3-5/10 on Q4.*
- **Skip**: Chapter 6 (Bootstrap Improvements — low yield), Chapter 10 (Invariant Tests — never tested).

Box Color Code

- **Green: Definition**

  — foundational concepts you must internalize.

- **Blue: Theorem**

  — main results, often invoked directly on the exam.

- **Purple: Lemma**

  — intermediate results.

- **Pink: Assumption**

  — listed conditions; cite by name on the exam.

- **Yellow: Answer Template**

  — copy verbatim under exam pressure.

- **Red: Exam Strategy**

  — where to spend (and not spend) your time.

Proof Tier System

Not every proof needs to be memorized. Patrik covers many proofs in lecture for completeness, but only a subset are worth your scarce study time. Each proof in this textbook is tagged with one of four tiers at its first line:

- **[REPRODUCE]** — you must be able to reproduce this proof on the exam, line by line. These are the proofs Patrik tests directly. Memorize the structure and key steps.
- **[STRUCTURE]** — know the high-level steps and the names of the tools used (CLT, Slutsky, USCON, etc.). Acceptable to fudge algebraic details under exam pressure; partial credit is generous if your structure is right.
- **[INTUITION ONLY]** — read once for understanding. Do not waste time memorizing. The intuition is enough to write a one-paragraph English answer if asked, but you will not need to write algebra.
- **[SKIP]** — included only for completeness because the result is cited. If you are time-constrained (you are), *skip these proofs entirely.* The result itself may matter, but the derivation will not appear on the exam.

Triage rule: if a section is Tier 1 (must master) AND the proof is [REPRODUCE], that is your highest-priority study target. If a section is Tier 3 (partial credit) and the proof is [SKIP], ignore it without guilt.

A Note on Asymptotics Notation

Throughout, we use:

- $X_n \xrightarrow{p} X$: convergence in probability. Means: for any $\varepsilon > 0$, $P(|X_n - X| > \varepsilon) \to 0$.
- $X_n \xrightarrow{d} X$: convergence in distribution. Means: $P(X_n \leq x) \to P(X \leq x)$ for every continuity point of $P(X \leq x)$.
- $X_n = O_p(a_n)$: X_n/a_n is bounded in probability. "Bounded in probability" means: for every $\varepsilon > 0$, there is M s.t. $P(|X_n/a_n| > M) < \varepsilon$ for all large n.
- $X_n = o_p(a_n)$: $X_n/a_n \xrightarrow{p} 0$. "Vanishingly small."
- Most common: $\sqrt{n}(\widehat{\theta} - \theta) = O_p(1)$ (the $\sqrt{n}$-consistent rate).

If these are uncomfortable, that is fine; you absorb them by using them. They appear in every chapter.

Contents

Chapter 1

The Q1 Template: Estimator to Inference

Why This Chapter Matters Most

Q1 of every ECON 510 final is the same template. Patrik gives you an estimator (OLS, IV, or GMM) and walks you through: probability limit → asymptotic distribution → variance estimator → Wald test → confidence interval. Memorizing this template is the single highest-ROI thing you can do for the exam.

Target: 9 out of 10 points on Q1. The midterm 2026 user only got 5/10. This chapter exists to make sure that doesn't happen again on the final.

1.1 The Universal Setup: Extremum Estimators

Almost every estimator we care about is an *extremum estimator*. They share a single asymptotic theory.

Remark (Why "extremum").

> An extremum estimator is just an estimator defined as the minimizer (or maximizer) of some sample objective function $Q_n(\theta)$. "Extremum" = "optimum" (max or min). MLE maximizes a log-likelihood; OLS minimizes a sum of squares; GMM minimizes a quadratic form in the moment averages. Casting all three under the same umbrella lets us prove consistency / asymptotic normality *once* and apply the result everywhere.

Definition 1.1: Extremum Estimator (EE)

A sequence $\{\widehat{\theta}_n : n \geq 1\}$ of random elements of $\Theta \subseteq \mathbb{R}^d$ is an extremum estimator if

$$\widehat{\theta}_n \in \Theta, \qquad Q_n(\widehat{\theta}_n) \leq \inf_{\theta \in \Theta} Q_n(\theta) + o_p(1),$$

where $Q_n(\theta)$ is a stochastic "criterion function" that depends on the data.

The three estimators we will see on the exam:

Type	$Q_n(\theta)$	Total limit $Q(\theta)$
ML	$-n^{-1}\sum_i \log f(W_i, \theta)$	$-\mathbb{E}(\log f(W_i, \theta))$
LS / OLS	$n^{-1}\sum_i (Y_i - g(X_i, \theta))^2/2$	$\mathbb{E}\big((Y_i - g(X_i, \theta))^2\big)/2$
GMM	$\left\|A_n n^{-1}\sum_i g(W_i, \theta)\right\|^2/2$	$\|A \cdot \mathbb{E}(g(W_i, \theta))\|^2/2$

For OLS the criterion function is just sum of squared residuals; for IV the moment function is $g(W_i, \theta) = Z_i(Y_i - X_i'\theta)$; for ML it is the (negative) log-likelihood. **Treat them uniformly.**

Remark.

Why uniform treatment matters: under the exam clock, you should not be re-deriving anything from scratch. You should write "this is GMM with $g(W, \theta) = \ldots$" and immediately know which sandwich formula applies.

1.2 Step One: Consistency (Lecture 11)

Theorem 1.2: Consistency of EE (Andrews, Theorem 11.1)

Under Assumption 1.3 (EE), Assumption 1.4 (ID), and Assumption 1.5 (U-WCON),

$$\widehat{\theta}_n \xrightarrow{p} \theta_0.$$

Assumption 1.3: EE

$\widehat{\theta}_n$ approximately minimizes Q_n: $Q_n(\widehat{\theta}_n) \leq \inf_{\theta \in \Theta} Q_n(\theta) + o_p(1)$. (This is just the definition of EE.)

Remark (Why "approximately" and not exactly minimizes).

In practice, numerical optimizers (Newton, BFGS, simulated annealing) rarely return the *exact* global minimizer of Q_n — they return something within $o_p(1)$ of it. The $o_p(1)$ slack in the definition is there so that any reasonable algorithmic output qualifies as an EE; if we required exact minimization, the theory would not cover real-world implementations. The key technical content is that this slack is asymptotically negligible: $o_p(1) \to 0$ in

probability, so the slack does not contaminate the asymptotic distribution.

Assumption 1.4: ID (Identifiable Uniqueness)

There exists $\theta_0 \in \Theta$ such that for all $\varepsilon > 0$,

$$\inf_{\theta \notin B(\theta_0, \varepsilon)} Q(\theta) > Q(\theta_0).$$

That is, θ_0 is a *strict and isolated* minimizer of Q.

Remark (Why "strict and isolated" rather than just "unique").

Plain uniqueness ($Q(\theta_0) < Q(\theta)$ for all $\theta \neq \theta_0$) is too weak to give consistency. Counterexample: if Q has a flat "valley" touching its minimum — $Q(\theta)$ uniformly close to $Q(\theta_0)$ for θ in some far-away region — then a finite-sample Q_n that perturbs Q slightly can have its minimizer somewhere in the valley, far from θ_0. ID rules out flat valleys: it forces a strictly positive "gap" $\delta = \inf_{\theta \notin B(\theta_0, \varepsilon)} Q(\theta) - Q(\theta_0) > 0$ outside any neighborhood of θ_0. With Q_n uniformly close to Q, the minimizer of Q_n cannot escape into the gap region. "Strict and isolated" is exactly what makes the contradiction proof in the Almost-Sure Variant template (above) go through.

Assumption 1.5: U-WCON (Uniform Weak Convergence)

$\sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{p} 0$ for some non-stochastic $Q(\theta)$.

Remark (Why *uniform* and not just pointwise).

Pointwise convergence ($Q_n(\theta) \xrightarrow{p} Q(\theta)$ for each fixed θ) is not enough. The estimator $\widehat{\theta}_n$ is itself random and changes with n, so we cannot just plug it into a pointwise statement that was valid only at fixed θ. The uniform supremum bound lets us "follow" $\widehat{\theta}_n$ in — whatever value it takes, $Q_n(\widehat{\theta}_n)$ is close to $Q(\widehat{\theta}_n)$. Same logic recurs for the Hessian step in asymptotic normality (Lemma 12.1 below).

Theorem 1.6: Sufficient Conditions for ID and U-WCON (Andrews, Theorem 11.3)

Suppose $\{W_i\}$ are i.i.d., $m(w, \theta)$ is continuous in θ on Θ for all $w \in \mathcal{W}$, $\mathbb{E}(\sup_{\theta \in \Theta} |m(W_i, \theta)|) < \infty$, and Θ is compact. Then U-WCON holds for $Q_n(\theta) = n^{-1} \sum_{i=1}^{n} m(W_i, \theta)$. The same dominating-function argument also yields continuity of $Q(\theta) = \mathbb{E}(m(W_i, \theta))$, hence Assumption ID1 (compact Θ + continuous Q + unique minimizer) which implies ID.

Consistency Three-Step (Q1(b) Style)

When asked "under what conditions is $\widehat{\theta}_n$ consistent?":

1. **Step A.** Write $Q_n(\theta)$ and its pointwise limit $Q(\theta)$.
2. **Step B.** Verify ID: state that θ_0 is the unique minimizer of $Q(\theta)$, and provide the standard short argument (Jensen's inequality for ML; expansion + iterated expectations for LS; full-rank moment matrix for GMM).
3. **Step C.** Verify U-WCON: invoke Theorem 1.2 with the explicit moment bound $\mathbb{E}(\sup_\theta |m(W_i, \theta)|) < \infty$.

Always cite Theorem 11.1. Patrik wrote it; he wants to see it used.

Almost-Sure Variant of the Consistency Proof (HW9 Q4)

[STRUCTURE — know the steps, fudge details]

Patrik occasionally asks for the *almost sure* version: assume USCON ($\sup_\theta |Q_n(\theta) - Q(\theta)| \to 0$ a.s.) and ID, conclude $\widehat{\theta}_n \to \theta_0$ a.s.

Proof. Take a sample path $\omega \in \Omega$ for which USCON holds (occurs with probability 1). Proceed by contradiction: suppose $\widehat{\theta}_n \to \theta_0$ fails along this path. Then there exists $\varepsilon > 0$ and a subsequence n_i with $\left\|\widehat{\theta}_{n_i} - \theta_0\right\| > \varepsilon$. By Assumption ID,

$$\delta := \inf_{\theta \notin B(\theta_0, \varepsilon)} Q(\theta) - Q(\theta_0) > 0.$$

By USCON, choose M so large that for all $n \geq M$, $\sup_\theta |Q_n(\theta) - Q(\theta)| < \delta/3$. Then for $n_i \geq M$,

$$Q(\widehat{\theta}_{n_i}) - \delta/3 < Q_{n_i}(\widehat{\theta}_{n_i}) \leq Q_{n_i}(\theta_0) < Q(\theta_0) + \delta/3,$$

where the middle inequality uses that $\widehat{\theta}_{n_i}$ minimizes Q_{n_i}. Hence $Q(\widehat{\theta}_{n_i}) - Q(\theta_0) < 2\delta/3 < \delta$, contradicting the definition of δ. ■

Why this matters: the standard Theorem 11.1 gives convergence in probability under U-WCON (uniform $\xrightarrow{p}$); the same argument with U-S-CON (uniform $\xrightarrow{a.s.}$) gives almost-sure convergence. The structure of the proof is identical.

1.3 Step Two: Asymptotic Normality (Lecture 12)

Theorem 1.7: Asymptotic Normality of EE (Andrews, Theorem 12.1)

Under Assumption 1.8 and Assumption 1.9,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0,\ {B_0}^{-1}\,\Omega_0\,{B_0}^{-1}\right).$$

The matrix ${B_0}^{-1}\,\Omega_0\,{B_0}^{-1}$ is the famous *sandwich* or *Eicker–White* matrix.

Assumption 1.8: EE2

(i) $\widehat{\theta}_n \xrightarrow{p} \theta_0$ (consistency, established in the previous step).
(ii) $\frac{\partial}{\partial\theta} Q_n(\widehat{\theta}_n) = o_p(\sqrt{n}^{-1})$ (the FOC holds approximately).

Assumption 1.9: CF (Criterion Function)

(i) θ_0 is in the interior of Θ.
(ii) Q_n is twice continuously differentiable on a neighborhood Θ_0 of θ_0.
(iii) $\sqrt{n}\frac{\partial}{\partial\theta} Q_n(\theta_0) \xrightarrow{d} N(0, \Omega_0)$ (CLT for the score).
(iv) $\sup_{\theta\in\Theta_0} \left\| \frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\theta) - B(\theta) \right\| \xrightarrow{p} 0$, with $B(\theta)$ continuous at θ_0 and $B_0 := B(\theta_0)$ nonsingular.

Recipe to identify B_0 and Ω_0.

Estimator	B_0	Ω_0
ML (correct)	$-\mathbb{E}\left(\frac{\partial^2 \log f(W_i,\theta_0)}{\partial\theta\partial\theta'}\right)$	$\mathbb{E}\left(\frac{\partial \log f}{\partial\theta}\frac{\partial \log f}{\partial\theta'}\right)$
ML (correct, info matrix eq.)	$B_0 = \Omega_0 \implies \text{var} = {B_0}^{-1}$	
LS (correct, possibly heterosk)	$\mathbb{E}\left(\frac{\partial g}{\partial\theta}\frac{\partial g}{\partial\theta'}\right)$	$\mathbb{E}\left(U_i^2\frac{\partial g}{\partial\theta}\frac{\partial g}{\partial\theta'}\right)$
GMM ($k = d$, just-id)	$\Gamma_0' A' A\,\Gamma_0$	$\Gamma_0' A' A\,V_0\,A' A\,\Gamma_0$
GMM ($k = d$, simplified)	$\text{var} = {\Gamma_0}^{-1}\,V_0\,{\Gamma_0}^{-1\prime}$	

Here $\Gamma_0 = \mathbb{E}\left(\frac{\partial g(W_i,\theta_0)}{\partial\theta'}\right)$ and $V_0 = \mathbb{E}(g(W_i,\theta_0)g(W_i,\theta_0)')$.

Remark (Why we expand the FOC, not the estimator).

> The standard "trick" below is to mean-value-expand the *first-order condition* $\frac{\partial}{\partial\theta} Q_n(\widehat{\theta}_n) \approx 0$, not $\widehat{\theta}_n$ itself. Reason: $\widehat{\theta}_n$ has no closed-form expression in terms of the data, so we cannot Taylor-expand it. But the FOC *is* a function of θ (parameter) and the sample, which we know how to differentiate. The expansion converts "$\widehat{\theta}_n - \theta_0$" into something proportional to a sample average ($\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial\theta}$), where CLT can take over.

The Four-Step Asymptotic Normality Proof (Critical Recipe)

[REPRODUCE — memorize this proof]

On Q1(c) you will be asked to derive this from scratch. **Memorize these four lines**:

(1) Mean-value expansion of the FOC about θ_0:

$$o_p(\sqrt{n}^{-1}) = \frac{\partial}{\partial\theta}Q_n(\widehat{\theta}_n) = \frac{\partial}{\partial\theta}Q_n(\theta_0) + \frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n^*)(\widehat{\theta}_n - \theta_0).$$

(2) Hessian converges to B_0:

$$\frac{\partial^2}{\partial\theta\partial\theta'}Q_n(\theta_n^*) = B_0 + o_p(1)$$

by U-WCON of the second derivative (Lemma 12.1) and consistency of $\widehat{\theta}_n$.

Lemma 12.1 (content). If $\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta'}$ satisfies U-WCON,

$$\sup_{\theta\in\Theta}\left\|\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta'} - B(\theta)\right\| \xrightarrow{p} 0, \quad B(\theta) := \frac{\partial^2 Q(\theta)}{\partial\theta\partial\theta'} \text{ continuous,}$$

and if $\theta_n^* \xrightarrow{p} \theta_0$ (which holds because θ_n^* lies between $\widehat{\theta}_n$ and θ_0, and $\widehat{\theta}_n \xrightarrow{p} \theta_0$), then

$$\frac{\partial^2 Q_n(\theta_n^*)}{\partial\theta\partial\theta'} \xrightarrow{p} B(\theta_0) =: B_0.$$

Why we need both conditions: pointwise convergence at the fixed point θ_0 is not enough — θ_n^* is itself random and varies with n, so we need uniform control over θ to "follow" θ_n^* in. The triangle inequality argument is

$$\left\|\frac{\partial^2 Q_n(\theta_n^*)}{\partial\theta\partial\theta'} - B_0\right\| \le \underbrace{\sup_\theta\left\|\frac{\partial^2 Q_n(\theta)}{\partial\theta\partial\theta'} - B(\theta)\right\|}_{o_p(1) \text{ by U-WCON}} + \underbrace{\|B(\theta_n^*) - B(\theta_0)\|}_{o_p(1) \text{ by continuity + consistency}}.$$

(3) Solve for $\sqrt{n}(\widehat{\theta}_n - \theta_0)$:

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -{B_0}^{-1} \cdot \sqrt{n}\frac{\partial}{\partial\theta}Q_n(\theta_0) + o_p(1).$$

(4) Apply CLT + Slutsky: $\sqrt{n}\frac{\partial}{\partial\theta}Q_n(\theta_0) \xrightarrow{d} N(0, \Omega_0)$, hence

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, {B_0}^{-1}\,\Omega_0\,{B_0}^{-1}).$$

That's it. Memorize the four steps; identify B_0 and Ω_0 for the specific estimator; you have 6 of the 10 points already.

1.4 Step Three: Variance Estimation (Lecture 13)

The asymptotic variance ${B_0}^{-1}\,\Omega_0\,{B_0}^{-1}$ is unknown. Replace expectations with sample averages, replace θ_0 with $\widehat{\theta}_n$.

Definition 1.10: Eicker–White Variance Estimator

$$\widehat{B}_n = \frac{\partial^2}{\partial\theta\partial\theta'} Q_n(\widehat{\theta}_n), \qquad \widehat{\Omega}_n = \widehat{\mathrm{Var}}\left(\sqrt{n}\frac{\partial}{\partial\theta} Q_n(\theta_0)\right)\Big|_{\theta=\widehat{\theta}_n}.$$

The variance estimator is the *sandwich*:

$$\widehat{\Sigma}_n = \widehat{B}_n^{-1}\,\widehat{\Omega}_n\,\widehat{B}_n^{-1} \xrightarrow{p} {B_0}^{-1}\,\Omega_0\,{B_0}^{-1}.$$

For OLS (the most common case):

$$\widehat{B}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i', \qquad \widehat{\Omega}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{U}_i^2 X_i X_i', \qquad \widehat{U}_i = Y_i - X_i'\widehat{\beta}_{LS}.$$

Remark (Why "sandwich" and why "Eicker–White").

The name "sandwich" refers to the three-layer structure $B^{-1}\cdot\Omega\cdot B^{-1}$ — the meat Ω between two slices of bread B^{-1}. The name "Eicker–White" honors Eicker (1967) and White (1980), who independently proved $\widehat{\Sigma}_n \xrightarrow{p} {B_0}^{-1}\Omega_0{B_0}^{-1}$ *without* requiring conditional homoskedasticity. Under homoskedasticity ($\mathbb{E}\big(U_i^2 \mid X_i\big) = \sigma^2$ a.s.), $\Omega_0 = \sigma^2 B_0$ and the sandwich collapses to the simpler $\sigma^2 {B_0}^{-1}$ — this is the variance you see in introductory OLS courses. With heteroskedasticity, the simpler form is biased; the sandwich is consistent. This is the variance estimator behind every empirical paper labeled "robust standard errors" or "HC0/HC1."

Proving $\widehat{\Omega}_n \xrightarrow{p} \Omega_0$ (Q1(c) Style)

[REPRODUCE — memorize this proof]

This is the *residual replacement trick*. Decompose:

$$\widehat{\Omega}_n - \Omega_0 = \underbrace{\frac{1}{n}\sum_{i=1}^{n}(\widehat{U}_i^2 - U_i^2)X_i X_i'}_{\text{(A)}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} U_i^2 X_i X_i' - \mathbb{E}\big(U_i^2 X_i X_i'\big)}_{\text{(B)}}.$$

Term (B) $\xrightarrow{p} 0$ by WLLN under $\mathbb{E}\Big(\|U_i X_i\|^2\Big) < \infty$.

Term (A): use $\widehat{U}_i = U_i + X_i'(\beta_0 - \widehat{\beta}) = U_i + X_i' O_p(\sqrt{n}^{-1})$, hence

$$\widehat{U}_i^2 - U_i^2 = 2U_i X_i' O_p(\sqrt{n}^{-1}) + (X_i' O_p(\sqrt{n}^{-1}))^2.$$

Each piece, multiplied by $X_i X_i'$ and averaged, converges in probability to zero by WLLN, using $\mathbb{E}(U_i X_{ij} X_i X_i') = \mathbb{E}(X_{ij} X_i X_i' \mathbb{E}(U_i \mid X_i)) = 0$ (LIE).

1.5 Step Four: Wald Test and Confidence Region (Lecture 14)

Definition 1.11: Wald Statistic

For testing $H_0 : h(\theta_0) = 0$ vs $H_1 : h(\theta_0) \neq 0$, where $h : \mathbb{R}^d \to \mathbb{R}^r$,

$$\mathcal{W}_n = n \cdot h(\widehat{\theta}_n)' \left[\widehat{H}_n \widehat{\Sigma}_n \widehat{H}_n'\right]^{-1} h(\widehat{\theta}_n),$$

where $\widehat{H}_n = \frac{\partial h}{\partial \theta'}(\widehat{\theta}_n)$.

Theorem 1.12: Asymptotic Null Distribution (Andrews, Theorem 14.1)

Under EE2, CF, R (rank conditions on h and Ω_0), and COV (consistent variance estimators),

$$\mathcal{W}_n \xrightarrow{d} \chi^2_r \quad \text{under } H_0.$$

The two simple cases that show up most:

Case 1: $H_0 : \theta = 0$ ($h =$ identity, so $H = I_d$).

$$\mathcal{W}_n = n\, \widehat{\theta}_n'\, \widehat{\Sigma}_n^{-1}\, \widehat{\theta}_n \xrightarrow{d} \chi^2_d.$$

Case 2: scalar parameter $H_0 : \beta_0 = 0$.

$$\mathcal{W}_n = \frac{n \cdot \widehat{\beta}_n^2}{\widehat{V}_n} \xrightarrow{d} \chi^2_1.$$

Remark (What Case 2 actually means).

This is the special case where you're testing only *one* coordinate of θ_0. Two equivalent ways to read it:

- **Either** θ_0 is a scalar to begin with — then $\beta_0 = \theta_0$ and $\widehat{V}_n = \widehat{\Sigma}_n$ (just a scalar variance estimator, the (1,1) sandwich).
- **Or** $\theta_0 = (\alpha_0, \beta_0)'$ is a vector and you only want to test the slope. Then $h(\alpha, \beta) = \beta$, the Jacobian is $H = (0, 1)$, and

$$\widehat{V}_n \;=\; H\widehat{\Sigma}_n H' \;=\; \left[\widehat{\Sigma}_n\right]_{2,2} \;=\; \text{the bottom-right entry of the sandwich.}$$

In the IV worked example below, we are in the second case: $\widehat{\beta}_n = \widehat{\beta}_{IV,n}$ is the IV slope and $\widehat{V}_n = n^{-1}\sum_{i=1}^n (Z_i \widehat{U}_i)^2 \,/\, (n^{-1}\sum_{i=1}^n Z_i X_i)^2$ is its (scalar) sandwich variance.

Wald Intuition Sentence (Q1(e))

Whenever asked to "describe the intuition", use this exact paragraph:

> *$\widehat{\theta}_n$ is an estimator of the true θ_0. If H_0 is true, $h(\widehat{\theta}_n)$ should be close to zero, so the quadratic form $\mathcal{W}_n$ in $h(\widehat{\theta}_n)$ should be small. If H_0 is false, $h(\widehat{\theta}_n)$ converges to $h(\theta_0) \neq 0$ and $\mathcal{W}_n$ blows up. The weighting matrix is a consistent estimator of the inverse asymptotic variance, so under H_0 the limit distribution is χ^2_r.*

Confidence region by inversion.

$$\mathrm{CR}_{1-\alpha} = \{\theta_0 \in \Theta : \mathcal{W}_n \text{ does not reject } H_0 : \theta = \theta_0\}.$$

CI Validity Argument (Q1(f))

Pointwise asymptotic validity:

$$\mathrm{P}(\theta_0 \in \mathrm{CR}_{1-\alpha}) = \mathrm{P}(\mathcal{W}_n \text{ fails to reject true } H_0) = 1 - \mathrm{P}(\mathcal{W}_n \text{ rejects true } H_0) \to 1-\alpha,$$

since by Theorem 1.5 the Wald test has correct asymptotic level. *Spelled out:* Theorem 1.5 says that under the (true) null $H_0 : h(\theta_0) = 0$, $\mathcal{W}_n \xrightarrow{d} \chi^2_r$. The test rejects when $\mathcal{W}_n > \chi^2_{r,\,1-\alpha}$. Since the limiting CDF of χ^2_r is continuous at the critical value, the portmanteau theorem gives

$$\mathrm{P}\big(\mathcal{W}_n > \chi^2_{r,\,1-\alpha} \,|\, H_0 \text{ true}\big) \to \mathrm{P}\big(\chi^2_r > \chi^2_{r,\,1-\alpha}\big) = \alpha.$$

Inverting (taking complements) yields the $1-\alpha$ coverage of the CR.
Caveat (which Patrik likes you to mention): this is *pointwise* validity — the DGP is fixed. If the DGP is allowed to change with n (e.g., weak instruments), the coverage can be much smaller than $1-\alpha$. This is what Chapter 8 addresses.

1.6 Worked Example: Midterm 2026 Q1 (IV from scratch)

Original Problem Statement (Midterm 2026 Q1, 10 pts)

Model. Consider $Y_i = \alpha_0 + \beta_0 X_i + U_i$ for $i = 1, \ldots, n$, where Z_i is an instrumental variable. The data (Y_i, X_i, U_i, Z_i) are i.i.d. with finite means and variances. The IV estimator of β_0 is

$$\widehat{\beta}_{IV,n} = \left(\frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)X_i\right)^{-1} \cdot \frac{1}{n}\sum_{i=1}^{n}(Z_i - \bar{Z}_n)Y_i.$$

Tasks.

(a) (2 pts) Find the probability limit of $\widehat{\beta}_{IV,n}$. State the assumptions you need.

(b) (1 pt) Give a sufficient condition for $\widehat{\beta}_{IV,n}$ to be consistent for β_0.

(c) (2 pts) Derive the asymptotic distribution of $\sqrt{n}(\widehat{\beta}_{IV,n} - \beta_0)$ from scratch.

(d) (1 pt) Propose a consistent estimator of the asymptotic variance.

(e) (2 pts) Construct the Wald statistic for $H_0 : \beta_0 = 0$ and give its asymptotic null distribution.

(f) (2 pts) Construct a 95% confidence interval for β_0 and prove its pointwise asymptotic validity.

Solution. The model and estimator are as above; (Y_i, X_i, U_i, Z_i) are i.i.d. with finite means and variances throughout.

(a) Probability Limit (2 pts)

$$\widehat{\beta}_{IV,n} \xrightarrow{p} \frac{\mathbb{E}(Z_iX_i) - \mathbb{E}(Z_i)\,\mathbb{E}(X_i)}{\mathbb{E}(Z_iX_i) - \mathbb{E}(Z_i)\,\mathbb{E}(X_i)} \cdot \beta_0 + \frac{\mathbb{E}(Z_iU_i) - \mathbb{E}(Z_i)\,\mathbb{E}(U_i)}{\mathbb{E}(Z_iX_i) - \mathbb{E}(Z_i)\,\mathbb{E}(X_i)}.$$

By WLLN and Slutsky's theorem, assuming $\mathbb{E}(Z_iX_i) - \mathbb{E}(Z_i)\,\mathbb{E}(X_i) \neq 0$ (instrument relevance), and finiteness of all relevant first moments.

(b) Consistency (1 pt)

If $\mathrm{Cov}(Z_i, U_i) = \mathbb{E}(Z_iU_i) - \mathbb{E}(Z_i)\,\mathbb{E}(U_i) = 0$ (instrument exogeneity), then $\widehat{\beta}_{IV,n} \xrightarrow{p} \beta_0$.

(c) Asymptotic Distribution (2 pts)

Plug in $Y_i = \alpha_0 + \beta_0 X_i + U_i$ and rearrange:

$$\sqrt{n}(\widehat{\beta}_{IV,n} - \beta_0) = \left(\frac{1}{n}\sum_{i=1}^{n} Z_iX_i - \bar{Z}_n\bar{X}_n\right)^{-1} \cdot \frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} Z_iU_i - \bar{Z}_n \cdot \frac{1}{n}\sum_{i=1}^{n} U_i \cdot \sqrt{n}\right).$$

Assume $\mathbb{E}(Z_iU_i) = \mathbb{E}(Z_i) = \mathbb{E}(U_i) = 0$ (simplification), $\mathbb{E}\left(\|Z_iU_i\|^2\right) < \infty$, $\mathbb{E}\big((Z_iU_i)^2\big) > 0$. By CLT,

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_iU_i - \left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} U_i\right) \xrightarrow{d} N(0,\, \mathbb{E}\big((Z_iU_i)^2\big)),$$

since the second product is $O_p(1) \cdot o_p(1) = o_p(1)$.

Remark (Why "$\mathbb{E}(Z_i) = \mathbb{E}(U_i) = 0$" is WLOG — not a substantive assumption).

> The only *substantive* exogeneity assumption is $\mathbb{E}(Z_iU_i) = 0$. The other two zero-mean conditions are notational simplifications that you can always achieve, for the following reasons:
>
> - $\mathbb{E}(U_i) = 0$ **is free** because the model contains an intercept α_0. Any non-zero $\mathbb{E}(U_i)$

would just be absorbed into α_0, so without loss of generality we redefine $U_i \leftarrow U_i - \mathbb{E}(U_i)$ and $\alpha_0 \leftarrow \alpha_0 + \mathbb{E}(U_i)$. The slope β_0 is unchanged.

- $\mathbb{E}(Z_i) = 0$ **is free** because the estimator already uses the demeaned instrument $(Z_i - \bar{Z}_n)$. Algebraically, replacing Z_i by $Z_i - \mu_Z$ for any constant μ_Z does not change $(Z_i - \bar{Z}_n)$ at all (the constant cancels), so $\widehat{\beta}_{IV,n}$ and $\sqrt{n}(\widehat{\beta}_{IV,n} - \beta_0)$ are *numerically identical* whether or not $\mathbb{E}(Z_i) = 0$.

What if you didn't assume them? Then in the algebra you would carry the cross-product term $\bar{Z}_n \cdot n^{-1}\sum_{i=1}^n U_i \cdot \sqrt{n}$. With $\mathbb{E}(Z_i) = \mu_Z$ and $\mathbb{E}(U_i) = \mu_U$, by WLLN $\bar{Z}_n \xrightarrow{p} \mu_Z$ and $n^{-1}\sum_{i=1}^n U_i \xrightarrow{p} \mu_U$, and a longer CLT-with-centering argument shows that the limiting variance is exactly $\mathrm{Var}(Z_iU_i) = \mathbb{E}\big((Z_i - \mu_Z)^2(U_i - \mu_U)^2\big)$ when $\mathrm{Cov}(Z_i, U_i) = 0$. So nothing actually goes wrong — you'd just write $\mathrm{Var}(Z_iU_i)$ instead of $\mathbb{E}\big((Z_iU_i)^2\big)$ in the answer, and the two coincide whenever $\mathbb{E}(Z_iU_i) = 0$ and one of μ_Z, μ_U is zero. Patrik accepts the simplification because it cuts a page of bookkeeping without changing the final answer.

By Slutsky,

$$\boxed{\sqrt{n}(\widehat{\beta}_{IV,n} - \beta_0) \xrightarrow{d} N\left(0,\ \frac{\mathbb{E}\big((Z_iU_i)^2\big)}{\big(\mathbb{E}(Z_iX_i)\big)^2}\right).}$$

(d) Variance Estimator + Consistency (1 pt)

$$\widehat{V}_n = \frac{n^{-1}\sum_{i=1}^n (Z_i\widehat{U}_i)^2}{\big(n^{-1}\sum_{i=1}^n Z_iX_i\big)^2}, \qquad \widehat{U}_i := Y_i - \widehat{\alpha} - \widehat{\beta}\, X_i.$$

Consistency follows by the residual replacement trick of the previous section.

(e) Wald Statistic (2 pts)

Test $H_0 : \beta_0 = 0$. Here $h(\alpha, \beta) = \beta$ and $H = (0, 1)$, so

$$\mathcal{W}_n = \frac{n\,\widehat{\beta}^2_{IV,n}}{\widehat{V}_n} \xrightarrow{d} \chi^2_1 \text{ under } H_0.$$

(f) 95% Confidence Interval (2 pts)

$$\mathrm{CI}_n(\beta_0) = \left\{\beta \in \mathbb{R} : \mathcal{W}_n(\beta_0 = \beta) \le \chi^2_{1,0.95}\right\}, \qquad \mathcal{W}_n(\beta_0 = \beta) = \frac{n(\widehat{\beta}_{IV,n} - \beta)^2}{\widehat{V}_n}.$$

Pointwise validity: $\mathrm{P}(\beta_0 \in \mathrm{CI}_n) \to 0.95$ for fixed DGP.
Caveat: this fails uniformly over weak-IV DGPs (Chapter 5).

1.7 One-Page Cheat Sheet

Q1 Cheat Sheet (memorize this verbatim)

1. **Plim:** write $\widehat{\theta}_n$ as a ratio of sample averages, apply WLLN + CMT, plug in DGP relations.
2. **Consistency:** requires exogeneity (e.g., $\mathbb{E}(Z_i U_i) = 0$).
3. **Asymptotic distribution:**
   (1) Mean-value expand FOC: $o_p(\sqrt{n}^{-1}) = \frac{\partial Q_n(\widehat{\theta}_n)}{\partial \theta} = \frac{\partial Q_n(\theta_0)}{\partial \theta} + \frac{\partial^2 Q_n(\theta_n^*)}{\partial \theta \partial \theta'}(\widehat{\theta}_n - \theta_0)$.
   (2) Hessian $\xrightarrow{p} B_0$ by U-WCON.
   (3) Rearrange: $\sqrt{n}(\widehat{\theta}_n - \theta_0) = -{B_0}^{-1}\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta} + o_p(1)$.
   (4) CLT: $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, {B_0}^{-1}\Omega_0 {B_0}^{-1})$.
4. **Variance estimator:** sandwich, residual replacement trick for consistency.
5. **Wald:** $\mathcal{W}_n = n\, h(\widehat{\theta}_n)'[\widehat{H}_n \widehat{\Sigma}_n \widehat{H}_n']^{-1} h(\widehat{\theta}_n) \xrightarrow{d} \chi^2_r$.
6. **CI by inversion:** $\mathrm{P}(\theta_0 \in \mathrm{CI}) \to 1 - \alpha$ pointwise; mention uniform caveat.

1.8 Self-Test Problems

Example (Self-Test 1: OLS from scratch).

$Y_i = X_i'\beta_0 + U_i$, $\mathbb{E}(U_i \mid X_i) = 0$ a.s., $\mathbb{E}\left(U_i^2 \mid X_i\right) < \infty$. Derive the asymptotic distribution of $\widehat{\beta}_{LS}$ from scratch.

Solution.

