Gov 51: Final Exam Review Guide

What to Know, What to Skip, and Practice Problems

Spring 2026

Exam: May 2026 · 75 minutes · Cheat sheet allowed

Exam format: 100 points, 75 minutes. You may use a **calculator**. You may bring **one cheat sheet: two standard pages ($8.5'' \times 11''$), front and back** — same as Exam 1. No computers, no internet, no AI. Show all work for calculations — partial credit is available.

A note on this review guide: This document covers *more* material than will appear on the exam. That is intentional — knowing which topics *could* appear forces you to understand all of them, and the exam will draw from this set without testing everything in it. The "What You Do NOT Need to Know" section tells you what is off the table entirely.

1 What You Need to Know

1.1 Potential Outcomes and the Fundamental Problem

This section is testable as pure calculation. You will be given a table of potential outcomes and asked to compute quantities by hand. Practice until the formulas are automatic.

- **Potential outcomes:** $Y_i(1)$ = outcome unit i would have if treated; $Y_i(0)$ = outcome if untreated. Only one is ever observed.
- **Individual treatment effect:** $\tau_i = Y_i(1) - Y_i(0)$ (never observed directly)
- **Simple Difference in Outcomes (SDO):** $\bar{Y}_{D=1} - \bar{Y}_{D=0}$ using observed data
- **ATT** (Average Treatment Effect on the Treated): $E[Y_i(1) - Y_i(0) \mid D_i = 1]$ — average effect for those who actually got treatment
- **ATC** (Average Treatment Effect on Controls): $E[Y_i(1) - Y_i(0) \mid D_i = 0]$ — average effect for the untreated group
- **ATE** (Average Treatment Effect): $E[Y_i(1) - Y_i(0)]$ — average over *all* units
- **ATE formula:** $\text{ATE} = p \cdot \text{ATT} + (1 - p) \cdot \text{ATC}$, where p is the share treated
- **Selection bias:** $E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]$ — would treated and control units have had the same baseline outcome if neither had been treated?

The decomposition (memorize this):

$$\underbrace{\text{SDO}}_{\text{observed}} = \underbrace{\text{ATT}}_{\text{causal}} + \underbrace{\big[E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]\big]}_{\text{selection bias}}$$

Selection bias is zero when treatment is as-good-as random. In observational data it is almost always nonzero.

- **Why SDO misleads:** treated units often differ from control units in ways that affect Y even without treatment — that gap *is* selection bias
- **Direction of selection bias:** depends on the context. In peacekeeping, conflict-prone countries receive missions $\Rightarrow E[Y(0) \mid D = 1] > E[Y(0) \mid D = 0] \Rightarrow$ positive selection bias makes missions look harmful when they aren't

1.2 Prediction and Machine Learning (PS3: COMPAS)

- **RMSE:** $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}$ — average prediction error in units of Y
- **In-sample vs. out-of-sample RMSE:** in-sample is computed on the training data (always lower); out-of-sample is computed on held-out test data (what actually matters for prediction)
- **Overfitting:** a model that memorizes training data (low train RMSE) but predicts new data poorly (high test RMSE). Adding predictors always improves in-sample fit but can hurt out-of-sample performance.
- **Train/test split:** randomly hold out ∼20% of data for evaluation; fit the model only on the remaining 80%
- **k-fold cross-validation:** divide the data into k equal groups (folds); for each fold, fit the model on the other $k-1$ folds and predict the held-out fold; average the k test RMSEs. More stable than a single train/test split because every observation is a test observation exactly once.
- **Why use CV to pick λ?** The optimal penalty λ is unknown before fitting. CV measures out-of-sample performance for each candidate λ and selects the one that minimizes average test error.

Three regularization methods:

Method	Penalty	Sets coefs to 0?	Key property
Ridge	L2: $\lambda \sum_j \beta_j^2$	No	Good when many small effects
LASSO	L1: $\lambda \sum_j \lvert\beta_j\rvert$	Yes	Sparse; performs variable selection
Elastic Net	αL1 $+(1-\alpha)$L2	Yes	Compromise; handles correlated predictors

