Gov 51: Exam 1 Review Guide

What to Know, What to Skip, and Practice Problems

Spring 2026

Exam: March 12, 2026 · 75 minutes · Cheat sheet allowed

Exam format: 100 points, 75 minutes, closed-resource except for one cheat sheet (two standard pages, front and back). No calculators, no computers. You will need 1.96 for confidence intervals. Show all work for partial credit.

1 What You Need to Know

1.1 Descriptive Statistics

- Mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Median: middle value (or average of two middle values) when data are sorted
- Variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
- Standard deviation: $s = \sqrt{s^2}$ (same units as data)
- Why $n-1$? Bessel's correction — using $\bar{x}$ instead of μ costs one degree of freedom
- Skewness: mean $>$ median $\Rightarrow$ right-skewed; mean $<$ median $\Rightarrow$ left-skewed

1.2 Covariance and Correlation

- Covariance: $s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
- Correlation: $r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$, always between -1 and $+1$
- Correlation measures *linear* association only
- $r = 0$ means no *linear* relationship (could still be nonlinear)
- Correlation $\neq$ causation — always state this when interpreting

1.3 Sampling and Confidence Intervals

- Standard error of a proportion: $\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Margin of error: MOE $= 1.96 \times$ SE
- 95% CI: $\hat{p} \pm 1.96 \times \text{SE}$
- **Correct interpretation**: "If we repeated this sampling procedure many times, 95% of the resulting intervals would contain the true parameter." NOT "95% probability the true value is in this interval."

- Square root rule: to halve the MOE, multiply n by 4
- Non-sampling errors: nonresponse bias, social desirability bias, question wording, selection bias

1.4 Regression: The Core Skill

This is the most important section. The exam emphasizes interpreting regression tables above all else. Practice reading tables, plugging in numbers, and writing interpretations.

- The regression equation: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \cdots$
- **Intercept** ($\hat{\beta}_0$): predicted Y when all X's equal zero
- **Slope** ($\hat{\beta}_1$): a one-unit increase in X_1 is associated with a $\hat{\beta}_1$ change in $\hat{Y}$, holding other variables constant
- **Binary X (0/1)**: intercept = mean of group 0; slope = difference in means between groups
- **t-statistic**: $t = \hat{\beta}/\text{SE}(\hat{\beta})$
- **Significance**: $|t| > 1.96 \Rightarrow$ significant at $\alpha = 0.05$ (for large n)
- R^2: fraction of variance in Y explained by the model (0 to 1)
- **Adding a covariate**: in a randomized experiment, adding a strong predictor of Y can reduce SE without changing $\hat{\beta}_1$ much — it "soaks up" residual variance

1.5 Prediction (Plug-in Calculations)

To predict $\hat{Y}$: Plug X values directly into the regression equation and do the arithmetic. This is the single most testable skill from weeks 5–6.

- Given $\hat{Y} = 82.00 - 7.50 \cdot \texttt{pid7} + 5.00 \cdot \texttt{female} + 1.80 \cdot (\texttt{pid7} \times \texttt{female})$
- For a male Strong Democrat: $\hat{Y} = 82.00 - 7.50(1) + 5.00(0) + 1.80(1)(0) = 74.50$
- For a female Strong Republican: $\hat{Y} = 82.00 - 7.50(7) + 5.00(1) + 1.80(7)(1) = 47.10$
- **Practice this.** Given any table, you should be able to write the equation and compute $\hat{Y}$ for any combination of X values.

1.6 Interactions

- **Interaction = the effect of one variable depends on another**
- Binary × continuous (e.g., `female` × `pid7`):
  - Slope for group 0 (men): $\hat{\beta}_1$
  - Slope for group 1 (women): $\hat{\beta}_1 + \hat{\beta}_3$
  - $\hat{\beta}_3$ = difference in slopes between the two groups
- Binary × binary: produces a 2×2 table of predicted means
  - $\hat{\beta}_3$ = difference-in-differences
- **Key**: always specify *for whom* a slope applies when there is an interaction

1.7 Overfitting and Prediction (Conceptual)

- R^2 **always increases** (or stays the same) when you add variables — it can never decrease
- A model that fits the training data perfectly may predict new data poorly — this is **overfitting**
- **RMSE** $= \sqrt{\frac{1}{n}\sum(\hat{Y}_i - Y_i)^2}$ = average prediction error in units of Y
- **Train/test split**: fit the model on training data, evaluate predictions on held-out test data
- Train RMSE always improves with more variables; test RMSE is what matters for prediction

1.8 Reading Research

- **Card et al. (PNAS 2022)**: Know the figure showing Congressional immigration speech tone over time. Key features: shift from negative to positive tone mid-20th century, partisan divergence, Trump's rhetoric compared to historical patterns.
- **LaCour & Green (retracted)**: Know the heaping evidence (responses at 50), the baseline distribution test (too-perfect match to national survey), the suspiciously low correlation across waves, and how adding noise to real data would eliminate heaping.
- **Broockman & Kalla (2016)**: Know their actual findings — conversations reduced prejudice, effects lasted at least 3 months. Know how their balance checks differed from LaCour's (small natural imbalance vs. implausibly perfect match).