$\widehat{\beta}_{LS} = (X'X)^{-1}X'Y$. Plug in $Y = X\beta_0 + U$:

$$\sqrt{n}(\widehat{\beta}_{LS} - \beta_0) = (X'X/n)^{-1} \cdot X'U/\sqrt{n}.$$

- By WLLN + CMT: $(X'X/n)^{-1} \xrightarrow{p} \mathbb{E}(X_i X_i')^{-1}$ (assuming $\mathbb{E}\left(\|X_i\|^2\right) < \infty$ and $\mathbb{E}(X_i X_i') > 0$).
- By CLT: $X'U/\sqrt{n} \xrightarrow{d} N(0, \mathbb{E}\left(U_i^2 X_i X_i'\right))$ (assuming $\mathbb{E}\left(\|U_i X_i\|^2\right) < \infty$; mean is $\mathbb{E}(U_i X_i) = \mathbb{E}(X_i \mathbb{E}(U_i \mid X_i)) = 0$ by LIE).
- By Slutsky: $\sqrt{n}(\widehat{\beta}_{LS} - \beta_0) \xrightarrow{d} N(0, \Sigma)$ with

$$\Sigma = \mathbb{E}(X_i X_i')^{-1}\mathbb{E}\left(U_i^2 X_i X_i'\right)\mathbb{E}(X_i X_i')^{-1}.$$

Example (Self-Test 2: Wald-CI Validity).

Construct a 95% Wald CI for β in the IV model and prove pointwise asymptotic validity.

Solution.

$$\mathrm{CI}_n = \left\{\beta_0 : \mathcal{W}_n(H_0 : \beta = \beta_0) \leq \chi^2_{1,0.95}\right\}.$$

$$\mathrm{P}(\beta_0 \in \mathrm{CI}_n) = 1 - \mathrm{P}(\mathcal{W}_n \text{ rejects true } H_0) \to 1 - 0.05 = 0.95$$

by Theorem 1.5. *Pointwise* because the DGP is fixed.

Chapter 2

Bootstrap

The Story (Read This First, No Math)

Suppose you have computed $\widehat{\beta}$ from your data and want to test whether the true value β_0 equals zero. The standard procedure runs as follows: you build a t-statistic $t_n = \sqrt{n}(\widehat{\beta} - 0)/\widehat{\sigma}$, you compare it to the critical value 1.96 (the 97.5th percentile of $N(0,1)$), and you reject if $|t_n| > 1.96$.

Why 1.96*?* Because asymptotic theory says $t_n \xrightarrow{d} N(0,1)$ when H_0 is true. So when n is large, t_n behaves *approximately* like a standard normal random variable, and 1.96 is what you would use if it were *exactly* normal.

What goes wrong with this? For small or moderate samples, the actual distribution of t_n may be quite different from $N(0,1)$. It might be skewed. It might have heavier tails. The normal approximation, while justified asymptotically, can produce confidence intervals that genuinely have 80% coverage when you advertised 95%.

Bootstrap is one way to do better. Instead of relying on the normal approximation, we estimate the distribution of t_n *directly from the data* and use a critical value computed from that estimated distribution.

The trick is this. The distribution of t_n depends on the unknown true distribution F of the data. If we knew F, we could compute the distribution of t_n exactly (in principle). We do not know F, but we have a sample drawn from F, and we can construct a sample-based estimate $\widehat{F}_n$ that is close to F. So we substitute $\widehat{F}_n$ for F, and use the distribution of the test statistic computed from samples drawn from $\widehat{F}_n$ as our estimate of the distribution of t_n.

That is the entire idea. Everything else in this chapter is detail: how exactly to estimate F, how exactly to draw the resamples, how to prove that the substitution is asymptotically valid, and what to do when it is not.

What You Need to Take Away

By the end of this chapter you should be able to do four things on the final exam:

1. **Implement** a parametric or nonparametric bootstrap test (Section 2.4).
2. **State** the formal definition of bootstrap consistency (Definition 2.6).

3. **Prove** bootstrap consistency in a location model using the four-step template (Section 2.7). This is the single most important page in the book.

4. **Recognize** settings where the bootstrap fails, e.g., parameter on the boundary (Section 2.10).

2.1 The Empirical Distribution Function

The first idea is what to use for $\widehat{F}_n$. The simplest choice, and the one Patrik uses everywhere, is the empirical distribution function (EDF).

Definition 2.1: Empirical Distribution Function

Given an i.i.d. sample $W_1, \ldots, W_n$ from an unknown distribution F, the EDF is

$$\widehat{F}_n(w) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(W_i \leq w).$$

$\widehat{F}_n$ is a step function on $\mathbb{R}$ that jumps by $1/n$ at each observation W_i. It is itself a CDF: it is the CDF of a discrete distribution that places equal mass $1/n$ on each of the observed sample points $W_1, \ldots, W_n$.

In words: $\widehat{F}_n$ asks "what fraction of my sample is at or below w?" For example, if you have $n = 10$ observations and exactly 7 of them are ≤ 2.0, then $\widehat{F}_n(2.0) = 0.7$.

The EDF estimates F in a strong sense:

Theorem 2.2: Glivenko–Cantelli

$\sup_w \left| \widehat{F}_n(w) - F(w) \right| \xrightarrow{a.s.} 0$ as $n \to \infty$. The convergence is uniform in w, almost surely.

You do not need to memorize the proof. The takeaway: *the EDF, viewed as a function of w, is uniformly close to the unknown F when n is large.* This is the foundation of bootstrap.

Remark.

> The bootstrap recipe in one sentence: do all your calculations as if $\widehat{F}_n$ were the true distribution, and report the answer.

2.2 The Three Bootstrap Procedures You Need

Patrik has shown the class three flavors of bootstrap. They all share the substitution idea; they differ in *how they generate* W_i^* (the bootstrap sample).

2.2.1 Nonparametric (NP) iid Bootstrap

Recipe. Given the original sample $\{W_1, \ldots, W_n\}$:

1. Draw n values *with replacement* from the original sample. Call these $W_1^*, \ldots, W_n^*$.
2. Equivalently: each W_i^* is chosen independently from $\widehat{F}_n$.

This is what people usually mean when they say "the bootstrap."

When to use. The default. Minimal assumptions. Works for i.i.d. data without a parametric model.

Example. If your sample is $\{2.1, 1.7, 4.3, 0.5\}$ ($n = 4$), one possible bootstrap sample is $\{1.7, 1.7, 4.3, 0.5\}$. Another is $\{2.1, 0.5, 0.5, 2.1\}$. Each draws each original observation with probability $1/4$, with replacement.

2.2.2 Parametric Bootstrap

Recipe. If you are willing to assume your data follows a parametric model $F(\cdot, \theta)$:

1. Estimate $\widehat{\theta}_n$ from the original sample (e.g., MLE).
2. Draw $W_1^*, \ldots, W_n^*$ i.i.d. from the parametric distribution $F(\cdot, \widehat{\theta}_n)$.

When to use. When the parametric model is credible. The parametric bootstrap squeezes more information out of the sample by using the assumed shape.

Example. Patrik's midterm 2025 Q3 used $W_i \sim N(\theta_1, \theta_2)$ i.i.d. The parametric bootstrap draws $W_{ib}^* \sim N(\widehat{\theta}_1, \widehat{\theta}_2)$.

2.2.3 Residual-Based Bootstrap (and Its Wild Variant)

Recipe (homoskedastic version). Given a regression $Y_i = X_i'\beta + U_i$:

1. Estimate $\widehat{\beta}$ and compute residuals $\widehat{U}_i = Y_i - X_i'\widehat{\beta}$.
2. Draw $\widehat{U}_i^*$ i.i.d. from the EDF of the residuals.
3. Set $Y_i^* = X_i'\widehat{\beta} + \widehat{U}_i^*$. The regressors X_i are kept fixed.

When to use. A regression model with i.i.d. *homoskedastic* errors. If errors are heteroskedastic, residuals do not capture the right variance.

Wild bootstrap. For heteroskedastic errors, replace step 2 with $\widehat{U}_i^* = \widehat{U}_i \cdot \varepsilon_i^*$ where ε_i^* is independent of the data with $\mathbb{E}(\varepsilon_i^*) = 0$, $\mathbb{E}\big((\varepsilon_i^*)^2\big) = 1$, $\mathbb{E}\big((\varepsilon_i^*)^3\big) = 1$ (the so-called Mammen distribution). The wild bootstrap preserves the heteroskedasticity in $\widehat{U}_i$.

2.3 The Centering Subtlety

Here is something that confuses everyone the first time. Suppose you want to estimate the distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$. You do not know θ_0. So how do you center the bootstrap statistic?

The bootstrap statistic must be centered at $\widehat{\theta}_n$, not at θ_0.

That is, you compute $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$, where $\widehat{\theta}_n^*$ is the same estimator applied to the bootstrap sample.

Why? In the bootstrap world, $\widehat{F}_n$ is the "true" distribution. The mean of $\widehat{F}_n$ is the sample mean $\bar{W}_n$, which is the value $\widehat{\theta}_n$ converges to in the bootstrap world. So $\widehat{\theta}_n$ plays the role of the "true parameter" inside the bootstrap.

What goes wrong if you fail to recenter? If you use $\sqrt{n}(\widehat{\theta}_n^* - \theta_0)$, that quantity contains $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ as a non-vanishing component. It does not approximate the limit distribution of $\sqrt{n}(\widehat{\theta}_n - \theta_0)$ at all.

If You Forget to Recenter, You Lose Points

This is the single easiest mistake in a bootstrap problem. Always write $\widehat{\theta}_n^* - \widehat{\theta}_n$, never $\widehat{\theta}_n^* - \theta_0$.

2.4 Bootstrap Tests: How to Run One

This section is the answer template for "explain how to implement the test." Memorize this structure.

Bootstrap Test Implementation (midterm 2025 Q3(a) Style)

To test $H_0 : \theta_1 = 0$ at nominal size[a] α:

Step 1: Pick a test statistic. For testing a single parameter, use the absolute t-statistic:

$$t_n := \left| \sqrt{n}\, \widehat{\theta}_1 / \mathrm{sd}(\widehat{\theta}_1) \right|,$$

where $\mathrm{sd}(\widehat{\theta}_1)$ is your variance estimator (the sandwich, or whatever the problem dictates).

Step 2: Generate B bootstrap samples. Use NP, parametric, or residual bootstrap as appropriate. Take $B = 999$.

Step 3: Compute the recentered statistic on each bootstrap sample.

$$t_n^* := \left| \sqrt{n}\, (\widehat{\theta}_1^* - \widehat{\theta}_1) / \mathrm{sd}(\widehat{\theta}_1^*) \right|.$$

Note the recentering: $\widehat{\theta}_1^* - \widehat{\theta}_1$, not $\widehat{\theta}_1^* - 0$.

Step 4: Bootstrap critical value. The $1 - \alpha$ sample quantile of $\{t_{n,b}^* : b = 1, \ldots, B\}$, denote $\widehat{k}_{1-\alpha,B}$.

Step 5: Reject H_0 if $t_n > \widehat{k}_{1-\alpha,B}$.

[a]The *nominal size* α is the rejection probability the test *advertises* (by convention, $\alpha = 0.05$). The *actual* rejection probability under H_0 may differ; the point of bootstrap is to bring the actual size closer to the nominal size in finite samples than the normal approximation does.

What we just did, intuitively. The plan above implements one idea repeatedly:

- t_n (Step 1) is a number that says "how far is $\widehat{\theta}_1$ from the null value 0, in units of standard error?" If H_0 is true, t_n should be small (somewhere in the bulk of its null distribution). If H_0 is false, t_n blows up.
- Under classical asymptotic theory you would compare t_n to a 1.96 critical value (the 0.975-quantile of $N(0,1)$), because $t_n \xrightarrow{d} N(0,1)$ under H_0. But $N(0,1)$ is just an *approximation* to the true finite-sample null distribution.
- Steps 2–3 build a *better* approximation to that null distribution. Each $t^*_{n,b}$ is the same statistic recomputed on a resample drawn from $\widehat{F}_n$ (the data's own EDF). Recentering at $\widehat{\theta}_1$ in Step 3 forces the bootstrap statistic to behave *as if* the null were true in the bootstrap world. So the cloud of $\{t^*_{n,b}\}$ is a Monte Carlo image of t_n's null distribution.
- Step 4 reads the upper $(1-\alpha)$-quantile of that cloud — this is where the bootstrap-derived critical value comes from, instead of the 1.96.
- Step 5 declares "unlikely under the simulated null" as the rejection rule.

So the workflow is: *measure* how far the original is from the null (t_n), *simulate* how far similar samples are when the null *is* true (t^*_n), *reject* if the original is more extreme than the simulated typical.

Why $1-\alpha$ and not $1-\alpha/2$? Because t_n is already the *absolute* t-statistic (note the $|\cdot|$ in Step 1). For a signed two-sided test you would compare $|\cdot|$ to the $z_{1-\alpha/2}$ critical value of the standard normal. Once you take absolute values up front, the two tails fold into one, and the relevant quantile of the (now non-negative) distribution is the $(1-\alpha)$ one. (If you used the signed $t_n = \sqrt{n}\widehat{\theta}_1/\text{sd}$, you would compare to the $1-\alpha/2$ quantile.)

Why $B = 999$, and where does $(B+1)$ come from? The $(1-\alpha)$-th sample quantile of B bootstrap draws is, by the standard order-statistic convention, the ν-th smallest value with $\nu := (1-\alpha)(B+1)$. We pick B so that ν is an integer — otherwise we would have to interpolate between two order statistics, which is messier and less standard.

Remark (Where does the "+1" come from?).

> It is *not* "B bootstraps plus the original sample." The original is never thrown into this set. The +1 comes from the standard convention for sample-quantile estimation. With B ordered values $t^*_{(1)} < t^*_{(2)} < \cdots < t^*_{(B)}$, you have $B+1$ "gaps" (one before $t^*_{(1)}$, one between each adjacent pair, one after $t^*_{(B)}$). Under a continuous distribution, each gap has equal probability $1/(B+1)$. So the i-th order statistic is treated as the empirical $i/(B+1)$-quantile, and to estimate the $(1-\alpha)$-quantile you pick the $\nu = (1-\alpha)(B+1)$-th order statistic. (This is "Type 6" in R's `quantile()` function.) For $\alpha = 0.05$ and $B = 999$, $\nu = 0.95 \cdot 1000 = 950$, integer — so the critical value is exactly the 950-th

order statistic of $\{t^*_{n,1}, \ldots, t^*_{n,999}\}$. (For the exam, just say "$B = 999$.")

2.5 Symmetric vs Equal-Tailed Confidence Intervals

The same idea gives you a CI. Two flavors.

What this CI is for. The CI is a random set built from the data that is supposed to cover the *true* (unknown) parameter θ_0 with prescribed probability — not the estimator $\widehat{\theta}_n$, which is observed and fixed once the data is in. Pointwise validity is the statement

$$P\left(\theta_0 \in \mathrm{CI}_{1-\alpha}\right) \to 1 - \alpha \quad \text{as } n \to \infty.$$

Symmetric two-sided CI.

$$\mathrm{CI}^{\mathrm{sym}}_{1-\alpha} = \left[\widehat{\theta}_n - n^{-1/2}\widehat{\omega}_n \widehat{k}_{1-\alpha,B},\ \widehat{\theta}_n + n^{-1/2}\widehat{\omega}_n \widehat{k}_{1-\alpha,B}\right],$$

where $\widehat{k}_{1-\alpha,B}$ is the $1-\alpha$ sample quantile of the absolute bootstrap statistic, $\widehat{\omega}_n$ is your standard error from the original sample.

Remark (What is $\widehat{\omega}_n$ exactly?).

$\widehat{\omega}_n = \mathrm{sd}(\widehat{\theta}_n)$ is computed *once*, from the original data — it is the same Eicker–White (or sandwich) standard error that appears in Step 1 of the test recipe. It does *not* change with the bootstrap sample. The bootstrap-sample standard error $\mathrm{sd}(\widehat{\theta}^*_n)$ enters the procedure only *inside* the bootstrap statistic t^*_n that produces the quantile $\widehat{k}_{1-\alpha,B}$.

Equal-tailed two-sided CI.

$$\mathrm{CI}^{\mathrm{eq}}_{1-\alpha} = \left[\widehat{\theta}_n - n^{-1/2}\widehat{\omega}_n \widehat{c}_{1-\alpha/2,B},\ \widehat{\theta}_n - n^{-1/2}\widehat{\omega}_n \widehat{c}_{\alpha/2,B}\right],$$

where $\widehat{c}_{q,B}$ is the q-th sample quantile of the (*not absolute*) bootstrap statistic.

Remark (Why both endpoints *subtract*, and why one quantile ends up adding).

The equal-tailed CI looks weird (both endpoints subtract!) but it is algebraically forced. Start from

$$P\left(c_{\alpha/2} \le \sqrt{n}(\widehat{\theta} - \theta_0)/\widehat{\omega} \le c_{1-\alpha/2}\right) = 1 - \alpha,$$

and solve the inequality for θ_0:

$$\widehat{\theta} - n^{-1/2}\widehat{\omega} \cdot c_{1-\alpha/2} \ \le\ \theta_0 \ \le\ \widehat{\theta} - n^{-1/2}\widehat{\omega} \cdot c_{\alpha/2}.$$

Both endpoints have the same minus sign — the algebra forces it. Where the user's intuition "one should be negative, the other positive" comes back: $c_{\alpha/2}$ is a *lower-tail* quantile of the (signed!) bootstrap statistic, so it is typically negative; $c_{1-\alpha/2}$ is the upper-tail quantile, typically positive. So in practice, $-c_{1-\alpha/2}$ pulls the lower endpoint *down* below $\widehat{\theta}$, and $-c_{\alpha/2}$ pulls the upper endpoint *up* above $\widehat{\theta}$. The signs work out, even though the formulas all say "subtract."

Remark (Sym vs equal-tailed: the two structural differences, side by side).

	Symmetric	Equal-tailed
Bootstrap statistic distribution	$\lvert t_n^* \rvert$ (folded)	t_n^* (signed)
# quantiles needed	1: $\widehat{k}_{1-\alpha,B}$	2: $\widehat{c}_{\alpha/2,B}$, $\widehat{c}_{1-\alpha/2,B}$
Quantile level	$1-\alpha$	$\alpha/2$ and $1-\alpha/2$
Endpoint construction	$\widehat{\theta}_n - \cdot$ and $\widehat{\theta}_n + \cdot$	both $\widehat{\theta}_n - \cdot$

Why the level is α vs $\alpha/2$. The symmetric CI works on the absolute statistic $|t_n^*|$, which has already folded both tails into one — so the relevant level is $1-\alpha$. The equal-tailed CI works on the signed t_n^*, splitting the rejection probability α into two equal pieces of $\alpha/2$ in each tail.

When do they coincide? If the bootstrap distribution of the signed t_n^* is exactly symmetric around 0, then $\widehat{c}_{\alpha/2,B} = -\widehat{c}_{1-\alpha/2,B}$ and the two CIs are numerically identical. In real samples this is almost never exactly true (skewness, finite-sample bias), so the two CIs differ slightly. Higher-order theory (Chapter 6) actually says the symmetric CI is *more accurate* when both are valid: error $O(n^{-3/2})$ vs $O(n^{-1})$ for equal-tailed.

2.6 Bootstrap Consistency: What It Means and How to Prove It

Now the proof material. This is the technical heart of the chapter.

What "consistency" is asking, in plain English. We have an original test statistic T_n (think: the t_n from Section 2.4) whose true null distribution we want to mimic. The bootstrap mimics it by computing T_n^* on each resample and treating the distribution of T_n^* as a substitute for the distribution of T_n ("stand-in" = substitute — same idea, just informal). *Bootstrap consistency* is the statement that this substitute is correct in the limit: the distribution of T_n^* converges to the same limit as the distribution of T_n. If consistency holds, the bootstrap critical value $\widehat{k}_{1-\alpha,B}$ converges to the correct $(1-\alpha)$-quantile of the null limit, and the test has correct asymptotic size. If consistency fails (Section 2.10), the bootstrap critical value targets the wrong distribution and the size advertisement is a lie.

What does "$T_n \xrightarrow{d} T$" even mean for a test statistic? Quick reminder: T_n is a *random number* computed from the random sample $W_1, \ldots, W_n$, and "$T_n \xrightarrow{d} T$" means that as n grows, the distribution (CDF) of T_n approaches the distribution of some limit random variable T. For our bootstrap t-statistic in Section 2.4, the limit T is just $|N(0,1)|$. So "$T_n \xrightarrow{d} T$" is shorthand for "$P(T_n \le x) \to P(T \le x)$ at every x." T_n is a sample-dependent number; T is the limiting law it follows.

Two sources of randomness, kept straight. Once we set up the bootstrap, there are *two* different sources of randomness in the picture, and consistency has to handle both:

1. **Original-sample randomness.** The original data $\mathcal{W} = (W_1, \ldots, W_n)$ is drawn from the unknown distribution F.

2. **Bootstrap-resampling randomness.** Given $\mathcal{W}$, each bootstrap sample $\mathcal{W}^* = (W_1^*, \ldots, W_n^*)$ is drawn from $\widehat{F}_n$ (or another bootstrap rule).

T_n^* depends on both. We cannot simply ask "does T_n^* converge in distribution to T?" without saying which randomness we are averaging over. The standard fix:

- **Condition on $\mathcal{W}$** (treat the original data as fixed). Then the only randomness left in T_n^* is the bootstrap resample. Call this conditional probability P^* (the "starred" measure — it's the bootstrap-world probability, holding the original sample fixed).

- Under P^*, ask whether $T_n^* \xrightarrow{d} T$. This is well-defined: the conditional distribution function $P^*(T_n^* \leq x \,|\, \mathcal{W})$ is now a deterministic function of $\mathcal{W}$ and n, and we ask whether it converges to $P(T \leq x)$.

- **For almost every $\mathcal{W}$.** Different realizations of $\mathcal{W}$ give different conditional distributions for T_n^*. Bootstrap consistency requires the conditional convergence to hold not just on average but for almost every realization (probability one) of $\mathcal{W}$.

That last "for almost every $\mathcal{W}$" is the strong requirement. A weaker bootstrap that only matched the limit *on average* could still misbehave on individual samples.

Definition 2.3: Bootstrap Consistency

Suppose the original test statistic T_n converges in distribution to a limit T under the null hypothesis: $T_n \xrightarrow{d} T$.
The bootstrap procedure for T_n is *consistent* if the conditional distribution of the bootstrap statistic T_n^*, given the original data, converges to the same limit T *with probability one*:

$$T_n^* \xrightarrow{d} T \quad \text{conditional on } \mathcal{W}, \text{ w.p. 1.}$$

2.7 The Four-Step Proof Template (This Is the One You Memorize)

This is the most important page in the chapter. The midterm 2025 Q3(b) was a direct application of this template. The final exam will almost certainly contain a question of the same shape.

Setting. Location model: $W_1, \ldots, W_n$ i.i.d. from F_θ, where $F_\theta(x) = F(x - \theta)$ for some unknown distribution F with mean zero and variance $\sigma^2 \in (0, \infty)$. Estimator $\widehat{\theta}_n = \bar{W}_n$. By CLT,

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{d} N(0, \sigma^2).$$

Question: is the NP bootstrap consistent here?

Remark (Notation: what the stars on $\mathbb{E}^*$, Var^*, P^* mean).

Throughout the proof, a star on a probability operator denotes the *bootstrap measure* — the probability over the resampling step alone, with the original sample $\mathcal{W}$ treated as fixed:

- $P^*(\cdot) := P(\cdot \mid \mathcal{W})$: probability under the bootstrap, given the original data.
- $\mathbb{E}^*[\cdot] := \mathbb{E}[\cdot \mid \mathcal{W}]$: expectation under P^*.
- $\text{Var}^*(\cdot) := \text{Var}(\cdot \mid \mathcal{W})$: variance under P^*.

P (no star) is the original probability over the data-generating process. The two measures live on different sample spaces: P averages over both the original sample and (any) downstream randomness; P^* averages over the bootstrap resampling only, with $\mathcal{W}$ fixed. *They are different probability measures.*

Four-Step Proof Template

[REPRODUCE — memorize this proof]

Step 1: Express the bootstrap statistic.
Under NP bootstrap, $W_i^* \sim \widehat{F}_n$ given the data. Computing the first two bootstrap moments:

$$\mathbb{E}^* W_1^* = \bar{W}_n \quad \text{and} \quad \text{Var}^*(W_1^*) = \widehat{\sigma}^2_{W,n} := n^{-1} \sum_{i=1}^{n} (W_i - \bar{W}_n)^2.$$

Note carefully: W_i in the RHS does not have a star, even though W_1^* on the LHS does. This is not a typo. $\text{Var}^*(W_1^*) = \int (w - \bar{W}_n)^2 \, d\widehat{F}_n(w)$, and $\widehat{F}_n$ puts mass $1/n$ on each of the original sample points $W_1, \ldots, W_n$. So integrating against $\widehat{F}_n$ forces the original (unstarred) W_i to appear in the answer.
The bootstrap statistic, recentered, is

$$\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) = \sqrt{n}(\bar{W}_n^* - \bar{W}_n) = n^{-1/2} \sum_{i=1}^{n} (W_i^* - \mathbb{E}^* W_i^*).$$

Step 2: Condition on a "good" sample path.
Goal of Step 2: legitimize treating $\widehat{\sigma}^2_{W,n}$ as "approximately" the constant σ^2 inside the next CLT step. By SLLN, $\widehat{\sigma}^2_{W,n} \xrightarrow{a.s.} \sigma^2$; that is, there is a measurable set of original-sample realizations $\Omega^{\text{good}} \subset \Omega$ with $P(\Omega^{\text{good}}) = 1$ on which $\widehat{\sigma}^2_{W,n}(\omega) \to \sigma^2$. We fix an arbitrary $\omega \in \Omega^{\text{good}}$ and run Step 3 conditional on that ω. From here on, $\widehat{\sigma}^2_{W,n}$ is a deterministic-converging-to-σ^2 sequence, and only the bootstrap resampling is random.
Step 3: Apply the Lindeberg CLT for triangular arrays.
Conditional on a good path, the bootstrap variables $\{W_i^* - \mathbb{E}^* W_i^*\}_{i=1}^n$ are row-wise i.i.d., mean zero, variance $\widehat{\sigma}^2_{W,n} \to \sigma^2$. The Lindeberg condition can be verified, so by the triangular array CLT,

$$\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \xrightarrow{d} N(0, \sigma^2) \quad \text{conditional on the good path.}$$

Equivalently, $P^*\left(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \le x\right) \to \Phi(x/\sigma)$ a.s.

Step 4: Pass to unconditional convergence by DCT.

Step 3 gave us conditional convergence: for almost every $\mathcal{W}$, $P^*(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \le x \mid \mathcal{W}) \to \Phi(x/\sigma)$. We want unconditional: $P(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \le x) \to \Phi(x/\sigma)$. The bridge is iterated expectations:

$$P\left(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \le x\right) = \mathbb{E}\left[P^*\left(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \le x \mid \mathcal{W}\right)\right].$$

Inside the expectation we have a sequence of conditional probabilities, each lying in $[0, 1]$, converging a.s. to $\Phi(x/\sigma)$. The Dominated Convergence Theorem (with constant dominator 1) lets us swap limit and expectation, giving the unconditional limit $\Phi(x/\sigma)$. This matches the limit of $\sqrt{n}(\widehat{\theta}_n - \theta)$, so the bootstrap is consistent. ■

Remark (What is a "triangular array" and why does it matter in Step 3?).

> In a textbook CLT, $X_1, X_2, \dots$ are i.i.d. from a single fixed distribution. A *triangular array* is a doubly-indexed sequence $\{X_{n,i} : i = 1, \dots, n,\ n = 1, 2, \dots\}$: row 1 has $X_{1,1}$; row 2 has $X_{2,1}, X_{2,2}$; row 3 has $X_{3,1}, X_{3,2}, X_{3,3}$; etc. The distribution of row n can depend on n. In our bootstrap, $W_1^*, \dots, W_n^*$ are drawn from $\widehat{F}_n$, which is computed from the original sample of size n — so the resampling distribution depends on n. Different rows have different distributions. The classical (Lindeberg–Lévy) CLT does not apply because it assumes a fixed distribution. The Lindeberg CLT for triangular arrays does apply, provided the variances stabilize and tails are not too fat (the "Lindeberg condition").

Remark (Exam triage on Step 4).

> The DCT manipulation is genuinely the most technical step, and it adds little to the substantive story (Steps 1–3 carry the content). Under exam pressure, write a single line: "By iterated expectations + DCT, the conditional convergence in Step 3 lifts to unconditional convergence; this matches the null limit of T_n, so the bootstrap is consistent." Patrik gives near-full credit for that.

Remark.

> Mnemonic: **Express, Condition, CLT, DCT**. Four steps. Drill them until you can write them in five minutes.

2.8 Worked Example: Midterm 2025 Q3 (Parametric Bootstrap Test)

This is what an actual exam answer looks like. Read this carefully and try to write it yourself afterward.

Setup. $W_i \overset{i.i.d.}{\sim} N(\theta_1, \theta_2)$. Test $H_0 : \theta_1 = 0$ vs $H_1 : \theta_1 \neq 0$.

(a) Implementation

- Test statistic: $t_n = \left|\sqrt{n}\widehat{\theta}_1/\text{sd}(\widehat{\theta}_1)\right|$ where $\widehat{\theta}_1 = \bar{W}_n$ and $\text{sd}(\widehat{\theta}_1)^2 = \widehat{\theta}_2 = n^{-1}\sum_{i=1}^n (W_i - \bar{W}_n)^2$.

  Remark (Why $\text{sd}(\widehat{\theta}_1)^2 = \widehat{\theta}_2$ here).

  > The notation looks weird at first because $\text{sd}(\widehat{\theta}_1)$ in this context is *not* the finite-sample standard error of $\bar{W}_n$ (which would be $\sqrt{\widehat{\theta}_2/n}$, with a $1/n$ inside). Rather, since the test statistic is written with $\sqrt{n}$ pulled out front, $\text{sd}(\widehat{\theta}_1)$ denotes the *asymptotic* standard deviation of $\sqrt{n}(\widehat{\theta}_1 - \theta_1)$. Three steps:
  >
  > – Model assumes $W_i \sim N(\theta_1, \theta_2)$, so by definition $\text{Var}(W_i) = \theta_2$.
  >
  > – For the sample mean, $\sqrt{n}(\bar{W}_n - \theta_1) \sim N(0, \theta_2)$ (exactly, by linearity of normals; or asymptotically by CLT). So the asymptotic variance of $\sqrt{n}\widehat{\theta}_1$ around θ_1 is θ_2.
  >
  > – The MLE for θ_2 is the sample variance $\widehat{\theta}_2 = n^{-1}\sum_{i=1}^n (W_i - \bar{W}_n)^2$. Plug it in: $\text{sd}(\widehat{\theta}_1)^2 = \widehat{\theta}_2$.
  >
  > The whole construction reduces to: $t_n = |\sqrt{n}\bar{W}_n/\sqrt{\widehat{\theta}_2}| \xrightarrow{d} |N(0,1)|$ under H_0, which is just the studentized z-statistic.

- Parametric bootstrap: draw $W_{i,b}^* \overset{i.i.d.}{\sim} N(\widehat{\theta}_1, \widehat{\theta}_2)$ for $i = 1, \ldots, n$, $b = 1, \ldots, B$.
- Bootstrap statistic (recentered): $t_n^* = \left|\sqrt{n}(\widehat{\theta}_1^* - \widehat{\theta}_1)/\text{sd}(\widehat{\theta}_1^*)\right|$.
- Reject H_0 if $t_n > (1-\alpha)$-quantile of $\{t_{n,b}^*\}_{b=1}^B$.

(b) Consistency

By the four-step template, applied to the parametric bootstrap.

Step 1. $W_{i,b}^* \sim N(\widehat{\theta}_1, \widehat{\theta}_2)$, so $\mathbb{E}^* W_{i,b}^* = \widehat{\theta}_1$ and $\text{Var}^*(W_{i,b}^*) = \widehat{\theta}_2$. Recentered: $t_n^* = \left|\sqrt{n}(\bar{W}_n^* - \widehat{\theta}_1)/\text{sd}(\bar{W}_n^*)\right|$.

Step 2. By SLLN, $\widehat{\theta}_2 \xrightarrow{a.s.} \theta_2 = \sigma^2$. Condition on a good path.

Step 3. The numerator $\sqrt{n}(\bar{W}_n^* - \widehat{\theta}_1) = n^{-1/2}\sum_{i=1}^n (W_{i,b}^* - \widehat{\theta}_1)$ where the summands are i.i.d. mean zero variance $\widehat{\theta}_2 \to \sigma^2$. By triangular CLT, this $\xrightarrow{d} N(0, \sigma^2)$. The denominator $\text{sd}(\bar{W}_n^*) \to \sigma$. So $t_n^* \xrightarrow{d} |N(0,1)|$ conditionally.

Step 4. DCT $\Rightarrow$ unconditional convergence. This matches the null limit of t_n, so the bootstrap is consistent. ■

(c) Power Under Fixed Alternatives

If the true $\theta_1 \neq 0$, what happens to the rejection probability as $n \to \infty$?

Decompose:

$$t_n = \left|\sqrt{n}\frac{\widehat{\theta}_1 - \theta_1}{\text{sd}(\widehat{\theta}_1)} + \sqrt{n}\frac{\theta_1}{\text{sd}(\widehat{\theta}_1)}\right|.$$

The first term $\xrightarrow{d} N(0,1)$ (the standard t for a consistent estimator). The second term $\to \pm\infty$ (because $\sqrt{n}\cdot$ constant). So $t_n \to \infty$.

Meanwhile, the bootstrap statistic $t_n^* = O_p(1)$ by part (b) (its limit distribution is $|N(0,1)|$). So the critical value is bounded.

Conclusion: rejection probability $\to 1$. Power tends to 1.

(d) Local Alternatives

If $\theta_1 = h/\sqrt{n}$ for some fixed h:

$$t_n = \left| \sqrt{n} \frac{\widehat{\theta}_1 - h/\sqrt{n}}{\text{sd}(\widehat{\theta}_1)} + \frac{h}{\text{sd}(\widehat{\theta}_1)} \right| \xrightarrow{d} |N(h/\sigma, 1)|,$$

since the first term $\xrightarrow{d} N(0,1)$ and the second $\to h/\sigma$. The bootstrap critical value $\to z_{1-\alpha/2}$. So

$$\text{Power} \to P(|N(h/\sigma, 1)| > z_{1-\alpha/2}).$$

The Pattern of Q3-Style Questions

Q3 problems tend to have four parts: (a) implementation, (b) consistency, (c) fixed-alternative power, (d) local-alternative power. Memorize this structure. Most of (a)–(b) is mechanical template-following. (c) and (d) require the simple "decompose the test statistic" trick shown above.

2.9 GMM Bootstrap: The Recentering Adjustment

One special case where the recentering needs a tweak.

In over-identified GMM ($k > d$, more moments than parameters), the standard bootstrap is *inconsistent* unless you make an additional adjustment. Specifically, the bootstrap criterion function should be

$$Q_n^*(\theta) := \left\| A_n^* \cdot \frac{1}{n} \sum_{i=1}^{n} \left[g(W_i^*, \theta) - \mathbb{E}^* g(W_i^*, \widehat{\theta}_n) \right] \right\|^2 /2,$$

where the subtracted term is $\mathbb{E}^* g(W_i^*, \widehat{\theta}_n) = n^{-1} \sum_j g(W_j, \widehat{\theta}_n)$ for the NP bootstrap.

Why? In the original sample, the moment condition is $\mathbb{E}(g(W_i, \theta_0)) = 0$. The corresponding statement in the bootstrap world should be $\mathbb{E}^* g(W_i^*, \widehat{\theta}_n) = 0$. But this latter quantity equals $n^{-1} \sum_j g(W_j, \widehat{\theta}_n)$, which is *not* zero in over-identified GMM (because $\widehat{\theta}_n$ minimizes a quadratic form, not the sample moment itself). To restore the analogy, we subtract this nonzero quantity off.

Just-identified case. If $k = d$, the GMM estimator solves $n^{-1} \sum g(W_j, \widehat{\theta}_n) = 0$ exactly, so the recentering term is automatically zero. No adjustment needed.

2.10 When the Bootstrap Fails: Boundary Example

Bootstrap is not a panacea. It can fail at *points of discontinuity* of the limit distribution. The classic example (Andrews 2000, HW8 Q2): a parameter on the boundary of its parameter space.

