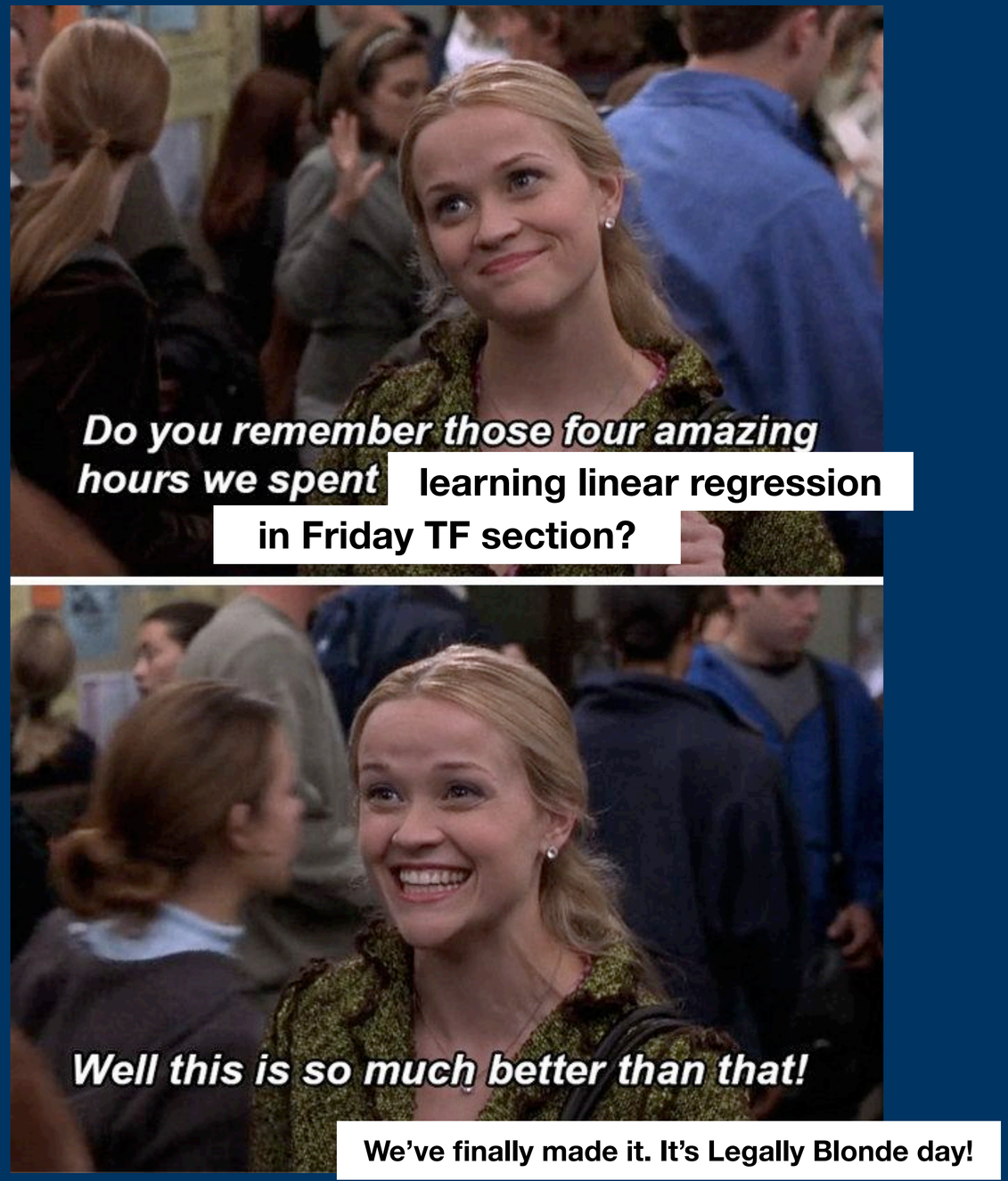
I’ll show you how valuable exam reviews can be!

API 202: TF Exam Review

ALL

Nolan M. Kavanagh
February 25, 2026



Practice set 1

Practice questions!

You work for the governor of Massachusetts. She plans to roll out free universal pre-K to 26 communities in the state. She wants to know if her plan will improve long-run educational attainment.

You have data on the educational attainment (measured in total years of formal education) of Boston students from 1996–2007 and run this regression, where pre-K attendance is coded either 0 or 1:

$$(years\_edu)_i = \hat{\beta}_0 + \hat{\beta}_1(pre\_K)_i + \hat{u}_i$$

Which is the "best" interpretation of $\hat{\beta}_1$ = 1.7 (SE = 1.5)?

