

Omitted Variable Bias (23 Questions)

Question 1

True, false or uncertain: One can still use a linear regression framework even if the relationship between a regressor and the dependent variable is not linear.

True ✓

False

Uncertain

Explanation

Polynomials or other transformations of your variables can be used to represent fit functional forms using the standard OLS linear regression framework.

Question 2

Which of the following statements is **not** true?

In kernel regression, a more flexible functional form (smaller bandwidth) leads to lower variance. Whereas, in series regression, a more flexible functional form (higher order polynomials) increases variance.

Kernel and series regression models are both ways modeling nonlinear functional forms.

In both the kernel and series regression framework, the optimal trade-off between bias and variance is found by minimizing the mean squared error, and/or other cross validation criterion. to balance the trade-off between bias and variance.

Like in kernel regression models, series regression models have a trade-off between bias and variance.



Explanation

The following table summarizes the key features of the kernel and series regression frameworks Prof. Duflo discussed in class. You should be have an intuition for why these are true, if not, we suggest that you rewatch the segment.

| | Kernel Regression | Series Regression |
|-------------------------|--|--|
| <u>Trade-off</u> | For fixed N , a more flexible functional form (smaller band-width) increases variance and decreases bias. | For fixed N , a more flexible functional form (more terms in the polynomial) increases variances and decreases bias. |
| <u>Promise</u> | To make your bandwidth smaller as N increases. | To increase the number of terms in your polynomial as your sample size increases. |
| <u>Cross Validation</u> | The trade-off between bias and variance by using cross validation criterion, usually by minimizing mean squared error. | |

Omitted Variable Bias (23 Questions)

Question 1

Going back to the impact of education on earnings. You regress some function of wages on years of education, and then feel like you should be controlling for age. You think about what wage would look like as a function of age: when you're young you make very little money, once you get a job your wage increases significantly, and then maybe increases very gradually thereafter until you retire. You are pretty sure wage as a function of age might look like x and x^2 . You do some research, and find evidence that this in fact the case.

Which of the following regression frameworks is most appropriate for your model?

series regression

dummy variable approximation

kernel regression

none of the above



Explanation

Polynomial regression and dummy variable approximation are useful when the functional form is unknown, they are ways to estimate the functional form. But if you know what functional form the relationship between age and wage looks like, you should transform your regressor, and simply include **Age** and **Age²** as a regressor in your model.

Question 2

When using dummies for approximation, why should you exclude the intercept or one of the dummies?

Multicollinearity

To lower the degrees of freedom

To minimize the mean squared error

To decrease R squared

Omitted variable bias



Explanation

Dummy variable approximation works by partitioning the range of your dependent variable, and including a dummy for each bin. If you include the full set of dummies, as well as the intercept, you will have a collinearity problem.

Omitted Variable Bias (23 Questions)

Question 1

Why might defining piece-wise linear variables perform better than approximation using dummy variables?

More accurately modeling complex/flexible functional forms

Accounts for omitted variable bias

Less concern about multicollinearity

Coefficients become easier to interpret



Explanation

Simply approximating with the dummies creates artificial discrete steps. Piece-wise linear variables will estimate lines at each interval, which can better reflect smooth complex functional forms.

Question 1

True or False: Local linear regression is the same as a kernel regression, except that instead of taking a weighted average of observations within an interval as your predicted value, you take the predicted value obtained by running a weighted regression over the set of observations in that interval. In both cases, the weights are given by a kernel.

True

False



Explanation

The statement above is true. If you are unclear about this, we recommend you go back and rewatch this lecture segment.

Omitted Variable Bias (23 Questions)

Question 2

Which of the following is true about local linear regressions? Select all that apply.

As the number of observations increases, the interval size can decrease without compromising your estimates.

They are always preferable to kernel regressions.

The derivative of the function on a given interval is the coefficient of the weighted linear regression on that interval.

They are preferable to kernel regression for producing predicted values at the boundaries.