Setup. $X_i \overset{i.i.d.}{\sim} N(\mu, 1)$, with $\mu \geq 0$ (the parameter space includes 0). The MLE is

$$\widehat{\mu}_n = \max\{\bar{X}_n, 0\}.$$

Distribution of the original statistic.

- If $\mu > 0$: $\bar{X}_n > 0$ with probability $\to 1$, so $\widehat{\mu}_n = \bar{X}_n$, and $\sqrt{n}(\widehat{\mu}_n - \mu) \xrightarrow{d} N(0, 1)$.
- If $\mu = 0$ (boundary): $\sqrt{n}\widehat{\mu}_n = \sqrt{n}\max\{\bar{X}_n, 0\} = \max\{\sqrt{n}\bar{X}_n, 0\} \xrightarrow{d} \max\{Z, 0\}$, where $Z \sim N(0, 1)$.

The limit distribution is *discontinuous in* μ *at* $\mu = 0$: it is a normal for $\mu > 0$ but a half-normal at $\mu = 0$.

Why the NP bootstrap fails at $\mu = 0$. Fix $\mu = 0$. Consider a sample path along which (occasionally, by the law of the iterated logarithm) $\sqrt{n}\bar{X}_n \leq -c$ for some $c > 0$. On such a path, $\widehat{\mu}_n = 0$. The bootstrap version is $\widehat{\mu}_n^* = \max\{\bar{X}_n^*, 0\}$. So

$$\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) = \sqrt{n}\widehat{\mu}_n^* = \sqrt{n}\max\{\bar{X}_n^*, 0\} = \max\{\sqrt{n}(\bar{X}_n^* - \bar{X}_n) + \sqrt{n}\bar{X}_n,\, 0\}.$$

Substituting $\sqrt{n}\bar{X}_n \leq -c$ and using the bootstrap CLT for the centered piece, the limit is $\max\{Z - c, 0\}$, which is *not* $\max\{Z, 0\}$. The bootstrap distribution is biased toward zero.

Subsampling (briefly). Subsampling — using subsamples of size $b \ll n$ instead of n — works in this example. Subsampling only requires a limit distribution to exist; it does not require the bootstrap conditional convergence. So when bootstrap fails, subsampling is the standard fix. (We do not derive subsampling here.)

The One Sentence Answer When Asked "Is the Bootstrap Always Consistent?"

"No. A standard example of bootstrap failure is the MLE of a Gaussian mean restricted to $\mu \geq 0$ at $\mu = 0$ (Andrews 2000). The limit distribution is discontinuous in μ at the boundary, and the bootstrap cannot mimic this discontinuity. Subsampling works in such settings."

2.11 Cheat Sheet

Bootstrap Cheat Sheet

Empirical CDF: $\widehat{F}_n(w) = n^{-1}\sum_{i=1}^n \mathbf{1}(W_i \le w)$. Consistent uniformly (Glivenko–Cantelli).
Three procedures: NP iid, parametric, residual / wild.
Centering: $\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n)$, NOT $\sqrt{n}(\widehat{\theta}_n^* - \theta_0)$.
Implementation: $B = 999$, compute test statistic on each, take $1-\alpha$ sample quantile.
Consistency definition: $T_n^* \xrightarrow{d} T$ conditional on $\mathcal{W}$, w.p. 1, where T is the null limit of T_n.
Four-step proof: Express–Condition–CLT–DCT.
Recentering for over-identified GMM: subtract $n^{-1}\sum g(W_j, \widehat{\theta}_n)$ from the bootstrap moment.
Failure case: parameter on boundary (Andrews 2000). Subsampling works instead.

2.12 Self-Test Problems

Example (Self-Test 1: Default the four-step template).

Without looking at this chapter, write the bootstrap consistency proof for the location model from scratch. Time yourself: under 10 minutes.

Example (Self-Test 2: Why GMM needs recentering).

Explain in three sentences why over-identified GMM needs recentering but just-identified does not.

Solution.

In over-identified GMM ($k > d$), the moment system has more equations than unknowns, and $\widehat{\theta}_n$ is found by minimizing a quadratic form, not by setting the moments to zero. Hence $n^{-1}\sum_{i=1}^n g(W_i, \widehat{\theta}_n) \neq 0$, breaking the analogy between bootstrap and original. Subtracting this term restores the analogy. Just-identified GMM ($k = d$) solves the moments exactly, so no adjustment needed.

Example (Self-Test 3: Power under fixed alternatives).

In midterm 2025 Q3(c), what happens to t_n and t_n^* as $n \to \infty$?

Solution.

$t_n \to \infty$ because the drift term $\sqrt{n}\theta_1/\text{sd}$ dominates. $t_n^* = O_p(1)$ because the bootstrap conditional limit is $|N(0,1)|$ regardless of θ_1. So the rejection probability $\to 1$.

Example (Self-Test 4: Recognizing failure).

Why does the bootstrap fail in the boundary example $\widehat{\mu} = \max(\bar{X}, 0)$ at $\mu = 0$?

Solution.

The limit distribution of $\sqrt{n}\widehat{\mu}_n$ is discontinuous in μ at $\mu = 0$: it is normal for $\mu > 0$ and half-normal at $\mu = 0$. The bootstrap conditional distribution sees only the (data-dependent) "effective μ" equal to $\bar{X}_n$, which can be slightly negative on some sample paths, producing a distribution like $\max\{Z - c, 0\} \neq \max\{Z, 0\}$.

Chapter 3

Identification

The Story (Read This First, No Math)

Suppose someone hands you a sample of measurements of the area of a table and asks: *what is the length of the table?* You stare at the data and quickly realize you cannot answer. Length and width together determine area, but neither one is determined by area alone. No matter how big your sample, the answer is the same: you cannot recover length from area. The data simply does not contain that information.

This is the problem of *identification*. It is a question about the model, not about the estimator. Before you can talk about consistency, asymptotic normality, or efficiency, you have to know whether the parameter you want is even *learnable from the data*. If it is not, no estimator will ever recover it, no matter how clever.

The two big questions in this chapter:

Question 1. What does it formally mean for a parameter to be "learnable from the data"? (Answer: *point identification*.)

Question 2. What if the parameter is not learnable, but we can at least narrow it down to a set? (Answer: *set / partial identification*.)

We then deploy these definitions in five contexts you might see on the exam: linear OLS, linear IV, IV with mixed regressors, the Heckit selection model, and the nonparametric control-function model. The Heckit and control-function results were exactly what HW9 tested, and Patrik has telegraphed that HW questions will recur on the final.

What You Need to Take Away

By the end of this chapter:

1. State the definition of point identification (Section 3.1).
2. Recall that linear IV requires the rank condition $\text{rk}(\mathbb{E}(Zx')) = d_x$ (Section 3.5).
3. Reproduce the Heckit identification proof, including the inverse Mills ratio derivation (Section 3.7). HW9-relevant.
4. Reproduce the control-function rank condition argument (Section 3.8). HW9-relevant.

5. Distinguish set identification from point identification using the table-width example (Section 3.3).

3.1 Point Identification: The Definition

The data we observe is a sample $\mathcal{W} = (W_1, \ldots, W_n)$. The model assumes the data was drawn according to some joint distribution $F_n(w, \theta)$, where θ is a parameter we want to learn. The problem: the same data could be generated by different values of θ.

Definition 3.1: Point Identification

A particular value $\theta_1 \in \Theta$ is *identified* in the parameter space Θ if no other parameter value $\theta_2 \in \Theta$ produces exactly the same distribution of the sample. Formally,

$$\forall \theta_2 \in \Theta, \quad F_n(\cdot, \theta_2) = F_n(\cdot, \theta_1) \implies \theta_2 = \theta_1.$$

The model is *identified* if every $\theta \in \Theta$ is identified.

In words. Two distinct parameter values $\theta_1 \neq \theta_2$ must produce *distinguishable* data distributions. If they cannot, the data cannot tell them apart.

The contrapositive. The same definition rephrased: if two parameters produce the same data distribution, they must be the same parameter. (F_n-equal $\Rightarrow$ θ-equal.)

Remark.

> Critical notational point. $F_n(w, \theta)$ is the *theoretical* joint distribution of W under parameter θ — a deterministic function of w and θ. It is **not** the empirical CDF $\widehat{F}_n(w)$ from Chapter 2. The empirical CDF is a random object built from observed data; the model's $F_n(w, \theta)$ is what the data *should* look like *if* parameter θ were true. Patrik will sometimes use the same letter F_n in both contexts; check by the arguments.

3.1.1 The Length-of-Table Example

The motivating example. You observe $W_i = (\text{area}_i)$ but the parameter you want is $\theta = \text{length}$. The model says length $\times$ width = area, but width is also unknown. So $\theta = \text{length}$ is *not* identified, because for any data distribution of areas, you can pair (length = 1, width = area) or (length = 2, width = area/2), and both fit the data equally well.

3.1.2 The Normal Mean Example

Let $W_i \overset{i.i.d.}{\sim} N(\mu_1 + \mu_2, \sigma^2)$ with $\theta = (\mu_1, \mu_2, \sigma)$.

Case 1: $\Theta_1 = \mathbb{R}^3$. Neither μ_1 nor μ_2 is identified, because $(\mu_1 + 1, \mu_2 - 1, \sigma)$ produces the same $N(\mu_1 + \mu_2, \sigma^2)$ as (μ_1, μ_2, σ). Only $\mu_1 + \mu_2$ is identified. σ is identified *except* at $\sigma = 0$ (where the variance becomes degenerate; some technical issue).

Case 2: $\Theta_2 = \mathbb{R}^2 \times \mathbb{R}_+$**.** Same as Case 1, but $\sigma > 0$ is identified everywhere.

Case 3: $\Theta_3 = \mathbb{R} \times \{0\} \times \mathbb{R}_+$**.** Restricting $\mu_2 = 0$ kills the ambiguity. Now μ_1 alone determines the mean, so μ_1 is identified. Every parameter vector in Θ_3 is identified.

Moral. Identification depends on the parameter space, not just on the model. Imposing restrictions can rescue identification.

3.2 Identified Features: When Parts of θ Are Identified

Often we do not need to identify all of θ. We only need a particular function $r(\theta)$, like an average treatment effect or a marginal effect.

Definition 3.2: Identified Feature

A function r of θ is *identified* at $\theta_1 \in \Theta$ if all parameter values that produce the same data distribution as θ_1 give the same value of r:

$$\forall \theta_2 \in \Theta, \quad F_n(\cdot, \theta_2) = F_n(\cdot, \theta_1) \implies r(\theta_2) = r(\theta_1).$$

r is identified *everywhere* if this holds for all $\theta_1 \in \Theta$.

Two facts that confuse students.

- θ does not need to be identified for $r(\theta)$ to be identified. In the normal mean example with $\Theta_1 = \mathbb{R}^3$, neither μ_1 nor μ_2 is identified, but $r(\theta) = \mu_1 + \mu_2$ is.
- r does not need to be one-to-one. The constant function $r \equiv 5$ is identified everywhere (vacuously: every θ_2 that is observationally equivalent to θ_1 trivially satisfies $r(\theta_2) = 5 = r(\theta_1)$).

3.3 Set / Partial Identification

When point identification fails, we may still narrow down θ to a set.

Definition 3.3: Set / Partial Identification

A parameter θ is *set identified* (or *partially identified*) if the data restricts θ to a strict subset $\Theta_0 \subsetneq \Theta$ but does not pin it down to a single point. The set Θ_0 is called the *identified set*.

3.3.1 Table Width with Functional-Form Knowledge

Suppose you observe area and length and want width.

- If you know the table is rectangular: width = area / length. Point identified.

- If the table could be rectangular or elliptical (and area of an ellipse is $\pi/4 \cdot$ length $\cdot$ width): width lies in the two-point set $\{\text{area/length},\ (4/\pi) \cdot \text{area/length}\}$. Set identified.

3.3.2 Table Width with Bound on Length

Suppose you observe area only, and you know length $\leq \ell_{UP}$ (the room dimensions).

Then width must satisfy two bounds: width $\geq$ area$/\ell_{UP}$ (because length $\leq \ell_{UP}$ and width $\cdot$ length = area), and width $\leq \ell_{UP}$ (because conventionally width $\leq$ length).

So the identified set for width is the interval

$$\Theta_0 = [\text{area}/\ell_{UP},\ \ell_{UP}].$$

3.3.3 Two Kinds of Confidence Interval for Set-Identified Parameters

If θ is point-identified, "95% CI" has one obvious meaning: $P(\theta_0 \in C_n) \to 0.95$. If θ is set-identified, there are two meaningful targets:

- **Cover the true** θ_0: $P(\theta_0 \in C_n) \geq 1 - \alpha$. Here θ_0 is an unknown point inside the identified set Θ_0.
- **Cover the entire identified set** Θ_0: $P(\Theta_0 \subseteq C_n) \geq 1 - \alpha$. The CI must contain every value compatible with the data.

The second is stricter (covering Θ_0 implies covering any single point inside it), so its CIs are wider. Patrik's research is largely on the first kind. We return to this topic in Chapter 8.

3.4 Linear OLS: Identification of β

The simplest case. The model is $y = x'\beta + \varepsilon$ with $y, \varepsilon \in \mathbb{R}$, $x, \beta \in \mathbb{R}^{d_x}$, i.i.d. data.

Theorem 3.4: OLS Identification

β is point identified if and only if

1. $\mathbb{E}(x\varepsilon) = 0$ (*exogeneity*: regressors are uncorrelated with the error), and
2. $\text{rk}(\mathbb{E}(xx')) = d_x$ (*rank condition*: regressors are not perfectly collinear).

Under these conditions, $\beta = (\mathbb{E}(xx'))^{-1}\mathbb{E}(xy)$.

Why these conditions? Multiply both sides of $y = x'\beta + \varepsilon$ by x and take expectations:

$$\mathbb{E}(xy) = \mathbb{E}(xx')\,\beta + \mathbb{E}(x\varepsilon) = \mathbb{E}(xx')\,\beta,$$

using exogeneity to drop the second term. If $\mathbb{E}(xx')$ is invertible (rank condition), we can solve uniquely for β. If $\mathbb{E}(xx')$ is singular, multiple values of β satisfy the same equation, and we cannot pin one down.

The contrapositive. $\text{rk}(\mathbb{E}(xx')) < d_x$ if and only if there exists nonzero c such that $c'x = 0$ a.s. — that is, some linear combination of the regressors is zero. This is what "perfect multicollinearity" means.

3.5 Linear IV: Identification of β When x Is Endogenous

The OLS model fails when x is correlated with ε (an "endogenous" regressor). The standard fix is instrumental variables: a vector of variables $z \in \mathbb{R}^{d_z}$ that is uncorrelated with ε but correlated with x.

Setup.

$$y = x'\beta + \varepsilon, \qquad x = \Pi z + \eta,$$

where $\Pi \in \mathbb{R}^{d_x \times d_z}$ is the matrix of reduced-form coefficients.

Theorem 3.5: Linear IV Identification

β is point identified if

1. $\mathbb{E}(z\varepsilon) = 0$ (instruments are exogenous),
2. $\mathbb{E}(z\eta') = 0$ (instruments are exogenous in the reduced form),
3. $\mathbb{E}(zz')$ is nonsingular,
4. $\text{rk}(\mathbb{E}(zx')) = d_x$ (*rank condition / relevance*).

The order condition. The matrix $\mathbb{E}(zx')$ has shape $d_z \times d_x$. Its rank can be at most $\min(d_z, d_x)$. So the rank condition $\text{rk}(\mathbb{E}(zx')) = d_x$ requires $d_z \geq d_x$. This is the *order condition*: at least as many instruments as endogenous regressors. The order condition is necessary but not sufficient for the rank condition.

Identifying formula. Assuming all four conditions, exogeneity gives $\mathbb{E}(zy) = \mathbb{E}(zx')\,\beta$. This is d_z equations in d_x unknowns. If $d_z = d_x$ (just-identified), we can simply invert. If $d_z > d_x$ (over-identified), the system is overdetermined and we use a GMM-style projection:

$$\beta = \left[(\mathbb{E}(zx'))'\mathbb{E}(zz')^{-1}\mathbb{E}(zx')\right]^{-1}(\mathbb{E}(zx'))'\mathbb{E}(zz')^{-1}\mathbb{E}(zy)\,.$$

Remark (Where the projection formula comes from).

> Read $\mathbb{E}(zy) = \mathbb{E}(zx')\,\beta$ as a (noiseless) linear regression of $\mathbb{E}(zy)$ on the columns of $\mathbb{E}(zx')$ in the inner product $\langle a, b\rangle = a'\mathbb{E}(zz')^{-1}b$. The minimum-distance / generalized-least-squares solution is exactly the formula above. "Just-identified" means the matrix is square and the projection collapses to ordinary inversion.

3.5.1 Failure of the Rank Condition: What It Means

Suppose $\text{rk}(\mathbb{E}(zx')) < d_x$. Then there is some nonzero $c \in \mathbb{R}^{d_x}$ with $\mathbb{E}(zx')\,c = 0$, i.e., $\mathbb{E}(z(x'c)) = 0$. Translation: the linear combination $x'c$ is uncorrelated with all instruments.

The instruments cannot "move" the direction c of x. So the data cannot tell apart β from $\beta + \lambda c$ for any λ, and the linear combination $c'\beta$ is unidentified.

Connection to Weak IV

The rank condition is a binary statement: either $\text{rk}(\mathbb{E}(zx')) = d_x$ or it is not. "Weak IV" is a continuous version of this concern: $\text{rk}(\mathbb{E}(zx')) = d_x$ but barely, with eigenvalues close to zero. Even when the rank condition technically holds, weak instruments wreak havoc on standard inference. Chapter 5 addresses this.

3.6 Mixed Regressors: When Some x Are Exogenous

A common practical setting: some regressors are exogenous (call them z_1, like demographics), and others are endogenous (like x, with separate instruments z_2).

Setup.

$$y = x'\beta + z_1'\gamma + \varepsilon, \qquad x = \Pi z + \eta,$$

where $z = (z_1', z_2')'$. The vector z_1 enters the structural equation; z_2 does not. Hence z_2 are the "excluded instruments."

Trick: stack and rewrite as standard IV. Let $w = (x', z_1')'$ and $\delta = (\beta', \gamma')'$. Then $y = w'\delta + \varepsilon$, and

$$w = \begin{pmatrix} \Pi_1 z_1 + \Pi_2 z_2 + \eta \\ z_1 \end{pmatrix} = \begin{pmatrix} \Pi_1 & \Pi_2 \\ I_{d_{z_1}} & 0 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} + \begin{pmatrix} \eta \\ 0 \end{pmatrix}.$$

Theorem 3.6: Mixed-IV Identification (HW9 Q2(a))

$\delta = (\beta', \gamma')'$ is point identified iff $\text{rk}(\Pi_2) = d_x$. That is: the excluded instruments must collectively shift each component of x.

Proof of Theorem 3.6

[REPRODUCE — memorize this proof]

Compute $\mathbb{E}(zw')$ using $\mathbb{E}(z\eta') = 0$:

$$\mathbb{E}(zw') = \mathbb{E}(zz') \begin{pmatrix} \Pi_1' & I_{d_{z_1}} \\ \Pi_2' & 0 \end{pmatrix}.$$

The right-hand side: $\mathbb{E}(zz')$ has full rank $d_{z_1} + d_{z_2}$ by assumption, and the right factor has full column rank $d_{z_1} + d_x$ if and only if $\text{rk}(\Pi_2) = d_x$. (The block structure means the columns are linearly independent precisely when Π_2's columns are.) So $\mathbb{E}(zw')$ has full column rank under the assumption.

Now apply the standard IV identifying argument: $\mathbb{E}(zy) = \mathbb{E}(zw')\,\delta$, and premultiply by $\mathbb{E}(wz')\,\mathbb{E}(zz')^{-1}$:

$$\delta = \left[\mathbb{E}(wz')\,\mathbb{E}(zz')^{-1}\mathbb{E}(zw')\right]^{-1}\mathbb{E}(wz')\,\mathbb{E}(zz')^{-1}\mathbb{E}(zy)\,.$$

Each piece is a function of observable distributions, so δ is identified. ■

Why the rank condition $\mathrm{rk}(\Pi_2) = d_x$? Intuitively: if $\Pi_2 = 0$, then $x = \Pi_1 z_1 + \eta$ depends only on z_1. But z_1 already enters the structural equation as an exogenous regressor, so it cannot serve as an instrument. With $\Pi_2 = 0$, there are no available instruments for x, and β is not identified.

3.7 Heckit Selection Model

This was HW9 Q2 (last page) and is highly likely to recur on the final.

Setup. You are studying wages, but you only observe wages for people who chose to work. People who did not work have "missing" wages. The decision to work is endogenous: people with high latent wages may be more likely to work.

Model.

$$\begin{aligned} y^* &= x'\theta + \varepsilon \quad \text{(latent wage)},\\ y &= d \cdot y^* \quad \text{(observed wage)},\\ d &= \mathbf{1}(x'\pi_1 + z'\pi_2 \geq -\eta) \quad \text{(participation)}, \end{aligned}$$

with $(\varepsilon, \eta) \perp (x, z)$ and $(\varepsilon, \eta) \sim N\left(0, \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\eta} \\ \sigma_{\varepsilon\eta} & 1 \end{pmatrix}\right)$. Variance of η is normalized to 1.

Remark (Why $\mathrm{Var}(\eta) = 1$ is a free normalization).

> The participation equation is $d = \mathbf{1}(x'\pi_1 + z'\pi_2 \geq -\eta)$. Multiplying both sides by any positive constant c gives the same d but rescaled coefficients $(c\pi_1, c\pi_2)$ and rescaled η. Without a scale restriction, the data cannot tell $(\pi, \mathrm{Var}(\eta))$ apart from $(c\pi, c^2\,\mathrm{Var}(\eta))$. Fixing $\mathrm{Var}(\eta) = 1$ pins the scale so π becomes identifiable. (Compare with multinomial probit: same trick.)

Object of interest. θ, the coefficient vector in the wage equation. We can only run regressions on the selected (working) subsample, where $d = 1$.

Theorem 3.7: Heckit Identifying Equation

$$\mathbb{E}(y \mid x, z, d = 1) = x'\theta + \sigma_{\varepsilon\eta}\,\lambda(x'\pi_1 + z'\pi_2),$$

where $\lambda(s) := \phi(s)/\Phi(s)$ is the *inverse Mills ratio*.

Five-Step Proof (HW9 Q2)

[REPRODUCE — memorize this proof]

Step 1: Decompose the conditional expectation.

$$\mathbb{E}(y \mid x, z, d = 1) = x'\theta + \mathbb{E}(\varepsilon \mid x, z,\, x'\pi_1 + z'\pi_2 \geq -\eta)\,.$$

We need to evaluate the conditional expectation of ε given the participation constraint.

Step 2: Compute $\mathbb{E}(\varepsilon \mid \eta = \bar{\eta})$. By bivariate normality of (ε, η) with $\mathbb{E}(\eta) = 0$, $\mathrm{Var}(\eta) = 1$, and $\mathrm{Cov}(\varepsilon, \eta) = \sigma_{\varepsilon\eta}$:

$$\mathbb{E}(\varepsilon \mid \eta = \bar{\eta}) = \sigma_{\varepsilon\eta}\,\bar{\eta}.$$

This is the standard formula for conditional means in joint normal: $\mathbb{E}(X_1 \mid X_2) = \mu_1 + \mathrm{Cov}(X_1, X_2) \,/\, \mathrm{Var}(X_2) \cdot (X_2 - \mu_2)$.

Step 3: Integrate over $\bar{\eta} \geq -c$ where $c = x'\pi_1 + z'\pi_2$.

$$\mathbb{E}(\varepsilon \mid \eta \geq -c) = \frac{1}{\mathrm{P}(\eta \geq -c)} \int_{-c}^{\infty} \mathbb{E}(\varepsilon \mid \eta = \bar{\eta})\ \phi(\bar{\eta})\, d\bar{\eta} = \frac{\sigma_{\varepsilon\eta}}{\Phi(c)} \int_{-c}^{\infty} \bar{\eta}\,\phi(\bar{\eta})\, d\bar{\eta},$$

using $P(\eta \geq -c) = P(-\eta \leq c) = \Phi(c)$ since $-\eta \sim N(0,1)$.

Step 4: Compute the integral. Use the identity $\bar{\eta}\phi(\bar{\eta}) = -\phi'(\bar{\eta})$:

$$\int_{-c}^{\infty} \bar{\eta}\,\phi(\bar{\eta})\, d\bar{\eta} = -[\phi(\bar{\eta})]_{-c}^{\infty} = \phi(-c) = \phi(c).$$

Hence

$$\mathbb{E}(\varepsilon \mid \eta \geq -c) = \sigma_{\varepsilon\eta}\,\phi(c)/\Phi(c) = \sigma_{\varepsilon\eta}\,\lambda(c).$$

Step 5: Combine.

$$\mathbb{E}(y \mid x, z, d = 1) = x'\theta + \sigma_{\varepsilon\eta}\,\lambda(x'\pi_1 + z'\pi_2). \qquad \blacksquare$$

3.7.1 Identification of $\pi_1, \pi_2, \theta, \sigma_{\varepsilon\eta}$

Identifying π_1, π_2 from probit.

$$\mathbb{E}(d \mid x, z) = \mathrm{P}(\eta \geq -(x'\pi_1 + z'\pi_2)) = \Phi(x'\pi_1 + z'\pi_2),$$

since $-\eta \sim N(0,1)$ and is independent of (x, z). Apply Φ^{-1}:

$$\Phi^{-1}(\mathbb{E}(d \mid x, z)) = x'\pi_1 + z'\pi_2.$$

This is a standard linear regression in $\Phi^{-1}(\mathbb{E}(d \mid x, z))$ on (x, z). So π_1, π_2 are identified (assuming $\mathbb{E}((x, z)(x, z)')$ is full rank).

Identifying $\sigma_{\varepsilon\eta}$. Take the derivative of the Heckit equation with respect to z:

$$\frac{\partial \mathbb{E}(y \mid x, z, d = 1)}{\partial z} = \sigma_{\varepsilon\eta}\, \lambda'(x'\pi_1 + z'\pi_2)\, \pi_2.$$

Postmultiply by π_2 and divide by $\|\pi_2\|^2$ (assumed nonzero, the *exclusion restriction*):

$$\sigma_{\varepsilon\eta}\, \lambda'(x'\pi_1 + z'\pi_2) = \frac{\partial \mathbb{E}(y \mid x, z, d = 1)}{\partial z} \pi_2 / \|\pi_2\|^2 .$$

The right-hand side is identified (everything observable). Since $\lambda'(\cdot)$ is a known function, $\sigma_{\varepsilon\eta}$ is identified.

Identifying θ. Differentiate with respect to x:

$$\frac{\partial \mathbb{E}(y \mid x, z, d = 1)}{\partial x} = \theta + \sigma_{\varepsilon\eta}\, \lambda'(x'\pi_1 + z'\pi_2)\, \pi_1,$$

so

$$\theta = \frac{\partial \mathbb{E}(y \mid x, z, d = 1)}{\partial x} - \sigma_{\varepsilon\eta}\, \lambda'(\cdot)\, \pi_1,$$

where every term on the right is identified. Hence θ is identified.

Why the Exclusion Restriction $\pi_2 \neq 0$ Matters

Without an excluded instrument z, the participation index $x'\pi_1 + z'\pi_2$ would be a function of x alone. Then $\lambda(\cdot)$ would also be a function of x alone, and we could not separate $x'\theta$ from $\sigma_{\varepsilon\eta}\, \lambda(x'\pi_1)$ inside $\mathbb{E}(y \mid x, z, d = 1)$. The selection bias correction would be confounded with the main effect.

3.8 Control Function Method (Newey–Powell–Vella 1999)

This was HW9 Q2(b) and Q3, and is a standard tool for nonparametric IV.

Remark (Why we are bothering to introduce a second strategy at all).

Standard linear IV (Section 3.5) handles endogeneity by *instrumenting*: project x onto z, run the structural regression on the projection. That trick crucially relies on *linearity* — the projected x enters the structural equation through a single coefficient β. If the structural relationship is nonlinear or nonparametric (like $y = m(x, z_1) + \varepsilon$ where m is an unknown function), there is no obvious place to "substitute the projection in," because the conditional mean of m is not m at the projected x. Control function is the standard nonparametric alternative: instead of replacing x, you *add* a regressor (the reduced-form residual η) that absorbs the endogeneity. After conditioning on η, the residual variation in x is exogenous, and you can identify m by ordinary nonparametric regression of y on (x, z_1, η) then differencing out $\xi(\eta) := \mathbb{E}(\varepsilon \mid \eta)$. So: *IV is the linear/parametric tool; CF is the nonparametric/nonlinear tool.* They give the same answer in linear models when both apply, but CF extends to settings where IV alone would be stuck.

Setup. Endogenous regressor, additively separable model:

$$y = m(x, z_1) + \varepsilon, \qquad x = \pi(z) + \eta,$$

with $z = (z_1', z_2')'$. The first equation is the structural model with unknown function m; the second is the reduced form with unknown function π and noise η. We want to identify m.

Exogeneity assumption (weak): $\mathbb{E}(\eta \mid z) = 0$. This implies $\pi(z) = \mathbb{E}(x \mid z)$, so π is identified.

Control function assumption:

$$\boxed{\mathbb{E}(\varepsilon \mid z, \eta) = \mathbb{E}(\varepsilon \mid \eta)\,.}$$

After conditioning on η (the reduced-form residual), z no longer affects ε's conditional mean.

Remark (Why we want this assumption (the whole point of CF)).

> The reason x is endogenous in the structural equation is precisely that some piece of ε correlates with x — and through $x = \pi(z) + \eta$, that piece can "leak in" via either z or η. The reduced-form residual η packages up everything in x that is *not* explained by z. The CF assumption says: once you have observed η, knowing z on top of that gives no extra information about ε. In other words, *η is a sufficient statistic for the endogenous channel*. If true, including η as an additional regressor (as a "control function") in $\mathbb{E}(y \mid z, \eta)$ purges the endogeneity — z becomes effectively exogenous after this conditioning, because there is no remaining z-channel into ε.
>
> **When does this hold?** Concretely: any time $(\varepsilon, \eta) \perp z$ (we prove this in the next subsection). It does *not* hold whenever z has a separate direct effect on ε that doesn't go through x. Heckit (Section 3.7) is exactly an instance of CF with η being the latent participation shock.

Why the control function assumption works. Let $\xi(\eta) := \mathbb{E}(\varepsilon \mid \eta)$. Then

$$\mathbb{E}(y \mid z, \eta) = m(\pi(z) + \eta,\, z_1) + \xi(\eta) = m(x, z_1) + \xi(\eta),$$

because $x = \pi(z) + \eta$ is a deterministic function of (z, η).

The left-hand side is identified from the joint distribution of observables. The right-hand side has two unknowns: $m(x, z_1)$ and $\xi(\eta)$. Identification follows by separately recovering each.

3.8.1 Sufficient Condition: $(\varepsilon, \eta) \perp z$ Implies the CF Assumption (HW9 Q3(a))

Proof

[STRUCTURE — know the steps, fudge details]

Express conditional expectations using densities (assuming densities exist):

$$\mathbb{E}(\varepsilon \mid z, \eta) = \int_\Omega \varepsilon \, f_{\varepsilon \mid z,\eta}(\varepsilon \mid z, \eta) \, d\mu = \int_\Omega \varepsilon \cdot \frac{f_{\varepsilon,z,\eta}(\varepsilon, z, \eta)}{f_{z,\eta}(z, \eta)} \, d\mu.$$

Independence $(\varepsilon, \eta) \perp z$ factors: $f_{\varepsilon,z,\eta} = f_{\varepsilon,\eta} \cdot f_z$ and $f_{z,\eta} = f_z \cdot f_\eta$. Substitute:

$$\mathbb{E}(\varepsilon \mid z, \eta) = \int_\Omega \varepsilon \cdot \frac{f_{\varepsilon,\eta}(\varepsilon, \eta)}{f_\eta(\eta)} \, d\mu = \mathbb{E}(\varepsilon \mid \eta). \qquad \blacksquare$$

3.8.2 Counter-Example: CF Assumption Does Not Imply IV Exogeneity (HW9 Q3(b))

Subtle Point Patrik Likes to Test

The control function assumption $\mathbb{E}(\varepsilon \mid z, \eta) = \mathbb{E}(\varepsilon \mid \eta)$ does **not** imply the IV exogeneity $\mathbb{E}(\varepsilon \mid z) = 0$, even when $\mathbb{E}(\varepsilon) = 0$. This is HW9 Q3(b).

Counter-example (extreme: $z = \eta$). Let $(\varepsilon, z) \in \{-1, 0, 1\} \times \{-1, +1\}$ with joint probabilities:

(ε, z)	$z = -1$	$z = +1$	marginal of ε
$\varepsilon = -1$	1/12	1/4	1/3
$\varepsilon = 0$	1/4	1/12	1/3
$\varepsilon = +1$	1/4	1/12	1/3
marginal of z	7/12	5/12	1

Verification.

- Probabilities sum to 1: $1/12 + 1/4 + 1/4 + 1/12 + 1/4 + 1/12 = 1$. **(yes)**
- $\mathbb{E}(\varepsilon) = (-1)(1/3) + 0(1/3) + (1)(1/3) = 0$. **(yes)**
- Since $z = \eta$, $\mathbb{E}(\varepsilon \mid z, \eta) = \mathbb{E}(\varepsilon \mid \eta)$ trivially. **(yes, CF assumption holds)**
- $\mathbb{E}(\varepsilon \mid z = +1) = [(-1)(1/4) + 0(1/12) + (1)(1/12)] / (5/12) = (-2/12)/(5/12) = -2/5$. **(NO, IV exogeneity fails)**

Takeaway. CF and IV are *distinct* identification strategies. They impose different exogeneity conditions and are not implied by one another.

3.8.3 Identifying Derivatives: Rank Condition (HW9 Q2(b))

To identify $\frac{\partial m}{\partial x}$, we need a rank condition on the reduced form.

Theorem 3.8: NPV Rank Condition

$\frac{\partial m(x,z_1)}{\partial x}$ is identified at $(x, z_1) = (\pi(z) + \eta, z_1)$ if

$$\operatorname{rk}\left(\frac{\partial \pi(z)}{\partial z_2'}\right) = d_x.$$

The argument is the same chain-rule trick as for Heckit:

$$\frac{\partial \mathbb{E}(y \mid z, \eta)}{\partial z_2} = \left(\frac{\partial \pi(z)}{\partial z_2'}\right)' \frac{\partial m(x, z_1)}{\partial x},$$

since $\frac{\partial \xi(\eta)}{\partial z_2} = 0$. If $\frac{\partial \pi}{\partial z_2'}$ has full row rank d_x, we can invert and solve:

$$\frac{\partial m(x, z_1)}{\partial x} = \left[\left(\frac{\partial \pi}{\partial z_2'}\right)\left(\frac{\partial \pi}{\partial z_2'}\right)'\right]^{-1}\left(\frac{\partial \pi}{\partial z_2'}\right)\frac{\partial \mathbb{E}(y \mid z, \eta)}{\partial z_2}.$$

This expresses $\frac{\partial m}{\partial x}$ in terms of identified quantities.

3.9 Cheat Sheet

Identification Cheat Sheet

Point ID definition: $F_n(\cdot, \theta_2) = F_n(\cdot, \theta_1) \implies \theta_2 = \theta_1$.
Identified feature: r identified at θ_1 if all observationally equivalent θ_2 give the same r. F need not be identified; r need not be one-to-one.
Set ID: identified set $\Theta_0 \subsetneq \Theta$. Two CI flavors: cover θ_0 vs cover Θ_0.
Linear OLS: identified iff $\mathbb{E}(x\varepsilon) = 0$ and $\operatorname{rk}(\mathbb{E}(xx')) = d_x$.
Linear IV: identified iff $\mathbb{E}(z\varepsilon) = 0$, $\mathbb{E}(z\eta') = 0$, $\mathbb{E}(zz')$ nonsingular, $\operatorname{rk}(\mathbb{E}(zx')) = d_x$. Order condition $d_z \geq d_x$ is necessary.
Mixed IV: $\delta = (\beta, \gamma)$ identified iff $\operatorname{rk}(\Pi_2) = d_x$ (excluded instruments matter).
Heckit: $\mathbb{E}(y \mid x, z, d = 1) = x'\theta + \sigma_{\varepsilon\eta}\,\lambda(x'\pi_1 + z'\pi_2)$. Inverse Mills $\lambda(s) = \phi(s)/\Phi(s)$. Exclusion restriction: $\pi_2 \neq 0$.
NPV Control Function: $\mathbb{E}(\varepsilon \mid z, \eta) = \mathbb{E}(\varepsilon \mid \eta)$. Implied by $(\varepsilon, \eta) \perp z$ but does NOT imply $\mathbb{E}(\varepsilon \mid z) = 0$.
NPV rank condition: $\operatorname{rk}(\frac{\partial \pi}{\partial z_2'}) = d_x$.