A. Attending pre-K is associated with a 1.7-year increase in educational attainment. It is statistically significant.

B. A 1-unit difference in pre-K is associated with a 1.7-year increase in education. The difference is not statistically significant.

C. Attending pre-K causes a 1.7-year increase in educational attainment. The difference is not statistically significant.

D. Students who attended pre-K had 1.7 more years of education than those who didn't. It is not statistically significant.

Practice questions!

Original regression: $(years\_edu)_i = \hat{\alpha}_0 + \hat{\alpha}_1(pre\_K)_i + \hat{v}_i$

Original estimate: $\hat{\alpha}_1 = 1.7$ **(SE = 1.5)**

You're worried about omitted variable bias.

In particular, you know that children with higher household incomes are more likely to attend pre-K. You also know that children with higher household incomes are more likely to attain more education.

Using this information, sign the omitted variable bias for household income.

	γ_1	β_2	Bias sign	Bias size
A.	(+)	(+)	(+)	Understatement/sign flip
B.	(+)	(+)	(+)	Overstatement/sign flip
C.	(+)	(–)	(–)	Overstatement
D.	(+)	(–)	(+)	Understatement
E.	(–)	(–)	(+)	Understatement/sign flip
F.	(–)	(+)	(–)	Overstatement

Practice questions!

You decide to not only control for income but add an interaction term on the hunch that the association between pre-K and years of education might be different for low- and high-income families:

$$(years\_edu)_i = \widehat{\beta}_0 + \widehat{\beta}_1(pre\_K)_i + \widehat{\beta}_2(high\_inc)_i + \widehat{\beta}_3(pre\_K * high\_inc)_i + \widehat{u}_i$$

Both pre-K and high income are 0 or 1 dummy variables.

Write the combinations of betas that represent the predicted years of education for each of the following groups:

Low-income children who didn't go to pre-K:

Low-income children who went to pre-K:

High-income children who didn't go to pre-K:

High-income children who went to pre-K:

Practice questions!

$$(years\_edu)_i = \widehat{\beta}_0 + \widehat{\beta}_1(pre\_K)_i + \widehat{\beta}_2(high\_inc)_i + \widehat{\beta}_3(pre\_K * high\_inc)_i + \widehat{u}_i$$

Given this <u>exact</u> equation, for which of the following pairs of groups can we conduct hypothesis testing?

Select all that apply.

A. Low-income/no pre-K vs. low-income/pre-K

B. Low-income/no pre-K vs. high-income/no pre-K

C. Low-income/no pre-K vs. high-income/pre-K

D. Low-income/pre-K vs. high-income/no pre-K

E. Low-income/pre-K vs. high-income/pre-K

F. High-income/no pre-K vs. high-income/pre-K

Practice questions!

You're also worried about the external validity of your model.

Which of the following is true about external validity in this context:

A. Because there are many potential omitted variables, we aren't sure whether our model is externally valid.

B. Because we only have data on students in Boston, we aren't sure whether the same relationship is true in rural areas.

C. Because our data is 15 years old, we aren't sure whether the same relationship is true for Bostonian children today.

D. A. and B.

E. A. and C.

F. B. and C.

G. All the above.

Practice questions!

You propose that the governor randomly selects the 26 communities that will receive free, universal pre-K to evaluate its efficacy.

What of the following statements are true about a randomized trial?

Select all that apply.

A. A randomized trial guarantees balance on all potential omitted variables that could bias our study's estimates.

B. A randomized trial will produce balance in omitted variables *on average*, but with only 26 communities, we still could have bias.

C. A randomized trial allows us to make causal claims that we couldn't with our simple, cross-sectional estimate.

D. A randomized trial might be politically infeasible if communities that don't receive pre-K perceive it as unfair.

E. It would be unethical to conduct an RCT if we already knew that pre-K is beneficial (assuming we have the resources to implement it).

F. A randomized trial is always better than a high-quality, causal observational study, like a difference-in-differences.

Practice questions!

You search the literature and find one randomized trial for pre-K: the Tennessee Voluntary Pre-K Program in 1996, which randomized 3,000 low-income families to the opportunity to attend pre-K or not.

The study found that students who won the pre-K lottery performed better on achievement tests at the end of pre-K. However, by the end of kindergarten, the control group caught up to the pre-K group.

The study followed students into third grade and still found the groups to be comparable. The study did not follow them beyond grade 3.

Thinking about external validity, describe how this study does and does not inform the Massachusetts governor's plan to expand universal pre-K to all schools in 26 select communities.

Practice questions!

Let's say that several savvy families in the control group were able to get their children into other pre-K programs.

