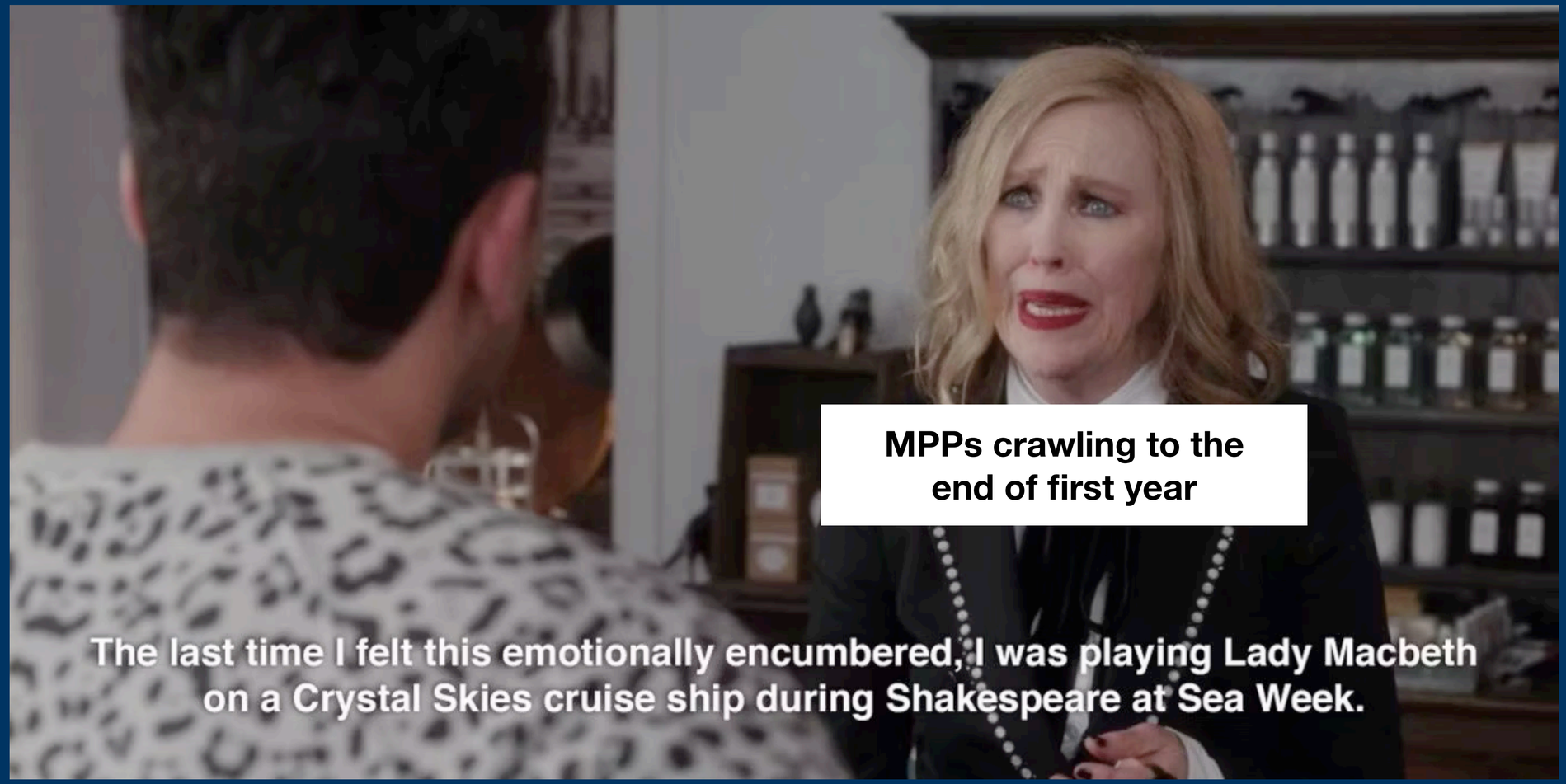

Ew, Exam Review!

API 203: TF Exam Review

R

Nolan M. Kavanagh
May 3, 2024

Practice questions!

You work for the governor of Texas, where the minimum wage is $7.25. He is considering raising it but needs more information.