2 What You Do NOT Need to Know

- R syntax or any coding
- Git, GitHub, or project workflow
- Weighted means or decompositions
- The KS test or any formal test for distributions
- How to *code* a train/test split (concept only — you should know *why* we do it)
- Adjusted R^2, AIC, BIC, Mallow's C_p, cross-validation, LASSO
- Continuous $\times$ continuous interactions
- Federalist Papers details (beyond the general idea of text-as-data)
- Proving any formulas — you just need to use them
- SE formula for a regression coefficient (you will be given SE's in any table)

3 Formula Reference (Cheat Sheet Suggestions)

Put these on your cheat sheet if you want them handy:

Mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Variance	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Std. Dev.	$s = \sqrt{s^2}$
Covariance	$s_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$
Correlation	$r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$
SE (proportion)	$\text{SE}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
95% CI	$\hat{p} \pm 1.96 \times \text{SE}$
t-statistic	$t = \frac{\hat{\beta}}{\text{SE}(\hat{\beta})}$
Prediction	$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots$
R^2	Fraction of variance in Y explained by model
RMSE	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_i - Y_i)^2}$
Square root rule	Halve MOE $\Rightarrow$ multiply n by 4

4 Practice Problems with Solutions

Work these with pencil and paper before checking the solutions. Time yourself — aim for about 1 minute per point.

Practice 1: Descriptive Statistics (8 points)

Five precincts report voter turnout (%):

$$62, \quad 48, \quad 55, \quad 71, \quad 44$$

(a) (2 pts) Calculate the mean.

(b) (2 pts) What is the median?

(c) (4 pts) Calculate the variance and standard deviation.

Solution:
(a) $\bar{x} = (62 + 48 + 55 + 71 + 44)/5 = 280/5 = 56.0$
(b) Sorted: 44, 48, 55, 62, 71. Median = 55 (middle value).
(c) Deviations from 56: $(6, -8, -1, 15, -12)$.
Squared deviations: $36 + 64 + 1 + 225 + 144 = 470$.
$s^2 = 470/(5-1) = 470/4 = 117.5$
$s = \sqrt{117.5} \approx 10.84$

Practice 2: Covariance and Correlation (8 points)

Four countries are observed:

Country	x (GDP growth %)	y (Approval %)
1	1	40
2	3	50
3	2	45
4	4	65

(a) (2 pts) Find $\bar{x}$ and $\bar{y}$.

(b) (3 pts) Calculate the covariance s_{xy}.

(c) (3 pts) Calculate s_x, s_y, and the correlation r_{xy}.

Solution:
(a) $\bar{x} = (1 + 3 + 2 + 4)/4 = 10/4 = 2.5$. $\bar{y} = (40 + 50 + 45 + 65)/4 = 200/4 = 50$.
(b) Deviations:

i	$(x_i - 2.5)$	$(y_i - 50)$	product
1	-1.5	-10	15
2	0.5	0	0
3	-0.5	-5	2.5
4	1.5	15	22.5

$s_{xy} = (15 + 0 + 2.5 + 22.5)/(4-1) = 40/3 \approx 13.33$
(c) $\sum(x_i - \bar{x})^2 = 2.25 + 0.25 + 0.25 + 2.25 = 5$, so $s_x^2 = 5/3 \approx 1.667$, $s_x \approx 1.291$.
$\sum(y_i - \bar{y})^2 = 100 + 0 + 25 + 225 = 350$, so $s_y^2 = 350/3 \approx 116.67$, $s_y \approx 10.80$.
$r_{xy} = 13.33/(1.291 \times 10.80) = 13.33/13.94 \approx 0.956$.
Strong positive linear relationship.

Practice 3: Confidence Interval (6 points)

A poll of $n = 400$ voters finds $\hat{p} = 0.47$ plan to vote yes on a referendum.

(a) (2 pts) Calculate the SE and construct a 95% CI.

(b) (2 pts) A reporter writes: "There is a 95% chance the true support is between 0.421 and 0.519." Fix this statement.