On average, these savvy families were better educated and higher-income than the rest of the families in the trial.

Assuming that pre-K actually has a positive effect on educational attainment, how will this affect our observed estimate?

A. Toward the null (i.e. toward no effect).

B. Away from the null (i.e. away from no effect).

C. No expected bias.

Practice questions!

Let's say that several families who won the lottery for pre-K ended up deciding not to send their children to it.

Which of the following are appropriate? Select all that apply.

A. **To avoid introducing bias in our estimate, we should keep these families in the "treatment" group of our analysis.**

B. **To avoid introducing bias in our estimate, we should drop these families from our analyses entirely.**

C. **To avoid introducing bias in our estimate, we can drop these families if the resulting groups still have similar demographics.**

D. **To avoid introducing bias in our estimate, we should move these families to the "control" group of our analysis.**

E. **To avoid introducing bias in our estimate, we should adjust these children's academic scores by what we think they would have been if the children had attended pre-K.**

Practice set 2

Practice questions!

You were just hired by Vot-ER, a non-profit organization that works to increase the political participation of less healthy people.

You have cross-sectional data on turnout (0 = didn't vote or 1 = voted) and self-reported health (1 = very bad to 5 = excellent).

You run this regression: $(turnout)_i = \hat{\beta}_0 + \hat{\beta}_1(health)_i + \hat{u}_i$

Interpret $\hat{\beta}_1$ = 0.04 (95% CI, –0.01 to 0.09).

A. **Every 1-unit improvement in health is associated with a 4% increase in the probability of voting. It's not statistically significant.**

B. **Every 1-unit improvement in health is associated with a 4 pp increase in the probability of voting. It's not statistically significant.**

C. **Every 1-unit improvement in health is associated with a 0.04% increase in the probability of voting. It's not statistically significant.**

D. **Every 1-unit improvement in health is associated with a 0.04 pp increase in the probability of voting. It's statistically significant.**

Practice questions!

Original regression: $(turnout)_i = \hat{\alpha}_0 + \hat{\alpha}_1(health)_i + \hat{v}_i$

Original estimate: $\hat{\alpha}_1$ = 0.04 (95% CI, –0.01 to 0.09)

You're worried about omitted variable bias. You know that older people tend to be less healthy and also tend to vote more.

Sign the bias with age as an omitted variable.

	γ_1	β_2	Bias sign	Bias size
A.	(+)	(+)	(+)	Understatement
B.	(+)	(–)	(–)	Overstatement/sign flip
C.	(–)	(+)	(–)	Understatement
D.	(–)	(+)	(–)	Overstatement/sign flip
E.	(–)	(–)	(+)	Understatement/sign flip
F.	(–)	(–)	(+)	Overstatement

Practice questions!

Original regression: $(turnout)_i = \hat{\alpha}_0 + \hat{\alpha}_1(health)_i + \hat{v}_i$

Original estimate: $\hat{\alpha}_1$ = 0.04 (95% CI, –0.01 to 0.09)

You're worried about other omitted variables. Propose a potential omitted variable and sign the likely bias.

Practice questions!

You collect data on people who *randomly* got sick, e.g. had a heart attack, between the 2018 and 2020 elections. You have access to their voting records and the voting records of their neighbors.

You estimate the following difference-in-differences model:

$$(turnout)_i = \widehat{\beta}_0 + \widehat{\beta}_1(got\_sick)_i + \widehat{\beta}_2(post\_2018)_i + \widehat{\beta}_3(got\_sick * post\_2018)_i + \widehat{u}_i$$

Which of the following omitted variables threaten(s) your internal validity?

Select all that apply.

A. People's genes

B. National changes in turnout during the 2020 election

C. People's jobs, which might change during the study period

D. The fact that 2020 was a presidential election and 2018 was a midterm

E. People's political upbringing

Practice questions!

You get the following output (see to the right).

$$(turnout)_i = \widehat{\beta}_0 + \widehat{\beta}_1(got\_sick)_i + \widehat{\beta}_2(post\_2018)_i + \widehat{\beta}_3(got\_sick * post\_2018)_i + \widehat{u}_i$$

What is the predicted turnout of…

The control group in 2018?

The control group in 2020?

The treatment group in 2018?

The treatment group in 2020?

What is the effect of getting sick on turnout? Is it significant?

	Model 1
Intercept	**0.40**
	(0.03)
got_sick	**0.02**
	(0.04)
post_2018	**0.21**
	(0.04)
got_sick * post_2018	**–0.15**
	(0.04)
Num.Obs.	1450
R2	0.042
R2 Adj.	0.041

Practice questions!

Which of the following statement(s) accurately interpret(s) the R-squared in this context?