- **Increasing λ:** more regularization ⇒ more shrinkage ⇒ higher bias, lower variance. In-sample RMSE increases; out-of-sample RMSE may improve or worsen depending on where you start relative to the optimum.
- **Bias-variance tradeoff:** models with too many parameters have low bias but high variance (overfit); models with too few have high bias but low variance (underfit); optimal is somewhere in between
- **LASSO and fairness (Baker 2025):** LASSO reduces subjective discretion in model building — the same data + penalty produces the same variable selection, unlike expert-driven variable choice which is manipulable
- **Prediction ≠ causation:** a LASSO coefficient on `priors_count` tells us prior offenses predict recidivism; it does NOT tell us that prior offenses cause future recidivism. Prediction models can contain selection bias.
- **Proxy discrimination:** race may be excluded from the model, but variables correlated with race (neighborhood, criminal history shaped by policing disparities) can produce racially disparate predictions

1.3 Omitted Variable Bias

- **OVB formula:** if you omit W from a regression of Y on D:

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

- OVB $= \beta_2 \times \delta_{DW}$ where δ_{DW} is the regression of W on D (the "auxiliary regression" coefficient)
- **Sign of OVB:** positive if the omitted variable has the same signed effect on D and Y; negative otherwise
- **Classic example:** ability (W) is omitted from a wage regression. Ability → more schooling (D) and ability → higher wages (Y), so both $\beta_2 > 0$ and $\delta_{DW} > 0 \Rightarrow$ OVB $> 0 \Rightarrow$ OLS overstates the return to education
- **Measurement error:** if D is measured with classical error, OLS is biased *toward zero* (attenuation bias). This can explain why OLS understates returns when 2SLS is larger.

1.4 Instrumental Variables: The Three Conditions

An instrument Z is valid if:

1. **Relevance:** $\text{Cov}(Z_i, D_i) \neq 0$ — Z shifts the treatment. *Testable: check first-stage $F > 10$.*
2. **Exclusion:** Z affects Y *only* through D — no direct path $Z \to Y$. *Not testable; argued from theory.*
3. **Independence:** $Z \perp U$ — no backdoor path from Z through unobservables. *Not testable; argued from theory.*

- **Strangeness principle:** a good instrument has a reduced-form correlation with Y that seems bizarre — until you hear what the treatment D is. Strangeness $\Leftrightarrow$ exclusion: if the path from Z to Y only runs through D, then knowing D makes everything make sense.
- **Card (1995):** college proximity $\to$ wages seems strange until you know $D =$ years of education
- **AJR (2001):** settler mortality in 1820 $\to$ GDP in 1995 seems strange until you know $D =$ institutional quality
- **Angrist & Krueger (1991):** birth quarter $\to$ earnings seems strange until you know $D =$ years of schooling (compulsory schooling laws)

1.5 The Wald Estimator and 2SLS

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

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

$$\text{Wald estimator:} \quad \hat{\delta}_{\text{Wald}} = \frac{\hat{\pi}_1}{\hat{\alpha}_1} = \frac{\text{Reduced form}}{\text{First stage}}$$

2SLS = Wald when one instrument, one endogenous variable (just-identified)

- **Intuition for Wald:** RF tells us how much Z shifts Y in total; FS tells us how much Z shifts D. Dividing rescales: "per unit of D caused by Z, how much does Y change?"
- **2SLS procedure:** Stage 1: regress D_i on Z_i, get $\hat{D}_i$. Stage 2: regress Y_i on $\hat{D}_i$. The second-stage coefficient is $\hat{\beta}_{\text{2SLS}}$.
- **Why not do Stage 2 by hand?** Standard errors from manual two-step are wrong — use `iv_robust()` which corrects for Stage 1 estimation error.
- **Card (1995) numbers:** FS = 0.327, RF = 0.041, Wald = 0.041/0.327 $\approx$ 0.125. OLS = 0.071. 2SLS > OLS.

1.6 The F-Statistic and Weak Instruments

- **General F-test:** compares unrestricted vs. restricted model. Measures fit lost by imposing a set of restrictions ($H_0 : \beta_{k-q+1} = \cdots = \beta_k = 0$).

$$F = \frac{(RSS_R - RSS_U)/q}{RSS_U/(n-k-1)}$$

- **IV first-stage F:** special case where the restriction is $H_0 : \alpha_1 = 0$ (instrument has no predictive power). With one instrument, $F = t^2$.
- **Stock–Yogo threshold:** $F > 10$ limits finite-sample 2SLS bias to less than 10% of OLS bias
- $p < 0.05$ **is not enough:** $p < 0.05 \Leftrightarrow F > 3.84$. Bias is severe even at $F = 5$ or $F = 8$. Statistical significance and instrument strength are different questions.
- **Finite-sample bias formula:**

