Exam Review: Regression, Inference, Prediction

Gov 51 Section — Week 7

George
Harvard University

March 10, 2026

Can you predict who wins a congressional election?

435 House races. Millions of dollars. One question.

If a challenger raises \$500,000 more than expected, how many additional percentage points of vote share does she gain?

This question requires regression, inference, *and* prediction

Interpreting Regression Coefficients

Three models of Democratic vote share

	Model 1	Model 2	Model 3
Intercept	44.1	40.2	33.8
	(1.2)	(1.8)	(2.4)
Dem. spending ($100K)	1.8	1.4	1.2
	(0.3)	(0.3)	(0.3)
Incumbent (1 = yes)		8.1	7.3
		(1.1)	(1.1)
District partisanship			0.31
			(0.08)
R^2	0.24	0.51	0.58
Observations	435	435	435

Each coefficient has a specific interpretation

Model 2: $\widehat{\text{Vote}} = 40.2 + 1.4 \times \text{Spending} + 8.1 \times \text{Incumbent}$

Coefficient	Interpretation
$\hat{\beta}_0 = 40.2$	Predicted vote share for a non-incumbent who spends \$0
$\hat{\beta}_1 = 1.4$	Each additional \$100K in spending is associated with 1.4 points higher vote share, *holding incumbency constant*
$\hat{\beta}_2 = 8.1$	Incumbents score 8.1 points higher on average, *holding spending constant*

Plug in values to predict

Model 2: $\widehat{\text{Vote}} = 40.2 + 1.4 \times \text{Spending} + 8.1 \times \text{Incumbent}$

Scenario: a Democratic incumbent who spent \$500K

$$\widehat{\text{Vote}} = 40.2 + 1.4(5) + 8.1(1) = 40.2 + 7.0 + 8.1 = \mathbf{55.3}\%$$

Your turn: a non-incumbent challenger who spent \$800K?

$$\widehat{\text{Vote}} = 40.2 + 1.4(\ \ ) + 8.1(\ \ ) = ?$$

Adding variables changes coefficients — and that is normal

	Model 1	Model 2
Dem. spending ($100K)	1.8	1.4

The spending coefficient fell from 1.8 to 1.4 when we added incumbency

Model 1 attributed some of the incumbency advantage to spending — incumbents raise more money. Model 2 separates the two.

"Holding constant" means comparing districts with the *same* incumbency status

Hypothesis Testing and Inference

We want to learn about a population, not just our sample

Estimand

The *population* parameter

β_1 = true effect of spending on vote share across *all* elections

We never observe this directly

Estimator

The *sample* statistic

$\hat{\beta}_1 = 1.4$ from our sample of 435 races

Our best guess, but it varies from sample to sample

Inference = using $\hat{\beta}_1$ to say something about β_1

The standard error measures estimation uncertainty

$$\hat{\beta}_1 = 1.4 \quad \text{with} \quad \text{SE} = 0.3$$

If we could repeat the study on many different samples of 435 races, $\hat{\beta}_1$ would vary from sample to sample. The SE estimates the standard deviation of that variation.

✓ Correct

"The SE tells us how much $\hat{\beta}_1$ would typically vary across repeated samples"

× Incorrect

"The SE tells us how spread out the vote share data is"

The t-statistic asks: is this estimate surprisingly far from zero?

$$t = \frac{\hat{\beta}_1 - 0}{\text{SE}} = \frac{1.4}{0.3} = 4.67$$

Logic: under $H_0: \beta_1 = 0$ (spending has no relationship with vote share), how many SEs away from zero is our estimate?

✓ **Correct**

"$t = 4.67$ means the estimate is 4.67 standard errors from zero"

× **Incorrect**

"$t = 4.67$ means spending causes a 4.67-point increase in vote share"

The p-value is a probability about the data, not the hypothesis

With $t = 4.67$: $p < 0.001$

✓ Correct

"If spending truly had no relationship with vote share, there is less than a 0.1% chance of seeing an estimate this large or larger"

× Incorrect

"There is a 0.1% probability that spending has no effect"

"There is a 99.9% probability that spending increases vote share"

The p-value is $P(\text{data this extreme} \mid H_0)$, **not** $P(H_0 \mid \text{data})$

The 95% CI gives a range of plausible values for β_1

$$\hat{\beta}_1 \pm 1.96 \times \text{SE} \;=\; 1.4 \pm 1.96(0.3) \;=\; [0.81,\;\; 1.99]$$

✓ Correct

"If we repeated this study many times, about 95% of the resulting intervals would contain the true β_1"

× Incorrect

"There is a 95% probability that β_1 falls in $[0.81, 1.99]$"

(β_1 is fixed — the interval is what varies)

Since the CI excludes zero, we reject H_0 at $\alpha \;=\; 0.05$ — same conclusion as the p-value

Description vs. Prediction

Same model, different questions

Description

What patterns exist in the 435 races we already observed?