He asks you to evaluate how minimum wage increases in other states have affected their labor markets. You have data on all states' minimum wages and labor markets from 2011 to 2022, and you run a regression with year and state fixed effects.

Which omitted variables does this approach eliminate?

A. State demographics that don't change over time.

B. State demographics that do change over time.

C. National trends in economic growth.

D. National economic policies, e.g. labor laws.

E. Pennsylvania economic policies that were passed in 1998.

F. Pennsylvania economic policies that were passed in 2014.

Practice questions!

You work for the governor of Texas, where the minimum wage is $7.25. He is considering raising it but needs more information.

He asks you to evaluate how minimum wage increases in other states have affected their labor markets. You have data on all states' minimum wages and labor markets from 2011 to 2022, and you run a regression with year and state fixed effects.

Which omitted variables does this approach eliminate?

A. State demographics that don't change over time.

B. State demographics that do change over time.

C. National trends in economic growth.

D. National economic policies, e.g. labor laws.

E. Pennsylvania economic policies that were passed in 1998.

F. Pennsylvania economic policies that were passed in 2014.

Fixed effects allow us to eliminate entire classes of omitted variables.

Year fixed effects allow us to eliminate omitted variables that affect all states at the same time. These include national policies (D.) and economic trends (C.).

Meanwhile, state fixed effects allow us to control for characteristics of each state that do not change over time, or at least during our study period. Here, that is (A.) and (E.) (which predates 2011).

Anything that changes at the state level during our study period (i.e. unit-time-varying characteristics) can be an omitted variable: (B.) and (F.).

Practice questions!

You take a closer look at the minimum wages of each state to better understand your data.

You notice the following minimum wages (see table). You can assume that if two years are the same, the minimum wage didn't change between them.

Which of the following state(s) is/are contributing variation to your fixed effects regression?

A. State 1

B. State 2

C. State 3

D. State 4

	2011	2016	2022
State 1	$7.25	$7.25	$7.25
State 2	$8.25	$10.35	$10.35
State 3	$12.75	$12.75	$12.75
State 4	$7.25	$9.75	$10.35

Practice questions!

You take a closer look at the minimum wages of each state to better understand your data.

You notice the following minimum wages (see table). You can assume that if two years are the same, the minimum wage didn't change between them.

Which of the following state(s) is/are contributing variation to your fixed effects regression?

A. State 1

B. State 2

C. State 3

D. State 4

	2011	2016	2022
State 1	$7.25	$7.25	$7.25
State 2	$8.25	$10.35	$10.35
State 3	$12.75	$12.75	$12.75
State 4	$7.25	$9.75	$10.35

State fixed effects will fully explain (or "identify") all states that do not experience a change in their minimum wage during the study period.

Put differently, fixed effects allow for "within-unit" estimation, so a state has to change its minimum wage at least once to contribute variation to our model.

Hence, only States 2 and 4 end up contributing. Note that we don't need a change in every year, so State 2 is OK!


Oh, look at
fixed effects
Smart enough to get that
unit-invariant OVB
but not smart enough to stop
that
time-unit-variant omitted variable bias

Practice questions!

You run a regression of log(labor market size) (with the market size measured in millions of people) on a state's minimum wage (in $USD) with state and year fixed effects from 2011 to 2022.

You get an estimate of $\beta_1 = 0.06$ (95% CI, 0.02 to 0.10).

Which of the following is the best interpretation of this estimate?

A. A $1 increase in a state's minimum wage was associated with a 6 percentage point increase in the state's labor market.

B. A $1 increase in a state's minimum wage was associated with a 60,000-person increase in the state's labor market.

C. A $1 increase in a state's minimum wage was associated with a 0.06 percent increase in the state's labor market.

D. A $1 increase in a state's minimum wage was associated with a 6% increase in the state's labor market.

E. A 1% increase in a state's minimum wage was associated with a 600-person increase in the state's labor market.

Practice questions!

You run a regression of log(labor market size) (with the market size measured in millions of people) on a state's minimum wage (in $USD) with state and year fixed effects from 2011 to 2022.

You get an estimate of $\beta_1 = 0.06$ (95% CI, 0.02 to 0.10).

Which of the following is the best interpretation of this estimate?

