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!

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.

This study lacks the hallmarks of a causal design, as there is no evidence of random variation. As such, we want to avoid causal language (e.g. "causes," "leads to," "results in," etc.) (C.).

With dummy variables, saying a "1-unit change" is technically correct but clunky. Pre-K doesn't really have units. You either attended or didn't (B.).

By contrast, we'd rather say: "Attending pre-K is associated with a 1.7-year increase in education" or "students who attended pre-K had 1.7 more years of education than those who didn't."

Neither implies a causal effect, as we're just reporting the differences between the groups.

To distinguish A. and D., we must assess the statistical significance of the estimate. It is not statistically significant because the t-statistic = |1.7/1.5| = |1.1| ≤ 2. Also, the 95% confidence interval (1.7 ± 1.96*1.5 = –1.3 to 4.7) includes "0."

Thus, D. is the best response.

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!

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

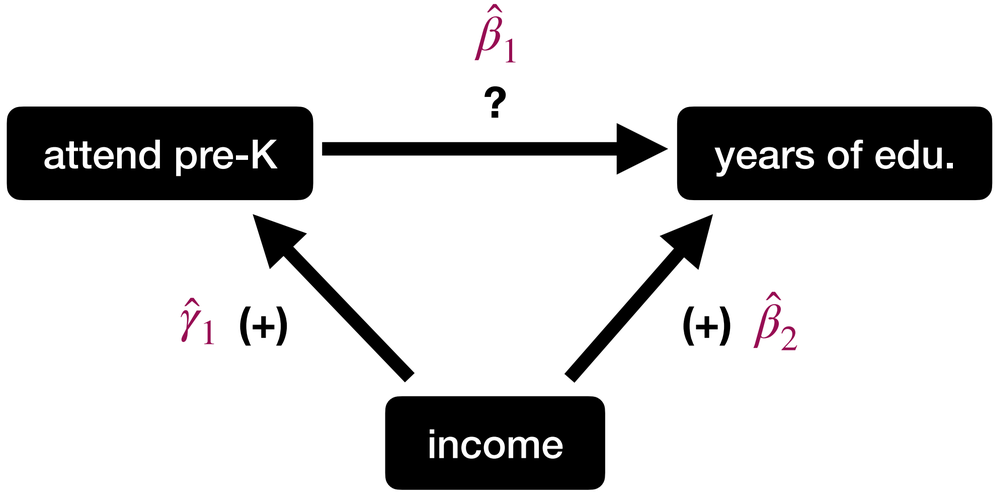
Original estimate: $\hat{\alpha}_1 = 1.7$ **(SE = 1.5)**

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.

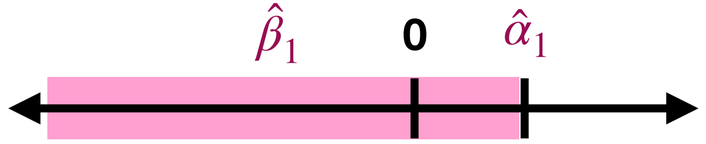
	γ_1	β_2	Bias sign	Bias size
A.	(+)	(+)	(+)	Understatement/sign flip
B.	**(+)**	**(+)**	**(+)**	**Overstatement/sign flip**
C.	(+)	(–)	(–)	Overstatement
D.	(+)	(–)	(+)	Understatement
E.	(–)	(–)	(+)	Understatement/sign flip
F.	(–)	(+)	(–)	Overstatement



Bias formula

$$\hat{\alpha}_1 - \hat{\beta}_1 = \hat{\beta}_2 * \hat{\gamma}_1 = (+)(+) = (+)$$

We know that $\hat{\alpha}_1 > 0$ and must be to the right of $\hat{\beta}_1$. However, we don't know exactly where $\hat{\beta}_1$ is.



So we have an overestatement or sign flip, depending on the degree of bias.

Bias: sign or size?