3.10 Self-Test Problems

Example (Self-Test 1: Heckit derivation).

> Reproduce the inverse Mills ratio derivation (Section 3.7) without looking. Time yourself: under 12 minutes.

Example (Self-Test 2: IV rank condition).

Suppose $d_x = 2$, $d_z = 2$, and $\Pi = \begin{pmatrix} 1 & 0 \\ 2 & 0 \end{pmatrix}$. Is β identified?

Solution.

$\operatorname{rk}(\Pi) = 1 < 2 = d_x$. The rank condition fails. So β is not identified. The order condition $d_z = 2 \geq 2 = d_x$ holds but is not sufficient.

Example (Self-Test 3: When the CF assumption fails to give IV exogeneity).

Construct a small distribution where $\mathbb{E}(\varepsilon \mid z, \eta) = \mathbb{E}(\varepsilon \mid \eta)$ holds but $\mathbb{E}(\varepsilon \mid z) = 0$ fails.

Solution.

The HW9 Q3(b) example: take $z = \eta$ and the joint distribution of (ε, z) from Section 3.8. CF holds trivially since $z = \eta$, but $\mathbb{E}(\varepsilon \mid z = +1) = -2/5 \neq 0$.

Example (Self-Test 4: Identify a function but not the parameter).

Give an example where θ is not identified but a function $r(\theta)$ is.

Solution.

$W_i \overset{i.i.d.}{\sim} N(\mu_1 + \mu_2, \sigma^2)$ with $\Theta = \mathbb{R}^3$. Neither μ_1 nor μ_2 is identified, but $r(\theta) = \mu_1 + \mu_2$ is.

Chapter 4

Hypothesis Tests: Wald, LM, QLR

Why This Chapter Matters

Patrik tested LM under local alternatives on midterm 2025 Q2. Q2 of the final is most likely to be an analogous question on one of the three test statistics. **All three converge to the same χ^2_r distribution under H_0 and to the same noncentral χ^2 under local alternatives**, so once you know one you essentially know all three. Target: 6/10.

4.1 The Setup

We test

$$H_0 : h(\theta_0) = 0 \quad \text{vs.} \quad H_1 : h(\theta_0) \neq 0,$$

where $h : \mathbb{R}^d \to \mathbb{R}^r$ is a function defining r restrictions.

Remark (Why writing the null as "$h(\theta_0) = 0$" is the right canonical form).

Almost any hypothesis you might want to test can be cast in this form, which is why the chapter focuses on this single template instead of treating each test type separately. Examples:

- Test that a single coefficient equals a specific value: $H_0 : \theta_2 = 5$ becomes $h(\theta) = \theta_2 - 5$, $r = 1$.
- Test that several coefficients equal zero: $H_0 : \theta_2 = \theta_3 = 0$ becomes $h(\theta) = (\theta_2, \theta_3)'$, $r = 2$.
- Test a nonlinear restriction: $H_0 : \theta_1\theta_2 = 1$ becomes $h(\theta) = \theta_1\theta_2 - 1$, $r = 1$.

The Jacobian $H = \frac{\partial h}{\partial \theta'}(\theta_0)$ then captures all the geometry: full row rank r means the r restrictions are non-redundant. By delta method, $\sqrt{n}h(\widehat{\theta}_n) \xrightarrow{d} N(0, HB_0{}^{-1}\Omega_0 B_0{}^{-1}H')$ regardless of what h looks like, so the same test theory covers all three examples above.

Assumption 4.1: R (Rank Conditions)

(i) $h(\theta)$ is continuously differentiable on a neighborhood of θ_0.
(ii) $H = \frac{\partial h}{\partial \theta'}(\theta_0)$ has full row rank r ($\leq d$).
(iii) Ω_0 is positive definite.

Two estimators play roles:

- Unrestricted estimator $\widehat{\theta}_n$: minimizes $Q_n(\theta)$ over Θ.
- Restricted estimator $\widetilde{\theta}_n$: minimizes Q_n subject to $h(\theta) = 0$.

Remark (Why three tests for the same job).

Wald, LM, and QLR all test the *same* hypothesis and have the *same* χ^2_r null distribution. They differ only in which estimator they require:

- **Wald** uses $\widehat{\theta}_n$ only — "how far is $h(\widehat{\theta}_n)$ from 0?"
- **LM** uses $\widetilde{\theta}_n$ only — "how far is $\frac{\partial Q_n(\widetilde{\theta}_n)}{\partial \theta}$ from 0?" (At the true unrestricted FOC, the score should be zero; the LM statistic measures this.)
- **QLR** uses both — "how much worse is $Q_n(\widetilde{\theta}_n)$ than $Q_n(\widehat{\theta}_n)$?" (The likelihood-ratio idea: imposing a true restriction shouldn't hurt fit much.)

The test you pick depends on which estimator is computationally cheaper to obtain. Patrik usually tests *LM* because the restricted estimator is often a simple linear model.

4.2 The Wald Statistic

Definition 4.2: Wald Statistic

$$\mathcal{W}_n = n \cdot h(\widehat{\theta}_n)' \left[\widehat{H}_n \widehat{B}_n^{-1} \widehat{\Omega}_n \widehat{B}_n^{-1} \widehat{H}_n'\right]^{-1} h(\widehat{\theta}_n),$$

where $\widehat{H}_n = \frac{\partial h}{\partial \theta'}(\widehat{\theta}_n)$.

Theorem 4.3: Wald Asymptotic Null Distribution

Under EE2, CF, R, and COV (consistent estimators of B_0, Ω_0), $\mathcal{W}_n \xrightarrow{d} \chi^2_r$ under H_0.

Intuition. Mean-value expansion of h around θ_0:

$$\sqrt{n}h(\widehat{\theta}_n) = \sqrt{n}h(\theta_0) + H\sqrt{n}(\widehat{\theta}_n - \theta_0) + o_p(1) = H\sqrt{n}(\widehat{\theta}_n - \theta_0) + o_p(1) \xrightarrow{d} N(0, H{B_0}^{-1}\Omega_0{B_0}^{-1}H').$$

The Wald statistic is a quadratic form in this normal limit with the inverse of its covariance matrix as weight, hence χ^2_r.

4.3 The LM (Lagrange Multiplier) Statistic

Definition 4.4: LM Statistic

$$\text{LM}_n = n \frac{\partial Q_n(\widetilde{\theta}_n)}{\partial \theta'} \widetilde{B}_n^{-1} \widetilde{H}_n' \left[\widetilde{H}_n \widetilde{B}_n^{-1} \widetilde{\Omega}_n \widetilde{B}_n^{-1} \widetilde{H}_n' \right]^{-1} \widetilde{H}_n \widetilde{B}_n^{-1} \frac{\partial Q_n(\widetilde{\theta}_n)}{\partial \theta},$$

where $\widetilde{H}_n = \frac{\partial h}{\partial \theta'}(\widetilde{\theta}_n)$ and tilde versions are evaluated at the restricted estimator.

Intuition. If H_0 is true, the restricted and unrestricted estimators are close, so $\frac{\partial Q_n(\widetilde{\theta}_n)}{\partial \theta}$ should be close to zero (since the FOC at $\widehat{\theta}_n$ is exactly zero up to $o_p(\sqrt{n}^{-1})$). The LM statistic measures how far $\frac{\partial Q_n(\widetilde{\theta}_n)}{\partial \theta}$ is from zero, normalized by an estimate of its asymptotic variance.

Computational Advantage of LM

LM only requires the restricted estimator $\widetilde{\theta}_n$, which is often simpler to compute (e.g., a linear regression instead of a nonlinear one). On the exam, if the restricted model is much simpler than the unrestricted, LM is the test of choice computationally.

4.4 The QLR (Quasi-Likelihood Ratio) Statistic

Definition 4.5: QLR Statistic

Under Assumption QLR ($\Omega_0 = cB_0$ for some scalar $c \neq 0$, with consistent $\widehat{c}_n \xrightarrow{p} c$):

$$\text{QLR}_n = 2n \left(Q_n(\widetilde{\theta}_n) - Q_n(\widehat{\theta}_n) \right) / \widehat{c}_n.$$

Remark.

When does Assumption QLR hold?

- ML correctly specified: $\Omega_0 = B_0$ (information matrix equality), $c = \widehat{c}_n = 1$. QLR is the standard likelihood ratio.
- LS with conditional homoskedasticity and correct specification: $c = \sigma^2$, $\widehat{c}_n = n^{-1} \sum_{i=1}^n \widehat{U}_i^2$.
- GMM/MD/TS with optimal weight matrix $A'A = {V_0}^{-1}$: $c = \widehat{c}_n = 1$.

Remark (Why QLR needs $\Omega_0 = cB_0$ in the first place).

Assumption QLR is the *information-matrix equality* in disguise. The QLR statistic uses only the raw difference $Q_n(\widetilde{\theta}_n) - Q_n(\widehat{\theta}_n)$, with no explicit estimate of the asymptotic variance. For the limit to be a clean χ^2_r, the difference must scale by a single scalar c that links the score variance Ω_0 to the Hessian curvature B_0. Each "when this holds" bullet above is exactly a setting where the link is automatic:

- ML correctly specified $\Rightarrow \Omega_0 = B_0$ ($c = 1$): this is the textbook information equality.

- LS homoskedastic $\Rightarrow \Omega_0 = \sigma^2 \mathbb{E}(XX') = \sigma^2 B_0$: same shape, scaled by error variance.
- Optimal-weight GMM $\Rightarrow$ pre- and post-multiplying by $V_0^{-1/2}$ aligns the two matrices, so again $c = 1$.

Outside these special cases (heteroskedasticity, misspecified ML, suboptimal GMM weighting), Ω_0 and B_0 are not proportional — the scalar c is no longer well-defined — and the QLR χ^2 approximation breaks. **Diagnostic for the exam:** "if information matrix equality is wrong, QLR is wrong; use Wald or LM, both of which estimate Ω_0 and B_0 separately and avoid the proportionality assumption."

Theorem 4.6: Joint Asymptotic Distribution Theorem (Andrews 14.1)

Under appropriate assumptions:

(a) $\mathcal{W}_n \xrightarrow{d} \chi^2_r$ under H_0.

(b) $\mathrm{LM}_n \xrightarrow{d} \chi^2_r$ under H_0.

(c) $\mathrm{QLR}_n \xrightarrow{d} \chi^2_r$ under H_0 (when Assumption QLR holds).

4.5 Local Power: All Three Have the Same

Remark (Why we even bother with "local" alternatives).

Under any *fixed* alternative $\theta_n \equiv \theta_1 \neq \theta_0$, every consistent test rejects with probability $\to 1$. So fixed-alternative power can't rank tests — everything goes to 1. The standard fix (Pitman) is to consider alternatives that drift toward the null at the same rate the estimator concentrates: $\theta_n = \theta_0 + \lambda/\sqrt{n}$. Under this drift, the test statistic stays $O_p(1)$ and converges to a *noncentral* χ^2_r, with noncentrality $\delta(\lambda)$. Different tests can have different δ's for the same λ — that gap is what we use to compare them. The headline result below is that Wald, LM, and QLR have *identical* δ, so first-order local power cannot distinguish them; choice between them is computational, not inferential.

Consider local alternatives $\theta_n = \theta_0 + \lambda/\sqrt{n}$.

Theorem 4.7: Local Power

Under θ_n, all three statistics converge to a noncentral χ^2 with the same noncentrality parameter:

$$\delta = \lambda' H' \left[H {B_0}^{-1} \Omega_0 {B_0}^{-1} H' \right]^{-1} H \lambda.$$

Hence the three tests have identical first-order local power. Choice between them rests on computational considerations and finite-sample behavior (folklore: Wald over-rejects more often than LM/QLR).

Local Power Derivation (midterm 2025 Q2 Style)

[REPRODUCE — memorize this proof]

Step 1. Mean-value expansion of $\sqrt{n}\frac{\partial Q_n(\widetilde{\theta}_n)}{\partial\theta}$ about θ_n:

$$\sqrt{n}\frac{\partial Q_n}{\partial\theta}(\widetilde{\theta}_n) = \sqrt{n}\frac{\partial Q_n}{\partial\theta}(\theta_n) + \frac{\partial^2 Q_n}{\partial\theta\partial\theta'}(\theta_n^*)\sqrt{n}(\widetilde{\theta}_n - \theta_n).$$

Step 2. Use the constraint $h(\widetilde{\theta}_n) = 0$ and Taylor expand h at θ_n:

$$0 = \sqrt{n}h(\widetilde{\theta}_n) = \sqrt{n}h(\theta_n) + H\sqrt{n}(\widetilde{\theta}_n - \theta_n) + o_p(1).$$

Since $\sqrt{n}h(\theta_n) = \sqrt{n}(h(\theta_n) - h(\theta_0)) = H\lambda + o(1)$, this gives $\sqrt{n}(\widetilde{\theta}_n - \theta_n) \to_d$ something, and the algebra can be solved.

Step 3. Combining and using CLT $\sqrt{n}\frac{\partial Q_n(\theta_n)}{\partial\theta} \xrightarrow{d} N(0, \Omega_0)$:

$$\widetilde{H}_n\widetilde{B}_n^{-1}\sqrt{n}\frac{\partial Q_n(\widetilde{\theta}_n)}{\partial\theta} \xrightarrow{d} Z_0 + H\lambda, \quad Z_0 \sim N(0, HB_0{}^{-1}\Omega_0 B_0{}^{-1}H').$$

Step 4. The LM statistic is a quadratic form in this with weight $(HB_0{}^{-1}\Omega_0 B_0{}^{-1}H')^{-1}$, so

$$\text{LM}_n \xrightarrow{d} \chi^2_r(\delta), \quad \delta = \lambda' H'(HB_0{}^{-1}\Omega_0 B_0{}^{-1}H')^{-1}H\lambda.$$

4.6 What Each Quantity Estimates

The midterm 2025 Q2(b) explicitly asked for this. Standard answers:

- $\widetilde{\theta}_n$: restricted EE for θ_0 (the parameter under the null).
- $\widetilde{\Omega}_n$: consistent estimator under H_0 of Ω_0 that appears in CF(iii) (the asymptotic variance of $\sqrt{n}\frac{\partial Q_n}{\partial\theta}(\theta_0)$), evaluated at $\widetilde{\theta}_n$.
- $\widetilde{B}_n$: consistent estimator under H_0 of B_0 that appears in CF(iv) (the limit of the Hessian $\frac{\partial^2 Q_n(\theta_0)}{\partial\theta\partial\theta'}$), evaluated at $\widetilde{\theta}_n$.
- $\widetilde{H}_n$: consistent estimator under H_0 of $H = \frac{\partial h}{\partial\theta'}(\theta_0)$.

4.7 Comparison Table

	Wald	LM	QLR
Estimator used	Unrestricted ($\widehat{\theta}$)	Restricted ($\widetilde{\theta}$)	Both
Computational cost	Medium	Low (if restriction simplifies)	High (need both)
Limit under H_0	χ^2_r	χ^2_r	χ^2_r
Local power	$\chi^2_r(\delta)$	$\chi^2_r(\delta)$	$\chi^2_r(\delta)$
Special assumption needed	—	—	Asm QLR ($\Omega_0 = cB_0$)
Finite-sample issues	Often over-rejects	Stable	Often best

4.8 Confidence Region by Test Inversion

CI by Inversion

For any of the three tests at nominal level α,

$$\mathrm{CR}_{1-\alpha} = \{\theta_0 \in \Theta : \text{test does not reject } H_0 : \theta = \theta_0\}.$$

By Theorem 4.4, this CR has asymptotic coverage $1-\alpha$ pointwise.
Caveat: in problems where the limit distribution depends discontinuously on nuisance parameters (weak IV, partial identification), pointwise validity may fail uniformly. See Chapter 8.

4.9 Self-Test Problems

Example (Self-Test 1: Reproduce midterm 2025 Q2).

Derive the local-power asymptotic distribution of LM under $\theta_n = \theta_0 + \lambda/\sqrt{n}$ from scratch. Time yourself: should take 15 minutes.

Example (Self-Test 2: Identify Failure Mode for QLR).

In the LS regression with heteroskedastic errors, why does Assumption QLR fail?

Solution.

Under heteroskedasticity, $\Omega_0 = \mathbb{E}\big(U_i^2 X_i X_i'\big)$ but $B_0 = \mathbb{E}(X_i X_i')$. These are not proportional in general; the proportionality constant depends on the conditional distribution of U_i given X_i. So $\Omega_0 \neq cB_0$ for any scalar c, and Assumption QLR fails. In this setting, use Wald (with Eicker–White SE) or LM. QLR's χ^2 approximation is wrong.

Example (Self-Test 3: When Wald and LM Differ Most).

Suppose the restricted model is linear OLS but the unrestricted model is a hard nonlinear regression. Which test would you pick computationally?

Solution.

LM: it only needs the restricted estimator $\widetilde{\theta}_n$, which is the (easy) linear OLS. No need to solve the (hard) nonlinear minimization.

Chapter 5

Weak Instruments

The Story (Read This First, No Math)

You ran a 2SLS regression. You got an estimate $\widehat{\beta} = 0.4$ and a standard error of 0.1. You report "β is between 0.2 and 0.6 with 95% confidence."

Are you sure?

The 2SLS asymptotic theory says that $\sqrt{n}(\widehat{\beta} - \beta_0) \xrightarrow{d} N(0, V)$, so the 1.96 critical value gives the right coverage in the limit. Yes, in the limit. As $n \to \infty$ with everything else fixed.

But here is what nobody tells you in the basic IV class. The 2SLS distribution is *very* sensitive to how strongly your instruments are correlated with your endogenous regressor. If that correlation is weak, $\widehat{\beta}$ can be substantially biased *and* its actual finite-sample distribution can be far from normal. The reported confidence interval can have actual coverage of 0% for some data-generating processes. Not 90%, not 50%. Zero.

This is not just a small-sample concern. Even in massive samples (Angrist–Krueger 1991 has hundreds of thousands of observations), the weak-IV problem persists if instruments are weak enough. The empirical literature is full of cautionary tales.

The fix — pursued by Patrik and his coauthors over the last 25 years — is to design tests that are *robust to weak instruments*: tests whose actual rejection probability matches the nominal level even when instruments are arbitrarily weak. Three such tests are central to this chapter:

- **Anderson–Rubin (AR) test**, which by clever design has a null distribution that does not depend on instrument strength at all.
- **Kleibergen's $\mathbf{LM_{CUE}}$ test**, which generalizes the LM idea to be valid under both strong and weak IV asymptotics.
- **Conditional likelihood ratio (CLR) test**, which optimizes power among the class of weak-IV-robust tests.

The chapter walks through each, with a focus on the proof structure of $\mathrm{LM}_{\mathrm{CUE}}$, which Patrik tested on HW6 and is the most likely Q2/Q3 candidate.

What You Need to Take Away

By the end of this chapter:

1. Understand why the standard 2SLS t-test fails under weak IVs (Section 5.2).
2. Be able to write down the AR test statistic and explain why it is robust (Section 5.3).
3. Reproduce the four-step proof of $\text{LM}_{\text{CUE}} \xrightarrow{d} \chi^2_{\dim \beta}$ (Section 5.4). HW6-relevant.
4. Recall the Hausman-pretest critique (Section 5.6).

5.1 The Linear IV Model and Concentration Parameter

Setup.

$$y_1 = y_2\beta + X\gamma_1 + u, \qquad y_2 = Z\pi + X\xi + v_2.$$

y_2 is the endogenous regressor; Z is a k-vector of instruments; X is a vector of exogenous controls. Errors (u, v_2) have mean zero and finite variances $\sigma_u^2, \sigma_{v_2}^2$. The correlation $\rho_{u,v_2} = \text{Corr}(u, v_2)$ measures the *degree of endogeneity*.

The concentration parameter. The key quantity measuring instrument strength is

$$\lambda := \pi' Z' Z \pi.$$

Large λ means the instruments collectively explain a lot of the variation in y_2 (strong IV). Small λ means weak IV. Under standard asymptotics with a fixed $\pi \neq 0$, $\lambda \to \infty$ as $n \to \infty$. Under weak-IV asymptotics, $\pi = C/\sqrt{n}$ for some constant C, and $\lambda \to C' D_Z C$ where $D_Z = \lim Z'Z/n$. The limit is a finite constant, capturing the "weakness" of the instruments.

Remark (Why $\pi' Z' Z \pi$ is the right "strength" measure).

> Read λ as a signal-to-noise ratio for the first-stage regression $y_2 = Z\pi + v_2$. The numerator $\pi' Z' Z \pi$ is (up to scaling) the variance of the explained part $Z\pi$; the implicit denominator (when you compare to $\sigma_{v_2}^2$) is the unexplained variance. So λ measures *how much of y_2's variation the instruments actually move* — which is exactly what you need IVs to do. The whole reason 2SLS works asymptotically is that the first-stage fit becomes near-perfect ($\lambda \to \infty$); the whole reason it fails under weak IVs is that this signal stays bounded ($\lambda = O(1)$) so the noise from estimating π is non-negligible relative to the signal.

Two regimes for asymptotics.

- **Strong IV asymptotics**: π fixed, $n \to \infty$. The standard 2SLS theory applies: $\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N(0, V)$.

- **Weak IV asymptotics**: $\pi = C/\sqrt{n}$, $n \to \infty$. The 2SLS distribution is non-normal, depends on C and ρ_{u,v_2}, and the standard t-test fails.

The weak-IV regime is meant to capture the practical reality where instruments may be weak. Standard asymptotics ignores this, which is why standard 2SLS inference can be misleading in practice.

Remark (Why this exact rate $\pi = C/\sqrt{n}$ and not faster or slower).

The rate is calibrated so that the concentration parameter $\lambda = \pi'(Z'Z)\pi$ stays $O_p(1)$ as $n \to \infty$: with $\pi = C/\sqrt{n}$, $\lambda \approx (C'(Z'Z)C)/n \to C'D_Z C$, a finite constant. This is the only rate that does so:

- Faster decay (e.g., $\pi = C/n$): then $\lambda = C'(Z'Z)C/n^2 \to 0$. Instruments effectively vanish; the model becomes degenerate and inference is hopeless.
- Slower decay (e.g., $\pi = C/n^{1/4}$): then $\lambda = C'(Z'Z)C/n^{1/2} \to \infty$. Instruments eventually become strong; standard 2SLS asymptotics kick in and the standard t-test becomes valid in the limit.
- Just right ($\pi = C/\sqrt{n}$): the noise from estimating π is of the same order as the signal from π. The system stays perpetually in the "not strong, not collapsed" middle regime — exactly the regime where weak-IV inference is needed and where Dufour's impossibility holds.

This rate is the analog of Pitman drift in the testing chapter (Chapter 4): we calibrate the parameter to drift at the rate the estimator concentrates, so the limiting object is non-degenerate and informative.

5.2 Why the Standard t-Test Fails (Dufour 1997)

Theorem 5.1: Dufour (1997) Impossibility

Consider the linear IV model with a parameter space allowing π to be arbitrarily close to zero. Then for the standard 2SLS t-test of $H_0 : \beta = \beta_0$ at any nominal level $\alpha \in (0, 1)$:

$$\text{AsySz} = \sup_n \sup_\pi P\left(|t_n| > z_{1-\alpha/2}\right) = 1.$$

The asymptotic size equals 1, not α.

Intuition. The proof builds drifting sequences $\pi_n \to 0$ at rate $\sqrt{n}$. Along these sequences, the t-statistic's limit distribution is a heavy-tailed ratio of normals, not $N(0, 1)$. The rejection probability under the standard normal critical value can approach 1.

Practical consequence. Standard 2SLS confidence intervals based on the 1.96 critical value can have actual coverage approaching 0% for weak-IV designs. Empirically: see Bound, Jaeger, Baker (1995) for the Angrist–Krueger reanalysis, where weak instruments invalidated the original t-test inference despite a sample size of hundreds of thousands.

The One-Sentence Diagnosis

"By Dufour (1997), the standard 2SLS t-test has asymptotic size 1, not the nominal level, when the instrument strength π can be arbitrarily close to zero. Use weak-IV-robust tests (AR, LM_{CUE}, CLR) instead."

5.3 The Anderson–Rubin (AR) Test

The first test designed to be robust to weak IVs.

Idea. Test $H_0 : \beta = \beta_0$ via the F-test of $H_0^* : \kappa = 0$ in the artificial regression

$$y_1 - y_2\beta_0 = Z\kappa + X\gamma + u.$$

If H_0 is true, then $y_1 - y_2\beta_0 = u$, which is uncorrelated with Z by exogeneity, so $\kappa = 0$. If H_0 is false, then $y_1 - y_2\beta_0 = u + (\beta - \beta_0)y_2 = u + (\beta - \beta_0)(Z\pi + \dots)$, which is correlated with Z, so $\kappa \neq 0$.

Definition 5.2: AR Statistic

$$\text{AR}(\beta_0) := \frac{(y_1 - y_2\beta_0)' P_Z (y_1 - y_2\beta_0)/k}{\widehat{\sigma}_u^2(\beta_0)},$$

where $\widehat{\sigma}_u^2(\beta_0) = (y_1 - y_2\beta_0)' M_{[Z:X]}(y_1 - y_2\beta_0)/(n-k-p)$, P_Z is the projection onto Z, and $M_{[Z:X]}$ is the residual maker.

Theorem 5.3: AR Robustness

Under $H_0 : \beta = \beta_0$:

$$\text{AR}(\beta_0) \xrightarrow{d} \chi^2_k/k \quad \text{or, in finite samples under normality, } \text{AR}(\beta_0) \sim F_{k,n-k-p}.$$

The null distribution does not depend on π. Hence the AR test is robust to weak IVs at any IV strength.

Why this works. Under H_0, the test statistic involves u and Z only, never π. So the null distribution is invariant to π. This is the entire point of the AR construction.

Limitation: power. For just-identified models ($k = 1$, one instrument), the AR test is essentially optimal. For over-identified models ($k > 1$), the AR test is a k-degree-of-freedom test for testing a single parameter β. This wastes power. In severely over-identified settings, AR-based confidence intervals can be very wide.

5.4 Kleibergen's $\mathrm{LM}_{\mathrm{CUE}}$ Statistic (HW6 Q2)

A more powerful weak-IV-robust test for over-identified models. Patrik tested this directly on HW6, so a Q2/Q3 question on $\mathrm{LM}_{\mathrm{CUE}}$ is highly likely.

Setup. Linear IV model in the form $y_i = x_i'\beta + \varepsilon_i$ with instruments Z_i. Define the moment function and its derivative:

$$g_i(\beta) := (y_i - x_i'\beta)Z_i, \qquad G_i := \frac{\partial g_i(\beta)}{\partial \beta'} = -Z_i x_i'.$$

Sample moments:

$$\bar{g}(\beta) := \frac{1}{n}\sum_{i=1}^{n} g_i(\beta), \qquad \widehat{\Omega}(\beta) := \frac{1}{n}\sum_{i=1}^{n} g_i(\beta)g_i(\beta)'.$$

Heteroskedasticity-corrected sample Jacobian:

$$D(\beta) := \frac{1}{n}\sum_{i=1}^{n}\left[\bar{g}(\beta)'\widehat{\Omega}(\beta)^{-1}g_i(\beta) - 1\right] G_i.$$

Definition 5.4: $\mathrm{LM}_{\mathrm{CUE}}$ Statistic

$$\mathrm{LM}_{\mathrm{CUE}}(\beta) := n\,\bar{g}(\beta)'\,\widehat{\Omega}(\beta)^{-1}\,D(\beta)\left[D(\beta)'\widehat{\Omega}(\beta)^{-1}D(\beta)\right]^{-1} D(\beta)'\,\widehat{\Omega}(\beta)^{-1}\,\bar{g}(\beta).$$

Remark (How to read this formula).

It looks intimidating but has the same shape as a standard LM statistic. Compare with the Wald-style quadratic form $T = a'Ma$ where $a =$ "thing that should be zero under H_0" and $M =$ "inverse of its asymptotic variance." Here:

- the "thing" is $\widehat{\Omega}(\beta)^{-1/2}D(\beta)'\widehat{\Omega}(\beta)^{-1/2}\cdot\sqrt{n}\bar{g}(\beta)$ (a projected, reweighted score),
- the inverse-variance weight is $\left[D'\widehat{\Omega}^{-1}D\right]^{-1}$.

Kleibergen's contribution was $D(\beta)$: it is a heteroskedasticity-corrected sample Jacobian designed so that $\sqrt{n}\bar{g}(\beta)$ and $\sqrt{n}\cdot\mathrm{vec}(D(\beta))$ are *asymptotically independent*, even under weak IV. That independence is what makes the χ^2 limit hold uniformly in IV strength.

Theorem 5.5: $\mathrm{LM}_{\mathrm{CUE}}$ Asymptotic Null Distribution

Under standard regularity conditions (and assuming $\mathbb{E}(Z_i x_i')$ has full rank under strong IV, or appropriate conditions under weak IV):

$$\mathrm{LM}_{\mathrm{CUE}}(\beta_0) \xrightarrow{d} \chi^2_{\dim\beta_0} \quad \text{under both strong and weak IV asymptotics.}$$

Four-Step Proof Under Strong IV (HW6 Q2(i))

[REPRODUCE — memorize this proof]

Step 1: CLT for $\sqrt{n}\bar{g}(\beta_0)$. Under exogeneity $\mathbb{E}(\varepsilon_i Z_i) = 0$,

$$\sqrt{n}\bar{g}(\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \varepsilon_i Z_i \xrightarrow{d} N(0, \Omega), \qquad \Omega := \mathbb{E}\left(\varepsilon_i^2 Z_i Z_i'\right).$$

Assuming this expectation is finite.

Step 2: WLLN for $\widehat{\Omega}(\beta_0)$.

$$\widehat{\Omega}(\beta_0) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 Z_i Z_i' \xrightarrow{p} \mathbb{E}\left(\varepsilon_i^2 Z_i Z_i'\right) = \Omega.$$

Step 3: Probability limit of $D(\beta_0)$. Plug in $g_i(\beta_0) = \varepsilon_i Z_i$:

$$D(\beta_0) = \underbrace{-\frac{1}{n}\sum_{i=1}^{n}\left[\bar{g}(\beta_0)'\widehat{\Omega}(\beta_0)^{-1}\varepsilon_i Z_i\right] Z_i x_i'}_{\text{(A)}} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} Z_i x_i'}_{\text{(B)}}.$$

For (A): $\bar{g}(\beta_0) = O_p(1/\sqrt{n})$, $\widehat{\Omega}(\beta_0)^{-1} = O_p(1)$, and the inner average is bounded by WLLN. So term (A) $= o_p(1) \cdot O_p(1) = o_p(1)$.
For (B): WLLN gives $n^{-1}\sum_{i=1}^{n} Z_i x_i' \xrightarrow{p} \mathbb{E}(Z_i x_i') =: G$, full rank by assumption.
So $D(\beta_0) \xrightarrow{p} G$.

Step 4: Combine via CMT.

$$\left[D(\beta_0)'\widehat{\Omega}(\beta_0)^{-1}D(\beta_0)\right]^{-1/2} D(\beta_0)'\widehat{\Omega}(\beta_0)^{-1}\sqrt{n}\,\bar{g}(\beta_0) \xrightarrow{d} N(0, I_{\dim \beta_0}).$$

The $\mathrm{LM}_{\mathrm{CUE}}$ statistic is the squared norm of this, hence

$$\mathrm{LM}_{\mathrm{CUE}}(\beta_0) \xrightarrow{d} \chi^2_{\dim \beta_0}. \qquad \blacksquare$$

What about weak IV? Under weak IV asymptotics, Step 3 is different: $D(\beta_0)$ does not converge to a constant. Instead, $\sqrt{n} \cdot D(\beta_0)$ has a normal limit. The trick is that $\sqrt{n}\bar{g}(\beta_0)$ and $\sqrt{n} \cdot \mathrm{vec}(D(\beta_0))$ are *asymptotically independent*, and conditional on D, the LM statistic has a χ^2 limit. Since the conditional limit does not depend on D, the unconditional limit is also χ^2. (HW6 Q3 worked through this.)

If $\mathrm{LM}_{\mathrm{CUE}}$ Appears on the Exam

The four steps above are the proof. Drill them. Each step has a one-sentence justification: CLT, WLLN, decomposition + Slutsky, CMT. Aim for 7–8/10.

5.5 The CLR Test (Brief Mention)

The conditional likelihood ratio (CLR) test of Moreira (2003) is the gold-standard weak-IV-robust test in the iid homoskedastic normal-error case. Its key property: among invariant similar tests, it is essentially *optimal* (lies on the power envelope). Under weak IV asymptotics, the test rejects when

$$\text{LR} > \kappa_\alpha^{\text{CLR}}(Q_T),$$

where Q_T is a sufficient statistic for π under H_0, and the critical value $\kappa_\alpha^{\text{CLR}}$ is computed by numerical integration of the conditional null distribution.

Remark (The CLR conditioning idea, in one sentence).

> Even under weak IV the joint distribution of the data factors into a part depending on β alone (the LR statistic) and a part depending on π alone (the sufficient statistic Q_T). *Conditioning on Q_T removes the nuisance dependence on π*, leaving a β-only test whose null distribution can be computed regardless of how weak the instruments are. The LM and AR tests are special cases of this construction; CLR optimizes power within the conditioning framework.

The HR-CLR and HAR-CLR variants extend CLR to heteroskedastic and autocorrelated errors.

What to Say About CLR on the Exam

You do not need to derive CLR. If asked "what is the most powerful weak-IV-robust test?", say: "the CLR test of Moreira (2003), which is on the power envelope for invariant similar tests in the iid homoskedastic normal model." That is enough.

5.6 The Hausman Pretest Trap (HW5 Q1-2)

A common applied practice: use a Hausman test to decide whether to use OLS or 2SLS. If Hausman doesn't reject exogeneity, use the more efficient OLS; otherwise use 2SLS. *This pretest itself fails under weak IVs.*

Remark (The Hausman idea, in one sentence).

> The Hausman test compares two estimators that should agree if a hypothesis (here: X is exogenous) is true: $H_n \propto (\widehat{\beta}_{OLS} - \widehat{\beta}_{2SLS})^2/\widehat{V}$. Under exogeneity both estimators converge to β_0, so H_n stays small; under endogeneity they diverge, so H_n blows up. With strong IVs the test has high power. With weak IVs, $\widehat{\beta}_{2SLS}$ itself is unstable, so even when OLS is biased the difference can be small relative to $\widehat{V}$ — and the test does not reject. That is the failure.

Setup. Test $H_0 : \beta = \beta_0$ via:

$$T_n^*(\beta_0) = T_{\text{OLS}}^*(\beta_0)\,\mathbf{1}(H_n \le \chi^2_{1,1-\beta}) + T_{\text{2SLS}}^*(\beta_0)\,\mathbf{1}(H_n > \chi^2_{1,1-\beta}),$$

where H_n is the Hausman test statistic for endogeneity, and $\beta = 5\%$ is the pretest size.

What goes wrong. Under weak-IV asymptotics with parameters h_1 = degree of endogeneity, h_2 = instrument strength, the Hausman statistic has noncentral χ^2_1 limit:

$$H_n \xrightarrow{d} \chi^2_1(h_1^2 h_2^2/(h_2^2+1)).$$

When h_2 is small (weak IV), the noncentrality is small, the Hausman test has low power, and the pretest typically chooses OLS — which is biased due to endogeneity.