A. **A $1 increase in a state's minimum wage was associated with a 6 percentage point increase in the state's labor market.**

B. **A $1 increase in a state's minimum wage was associated with a 60,000-person increase in the state's labor market.**

C. **A $1 increase in a state's minimum wage was associated with a 0.06 percent increase in the state's labor market.**

D. **A $1 increase in a state's minimum wage was associated with a 6% increase in the state's labor market.**

E. **A 1% increase in a state's minimum wage was associated with a 600-person increase in the state's labor market.**

Here, we have a logged variable. The next slide details the various options for logged regressions.

But in this case, we have a log-level model, or a model with a logged outcome and level predictor.

We interpret it as the association between a 1-unit change in the predictor and a 100*β_1% change in the outcome, which is given by (D.).

(Note that I'm saying "association" because our fixed effects model accounts for many omitted variables but not all so may not be causal.)

Model	Regression Equation (PRF)	Interpretation of β_1 (Math)	Interpretation of β_1 (Words)
Level-Level	$Y = \beta_0 + \beta_1(X_1) + u$	$\Delta Y = (\beta_1)\Delta X_1$	A 1 unit increase in X_1 is associated with a β_1 change in Y
Level-Log	$Y = \beta_0 + \beta_1(ln\ X_1) + u$	$\Delta Y = (\frac{\beta_1}{100})\%\Delta X_1$	A 1% increase in X_1 is associated with a $0.01\beta_1$ change in Y
Log-Level	$ln(Y) = \beta_0 + \beta_1(X_1) + u$	$\%\Delta Y = (100\beta_1)\Delta X_1$	A 1 unit increase in X_1 is associated with a $100\beta_1\%$ change in Y
Log-Log	$ln(Y) = \beta_0 + \beta_1(ln\ X_1) + u$	$\%\Delta Y = (\beta_1)\%\Delta X_1$	A 1% increase in X_1 is associated with a $\beta_1\%$ change in Y

Honestly, I have to reference this table all the time, and that's OK!

Practice questions!

The governor asks you to broaden your search for any evidence of increased take-home pay on worker well-being. You notice that people at 100–400% of the federal poverty level (FPL) are eligible for health insurance subsidies, but people living <100% of FPL are not.

You believe that this context is perfect for regression discontinuity. Which of the following lines of code estimates the causal effect of eligibility for insurance subsidies on workers' financial stress?

A. `feols(stress ~ FPL + above_FPL_100 | 0, data)`

B.
```
feols(stress ~ FPL + above_FPL_100 +
FPL*above_FPL_100 | FPL, data)
```

C. `feols(stress ~ above_FPL_100 | 0, data)`

D.
```
feols(stress ~ FPL + above_FPL_100 +
FPL*above_FPL_100 | 0, data)
```

Practice questions!

The governor asks you to broaden your search for any evidence of increased take-home pay on worker well-being. You notice that people at 100–400% of the federal poverty level (FPL) are eligible for health insurance subsidies, but people living <100% of FPL are not.

You believe that this context is perfect for regression discontinuity. Which of the following lines of code estimates the causal effect of eligibility for insurance subsidies on workers' financial stress?

A. `feols(stress ~ FPL + above_FPL_100 | 0, data)`

B. `feols(stress ~ FPL + above_FPL_100 + FPL*above_FPL_100 | FPL, data)`

C. `feols(stress ~ above_FPL_100 | 0, data)`

D. `feols(stress ~ FPL + above_FPL_100 + FPL*above_FPL_100 | 0, data)`

To estimate a regression discontinuity, we want to include the running variable (in this case, a measure of each person's federal poverty level), an indicator for which side of the eligibility threshold they fall on, and the interaction between those two variables.

This combination ensures that we not only estimate the level change at the threshold but allow for the slope to vary as well, thereby fitting our line(s) more exactly.

Practice questions!

You retrieve the following RD OLS estimates. Note that financial stress is measured as a binary, with 0 indicating no financial stress and 1 indicating financial stress.

Which is the most accurate interpretation of the findings?

A. Being eligible for the insurance subsidies causes a 10.2 percentage point decrease in the probability of financial stress. The result is statistically significant at the 5% level.

B. Being eligible for the insurance subsidies causes a 10.2 percentage point increase in the probability of financial stress. The result is statistically significant at the 5% level.

