

Exam 2 Review

Gov 51 Section — Week 13 | Everything Since the Midterm

George

Harvard University

April 22, 2026

Six acts, one goal: understand what your estimate actually measures

Act	Topic	Exam weight
I	Potential outcomes, ATT/ATE, selection bias	~16 pts
II	Prediction, RMSE, regularization, CV	~12 pts
III	Omitted variable bias	~6 pts
IV	IV conditions, Wald estimator, 2SLS	~8 pts
V	F-statistic, weak instruments	~6 pts
VI	LATE and compliers	~6 pts

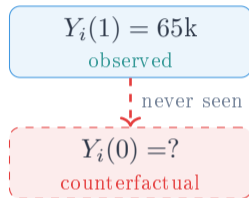
Exam: Thu Apr 23 | 75 min | one cheat sheet: two pages, front and back



Act I: Potential Outcomes
ATT, ATE, and Selection Bias

Maria went to college. Would she have earned more anyway?

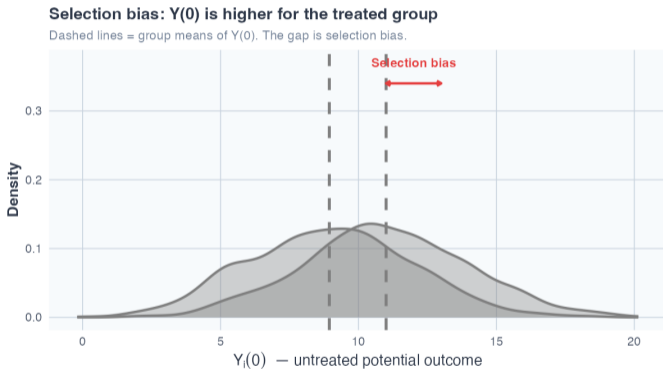
- ▶ **Observed:** Maria went to college ($D_i = 1$); earns \$65k
- ▶ **Missing:** What would she earn *without* college? — $Y_i(0)$
- ▶ **Individual effect:** $\tau_i = Y_i(1) - Y_i(0)$ — *never observed*



Without the counterfactual, the individual treatment effect is unobservable.

That is the **fundamental problem of causal inference**.

The simulation: ability creates selection bias



Even with *zero* treatment effect, the groups look different. That gap is selection bias.

The decomposition: $SDO \neq ATT$

$$\underbrace{\bar{Y}_{D=1} - \bar{Y}_{D=0}}_{\text{SDO (what you see)}} = \underbrace{E[Y_i(1) - Y_i(0) | D_i = 1]}_{\text{ATT (what you want)}} + \underbrace{E[Y_i(0) | D_i = 1] - E[Y_i(0) | D_i = 0]}_{\text{selection bias}}$$

- ▷ Selection bias = would treated units have fared differently *even without treatment*?
- ▷ Positive when high- $Y(0)$ units select into treatment (like ability \rightarrow college)
- ▷ Randomization sets it to zero: $E[Y_i(0) | D = 1] = E[Y_i(0) | D = 0]$

In an RCT: $SDO = ATT$. In observational data: almost never.

Three estimands: know which one you are computing

Estimand	Formula	Average over...
ATT	$E[Y_i(1) - Y_i(0) \mid D_i = 1]$	treated units only
ATC	$E[Y_i(1) - Y_i(0) \mid D_i = 0]$	control units only
ATE	$p \cdot \text{ATT} + (1 - p) \cdot \text{ATC}$	all units

p = share treated. ATE is a share-weighted average of ATT and ATC.

On the exam: compute ATT first, then ATC, then combine.

Practice 1 — Work this before checking

Practice 1: five workers, job training

i	D_i	$Y_i(1)$	$Y_i(0)$	Y_i^{obs}
A	1	6	10	6
B	1	8	12	8
C	1	4	8	4
D	0	3	5	5
E	0	4	6	6

- (a) Calculate the SDO.
- (b) Calculate the ATT and ATC.
- (c) Calculate the ATE using the weighted formula.
- (d) Calculate the selection bias.
Verify: $\text{SDO} = \text{ATT} + \text{selection bias}$.
- (e) Is selection bias positive or negative?
What does that say about who sought training?