(c) (2 pts) How large a sample is needed to cut the margin of error in half?

Solution:
(a) $\text{SE} = \sqrt{0.47 \times 0.53/400} = \sqrt{0.2491/400} = \sqrt{0.000623} \approx 0.0250$.
$\text{MOE} = 1.96 \times 0.0250 = 0.0490$.
$\text{CI} : 0.47 \pm 0.049 = (0.421,\ 0.519)$.
(b) **Wrong:** "95% chance the true value is in this interval." The true value is fixed. **Correct:** "If we repeated this sampling procedure many times, 95% of the resulting confidence intervals would contain the true proportion."
(c) Square root rule: multiply n by 4. New $n = 400 \times 4 = 1{,}600$.

Practice 4: Regression with Binary X (8 points)

A study randomly assigns 500 students to a tutoring program (`tutor` $= 1$) or control (`tutor` $= 0$) and measures final exam scores.

	Coef.	SE	t	p
Intercept	72.00	1.50	48.00	< 0.001
`tutor`	6.30	2.10		
$R^2 = 0.018$		$n = 500$		

(a) (3 pts) Interpret the intercept and slope in context.

(b) (2 pts) Compute the t-statistic. Is the coefficient significant at $\alpha = 0.05$?

(c) (3 pts) The R^2 is only 0.018. Does this mean the tutoring program is useless? Explain.

Solution:
(a) **Intercept (72.00)**: The average exam score for students in the control group (no tutoring) is 72 points. **Slope (6.30)**: Students who received tutoring scored 6.30 points higher on average than control students.
Because `tutor` is binary (0/1), the intercept equals the control group mean and the slope equals the difference in means.
(b) $t = 6.30/2.10 = 3.00$. Since $|3.00| > 1.96$, the coefficient is statistically significant at $\alpha = 0.05$.
(c) No. A low R^2 means the model explains little of the total variation in exam scores — many other factors affect scores (prior ability, study habits, etc.). But R^2 does not tell us whether the treatment effect is real or meaningful. The t-statistic (3.00) shows the effect is statistically significant, and 6.3 points may be practically important.

Practice 5: Interaction and Prediction (10 points)

A regression predicts job approval (0–100) from education level (`college` = 1 if BA or higher) and age (in years), with an interaction:

	Coef.	SE	t	p
Intercept	40.00	3.00	13.33	< 0.001
`college`	15.00	4.20	3.57	< 0.001
`age`	0.30	0.05	6.00	< 0.001
`college` × `age`	−0.20	0.08	−2.50	0.013
$R^2 = 0.15$	$n = 1{,}200$			

(a) (3 pts) Write the prediction equation. What is the slope of age for non-college respondents? For college respondents?

(b) (4 pts) Predict the approval rating for: (i) a 30-year-old without college, (ii) a 30-year-old with college, (iii) a 60-year-old without college, (iv) a 60-year-old with college.

(c) (3 pts) Interpret the interaction coefficient (−0.20) in context.

Solution:
(a) $\hat{Y} = 40 + 15 \cdot \texttt{college} + 0.30 \cdot \texttt{age} - 0.20 \cdot (\texttt{college} \times \texttt{age})$.
Slope of age for non-college ($\texttt{college} = 0$): 0.30.
Slope of age for college ($\texttt{college} = 1$): $0.30 + (-0.20) = 0.10$.
(b) Plug in:

- (i) $\hat{Y} = 40 + 15(0) + 0.30(30) - 0.20(0)(30) = 40 + 9 = 49.0$
- (ii) $\hat{Y} = 40 + 15(1) + 0.30(30) - 0.20(1)(30) = 40 + 15 + 9 - 6 = 58.0$
- (iii) $\hat{Y} = 40 + 15(0) + 0.30(60) - 0.20(0)(60) = 40 + 18 = 58.0$
- (iv) $\hat{Y} = 40 + 15(1) + 0.30(60) - 0.20(1)(60) = 40 + 15 + 18 - 12 = 61.0$

(c) The interaction coefficient (-0.20) means the age slope is 0.20 points *smaller* for college-educated respondents than for non-college respondents. In other words, approval rises with age for both groups, but the increase is steeper for those without a college degree. Among non-college respondents, each additional year of age is associated with 0.30 more approval points; among college respondents, the increase is only 0.10 per year.