Explanation

As the number of observations increases, you can make the intervals smaller without making your estimates imprecise. The coefficient on interval will be the slope of the line over that interval, so it will also be the derivative of the function. Local linear regression performs better than kernel regressions at the boundaries, because they predict a line to the boundary whereas in kernel regression the predicted value is the weighted mean of observations within that interval.

Question 3

True or False: If you have a jump" or discontinuity in your data, you are better off using a kernel regression as opposed to a local linear regression.

True

False



Explanation

If you have a discontinuity in your data, and you care about the performance of your model near that discontinuity, you should use a local linear regression, since kernel regression performs relatively poorly at the boundary.

Question 1

A regression discontinuity approach is appropriate in cases where:

An outcome variable shifts discontinuously

The hypothesized functional form is non-linear

The probability of a particular outcome being realized shifts discontinuously with a running variable

One uses a locally linear regression



Explanation

This discontinuous shift in treatment across some running variable creates an opportunity to test the effect of the treatment on a given outcome variable.

Omitted Variable Bias (23 Questions)

Question 2

What assumption underlies regression discontinuity designs?

- The running variable is not correlated with your outcome variable.
- The running variable has no effect on the outcome variable before or after the threshold.
- There are no variables that are correlated with both the running variable and the outcome variable.
- There is no other reason for a discontinuous jump in the outcome variable at the threshold other than that the treatment has changed.



Explanation

The running variable or other variables may have an effect on the outcome variable as long as the effect does not change discontinuously at the threshold. However, if there is a reason for the outcome variable to discontinuously jump at the threshold besides the treatment change, then this jump cannot be distinguished from the effect of the treatment change. In other words, all other variables that might be correlated with treatment, need to be the same on both sides of the jump.

Question 1

When running an RD, what problem might you run into if you try to impose a linear form on the data?

- There might be other variables that are correlated with the regressor.
- Non-linearities may disguise themselves as discontinuities
- Treatment does not affect the outcome variable
- There maybe omitted variables



Explanation

If you impose a linear functional form, you might mistake nonlinearities for discontinuities (like in the example from Mastering Metrics). Omitted variables should not be a problem if the assumptions associated with regression discontinuity are true. If treatment does not affect the outcome variable, then regression discontinuity should correctly estimate no effect.

Omitted Variable Bias (23 Questions)

Question 2

What else could you do to strengthen the credibility of your estimates from your RD? (Select all that apply)

Look at whether other variables vary discontinuously at the threshold. ✓

Decreasing the bandwidth ✓

Increase the bandwidth

Estimate polynomial functions of the running variables on both sides of the threshold ✓

Use a kernel regression

Explanation

In a parametric RD you center your variables and fit polynomial functions of your running variable on both sides of the cutoff to help distinguish discontinuities from nonlinearities. You can even allow for the coefficients to be different on both sides of the cutoff. These are things you can do in the simple parametric RD framework. To justify your assumption, you can look at whether other variables vary discontinuously at the threshold. In nonparametric RDs, usually local linear regression is used.

You can also run a nonparametric RD, which exploits the fact that the problem of distinguishing jumps from nonlinear trends is probably less bad as we zero in on points close to the cutoff. The drawback of this, is that by restricting your observations to a narrow window, you lose a lot of observations. So your promise in an RD is to decrease your bandwidth as the number of observations increases (remember, there is the bias vs. variance tradeoff!)

There is no reason for you to use a kernel regression, since local linear regression performs better at the boundaries.

Question 1

True or False: Any variable that is correlated with your regressor of interest is an omitted variable if you don't control for it in your model.

True

False ✓

Explanation

Variables that are correlated with your regressor are only a problem if they also affect your outcome. (Ex. If you are interested in the effect of years of schooling on your wages, and you fail to control for ability- it could be that schooling doesn't affect earnings, but earnings are partly determined by ability. Since high ability individuals are more likely to stay in school, if you don't control for ability, you might overstate the effect of schooling. However, if ability doesn't affect earnings, then this would not have an impact on your estimates.