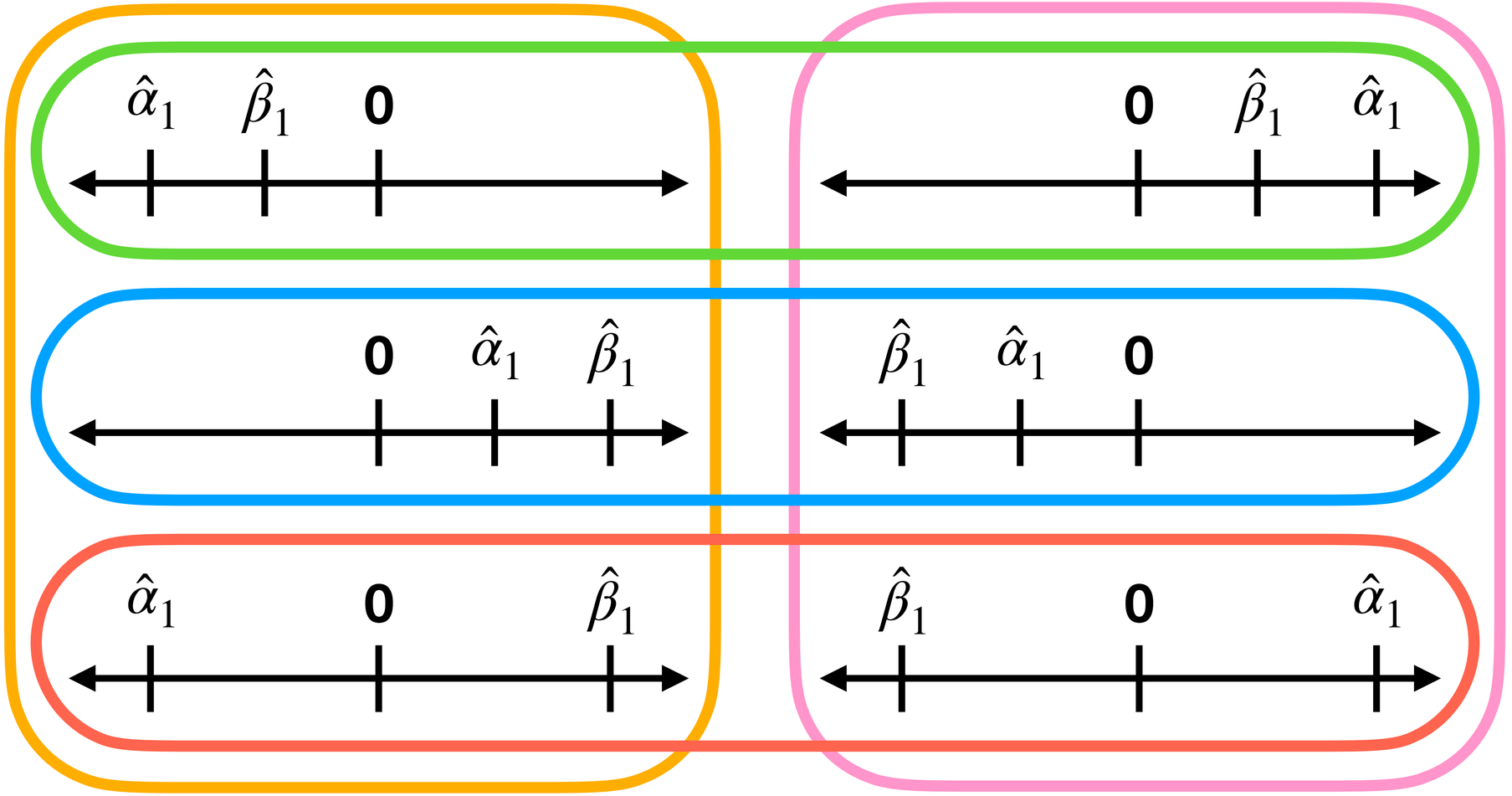
Overstatement

i.e. α_1 is farther from 0

Understatement

i.e. α_1 is closer to 0

Sign flip!



Negative bias (–)

i.e. α_1 is to the left of β_1

Positive bias (+)

i.e. α_1 is to the right of β_1

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!

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:

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

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: $\hat{\beta}_0$

Low-income children who went to pre-K: $\hat{\beta}_0 + \hat{\beta}_1$

High-income children who didn't go to pre-K: $\hat{\beta}_0 + \hat{\beta}_2$

High-income children who went to pre-K: $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

The secret sauce here is just setting each variable equal to 0 or 1 based on the groups we're interested in.

So if we want low-income children, we have to set "high_inc" = 0.

Meanwhile, if we want high-income children, we set "high_inc" = 1.

The same goes for "pre-K".

Then, we just drop the term(s) that have a "0" for any of the variables.

Yes, I kind of butchered this one.

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!

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

Given this <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 $\hat{\beta}_1$

B. Low-income/no pre-K vs. high-income/no pre-K $\hat{\beta}_2$

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

When we have a dummy-dummy interaction model, we can get the differences in predicted values between all combinations of groups.

