Notes 16: Omitted, superfluous, and proxy variables.

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Review of omitted variables

This is perhaps the most important type of model failure. In fact, many failures of other assumptions can be understood as omitted variable bias, as we will see later on. Remember the theorem:

Theorem 5. *Suppose that w is a confounder which was omitted from the regression (perhaps because it was unobservable), then*

$$\hat{\beta}_1 \approx \beta_1 + \hat{\beta}_w \hat{\theta}_1$$

where

- $\hat{\theta}_1 \approx \mathbb{E}[w|x_1 = 1, x_2, \ldots, x_k] - \mathbb{E}[w|x_1 = 0, x_2, \ldots, x_k]$ *is the average change in w when x_1 increases one unit, holding $x_2, \ldots, x_k$ constant.*

  (Formally, it is the slope coefficient of x_1 in a regression of w onto $x_1, \ldots, x_k$. This is mathematically equivalent to the following: define $\eta = w - d_0 - d_1 x_2 - \cdots - d_{k-1} x_k$, the residual of a regression of w onto the controls $x_2, \ldots, x_k$. Then θ_1 is the coefficient of a regression of η onto x_1).

- $\hat{\beta}_w \approx E[y|x_1, \ldots, x_k, w = 1] - E[y|x_1, \ldots, x_k, w = 0]$ *is the average change in y when w increases one unit, holding $x_1, \ldots, x_k$ constant.*

  (Formally, it is the coefficient of w if we could run a regression of y onto $x_1, \ldots, x_k, w$. This is mathematically equivalent to the slope coefficient of a regression of y on η, which is the average change in y when we increase η in one unit.)

Let's think of an example different from the last time, so that you can get the pattern. Consider the model:

$$birthweight = \beta_0 + \beta_1 cigarettes + \beta_2 income + \beta_3 education + \beta_4 married + u$$

Ideally, we would have $\mathbb{E}[u|cigarettes, income, education, married] = 0$. However, that is unlikely. There are way too many omitted variables in this problem. Let's think of one: $drinks$ = number of drinks per week. We want β_1, the physical effect of smoking. We don't want it mixed up with the effect of drinking. Do you expect the bias to be positive or negative? _____

Now, let's take a look at the formula. $\hat{\theta}_1$ is the average change in *drink* when we increase 1 unit of cigarettes, holding the covariates constant. What do you expect $\hat{\theta}_1$ to be? _____.
However, to go deeper into the issues of the problem, it is always advisable to check the alternative explanation using the residuals. η is the part of drinking which isn't explained by income, education, or marital status. It is a mix of the actual taste for drinking, the amount of risky behavior in which the person engages, addiction, social activity, peer pressure, etc. More smoking should mean more or less of the things? _____.

Now, think about $\hat{\beta}_w$. That is the expected change in *birthweight* when we increase *drinks* by one unit, holding *cigarettes*, *income*, *education*, and *married* constant. What do you expect $\hat{\beta}_w$ to be? _____.
Now, let's examine the residual explanation just to be sure. What is the effect of taste for drinking, risky behavior, addiction, social activity, and peer pressure on birth weight? Should it increase or decrease the birth weight? _____

Hence, the bias is? _____

Remark 1: Observe that I never said that w is unobservable. Sometimes a variable is omitted because it is unobservable, or at least is not in the data set. However, the same analysis can be done with any set of variables, including those that we do observe in our data set.

Remark 2: I recommend that you always think of one omitted variable at a time. However, if you must look at more than one, then the formula is

$$\hat{\beta}_1 \approx \beta_1 + \hat{\beta}_{w_1}\hat{\theta}_{11} + \hat{\beta}_{w_2}\hat{\theta}_{12}$$

where

- $\hat{\theta}_{11} \approx \mathbb{E}[w_1|x_1 = 1, x_2, \ldots, x_k, w_2] - \mathbb{E}[w|x_1 = 0, x_2, \ldots, x_k, w_2]$ is the average change in w_1 when x_1 increases one unit, holding $x_2, \ldots, x_k, w_2$ constant.
- $\hat{\theta}_{12} \approx \mathbb{E}[w_2|x_1 = 1, x_2, \ldots, x_k, w_1] - \mathbb{E}[w_2|x_1 = 0, x_2, \ldots, x_k, w_1]$ is the average change in w_2 when x_1 increases one unit, holding $x_2, \ldots, x_k, w_1$ constant.

- $\hat{\beta}_{w_1} \approx E[y|x_1, \ldots, x_k, w_1 = 1, w_2] - E[y|x_1, \ldots, x_k, w_1 = 0, w_2]$ is the average change in y when w_1 increases one unit, holding $x_1, \ldots, x_k, w_2$ constant.
- $\hat{\beta}_{w_2} \approx E[y|x_1, \ldots, x_k, w_1, w_2 = 1] - E[y|x_1, \ldots, x_k, w_1, w_2 = 0]$ is the average change in y when w_2 increases one unit, holding $x_1, \ldots, x_k, w_1$ constant.