C. Being eligible for the subsidies causes a 2.3 percentage point decrease in the probability of financial stress, but the difference is not statistically significant at the 5% level.

D. Every 1 percentage point increase in FPL causes a 0.3 percent decline in the probability of being financially stressed. The result is not statistically significant at the 5% level.

	Model 1
Intercept	**0.601**
	(0.023)
FPL	**–0.003**
	(0.009)
above_FPL_100	**–0.102**
	(0.035)
FPL * above_FPL_100	**–0.023**
	(0.029)
Num.Obs.	2,345
R2	0.034
R2 Adj.	0.032

Practice questions!

You retrieve the following RD OLS estimates. Note that financial stress is measured as a binary, with 0 indicating no financial stress and 1 indicating financial stress.

Which is the most accurate interpretation of the findings?

A. Being eligible for the insurance subsidies causes a 10.2 percentage point decrease in the probability of financial stress. The result is statistically significant at the 5% level.

B. Being eligible for the insurance subsidies causes a 10.2 percentage point increase in the probability of financial stress. The result is statistically significant at the 5% level.

C. Being eligible for the subsidies causes a 2.3 percentage point decrease in the probability of financial stress, but the difference is not statistically significant at the 5% level.

D. Every 1 percentage point increase in FPL causes a 0.3 percent decline in the probability of being financially stressed. The result is not statistically significant at the 5% level.

	Model 1
Intercept	**0.601**
	(0.023)
FPL	**–0.003**
	(0.009)
above_FPL_100	**–0.102**
	(0.035)
FPL * above_FPL_100	**–0.023**
	(0.029)
Num.Obs.	2,345
R2	0.034
R2 Adj.	0.032

When we evaluate RD estimates, we're usually most interested in the level change at the threshold, represented by the indicator for the threshold.

Here, that's "above_FPL_100".

Because the outcome is binary, we interpret it in percentage points. It's statistically significant because our CI (±1.96*SE) doesn't cross 0.

Practice questions!

Describe two criteria for the subsidy eligibility threshold that must be met to estimate a causal effect on financial stress. Briefly evaluate whether you believe these criteria were met here.

Practice questions!

Describe two criteria for the subsidy eligibility threshold that must be met to estimate a causal effect on financial stress. Briefly evaluate whether you believe these criteria were met here.

1. Whether someone falls above or below the threshold must be random, i.e. there must be similar demographic + other characteristics of people above and below the 100% FPL line.

You could make a good argument either way. We might be worried about other social programs that turn on or off at 100% FPL, but in the absence of more data, I have no explicit worry.

2. Workers cannot manipulate whether they fall above or below the threshold. Put differently, there cannot be "bunching" of workers with wages just above or below 100% FPL.

This strikes me as unlikely since manipulating your wage to fall above 100% FPL would be hard. We could use data to verify.


model assumptions
every applied researcher
Trust me, people aren't thinking about you the way that you're thinking about you.

Practice questions!

To whom does your causal estimate apply?

Recall that you work for the governor of Texas, who is interested in raising the minimum wage.

Describe how your results may or may not generalize to your governor's constituents.

Practice questions!

To whom does your causal estimate apply?

Recall that you work for the governor of Texas, who is interested in raising the minimum wage.

Describe how your results may or may not generalize to your governor's constituents.

Recall that a regression discontinuity only estimates a causal effect for people near the eligibility threshold (this is the "local average treatment effect"). Here, that's 100% FPL.

Workers near 100% FPL are also probably those who are most likely to be earning a minimum wage and benefit from a raise. This enhances the estimate's generalizability.

That said, it would be harder to generalize these estimates to workers at higher wages, e.g. if there is spillover wage growth for higher-income workers. The governor may also care about the impacts on these folks.

It's also worth considering the magnitude of the policy intervention. We aren't given information about the size of the insurance subsidy, but we would also want to see if it's comparable to the change in take-home pay that a worker would see if the minimum wage rises. If there's a large mismatch, we're less sure about the estimate's applicability.

Practice questions!

The governor (surprisingly) decides to pilot a universal basic income for households earning less than the federal poverty level. The available funds are limited, so you randomize it among those who apply. However, only about 50% of lottery winners end up getting the money.