However, we can only conduct hypothesis testing (i.e. get p-values) for comparisons that vary by one — and only one — beta, at least not without changing the regression.

In this case, that only applies to a few potential comparisons.

Low-income children who didn't go to pre-K: $\hat{\beta}_0$
Low-income children who went to pre-K: $\hat{\beta}_0 + \hat{\beta}_1$
High-income children who didn't go to pre-K: $\hat{\beta}_0 + \hat{\beta}_2$
High-income children who went to pre-K: $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

We could get the other comparisons by re-defining our dummy variables!

(OK look, making these memes isn't easy and they aren't all winners.)

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'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.

External validity is about the generalizability of our model, or how well it applies to contexts other than the one we studied.

We might think about whether another population is similar to our study in geography, demography, economics, and even time.

As a result, we might worry about different factors in non-urban areas that could change the relationship between pre-K and edu. attainment (B.), as well as how the relationship might have changed over time (C.).

Omitted variables are a problem for internal validity or causality, less so for external validity (A.).

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 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.

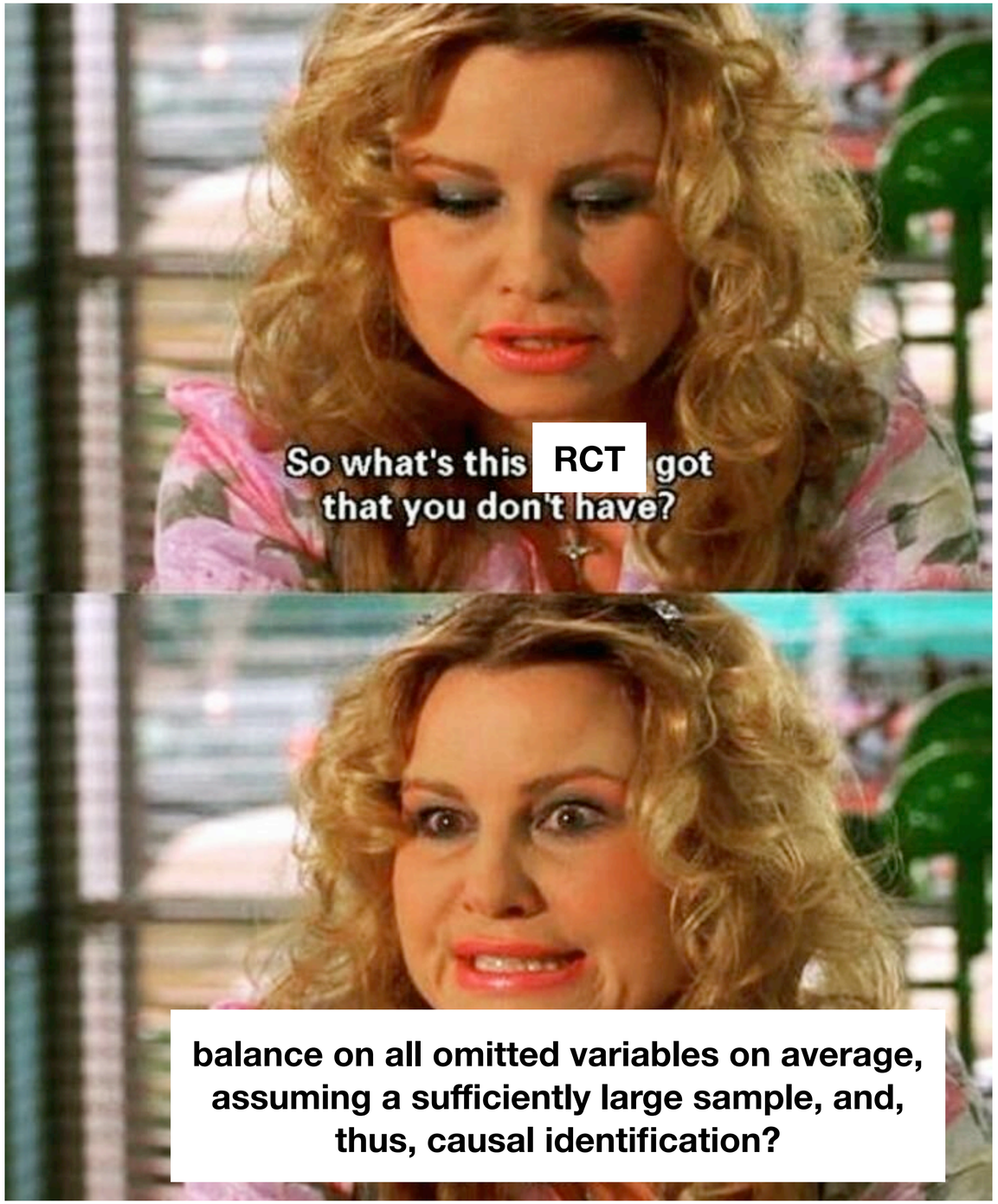
Randomization produces balance in omitted variables *on average*, assuming a reasonably large sample (B.). With 26 communities, we could end up with imbalance just due to chance (A.).