2 Superfluous variables

What happens if a variable which should not be included in the model is included by mistake? First, what is a superfluous variable?

- **A variable is superfluous** _____

Let's give it a notation, so we can do the math. Say that x_k is superfluous. Then this translates into three possibilities:

1. _____
2. _____
3. _____

_____

What do these things mean mathematically?

1. The first one is easy. Try it: _____
2. The second one is a bit more complex. It implies that _____.

   Try to figure out why this is the case. Hint: x_k not related to x_1 is the same as x_1 not related to x_k.
3. The third reason is very complex. It means that _____.

   It also implies that _____.

   Try to think why this is the case. Hint: the other covariates rendered x_k redundant. Alternatively (you will understand this later), the other covariates are proxies for x_k.

So, what happens if we include a superfluous variable in the model (meaning, what happens if we use x_k in the regression, when it is superfluous?) The answer depends on the reason why the variable is superfluous.

- Suppose that the reason is because $\beta_k = 0$, in this case:

- What happens to $\hat{\beta}_1$? _____
- What happens to $\hat{\beta}_2, \ldots, \hat{\beta}_{k-1}$? _____

_____

- What happens to $Var(\hat{\beta}_1), \ldots, Var(\hat{\beta}_{k-1})$? _____.

  Remember the variance formula:

$$Var(\hat{\beta}_1) = \frac{\sigma^2/n}{\widehat{Var}(x_1)(1 - R_1^2)}$$

  Since x_k is not part of u, σ^2 does not change, and of course neither do n and $\widehat{Var}(x_1)$. However, $1 - R_1^2$ will be the same or smaller.

- Now, if the variable is superfluous because $\hat{\theta}_1 = 0$, then
  - What happens to $\hat{\beta}_1$? _____
  - What happens to $\hat{\beta}_2, \ldots, \hat{\beta}_{k-1}$? _____

    _____

    However, we don't care very much, since the model is written for the variable of interest. To understand the intuition of this point, think that x_2 is income, and x_k is wealth. Perhaps wealth is superfluous (once we have income in the model), so including it will not change the calculated effect of smoking. However, the coefficient of income will probably decrease. Think why this is the case.
  - What happens to $Var(\hat{\beta}_1), \ldots, Var(\hat{\beta}_{k-1})$? _____

    _____

Hence, the takeaway is that including a superfluous variable does not affect the bias of $\hat{\beta}_1$, but may affect the bias of the other coefficients (which we don't mind). It will likely increase the variance of the estimators. Hence, if you know for sure that a variable is superfluous, you should not use it. If you have doubts, it doesn't hurt very much to include some potentially superfluous variables anyway, so go ahead. However, don't include too many superfluous variables, since the variance can become extremely large.

3 Proxy variables

A proxy variable tries to solve the problem of endogeneity by approximating the omitted variables. The whole point is to use one or more variables (usually many variables) to try to render the omitted variable redundant. If we manage to do this, then the variable which was originally a confounder is no longer. It becomes a redundant variable (and therefore

superfluous), and thus does not need to be included in the model. Don't worry if you still don't understand this, we will examine this point in more detail.

Intuitively, you should think about what is the true omitted variable for which you would like to control. In the case of the effect of smoking on birth weight, we would like to control for all the other substances that the mother consumed, as well as her health. We can control for some of these things. For example, data sets on this topic often contain very good information on the mother's health, for example if she has diabetes or other chronic conditions. We can know her age, and even sometimes her height, weight, and the father's height and weight. However, the information on her nutrition and substance exposition is harder to find. We sometimes see information on whether the woman drinks, takes vitamins and other supplements, and sometimes even if she takes controlled medication. However, we never see how many meals she has per day, nor the quality of the food she consumes. We never see if she takes drugs, if she is exposed to environmental chemicals, etc. All of these are the true omitted variables in this model. Unfortunately we may never observe them.

Instead, we try to predict them so that they become redundant. The predictors are what we call proxy variables. Considering the example above, some controls in the model are the true omitted variables we would like to have, such as drinking. However, income, education, marital status and race don't cause birth weight. They are not the variables that directly determine birth weight. They are used as controls to serve as proxy variables for the things we cannot observe. The mother's income, education, marital status, and race are trying to predict the quality of her nourishment, the types of substances to which she exposes herself, the type of environment where she lives. Are they doing a good job? First let's formally define a proxy variable.