Simulation results. With $h_2 = 0.1$ (weak IV) and $h_1 \in \{3, 8\}$:

- Two-stage test null rejection probability for $n = 100$: 75.2% and 54.7%.
- For $n = 500$: 68.1% and 40.4%.
- Far above the nominal 5%.

Theoretical AsySz. Computed in Guggenberger (2009): can be much larger than the nominal 5%. The pretest does not solve the problem; it just makes the failure mode harder to detect.

Hausman Pretest Critique

"By Guggenberger (2009), the two-stage test based on a Hausman pretest has asymptotic size that substantially exceeds the nominal level under weak IVs. The Hausman statistic has noncentral χ^2_1 limit with parameter $h_1^2 h_2^2/(h_2^2+1)$, which is small when IVs are weak (small h_2). The pretest then typically chooses OLS even when endogeneity is severe. **Recommendation**: use weak-IV-robust tests (AR, LM_{CUE}, CLR) directly."

5.7 Estimation Under Weak IVs

When IVs are weak, the 2SLS estimator itself is biased and unstable. Some alternatives:

- **Fuller's modified LIML** (with $a = 1$ or $a = 4$): takes the LIML eigenvalue $\widehat{\kappa}_{\text{LIML}}$ and replaces it by $\widehat{\kappa}_{\text{LIML}} - a/(n-k-p)$ before plugging into the LIML formula. This ridge-like correction (a small additive shift) shrinks LIML toward 2SLS just enough to give the estimator *finite moments* and small median bias even under weak IVs — LIML alone has no finite moments at all.
- **Jackknife 2SLS** (JIVE): replaces the projection $P_Z = Z(Z'Z)^{-1}Z'$ with a leave-one-out version that removes own-observation contributions. The leading-order 2SLS bias under weak IVs comes from each i's own z_i projecting onto its own u_i via P_Z; deleting the diagonal entries of P_Z kills exactly this bias term.
- **Andrews–Armstrong (2017)**: when the *sign* of the first-stage coefficient π is known a priori (e.g., theory says "higher tax rate must reduce demand"), they construct an

estimator that is uniformly unbiased for any π with that sign. The sign restriction is the lever that buys uniform unbiasedness; without it, no such estimator exists (Hirano–Porter impossibility).

The AR or CLR *confidence intervals* can be infinite-length under weak IVs — which correctly reflects the inability of weak instruments to identify β precisely. This is a feature, not a bug.

Honest Empirical Reporting

The recommended practice when instruments may be weak: report point estimates accompanied by AR or CLR confidence intervals. Do not rely on 2SLS standard errors and the 1.96 critical value.

5.8 Cheat Sheet

Weak IV Cheat Sheet

Concentration parameter: $\lambda = \pi' Z' Z \pi$. Strong if $\to \infty$, weak if bounded.
Dufour (1997): standard 2SLS t-test has AsySz $= 1$ under weak IVs.
AR test: $\text{AR}(\beta_0) = (y_1 - y_2\beta_0)' P_Z (y_1 - y_2\beta_0)/(k\,\widehat{\sigma}_u^2)$. Null distribution does not depend on π. Robust at any IV strength. UMP for $k = 1$.
LM_{CUE} proof structure: CLT on $\bar{g}$, WLLN on $\widehat{\Omega}$, decompose $D(\beta_0)$ into op + WLLN, combine. $\xrightarrow{d} \chi^2_{\dim \beta_0}$.
CLR test (Moreira 2003): power-optimal among invariant similar tests under iid homoskedastic normal errors. HR-CLR and HAR-CLR extend to heteroskedastic and autocorrelated cases.
Hausman pretest: *fails* under weak IVs (Guggenberger 2009). Do not use.
Estimation alternatives: Fuller's modified LIML, jackknife 2SLS, Andrews–Armstrong (2017).

5.9 Self-Test Problems

Example (Self-Test 1: One-sentence diagnosis).

Why does the standard 2SLS t-test fail under weak IVs?

Solution.

By Dufour (1997), if the parameter space allows the instrument strength π to be arbitrarily close to zero, the asymptotic size of the standard t-test equals 1, not the nominal level α. The reason is that the t-statistic's limit distribution is non-normal under weak-IV asymptotics ($\pi = C/\sqrt{n}$), and the standard 1.96 critical value can lead to rejection probabilities approaching 1.

Example (Self-Test 2: Why is AR robust?).

Why is the null distribution of the AR statistic invariant to π?

Solution.

Under $H_0 : \beta = \beta_0$, $y_1 - y_2\beta_0 = u$, which depends only on the structural error u, not on the reduced-form coefficient π. The AR statistic is a function of u and Z only. So its distribution under H_0 does not involve π, regardless of instrument strength.

Example (Self-Test 3: $\mathbf{LM_{CUE}}$ four-step proof).

Reproduce the four-step proof of $\mathrm{LM}_{\mathrm{CUE}}(\beta_0) \xrightarrow{d} \chi^2_{\dim \beta_0}$ under strong IV asymptotics. Time yourself: under 12 minutes.

Example (Self-Test 4: Hausman critique).

Why does the Hausman pretest fail under weak IVs?

Solution.

The Hausman statistic has noncentrality parameter $h_1^2 h_2^2/(h_2^2 + 1)$, where h_2 is instrument strength. When h_2 is small (weak IV), the noncentrality is small, and the Hausman test has low power to detect endogeneity. The pretest typically chooses OLS, which is biased by the endogeneity that the weak instruments could not detect. The result is a two-stage test whose actual size substantially exceeds the nominal level (Guggenberger 2009).

Chapter 6

Bootstrap Improvements (Edgeworth Expansion)

Defensive Chapter

This is short. Lecture 23 details (Edgeworth coefficient calculations, cumulant generating functions) are unlikely to be tested in proof-form. What *is* testable: the qualitative result that bootstrap is higher-order accurate for pivotal statistics. If a Q3-style problem asks "why bootstrap?" or "how does bootstrap compare to delta method?", you need to be able to state the rates.
Target: 1–2 sentences worth of points if asked.

Why This Chapter Exists at All

Chapter 2 showed that bootstrap is asymptotically valid: $\mathrm{P}(\theta_0 \in \mathrm{CI}_n) \to 1-\alpha$. So is the standard normal-approximation CI: $\mathrm{P}\Big(\theta_0 \in \Big[\widehat{\theta} \pm 1.96\widehat{\sigma}/\sqrt{n}\Big]\Big) \to 1-\alpha$. Both have the same first-order limit $1-\alpha$. So why bother with bootstrap, which is computationally more expensive?

The answer is that "$\to 1-\alpha$" hides a *rate*: how fast the actual coverage approaches the nominal level as n grows. The standard normal CI has coverage error $O(n^{-1/2})$. The bootstrap CI, when applied to a *pivotal* (studentized) statistic, has coverage error $O(n^{-1})$ (equal-tailed) or $O(n^{-3/2})$ (symmetric) — one to two orders of magnitude better in n.

The reason is the Edgeworth expansion: the true CDF of a t-statistic has corrections beyond the leading $\Phi(x)$ term, scaled by powers of $n^{-1/2}$. The normal approximation discards all these corrections. The bootstrap, by resampling from the data's own empirical distribution, automatically captures the leading correction (and sometimes more). This is the entire content of the chapter — a higher-order rate justification for using bootstrap CIs in finite samples, even though both procedures are asymptotically valid.

6.1 The Edgeworth Expansion

For any well-behaved t-statistic $T_n = \sqrt{n}(\widehat{\theta}_n - \theta_0)/\sigma$, its CDF admits an asymptotic expansion:

$$\mathrm{P}(T_n \le x) = \Phi(x) + n^{-1/2} p_1(x)\phi(x) + n^{-1} p_2(x)\phi(x) + O(n^{-3/2}),$$

where Φ, ϕ are the standard normal CDF and PDF, and p_1, p_2 are polynomials whose coefficients depend on cumulants of the underlying distribution (skewness, excess kurtosis, etc.).

Delta method approximates only the first term. The standard normal approximation, used in delta-method CIs, ignores the $n^{-1/2}$ and n^{-1} corrections, hence has error $O(n^{-1/2})$.

6.2 Bootstrap as a Higher-Order Approximation

Theorem 6.1: Hall (1992) Bootstrap Improvement

For a *pivotal* statistic (one whose limit distribution does not depend on nuisance parameters — e.g., a fully studentized t-statistic), the bootstrap captures the $n^{-1/2}$ Edgeworth correction. Consequently:

- Two-sided equal-tailed bootstrap CI: error $O(n^{-1})$.
- Symmetric two-sided bootstrap CI: error $O(n^{-3/2})$.

For comparison, the delta-method CI has error $O(n^{-1/2})$ (equal-tailed) or $O(n^{-1})$ (symmetric).

Remark.

> Why bootstrap captures the correction. The bootstrap distribution of T_n^* matches the original distribution of T_n up to terms of order $n^{-1/2}$, because $\widehat{F}_n \xrightarrow{a.s.} F$ uniformly and the polynomials p_1, p_2 are continuous in F. Pivoting (studentizing) ensures that the leading-order dependence on nuisance parameters cancels.

6.3 Why Symmetric Beats Equal-Tailed

Symmetric CI cancels odd-power Edgeworth terms. The polynomial p_1 is odd ($p_1(-x) = -p_1(x)$). When you symmetrize the test, the p_1 contributions cancel, leaving the next-order p_2 term. Bootstrap captures *both* p_1 and p_2 corrections, so symmetric bootstrap CIs have error $O(n^{-3/2})$.

Quick Comparison Table

Statistic	Delta-method error	Bootstrap error
Pivotal, equal-tailed CI	$O(n^{-1/2})$	$O(n^{-1})$
Pivotal, symmetric CI	$O(n^{-1/2})$	$O(n^{-3/2})$
Non-pivotal	$O(n^{-1/2})$	$O(n^{-1/2})$ (no improvement)

6.4 When Bootstrap Does *Not* Help

- Non-pivotal statistics (e.g., the unstudentized estimator $\widehat{\theta}_n$ itself, whose limit distribution depends on σ^2).
- Discontinuities in the limit distribution (boundary cases, weak IVs, partial identification).
- Bootstrap inconsistency cases: e.g., parameter on boundary of parameter space.

6.5 Self-Test

Example (Self-Test: Two-Sentence Answer).

"Why is the bootstrap better than the asymptotic normal approximation?"

Solution.

For a pivotal statistic, the bootstrap automatically incorporates the $n^{-1/2}$ Edgeworth correction that the standard normal approximation misses. As a result, equal-tailed bootstrap CIs have error $O(n^{-1})$ and symmetric ones $O(n^{-3/2})$, both better than the delta-method's $O(n^{-1/2})$.

Chapter 7

Nonsmooth GMM (Quantile Regression)

Why This Chapter Is Worth a Pass

Midterm 2025 Q4 and midterm 2026 Q2 *both* tested the nonsmooth GMM material. So Q2 of the final could very plausibly be a V_0/Γ identification problem. The good news: the answer is mostly formula-recall.
Target: 4–6/10 by writing down the V_0 estimator, Γ estimator, and the quantile-regression specialization.

7.1 The Setup: Nonsmooth Sample Moments, Smooth Population Moments

In some models, the moment function $g(W_i, \theta)$ is *nonsmooth* in θ (e.g., contains indicator functions or kinks), but its expectation $g(\theta) = \mathbb{E}(g(W_i, \theta))$ is smooth. Why? Integration is a smoothing operation. Example:

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(Y_i \leq \theta) \quad \text{nonsmooth in } \theta,$$

but

$$\mathbb{E}(\mathbf{1}(Y_i \leq \theta)) = F_Y(\theta) \quad \text{smooth in } \theta.$$

7.2 Assumptions and Main Theorem

Assumption 7.1: EE3 (Nonsmooth EE)

(i) $Q_n(\widehat{\theta}_n) = \inf_\theta Q_n(\theta) + o_p(\sqrt{n}^{-1})$.
(ii) $\widehat{\theta}_n \xrightarrow{p} \theta_0$.

Assumption 7.2: CF-NS (Nonsmooth Criterion Function)

(i) $\theta_0 \in \text{int}(\Theta)$.
(ii) $g(\theta)$ is differentiable at θ_0 with $\Gamma := \frac{\partial g}{\partial \theta'}(\theta_0)$ of full rank $d \leq k$.
(iii) $g(\theta_0) = 0$.
(iv) $\sqrt{n}\bar{g}_n(\theta_0) \xrightarrow{d} N(0, V_0)$ (CLT for the moment).
(v) (Stochastic equicontinuity) For every $\delta_n \downarrow 0$,

$$\sup_{\theta \in \Theta : \|\theta - \theta_0\| < \delta_n} \sqrt{n}\, \|\bar{g}_n(\theta) - g(\theta) - \bar{g}_n(\theta_0)\| \xrightarrow{p} 0.$$

Remark (What stochastic equicontinuity buys us, and why it is the right replacement for U-WCON).

Stochastic equicontinuity is the nonsmooth replacement for the U-WCON condition that powered consistency in Chapter 1. There the moment function $m(W_i, \theta)$ was continuous in θ, so we could take a uniform sup over θ; here, $g(W_i, \theta)$ has kinks/indicators, so the empirical process $\nu_n(\theta) := \sqrt{n}(\bar{g}_n(\theta) - g(\theta))$ can jump as θ moves by an infinitesimal amount.

What the condition says, in words. Near θ_0, the jumps shrink fast enough that we can still treat $\bar{g}_n(\theta)$ as approximately $g(\theta) + \nu_n(\theta_0)/\sqrt{n}$ for any θ close to θ_0 — i.e., the noise around $g(\theta)$ is well-approximated by the noise at the single fixed point θ_0. The expansion $\sqrt{n}\bar{g}_n(\widehat{\theta}_n) = \sqrt{n}g(\widehat{\theta}_n) + \nu_n(\theta_0) + o_p(1)$ relies on exactly this — the equicontinuity controls the $\nu_n(\widehat{\theta}_n) - \nu_n(\theta_0)$ remainder. Without it, the remainder could be $O_p(1)$, and the asymptotic distribution would not be normal.

Why "technically hard." Verifying it requires bracketing-entropy or VC-class machinery (empirical-process theory). For exam purposes you do not derive it; you cite it and move on.

Theorem 7.3: Pakes–Pollard / Andrews Theorem 16.1

Under EE3 and CF-NS with $A_n = I_k$,

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0,\ (\Gamma'\Gamma)^{-1}\Gamma' V_0 \Gamma (\Gamma'\Gamma)^{-1}\right).$$

The linear expansion is $\sqrt{n}(\widehat{\theta}_n - \theta_0) = -(\Gamma'\Gamma)^{-1}\Gamma'\sqrt{n}\bar{g}_n(\theta_0) + o_p(1)$.

Remark (How Theorem 7.2 generalizes the smooth-GMM Theorem 1.3).

Compare to the smooth case from Chapter 1: there we proved $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, {B_0}^{-1}\Omega_0 {B_0}^{-1})$ via the four-step mean-value expansion of the FOC. Here the FOC itself cannot be expanded (it's nonsmooth in θ), so we expand the *population* moment $g(\theta) = \mathbb{E}(g(W_i, \theta))$ instead — expectation smooths out the indicators/kinks, so $g(\theta)$ has a derivative even when $g(W_i, \theta)$ does not. Match the pieces:

- Smooth-case Hessian $B_0 = \frac{\partial^2 Q(\theta_0)}{\partial \theta \partial \theta'}$ $\leftrightarrow$ nonsmooth-case Jacobian $\Gamma = \frac{\partial g(\theta_0)}{\partial \theta'}$.

- Smooth-case score variance $\Omega_0 = \mathrm{Var}\left(\sqrt{n}\frac{\partial Q_n(\theta_0)}{\partial \theta}\right)$ $\leftrightarrow$ nonsmooth-case moment variance $V_0 = \mathbb{E}(g(W_i,\theta_0)g(W_i,\theta_0)')$.

The sandwich $(\Gamma'\Gamma)^{-1}\Gamma' V_0 \Gamma(\Gamma'\Gamma)^{-1}$ collapses to $\Gamma^{-1}V_0\Gamma^{-1\prime}$ when $k = d$ (just-identified) — exactly the smooth-GMM formula in that case. So nonsmooth GMM adds one extra ingredient (stochastic equicontinuity to control the empirical process near θ_0) and otherwise mirrors smooth GMM step for step.

7.3 Estimation of V_0 and Γ

7.3.1 Estimating V_0 (Lecture 16, Eq. 16.24)

In the i.i.d. case, $V_0 = \mathrm{Var}(g(W_i,\theta_0))$, estimated by

$$\widehat{V}_n = \frac{1}{n}\sum_{i=1}^{n}(g(W_i,\widehat{\theta}_n) - \bar{g}_n(\widehat{\theta}_n))(g(W_i,\widehat{\theta}_n) - \bar{g}_n(\widehat{\theta}_n))'.$$

Consistent under uniform WLLNs on g and gg' (a manageable assumption).

7.3.2 Estimating Γ (Lecture 16, Eq. 16.25)

Because $g(W_i,\theta)$ is *nonsmooth* in θ, we cannot just take a sample derivative. Use *finite differences*:

$$\widehat{\Gamma}_{nj} := \frac{1}{\varepsilon_n}\left[\bar{g}_n(\widehat{\theta}_n + \varepsilon_n e_j) - \bar{g}_n(\widehat{\theta}_n)\right],$$

where e_j is the j-th elementary vector and ε_n is a positive bandwidth satisfying

$$\varepsilon_n \to 0 \quad \text{and} \quad 1/(\sqrt{n}\varepsilon_n) = O_p(1).$$

A common choice: $\varepsilon_n = n^{-\delta}$ for $0 < \delta \leq 1/2$.

Remark (Why these two bandwidth conditions: bias-variance trade-off).

The two bandwidth conditions encode a bias-variance trade-off familiar from kernel estimation:

- **Bias** (from finite differencing): the ratio $\left[\bar{g}_n(\widehat{\theta}_n + \varepsilon_n e_j) - \bar{g}_n(\widehat{\theta}_n)\right]/\varepsilon_n$ approximates the true derivative $\nabla g(\theta_0)$ with error $O(\varepsilon_n)$ by Taylor's theorem. This bias vanishes only if $\varepsilon_n \to 0$.
- **Variance** (from sample noise): the numerator has stochastic fluctuations of order $1/\sqrt{n}$ (CLT scale of the empirical moment). Dividing by ε_n amplifies the noise to $1/(\sqrt{n}\varepsilon_n)$. This vanishes only if $\sqrt{n}\varepsilon_n \to \infty$.

Together: $\varepsilon_n \to 0$ kills bias, $\sqrt{n}\varepsilon_n \to \infty$ kills variance. The valid window $1/\sqrt{n} \ll \varepsilon_n \ll 1$ is exactly $\varepsilon_n = n^{-\delta}$ for $\delta \in (0, 1/2]$. Pick δ in the middle (e.g., $\delta = 1/4$, so $\varepsilon_n = n^{-1/4}$) to balance the two errors at order $n^{-1/4}$.

Theorem 7.4: Consistency of $\widehat{\Gamma}_n$

$\widehat{\Gamma}_{nj} \xrightarrow{p} \Gamma_j$ (the j-th column of Γ) under stochastic equicontinuity (CF-NS(v)).

V_0 and Γ Estimators (Q2 / Q4 of Midterms)

- $\widehat{V}_n = n^{-1}\sum_{i=1}^n (g(W_i,\widehat{\theta}_n) - \bar{g}_n(\widehat{\theta}_n))(g(W_i,\widehat{\theta}_n) - \bar{g}_n(\widehat{\theta}_n))'$.
- $\widehat{\Gamma}_{nj} = \varepsilon_n^{-1}(\bar{g}_n(\widehat{\theta}_n + \varepsilon_n e_j) - \bar{g}_n(\widehat{\theta}_n))$, with $\varepsilon_n \to 0$ and $1/(\sqrt{n}\varepsilon_n) = O_p(1)$.

7.4 Quantile Regression as the Canonical Example

Setup. For some known $\tau \in (0,1)$:

$$Y_i = X_i'\theta_0 + U_i, \quad q(\tau \mid X_i) = 0 \text{ a.s.},$$

where $q(\tau \mid x)$ is the τ-quantile of U_i given $X_i = x$. We assume U_i has a continuous conditional density $f_{U \mid X}$ that is continuous in a neighborhood of zero, and $\mathbb{E}\left(\|X_i\|^2\right) < \infty$.

Remark (What "$q(\tau \mid X_i) = 0$" really says, and why it identifies θ_0).

The condition is a *quantile* version of the OLS exogeneity condition $\mathbb{E}(U_i \mid X_i) = 0$. Whereas OLS pins down the conditional *mean* of Y_i at $X_i'\theta_0$, the quantile-regression assumption pins down the conditional τ-*quantile* of Y_i at $X_i'\theta_0$:

$$q(\tau \mid X_i) = 0 \iff P(U_i \le 0 \mid X_i) = \tau \iff P(Y_i \le X_i'\theta_0 \mid X_i) = \tau.$$

So $X_i'\theta_0$ is the regression line that has exactly $100\tau\%$ of the Y_i's below it (conditional on X_i). For $\tau = 1/2$ this is median regression: the line bisects Y_i's conditional distribution. The assumption is what *makes* θ_0 identifiable from quantile regression — without it, no specific line plays the role of the conditional τ-quantile.

7.4.1 Two Equivalent Formulations

Check function. The estimator $\widehat{\theta}_n$ minimizes

$$n^{-1}\sum_{i=1}^n \rho_\tau(Y_i - X_i'\theta), \quad \rho_\tau(u) = u(\tau - \mathbf{1}(u < 0)).$$

ρ_τ is the "check function": linear with slope τ for $u > 0$ and slope $-(1-\tau)$ for $u < 0$.

GMM moment. Equivalently, $\widehat{\theta}_n$ solves

$$\bar{g}_n(\theta) := n^{-1}\sum_{i=1}^n [\tau \mathbf{1}(Y_i - X_i'\theta > 0) - (1-\tau)\mathbf{1}(Y_i - X_i'\theta < 0)]\, X_i = 0_d.$$

7.4.2 Population Moments

Compute:

$$g(\theta) = \mathbb{E}(X_i \left[\tau \cdot \mathrm{P}(U_i > X_i'(\theta - \theta_0) \,|\, X_i) - (1-\tau) \cdot \mathrm{P}(U_i < X_i'(\theta - \theta_0) \,|\, X_i)\right]).$$

At $\theta = \theta_0$: $\mathrm{P}(U_i > 0 \,|\, X_i) = 1 - \tau$, $\mathrm{P}(U_i < 0 \,|\, X_i) = \tau$, so

$$g(\theta_0) = \mathbb{E}(X_i \left[\tau(1-\tau) - (1-\tau)\tau\right]) = 0.$$

This verifies CF-NS(iii).

7.4.3 Computing Γ

By the fundamental theorem of calculus:

$$\frac{\partial}{\partial \theta'} \mathrm{P}(U_i \le X_i'(\theta - \theta_0) \,|\, X_i = x) = f_{U \,|\, X}(x'(\theta - \theta_0) \,|\, x)x'.$$

Combining,

$$\Gamma = \frac{\partial g(\theta_0)}{\partial \theta'} = -\mathbb{E}\big(f_{U \,|\, X}(0 \,|\, X_i) X_i X_i'\big).$$

If $U_i \perp X_i$ (homogeneous quantile density), this simplifies to

$$\Gamma = -f_U(0)\mathbb{E}(X_i X_i').$$

7.4.4 Computing V_0

$$V_0 = \mathbb{E}\Big(\left[\tau \mathbf{1}(U_i > 0) - (1-\tau)\mathbf{1}(U_i < 0)\right]^2 X_i X_i'\Big).$$

Using $\mathbf{1}(U_i > 0) \cdot \mathbf{1}(U_i < 0) = 0$:

$$V_0 = \mathbb{E}\big(X_i X_i' \left[\tau^2 \,\mathrm{P}(U_i > 0 \,|\, X_i) + (1-\tau)^2 \,\mathrm{P}(U_i < 0 \,|\, X_i)\right]\big) = \tau(1-\tau)\mathbb{E}(X_i X_i').$$

7.4.5 Combining: Asymptotic Variance

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0,\ \Gamma^{-1} V_0 \left(\Gamma^{-1}\right)'\right).$$

Under $U \perp X$:

$$\Gamma^{-1} V_0 \Gamma^{-1} = \frac{\tau(1-\tau)}{f_U(0)^2} \left(\mathbb{E}(X_i X_i')\right)^{-1}.$$

For median regression ($\tau = 1/2$):

$$\Gamma^{-1} V_0 \Gamma^{-1} = \frac{1}{4 f_U(0)^2} \left(\mathbb{E}(X_i X_i')\right)^{-1}.$$

Remark.

> Compare to LS: under homoskedasticity, the LS variance is $\sigma_U^2 (\mathbb{E}(X_i X_i'))^{-1}$. Median regression beats LS iff $1/(4f_U(0)^2) < \sigma_U^2$. For Cauchy or other heavy-tailed errors, median regression is overwhelmingly more efficient. For Gaussian errors, LS is more efficient.

7.4.6 V_0 Estimator (Doesn't Need $U \perp X$)

$$\widehat{V}_n = \tau(1-\tau) \cdot \frac{1}{n} \sum_{i=1}^{n} X_i X_i'.$$

Γ estimation requires either nonparametric density estimation or the finite-difference scheme.

7.5 Cheat-Sheet Summary

Nonsmooth GMM Cheat Sheet

Asymptotic distribution: $\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{d} N(0, (\Gamma'\Gamma)^{-1}\Gamma' V_0 \Gamma (\Gamma'\Gamma)^{-1})$.
V_0 **estimator**: $\widehat{V}_n = n^{-1} \sum_{i=1}^{n} (g(W_i, \widehat{\theta}_n) - \bar{g}_n)(\cdot)'$.
Γ **estimator**: finite difference $\widehat{\Gamma}_{nj} = \varepsilon_n^{-1}(\bar{g}_n(\widehat{\theta}_n + \varepsilon_n e_j) - \bar{g}_n(\widehat{\theta}_n))$.
Quantile regression V_0: $\tau(1-\tau)\mathbb{E}(X_i X_i')$.
Quantile regression Γ: $-\mathbb{E}\big(f_{U \mid X}(0 \mid X_i) X_i X_i'\big)$. Under $U \perp X$: $-f_U(0)\mathbb{E}(X_i X_i')$.
Asymptotic variance (independence): $\frac{\tau(1-\tau)}{f_U(0)^2}(\mathbb{E}(X_i X_i'))^{-1}$.
Median regression: $\tau = 1/2$, variance $= \frac{1}{4 f_U(0)^2}(\mathbb{E}(X_i X_i'))^{-1}$.

7.6 Self-Test Problems

Example (Self-Test 1: Verify $g(\theta_0) = 0$ for Quantile Regression).

Show that $g(\theta_0) = E[X_i \cdot (\tau \mathbf{1}(U_i > 0) - (1-\tau)\mathbf{1}(U_i < 0))] = 0$.

Solution.

Use iterated expectations:

$$g(\theta_0) = \mathbb{E}(X_i \cdot [\tau \, \mathrm{P}(U_i > 0 \mid X_i) - (1-\tau)\, \mathrm{P}(U_i < 0 \mid X_i)]) \, .$$

By the τ-quantile condition, $\mathrm{P}(U_i \le 0 \mid X_i) = \tau$ a.s. (for continuous U), so $\mathrm{P}(U_i > 0 \mid X_i) = 1 - \tau$ and $\mathrm{P}(U_i < 0 \mid X_i) = \tau$. Substitute:

$$\mathbb{E}(X_i \cdot [\tau(1-\tau) - (1-\tau)\tau]) = \mathbb{E}(X_i \cdot 0) = 0.$$

Example (Self-Test 2: Quantile Regression vs OLS).

Under homoskedastic Cauchy errors, which estimator has smaller asymptotic variance: median regression or OLS?

Solution.

OLS variance is infinite under Cauchy (no second moment). Median regression has variance $\frac{1}{4 f_U(0)^2}(\mathbb{E}(X_i X_i'))^{-1}$, which is finite (Cauchy has finite density at 0). Median wins decisively. (More generally, median regression is preferred to OLS for heavy-tailed errors where the variance may not exist or be very large.)

Chapter 8

Asymptotic Size (AsyCS)

The Story (Read This First, No Math)

You have built a test of $H_0 : \theta = \theta_0$. You designed it so that under H_0, the rejection probability tends to 5% as n grows. The lecture notes tell you "the test is asymptotically valid at level 5%." But *is* it really 5%?

Here is the catch. The phrase "rejection probability tends to 5%" usually means: *for each fixed* parameter combination, the rejection probability tends to 5%. Different parameter values may approach 5% at different speeds. So when you actually have a finite sample n, the rejection probability could be far from 5% for *some* parameter values, even though it would converge to 5% if you fixed those values and let $n \to \infty$.

The classic example is Patrik's bread and butter: the linear IV t-test under weak instruments. The asymptotic distribution of the t-statistic is $N(0, 1)$ when the instruments are strong, so the 1.96 critical value gives 5% rejection. But when the instruments are arbitrarily weak — which is allowed in the parameter space — the actual rejection probability for any fixed sample size can be *up to* 100%. Dufour (1997) proved this. The standard t-test based on 2SLS has *actual* size 1, not 5%.

This chapter formalizes the gap. The key concept is *asymptotic size*: the worst-case rejection probability over the parameter space, taken first, then with $n \to \infty$. "Worst case first" is the entire point: we want a uniform statement, not a pointwise one.

What You Need to Take Away

This chapter is in defensive territory. Patrik will test it on Q4, but the entire class typically gets very few points on Q4. Aim for partial credit by:

1. State the AsySz definition correctly (Section 8.1).
2. Distinguish pointwise convergence from uniform convergence (Section 8.2).
3. Cite the linear IV t-test as the canonical AsySz = 1 example (Section 8.3).
4. Set up the drifting-sequence framework if asked, but do not overinvest (Section 8.4).

Do not spend more than 10 minutes on Q4. Skip parts (ii) and (iii). Move on.

8.1 Asymptotic Size: The Definition

The null hypothesis is $H_0 : \theta = \theta_0$. The test statistic is $T_n(\theta_0)$. The critical value is $c_{1-\alpha}$. We *want* the rejection probability to be α in finite samples, but we have to settle for an asymptotic version.

There are two natural asymptotic notions, and they are different.

Pointwise asymptotic null rejection probability. For each fixed parameter combination $\gamma \in \Gamma$ (where Γ is the parameter space):

$$\text{NRP}(\gamma) := \lim_{n\to\infty} P_{\theta_0,\gamma}\left(T_n > c_{1-\alpha}\right).$$

You compute the limit *after* fixing γ. The textbook claim "the test has asymptotic level α" usually means $\text{NRP}(\gamma) = \alpha$ for each γ.

Asymptotic size (uniform).

$$\boxed{\text{AsySz}(\theta_0) := \limsup_{n\to\infty} \sup_{\gamma\in\Gamma} P_{\theta_0,\gamma}\left(T_n > c_{1-\alpha}\right).}$$

You take the worst case over γ *first*, then let $n \to \infty$.

Why are these different? The pointwise version asks, "for each γ, eventually the rejection probability is close to α." But "eventually" might depend on γ: maybe for γ_1, it takes $n = 1000$, while for γ_2, it takes $n = 10^9$. If you fix any finite n, you can always find a γ for which the rejection probability is far from α.

The asymptotic size says: "there is some sample size n_0 from which on, the rejection probability is uniformly close to α across the entire parameter space." This is a much stronger statement.

Order of Operations Matters

The supremum is taken *before* the limit superior. Switching the order would give the (much weaker) pointwise notion. This is the entire point of the definition.

8.2 Pointwise vs Uniform: A Picture

Imagine plotting the rejection probability $P_{\theta_0,\gamma}(T_n > c_{1-\alpha})$ as a function of γ, for a fixed sample size n. As $n \to \infty$:

- Pointwise convergence to α: for each γ, the value at γ tends to α. The graph passes through α at every point.
- Uniform convergence (AsySz $= \alpha$): the entire graph approaches the horizontal line α uniformly. There are no "spikes" anywhere.

It is possible to have pointwise convergence without uniform convergence. The classic shape: the graph has a tall "spike" at some γ^*. As $n \to \infty$, the spike narrows but stays tall.

For each $\gamma \neq \gamma^*$, the rejection probability eventually leaves the spike and tends to α, so pointwise convergence holds. But the supremum (the spike's height) never decreases. The asymptotic size remains at the spike height.

8.3 The Canonical Example: Linear IV t-Test (Dufour 1997)

This is the example everyone references. You should be able to write a one-paragraph version on the exam.

Setup. Linear IV: $y_1 = y_2\theta + u$, $y_2 = z\pi + v$. Test $H_0 : \theta = \theta_0$ with the standard 2SLS t-statistic and normal critical value $z_{1-\alpha/2}$.

Parameter space. The parameter π measures instrument strength. The standard parameter space allows π to be any real number, including arbitrarily close to zero. "Weak instruments" means π near zero.

Theorem 8.1: Dufour (1997) Impossibility

For the linear IV t-test with parameter space allowing π to be arbitrarily close to zero,

$$\text{AsySz}(\theta_0) = 1.$$

The asymptotic size equals 1, regardless of the nominal level $\alpha \in (0, 1)$.

Why? Take a sequence of parameters π_n approaching zero in such a way that the concentration parameter $n\pi_n^2\mathbb{E}(z^2)$ stays bounded (the "weak IV asymptotics" regime). Under such a sequence, the limit distribution of the t-statistic is *not* $N(0, 1)$; it is a complicated ratio-of-normals object with extreme tail behavior. The 1.96 critical value, designed for $N(0, 1)$, can produce rejection probabilities approaching 1 along these sequences.

In contrast, the pointwise NRP at any *fixed* $\pi \neq 0$ is exactly α, because along that fixed value, eventually the standard $N(0, 1)$ asymptotic kicks in. The failure is in the corner where π is allowed to drift toward zero with n.

Standard Q4 Answer (Memorize)

"The standard 2SLS t-test has asymptotic size equal to 1 (Dufour 1997), not the nominal level α, when the parameter space allows the instrument strength π to be arbitrarily close to zero. The reason is that the limit distribution of the t-statistic depends discontinuously on π at $\pi = 0$. For each fixed $\pi \neq 0$, the pointwise asymptotic NRP equals α, but along sequences where $\pi_n \to 0$ at rate $n^{-1/2}$, the rejection probability can approach 1. The cure is to use weak-IV-robust tests like the AR test or Kleibergen's LM_{CUE} (Chapter 5)."

8.4 The Drifting Sequence Framework

This Section Is Optional

The full Andrews–Guggenberger framework is complicated. Read this section once for understanding, then move on. *Do not* try to memorize the details; on the exam, just write the standard Q4 answer above.

The technical machinery for computing AsySz involves *drifting parameter sequences* that approach a point of discontinuity in the limit distribution.

Reparametrization. Decompose the parameter γ into three pieces:

$$\gamma = (\gamma_1, \gamma_2, \gamma_3),$$

where γ_1 is the "distance to discontinuity," γ_2 is a continuous nuisance parameter that affects the limit distribution, and γ_3 is everything else (does not affect the limit).

For the IV t-test: $\gamma_1 \propto \pi$ (instrument strength), γ_2 = correlation between structural and reduced-form errors, γ_3 = error distribution.

Drifting sequence. Consider a sequence of parameter values γ_n such that

$$\sqrt{n}\,\gamma_{n,1} \to h_1, \qquad \gamma_{n,2} \to h_2.$$

The localization parameter $h = (h_1, h_2)$ tells us *how fast* we approach the discontinuity. Different rates of approach can give different limit distributions J_h.

The AsySz formula.

$$\text{AsySz}(\theta_0) = \sup_{h \in H} \left[1 - J_h(c_{1-\alpha})\right],$$

where H is the set of localization parameter values h achievable along some drifting sequence.

What this says. The asymptotic size is the worst rejection probability over the limit distributions J_h, evaluated at the fixed critical value $c_{1-\alpha}$. If for some h, the limit distribution J_h has a fat tail at $c_{1-\alpha}$, the asymptotic size will be much larger than α.