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

  Small $F \Rightarrow$ 2SLS inherits most of OLS bias. At $F = 3$: 2SLS bias $\approx$ OLS bias /4.
- **Adding weak instruments makes it worse:** 180 weak AK instruments $\Rightarrow F \approx 1 \Rightarrow$ 2SLS bias $\approx$ OLS bias. More instruments spread the same total predictive power thinner.
- **Consequences of weak instruments:** (1) bias toward OLS; (2) inflated standard errors; (3) invalid confidence intervals (non-normal sampling distribution of $\hat{\beta}_{\text{2SLS}}$)

1.7 LATE: What IV Estimates

- IV does not estimate the ATE. It estimates the **Local Average Treatment Effect (LATE)** — the average treatment effect for **compliers**.
- **Compliers:** units whose treatment status changes because of the instrument. $D_i(Z = 1) = 1$ but $D_i(Z = 0) = 0$.
- **Always-takers:** get treated regardless of Z. **Never-takers:** never get treated regardless of Z.
- **Card (1995) compliers:** men who went to college because they happened to grow up near one, but would not have gone otherwise. Not always-takers (who would have attended regardless) and not never-takers.
- **Why LATE may differ from ATE:** compliers in Card tend to be from lower-income families on the margin of attending college; returns to education may be higher precisely for those at the margin (they have the most to gain)
- **Writing a LATE sentence:** name (i) the effect (return to a year of education on log wages), (ii) the population (compliers: men who attend college only because they live near one), (iii) the source of variation (whether a 4-year college is in one's county)

1.8 The AJR Paper and Its Critique ("Extra" Material)

- **AJR instrument:** Z = log settler mortality (1817–1848). High mortality → few settlers → extractive institutions → lower GDP today.
- **Exclusion argument:** mortality in 1820 cannot directly affect GDP in 1995; the only channel is institutional quality
- **Independence argument:** mortality rates were determined by disease environment, not by pre-existing economic conditions
- **Albouy (2012) critique:** mortality data assigned to wrong countries; corrected data drops F from 16 to $\approx 4 \Rightarrow$ results lose significance. Nobel committee sided with AJR.
- **Lesson:** a high F from noisy or mismeasured instrument data is not evidence of a strong instrument — it is evidence that the noisy measure correlates with D

1.9 Anderson-Rubin Confidence Intervals ("Extra" Material)

- Standard 2SLS CIs ($\hat{\beta} \pm 1.96 \cdot \widehat{SE}$) break down when $F < 10$: the normal approximation fails because $\hat{\pi}/\hat{\alpha}$ has heavy tails when $\hat{\alpha} \approx 0$
- **Anderson-Rubin CI:** invert a test of $H_0 : \beta = \beta_0$. For each candidate β_0, check whether Z can explain residual $Y - \beta_0 D$. The CI is the set of β_0 values we cannot reject.
- AR CI is valid regardless of instrument strength. When $F > 10$: AR CI $\approx$ standard CI. When F is small: AR CI may be very wide or even unbounded.

2 What You Do NOT Need to Know

- R syntax or code — you will not be asked to write or debug code
- The exact `cv.glmnet()` or `iv_robust()` syntax
- Proof of any formula (you need to use formulas, not derive them)
- Adjusted R^2, AIC, BIC, or Mallow's C_p
- The KS test or any distribution test
- Weighted means, sampling weights, or survey design
- Continuous × continuous interactions
- Standard error formula for regression coefficients (you will be given SEs)
- The specific technical details of how LASSO coordinates descent is computed
- Anderson-Rubin calculations (conceptual understanding only)
- Miguel, Satyanath & Sergenti (2004) in detail (though the idea that rainfall instruments for growth is fair game as an example)
- The Albouy critique in depth (know the headline: contested data weakens the first stage)

3 Formula Reference (Cheat Sheet Suggestions)

You have **one side** of one page. Be selective. These are the formulas most likely to matter:

SDO	$\bar{Y}_{D=1} - \bar{Y}_{D=0}$
ATT	$E[Y_i(1) - Y_i(0) \mid D_i = 1]$
ATC	$E[Y_i(1) - Y_i(0) \mid D_i = 0]$
ATE	$p \cdot \text{ATT} + (1-p) \cdot \text{ATC}$
Selection bias	$E[Y_i(0) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0]$
SDO decomposition	$\text{SDO} = \text{ATT} + \text{selection bias}$
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}$
OVB formula	$\hat{\beta}_1^{\text{OLS}} \to \beta_1 + \beta_2 \cdot \delta_{DW}$ where $\delta_{DW} = \text{Cov}(D,W)/\text{Var}(D)$
First stage	$\hat{\alpha}_1 = E[D \mid Z = 1] - E[D \mid Z = 0]$
Reduced form	$\hat{\pi}_1 = E[Y \mid Z = 1] - E[Y \mid Z = 0]$
Wald	$\hat{\delta} = \hat{\pi}_1 / \hat{\alpha}_1$
F-test (IV)	$F = t^2$ (one instrument); Stock-Yogo: $F > 10$
2SLS bias	$\text{Bias}(\hat{\beta}^{\text{2SLS}}) \approx \text{Bias}(\hat{\beta}^{\text{OLS}}) \times \frac{1}{F+1}$
IV variance	$\text{Var}(\hat{\beta}_{\text{IV}}) = \frac{\text{Var}(\hat{\beta}_{\text{OLS}})}{\rho_{ZD}^2} \geq \text{Var}(\hat{\beta}_{\text{OLS}})$

4 Practice Problems with Solutions

Work these with pencil and paper before checking the solutions. Aim for about 1 minute per point — a 10-point problem should take roughly 10 minutes.

Practice 1: Potential Outcomes (16 points)

A researcher studies whether job training reduces unemployment duration. She observes five workers. Three receive job training ($D_i = 1$); two do not. The outcome Y_i is weeks until reemployment (lower is better). Both potential outcomes are given.

Worker	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) (3 pts) Calculate the SDO.

(b) (4 pts) Calculate the ATT and ATC.

(c) (3 pts) Calculate the ATE using the weighted formula.

(d) (4 pts) Calculate the selection bias. Verify: SDO = ATT + selection bias.

(e) (2 pts) Is the selection bias positive or negative here? What does it mean about which workers seek out job training?

Solutions:
(a) SDO: Treated observed outcomes: $\{6, 8, 4\}$, mean $= 6$. Control observed outcomes: $\{5, 6\}$, mean $= 5.5$. SDO $= 6 - 5.5 = +0.5$.
(b) ATT: Effect for treated workers: $(6-10) + (8-12) + (4-8) = -4-4-4 = -12$. ATT $= -12/3 = -4$.
ATC: Effect for control workers: $(3-5) + (4-6) = -2-2 = -4$. ATC $= -4/2 = -2$.
(c) ATE: $p = 3/5 = 0.6$. ATE $= 0.6 \times (-4) + 0.4 \times (-2) = -2.4 - 0.8 = -3.2$.
(d) Selection bias: $E[Y(0) \mid D=1] - E[Y(0) \mid D=0]$. $E[Y(0) \mid D=1] = (10+12+8)/3 = 10$. $E[Y(0) \mid D=0] = (5+6)/2 = 5.5$. Selection bias $= 10 - 5.5 = +4.5$. Verify: SDO = ATT + selection bias $= -4 + 4.5 = +0.5$. ✓
(e) Selection bias is positive: trained workers would have had *higher* baseline unemployment duration even without training. Workers who need training the most (hardest cases) select into it, making the program look ineffective in raw comparisons.

Practice 2: Regularization and Cross-Validation (12 points)

(a) (3 pts) A LASSO model at $\lambda_{\min}$ has CV MSE $= 0.198$. What is the RMSE?

(b) (3 pts) You run 5-fold CV on a LASSO model and observe these fold-level test RMSEs: 0.452, 0.448, 0.461, 0.455, 0.444. What is the 5-fold CV RMSE?

(c) (3 pts) A student says: "LASSO set 60 of my 100 coefficients to exactly zero, which means those variables have no effect on the outcome." What is wrong with this interpretation?

(d) (3 pts) You fit a Ridge model and a LASSO model on the same dataset with the same λ. The Ridge model retains all 100 predictors with small but nonzero coefficients. The LASSO model retains 22 predictors. Both have nearly identical out-of-sample RMSE. Which model would you prefer if the goal is prediction? Which if the goal is explanation? Explain briefly.