Practice 6: Binary × Binary Interaction (6 points)

A model predicts civic knowledge (0–100) from college attendance and gender:

$$\hat{Y} = 46 + 6 \cdot \texttt{college} - 1 \cdot \texttt{female} + 9 \cdot (\texttt{college} \times \texttt{female})$$

(a) (4 pts) Fill in the 2×2 table of predicted means:

	$\texttt{college} = 0$	$\texttt{college} = 1$
Male		
Female		

(b) (2 pts) What is the "college effect" for men? For women? What is the difference-in-differences?

Solution:
(a) Plug in all four combinations:

	$\texttt{college} = 0$	$\texttt{college} = 1$
Male ($\texttt{female} = 0$)	46	$46 + 6 = 52$
Female ($\texttt{female} = 1$)	$46 - 1 = 45$	$46 + 6 - 1 + 9 = 60$

(b) College effect for men: $52 - 46 = 6$. College effect for women: $60 - 45 = 15$.
Difference-in-differences: $15 - 6 = 9 = \hat{\beta}_3$. The college boost to civic knowledge is 9 points larger for women than for men.

Practice 7: True/False (10 points, 2 each)

For each statement, say **True** or **False** and explain in 1–2 sentences.

(a) A correlation of $r = -0.95$ indicates a weaker relationship than $r = +0.80$.

(b) In a randomized experiment, adding a control variable to the regression cannot change $\hat{\beta}_1$ because treatment is independent of everything.

(c) If a regression has $R^2 = 0.92$, we can conclude it will predict well on new data.

(d) When we say a coefficient is "statistically significant at $\alpha = 0.05$," we mean there is only a 5% chance the coefficient is zero.

(e) In a regression with an interaction $\hat{Y} = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (X \times D)$, the coefficient β_1 is the effect of X for *everyone* in the sample.

Solutions:
(a) **False.** The *strength* of a linear relationship is measured by $|r|$. Since $|-0.95| = 0.95 > 0.80$, the $r = -0.95$ relationship is actually *stronger*. The sign tells you the direction, not the strength.
(b) **False.** In theory, adding a covariate should not change $\hat{\beta}_1$ much in a randomized experiment (because treatment is independent of covariates). But in any finite sample, it can change slightly due to sampling variability. The main benefit is reducing the standard error.
(c) **False.** A high R^2 on the training data does not guarantee good prediction on new data. This could be a case of overfitting — the model may be fitting noise in the training data that does not generalize.
(d) **False.** Statistical significance means: if the true coefficient were zero, there would be less than a 5% probability of observing a t-statistic this extreme. It does *not* mean there is a 5% probability the coefficient is zero.
(e) **False.** When there is an interaction, β_1 is the effect of X only when $D = 0$. For the group with $D = 1$, the effect of X is $\beta_1 + \beta_3$. The interaction means the effect varies by group.

Practice 8: Spotting Suspicious Data (4 points)

A researcher claims to have conducted three follow-up surveys after an initial baseline. Below are the correlations between the baseline feeling thermometer and each follow-up wave:

	Correlation with baseline
Wave 2 (2 weeks later)	0.35
Wave 3 (6 weeks later)	0.33
Wave 4 (12 weeks later)	0.30

A different (verified) survey using the same instrument typically finds correlations of 0.75–0.85 between waves. In 2–3 sentences, explain why these low correlations are suspicious and what they might indicate about the data.

Solution:
People's attitudes are fairly stable over short periods. When the same people take the same survey weeks apart, their responses should be highly correlated ($r \approx 0.75$–0.85). Correlations of only 0.30–0.35 suggest the follow-up data are *not* from the same respondents, or the data were fabricated. If someone generated fake follow-up data by adding large random noise to baseline values (rather than collecting real responses), the correlation would be much lower than expected — exactly what we see here.

5 Study Tips

1. **Build your cheat sheet as you study.** The act of deciding what goes on it IS studying. Do not just photocopy your notes.
2. **Practice plug-in predictions from any regression table.** Given coefficients, you should be able to compute $\hat{Y}$ for any combination of X values in under 60 seconds.
3. **Interpret coefficients IN CONTEXT.** "The slope is 5.94" earns partial credit. "Students in the treatment group scored 5.94 points higher on the feeling thermometer, on average, than control students" earns full credit.
4. **Time yourself: ∼1 minute per point.** The exam is 100 points in 75 minutes, so you cannot spend 10 minutes on a 3-point question. Move on and come back.
5. **Show your work.** Partial credit is generous for correct setup with arithmetic mistakes. No work shown = no partial credit.
6. **Review the studies.** You will be asked about Card et al., LaCour & Green (the fraud), and Broockman & Kalla (the real study). Know the key findings and evidence — you do not need to memorize exact numbers.

Good luck on the exam!