Practice 1: solutions

(a) **SDO:**

Treated obs: $\{6, 8, 4\}$, mean = 6

Control obs: $\{5, 6\}$, mean = 5.5

SDO = $6 - 5.5 = +0.5$ (looks bad)

(b) **ATT:** $\frac{(6-10)+(8-12)+(4-8)}{3} = -4$

ATC: $\frac{(3-5)+(4-6)}{2} = -2$

(c) **ATE:** $0.6(-4) + 0.4(-2) = -3.2$
(actually works)

(d) **Selection bias:**

$E[Y(0) | D = 1] = (10 + 12 + 8)/3 = 10$

$E[Y(0) | D = 0] = (5 + 6)/2 = 5.5$

Bias = $10 - 5.5 = +4.5$

Verify: $-4 + 4.5 = +0.5 \checkmark$

(e) Positive bias: workers with *highest* untreated potential took training. Hardest cases select in — making a successful program look useless.

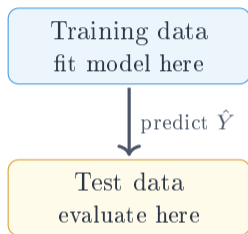


Act II: Prediction and Machine Learning
RMSE, Regularization, Cross-Validation

Prediction goal: minimize error on data you have not seen

The story:

- ▷ You have a training dataset
- ▷ You fit a model on it
- ▷ Your model will be applied to *new* data
- ▷ In-sample fit doesn't guarantee out-of-sample fit

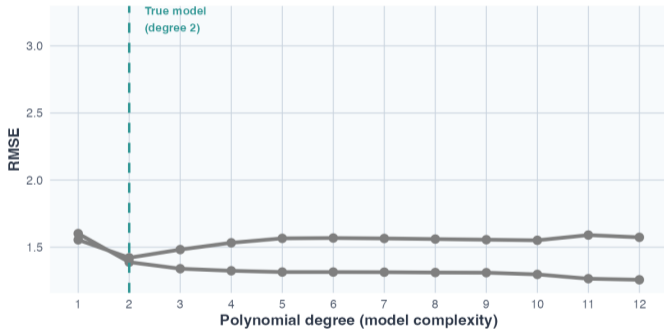


$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{Y}_i - Y_i)^2}$$

The simulation shows the tradeoff clearly

Overfitting: train RMSE keeps falling; test RMSE rises

Adding polynomial terms always improves in-sample fit — but hurts out-of-sample



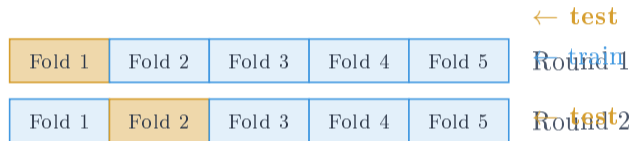
Train RMSE falls monotonically. Test RMSE has a minimum. Regularization (λ) keeps you near it.

Three regularization methods — know their differences

Method	Penalty	Zeros?	Key use
Ridge	L2: $\lambda \sum_j \beta_j^2$	No	many small effects; correlated X 's
LASSO	L1: $\lambda \sum_j \beta_j $	Yes	variable selection; sparse models
Elastic Net	α L1 + $(1-\alpha)$ L2	Yes	correlated predictors; compromise

- ▷ **Larger** $\lambda \Rightarrow$ more shrinkage \Rightarrow higher bias, lower variance
- ▷ **LASSO zeros** = penalty cost exceeded predictive benefit — *not* proof of no effect
- ▷ **Proxy discrimination**: excluding race doesn't stop racially biased predictions if correlated variables remain

k -fold CV: use every observation as a test point



1. Divide data into k folds; for each fold: train on remaining $k - 1$, test on this one
2. Average k test RMSEs \Rightarrow CV RMSE for this λ
3. **Pick** λ that minimizes CV RMSE — never use in-sample RMSE to choose

Practice 2 — Quick calculations

Practice 2: regularization and CV

- (a) A LASSO model at λ_{\min} has CV MSE = 0.198. What is the RMSE?
- (b) You run 5-fold CV and observe fold-level test RMSEs:
0.452, 0.448, 0.461, 0.455, 0.444
What is the 5-fold CV RMSE?
- (c) “LASSO set 60 of my 100 predictors to zero — those variables have no effect.”
What is wrong?
- (d) Ridge and LASSO give nearly identical out-of-sample RMSE. Which do you prefer for *explanation*? Why?

Practice 2: solutions

(a) $\text{RMSE} = \sqrt{0.198} \approx \mathbf{0.445}$

(b) Sum = 2.260, mean = **0.452**

(c) LASSO zeros reflect the *penalty cost*, not true zero effects. A zero coefficient means LASSO decided the predictor wasn't worth including given the regularization budget. The true effect may be nonzero.