8.5 Worked Example: Midterm 2025 Q4

This was on the midterm 2025. Q4 is famously hard; the goal is partial credit, not a complete answer.

Setup. $(X_i, Y_i) \in \mathbb{R}^2$ i.i.d., parameter of interest $\theta = \max\{\mu_X, \mu_Y\}$. Construct a 95% confidence interval for θ as follows:

$$C_n = \begin{cases} C_{X_n} & \text{if } \bar{X}_n \geq \bar{Y}_n, \\ C_{Y_n} & \text{otherwise,} \end{cases}$$

where C_{X_n}, C_{Y_n} are the standard t-CIs for μ_X, μ_Y.

Part (i): Coverage for fixed F

Case 1: $\mu_X \neq \mu_Y$ (say $\mu_X > \mu_Y$). By LLN, $\bar{X}_n > \bar{Y}_n$ with probability $\to 1$, so $C_n = C_{X_n}$ eventually. By CLT, $P(\mu_X \in C_{X_n}) \to 1 - \alpha$. The true θ is μ_X, so coverage tends to $1 - \alpha$. Good.

Case 2: $\mu_X = \mu_Y = \theta$. Now $\bar{X}_n - \bar{Y}_n$ is centered at zero and has variance shrinking like $1/n$, so the choice between C_{X_n} and C_{Y_n} becomes random. The coverage probability converges to a limit involving a joint normal distribution of $(\bar{X}_n, \bar{Y}_n)$. The exact formula is messy; the key observation is that this limit is generally *less than* $1-\alpha$ because the random choice between two CIs degrades coverage.

Q4(i) Tip

Write Case 1 carefully (this is straightforward). For Case 2, write down "the coverage converges to a joint-normal probability that may be less than $1 - \alpha$" and stop. Do not try to compute the exact formula. You will get most of the available points for Case 1.

Parts (ii) and (iii): Drifting sequences and AsyCS formula

Skip These Parts

Computing the limit under drifting sequences and the AsyCS formula requires the full Andrews–Guggenberger machinery. Even strong students typically score zero here. Spend the time elsewhere on the exam.

8.6 Sharper Critical Values: GMS and Bonferroni (Background Reading)

Read Once for Awareness, Then Move On

This section describes how researchers in this area construct critical values that are less conservative than the worst-case plug-in. Patrik tested this on HW5 and HW10. If asked on the exam, mention the names; do not try to derive.

The plug-in critical value $c_{(0,\widehat{\gamma}_{2,n})}(1-\alpha)$ is conservative because it assumes the worst-case localization $h_1 = 0$. Two refinements use the data more aggressively.

Andrews–Soares (2010) Generalized Moment Selection (GMS)

Idea. Although h_1 cannot be consistently estimated, the data does carry *some* information about h_1. Use a pre-test to decide which moment inequalities are slack.

For a user-chosen sequence $\kappa_n \to \infty$ with $\kappa_n/\sqrt{n} \to 0$ (e.g., $\kappa_n = \sqrt{\log \log n}$), let

$$\xi_n(\theta) := \kappa_n^{-1} \, \widehat{D}_n^{-1/2}(\theta) \, \sqrt{n} \, \bar{m}_n(\theta),$$

an inconsistent estimator of h_1. Then construct

$$\widehat{h}_{1,j} := \infty \cdot \mathbf{1}\left(\xi_{n,j}(\theta_0) > 1\right).$$

That is, $\widehat{h}_{1,j} = \infty$ if the j-th moment "looks slack" (large ξ), else 0. Use this $\widehat{h}_1$ as the localization parameter for the critical value.

Why it works. For genuinely binding moments (small true $h_{1,j}$), $\xi_{n,j}$ stays small and $\widehat{h}_{1,j} = 0$. For genuinely slack moments (large true $h_{1,j}$), $\xi_{n,j}$ grows and $\widehat{h}_{1,j} = \infty$, removing that moment from the critical value calculation. The result is asymptotically valid *and* less conservative than worst-case.

Bonferroni-Style Critical Values

Idea. Build a $(1-\beta)$-confidence region $\mathrm{CR}_{h_1,n}$ for the unknown h_1, then take the worst-case critical value over that region:

$$\mathrm{cv}(1-\alpha) = \sup_{\bar{h}_1 \in \mathrm{CR}_{h_1,n}(1-\beta)} c_{(\bar{h}_1, \widehat{h}_{2,n})}(1-\alpha).$$

Bonferroni adjustment. The naive version (using $\delta = \alpha$) gives asymptotic size $\leq \alpha+\beta$. To get true size α, use $\delta = \alpha - \beta$ (assuming $\alpha > \beta$). The Bonferroni inequality:

$$P(T_n > \mathrm{cv}(1-\delta)) \leq P(T_n > c_{(h_1, \widehat{h}_{2,n})}(1-\delta)) + P(h_1 \notin \mathrm{CR}_n) \to \delta + \beta = \alpha.$$

Power comparison. The $\mathrm{CR}_{h_1,n}$ is typically a strict subset of the parameter space, so the Bonferroni critical value is often *smaller* than the worst-case plug-in critical value. Hence Bonferroni gives a more powerful test.

8.7 Cheat Sheet

AsyCS Cheat Sheet

Definition: $\mathrm{AsySz}(\theta_0) = \limsup_n \sup_\gamma P_{\theta_0,\gamma}(T_n > c_{1-\alpha})$. Sup before limsup.
Pointwise vs uniform: pointwise NRP $= \alpha$ does not imply AsySz $= \alpha$. Uniform requires the limit to be reached at the same rate across the entire parameter space.
Canonical example: Linear IV t-test with weak IV (Dufour 1997). AsySz $= 1$ regardless of nominal α.
Drifting sequence: $\sqrt{n}\,\gamma_{n,1} \to h_1$, $\gamma_{n,2} \to h_2$. Limit distribution J_h depends on h.
AsySz formula: $\sup_{h \in H}(1 - J_h(c_{1-\alpha}))$. Worst case over the localization parameter space.
Worst-case plug-in critical value: $c_{(0,\widehat{\gamma}_{2,n})}(1-\alpha)$. Conservative.
GMS (Andrews–Soares): $\widehat{h}_{1,j} = \infty\,\mathbf{1}(\xi_{n,j} > 1)$ with $\xi_n = \kappa_n^{-1}\widehat{D}_n^{-1/2}\sqrt{n}\,\bar{m}_n$. Less conservative.
Bonferroni: $(1-\beta)$-CR for h_1, sup critical value over CR, use $\delta = \alpha - \beta$ for size.

8.8 Self-Test

Example (Self-Test 1: AsySz vs pointwise).

Define AsySz and explain in 4 sentences why pointwise NRP $= \alpha$ does not imply AsySz $= \alpha$.

Solution.

Definition. $\text{AsySz}(\theta_0) = \limsup_n \sup_{\gamma \in \Gamma} P_{\theta_0,\gamma}(T_n > c_{1-\alpha})$.

Order matters. The supremum over γ is taken before the limit, so AsySz captures finite-sample worst-case rejection.

Pointwise weaker. Pointwise NRP says that for each fixed γ, the rejection probability eventually approaches α, but the sample size at which this happens may depend on γ.

Failure mode. If the limit distribution depends discontinuously on γ, sequences γ_n approaching the discontinuity at the right rate can keep the rejection probability far from α for every n, e.g., the linear IV t-test under weak instruments (Dufour 1997).

Example (Self-Test 2: Why does Dufour's result hold?).

Briefly: why does the IV t-test have AsySz $= 1$?

Solution.

The limit distribution of the t-statistic depends discontinuously on the instrument strength π at $\pi = 0$. The standard $N(0,1)$ critical value is correct for fixed $\pi \neq 0$ but wrong along sequences $\pi_n \to 0$ at rate $n^{-1/2}$. Along these sequences the limit is a heavy-tailed ratio-of-normals, and the rejection probability can approach 1.

Chapter 9

Ridge, Lasso, and Thresholding

Status Update: This Chapter Is No Longer Optional

April 28 update. Patrik announced that final-exam HW questions will be drawn from **HW5–HW10**. HW7 is entirely on Ridge / Lasso / soft-hard thresholding. Therefore this chapter is now a *must-cover* topic.

Targets:

- Ridge bias and covariance formulas (HW7 Q1): high probability of clean derivation question.
- Sub-Gaussian inequality and its consequences for $\mathbb{E}(X)$ and $\mathrm{Var}(X)$ (HW7 Q2): clean derivation.
- Hard vs soft thresholding via ℓ_0/ℓ_1 penalties (HW7 Q3): closed-form derivation.
- Lasso oracle rate / sparsity intuition: lower priority but worth recognizing.

These are clean closed-form questions — exactly the kind of Q2/Q3 material Patrik likes.

The Story (Read This First, No Math)

The estimators in this chapter all answer the same practical question: *what do you do when OLS breaks?*

OLS breaks in two related ways:

- $X'X$ **is nearly singular** (collinearity, p close to n). Then $(X'X)^{-1}$ has huge eigenvalues; the OLS estimator is unbiased but its variance explodes. Standard errors become uninformative; predictions become wild.
- $p > n$ (more regressors than observations, "high-dimensional"). Then $X'X$ is genuinely singular; OLS is not even defined.

Two classical fixes correspond to two penalties added to the OLS objective:

- **Ridge** (penalty $\lambda\|\beta\|_2^2$). Adds λI_p to $X'X$ before inverting — guarantees invertibility and shrinks every coefficient toward zero. *Smooths* the OLS estimator. Trades a little bias for a big variance reduction. But every $\widehat{\beta}_j$ stays nonzero — ridge does not select variables.
- **Lasso** (penalty $\lambda\|\beta\|_1$). Same idea, but the ℓ_1 penalty has a sharp corner at zero, so the optimum drives some coefficients exactly to zero. *Selects* variables. Useful when you believe the truth is sparse (only a few of the p regressors actually matter).

The orthogonal-design version ($X = I$) gives clean closed forms: ridge becomes simple shrinkage, lasso becomes *soft thresholding*. A cousin, *hard thresholding*, sets small entries to zero but leaves large ones alone — this corresponds to the (nonconvex) ℓ_0 penalty.

Last piece of background: *sub-Gaussian* random variables. For Lasso theory you need to control the maximum, over p coordinates, of a sample average of errors. A union bound delivers $\left\|n^{-1}X'e\right\|_\infty = O_p(\sqrt{\log p/n})$ *provided* the errors have light enough tails. "Sub-Gaussian" is the standard tail condition that makes this work. HW7 Q2 walks you through what sub-Gaussian implies.

Remark (Map to the HW7 questions).

> HW7 Q1 = Ridge bias/variance (Section 9.1). HW7 Q2 = Sub-Gaussian implies $\mathbb{E}(X) = \mu$, $\mathrm{Var}(X) \le \sigma^2$ (Section 9.2). HW7 Q3 = Hard vs soft thresholding from ℓ_0/ℓ_1 penalties (Section 9.3). The other Q (simulation comparing Lasso/Ridge) is Section 9.4.

9.1 Ridge Estimator: Bias and Variance (HW7 Q1)

Why ridge in the first place? The OLS estimator $\widehat{\beta}_{\text{OLS}} = (X'X)^{-1}X'Y$ is unbiased and has variance $(X'X)^{-1}X'DX(X'X)^{-1}$. Two failure modes motivate a different estimator:

- **Multicollinearity.** If columns of X are nearly linearly dependent, the smallest eigenvalue of $X'X$ is tiny, so $(X'X)^{-1}$ has a huge eigenvalue, so the variance is huge. OLS becomes unstable: small changes in the data produce huge changes in $\widehat{\beta}$.
- **High dimension.** If $p > n$, the $p \times p$ matrix $X'X$ has rank at most $n < p$ and is singular. OLS is undefined.

Ridge adds λI_p to $X'X$ before inverting, which (i) guarantees $X'X + \lambda I_p$ is positive definite (eigenvalues bounded below by $\lambda > 0$), so the inverse exists, and (ii) shrinks the eigenvalues of the inverse, deflating the variance. The price is bias: $\widehat{\beta}_{\text{ridge}}$ is no longer centered at β. The bias-variance trade-off (Theorem 9.1 below) makes this trade explicit.

Remark (Equivalent characterization as a penalized objective).

> Ridge has a clean optimization interpretation: $\widehat{\beta}_{\text{ridge}} = \arg\min_\beta \left\{\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2\right\}$. Differentiating and setting to zero gives the FOC $-2X'(Y - X\beta) + 2\lambda\beta = 0$, which rearranges to $(X'X + \lambda I_p)\beta = X'Y$. So the closed-form estimator below is the FOC of a penalized least-squares problem.

Setup. Linear regression $Y = X\beta + e$ with $\beta \in \mathbb{R}^p$, $\mathbb{E}(e \mid X) = 0$, i.i.d. observations. Define

$$D := \mathbb{E}(ee' \mid X) \in \mathbb{R}^{n \times n}.$$

The *ridge estimator* with fixed penalty $\lambda > 0$:

$$\widehat{\beta}_{\text{ridge}} := (X'X + \lambda I_p)^{-1} X'Y.$$

Theorem 9.1: Ridge Bias and Variance (HW7 Q1)

$$\begin{aligned}\text{bias}(\widehat{\beta}_{\text{ridge}} \,|\, X) &= -\lambda (X'X + \lambda I_p)^{-1} \beta, \\ \text{Var}\left(\widehat{\beta}_{\text{ridge}} \,|\, X\right) &= (X'X + \lambda I_p)^{-1} (X'DX) (X'X + \lambda I_p)^{-1}.\end{aligned}$$

Proof Template (HW7 Q1)

[REPRODUCE — memorize this proof]

Step 1: Decompose $\widehat{\beta}_{\text{ridge}} - \beta$.

$$\begin{aligned}\widehat{\beta}_{\text{ridge}} - \beta &= (X'X + \lambda I_p)^{-1} X'Y - \beta \\ &= (X'X + \lambda I_p)^{-1} \left[X'Y - (X'X + \lambda I_p)\beta\right] \\ &= (X'X + \lambda I_p)^{-1} (X'e - \lambda\beta),\end{aligned}$$

using $X'Y = X'X\beta + X'e$ and rearranging.

Step 2: Bias. Take conditional expectation:

$$\text{bias} = (X'X + \lambda I_p)^{-1} \mathbb{E}(X'e - \lambda\beta \,|\, X) = (X'X + \lambda I_p)^{-1}(-\lambda\beta),$$

using $\mathbb{E}(e \,|\, X) = 0$.

Step 3: MSE.

$$\text{MSE}(\widehat{\beta}_{\text{ridge}} \,|\, X) = (X'X + \lambda I_p)^{-1} \mathbb{E}((X'e - \lambda\beta)(X'e - \lambda\beta)' \,|\, X) \, (X'X + \lambda I_p)^{-1}.$$

Expand the inner expectation:

$$\mathbb{E}\left(X'ee'X - \lambda X'e\beta' - \lambda\beta e'X + \lambda^2 \beta\beta' \,|\, X\right) = X'DX + \lambda^2 \beta\beta',$$

again using $\mathbb{E}(e \,|\, X) = 0$.

Step 4: Variance = MSE − bias bias′.

$$\begin{aligned}\text{Var}\left(\widehat{\beta}_{\text{ridge}} \,|\, X\right) &= \text{MSE} - \text{bias} \cdot \text{bias}' \\ &= (X'X + \lambda I_p)^{-1} \left[X'DX + \lambda^2\beta\beta' - \lambda^2\beta\beta'\right] (X'X + \lambda I_p)^{-1} \\ &= (X'X + \lambda I_p)^{-1} (X'DX) (X'X + \lambda I_p)^{-1}. \qquad \blacksquare\end{aligned}$$

Remark.

Bias-variance trade-off intuition. As $\lambda \to 0$, ridge reduces to OLS: bias vanishes but variance grows when $X'X$ is ill-conditioned. As $\lambda \to \infty$, $\widehat{\beta}_{\text{ridge}} \to 0$: bias dominates. Cross-validation chooses λ to balance.

9.2 Sub-Gaussian Random Variables (HW7 Q2)

Definition 9.2: Sub-Gaussian Property

A random variable X is *sub-Gaussian with parameter* σ^2 if for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}(\exp(\lambda(X-\mu))) \leq \exp(0.5\,\lambda^2\sigma^2).$$

Remark (Why sub-Gaussian instead of Gaussian).

Saying "X is Gaussian" is restrictive. "Sub-Gaussian" weakens the requirement: X's moment generating function is *bounded above* by that of a normal. This includes Gaussian, bounded random variables (which are automatically sub-Gaussian), Bernoulli, and many empirical errors — but excludes heavy-tailed things like Cauchy or t_2. The reason we care: the sub-Gaussian inequality directly delivers exponential tail bounds $P(|X-\mu|>t) \leq 2\exp(-t^2/(2\sigma^2))$, which combined with a union bound over p coordinates gives $\left\|n^{-1}X'e\right\|_\infty = O_p(\sqrt{\log p/n})$ — the rate that makes Lasso oracle theory work.

This is the key concentration property used in Lasso theory: sub-Gaussian errors give exponential tail bounds.

Theorem 9.3: Sub-Gaussian Implies Mean and Variance Bounds (HW7 Q2)

If X is sub-Gaussian with parameter σ^2 around μ:

(a) $\mathbb{E}(X) = \mu$.

(b) $\mathrm{Var}(X) \leq \sigma^2$.

(c) If σ^2 is the smallest such parameter, $\mathrm{Var}(X)$ may be *strictly less* than σ^2.

Proof of (a) (HW7 Q2(a))

[REPRODUCE — memorize this proof]

Let $g(\lambda) := \exp(0.5\lambda^2\sigma^2 + \lambda\mu)$ and $M(\lambda) := \mathbb{E}(\exp(\lambda X))$ (the moment generating function). Note $g(0) = M(0) = 1$ and $M(\lambda) \leq g(\lambda)$ for all λ.
Claim: $g'(0) = M'(0)$.
Proof by contradiction: if $M'(0) \neq g'(0)$, then $\theta := \lim_{\lambda\to 0}(M(\lambda)-g(\lambda))/\lambda \neq 0$.

- If $\theta > 0$, taking $\lambda \to 0^+$ gives $\lim_{0<\lambda\to 0}(M(\lambda)-g(\lambda))/\lambda > 0$, contradicting $M(\lambda)-g(\lambda) \leq 0$.
- If $\theta < 0$, taking $\lambda \to 0^-$ gives a similar contradiction.

Hence $M'(0) = g'(0)$, i.e., $\mathbb{E}(X) = \mu$. ■

Proof of (b) (HW7 Q2(b))

[REPRODUCE — memorize this proof]

Expand both sides of the sub-Gaussian inequality in Taylor series:

$$M(\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^n \mathbb{E}((X-\mu)^n)}{n!} \leq \exp(0.5\lambda^2\sigma^2) = \sum_{n=0}^{\infty} \frac{\lambda^{2n}\sigma^{2n}}{2^n n!}.$$

Since $\mathbb{E}(X-\mu) = 0$ by part (a), the $n=1$ term on the LHS vanishes. The $n=0$ terms cancel. So

$$\sum_{n=2}^{\infty} \frac{\lambda^n \mathbb{E}((X-\mu)^n)}{n!} \leq \sum_{n=1}^{\infty} \frac{\lambda^{2n}\sigma^{2n}}{2^n n!}.$$

Divide by $\lambda^2 \neq 0$ and let $\lambda \to 0$: the leading terms are

$$\frac{\mathbb{E}\big((X-\mu)^2\big)}{2} + O(\lambda) \leq \frac{\sigma^2}{2} + O(\lambda^2).$$

Hence $\mathrm{Var}(X) = \mathbb{E}\big((X-\mu)^2\big) \leq \sigma^2$. ■

Remark.

> Part (c): the inequality can be strict. Take $X = 1-p$ w.p. p and $X = -p$ w.p. $1-p$ (Bernoulli centered). Then $\mathbb{E}(X) = 0$ and $\mathrm{Var}(X) = p(1-p)$. For $p = 1/4$: $\mathrm{Var}(X) = 3/16 = 0.1875$. The smallest sub-Gaussian σ^2 is computed numerically as $\geq 0.1878 > \mathrm{Var}(X)$.

9.3 Hard vs Soft Thresholding (HW7 Q3)

Why care about thresholding? If the truth is *sparse* — only a few of the p candidate regressors actually matter — then the right answer is to *set the others to exactly zero*, not just shrink them small. Ridge cannot do this: its closed form $\widehat{\beta}_j = (\text{scalar shrinkage}) \cdot \widehat{\beta}_j^{OLS}$ never produces a clean zero. Thresholding is the simplest device that does. Two natural choices:

- **Hard threshold**: kill anything below the cutoff λ, keep everything above unchanged. "If $|y| < \lambda$, set to 0; else keep y."
- **Soft threshold**: kill anything below λ, and *also shrink* larger entries by λ toward zero. "$\mathrm{sgn}(y)(|y| - \lambda)_+$."

The miraculous fact (Theorems 9.3 and 9.3 below) is that these two informal recipes are exactly the solutions of two penalized minimization problems:

- Hard $\Leftrightarrow$ ℓ_0-penalty, which counts the number of nonzero coefficients. Combinatorial, nonconvex.
- Soft $\Leftrightarrow$ ℓ_1-penalty, which is the Lasso. Convex, computable in polynomial time.

This connection is the engine that makes Lasso theory go: the ℓ_1 penalty is a *convex relaxation* of the (intractable) ℓ_0 penalty, and in the orthogonal-design case the two relaxations differ only by an additive shrinkage term.

Setup. Consider the model $y = \theta + w$ for $y \in \mathbb{R}^n$, with a fixed threshold $\lambda > 0$.

Theorem 9.4: Hard Thresholding via ℓ_0 Penalty (Non-convex)

The *hard-thresholding estimator*

$$\widehat{\theta}_i^H(y) := \begin{cases} y_i & \text{if } |y_i| \geq \lambda, \\ 0 & \text{otherwise} \end{cases}$$

solves the (**non-convex**) optimization

$$\min_{\theta \in \mathbb{R}^n} 0.5 \, \|y - \theta\|_2^2 + 0.5\,\lambda^2 \, \|\theta\|_0 \,,$$

where $\|\theta\|_0 := \#\{i : \theta_i \neq 0\}$.

Theorem 9.5: Soft Thresholding via ℓ_1 Penalty = Lasso (Convex)

The *soft-thresholding estimator*

$$\widehat{\theta}_i^S(y) := \operatorname{sgn}(y_i)\,(|y_i| - \lambda)_+$$

solves the **convex** optimization

$$\min_{\theta \in \mathbb{R}^n} 0.5 \, \|y - \theta\|_2^2 + \lambda \, \|\theta\|_1 \,.$$

This is the orthogonal-design Lasso problem.

Derivation of Soft Thresholding (HW7 Q3)

[REPRODUCE — memorize this proof]

Reduction to scalar. Both objectives are additively separable, so consider $n = 1$. Minimize $c(\theta) = 0.5(y - \theta)^2 + \lambda|\theta|$.
For $\theta \neq 0$, $c'(\theta) = -(y - \theta) + \lambda \operatorname{sgn}(\theta)$. Three cases:
Case 1: $|y| < \lambda$. If $\theta > 0$: $c'(\theta) = \theta - y + \lambda > 0$. If $\theta < 0$: $c'(\theta) = \theta - y - \lambda < 0$. Moving away from $\theta = 0$ in either direction increases c. By continuity of c, the minimum is at $\theta = 0$.
Case 2: $y \geq \lambda > 0$. The slope $c'(\theta)$ is negative for $\theta < y - \lambda$ and positive for $\theta > y - \lambda$. So c is minimized at $\theta = y - \lambda = \operatorname{sgn}(y)(|y| - \lambda)$.
Case 3: $y \leq -\lambda$. Analogous: c minimized at $\theta = y + \lambda = \operatorname{sgn}(y)(|y| - \lambda)$.
Combining: $\widehat{\theta}^S(y) = \operatorname{sgn}(y)(|y| - \lambda)_+$. ■
Hard thresholding (sketch). For $n = 1$: at $\theta = 0$, $c(0) = 0.5y^2$. At $\theta = y$,

$c(y) = 0.5\lambda^2$. So the minimum is $\theta = 0$ if $|y| < \lambda$ (giving $0.5y^2 < 0.5\lambda^2$), and $\theta = y$ if $|y| > \lambda$.

Remark.

Convexity. The soft objective is convex (sum of two convex terms). The hard objective is non-convex — the line segment from $(0, 0.5y^2)$ to $(\theta^*, 0.5\lambda^2)$ for any θ^* on the curve lies *below* the curve (the function value is $0.5(y-\theta)^2$ between, which can dip below the line). Lasso (soft) is convex and computable; hard thresholding requires combinatorial search.

9.4 Lasso Theory: Oracle Rate (Lec 25 Brief)

For high-dimensional regression $Y = X\beta_0 + e$ with $p \gg n$, suppose β_0 has only s nonzero components (*sparsity*). The Lasso estimator

$$\widehat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \left\{ n^{-1} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

under BRT (Bickel–Ritov–Tsybakov) regularity conditions (sub-Gaussian errors + restricted eigenvalue) achieves the *oracle rate*:

$$\left\| \widehat{\beta}_{\text{Lasso}} - \beta_0 \right\|_2 = O_p\left(\sqrt{s \log p / n} \right).$$

Lemma 1 (union bound). A key intermediate result:

$$\left\| n^{-1} X' e \right\|_\infty = O_p\left(\sqrt{\log p / n} \right).$$

Proof uses the sub-Gaussian property and a union bound over p coordinates — this is where Section 9.6 matters.

Remark.

Comparison Lasso vs Ridge in simulations (HW7 Q4):

- Lasso correctly assigns $\theta_i = 0$ for genuinely zero coefficients (oracle property).
- Ridge always gives nonzero estimates — no variable selection.
- Lasso usually wins in ℓ_2 loss.
- Exception: when small nonzero coefficients are mistakenly thresholded to zero, Ridge can outperform.

9.5 Cheat-Sheet Summary

Ridge, Lasso, Thresholding Cheat Sheet

Ridge bias: $-\lambda(X'X+\lambda I)^{-1}\beta$.
Ridge variance: $(X'X+\lambda I)^{-1}X'DX(X'X+\lambda I)^{-1}$.
Sub-Gaussian inequality: $\mathbb{E}(\exp(\lambda(X-\mu))) \leq \exp(0.5\lambda^2\sigma^2)$.
Sub-Gaussian implies: $\mathbb{E}(X)=\mu$ and $\mathrm{Var}(X)\leq\sigma^2$.
Soft thresholding (Lasso, convex): $\widehat{\theta}^S = \mathrm{sgn}(y)(|y|-\lambda)_+$ from ℓ_1-penalty.
Hard thresholding (non-convex): $\widehat{\theta}^H = y\cdot\mathbf{1}(|y|\geq\lambda)$ from ℓ_0-penalty.
Lasso oracle rate: $\left\|\widehat{\beta}-\beta_0\right\|_2 = O_p\left(\sqrt{s\log p/n}\right)$ under BRT and sparsity.

9.6 Self-Test Problems

Example (Self-Test 1: Ridge bias-variance from scratch).

Reproduce the proof of Theorem 9.1 (Ridge bias and variance) without looking. Time yourself: should take less than 10 minutes.

Example (Self-Test 2: Sub-Gaussian implies mean).

Why does $\mathbb{E}(\exp(\lambda(X-\mu))) \leq \exp(0.5\lambda^2\sigma^2)$ for all λ force $\mathbb{E}(X)=\mu$?

Solution.

The inequality is between two MGFs that agree at $\lambda=0$. If $\mathbb{E}(X)\neq\mu$, then $M'(0)=\mathbb{E}(X-\mu)+\mu-\mu=\mathbb{E}(X)-\mu\neq 0=g'(0)$. By a contradiction argument considering left/right limits, this violates the inequality near 0.

Example (Self-Test 3: Soft vs hard thresholding).

Show that the soft-thresholding estimator is $\mathrm{sgn}(y)(|y|-\lambda)_+$ by minimizing $0.5(y-\theta)^2+\lambda|\theta|$ over θ.

Solution.

Three cases: $|y|<\lambda$ gives $\theta=0$; $y\geq\lambda$ gives $\theta=y-\lambda$; $y\leq-\lambda$ gives $\theta=y+\lambda$. Combine using sgn and $(\cdot)_+$.

Chapter 10

Invariant Tests (Skip)

Why You Should Skip This Chapter

Invariant tests have never appeared on a midterm. The lecture notes (`InvariantTestsLectureNotes.pdf`) are dense and the topic is highly technical (Hunt–Stein theorem, maximal invariant statistics, group-theoretic optimality).

Estimated probability of appearing on final: very low ($< 5\%$).

Recommendation: *do not study* this for the final. If anything related to invariance comes up, it will be inside the Weak IV chapter (Andrews–Moreira–Stock 2006 invariance argument for the CLR test, which is already partially covered in Chapter 5). For your purposes: this chapter is a placeholder. Allocate the time you would have spent here to Chapters 1 and 2.

Chapter 11

HW5–HW10 Problem-and-Answer Compendium

Why This Chapter Exists and How to Use It

Patrik announced that final-exam questions will be *drawn from HW5–HW10*. So this is the single most actionable chapter in the book.

Format for each problem. Every question is presented in four parts:

- **Problem (verbatim)**: the problem as Patrik wrote it. Read this first.
- **What this tests**: 1–3 sentences on which concept the question targets and how it relates to the textbook.
- **Approach**: the high-level strategy — how to think about the problem before diving into algebra.
- **Solution**: every step labelled with *why we do it this way*.

Tag system. Each problem has a priority tag. With 4 days to the exam, drilling everything is impossible; drilling the high-ROI ones and skipping the low-ROI ones is.

11.1 Tag System

- **[CORE — must master]**

  — exactly the kind of problem Patrik likes for the final. Reproduce the proof structure verbatim.

- **[MEMORIZE — not in main text, just remember the formula]**

— the result is testable but the derivation is not in the textbook body.

- **[EXAM-WRITE — full proof too hard, but write *this* much for credit]**

  — the full derivation is technical/long. Write the abbreviated version shown.

- **[LOW ROI — do not study]**

  — low probability of being tested in derivation form.

- **[REFERENCE ONLY]**

  — the official solution is just a pointer to a paper.

11.2 HW5: Hausman Pretest, AsyCS, GMS, Sufficient Statistics

11.2.1 HW5 Q1: Simulations Involving a Hausman Pretest

[LOW ROI — simulation; only the conclusion matters]

Problem. In the model of problem set 4 question 2, using the same notation as there, assume the objective is to test the null hypothesis $H_0 : \theta = \theta_0$ (against a one-sided or two-sided alternative). Define the t-test statistic

$$T_l^*(\theta_0) = n^{1/2}(\widehat{\theta}_l - \theta_0)/\widehat{V}_l^{1/2} \tag{1}$$

for $l =$ OLS and 2SLS. Often applied researchers first employ a Hausman pretest in order to determine whether (1) should be tested based on a t-test based on OLS or 2SLS. The definition of the two-stage test statistic is

$$T_n^*(\theta_0) = T_{\text{OLS}}^*(\theta_0)\,\mathbf{1}(H_n \leq \chi^2_{1,1-\beta}) + T_{\text{2SLS}}^*(\theta_0)\,\mathbf{1}(H_n > \chi^2_{1,1-\beta}), \tag{2}$$

where β is the nominal size of the Hausman pretest, $\mathbf{1}$ is the indicator function, and $\chi^2_{1,1-\beta}$ is the $1-\beta$ quantile of a chi-square random variable with one degree of freedom. Define the two-stage test statistic $T_n(\theta_0)$ as $T_n^*(\theta_0)$ or $|T_n^*(\theta_0)|$ depending on whether the test is upper one-sided or symmetric two-sided. The nominal size α test rejects H_0 if

$$T_n(\theta_0) > c_\infty(1-\alpha), \tag{3}$$

where $c_\infty(1-\alpha) = z_{1-\alpha}$ and $z_{1-\alpha/2}$ for the upper one-sided and symmetric two-sided test, respectively. Perform finite-sample simulations of the null rejection probability of the two-stage test using the specifications: $\alpha = \beta = 5\%$, $n = 100, 500$, $k = 5$, $(u_i, v_i, Z_i) \sim N(0, V)$, where V equals the $k+2$-dimensional identity matrix except that the $(1,2)$ and $(2,1)$ elements equal $h_1/n^{1/2}$ for $h_1 = 0, 3, 8$. Take the reduced-form coefficient vector π as a k-vector of ones multiplied by h_2 for $h_2 = 0.1, 1, 10$, and set $\theta = 0$. Discuss your findings. What would your recommendation be to applied researchers that follow this two-stage practice?

What this tests. A Monte Carlo investigation of *why the Hausman pretest fails under weak IVs*. The key concept being tested: the noncentral $\chi^2_1(h_1^2 h_2^2/(h_2^2+1))$ limit of the Hausman statistic, and what happens when noncentrality is small (= the test has low power, so OLS is incorrectly chosen even when endogeneity is severe). Connects to Section 5.6 of Chapter 5.

Approach. You don't need the simulation numbers themselves — you need the qualitative finding. Chain of reasoning: (i) write the noncentrality from PS4; (ii) note it's small when h_2 is small (weak IVs) *or* h_1 is small (no endogeneity); (iii) when h_2 small but h_1 large, pretest has low power, picks OLS, but OLS is biased — so the two-stage test rejects too often; (iv) recommend weak-IV-robust tests instead.

Solution.

Step 1. Recall from PS4 that under "weak endogeneity,"

$$H_n \xrightarrow{d} \chi^2_1\left(h_1^2 h_2^2/(h_2^2+1)\right).$$

Noncentrality is small when h_1^2 *or* h_2^2 is small. *Why this matters.* Hausman's test has low power exactly when noncentrality is small.

Step 2. In the regime h_2 small (weak IVs), h_1 large (real endogeneity), noncentrality is still small (because $h_2^2 \ll 1$ pulls it down), so Hausman fails to reject exogeneity — and the two-stage test then uses OLS, which is biased. *Why OLS is biased.* h_1 encodes the structural-error vs reduced-form-error correlation; $h_1 \neq 0$ means OLS converges to a wrong value. The standard t-test on a biased estimator over-rejects.

Step 3. Simulation: with $h_2 = 0.1$ (weak), $h_1 \in \{3, 8\}$, NRP for $n = 100$ is $75.2\%, 54.7\%$; for $n = 500$ is $68.1\%, 40.4\%$ — far above the nominal 5%.

Recommendation. Do not use the Hausman pretest; use weak-IV-robust tests (AR, LM_{CUE}, CLR), valid regardless of instrument strength.

11.2.2 HW5 Q2: Asymptotic Distribution + AsySz of the Two-Stage Test

[EXAM-WRITE — state the limits, cite Guggenberger 2009 for AsySz]

Problem. Reconsider the setup in problem set 4, question 2 about the Hausman statistic. It was shown there that when $n^{1/2}\text{Corr}_{F_n}(u_i, v_i) \to h_1$ for finite h_1, and $\|E_{F_n} Z_i Z_i'^{1/2} \pi_n/\sigma_{vn}\| \to h_2$,

$$H_n \xrightarrow{d} \eta_h \sim (1+h_2^2)[s_k'\psi_{u,0} - h_2(1+h_2^2)^{-1}\xi_{2,h}]^2 \sim \chi^2_1(h_1^2 h_2^2 (h_2^2+1)^{-1}),$$

and when $|h_1| = \infty$ the Hausman statistic diverges to infinity. In question 1 we consider tests of $H_0 : \theta = \theta_0$ using the two-stage test statistic $T_n^*(\theta_0)$ based on $T_l^*(\theta_0)$.

(a) Show that $T^*_{\text{OLS}}(\theta_0) \xrightarrow{d} (1+h_2^2)^{-1/2}\xi_{2,h}$ and $T^*_{\text{2SLS}}(\theta_0) \xrightarrow{d} s_k'\psi_{u,0}$.