Randomized trials have several benefits over other study designs, like our cross-sectional regression. The most important is that we can make causal claims that we couldn't before (C.).

However, randomized trials can be imperfect. Failures in randomization or other design flaws can introduce bias and threaten their causal claims, potentially enough to make them "worse" than high-quality observational studies (F.).

Meanwhile, randomized trials can be expensive or politically or ethically infeasible (D.). One important ethical issue is "equipoise," meaning we shouldn't experiment on people when we already know what works or doesn't* (E.).

*There are some circumstances for which we might do a randomized trial even when we "know" an intervention is beneficial, e.g. we don't have enough resources to give it to everyone. But if we do have the resources, we should avoid putting people in control groups when there's an available treatment known to improve their well-being.


So what's this RCT got
that you don't have?
balance on all omitted variables on average, assuming a sufficiently large sample, and, thus, causal identification?

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!

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.

Lots of possible answers here.

Top of my mind are:

1. **The effects were null, so at baseline, I would expect relatively small effects of pre-K on student achievement in MA communities.**

2. **Educational attainment is likely related to test scores, but they're different outcomes.**

3. **The study didn't follow families beyond third grade. The fact that the results were null for several years makes it unlikely that we'd see an effect on long-run educational attainment (even further in the future), but we can't be sure.**

4. **The study randomized families, not entire communities. Perhaps directing resources to an entire community would increase the effect?**

5. **The study is fairly old. Perhaps pre-K instruction is more impactful now than in 1996?**

6. **Is Tennessee similar to Massachusetts?**

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 behavior affect our estimate of the policy's apparent effect?

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 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 behavior affect our estimate of the policy's apparent effect?

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

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

C. No expected bias.

This is an example of noncompliance in a randomized trial: Some families in one of the groups didn't "comply" with their assigned group. In this case, some of the control children ended up getting "treated" by pre-K.

We were told to assume that pre-K has a positive effect on edu. attainment. Since some of the control group was also treated, our estimate will probably be smaller than the "true" effect of pre-K on edu. attainment.

This is akin to "raising the floor" of the control group estimate, making the gap between the two groups smaller.

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 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.**

This is an "intention to treat" question.

Randomized trials only solve the issue of omitted variable bias when they preserve the original randomization.

Even though some of the "treatment" children didn't actually get treated, we still have to include them in their assigned group for our analyses (hence, A.).

The inclination not to attend is, itself, an omitted variable, so dropping, moving, or altering those children's values would introduce bias in our estimate.

We might be tempted to drop them if the demographics look similar (D.), but what if there isn't balance on unobservable characteristics? We can't really know.

In this scenario, the causal effect we're estimating is now that of winning a lottery for pre-K, not of attending pre-K! This estimate is unbiased. To get an unbiased estimate of attending pre-K itself, we'd have to do more advanced statistics (namely, an instrumental variable).


Whoever said you should analyze an RCT
“as treated” was seriously disturbed.
(aka based on whatever people choose for themselves)

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!

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.**

Here, we have a linear probability model with a binary outcome.