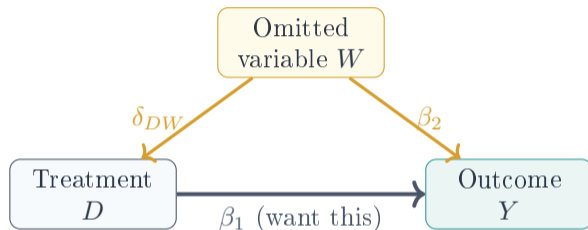
(d) LASSO — produces a *sparse* model (fewer predictors) that is easier to interpret and communicate. If prediction accuracy is identical, parsimony wins for explanation.

Remember: prediction \neq causation. A LASSO coefficient on prior arrests predicts recidivism; it does *not* say prior arrests *cause* future crime.



Act III: Omitted Variable Bias
The Formula, the Sign, and Why It Persists

Omitted variable bias has a formula — and a sign



$$\hat{\beta}_1^{\text{OLS}} \xrightarrow{p} \beta_1 + \underbrace{\beta_2 \cdot \delta_{DW}}_{\text{OVB}} \quad \text{where } \delta_{DW} = \frac{\text{Cov}(D, W)}{\text{Var}(D)}$$

The sign is everything: predict it before you run the regression

β_2 ($W \rightarrow Y$)	δ_{DW} ($W \rightarrow D$)	OVB	OLS is...
+	+	+	too large
-	-	+	too large
+	-	-	too small
-	+	-	too small

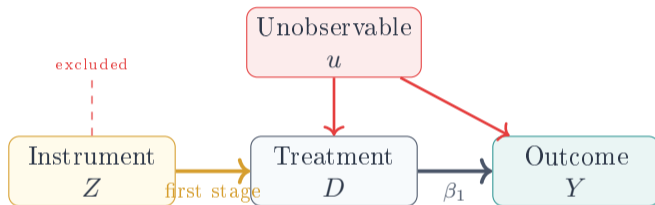
Classic case: Ability \rightarrow schooling (+) and ability \rightarrow wages (+) \Rightarrow OVB $>$ 0, OLS too large.

Counterforce: measurement error in D causes attenuation bias toward zero. Both can operate at once.



Act IV: Instrumental Variables
Three Conditions, Wald, and 2SLS

The IV idea: find variation in D that is not contaminated by u



Z must move D (**relevance**) but affect Y *only through* D (**exclusion**) and be uncorrelated with u (**independence**).

A valid instrument satisfies three conditions

1. **Relevance:** $\text{Cov}(Z_i, D_i) \neq 0$ testable: first-stage $F > 10$
2. **Exclusion:** $Z \rightarrow Y$ only through D not testable — argued from theory
3. **Independence:** $Z \perp U$ not testable — argued from theory

Strangeness principle: a good instrument has a reduced-form correlation with Y that seems *bizarre* — until you hear what D is.

Z	D	Y
College proximity	Years of schooling	Wages (Card 1995)
Settler mortality	Institutional quality	GDP (AJR 2001)
Birth quarter	Years of schooling	Earnings (AK 1991)

The Wald estimator rescales the reduced form by the first stage

First stage (FS): $\hat{\alpha}_1 = E[D \mid Z = 1] - E[D \mid Z = 0]$

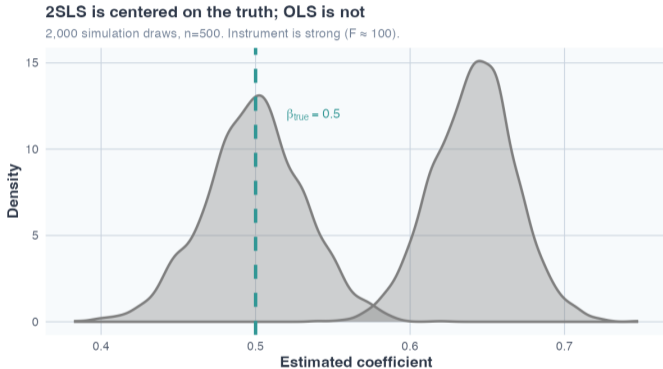
Reduced form (RF): $\hat{\pi}_1 = E[Y \mid Z = 1] - E[Y \mid Z = 0]$

$$\text{Wald} = \frac{\hat{\pi}_1}{\hat{\alpha}_1} = \frac{\text{how much } Z \text{ shifts } Y}{\text{how much } Z \text{ shifts } D}$$

- ▷ Interpretation: per unit of D caused by Z , how much does Y change?
- ▷ With one instrument, one endogenous variable: Wald = 2SLS

Card (1995): FS = 0.327, RF = 0.041,
Wald = 0.125. OLS = 0.071.
2SLS > OLS — next act explains why.

The simulation: 2SLS is centered on truth; OLS is not



2,000 draws. OLS shifted right (ability bias). 2SLS centered on truth but wider — that width is the variance cost of IV.