(b) Using the asymptotic size formula derived in class and the results above, write down a formula for the asymptotic size of the upper one-sided and symmetric two-sided test and then "calculate" the asymptotic size by simulation (using 20,000 simulation draws for each value of (h_1, h_2) on a fine grid in H). Take $\alpha = \beta = 5\%$, parameter space $H = [-\infty, \infty] \times [\kappa, \overline{\kappa}]$ with $\overline{\kappa} = 1000$ and five choices for κ: 0.0001, 0.1, 0.5, 1, 2, 10. Provide a verbal description of what h_1 and h_2 measure.

What this tests. Two skills: (a) deriving limit distributions of T^*_{OLS} and T^*_{2SLS} under weak-IV asymptotics; (b) applying the AsySz formula from Chapter 8 to the two-stage test; understanding the meaning of h_1 (degree of endogeneity) and h_2 (instrument strength).

Approach. (a) Both limits come from PS4's joint limit theorem. T^*_{OLS} inherits the OLS bias under endogeneity (the $\xi_{2,h}$ piece, with $(1+h_2^2)^{-1/2}$ factor from variance estimator behavior); T^*_{2SLS} becomes a non-normal weak-IV ratio. (b) Write rejection probability as $\sup_h$. *Skip simulation; cite Guggenberger 2009.*

Solution.

(a) Limits. Under weak-IV asymptotics with h_1, h_2 finite:

$$T^*_{\text{OLS}}(\theta_0) \xrightarrow{d} (1+h_2^2)^{-1/2}\xi_{2,h}, \quad T^*_{\text{2SLS}}(\theta_0) \xrightarrow{d} s_k'\psi_{u,0}.$$

Why. The numerator of T^*_{OLS} is $\sqrt{n}(\widehat{\theta}_{\text{OLS}} - \theta_0) = \sqrt{n}\cdot(\text{OLS bias}) + o_p(1)$; the denominator (OLS standard error) under weak IVs has factor $(1+h_2^2)^{1/2}$ (the noise-variance estimator absorbs first-stage variance dependence). Ratio is $\xi_{2,h}/(1+h_2^2)^{1/2}$. For T^*_{2SLS}, both numerator and denominator are non-normal; AR-style projection $s_k'\psi_{u,0}$ survives.

(b) AsySz formula. Upper one-sided:

$$\text{AsySz} = \sup_{h\in H}\left[1 - J_h(z_{1-\alpha})\right], \quad J_h := \text{CDF of } T_n^* \text{ under } h.$$

Two-sided: replace $z_{1-\alpha}$ with $z_{1-\alpha/2}$ and use the absolute statistic. *Why this formula.* Standard AsySz definition (Section 8.1): sup over h first, then lim sup. Under fixed h, statistic limit is J_h; rejection probability is $1 - J_h(c_{1-\alpha})$. Worst case is the asymptotic size.

Simulation result. Guggenberger (2009, Table 1, p. 377): AsySz substantially exceeds 5% for all κ. Two-stage test does **not** control asymptotic size.

Verbal description.

- $h_1 = \lim n^{1/2}\text{Corr}_{F_n}(u_i, v_i)$: *degree of endogeneity*, on the $\sqrt{n}$ scale.
- $h_2 = \lim \|EZZ'^{1/2}\pi/\sigma_v\|$: *strength of instruments.*

Why these scalings. $\sqrt{n}$ for h_1 matches Pitman drift (Chapter 4); h_2 is bounded but possibly small (weak IVs).

11.2.3 HW5 Q3: Andrews–Soares (2010) GMS Critical Values

[EXAM-WRITE — state the result, write the 4-step argument]

Problem. In the model defined by moment inequalities/equalities

$$E_{F_0}m_j(W_i,\theta_0) \geq 0 \text{ for } j = 1,\ldots,p \text{ and } E_{F_0}m_j(W_i,\theta_0) = 0 \text{ for } j = p+1,\ldots,p+v, \quad (5)$$

define $\gamma_1 = (\gamma_{1,1},\ldots,\gamma_{1,p})' \in \mathbb{R}^p_+$, where

$$\gamma_{1,j} = \sigma_{F,j}^{-1}(\theta)E_F m_j(W_i,\theta) \text{ for } j = 1,\ldots,p, \quad (6)$$

$\sigma_{F,j}(\theta) = \text{Var}(m_j(W_i,\theta))^{1/2}$, F is the distribution of the data. Let $\gamma_2 = \Omega = \text{Corr}_F(m(W_i,\theta))$

denote the $k \times k$ correlation matrix, $\gamma_3 = (F, \theta)$. Recall

$$T_n(\theta) = S(n^{1/2}\bar{m}_n(\theta), \widehat{\Sigma}_n(\theta))$$

with S and $\widehat{\Sigma}_n(\theta) = n^{-1}\sum_{i=1}^n (m(W_i,\theta) - \bar{m}_n(\theta))(m(W_i,\theta) - \bar{m}_n(\theta))'$ as in lecture, S satisfying (i) non-increasing in m_1 and (ii) $S(\Delta m, \Delta\Sigma\Delta) = S(m,\Sigma)$ for diagonal pd Δ.

(a) In class we showed under sequences $\{\gamma_{n,h} : n \geq 1\}$ such that $n^{1/2}\gamma_{n,h,1} \to h_1 \in [0,\infty]^p$, $\gamma_{n,h,2} \to h_2$:

$$T_n(\theta_{n,h}) \xrightarrow{d} S(h_2^{1/2}Z + (h_1, 0_v), h_2),$$

where $Z \sim N(0, I_k)$. The argument was sloppy when components of h_1 equal infinity (CMT not directly applicable). Provide a solid argument.

(b) Andrews–Soares (2010): for $\widehat{D}_n(\theta) := \text{Diag}(\widehat{\Sigma}_n(\theta))$, define

$$\xi_n(\theta) := \kappa_n^{-1}\widehat{D}_n^{-1/2}(\theta)n^{1/2}\bar{m}_n(\theta) \in \mathbb{R}^k$$

for some $\kappa_n \to \infty$ slower than $n^{1/2}$. The proposed critical value is $c_{(\widehat{h}_1, \widehat{\gamma}_{2,n})}(1-\alpha)$ where $\widehat{h}_{1,j} = \infty \cdot \mathbf{1}(\xi_{n,j}(\theta_0) > 1)$ (with $\infty \cdot 0 = 0$). Show that the resulting test has correct asymptotic size. Compare its power to the worst-case/plug-in test.

What this tests. Tests two skills: (a) careful CMT-with-infinite-limit argument for moment inequalities; (b) asymptotic-size proof for GMS. Section 8.6 of Chapter 8.

Approach. (a) Technical, defer to lecture-notes appendix. (b) Show $\widehat{h}_{1,j}$ correctly identifies binding moments ($h_{1,j} < \infty$, get $\widehat{h}_{1,j} = 0$) vs slack ones ($h_{1,j} = \infty$, get $\widehat{h}_{1,j} = \infty$). Then critical value at least as large as ideal $c_h(1-\alpha)$. Power gain: GMS drops genuinely-slack moments, shrinking critical value relative to worst-case plug-in.

Solution.

(a) See lecture-notes appendix on Asymptotic Size, p. 28–29 from (4.3). Issue: $S(h_2^{1/2}Z + (h_1, 0_v), h_2)$ involves ∞ inputs; CMT fails for discontinuous limits. Fix: separate continuity argument for each "$h_{1,j} = \infty$" coordinate — the moment inequality $Em_j \geq 0$ becomes "infinitely slack" and drops out by non-increasing property (i). *Skip details on the exam; cite the appendix.*

(b) Decompose $\xi_{n,j}(\theta_0)$.

$$\xi_{n,j}(\theta_0) = \underbrace{\kappa_n^{-1}\widehat{\sigma}_{n,jj}^{-1/2}n^{1/2}(\bar{m}_{n,j} - \mathbb{E}_{F_n}[m_j])}_{(A)} + \underbrace{\kappa_n^{-1}\widehat{\sigma}_{n,jj}^{-1/2}n^{1/2}\mathbb{E}_{F_n}[m_j]}_{(B)}.$$

Step 1: $(A) \xrightarrow{p} 0$**.** CLT $\Rightarrow n^{1/2}(\bar{m}_{n,j} - Em_j) = O_p(1)$; $\widehat{\sigma}_{n,jj}^{-1/2} = O(1)$; dividing by $\kappa_n \to \infty$ gives $o_p(1)$. *Why.* CLT-rate noise divided by exploding κ_n.

Step 2: (B) **behavior depends on** $h_{1,j}$**.**

- $h_{1,j} < \infty$: $n^{1/2}Em_j \to h_{1,j}$ finite, divided by $\kappa_n \to \infty$ gives 0. So $\xi_{n,j} \xrightarrow{p} 0$, $\widehat{h}_{1,j} \xrightarrow{p} 0$.
- $h_{1,j} = \infty$: $n^{1/2}Em_j \to \infty$ faster than κ_n, so $\xi_{n,j} > 1$ eventually, $\widehat{h}_{1,j} = \infty$.

Why this is the right inference. Binding $\to$ pessimistic $h_{1,j} = 0$ (conservative). Slack $\to$ drop the moment ($h_{1,j} = \infty$).

Step 3: Critical value monotonicity. $h_1 \mapsto c_{(\bar{h}_1,h_2)}(1-\alpha)$ is non-increasing in $\bar{h}_1$: as $\bar{h}_{1,j}$ goes $0 \to \infty$, the moment becomes more slack and drops out, so the $1-\alpha$ quantile shrinks.

Step 4: Conclude. $\operatorname{plim} c_{(\widehat{h}_1,\widehat{\gamma}_{2,n})}(1-\alpha) \geq c_h(1-\alpha)$ (Step 3 monotonicity + Step 2 lower bound). Test has rejection rate $\leq \alpha$ under any $\gamma_{n,h}$. AsySz controlled. ■

Power. Worst-case plug-in uses $\bar{h}_1 = 0_p$ always (most pessimistic). GMS sets $\widehat{h}_{1,j} = \infty$ for slack moments, dropping them from the critical value calculation. Smaller critical value $\Rightarrow$ higher power. *Why "at least as powerful, sometimes strictly."* On sequences with $h_1 = 0$ everywhere (all binding), GMS reduces to worst-case. Strict gain only when some moments are slack.

11.2.4 HW5 Q4: Sufficient Statistic for Bernoulli

[MEMORIZE — basic stat; remember the factorization argument]

Problem. By definition, $T(X)$ is a sufficient statistic for θ if the conditional distribution of the sample $X = x_1, \ldots, x_n$ of iid data given $T(X)$ does not depend on θ. One can show that if $p(x|\theta)$ is the joint pmf of X and $q(t|\theta)$ is the pmf of $T(X)$, then $T(X)$ is sufficient iff for every $\widetilde{X}$ in the sample space the ratio $p(\widetilde{X}|\theta)/q(T(\widetilde{X})|\theta)$ is constant in θ.

(a) Prove that statement for the discrete case.

(b) Let X be an iid Bernoulli(θ) sample. Show $T(X) = x_1 + \cdots + x_n$ is sufficient.

What this tests. Basic mathematical statistics — Fisher–Neyman factorization. Background for invariant testing (Chapter 10, low priority). Bernoulli example is standard.

Approach. (a) Use definition of sufficiency + the inclusion $\{X = x\} \subseteq \{T(X) = T(x)\}$. (b) Compute the ratio for Bernoulli; the θ exponents cancel because $\sum x_i = t$.

Solution.

(a) Sufficiency means $P_\theta(X = x \mid T(X) = t)$ doesn't depend on θ.

- If $T(x) \neq t$: probability $= 0$ trivially.
- If $T(x) = t$: $\{X = x\} \subseteq \{T(X) = T(x)\}$ (applying T to fixed input gives fixed output), so

$$P_\theta(X = x|T(X) = T(x)) = \frac{P_\theta(X = x)}{P_\theta(T(X) = T(x))} = \frac{p(x|\theta)}{q(T(x)|\theta)}.$$

θ-free conditional $\Leftrightarrow$ θ-free ratio.

(b) Bernoulli. $p(x|\theta) = \theta^{\sum x_i}(1-\theta)^{n-\sum x_i}$; $q(t|\theta) = \binom{n}{t}\theta^t(1-\theta)^{n-t}$. With $t = \sum x_i$:

$$\frac{p(x|\theta)}{q(t|\theta)} = \frac{\theta^t(1-\theta)^{n-t}}{\binom{n}{t}\theta^t(1-\theta)^{n-t}} = \binom{n}{t}^{-1},$$

θ-free. So T is sufficient. ■ *Why the cancellation.* Bernoulli likelihood factors entirely through the count $\sum x_i$ — only "how many" matters, not "which."

11.3 HW6: Kleibergen's $\mathrm{LM}_{\mathrm{CUE}}$, Newey–Smith

11.3.1 HW6 Q1: AMS (2006) Lemma 1 + Theorem 1

[REFERENCE ONLY]

Problem. See Andrews, Moreira, and Stock (Econometrica, 2006). For (a), see Lemma 1; for (b), see Theorem 1, proven on p. 742.

Solution. If asked: "By AMS (2006), the CLR test is power-optimal among invariant similar tests in the iid homoskedastic normal-error linear IV model. Proof uses Hunt–Stein on the orbit of the rotation group." Skip; do not memorize.

11.3.2 HW6 Q2: $\mathrm{LM}_{\mathrm{CUE}} \xrightarrow{d} \chi^2$ Under Strong IV

[CORE — must master, four-step proof]

Problem. On the previous problem set it was found that a t-test can have finite-sample null rejection probabilities that differ substantially from the nominal level, especially when instruments are weak and the correlation between the error terms is large. Here we investigate the finite-sample properties of an LM-type test based on the so-called $\mathrm{LM}_{\mathrm{CUE}}$ statistic that overcomes this problem. In the model $y_i = x_i'\beta + \varepsilon$, $i = 1, \ldots, n$, where x_i is potentially endogenous and Z_i is a vector of IVs, define

$$\begin{aligned} g_i(\beta) &:= (y_i - x_i'\beta)Z_i, \quad G_i := (\partial g_i/\partial\beta)(\beta) = -Z_i x_i', \\ \widehat{g}(\beta) &:= \sum_{i=1}^n g_i(\beta)/n, \quad \widehat{\Omega}(\beta) := \sum_{i=1}^n g_i(\beta)g_i(\beta)'/n, \\ D(\beta) &:= \sum_{i=1}^n (\widehat{g}(\beta)'\widehat{\Omega}(\beta)^{-1}g_i(\beta) - 1)G_i/n, \\ \mathrm{LM}_{\mathrm{CUE}}(\beta) &:= n\widehat{g}(\beta)'\widehat{\Omega}(\beta)^{-1}D(\beta)[D(\beta)'\widehat{\Omega}(\beta)^{-1}D(\beta)]^{-1}D(\beta)'\widehat{\Omega}(\beta)^{-1}\widehat{g}(\beta). \end{aligned}$$

(i) Show that for $n \to \infty$, $\mathrm{LM}_{\mathrm{CUE}}(\beta_0) \xrightarrow{d} \chi^2_{\dim\beta_0}$, where β_0 is the true structural parameter vector. (Hint: First derive the asymptotic distribution of $n^{1/2}\widehat{g}(\beta_0)$ and then probability limits of $D(\beta_0)$ and $\widehat{\Omega}(\beta_0)$. Assume iid observations but not conditional homoskedasticity. State any additional assumptions. Note one can show the same result under weak instrument asymptotics.)

(ii) Redo the Monte Carlo exercise in question 1 PS 5, using $\mathrm{LM}_{\mathrm{CUE}}$ instead of the t-test. Compare.

What this tests. Core proof of the chapter. Section 5.4 of Chapter 5.

Approach. Standard four-step structure. The non-trivial step is decomposing $D(\beta_0)$: the -1 correction kills a cross-term that would otherwise prevent independence between D and $\widehat{g}$. Once $D \xrightarrow{p} G := E[Zx']$, plug into the quadratic form and apply CMT.

Solution.

Step 1: CLT for $n^{1/2}\widehat{g}(\beta_0)$. Plug in $g_i(\beta_0) = \varepsilon_i Z_i$:

$$n^{1/2}\widehat{g}(\beta_0) = \frac{1}{\sqrt{n}}\sum_i \varepsilon_i Z_i \xrightarrow{d} N(0, \Omega), \quad \Omega := E[\varepsilon_i^2 Z_i Z_i'].$$

Why. $E[\varepsilon_i Z_i] = 0$ by exogeneity. iid mean-zero with finite variance Ω (assuming $E[\varepsilon_i^2 \|Z_i\|^2] < \infty$). Lindeberg–Lévy CLT. *Note.* No homoskedasticity assumed: $\Omega \neq \sigma^2 E[ZZ']$ in general.

Step 2: WLLN for $\widehat{\Omega}(\beta_0)$.

$$\widehat{\Omega}(\beta_0) = \frac{1}{n}\sum_i \varepsilon_i^2 Z_i Z_i' \xrightarrow{p} E[\varepsilon_i^2 Z_i Z_i'] = \Omega.$$

Step 3: $D(\beta_0) \xrightarrow{p} G$. Substitute $g_i = \varepsilon_i Z_i$, $G_i = -Z_i x_i'$:

$$D(\beta_0) = -\frac{1}{n}\sum_i \left[\widehat{g}'\widehat{\Omega}^{-1}\varepsilon_i Z_i\right] Z_i x_i' + \frac{1}{n}\sum_i Z_i x_i'.$$

First term: $\widehat{g} = O_p(n^{-1/2})$, $\widehat{\Omega}^{-1} = O_p(1)$, inner average bounded $\Rightarrow o_p(1)$. Second term: WLLN $\to E[Zx'] =: G$, full rank. So $D(\beta_0) \xrightarrow{p} G$. *Why the -1 in D matters.* Without it, D converges to a different limit involving cross-moments of ε and reduced-form residual. The -1 shifts the limit to G, making D asymptotically independent of $\widehat{g}$ — engine of pivotality under weak IV.

Step 4: Combine via CMT.

$$[D'\widehat{\Omega}^{-1}D]^{-1/2}D'\widehat{\Omega}^{-1}\cdot n^{1/2}\widehat{g} \xrightarrow{d} N(0, I_{\dim\beta_0}).$$

$\mathrm{LM}_{\mathrm{CUE}}$ is the squared norm $\xrightarrow{d} \chi^2_{\dim\beta_0}$. ■

(ii) Monte Carlo. NRPs across all (h_1, h_2) are close to nominal 5%, even under weak IVs and high endogeneity. Reflects size control under both strong and weak IV.

11.3.3 HW6 Q3: $\mathrm{LM}_{\mathrm{CUE}}$ Under Weak-IV Asymptotics

[EXAM-WRITE — state independence + conditioning argument]

Problem. The model is $y_i = x_i'\beta + \varepsilon_i$, $i = 1, \ldots, n$, with reduced form $x_i' = Z_i'\Pi + v_i'$. Show that for $n \to \infty$, $\mathrm{LM}_{\mathrm{CUE}}(\beta_0) \xrightarrow{d} \chi^2_{\dim\beta_0}$ under weak-IV asymptotics where $\Pi = n^{-1/2}C$.

(a) Work out the joint limiting normal distribution $(\bar{g}', \mathrm{vec}(\bar{D})')'$ of $n^{1/2}(\widehat{g}(\beta_0)', \mathrm{vec}(D(\beta_0))')'$ and show $n^{1/2}\widehat{g}(\beta_0)$ and $n^{1/2}\mathrm{vec}(D(\beta_0))$ are asymptotically independent.

(b) Show the limiting distribution of $\mathrm{LM}_{\mathrm{CUE}}(\beta_0)$ conditional on $\mathrm{vec}(\bar{D})$ is $\chi^2_{\dim\beta_0}$. Conclude unconditionally.

What this tests. Harder weak-IV version. Tests joint asymptotic normality with bookkeeping that establishes asymptotic independence; conditioning argument.

Approach. Under weak IV, $D(\beta_0)$ stays $O_p(n^{-1/2})$ with normal limit. Step (a): the -1 correction in D produces exact cancellation in the cross-covariance. Step (b): conditional on $\bar{D}$, LM is a quadratic form in fresh independent normal $\Rightarrow \chi^2$ regardless of $\bar{D} \Rightarrow$ unconditionally χ^2.

Solution.

(a) Joint limit + independence.

Rewriting $n^{1/2}D(\beta_0)$ under $\Pi = n^{-1/2}C$, reduced form $x_i' = Z_i'\Pi + v_i'$:

$$n^{1/2}D(\beta_0) = -\Lambda \cdot n^{1/2}\widehat{g}(\beta_0) + n^{-1/2}\sum_j Z_j x_j' + o_p(1),$$

where $\Lambda := n^{-1}\sum Z_i v_i' Z_i' \varepsilon_i \widehat{\Omega}^{-1} \xrightarrow{p} E[Z_i v_{il} Z_i' \varepsilon_i]\Omega^{-1}$. The vector $(n^{-1/2}\sum Z_j\varepsilon_j, n^{-1/2}\sum Z_i v_{il})'$ is jointly normal by CLT.

After algebra, the joint limit covariance has off-diagonal block

$$-E[\varepsilon_i^2 Z_i Z_i']\Omega^{-1}E[\varepsilon_i v_{il} Z_i Z_i'] - E[\varepsilon_i v_{il} Z_i Z_i'] = 0,$$

using $\Omega = E[\varepsilon_i^2 Z_i Z_i']$ in the first term, which exactly cancels the second. *Why this cancellation.* Kleibergen's -1 correction was designed precisely so that cross-covariance vanishes. Joint normality + zero covariance $\Rightarrow$ independence.

(b) Conditioning. Limit:

$$\mathrm{LM}_{\mathrm{CUE}}(\beta_0) \xrightarrow{d} \overline{\mathrm{LM}} := \xi'\widehat{\Omega}^{-1/2}\bar{D}[\bar{D}'\widehat{\Omega}^{-1}\bar{D}]^{-1}\bar{D}'\widehat{\Omega}^{-1/2}\xi,$$

where $\xi := \widehat{\Omega}^{-1/2}\bar{g} \sim N(0, I)$ independent of $\bar{D}$.

Conditional on $\bar{D}$: $\xi \sim N(0, I)$ unchanged. The matrix $P := \widehat{\Omega}^{-1/2}\bar{D}[\bar{D}'\widehat{\Omega}^{-1}\bar{D}]^{-1}\bar{D}'\widehat{\Omega}^{-1/2}$ is an orthogonal projection of rank $\dim\beta_0$ ($\bar{D}$ full rank a.s.). So $\overline{\mathrm{LM}}|\bar{D} \sim \chi^2_{\dim\beta_0}$. *Why $\xi'P\xi \sim \chi^2_r$ for orthogonal projection P of rank r.* Standard fact: $\xi \sim N(0, I_k)$, $P^2 = P$, $P' = P$, rank $r \Rightarrow \xi'P\xi \sim \chi^2_r$.

Conditional law $\chi^2_{\dim\beta_0}$ regardless of $\bar{D} \Rightarrow$ unconditional law $\chi^2_{\dim\beta_0}$. ■

11.3.4 HW6 Q4: Newey–Smith (2004) Empirical Likelihood

[REFERENCE ONLY]

Problem. See Newey and Smith (2004, Econometrica), namely (a) Lemma A1 (p. 239), (b) Lemma A2 (p. 239), (c) Lemma A3 and proof of Theorem 3.1 (p. 239–240).

Solution. Skip. If asked: "By Newey–Smith (2004), EL achieves higher-order efficiency in iid GMM with overidentification."

11.4 HW7: Ridge, Sub-Gaussian, Thresholding, Lasso/Ridge Simulation

11.4.1 HW7 Q1: Bias and Variance of the Ridge Estimator

[CORE — clean closed-form derivation, must master]

Problem. Bias/variance of ridge estimator. Consider the model $Y = X\beta + e$, $\beta \in \mathbb{R}^p$, $E(e|X) = 0$ with iid observations. Show that the bias and covariance matrix of the ridge estimator

$$\widehat{\beta}_{\mathrm{ridge}} = (X'X + \lambda I_p)^{-1}X'Y$$

with fixed parameter $\lambda > 0$ are given by

$$\begin{aligned}\text{bias}(\widehat{\beta}_{\text{ridge}}|X) &= -\lambda(X'X + \lambda I_p)^{-1}\beta,\\ \text{Var}\left(\widehat{\beta}_{\text{ridge}}|X\right) &= (X'X + \lambda I_p)^{-1}(X'DX)(X'X + \lambda I_p)^{-1},\end{aligned}$$

respectively, where $D := E(ee'|X)$.

What this tests. Cleanest closed-form derivation in the course. Section 9.1 of Chapter 9.

Approach. Decompose $\widehat{\beta}_{\text{ridge}} - \beta = (X'X + \lambda I)^{-1}(X'e - \lambda\beta)$ first; bias and MSE follow by conditional expectations. Variance is MSE minus bias-bias-prime; $\lambda^2\beta\beta'$ terms cancel.

Solution.

Step 1: Decompose.

$$\begin{aligned}\widehat{\beta}_{\text{ridge}} - \beta &= (X'X + \lambda I)^{-1}X'Y - \beta\\ &= (X'X + \lambda I)^{-1}\left[X'Y - (X'X + \lambda I)\beta\right]\\ &= (X'X + \lambda I)^{-1}(X'e - \lambda\beta),\end{aligned}$$

using $X'Y = X'X\beta + X'e$. *Why.* Pulling the inverse to the front isolates "raw error" $X'e - \lambda\beta$. The $-\lambda\beta$ is the shrinkage signature.

Step 2: Bias.

$$\text{bias}(\widehat{\beta}_{\text{ridge}}|X) = (X'X + \lambda I)^{-1}E[X'e - \lambda\beta|X] = -\lambda(X'X + \lambda I)^{-1}\beta,$$

using $E[e|X] = 0$. *Why X' pulls out.* Conditioning on X treats it as constant.

Step 3: MSE.

$$\text{MSE}(\widehat{\beta}_{\text{ridge}}|X) = (X'X + \lambda I)^{-1}E[(X'e - \lambda\beta)(X'e - \lambda\beta)'|X](X'X + \lambda I)^{-1}.$$

Expand inner expectation:

$$E[X'ee'X - \lambda X'e\beta' - \lambda\beta e'X + \lambda^2\beta\beta'|X] = X'DX + \lambda^2\beta\beta',$$

cross-terms vanish by $E[e|X] = 0$.

Step 4: Variance. $\text{Var}\left(\widehat{\beta}|X\right) = \text{MSE} - \text{bias}\cdot\text{bias}'$:

$$= (X'X+\lambda I)^{-1}\left[X'DX + \lambda^2\beta\beta' - \lambda^2\beta\beta'\right](X'X+\lambda I)^{-1} = (X'X+\lambda I)^{-1}(X'DX)(X'X+\lambda I)^{-1}. \quad \blacksquare$$

Why this decomposition. Standard identity $E[(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)'] = \text{Var}\left(\widehat{\theta}\right) + \text{bias}\cdot\text{bias}'$.

11.4.2 HW7 Q2: Sub-Gaussian Bounds

[CORE — clean closed-form, must master]

Problem. Sub-Gaussian bounds and means/variances. Consider a random variable X such that for all $\lambda \in \mathbb{R}$,

$$E\exp(\lambda(X - \mu)) \leq \exp(0.5\lambda^2\sigma^2). \tag{1}$$

(a) Show $EX = \mu$.

(b) Show $\text{Var}(X) \leq \sigma^2$.

(c) Suppose the smallest possible σ^2 satisfying (1) is chosen. Is it then true that $\text{Var}(X) = \sigma^2$?

What this tests. MGF-based contradiction (a) + Taylor expansion (b) + counter-example (c). Section 9.2 of Chapter 9.

Approach. The inequality says "X's MGF $\leq$ Gaussian's MGF." Match derivatives at $\lambda = 0$ for (a). Taylor + divide by λ^2 + take limit for (b). Asymmetric distribution example for (c).

Solution.

(a) $EX = \mu$**.** Let $g(\lambda) := \exp(0.5\lambda^2\sigma^2 + \lambda\mu)$, $M(\lambda) := E\exp(\lambda X)$. $g(0) = M(0) = 1$, $M(\lambda) \leq g(\lambda)$.

Claim: $M'(0) = g'(0)$. Proof by contradiction: if $M'(0) \neq g'(0)$, then $\theta := \lim_{\lambda\to 0}(M(\lambda) - g(\lambda))/\lambda \neq 0$.

- $\theta > 0$: $\lim_{\lambda\to 0^+}(M - g)/\lambda > 0$ requires $M > g$ for small $\lambda > 0$, contradicting $M \leq g$.
- $\theta < 0$: same contradiction at $\lambda < 0$.

So $M'(0) = g'(0) = \mu$. But $M'(0) = E[X]$. Hence $E[X] = \mu$. ■

(b) $\text{Var}(X) \leq \sigma^2$**.** Taylor:

$$M(\lambda) = \sum_{n=0}^{\infty} \frac{\lambda^n E[(X-\mu)^n]}{n!} \leq \exp(0.5\lambda^2\sigma^2) = \sum_{n=0}^{\infty} \frac{\lambda^{2n}\sigma^{2n}}{2^n n!}.$$

$n = 0$ cancels; $n = 1$ on LHS is 0 by (a). Divide by $\lambda^2 > 0$:

$$\frac{E[(X-\mu)^2]}{2} + O(\lambda) \leq \frac{\sigma^2}{2} + O(\lambda^2).$$

Send $\lambda \to 0$: $\text{Var}(X) \leq \sigma^2$. ■ *Why divide by λ^2 first.* Both sides $\to 0$ otherwise. Dividing isolates leading λ^2-coefficients.

(c) Strict inequality possible. Take $X = 1 - p$ w.p. p, $X = -p$ w.p. $1 - p$. $EX = 0$, $\text{Var}(X) = p(1-p)$. MGF $M(\lambda) = p\exp((1-p)\lambda) + (1-p)\exp(-p\lambda)$.

For $p = 1/4$: $\text{Var}(X) = 3/16 = 0.1875$. Numerically at $\lambda = 0.01$: $\sigma^2_{\min} \geq (2/\lambda^2)\ln M(\lambda) = 0.1878 > 0.1875$.

Answer: NO, $\text{Var}(X) < \sigma^2_{\min}$ strictly. *Why.* Sub-Gaussian is one-sided MGF bound; for asymmetric distributions one tail dominates, forcing $\sigma^2 > \text{Var}(X)$.

11.4.3 HW7 Q3: Hard vs Soft Thresholding

[CORE — must master soft-threshold case-by-case]

Problem. Consider the model $y = \theta + w \in \mathbb{R}^n$. Denote by $\lambda > 0$ a user-chosen threshold. Show that the hard-thresholding estimator $\widehat{\theta}^H_i(y) := y_i$ if $|y_i| \geq \lambda$ and 0 otherwise corresponds to the solution $\widehat{\theta}$ of

$$\min_{\theta\in\mathbb{R}^n} 0.5\|y - \theta\|_2^2 + 0.5\lambda^2\|\theta\|_0,$$

while the soft-thresholding estimator $\widehat{\theta}_i^S(y) := \text{sgn}(y_i)(|y_i| - \lambda)$ if $|y_i| \geq \lambda$ and 0 otherwise solves

$$\min_{\theta \in \mathbb{R}^n} 0.5\|y - \theta\|_2^2 + \lambda\|\theta\|_1.$$

Show the first problem is non-convex while the second is convex.

What this tests. Standard scalar derivation; structural insight that ℓ_0 is non-convex, ℓ_1 is convex. Section 9.3 of Chapter 9.

Approach. Reduce to $n = 1$ by additive separability. Hard: only two candidates ($\theta = 0$ or $\theta = y$). Soft: case-by-case on sign of y and whether $|y| \gtrless \lambda$.

Solution.

Reduction to scalar. Both objectives separable across i; minimize each summand. Take $n = 1$.

Hard. For $\theta \neq 0$: minimized at $\theta = y$ with value $0.5\lambda^2$. For $\theta = 0$: value $0.5y^2$.

- $|y| < \lambda$: $0.5y^2 < 0.5\lambda^2$, choose $\theta = 0$.
- $|y| > \lambda$: $0.5y^2 > 0.5\lambda^2$, choose $\theta = y$.

$\widehat{\theta}^H = y \cdot \mathbf{1}(|y| \geq \lambda)$. ■ *Why only two candidates.* ℓ_0 jumps by $0.5\lambda^2$ at $\theta = 0$; conditional on $\theta \neq 0$, smooth optimum is $\theta = y$.

Soft. $c(\theta) = 0.5(y - \theta)^2 + \lambda|\theta|$. For $\theta \neq 0$: $c'(\theta) = -(y - \theta) + \lambda\,\text{sgn}(\theta)$.

Case 1: $|y| < \lambda$**.** $\theta > 0$: $c'(\theta) > 0$. $\theta < 0$: $c'(\theta) < 0$. Both push to $\theta = 0 \Rightarrow$ minimum at $\theta = 0$.

Case 2: $y \geq \lambda > 0$**.** $c'(\theta) = 0$ at $\theta = y - \lambda > 0 \Rightarrow$ minimum at $\theta = y - \lambda = \text{sgn}(y)(|y| - \lambda)$.

Case 3: $y \leq -\lambda$**.** Analogous: minimum at $\theta = y + \lambda = \text{sgn}(y)(|y| - \lambda)$.

Combine: $\widehat{\theta}^S = \text{sgn}(y)(|y| - \lambda)_+$. ■ *Why λ is additive shrinkage.* ℓ_1 has subdifferential $[-\lambda, \lambda]$ at 0, "catching" values within $|y| < \lambda$. For larger y, optimum shifts by exactly λ toward zero.

Convexity. Soft: convex (sum of convex). Hard: non-convex — for $|y| > \lambda$, line segment from $(0, 0.5y^2)$ to $(y, 0.5\lambda^2)$ lies above the curve at $\theta = y/2$. (Check: function value at $y/2$ is $0.125y^2 + 0.5\lambda^2$, less than linear interpolation $0.25y^2 + 0.25\lambda^2$ when $y^2 > 2\lambda^2$.) *Why this matters.* Convex $\Rightarrow$ poly-time solvable. Lasso (soft) computable; hard thresholding NP-hard in general.

11.4.4 HW7 Q4: Lasso vs Ridge Simulation

[LOW ROI — simulation; remember the qualitative pattern]

Problem. Simulation exercise involving LASSO and Ridge. Consider the same model as in question 2 with $n = 500$ and $p = 600$. Assume all components of x_i and ε_i are iid $N(0, 1)$. Consider different choices for $\beta = (\beta_1', \beta_2')' \in \mathbb{R}^{p_1 + p_2}$ where $\beta_1 = c_1 \mathbf{1}_{p_1}$, $\beta_2 = c_2 \mathbf{1}_{p_2}$. Consider all combinations of $p_1 = 400, 550, 570$ ($p_2 = 600 - p_1$), $c_1 = 0, 0.1$, $c_2 = 0.1, 0.5$. Report results on 1,000 simulation repetitions.

(a) Report sample percentage $P(\widehat{\beta}_{\text{ridge}}(1) = 0)$ and $P(\widehat{\beta}_{\text{ridge}}(600) = 0)$. Same for LASSO.

(b) For both estimators, report average prediction errors in ℓ_1, ℓ_2, and $\ell_{2,n}$ norm.

(c) Interpret your results.

Note: STATA `lasso` command (default $K = 10$ CV); Ridge in `lassopack`.

What this tests. Empirical demonstration of Lasso oracle property vs Ridge no-thresholding.

Approach. Don't memorize numbers; remember pattern. Lasso → exact zeros for genuinely zero coefficients; Ridge → all nonzero. Lasso wins generally, except in small-nonzero regime where it incorrectly zeros out signal.

Solution.

Pattern from simulation:

- $P(\widehat{\beta}_{\text{ridge}}(\cdot) = 0) = 0$ *always.*
- $P(\widehat{\beta}_{\text{Lasso}}(j) = 0)$ high for $c_j = 0$ (oracle).
- $P(\widehat{\beta}_{\text{Lasso}}(j) = 0)$ also high for $c_j = 0.1$ small but nonzero (small signal thresholded).
- ℓ_2 loss: Lasso < Ridge generally; Ridge < Lasso when $c_1 = c_2 = 0.1$ (small nonzero).