Omitted Variable Bias (23 Questions)

Question 2

Why would you control for SAT, parental income, and group fixed effects in a model that estimates the impact of attending a selective school on earnings?

- To control for omitted variables.
- To reduce selection bias.
- To increase the likelihood that potential outcomes would have been the same for those who attended a private college and those who didn't.
- All of the above.



Explanation

Underlying your interpretation of your model is the assumption that potential outcomes would have been the same for those who attended a private college and those who didn't. Both selection bias or omitted variable bias would violate that assumption. In this case, if you just regressed earnings on private school attendance, your coefficient might be biased, because it might be that smarter kids are more likely to be richer, and also more likely to attend a private school. The authors argue these measures are sufficient proxies for these omitted variables, therefore including them in the model would reduce selection bias and control for omitted variables. In turn, this would also increase the likelihood that potential outcomes would have been the same for those who attended a private college and those who didn't.

Question 3

Why should the group dummies be included in the regression?

- They directly affect future earnings
- They control for unobservable factors
- They prevent multicollinearity
- They are a non-linear transformation



Explanation

The group dummies can control for unobservable factors, like the desirability or motivation of the student, which would affect their future earnings and be correlated with whether they attended a private college.

Omitted Variable Bias (23 Questions)

Question 1

What, if anything, can you do to preserve your number of observations, while avoiding introducing bias, if you have missing data in some of your regressors? (Select all that apply)

Drop the observations with missing data completely.

Replace missings with 0 and include a dummy for missing.

Omit these regressors from your model if they don't have explanatory power.

Make educated guesses on what the value for this missing data would be.



Explanation

One way to handle missing data, and preserve the number of observations, is to replace missings with 0s, and include a dummy variable for missing as a control in your regression, assuming this variable does not take the value 0. The intuition for this is that you can correctly estimate your coefficient by doing this, since your indicator variable controls for any differences in outcomes between observations with missing data, and observations without. Another thing you can do is omit the regressors from your model if they don't have much explanatory power, as in the case of the group dummies discussed in lecture.

Question 2

True or False: In the Dale and Krueger study on the returns to private school you saw in lecture, the private school effect on earnings was much smaller in the model without controls.

True

False ✓

Explanation

The results showed that the private school effect was around 13% in the model without controls, relative to 1.2% in the model with controls. This suggests that the controls are proxying for some unobservable characteristics that are correlated with private school attendance and also affect earnings.

Omitted Variable Bias (23 Questions)

Question 3

According to the true model, what is the effect of private school on future earnings?

- a. There is no significant effect of private school on earnings
- b. There is a significant positive effect of private school on earnings
- c. There is a significant negative effect of private school on earnings
- d. There is an ambiguous effect of private school on earnings.



Explanation

The coefficient is **.013**. The standard deviation is **.025**. This means the t-stat (approx. $.013 / .025$) will be far less than **1.96** (the minimum t-stat that would be significant at the 5% level).

Question 4

In the model with only a dummy for private college, the coefficient on private college is:

- a. unbiased
- b. upward biased
- c. downward biased
- d. the direction of the bias is ambiguous



Explanation

The coefficient is upward biased, since it is smaller and insignificant in the true model.

Question 5

True or False: If you have more variables, you should always include them as controls in order to reduce the chances of OVB.

- a. True
- b. False



Explanation

Remember, there's no free lunch! Although you should control for omitted variables, your standard errors pay a price for each additional term you include. So there is a tradeoff between the number of controls you can include and your precision.

Omitted Variable Bias (23 Questions)

Question 1

Suppose Mark is analyzing an RCT randomizing the introduction of a one-on-one after-school tutoring program (T) on test scores (Y) for students in a certain school in a certain developing country. The tutoring program takes place every day and is randomized at the individual level.