2SLS in two steps — and one important warning

Step 1: Regress D_i on Z_i \Rightarrow get \hat{D}_i (predicted treatment)

Step 2: Regress Y_i on \hat{D}_i \Rightarrow coefficient is $\hat{\beta}_{2SLS}$

Do **NOT** run the two stages by hand and read off the SE.
Stage 2 SEs from manual OLS ignore estimation error in Stage 1 — they are too small and wrong.
Always use `iv_robust(Y ~ D | Z, data = df)` to get correct SEs.

```
library(estimatr)
fit_iv <- iv_robust(log_wage ~ college | proximity, data =
  card)
tidy(fit_iv)
```

Practice 3 — OVB and IV

Practice 3: OVB and IV logic

A researcher regresses **log income** (Y) on **years of college** (D). Family wealth (W) is omitted.

- (a) Write the OVB formula. Define every term.
- (b) Wealth \rightarrow more schooling ($\delta_{DW} > 0$) and wealth \rightarrow higher income ($\beta_2 > 0$). Sign of OVB?
- (c) The researcher uses distance to nearest 4-year college as an instrument. State the three IV conditions and assess each briefly.
- (d) FS = 0.31, RF = 0.038. Calculate the Wald estimate and interpret it in one sentence.

Practice 3: solutions

(a) $\hat{\beta}_1^{\text{OLS}} \rightarrow \beta_1 + \beta_2 \cdot \delta_{DW}$

β_1 : true return to schooling

β_2 : effect of wealth on income

δ_{DW} : how schooling predicts wealth

(b) OVB = (+)(+) > 0 OLS
overstates return to schooling

(c) Relevance: $F = 16.5$ — passes

Exclusion: possible violation: wealthy families cluster near colleges

Independence: same concern

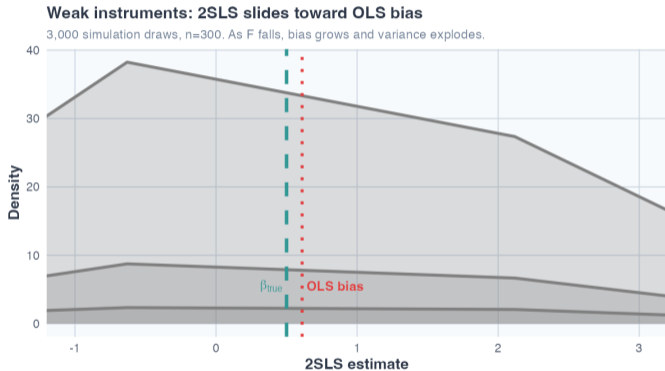
(d) Wald = $0.038/0.31 \approx \mathbf{0.123}$

Among men who attend college *only* because they grew up near one, an additional year of college raises log wages by ≈ 0.123 (about 13%).



**Act V: F-Statistic and Weak Instruments
When IV Bias Sneaks Back In**

The simulation: as F falls, 2SLS slides toward OLS bias



Strong instrument (blue): 2SLS is near β_{true} . Weak instrument (red): distribution slides right toward the OLS bias *and* explodes in width. Two problems at once.

Weak instruments inherit OLS bias — quantified

$$\text{Bias}(\hat{\beta}^{2SLS}) \approx \text{Bias}(\hat{\beta}^{OLS}) \times \frac{1}{F + 1}$$

F	p-value	2SLS bias (of OLS)	Stock-Yogo verdict
3.84	= 0.05	21%	Fail
5	< 0.05	17%	Fail
10	threshold	9%	Pass
16	AJR (orig.)	6%	Strong

$p < 0.05 \neq$ strong instrument. Statistical significance is a *different* question from instrument strength. Stock-Yogo threshold: $F > 10$.

Three consequences of a weak first stage

1. Bias toward OLS

2SLS inherits the contamination it was designed to cure

2. Inflated variance

$$\text{Var}(\hat{\beta}_{IV}) = \frac{\text{Var}(\hat{\beta}_{OLS})}{\rho_{ZD}^2} \quad \text{— explodes as } \rho_{ZD} \rightarrow 0$$

3. Invalid confidence intervals

Normal approximation breaks down; 2SLS SE is not reliable

Fix: use **Anderson-Rubin** confidence intervals (valid regardless of F)

Adding more weak instruments makes it *worse*: spreading predictive power thinner reduces F further. The AK (1991) result with 180 quarter-of-birth dummies is an example.