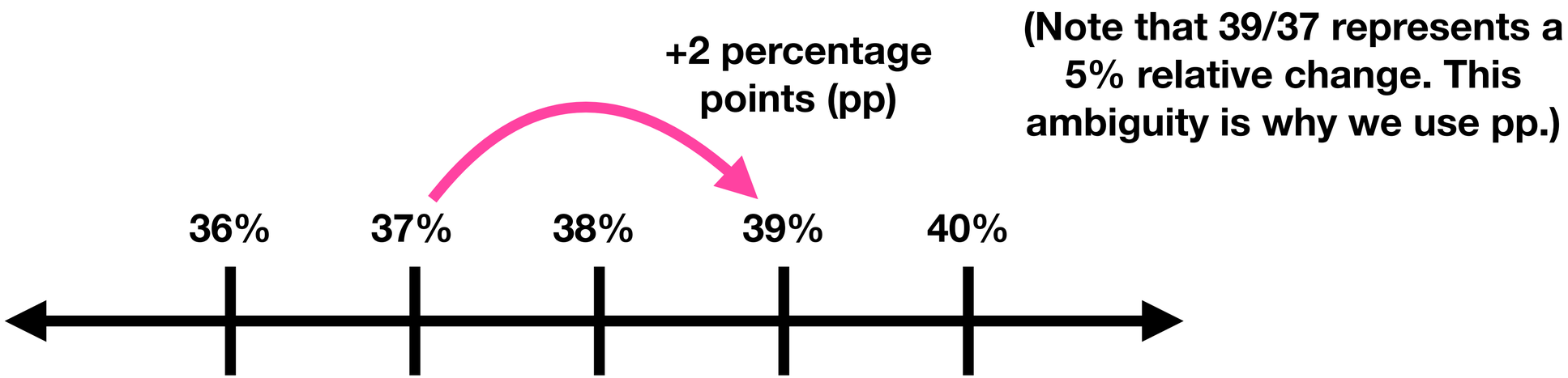
As a result, we have to interpret the outcome in percentage points.

When the outcome is 0 or 1, we have to multiply the coefficient by 100.

Lastly, because the 95% confidence interval includes "0", the result is not statistically significant (i.e. P>0.05).

A quick detour on percentage points

- When our outcome is measured in percents (%), we describe any movement along the number line using percentage points.



- If the outcome is measured from 0 to 100, you can interpret β_1 directly in pp. If measured 0 to 1, you must multiply by 100.

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!

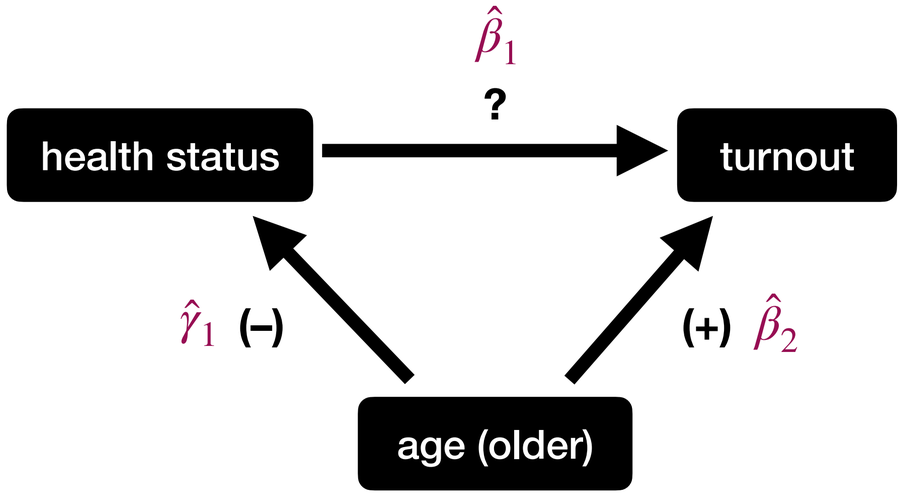
Original regression: $(turnout)_i = \hat{\alpha}_0 + \hat{\alpha}_1(health)_i + \hat{v}_i$

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

You're worried about 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.

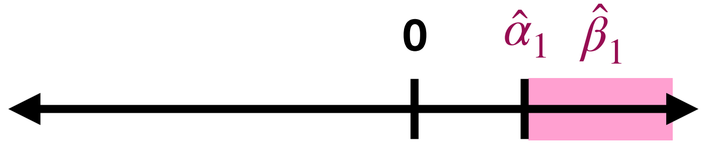
	γ_1	β_2	Bias sign	Bias size
A.	(+)	(+)	(+)	Understatement
B.	(+)	(–)	(–)	Overstatement/sign flip
C.	**(–)**	**(+)**	**(–)**	**Understatement**
D.	(–)	(+)	(–)	Overstatement/sign flip
E.	(–)	(–)	(+)	Understatement/sign flip
F.	(–)	(–)	(+)	Overstatement



Bias formula

$$\hat{\alpha}_1 - \hat{\beta}_1 = \hat{\beta}_2 * \hat{\gamma}_1 = (-)(+) = (-)$$

We know that $\hat{\alpha}_1 > 0$ and must be to the left of $\hat{\beta}_1$.