$\hat{Y}$ summarizes what happened

All covariate combinations are in the data

Prediction

What will vote share be in *next year's* races?

$\hat{Y}$ is a bet on the unseen

We plug in values we may have never observed together

Same $\hat{\beta}$'s, same formula — the difference is whether the observations are in-sample or out-of-sample

The intercept reveals when you are already predicting

Model 3: $\widehat{\text{Vote}} = 33.8 + 1.2 \cdot \text{Spend} + 7.3 \cdot \text{Inc} + 0.31 \cdot \text{Partisan}$

$\hat{\beta}_0 = 33.8 =$ predicted vote share when Spend $= 0$, Inc $= 0$, Partisan $= 0$

A non-incumbent in a perfectly neutral district who spends nothing

Does any such district exist in the data? Maybe not.
The intercept is already predicting beyond the observed data.

Overfitting and Prediction

R^2 always goes up when you add variables

Remember our three models of vote share?

	Model 1	**Model 2**	**Model 3**
Variables	1	2	3
R^2	0.24	0.51	0.58

OLS *guarantees* R^2 cannot decrease when you add a variable

So why not add everything?

Because more variables can make predictions *worse*

Imagine: you add 200 variables to predict vote share — district median income, rainfall, local sports team wins...

On training data

R^2 keeps climbing
Every added variable explains a tiny bit more variation
Looks great

On new elections

Predictions get wild
The model memorized noise — quirks of *these* 435 races
Looks terrible

Overfitting = the model fits the training data so well that it fails on new data

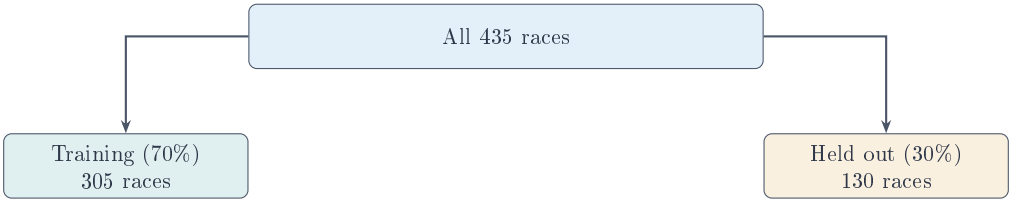
RMSE measures prediction error in real units

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

- $Y_i - \hat{Y}_i$ = prediction error for observation i
- Square, average, take the root $\rightarrow$ back in original units
- **RMSE = 4.2** means predictions are off by about 4.2 percentage points on average

R^2 rising = better fit. RMSE falling = better fit. Two sides of the same coin.

Train/test splits detect overfitting



The model never sees the held-out data — so RMSE there is an honest measure of prediction quality

The overfitting pattern: train improves, new data suffers

	3 vars	20 vars	200 vars
Train RMSE	8.4	5.9	2.9
New-data RMSE	8.7	7.2	18.0
Train R^2	0.58	0.79	0.95
New-data R^2	0.55	0.39	-0.92

200 variables: $R^2 = 0.95$ on training data, $R^2 = -0.92$ on new data.
Negative R^2 = predictions are worse than just guessing the average vote share.

Wait — SSR and SST are both sums of squares, so both are positive. How can R^2 be negative?

Putting It All Together

From regression table to exam answer

Model 2: $\hat{\beta}_1 = 1.4, \quad \text{SE} = 0.3, \quad t = 4.67, \quad p < 0.001$

1. **Interpret the coefficient:** each additional \$100K in Democratic spending is associated with 1.4 points higher vote share, holding incumbency constant
2. **State the hypothesis:** $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$
3. **Use the evidence:** $t = 4.67$ with $p < 0.001$; reject H_0 at $\alpha = 0.05$
4. **Give the CI:** $1.4 \pm 1.96(0.3) = [0.81,\ 1.99]$ — excludes zero

Common exam mistakes to avoid

Wrong	Right
"$p = 0.001$ means 0.1% chance H_0 is true"	"If H_0 were true, data this extreme would occur $< 0.1\%$ of the time"
"95% probability β_1 is in the CI"	"95% of intervals constructed this way contain β_1"
"SE measures spread of Y"	"SE measures spread of $\hat{\beta}$ across repeated samples"
"Higher R^2 = better model"	"Higher R^2 on *new data* = better predictions"

Same $\hat{\beta}$'s power both description and prediction — but only honest evaluation on new data tells you whether the model actually works