

Q&A Chat Webinar 1 (Potential Outcomes and Omitted Variable Bias I (Theory))

February 29th 2024

	Question	Answer
1.	How do you find non renters?	Non renters = people who own a house. Renters = people who rent a house
2.	Does it mean the outcome variable is 1 renter, and 0 for non-renter? why did you use OLS?	The outcome is log (housing expenditure). The key predictor of interest is a dummy variable equal to 1 if a person rents their home, 0 if they own.
3.	What happens in case of more than two outcomes?	Each regression only has one outcome. Perhaps you mean what happens if the X variable of interest, in this case 'renter', is not binary. The logic is the same. The Bias Formula still holds and can be used.
4.	Where is the difference between ui and ei?	There is no difference
5.	Would the omitted variable be captured by the error term?	Yes, the omitted variable ends up in the error term. But, if the omitted variable is correlated with X, and it is in the error, then the error is correlated with X! This leads to bias in your OLS coefficient. This is exactly the issue we are discussing--bias in the OLS regression framework caused by omitting an important variable.
6.	Following the lesson keenly, I can deduce that basically the Z will serve as the instrumental variable in the model, right?	Not really. We will have a separate session on IV. What we are talking about is an omitted variable that belongs in the regression itself. Remember an instrument does NOT belong in the regression itself, rather, it is used to predict X. That is different model.
7.	So, by induction, the more variables are omitted the bigger the bias, and at what point do you cut the bias?	Not necessarily. You need to use theory to deduce whether there is potential selection bias in your model. The Bias Formula can help you. Remember omitted variable will lead to bias ONLY if the omitted variable correlates with your explanatory variable of interest AND the outcome variable. On Slide 8, we added the variable 'URBAN', it is clearly an important predictor of housing expenditure, yet when we added it, the coefficient of 'renter' hardly changed. Hence, omitting this variable did not really bias our coefficient of 'renter'! Why is that? Because while 'URBAN' is highly correlated with the outcome, it is NOT highly correlated with 'renter'. So not all omitted variables lead to bias. They need to be correlated with both the outcome AND the included variable.
8.	Should we call the first term of the selection bias term covariance? I	Yes, this is correct, it is the OLS coefficient estimate of the excluded (omitted) variable. Looking at the formula, you see

	thought it is a parameter of the endogenous variable with the excluded variable.	that it consists of $COV(Z,Y)/VAR(Z)$. The sign of this depends on the numerator only, so when trying to derive the sign, all we need to do is focus on the numerator, which is $COV(Z,Y)$!
9.	Hello, can you please explain how you got the covariance between income and renters?	On slide 31 , I 'backed out' the sign. On that slide, in the top text box third line, I know -0.093, I know 0.119, and I know that $COV(\text{income, housing spending})$ is positive--look at the coefficient of LOG_PCINC in the table below, it is positive (+0.399). Knowing all this, and knowing that the term in curly brackets {} must be negative, then it follows that $COV(\text{income, renter})$ must be negative. This is how you use the Bias Formula.
10.	I did not get how you got the covariance of income and housing and income and renter. Please can you re-explain. Thank you	See the explanation for Q9 above. It may be good to review the video around slide 30-31 .
11.	How would we make sure 0.119 is the true coefficient, what if we missed to include another potential confounder?	Yes indeed, there could be other potential confounders, that is up to you to decide based on theory. Here we are showing an example of one potential confounder to illustrate how to use the Bias Formula. In the real world you may have multiple omitted variables, and you need to figure out the direction of bias for each one. And if you so many of them, then unfortunately it probably isn't worth moving ahead with the study! Your data isn't up to the job.
12.	Yes, I think the omitted variable is lumping together with the error term but that would violate the basic assumption regression, exogeneity condition.	Yes, I now understand what you mean. The omitted variable and anything you are not including ends up in the error term. See response in Q5 .
13.	How do I know if I have omitted variable/ I encounter omitted variable in the model?	Your first guide is theory--use theory to help you. Then hopefully you have the variable in your data and can include it and see if the coefficient of interest changes.
14.	How do we know whether z is omitted or not from the very beginning?	See response to Q13
15.	Suppose data is not available even though you would like to include an important variable?	This is usually the case. You have to think about the direction of the bias and how that would impact your regression coefficient. This is the point of this session, using the Bias Formula to help you understand the direction of bias when you know you have omitted an important variable.
16.	Pls how did you get the 5.6	In Slide 34 I just solved the equation: $-40.8 - (-35.2) = -5.6$.
17.	So, the Selection bias for this final example is negative selection bias. When can we determine when it is large or not?	Usually, you only need to know the direction of the bias. You can't really know the magnitude of the bias because you would need to have the variable that your omitting, which if you do, then you should include it in your regression.