Because this is an RCT, he would usually run $Y \sim T + \varepsilon$ to determine the causal impact of T through its coefficient. However, he remembers from doing fieldwork in that country that gender (G) is highly determinant on outcome; in particular, a widespread phenomenon is that girls are often required to stay at home to help with housework instead of attending school on several days of the week. Thus he argues he must run $Y \sim T + G + \varepsilon$ to avoid omitted variable bias in estimating the average effect of T .

However, Mark is told by Prof. Duflo that since the sample size is very large, it usually shouldn't matter which specification he uses. What does she mean?

For a randomized experiment, G is uncorrelated with Y in expectation (it has a small coefficient in the "long" model).

For a randomized experiment, G is uncorrelated with T in expectation ✓

Explanation

Mark's instinct that gender is determinant of outcome is well taken. However, the randomization ensures that gender is uncorrelated with treatment assignment, which means that $G \sim T$ would reveal a negligible coefficient for a sufficiently large sample. Recalling the OVB formula,

$OVB = \text{Effect of the omitted variable on the included variables} * \text{Effect of omitted in "long" (true) model}$

While the second term is significant, the first term is zero in expectation.

Another, equivalent way of saying this is that the proportion of girls and boys for $T = 1$ and $T = 0$ should converge in expectation (i.e. for a sufficiently large sample). Thus, not explicitly controlling for gender shouldn't bias the estimate on T .

Question 2

For the same randomized experiment, under what circumstances does $Y \sim T + G + \varepsilon$ control possible bias in the vanilla specification $Y \sim T + \varepsilon$?

As mentioned in the previous question, controlling for G via $Y \sim T + G + \varepsilon$ is never necessary.

If in the particular sample, G and T are "accidentally" correlated (i.e. a low probability event occurs), e.g. somehow more boys were assigned to the tutoring program than girls in the randomized assignment. ✓

Explanation

The previous question is a statement of what happens in expectation. However, for a particular sample (especially a smaller sample), one could be unlucky and have an accidental correlation between T and G , which *would* bias the estimate of T in $Y \sim T + \varepsilon$.

Several techniques can be deployed for this: [stratified random sampling](#), baseline balance checks to check for these imbalances empirically, and, more recently, ex-post controlling for possible confounders through techniques such as [double post lasso](#).

Omitted Variable Bias (23 Questions)

Question 1

Why is the fact that the coefficients on SAT scores, and parent's income disappear once you control for the group dummies?

- It suggests that the group dummies are probably a good proxy for unobserved background variables (since they are a good proxy for the observed variables).
- It suggests that there may be other omitted variables.
- It suggests that our measures of SAT score and parent's income is very imprecise.
- It tells us that SAT score and parental income have no impact on earnings.



Explanation

Because our estimated private school effect is insensitive to the inclusion of the available ability and family background variables once the group controls are included, other control variables, including those for which we have no data, might matter little as well. In other words, OVB due to uncontrolled differences is probably modest.

Question 2

Suppose that those who eat healthier also exercise and their exercise makes them more alert during the day.

In what direction is your estimate likely to be biased if you fail to control for exercise given the relationship described above?

- downward biased
- not biased
- upward biased
- cannot determine before looking at the estimates.



Explanation

The relationship described above suggests that exercise is negatively correlated with a high fat diet, and negatively correlated with day time sleepiness. So going back to our OVB formula, since we expect both terms to be negative, the OVB will be positive and therefore the coefficient is biased upward.

Omitted Variable Bias (23 Questions)

Question 1

Which of the following is **not** an example of advanced techniques you can use to get at a reasonable causal model?

Dummy variable approximation

Propensity score matching

Machine Learning techniques

Matching methods



Explanation

Matching methods can help reduce omitted variable bias by allowing you to control flexibly for a set of covariates which are known to be correlated with treatment, without sacrificing too much precision. (propensity score matching is just one way of doing this) Machine learning techniques if you have a lot of variables and don't know which ones are relevant (which ones might be collinear) etc. On the other hand, dummy variable approximation does not help with omitted variable bias, it is just another way of fitting a curve to your data.