Interpretation.

- Lasso has *oracle property*: correctly identifies zeros with high probability.
- Lasso wrongly zeros out small nonzero coefficients (failure under small signal). Ridge's smooth shrinkage preserves them.
- Generally Lasso wins on ℓ_p losses, except small-signal regime.

Why. Lasso = OLS soft-thresholded at λ. Below λ killed; above λ shrunk by λ; zero correctly zeroed (oracle); small-but-nonzero ($c < \lambda$) wrongly zeroed.

11.5 HW8: Edgeworth Lemmas, Bootstrap Failure, Subsampling

11.5.1 HW8 Q1: Lecture 23 Technical Lemmas

[LOW ROI — Edgeworth machinery]

Problem. Some details in Lecture 23:

(a) Show how (11) implies (12). Knowing the difference between coefficients of $\widehat{q}_1(x)$ and $q_1(x)$ is $O_p(n^{-1/2+\delta}, n^{-j_2})$, show $n^{-1/2}(\widehat{q}_1(x) - q_1(x))\phi(x) = O_p(n^{-1+\delta}, n^{-j_2})$ uniformly over $x \in C$ for any compact $C \subset \mathbb{R}$.

(b) In (34) we use: if X has cdf F, $F(x) = P(X \leq x)$, then $U = F(X) \sim U[0, 1]$. Prove.

(c) Show that if $X_n = O_p(n^{-1+\delta}, n^{-1})$ for every $\delta > 0$, then also $X_n = o_p(n^{-1+\delta}, n^{-1})$ for every $\delta > 0$.

What this tests. Edgeworth expansion bookkeeping. Background for Chapter 6.

Approach. Each part is 2–3 lines. (a) Polynomials and ϕ are $O(1)$ on compacts; multiplication preserves O_p. (b) Probability-integral transform: $P(F(X) \leq u) = u$. (c) Apply assumption with $\delta/2$, use the gap to absorb constants.

Solution.

(a) $\widehat{q}_1 - q_1$ has coefficients $O_p(n^{-1/2+\delta}, n^{-j_2})$; on compact C, $|x^i| = O(1)$; $\phi(x) = O(1)$. Multiply: $(\widehat{q}_1 - q_1)\phi(x) = O_p(n^{-1/2+\delta}, n^{-j_2})$. Multiply by $n^{-1/2}$: $O_p(n^{-1+\delta}, n^{-j_2})$ uniformly. ■

(b) For $u \in [0,1]$:

$$P(U \leq u) = P(F(X) \leq u) = P(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u,$$

using monotonicity of F. ■

(c) Given $\varepsilon > 0$, $v > 0$. Apply assumption with $\delta/2$: K_v such that $\limsup nP(\|X_n\| > n^{-1+\delta/2}K_v) \leq v/2$. Pick n_v such that $n^{-\delta/2}K_v < \varepsilon$ for $n \geq n_v$. Then for such n,

$$v/2 \geq \limsup nP(\|X_n\| > n^{-1+\delta/2}K_v) \geq \limsup nP(\|X_n\| > n^{-1+\delta}\varepsilon),$$

so $X_n = o_p(n^{-1+\delta}, n^{-1})$. ■

11.5.2 HW8 Q2: Bootstrap Failure at the Boundary (Andrews 2000)

[CORE — must master the failure mechanism]

Problem. The bootstrap is not a panacea. Let $X_i \overset{i.i.d.}{\sim} N(\mu, 1)$, $i = 1, \ldots, n$, with $\mu \in \mathbb{R}^+ = \{y \geq 0\}$. The MLE of μ is $\widehat{\mu}_n = \max\{\bar{X}_n, 0\}$.

(a) Show $n^{1/2}(\widehat{\mu}_n - \mu) \xrightarrow{d} Z$ if $\mu > 0$, where $Z \sim N(0,1)$, and $n^{1/2}(\widehat{\mu}_n - \mu) \xrightarrow{d} \max\{Z, 0\}$ if $\mu = 0$.

(b) Now consider the nonparametric bootstrap. Let $X_i^* \overset{i.i.d.}{\sim} \widehat{F}_n$, define $\widehat{\mu}_n^* = \max\{\bar{X}_n^*, 0\}$. Asymptotic validity requires conditional (on $\widehat{F}_n$) limit of $n^{1/2}(\widehat{\mu}_n^* - \widehat{\mu}_n)$ equals limit of $n^{1/2}(\widehat{\mu}_n - \mu)$ w.p. 1. Show this is NOT the case at $\mu = 0$. Let $A_c := \{\omega : \liminf_n n^{1/2}\bar{X}_n(\omega) < -c\}$ for $0 < c < \infty$. By LIL, $P(A_c) = 1$. For $\omega \in A_c$ pick a subsequence n_k with $n_k^{1/2}\bar{X}_{n_k}(\omega) \leq -c$. Show:

   (i) $n_k^{1/2}(\widehat{\mu}_{n_k}^* - \widehat{\mu}_{n_k}(\omega)) \leq \max\{n_k^{1/2}(\bar{X}_{n_k}^*(\omega) - \bar{X}_{n_k}(\omega)) - c, 0\}$.

   (ii) $\max\{n_k^{1/2}(\bar{X}_{n_k}^*(\omega) - \bar{X}_{n_k}(\omega)) - c, 0\} \xrightarrow{d} \max\{Z - c, 0\}$ as $k \to \infty$ conditional on $\widehat{F}_n$.

   (iii) $\max\{Z - c, 0\} \leq \max\{Z, 0\}$ with strict inequality with positive probability. Conclude bootstrap inconsistency.

(c) In this setup: (i) Would the parametric bootstrap be consistent? (ii) Would subsampling be consistent?

What this tests. Textbook example of bootstrap failure (Andrews 2000). Tests: discontinuity of limit at boundary + LIL property. Section 2.10 of Chapter 2.

Approach. (a) For $\mu > 0$: SLLN $\Rightarrow \widehat{\mu}_n = \bar{X}_n$ eventually $\Rightarrow$ CLT gives $N(0,1)$. For $\mu = 0$: CMT applied to max. (b) On LIL event, $\bar{X}_n$ dips below $-c/\sqrt{n}$ infinitely often; bootstrap thinks "$\mu < 0$" regime, gets wrong limit. (c) Parametric also fails; subsampling works.

Solution.

(a) If $\mu > 0$: SLLN $\Rightarrow \bar{X}_n \to \mu > 0$ a.s., so $\widehat{\mu}_n = \bar{X}_n$ eventually. CLT $\Rightarrow n^{1/2}(\widehat{\mu}_n - \mu) \xrightarrow{d} Z$.

If $\mu = 0$:

$$n^{1/2}\widehat{\mu}_n = \max\{n^{1/2}\bar{X}_n, 0\} \xrightarrow{d} \max\{Z, 0\}$$

by CMT applied to $x \mapsto \max\{x, 0\}$, with $n^{1/2}\bar{X}_n \xrightarrow{d} Z$. *Why discontinuous in μ.* $\mu > 0$: $N(0,1)$ symmetric. $\mu = 0$: half-normal. Different distributions.

(b)(i) On $\omega \in A_c$, $n_k^{1/2}\bar{X}_{n_k}(\omega) \le -c$ so $\widehat{\mu}_{n_k}(\omega) = 0$. Then

$$\begin{aligned} n_k^{1/2}(\widehat{\mu}^*_{n_k} - \widehat{\mu}_{n_k}(\omega)) = n_k^{1/2}\widehat{\mu}^*_{n_k} &= \max\{n_k^{1/2}\bar{X}^*_{n_k}, 0\} \\ &= \max\{n_k^{1/2}(\bar{X}^*_{n_k} - \bar{X}_{n_k}(\omega)) + n_k^{1/2}\bar{X}_{n_k}(\omega), 0\} \\ &\le \max\{n_k^{1/2}(\bar{X}^*_{n_k} - \bar{X}_{n_k}(\omega)) - c, 0\}. \end{aligned}$$

(ii) Conditional bootstrap CLT: $n_k^{1/2}(\bar{X}^*_{n_k} - \bar{X}_{n_k}(\omega)) \xrightarrow{d} Z$. By CMT: $\max\{n_k^{1/2}(\bar{X}^*_{n_k} - \bar{X}_{n_k}) - c, 0\} \xrightarrow{d} \max\{Z - c, 0\}$.

(iii) $\max\{Z - c, 0\} \le \max\{Z, 0\}$ pointwise, strict inequality when $0 < Z < c$ (positive probability). So bootstrap conditional limit is stochastically smaller than true limit $\max\{Z, 0\}$. Inconsistent. ■ *Why this happens.* Bootstrap thinks $\widehat{\mu}_n$ is the "true μ" in bootstrap world; at $\mu = 0$, $\bar{X}_n$ wanders below 0 along LIL subsequences, misleading the bootstrap.

(c)(i) Parametric bootstrap: also **inconsistent** (Andrews 2000, p. 401–402). Same boundary issue.

(c)(ii) Subsampling: **consistent** (Andrews 2000, p. 403–404). Subsampling only requires test statistic to have a limit, not bootstrap-conditional matching. *Why subsampling escapes.* Doesn't plug in $\widehat{\mu}_n$; uses subsamples of size $b \ll n$. Subsample statistic has same limit law as full-sample; subsampling tracks correct quantile.

11.5.3 HW8 Q3: Equivalence of Two Bootstrap Procedures

[MEMORIZE — quick algebra]

Problem. Consider the following bootstrap for $y_i \in \mathbb{R}$ on $x_i \in \mathbb{R}^k$. Let $\widehat{\beta}$ denote OLS of $Y = (y_1, \ldots, y_n)'$ on $X = (x_1, \ldots, x_n)'$ and $\widehat{e} = Y - X\widehat{\beta} \in \mathbb{R}^n$ the residuals.

(a) Draw a random vector (x'^*, e^*) from the pair $\{(x_i', \widehat{e}_i) : i = 1, \ldots, n\}$. That is, draw a random integer i' from $\{1, \ldots, n\}$ and set $x^* = x_{i'}$ and $e^* = \widehat{e}_{i'}$. Set $y^* = x'^*\widehat{\beta} + e^*$. Draw with replacement n such vectors, creating a random bootstrap data set (Y^*, X^*).

(b) Regress Y^* on X^*, yielding OLS estimates $\widehat{\beta}^*$ and any other statistic of interest.

Show that the bootstrap procedure is numerically identical to the nonparametric bootstrap.

What this tests. Short algebraic identity showing residual-resampling equals iid-pair resampling in linear models.

Approach. Plug $\widehat{e}_{i'} = y_{i'} - x_{i'}'\widehat{\beta}$ into y^*.

Solution. For drawn index i':

$$y^* = x'^*\widehat{\beta} + e^* = x_{i'}'\widehat{\beta} + \widehat{e}_{i'} = x_{i'}'\widehat{\beta} + (y_{i'} - x_{i'}'\widehat{\beta}) = y_{i'}.$$

So $(y^*, x^*) = (y_{i'}, x_{i'})$, identical to iid pair drawing from empirical distribution. ■ *Why useful.* NP bootstrap consistency ⇒ residual bootstrap consistency for free in linear models. Nonlinear or non-iid: equivalence breaks.

11.5.4 HW8 Q4: Subsampling AsySz at the Boundary

[EXAM-WRITE — state calculation, do not simulate]

Problem. Continuing the boundary setup. Subsampling with subsample size b: $b \to \infty$, $b/n \to 0$. Show subsampling-based CI has asymptotic size $1 - \alpha$.

What this tests. Subsampling robustness to discontinuities.

Approach. Under sequences $b^{1/2}\mu \to g$: full-sample statistic limit $\xrightarrow{d} Z$ (since $n^{1/2}\mu \to \infty$); subsample limit $\xrightarrow{d} \max\{Z, -g\}$. Worst-case coverage at $g = \infty$ gives exactly $1 - \alpha$.

Solution.

(a) Test statistic limit. If $b^{1/2}\mu \to g$ and $b/n \to 0$, then $n^{1/2}\mu = (n/b)^{1/2} \cdot b^{1/2}\mu \to \infty$. So $\bar{X}_n > 0$ eventually:

$$n^{1/2}(\widehat{\mu}_n - \mu) = \max\{n^{1/2}(\bar{X}_n - \mu), -n^{1/2}\mu\} \xrightarrow{d} Z.$$

(b) Subsampling critical value. Subsample statistic $b^{1/2}(\widehat{\mu}_b^* - \mu)$:

- $g = \infty$: $\xrightarrow{d} Z$.
- g finite: $\xrightarrow{d} \max\{Z, -g\}$.

Subsampling CV $\widehat{c}_n(1-\alpha) \to c_g(1-\alpha) = 1 - \alpha$ quantile of $\max\{Z, -g\}$.

(c) Coverage. $\mathrm{P}(Z \le c_g(1-\alpha))$. As g grows from 0 to ∞: $\max\{Z, -g\}$ shifts toward Z, so $c_g(1-\alpha)$ decreases. Smallest at $g = \infty$: $c_\infty(1-\alpha) = z_{1-\alpha}$. $\inf_g \mathrm{P}(Z \le c_g(1-\alpha)) = 1-\alpha$.

So $\mathrm{AsySz}_{\text{coverage}} = 1-\alpha$. ■ *Why subsampling controls size while bootstrap fails.* Subsampling tracks the limit law directly via subsample-scale statistics, sidestepping the bootstrap-conditional-mimicry requirement that fails at discontinuities.

11.6 HW9: Identification (Heckit, Mixed IV, Control Function), USCON

11.6.1 HW9 Q1: Heckit Selection Model

[CORE — Heckit identifying equation; must master]

Problem. In the Heckit model (Heckman, 1979),

(a) Show $E(y|x, z, d = 1) = x'\theta + \sigma_{\varepsilon\eta}\lambda(x'\pi_1 + z'\pi_2)$, where $\lambda(s) = \phi(s)/\Phi(s)$.

(b) Identify π_1 and π_2 from a probit regression of d onto x and z.

(c) Specify conditions under which one can identify θ and $\sigma_{\varepsilon\eta}$ (consider the LS regression of y on x and $\lambda(x'\pi_1 + z'\pi_2)$).

The model: $y^* = x'\theta + \varepsilon$ (latent), $y = d \cdot y^*$ (observed), $d = \mathbf{1}(x'\pi_1 + z'\pi_2 \geq -\eta)$, $(\varepsilon, \eta) \perp (x, z)$ and $(\varepsilon, \eta) \sim N(0, \Sigma)$ with $\Sigma = \begin{pmatrix} \sigma_\varepsilon^2 & \sigma_{\varepsilon\eta} \\ \sigma_{\varepsilon\eta} & 1 \end{pmatrix}$ (variance of η normalized to 1).

What this tests. Tests inverse-Mills-ratio derivation. Section 3.7 of Chapter 3.

Approach. (a) Decompose $E(y|x, z, d = 1) = x'\theta + E(\varepsilon|x, z, x'\pi_1 + z'\pi_2 \geq -\eta)$. Bivariate normal: $E(\varepsilon|\eta = \bar{\eta}) = \sigma_{\varepsilon\eta}\bar{\eta}$. Integrate over $\bar{\eta} \geq -c$. Use $\bar{\eta}\phi(\bar{\eta}) = -\phi'(\bar{\eta})$. (b) Probit: $E(d|x, z) = \Phi(x'\pi_1 + z'\pi_2)$. (c) Differentiate w.r.t. z and x.

Solution.

(a) Step 1. $E(y|x, z, d = 1) = x'\theta + E(\varepsilon|x, z, \eta \geq -c)$, $c = x'\pi_1 + z'\pi_2$.

Step 2. Bivariate normal with $E\eta = 0$, $\mathrm{Var}(\eta) = 1$, $\mathrm{Cov}(\varepsilon, \eta) = \sigma_{\varepsilon\eta}$:

$$E(\varepsilon|\eta = \bar{\eta}) = \sigma_{\varepsilon\eta}\bar{\eta}.$$

Step 3.

$$E(\varepsilon|\eta \geq -c) = \frac{1}{\mathrm{P}(\eta \geq -c)} \int_{-c}^{\infty} \sigma_{\varepsilon\eta}\bar{\eta}\phi(\bar{\eta})d\bar{\eta} = \frac{\sigma_{\varepsilon\eta}}{\Phi(c)} \int_{-c}^{\infty} \bar{\eta}\phi(\bar{\eta})d\bar{\eta},$$

using $\mathrm{P}(\eta \geq -c) = \Phi(c)$ since $-\eta \sim N(0, 1)$.

Step 4. $\bar{\eta}\phi(\bar{\eta}) = -\phi'(\bar{\eta})$:

$$\int_{-c}^{\infty} \bar{\eta}\phi(\bar{\eta})d\bar{\eta} = -[\phi(\bar{\eta})]_{-c}^{\infty} = \phi(c).$$

Step 5. $E(\varepsilon|\eta \geq -c) = \sigma_{\varepsilon\eta}\lambda(c)$, $\lambda(s) = \phi(s)/\Phi(s)$. So $E(y|x, z, d = 1) = x'\theta + \sigma_{\varepsilon\eta}\lambda(x'\pi_1 + z'\pi_2)$. ■

(b) $E(d|x, z) = \mathrm{P}(\eta \geq -(x'\pi_1 + z'\pi_2)) = \Phi(x'\pi_1 + z'\pi_2)$ $(-\eta \sim N(0, 1))$. Invert: $\Phi^{-1}(E(d|x, z)) = x'\pi_1 + z'\pi_2$. Linear regression of $\Phi^{-1}(E(d|x, z))$ on (x, z); identified if $E[(x', z')'(x', z')]$ full rank.

(c) Differentiate w.r.t. z:

$$\frac{\partial E(y|x, z, d = 1)}{\partial z} = \sigma_{\varepsilon\eta}\lambda'(x'\pi_1 + z'\pi_2)\pi_2.$$

Postmultiply by π_2 (assuming $\neq 0$, exclusion restriction):

$$\sigma_{\varepsilon\eta}\lambda'(\cdot) = \frac{\partial E(y|x, z, d = 1)}{\partial z}\pi_2 / \left\|\pi_2\right\|^2.$$

RHS identified, λ' known $\Rightarrow$ $\sigma_{\varepsilon\eta}$ identified.

For θ: differentiate w.r.t. x:

$$\frac{\partial E(y|x, z, d = 1)}{\partial x} = \theta + \sigma_{\varepsilon\eta}\lambda'(\cdot)\pi_1 \Rightarrow \theta = \partial E(y|\cdot)/\partial x - \sigma_{\varepsilon\eta}\lambda'\pi_1.$$

Each piece identified $\Rightarrow$ θ identified. *Why exclusion restriction.* $\pi_2 = 0 \Rightarrow \lambda(x'\pi_1)$ is a function of x alone, indistinguishable from $x'\theta$ in $E(y|x, z, d = 1)$.

11.6.2 HW9 Q2: Mixed IV + NPV Derivative Identification

[CORE — must master both parts]

Problem.

(a) Consider the model defined by $y = x'\beta + z_1'\gamma + \varepsilon$ and $x = \Pi z + \eta$, where $E[z\varepsilon] = 0$, $E[z\eta'] = 0$, $E[zz']$ is nonsingular. Show that $\delta = (\beta', \gamma')'$ is point identified under the rank condition $\text{rk}(\Pi_2) = d_x$.

(b) Show $\partial m/\partial x$ is identified in the second model defined in "Identification in the Additively Separable Nonparametric Model."

What this tests. (a) Rank-condition argument for mixed IV. (b) Chain-rule for nonparametric IV with control function. Sections 3.5 and 3.8.

Approach. (a) Compute $E[zw']$ in block-matrix form. Right factor full column rank iff $\text{rk}(\Pi_2) = d_x$. (b) Differentiate $E(y|z,\eta) = m(\pi(z) + \eta, z_1) + \xi(\eta)$ w.r.t. z_2.

Solution.

(a) With $w = (x', z_1')'$:

$$E[zw'] = E[zz'] \begin{pmatrix} \Pi_1' & I_{d_{z_1}} \\ \Pi_2' & 0 \end{pmatrix},$$

using $E[z\eta'] = 0$.

Right factor full column rank $\Leftrightarrow$ $\text{rk}(\Pi_2) = d_x$:

- First d_x columns: $(\Pi_1', \Pi_2')'$. Linearly independent iff Π_2 full column rank.
- Next d_{z_1} columns: $(I, 0)'$. Always linearly independent.
- Joint independence: first block has nonzero bottom rows (Π_2'); second block has zero bottom rows. Cannot combine to give zero.

So $E[zw']$ full column rank.

Identify δ via $E[zy] = E[zw']\delta + 0$, premultiply by $E[wz']E[zz']^{-1}$:

$$\delta = \left[E[wz']E[zz']^{-1}E[zw']\right]^{-1} E[wz']E[zz']^{-1}E[zy]. \quad \blacksquare$$

(b) Under CF $E(\varepsilon|z,\eta) = E(\varepsilon|\eta) =: \xi(\eta)$:

$$E(y|z,\eta) = m(\pi(z) + \eta, z_1) + \xi(\eta).$$

Differentiate w.r.t. z_2 (exclusion: z_2 enters only π, not z_1 or ξ):

$$\frac{\partial E(y|z,\eta)}{\partial z_2} = \frac{\partial m}{\partial x} \cdot \frac{\partial \pi(z)}{\partial z_2'}.$$

If $\partial\pi/\partial z_2'$ has full row rank d_x:

$$\frac{\partial m}{\partial x} = \left[\left(\frac{\partial \pi}{\partial z_2'}\right)\left(\frac{\partial \pi}{\partial z_2'}\right)'\right]^{-1} \left(\frac{\partial \pi}{\partial z_2'}\right) \frac{\partial E(y|z,\eta)}{\partial z_2'}.$$

RHS observable $\Rightarrow$ $\partial m/\partial x$ identified. $\blacksquare$ *Why differentiate w.r.t. z_2.* Excludes z_1-channel and $\xi(\eta)$-channel, isolating $\partial m/\partial x$.

11.6.3 HW9 Q3: Control Function Conditions

[CORE — must master discrete counter-example]

Problem. The control function assumption is $E(\varepsilon|z,\eta) = E(\varepsilon|\eta)$ a.s.

(a) Show that a sufficient condition is that (ε,η) and z are independent (assuming $E|\varepsilon| < \infty$).

(b) Show that the CF assumption does not necessarily imply $E(\varepsilon|z) = 0$ a.s., even if $E\varepsilon = 0$.

What this tests. CF $\not\Rightarrow$ IV exogeneity counter-example. Section 3.8.

Approach. (a) Density factorization with f_z canceling. (b) Take $z = \eta$: CF holds trivially, but joint distribution of (ε, z) can violate IV exogeneity.

Solution.

(a) Density factorization.

$$\begin{aligned} E(\varepsilon|z,\eta) &= \int \varepsilon \frac{f_{\varepsilon,z,\eta}(\varepsilon,z,\eta)}{f_{z,\eta}(z,\eta)} d\varepsilon \\ &= \int \varepsilon \frac{f_{\varepsilon,\eta}(\varepsilon,\eta) f_z(z)}{f_z(z) f_\eta(\eta)} d\varepsilon \quad \text{(by } (\varepsilon,\eta) \perp z\text{)} \\ &= \int \varepsilon \frac{f_{\varepsilon,\eta}(\varepsilon,\eta)}{f_\eta(\eta)} d\varepsilon = E(\varepsilon|\eta). \quad \blacksquare \end{aligned}$$

(b) Counter-example. Take $z = \eta$. CF holds trivially since $\eta = z$ is no more information than z alone. Now construct $(\varepsilon, z) \in \{-1, 0, 1\} \times \{-1, 1\}$:

(ε, z)	$z = -1$	$z = +1$	marginal of ε
$\varepsilon = -1$	1/12	1/4	1/3
$\varepsilon = 0$	1/4	1/12	1/3
$\varepsilon = +1$	1/4	1/12	1/3
marginal of z	7/12	5/12	1

Check: probabilities sum to 1. $E\varepsilon = 0$. $E(\varepsilon|z,\eta) = E(\varepsilon|\eta)$ trivially. But:

$$E(\varepsilon|z=+1) = \frac{(-1)(1/4) + 0 \cdot (1/12) + (1)(1/12)}{5/12} = \frac{-2/12}{5/12} = -\frac{2}{5} \neq 0.$$

So CF holds + $E\varepsilon = 0$ but IV exogeneity fails. $\blacksquare$ *Take-away.* CF and IV are distinct. Patrik likes this distinction.

11.6.4 HW9 Q4: USCON + ID ⇒ Almost-Sure Consistency

[CORE — almost-sure variant of Theorem 11.1]

Problem. Let $\widehat{\theta}_n$ minimize $Q_n(\theta)$ over $\theta \in \Theta$. Suppose:

- **USCON:** $\sup_{\theta\in\Theta} |Q_n(\theta) - Q(\theta)| \to 0$ a.s. as $n \to \infty$.

- **ID:** $\forall \varepsilon > 0$, $\inf_{\theta \notin B(\theta_0,\varepsilon)} Q(\theta) > Q(\theta_0)$.

Show $\widehat{\theta}_n \to \theta_0$ a.s.

What this tests. Standard contradiction proof for almost-sure consistency. Section "Almost-Sure Variant" in Chapter 1.

Approach. Pick good ω, suppose $\widehat{\theta}_n \not\to \theta_0$, extract subsequence outside $B(\theta_0, \varepsilon)$, use ID for gap δ, USCON with $\delta/3$ tolerance, contradiction.

Solution.

Step 1. Take $\omega \in \Omega^{\text{good}}$ where USCON holds ($P(\Omega^{\text{good}}) = 1$).

Step 2. Suppose $\widehat{\theta}_n(\omega) \not\to \theta_0$. Then $\exists \varepsilon > 0$ and subsequence n_i with $\left\|\widehat{\theta}_{n_i}(\omega) - \theta_0\right\| > \varepsilon$.

Step 3. By ID: $\delta := \inf_{\theta \notin B(\theta_0,\varepsilon)} Q(\theta) - Q(\theta_0) > 0$.

Step 4. Pick M such that for $n \geq M$, $\sup_\theta |Q_n - Q|(\omega) < \delta/3$. For $n_i \geq M$:

$$Q(\widehat{\theta}_{n_i}) - \delta/3 < Q_{n_i}(\widehat{\theta}_{n_i}) \leq Q_{n_i}(\theta_0) < Q(\theta_0) + \delta/3,$$

where middle uses minimization. Subtract: $Q(\widehat{\theta}_{n_i}) - Q(\theta_0) < 2\delta/3 < \delta$. But Step 3 says $\geq \delta$. **Contradiction.** ■

11.7 HW10: Bonferroni Critical Values, Bootstrap CI Simulation

11.7.1 HW10 Q1: Newey–Smith Theorem 3.2

[REFERENCE ONLY]

Problem. See Newey and Smith, Theorem 3.2 (2004, Econometrica).

Solution. Skip; cite if it appears.

11.7.2 HW10 Q2: Bonferroni-Style Critical Values (2024 Final)

[EXAM-WRITE — follow size-bound argument]

Problem. This question appeared on the final exam of 2024.

(a) By definition $n^{1/2}\gamma_{n,h,1} = (n^{1/2}\sigma^{-1}_{F_n,j}(\theta_{n,h})E_{F_n}m_j(W_i, \theta_{n,h}))_{j=1,\ldots,p} \to h_1$. Construct an (inconsistent) estimator $\widehat{\gamma}_{n,h,1}$ for $\gamma_{n,h,1}$ using sample analogues. Build a Wald-type $1-\beta$ CR for $\gamma_{n,h,1}$. Translate to a CR for h_1.

(b) $h_{2,n} = \lim \gamma_{n,h,2} = \lim \operatorname{Corr}_{F_n}(m(W_i, \theta_{n,h}))$. Estimate by sample correlation $\widehat{h}_{2,n}$.

(c) Define the proposed critical value $cv(1-\alpha) := \sup_{\bar{h}_1 \in \mathrm{CR}_{h_1,n}(1-\beta)} c_{(\bar{h}_1, \widehat{h}_{2,n})}(1-\alpha)$. Show that the bound on the asymptotic size with this critical value is $\alpha + \beta$, so the procedure as defined does not in general control size.

(d) Show that implementing the test with $\delta = \alpha - \beta$ in place of α gives asymptotic size bounded by α.

(e) Compare the power merits of the resulting test relative to the worst-case plug-in test.

What this tests. Bonferroni-style construction. Tests size-bound argument $P(A \cup B) \leq P(A) + P(B)$. Section 8.6 of Chapter 8.

Approach. (a) Wald CR using $\widehat{h}_{2,n}$ as variance. (b) Plug-in correlation. (c) Decompose $P(\text{reject}) \leq P(\text{reject} \cap h_1 \in \text{CR}) + P(h_1 \notin \text{CR}) \leq \alpha + \beta$. (d) Solve $\delta + \beta = \alpha$. (e) Smaller sup $\Rightarrow$ smaller CV but larger quantile (1-δ vs 1-α); typically more powerful.

Solution.

(a) Estimator + CR. Replace expectations with sample analogues:

$$\widehat{\gamma}_{n,h,1} := \widehat{D}_n^{-1/2}(\theta_{n,h})\bar{m}_n(\theta_{n,h}),$$

$\widehat{D}_n = \text{Diag}(\widehat{\sigma}_{n,j}^2)$. By CLT/WLLN/Slutsky, for $\|h_1\| < \infty$:

$$n^{1/2}(\widehat{\gamma}_{n,h,1} - \gamma_{n,h,1}) = n^{1/2}\widehat{D}_n^{-1/2}(\bar{m}_n - Em) + o_p(1) \xrightarrow{d} \widetilde{Z} \sim N(0, h_2),$$

so $\widehat{h}_{2,n}^{-1/2} n^{1/2}(\widehat{\gamma}_{n,h,1} - \gamma_{n,h,1}) \xrightarrow{d} N(0, I_p)$.

Wald CR for $\gamma_{n,h,1}$:

$$\text{CR}_{\gamma_{n,h,1},n}(1-\beta) = \left\{\gamma_1 \in \mathbb{R}^p : n(\widehat{\gamma}_{n,h,1} - \gamma_1)'\widehat{h}_{2,n}^{-1}(\widehat{\gamma}_{n,h,1} - \gamma_1) \leq \chi^2_{p,1-\beta}\right\}.$$

Translate: $\text{CR}_{h_1,n}(1-\beta) := \{h_1 = n^{1/2}\gamma_1 : \gamma_1 \in \text{CR}_{\gamma_{n,h,1},n}(1-\beta)\}$.

(b) $\widehat{h}_{2,n} = \widehat{D}_n^{-1/2}\widehat{\Sigma}_n\widehat{D}_n^{-1/2}$.

(c) Size bound $\alpha + \beta$. Total probability:

$$\begin{aligned} P(T_n > cv(1-\alpha)) &= P(T_n > cv \cap h_1 \in \text{CR}) + P(T_n > cv \cap h_1 \notin \text{CR}) \\ &\leq P(T_n > cv \cap h_1 \in \text{CR}) + P(h_1 \notin \text{CR}). \end{aligned}$$

If $h_1 \in \text{CR}$: $cv(1-\alpha) = \sup_{\bar{h}_1 \in \text{CR}} c_{(\bar{h}_1, \widehat{h}_{2,n})}(1-\alpha) \geq c_{(h_1, \widehat{h}_{2,n})}(1-\alpha)$. So

$$P(T_n > cv \cap h_1 \in \text{CR}) \leq P(T_n > c_{(h_1, \widehat{h}_{2,n})}(1-\alpha)) \to \alpha.$$

$P(h_1 \notin \text{CR}_{h_1,n}(1-\beta)) \to \beta$. Total: $\to \alpha + \beta$. ■

(d) Use $\delta = \alpha - \beta$. Replace α in CV with δ. By (c), rejection prob $\leq \delta + \beta = \alpha$. AsySz $\leq \alpha$.

(e) Power. Bonferroni CV uses sup over $\text{CR}_{h_1,n}(1-\beta)$ (subset of $\mathbb{R}^p$). Worst-case uses sup over all of $\mathbb{R}^p$. So Bonferroni CV $\leq$ worst-case CV $\Rightarrow$ Bonferroni more powerful in general.

Trade-off. Bonferroni uses $1-\delta$ quantile (with $\delta < \alpha$), which is larger than $1-\alpha$ quantile. Net effect: typically more powerful, but not uniformly. *Why "Bonferroni".* The decomposition in (c) is the Bonferroni inequality $P(A \cup B) \leq P(A) + P(B)$. Splits size budget $\alpha = \delta + \beta$.

11.7.3 HW10 Q3: NP and Residual Bootstrap CI Simulation

[LOW ROI — simulation; remember the qualitative comparison]

Problem. Compute and compare three confidence intervals for $\beta_{(1)}$ in the linear regression model with $\beta = (1,1,1,1)'$, n iid $N(0,1)$ regressors and errors. Use $B = 20 \cdot 50 - 1 = 999$.

(i) NP iid bootstrap; (ii) homoskedastic residual-based bootstrap; (iii) standard CI based on delta method. Report critical values, CIs, coverage probabilities for (i) and (ii).

What this tests. Bootstrap higher-order accuracy demonstration.

Approach. Don't memorize numbers; remember pattern. Bootstrap CIs slightly wider than delta CI; coverage closer to nominal.

Solution.

Reported values:

- Critical values: (i) 2.0654, (ii) 2.2045, (iii) 1.9600 ($= z_{0.975}$).
- CIs around $\beta_{(1)} = 1$: (i) $[0.7645, 1.2471]$, (ii) $[0.7483, 1.2633]$, (iii) $[0.7768, 1.2348]$.
- Coverage: (i) 0.9466, (ii) 0.9474. Both close to nominal 0.95.

Interpretation. Bootstrap CIs slightly wider (CV > 1.96) but coverage closer to nominal in finite samples — captures the $n^{-1/2}$ Edgeworth correction missed by the normal approximation (Chapter 6).

Why residual differs from NP. Residual: condition on X, resample errors only. NP: resample (Y, X) pairs. Both valid in homoskedastic linear model; finite-sample distributions slightly differ.

11.8 Quick Self-Test Index

Use this as your final-week drill checklist. Rate yourself: can you do each in ≤ 10 minutes without notes?

1. HW6 Q2: Four-step proof of $\text{LM}_{\text{CUE}} \xrightarrow{d} \chi^2$ under strong IV. *(CORE)*
2. HW7 Q1: Ridge bias and variance from scratch. *(CORE)*
3. HW7 Q2: Sub-Gaussian implies $\mathbb{E}(X) = \mu$ and $\text{Var}(X) \leq \sigma^2$. *(CORE)*
4. HW7 Q3: Soft thresholding via ℓ_1 penalty (case-by-case). *(CORE)*
5. HW8 Q2: Bootstrap failure at the boundary (with the $\max\{Z - c, 0\}$ argument). *(CORE)*
6. HW9 Q1: Heckit inverse Mills derivation. *(CORE)*
7. HW9 Q2(a): Mixed IV identification under $\text{rk}(\Pi_2) = d_x$. *(CORE)*
8. HW9 Q3(b): CF $\not\Rightarrow$ IV exogeneity counter-example (the discrete table). *(CORE)*
9. HW9 Q4: USCON + ID $\Rightarrow$ a.s. consistency (contradiction proof). *(CORE)*

If you can do all 9 in under 10 minutes each, with the right "why each step" verbal annotations, you have $\geq 7/10$ on Q3 of the final locked in. That is your Tier-1 deliverable.

Final Drill Schedule (4 Days)

- **Day −4 (May 2 today)**: Drill HW7 Q1, Q2, Q3 (Ridge, sub-Gaussian, thresholding). All three are clean closed-form, high probability of appearing.
- **Day −3 (May 3)**: Drill HW6 Q2 (LM_CUE strong IV) + HW9 Q4 (USCON consistency) + HW9 Q3(b) (counter-example).
- **Day −2 (May 4)**: Drill HW9 Q1 (Heckit) + HW9 Q2 (mixed IV + NPV) + HW8 Q2 (bootstrap failure). Mock midterm 2025 for timing.
- **Day −1 (May 5)**: Light review of cheat sheets in each chapter. *Do not study new material.* Sleep early.