		From that we can then deduce whether our regression coefficient is overestimated or underestimated because of the bias. If the omitted variable is biasing your regression coefficient downward and you still were able to estimate a meaningfully significant coefficient, then bias would not really be a big deal (so to speak). On the other hand, if the omitted variable is biasing your estimate upward--meaning your regression coefficient is inflated (higher than it really should be) because of the omitted variable bias, then your reviewers/colleagues would not really trust your estimates.
18.	How do you suspect which omitted variable is potentially likely to induce selection bias?	See response to Q13
19.	Omitting variable and not doing interaction of variables is it the same? not sure if am asking the right qn	An interaction term can be considered an additional variable. But usually, we don't think of that in terms of omitted variable bias, it's more a specification error.
20.	Where did you get Data to estimate the omitted bias with the omitted variable. Since the omitted variable was not captured	Yes, this is the typical problem, we don't have the variable, so we have to try and understand the consequences of omitting it. That is where we use the Bias Formula to help us understand the consequences of omitting the variable, whether the bias is upward or downward.
21.	How do we know the research output has a selection bias?	Theory is your best guide. For example, if you are interested in understanding the effect of attending private school on student academic performance and you fail to control for, among other things, say parents' income, the regression coefficient on private school will be biased. Because we know from theory that parental income is correlated with private school attendance and parents' income is also correlated with student's academic performance. You need to have a good understanding of the literature and theories in your field of study.
22.	How did you get the -40.8 in the last example?	It comes from Slide 32 , it is the 'estimated' coefficient of ATTENDING_NOW, when we omitted age from the regression.
23.	When will you do Panel Data Analysis? We have some papers utilizing this	WE will cover fixed-effects May 23/30, and difference-in-differences March 21/28.

--	--

Q&A Chat Webinar 1 (Potential Outcomes and Omitted Variable Bias I (Application))

March 7th 2024

	Questions	Answers
1	If we have a categorical variable (which have value 1 2 3 for ex) or a binary variable among regressors in an OLS regression what command to use	Let's focus on selection bias. Whether you have a categorical variable or not, it should not really matter.
2	Will you have sessions on CGE analysis?	I do not think so for this webinar series
3	What if the selection bias is correlated with other variables in your model other than your interest variable?	Then they will also be subject to omitted variable bias.
4	How can someone notice there is omitted bias or selection?	You must use theory to guide you
5	Is B1 not the _cons?	no
6	What of a situation you don't have the omitted variable in your data, or it is difficult to get a proxy for it.	Indeed, this is the usual case--we know there is an important variable affecting the regression, but we do not have it in our data. And so, we use the bias formula to help us understand how this will bias our coefficient of interest.
7	Given the number of independent variables that can be used in any model how do you determine one specific variable as being the most important omitted variable? Do you test any variable you consider as being omitted?	Some omitted variables are known (from theory and past research) --e.g., if you fail to control for skills or experience in a wage equation, your regression will be biased. We know that. In general, you use theory to guide the specification. In the example on the returns to tertiary education, it is well-known that there is 'ability bias' In that equation. Similarly, when estimating health regressions, we know that the health endowment is important, and typically unobserved.
8	What if we also include additional explanatory variables?	So, the key is not to include all explanatory variables you can find, but rather to think carefully about whether there is selection bias in your model by looking at whether there is a variable that you omitted that influences both your treatment variable and your outcome. You should not throw as many variables into your regression as you can--theory must guide the choice of variables. And some of the variables that are dictated by theory may not be in your data set--this is where omitted variable bias comes in.