Practice 4 — Weak instruments and LATE

Practice 4: weak instrument calculation

- (a) An IV study reports first-stage $F = 4$. OLS has omitted-variable bias of $+0.050$.
 - (i) Calculate the approximate 2SLS bias.
 - (ii) If the true effect is 0.15, what does OLS estimate? What does 2SLS estimate?
- (b) Explain in 2–3 sentences why 2SLS is *consistent* as $n \rightarrow \infty$ but *biased* in finite samples.
- (c) Albouy (2012) shows AJR's settler mortality data has errors that drop F from 16 to ≈ 4 . What happens to the 2SLS estimate? What happens to the confidence intervals?

Practice 4: solutions

(a)(i) $\text{Bias}(2\text{SLS}) \approx 0.050 \times \frac{1}{5} = \mathbf{0.010}$

(a)(ii) $\text{OLS} \approx 0.15 + 0.05 = \mathbf{0.200}$

$2\text{SLS} \approx 0.15 + 0.01 = \mathbf{0.160}$

(b) Exclusion makes $\text{Cov}(Z, u) = 0$, so the probability limit of Wald goes to the true β . In finite samples, a noisy $\hat{\alpha}_1$ in the denominator introduces bias. $\text{Bias} \propto 1/F$, so it vanishes as the first stage strengthens with n .

(c) At $F \approx 4$:

- ▶ 2SLS inherits $\approx 20\%$ of OLS bias — no longer credibly causal
- ▶ Standard errors inflate
- ▶ CIs based on normal 2SLS distribution are invalid — results lose significance
- ▶ Lesson: a contested F from noisy data is not evidence of instrument strength



Act VI: LATE
What IV Actually Estimates — and for Whom

IV does not estimate the population average treatment effect

Type	When Z changes	Card example
Compliers	$D(Z = 0) = 0 \rightarrow D(Z = 1) = 1$	attend college <i>because</i> nearby
Always-takers	$D = 1$ regardless	attend regardless of proximity
Never-takers	$D = 0$ regardless	never attend regardless

$$\mathbf{LATE} = E[Y_i(1) - Y_i(0) \mid \text{complier}]$$

IV identifies the effect *only* for compliers. Always-takers and never-takers never change D when Z changes — they contribute nothing to the first stage.

LATE exceeded ATE in Card (1995) — and here is why

What OVB predicted:

Ability bias pushes OLS up \Rightarrow IV should come in *below* OLS.

What Card found:

OLS = 0.071, 2SLS = 0.125
2SLS is almost *twice* as large.

Card's explanation:

College proximity selects *credit-constrained* students — not low-ability, but low-income, on the margin. That group may have the *highest* returns.

If compliers have *above-average* returns to education, then $LATE > ATE$ — and $2SLS > OLS$ makes perfect sense.

Measurement error in D also causes attenuation in OLS — another reason 2SLS might exceed OLS even if $LATE = ATE$.

Write a LATE sentence: three components, always

A LATE interpretation sentence must include:

1. **The effect:** what outcome, what treatment change
2. **The population:** who are the compliers?
3. **The source of variation:** what does the instrument move?

Card example: 2SLS estimates the causal return to an additional year of college on log wages for men who attend college only because they grew up near one (credit-constrained men on the margin of attendance), where variation comes from whether a 4-year college was located in their county of residence.


What to put on your cheat sheet

Must-have formulas:

- ▷ SDO = ATT + selection bias
- ▷ ATE = $p \cdot \text{ATT} + (1 - p) \cdot \text{ATC}$
- ▷ $\text{RMSE} = \sqrt{\frac{1}{n} \sum (\hat{Y}_i - Y_i)^2}$
- ▷ OVB: $\hat{\beta}_1 \rightarrow \beta_1 + \beta_2 \cdot \delta_{DW}$
- ▷ Wald = $\hat{\pi}_1 / \hat{\alpha}_1$
- ▷ 2SLS bias $\approx \text{OLS bias} \times \frac{1}{F+1}$
- ▷ IV variance = $\frac{\text{OLS variance}}{\rho_{ZD}^2}$
- ▷ LATE = $E[Y(1) - Y(0) \mid \text{complier}]$

Common exam mistakes:

- ▷ Confusing SDO with ATT
- ▷ Wrong sign on OVB
- ▷ $p < 0.05 \neq$ strong instrument
- ▷ Interpreting 2SLS as population ATE
- ▷ Running 2SLS by hand and using wrong SEs
- ▷ Saying LASSO zeros mean true zeros
- ▷ Forgetting $n - 1$ in variance formulas



The exam asks one question over and over:

does your number measure a *causal* effect?

Selection bias says no.

Randomization, IV, and DiD say how to get there.

Questions?

Exam: Thursday April 23 | Good luck