You're still interested in the effect of the universal basic income on financial stress.

What's a statistical approach you could use to causally answer this question? Explain the strengths and weaknesses of your approach.

Practice questions!

The governor (surprisingly) decides to pilot a universal basic income for households earning less than the federal poverty level. The available funds are limited, so you randomize it among those who apply. However, only about 50% of lottery winners end up getting the money.

You're still interested in the effect of the universal basic income on financial stress.

What's a statistical approach you could use to causally answer this question? Explain the strengths and weaknesses of your approach.

Our first instinct here might be to treat it like a simple RCT and compare lottery winners and losers. However, this will give us the intention-to-treat estimate, or the causal effect of winning the lottery, not necessarily the causal effect of getting the money.

Instead, we can augment our approach with an instrumental variable, using winning the lottery as an instrument for getting a UBI.

In this setup, Z (our instrument) is winning the lottery, D (our endogenous variable) is getting the money, and Y is financial stress.

This approach has high internal validity but ONLY for the compliers, i.e. people who got the money because they won the lottery. This is our local average treatment effect.

External validity would depend on who, exactly, the compliers are (and relatedly, who entered the lottery in the first place).

Recall the LATE for IVs.

The “local average treatment effect” (LATE) is only for compliers.

We must ask: Is this group generalizable?

		Offered (Z = 1)	
		Don’t get treatment (D = 0)	**Get treatment (D = 1)**
Not offered (Z = 0)	**Don’t get treatment (D = 0)**	Never takers	Compliers
	Get treatment (D = 1)	Defiers	Always takers

We assume this group doesn’t exist.


Compliers
In case you didn't know, I'm wildly popular.
Some might even venture to call me beloved.

Practice questions!

You decide to pursue an instrumental variables approach to answer this causal question.

Write out the regressions for the following three models, where the Y variable is “stress”, D is “got_money”, and Z is “won_lottery”:

1. The reduced form

2. The first stage

3. The second stage

Indicate which coefficient(s) give the causal estimate that you’re interested in.

Practice questions!

You decide to pursue an instrumental variables approach to answer this causal question.

Write out the regressions for the following three models, where the Y variable is "stress", D is "got_money", and Z is "won_lottery":

1. The reduced form stress = $\gamma_0 + \gamma_1$*won_lottery + ε

2. The first stage got_money = $\alpha_0 + \alpha_1$*won_lottery + ε

3. The second stage stress = $\beta_0 + \beta_1$*$\widehat{\text{got\_money}}$ + ε

Indicate which coefficient(s) give the causal estimate that you're interested in.

Note: The second stage uses the predicted values of got_money, so yes, the hat is essential!

Note that our three instrumental variables equations are given by:

Reduced form

$$Y \sim Z$$

First stage

$$D \sim Z$$

Second stage

$$Y \sim \widehat{D}$$

We can get our causal estimate of interest in two ways:

1. β_1 from the second stage, or

2. γ_1/α_1, or the reduced form scaled by the first stage, also known as the Wald estimator.

Practice questions!

Which of the following are potential threats to the internal validity of your IV estimate?

Select all that apply.

A. The group of people who enrolled in the lottery were more white and lower-income than the people who would get the policy once it's fully rolled out.

B. There is evidence that the administrators of the program gave preference to people living in major cities when selecting the lottery winners.

C. Some people won the lottery and didn't end up getting the money, but they did enroll in other government programs they were eligible for.

D. None of the above are threats to inference.

Practice questions!

Which of the following are potential threats to the internal validity of your IV estimate?

Select all that apply.

A. The group of people who enrolled in the lottery were more white and lower-income than the people who would get the policy once it's fully rolled out.

B. There is evidence that the administrators of the program gave preference to people living in major cities when selecting the lottery winners.

C. Some people won the lottery and didn't end up getting the money, but they did enroll in other government programs they were eligible for.

D. None of the above are threats to inference.

Lots of things can go wrong with instrumental variables. Recall that we need:

1. A relevant instrument, i.e. the instrument has to be related to the endogenous variable.

2. The instrument must be random or "exogenous" with respect to the outcome.

3. There can be no other pathway except via the endogenous variable for the instrument to affect the outcome (this is the "exclusion restriction," and it's related to #2).