9	How can one deal with selection bias in the case of dependent variable is binary i.e., when using the logistics regression	You can use the same formula!
10	What is the exact difference between omitted variable and unobservable variable?	Observable variable: can't be measured. Omitted variable: can be observable (e.g., age) or unobservable (e.g., ability)
11	The material being used is practical and interesting. I would like to know some of the book references that match the contents of the course, especially on the theory of the topic(s)?	WE are compiling a bibliography that we will post to the website
12	Aside using the formula to check if there's omitted bias, is there a test that you can do in STATA to detect that? I am thinking it's easy to have OVB than it is to have a perfect model	Unfortunately, we do not have a simple test for OVB in STATA. That is because typically we do not have the 'omitted' variable in our data set--that is why it is omitted in the first place. Now omitting an important variable can lead to other problems in the regression, like heteroskedasticity. But this is outside our focus.
13	In a case where there is the ongoing debate about whether the recent fall in inflation (following the 2022 price spirals), is due to the lagged effects of monetary policy tightening or due to the natural fall, as the exogenous drivers fall off, can we examine some bias of associating this with monetary policy?	Yes, just think through the bias formula
14	Is it possible to notice that there is bias in your data immediately after finding regression results? I mean what is the highest number you can find and decide that you are over estimating? or low number indicating that you are under estimating?	Maybe. but there is no rule of thumb. you just have the selection bias framework to figure that out. In general, we don't ever calculate the actual bias because we do not have the omitted variable. Also, sometimes you might estimate a coefficient that looks 'wrong' to you, maybe it is very high, higher than you expected. This ought to trigger a deeper investigation into omitted variable bias--could you have left out an important variable which is leading to large upward bias in your coefficient?
15	In the case of unobserved variable, what if you employ confirmatory factor analysis, for latent variable leading to SEM structural equation modeling	If you can get a proxy for an unobservable variable, you will be fine so yes, this can work

16	How can we measure the innate ability	Unfortunately, we cannot measure innate ability, hence it is omitted.
17	How can estimate the exact figure of the bias?	We really do not need to estimate the exact figure. Ask yourself this question: Is your coefficient of interest under or overestimated because of the selection bias. That is the most important question.
18	What if the omitted variable correlated with multiple included variables?	That's a problem and often is the case. You would go through the bias formula for each of the variables. However, in most empirical analyses, we are only really interested in one main variable.
19	is it possible to have several omitted variables?? if yes, how do we apply the selection bias formula??	Selection bias formula is just a tool to help you figure out the direction of the bias. Whether you have negative or positive selection bias. If you have omitted more than one variable, then you have to look independently how each omitted variable biases your estimates. And they might cancel each other out!
20	How do I critically determine that a significant variable has been omitted in an empirical model?	Theory and your understanding of the literature. Some confounders are well known. If you are familiar with the literature and theory, you will also know them. In the example we gave on the returns to schooling, for these models, 'ability-bias' is a well-known issue.
21	How do we know unobservable variable is there? or do we need to start with, there might be there? I mean in default	A general understanding of the theory and the literature will guide you to identify omitted variables. See response to the previous question.
22	That is true, my question is will the result change if we include additional independent variable in addition to the omitted variable? Sorry if my question is silly	there are no silly questions. Thank you for your question. let me think... If the additional independent variables are proxies for the omitted variable, then that helps reduce the bias. So, it COULD help.
23	Treated group =1 however, control group =0	yes
24	If impact means causality. Does relationship and causality mean the same.	definitely not! relationship just means correlation.
25	With the just ended example, ability was identified as omitted but there is no proxy for it to be controlled for. Does that mean the estimates should not be taken seriously?	By failing to control for ability your result would be biased. At minimum, you must employ the bias formula and comment on the direction of bias. But many people would not consider that to be enough. Then yes, those people would probably not take the analysis seriously! :)