Solutions:
(a) RMSE $= \sqrt{0.198} \approx 0.445$.
(b) Mean of $\{0.452, 0.448, 0.461, 0.455, 0.444\}$: sum $= 2.260$, mean $= 2.260/5 = 0.452$.
(c) LASSO sets coefficients to zero because of the penalty, not because those variables have no effect. A zero coefficient means LASSO decided the predictor wasn't worth its regularization cost given the other predictors in the model. The true effect of an excluded variable may be nonzero — LASSO is making a prediction trade-off, not an oracle causal claim.
(d) If the *goal is prediction*: either model works equally well (same RMSE), so prefer whichever is more computationally tractable. If the *goal is explanation*: LASSO is preferable because it produces a sparse, interpretable model — 22 predictors are easier to discuss and communicate than 100, and the selection reduces the risk that an analyst is cherry-picking which predictors to highlight.

Practice 3: OVB and IV Logic (14 points)

A researcher regresses **log income** (Y) on **years of college** (D). The true model also includes **family wealth** (W), which is omitted.

(a) (3 pts) Write the OVB formula. Define every term.

(b) (3 pts) Family wealth increases both the years of college a person completes ($\delta_{DW} > 0$) and their income independently ($\beta_2 > 0$). What is the sign of OVB? Is OLS biased upward or downward?

(c) (4 pts) A researcher uses **distance to the nearest 4-year college** as an instrument for years of college. State the three IV conditions and explain briefly whether each is satisfied (or where it might be violated).

(d) (4 pts) The first stage gives $\hat{\alpha}_1 = 0.31$ (one additional year of college education for those who live close). The reduced form gives $\hat{\pi}_1 = 0.038$ (living near a college raises log income by 0.038). Calculate the Wald estimate. Interpret it in one sentence.

Solutions:
(a) $\hat{\beta}_1^{\text{OLS}} \to \beta_1 + \underbrace{\beta_2 \cdot \delta_{DW}}_{\text{OVB}}$. β_1: true effect of schooling on income. β_2: effect of wealth on income. δ_{DW}: regression of wealth on schooling (how much schooling predicts wealth).
(b) OVB $= \beta_2 \times \delta_{DW} = (+)(+) > 0$. OLS is biased *upward*: the estimated return to college exceeds the true return because we're partly capturing the effect of family wealth.
(c)

- **Relevance:** people near colleges get more schooling. Plausible; testable with first-stage F. Card (1995): $F = 16.5$ — passes.
- **Exclusion:** proximity affects income only through schooling. Potential violation: wealthy families may live in areas with more colleges (proximity correlated with family wealth directly affecting income).
- **Independence:** proximity is uncorrelated with omitted determinants of income. Same potential violation as above.

(d) Wald $= 0.038/0.31 \approx 0.123$. Interpretation: among men who attend college only because they grew up near one, an additional year of college raises log wages by approximately 0.123 (about 13%).

Practice 4: Weak Instruments and LATE (10 points)

(a) (4 pts) An IV study reports a first-stage $F = 4$. OLS has an omitted-variable bias of $+0.050$.

(i) Calculate the approximate 2SLS bias using the finite-sample formula.

(ii) If the true effect is 0.15, what does OLS estimate? What does 2SLS estimate?

(b) (3 pts) Explain in 2–3 sentences why the 2SLS estimator is *consistent* as $n \to \infty$ but *biased* in finite samples. What conditions make it consistent?

(c) (3 pts) Consider Angrist & Krueger (1991) with quarter of birth as the instrument for schooling. Who are the compliers in this study? Describe them in a sentence, then write a one-sentence LATE interpretation of the 2SLS estimate.

Solutions:
(a)(i) Bias(2SLS) $\approx 0.050 \times \frac{1}{4+1} = 0.050/5 = 0.010$.
(a)(ii) OLS estimate $\approx 0.15 + 0.050 = 0.200$. 2SLS estimate $\approx 0.15 + 0.010 = 0.160$.
(b) 2SLS is consistent because $\text{Cov}(Z_i, \varepsilon_i) = 0$ (exclusion restriction) makes the numerator of the probability limit of $(\hat{\pi}/\hat{\alpha})$ go to zero as $n \to \infty$. In finite samples, however, the estimated first stage $\hat{\alpha}_1$ is imprecise; random variation in $\hat{\alpha}_1$ introduces bias through the ratio. The bias shrinks at rate $1/F$, so it vanishes as the first stage becomes stronger.
(c) Compliers are men who stayed in school longer because they happened to be born in the fourth quarter (entering school early, accumulating more schooling by the mandatory dropout age), but who would have dropped out earlier if they had been born in the first quarter. LATE interpretation: the 2SLS estimate is the causal return to an additional year of schooling *for men whose schooling was determined by compulsory attendance laws interacting with their birth quarter* — those who would not have completed that year of schooling absent the accident of being born late in the year.