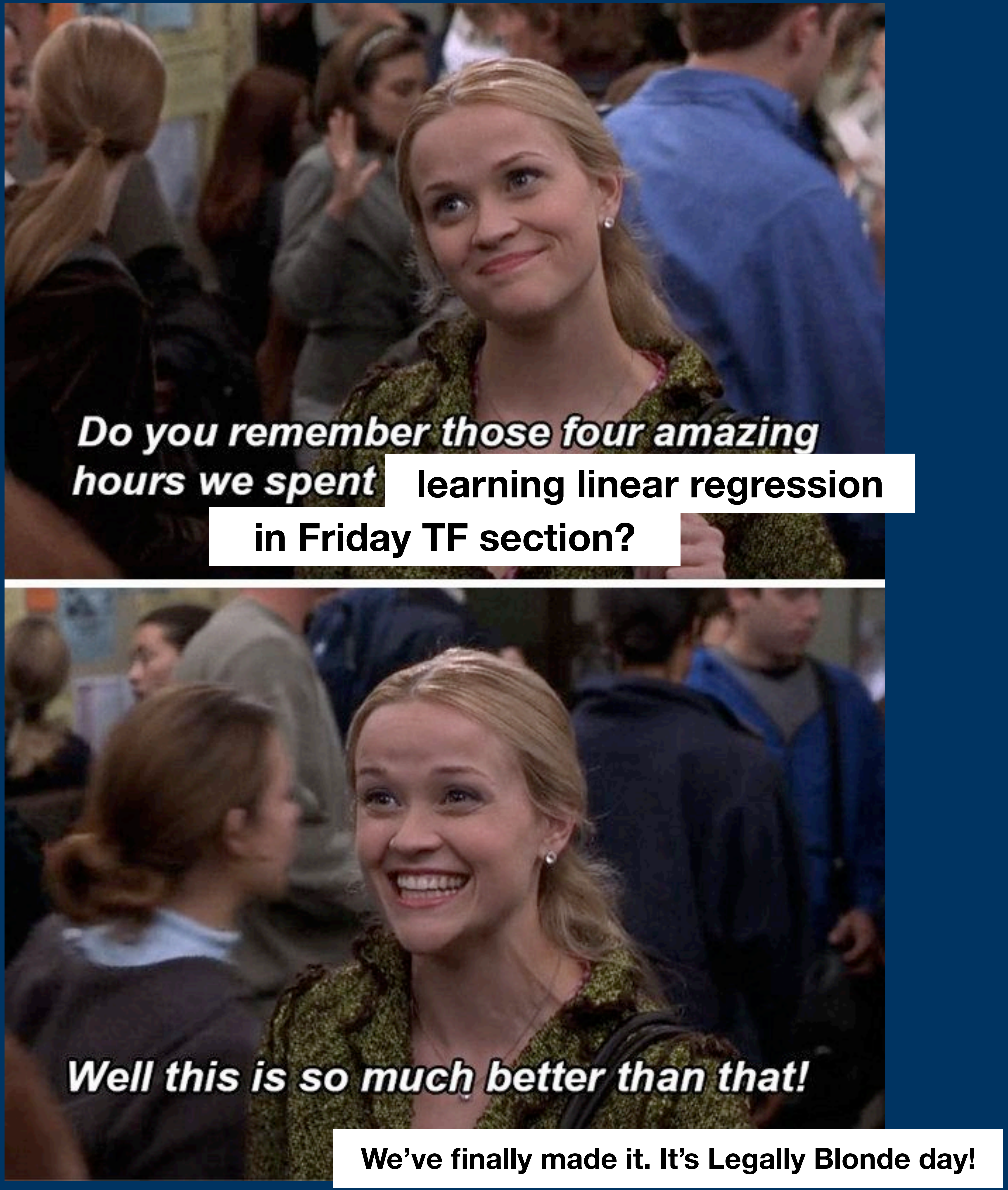
I’ll show you how valuable exam reviews can be!

API 202: TF Exam Review

ALL

Nolan M. Kavanagh
February 21, 2024



Practice set 1

Practice questions!

You work for the governor of Massachusetts. She plans to roll out <u>free universal pre-K</u> to 26 communities in the state by 2026. She wants to know if her plan will improve high school (H.S.) graduation rates.

You have data on Boston students who graduated H.S. from 1996–2007 and run this regression, where graduation and pre-K are coded as 0/1:

$$(graduation)_i = \widehat{\beta}_0 + \widehat{\beta}_1 (pre\_K)_i + \widehat{u}_i$$

Which is the "best" interpretation of β_1 = 0.17 (SE = 0.05, P<0.001)?

A. **Attending pre-K is associated with a 17% higher high school graduation rate. It is statistically significant.**

B. **A 1-unit change in pre-K is associated with a 0.17 pp higher high school graduation rate. It is statistically significant.**

C. **Attending pre-K is associated with a 0.17% higher high school graduation rate. It is statistically significant.**

D. **Attending pre-K is associated with a 17 pp higher high school graduation rate. It is statistically significant.**

Practice questions!

Original regression: $(graduation)_i = \widehat{\alpha}_0 + \widehat{\alpha}_1(pre\_K)_i + \widehat{v}_i$

Original estimate: α_1 = 0.17 (SE = 0.05, P<0.001)

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 graduate high school.

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

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

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 high school graduation might be different for low- and high-income families:

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

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

Write the combinations of betas that represent the predicted high school graduation rates 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!

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

Given this exact 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 estimate.

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 estimate 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. We should only conduct randomized trials when we are genuinely unsure whether pre-K will help or hurt participants.

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.

What is the expected direction of the bias?

A. **Toward the null.**

B. **Away from the null.**

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 is true about this scenario?

A. **To estimate an unbiased causal effect, we should include these families in the "treatment" group of our analysis.**

B. **To estimate an unbiased causal effect, we should include these families in the "control" group of our analysis.**

C. **To estimate an unbiased causal effect, we should drop these families from our analyses entirely.**

D. **To estimate an unbiased causal effect, 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 participation of "sicker" people in politics.

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: $(turnout)_i = \widehat{\beta}_0 + \widehat{\beta}_1 (health)_i + \widehat{u}_i$

Interpret β_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: $(turnout)_i = \widehat{\alpha}_0 + \widehat{\alpha}_1 (health)_i + \widehat{v}_i$

Original estimate: α_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 of age as an omitted variable.

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

Practice questions!

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

Original estimate: α_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. got in a car crash, 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 your estimate?

Select all that apply.

A. A person's genes

B. National changes in turnout during the 2020 election

C. A person's job, which might change during the study period

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

E. A person'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) most 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 any given 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

Good luck!


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