So we have an understatement.

Bias: sign or size?

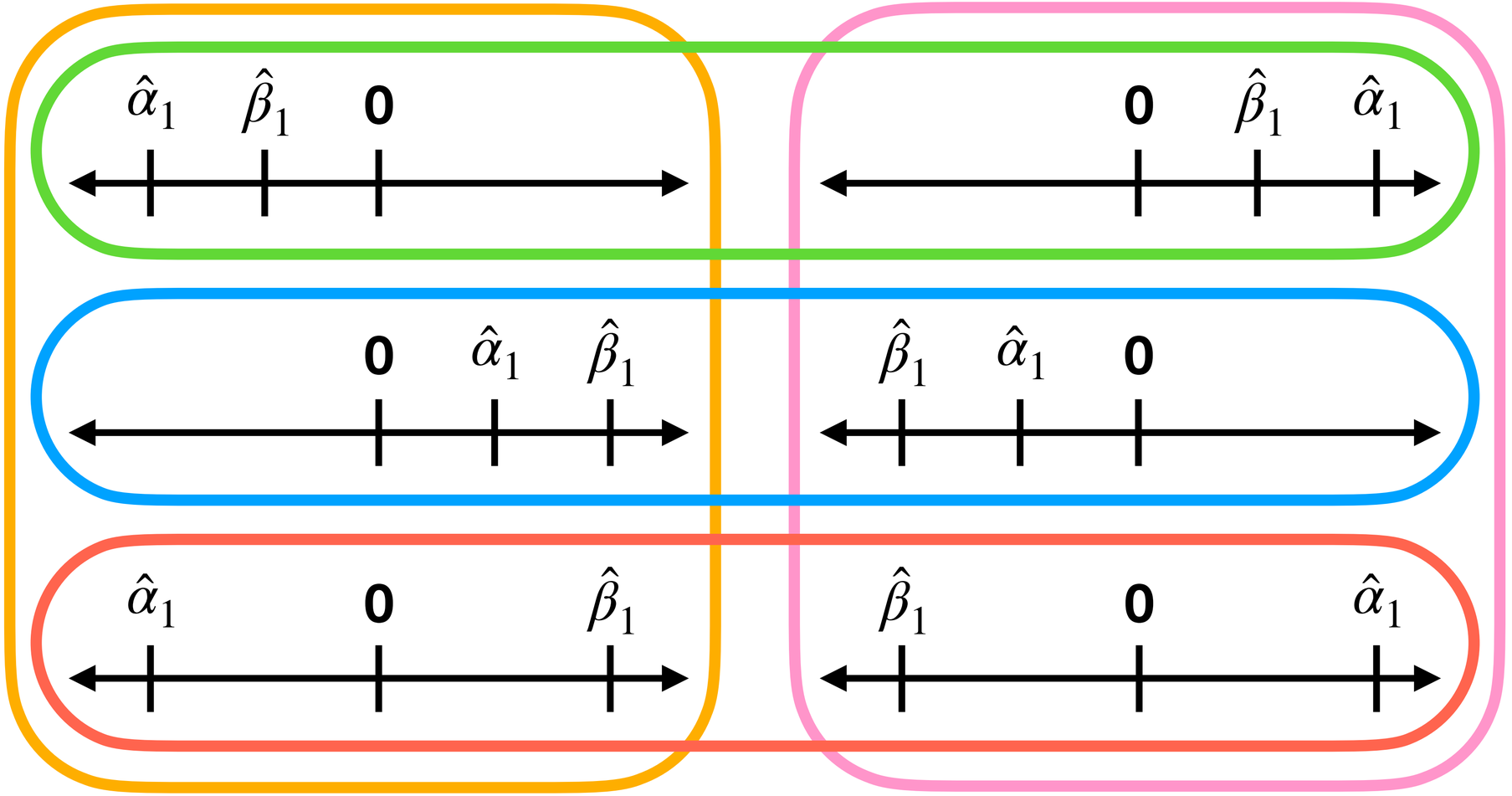
Overstatement

i.e. α_1 is farther from 0

Understatement

i.e. α_1 is closer to 0

Sign flip!



Negative bias (–)

i.e. α_1 is to the left of β_1

Positive bias (+)

i.e. α_1 is to the right of β_1


You signed… an omitted variable?
What, like it's hard?

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!

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.