26	How does randomized control trials (RCTs) comes in this?	Remember you have selection bias when you omit a variable that influences both your key explanatory variable (call it "treatment variable") and your outcome of interest. So, if you think for a minute, if there is a way to make sure that your "treatment variable" is not influenced by an omitted variable (any omitted variable out there in the wild), then you would not have selection bias, right? what if treatment variable is random? if it is random, then it wouldn't be correlated with any other "omitted variables" out there in the wild. And there you go. RCT!
27	So how can I know by how much that my estimate is over or underestimated? or it mean that knowing if my estimate is under or overestimated is enough for the whole issue of the selection bias?	Look at the selection bias formula and think through it. Remember, scientific research is an art of persuasion. You (as a researcher) must convince your peers that your estimates are not spurious.
28	What I'm thinking is, is there a way to obtain the estimate for the counterfactual in the treatment effect case?	Yes we have experimental and non-experimental methods, see the last slide
29	Effect then can also be considered to imply causality, right?	Not everyone uses the term 'effect to mean impact. For example, if a person is interested in the effect of gender on income. They could simply be interested in the relationship between those two variables (gender and income). Or they might be interested in the causal effect of gender on income. Many times, 'effect is used in a general sense to mean correlation or causation. If a person is interested in the impact of gender on income, they are looking for a causal relationship. Some people do use the word 'effect' to mean causation, as in cause and effect'. You should always try and clarify.
30	Is there a link between the selection bias and the placebo effect?	Selection bias could be due to a placebo effect. People may voluntarily participate in a training program and even though the training program itself doesn't teach them anything, they FEEL like

		they know more, and are more confidence, and perform better.
31	The selection bias term equation is wrong in the example but the calculation are correct	I checked slide 19 and it looks correct, it's the same as in slide 18
32	RCTs offer a structured method to test hypotheses and mitigate the impact of omitted variables through random assignment.	Potentially yes. The problem with observational studies ("non-RCT") is that individuals choose/select, using a number of factors that we may or may not observe, their own "treatment"/"program" assignment ($P=1$ or $P=0$, in our potential outcome framework). Therefore, your treatment variable (P) in observational studies, by default, is influenced by all the factors that made individuals select into $P=1$ or $P=0$. And if you do not know these factors (variables) and/or you cannot measure them, then you will potentially be introducing bias in your estimates. That is to say, your regression coefficient of P (treatment variable) will be biased by all and any variables that are correlated with P and Y (outcome). The RCT removes this bias through random assignment to treatment.
33	Please share the slides	https://www.airp3-africa.org/applied-quantitative-analysis-webinar-series
34	What about the Exogenous switching equations?	that is a good one
35	Would it be appropriate to use methods such as LASSO to help us identify the variables? So as to reduce the selection bias?	Lasso is a data-driven method to identify the best set of variables that predict Y . We are not discussing prediction here; we are discussing causal inference. Does X cause Y ? Lasso is used for other purposes.
36	Please when the R Squared is small for OLS analysis. Please what do you do?	See response above, we do not really concern ourselves with R-squared when doing causal inference. For prediction models then yes, R-squared will be relevant.
37	Is it feasible to get downloadable recording to allow those without strong internet to come back to the course after, offline?	We are checking on this.
38	Is it really true that we can observe Y zero for $P=1$?	No, that is the problem of the 'missing counterfactual', we do not observe the person's outcome without the program if they are in the

		program! So, we have to estimate it (using a comparison group).
39	Are the values estimated or gotten through surveys?	You can get them through a survey
40	What if the selection bias is correlated with other variables in your model other than your interest variable?	Then the coefficients of those other variables will also be biased, and the direction of bias can be figured from the bias formula
41	Is it really true that we can observe Y zero for P=1 in an RCT?	NO. You cannot observe the same individual with the program and without the program. Because once a person receives the program, you only see that person's outcome with the program. You would need a to go back in time to be able to see that same person without the program. But that requires some sort of time machine (to travel back in time). I do not believe we have that technology yet. RCT, unlike observational studies, allows us to cleanly disentangle the effects of other variables that are omitted in our model--biasing our results. RCT achieves this by ensuring that individuals are not choosing/selecting their own "treatment" (program). Instead, RCT distributes individuals into a treatment or control condition using a random change (say a flip of coin). A flip of coin, by design, cannot correlate with any variable that we might be omitting in our model (provided your samples are large enough).
42	How do you know the omitted or bias variable	By developing a good understanding of the theories and literature in your field.