Option (B.) suggests that the instrument wasn't fully exogenous, i.e. that race and income may be omitted variables.

Option (C.) is a violation of the exclusion restriction since there are other pathways whereby winning the lottery helped people's financial stress not due to getting the UBI.

Meanwhile, (A.) is a problem of external validity, not internal validity.

Practice questions!

The governor asks you to identify Texans with the most financial stress so that the administration can design other programs to help them.

You don't have the resources to collect a survey of all Texans, so you decide to make a predictive model using your existing data.

You take two approaches: an OLS model with all predictor variables that you have available, and a LASSO model using the same predictors.

Which of the following statements is most likely to be true?

A. The OLS model will probably have the greatest out-of-sample prediction because it uses all available variables.

B. The LASSO model uses a tuning parameter to select highly correlated variables since including predictors that are highly correlated with each other tends to improve a model's predictive power.

C. We can't be sure which model is necessarily more predictive without evaluating recall and precision, preferably in the test set.

D. The OLS model is likely to be overfitted, but we can evaluate overfitting using the model's recall and precision in our training set.

Practice questions!

The governor asks you to identify Texans with the most financial stress so that the administration can design other programs to help them.

You don't have the resources to collect a survey of all Texans, so you decide to make a predictive model using your existing data.

You take two approaches: an OLS model with all predictor variables that you have available, and a LASSO model using the same predictors.

Which of the following statements is most likely to be true?

A. The OLS model will probably have the greatest out-of-sample prediction because it uses all available variables.

B. The LASSO model uses a tuning parameter to select highly correlated variables since including predictors that are highly correlated with each other tends to improve a model's predictive power.

C. We can't be sure which model is necessarily more predictive without evaluating recall and precision, preferably in the test set.

D. The OLS model is likely to be overfitted, but we can evaluate overfitting using the model's recall and precision in our training set.

When we're evaluating predictive models, the proof is generally in the pudding, specifically the model's out-of-sample performance.

This means that the most useful information is going to be the test sample's precision and recall (or whichever metrics are relevant) (C.). In-sample predictions are less useful (D.).

The OLS model is probably going to be overfit, which would lower its out-of-sample predictive power (A.). That said, we can't be sure until we actually look at the data. After all, the best predictive model is the one that's, well, the most predictive out-of-sample!

Lastly, we tend to improve our predictive power when we include uncorrelated, or very different, variables. LASSO is likely to drop highly correlated ones because they don't explain much more variation (B.).


LASSO
OLS

Practice questions!

The conservative governor is worried about the political consequences of giving “handouts” to people who don’t actually need help.

As a result, he asks you to select a model based on how well it correctly identifies people in need — without including financially healthy people.

Which of the following measures is most directly useful for this purpose?

A. Specificity

B. Recall (a.k.a. sensitivity)

C. Precision (a.k.a. positive predictive value)

D. Negative predictive value

Practice questions!

The conservative governor is worried about the political consequences of giving "handouts" to people who don't actually need help.

As a result, he asks you to select a model based on how well it correctly identifies people in need — without including financially healthy people.

Which of the following measures is most directly useful for this purpose?

A. Specificity

B. Recall (a.k.a. sensitivity)

C. Precision (a.k.a. positive predictive value)

D. Negative predictive value

Each of these measures is useful for evaluating predictive models. But in this context, we want a model for which someone who is predicted to be in financial stress is very likely to truly be in financial stress.

This points us to precision (a.k.a. positive predictive value), which is the proportion of true positives over all predicted positives (C.). If this measure equals 1, then everyone who gets the funds is actually in financial stress.

Recall (or sensitivity) is useful for capturing all people in financial stress. However, doing so will likely result in more false positives, which the governor doesn't want (B.).

Meanwhile, specificity (A.) and negative predictive value (D.) focus our attention on the people who aren't in financial stress, which is less useful for our purposes.

A brief review of binary prediction metrics.

Calculation	Terminology
FP / (TN + FP)	**False positive rate**, 1–Specificity, Type I error
TP / (TP + FN)	**True positive rate, Recall, Sensitivity**, 1–Type II error, Power
TP / All predicted positives	**Precision**, Positive predicted value
TN / All predicted negatives	Negative predicted value

Good luck!