A. Our regression model "explains" about 4% of the variation in turnout for our 1,450 respondents.

B. Because the R-squared is so low, our model wouldn't do a great job predicting turnout for the average respondent.

C. Because the R-squared is so low, we shouldn't think of these estimates as causal, only as correlational.

D. A. and B.

E. A. and C.

F. B. and C.

G. All of the above.

	Model 1
Intercept	**0.40**
	(0.03)
got_sick	**0.02**
	(0.04)
post_2018	**0.21**
	(0.04)
got_sick * post_2018	**–0.15**
	(0.04)
Num.Obs.	1450
R2	0.042
R2 Adj.	0.041

Practice questions!

You gather data from additional years and graph the turnout for each group across several elections.

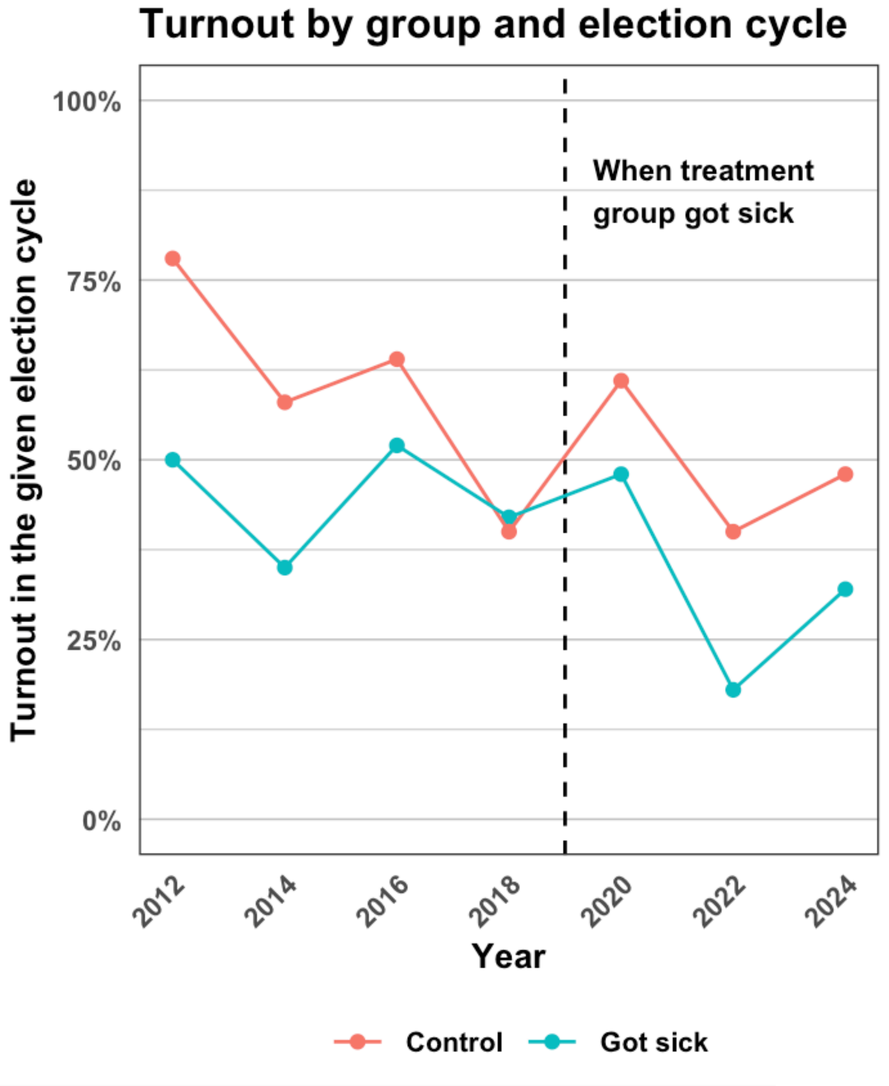
Which of the following statement(s) is (are) true about internal validity in a difference-in-differences? Select all that apply.

A. The parallel trends assumption is fundamentally unprovable because we cannot observe what the treatment group's trajectory would have been in the absence of treatment.

B. Although imperfect, one way to assess for parallel pre-trends is by inspecting the pre-treatment time points, and these groups have visually parallel pre-trends.

C. These groups have visually non-parallel pre-trends, meaning the control group is likely not an appropriate counterfactual, and we should not assert causality.

D. To have high internal validity, not only must the two groups have visually parallel pre-treatment lines, but they must also have similar pre-treatment levels, i.e. similar means.



Good luck!


Passing
gives you endorphins. Endorphins make you happy.
Happy people just don't
over-interpret bad studies as causal