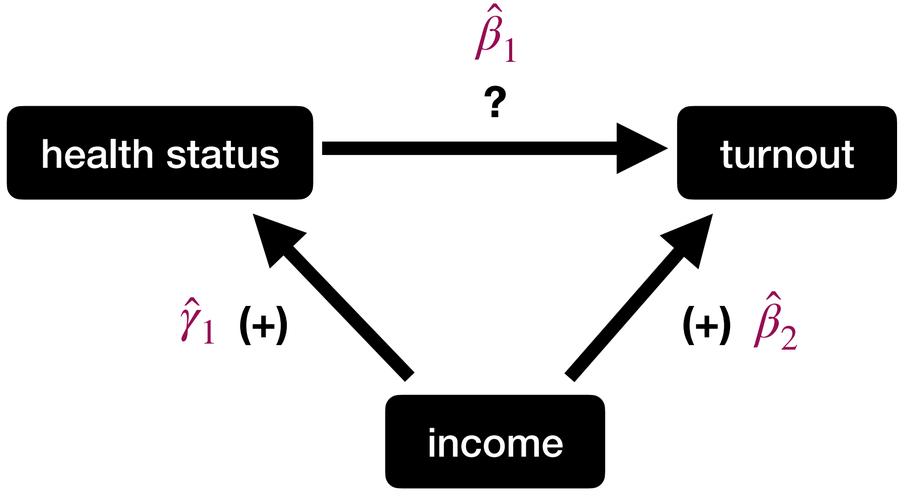
Let's say we're concerned about income.

Wealthier people tend to be healthier, so $\hat{\gamma}_1 > 0$.

Wealthier people tend to vote more often, so $\hat{\beta}_2 > 0$.

Thus, the bias is positive, meaning $\hat{\alpha}_1$ must be to the right of $\hat{\beta}_1$.

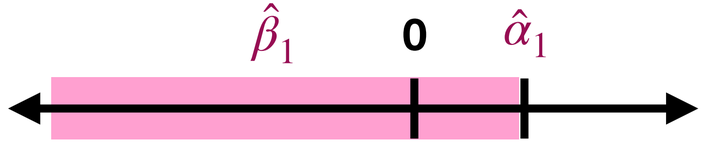
Since $\hat{\alpha}_1 > 0$, we can have an overestimate or sign flip.



Bias formula

$$\hat{\alpha}_1 - \hat{\beta}_1 = \hat{\beta}_2 * \hat{\gamma}_1 = (+)(+) = (+)$$

We know that $\hat{\alpha}_1 > 0$ and must be to the right of $\hat{\beta}_1$. However, we don't know exactly where $\hat{\beta}_1$ is.



So we have an overestimate or sign flip.

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 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:

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

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

In a diff-in-diff, we only worry about omitted variables that vary by time AND unit.

People's jobs (C.) is the only choice that fits this description, as treatment and control individuals can experience different job pressures throughout the study period.

Our model controls for omitted variables that differ between units but are fixed over time, like people's genes (A.) or upbringing (E.).

Our model also controls for time-varying omitted variables that affect all units equally, like the political environment (B. and D.).


how it feels to eliminate several classes of omitted variables with my kick ass study design

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!

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

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

What is the predicted turnout of…

The control group in 2018? $\hat{\beta}_0 = 40\%$

The control group in 2020? $\hat{\beta}_0 + \hat{\beta}_2 = 40 + 21 = 61\%$

The treatment group in 2018? $\hat{\beta}_0 + \hat{\beta}_1 = 40 + 2 = 42\%$

The treatment group in 2020? $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3 = 48\%$

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

$\hat{\beta}_3$ = –15 pp, and yes (95% CI, –15 ± 1.96*4 = –23 to –7)

	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

See the next slide for a review of the coefficients in diff-in-diffs!

Let's do a difference-in-differences!

$$(turnout)_i = \beta_0 + \beta_1(got\_sick)_i + \beta_2(post\_2018)_i + \beta_3(got\_sick * post\_2018)_i + u_i$$

dummy for getting sick
1 = Randomly got sick
0 = Didn't get sick

dummy for time
1 = 2020 election
0 = 2018 election

interaction
i.e. the diff-in-diff

	2018 election `post_2018 = 0`	**2020 election** `post_2018 = 1`	**Difference**
Didn't get sick `got_sick = 0`	β_0	$\beta_0 + \beta_2$	β_2
Randomly got sick `got_sick = 1`	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_2 + \beta_3$
Difference	β_1	$\beta_1 + \beta_3$	β_3

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!

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

By definition, R-squared quantifies the proportion of variation in our outcome variable "explained" or captured by our model.

It's a useful metric for evaluating predictions, but it doesn't confirm or deny whether our estimates are causal.

Causality is more about study design, and we have a quasi-random diff-in-diff, which is pretty good!


Elle, if I'm going to be
a senator...
I need to have a Jackie (causal study)
not a Marilyn. (just correlational)

Practice questions!

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

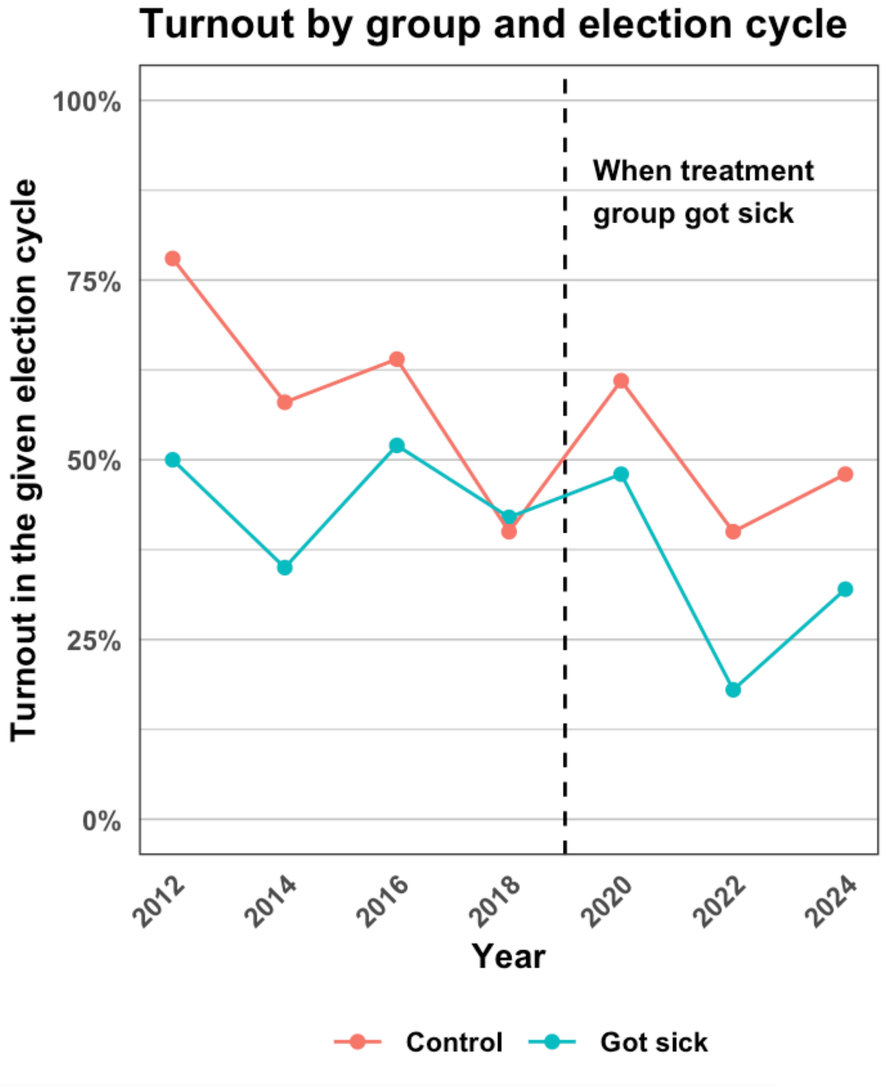
Which of the following statement(s) is (are) true about internal validity in a difference-in-differences? Select all that apply.

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.



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.

The fundamental identifying assumption of difference-in-differences is parallel trends. That is, we assume that the treatment group would experience a similar trajectory as the control group in the absence of treatment.

Because we can't actually observe what would've happened, parallel trends are fundamentally unprovable (A.).

Even so, we can reassure ourselves by assessing whether the two groups have similar trajectories in the pre-period. These trends are clearly not parallel (hence, C.).

Notably, the two groups can start at different levels. It's possible that the starting level affects the size of the treatment effect (for this reason, we are estimating the "treatment effect on treated units"). But our inferences are still intact. Just have to be parallel!

Good luck!


Passing API 202
gives you endorphins. Endorphins make you happy.
Happy people just don't
over-interpret bad studies as causal