- **The variable z_1 is a proxy variable for the omitted variable w if** _____

  _____

  _____

Finding a proxy variable for the true omitted variables is the holy grail. It never happens, it is virtually impossible to find one variable which will completely account for another which we do not observe. However, a set of variables can serve as proxy for a single variable. For example, nourishment cannot be explained by income, but perhaps it can be explained by income, education, marital status, race, age, employment status, profession, number of hours the woman works, number of prenatal visits, and whether the pregnancy was planned.

- **The variables $\{z_1, \ldots, z_L\}$ are a proxy for the omitted variable w if** _____

How do we use the proxy variables? We use them as controls. Throw them in the model as if they belonged there. In fact, we have been doing this all along throughout this course. What does the exogeneity condition mean? $\mathbb{E}[u|x_1, \ldots, x_k] = 0$ means that $x_1, \ldots, x_k$ are proxies for everything that we cannot observe.

Hence, the takeaway is that when you write a model, you should approach it in the following way: think about all the factors which cause the outcome and which are related to the variable of interest. For example, in the case of the effect of smoking in birth weight, think about all the things that are related to smoking and also cause birth weight: nutrition, substances, and health. Those are the fundamental confounders. If you can observe some of those, include them in the model as controls. The ones you cannot observe, try to approximate them with other variables. In other words, use other variables as proxy for them. Include them in the model as controls as well.

Observe that this is not always easy to do. In fact, in the exact example of smoking and birth weight, a much larger set of variables than the ones we mentioned here is still not enough to account for all the omitted variables in this problem. However, looking for proxies should still be the first thing you do. Even when the proxy is not perfect ($\mathbb{E}[w|z_1, \ldots, z_L] \neq \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_L z_L$), you can decrease the bias substantially using a good, but imperfect proxy (when $\mathbb{E}[w|z_1, \ldots, z_L] \approx \gamma_0 + \gamma_1 z_1 + \cdots + \gamma_L z_L$, so that $\varepsilon = w - \gamma_0 - \gamma_1 z_1 - \cdots - \gamma_L z_L$ is either not very correlated with y, or with x_1.) Intuitively, your proxies may not make w redundant, but they may make it almost redundant.

Should you throw in everything in the data set in the hope that it serves as proxy for things you don't even imagine? Well, depends. If your data set is really very large, this is sometimes a good idea. Just use every piece of information you possess about the mother, and throw them as controls. This approach is known as the **kitchen sink** regression. However, if you have indeed a lot of information, and not too many observations, then we are getting into a dangerous territory because of the variance of the estimators. How many observations? How many controls? Depends on two factors: (1) the confidence you want, and (2) the size of the confidence interval you think is informative enough. For example, if you are happy with 95% confidence, the confidence interval is

$$[\hat{\beta}_1 - 2SE(\hat{\beta}_1), \hat{\beta}_1 + 2SE(\hat{\beta}_1)].$$

Suppose that the size of the confidence interval you would consider informative is 50 grams or less. Then:

$$4SE(\hat{\beta}_1) \leqslant 50 \implies SE(\hat{\beta}_1) \leqslant 12.5.$$

If you have many observations, keeping the standard error below 12.5 is easy, even with many controls. However, as we increase the number of controls, R_1^2 increases. Hence, the kitchen sink approach (or, as people say, "throwing in the kitchen sink") could make your estimators very imprecise. However, if you don't use the right controls, or if you indeed need to use them all, then not using them could make your estimators biased. So, accuracy or precision? Don't answer this yet. Stop, and do things more carefully.

Here is what you do: start with the most important controls, and keep adding controls in order of importance. If the standard errors become unacceptable, and you still have important controls you would like to include, probably your data set is not appropriate for the question you are trying to answer, and you should look for one with more observations. The answer is that you should not choose between accuracy and precision. If you are faced with that choice, you should look for a better data set, and if you can't find one, then you should abstain from answering this question. You just don't have enough information.

However, sometimes the important controls you wanted to add are important as a set. Each one of them is not very important. For example, if your data set has information about every possible disease the woman can have, then perhaps you don't need to add them all. Add diabetes, a few more important ones, and then separate the remaining diseases into categories, and instead of adding the specific disease, add: respiratory, cardiac, hormonal, etc. That's it, 5 to 8 controls instead of 300. If you have 500,000 observations, you don't need to do this, but if you have 5,000, perhaps this would be a good compromise. The idea is that some controls are fundamental, others are an overkill, and you may get away with condensing those into a few categories.

Though this is a nice discussion, remember that the problem you will usually face is not that of having too much information, and wondering how to include it all without blowing up the standard errors. Often the problem is that you don't have enough information. For example, although I am used to data sets that have very detailed information about the health of the woman (and you may need to condense all of it into categories), there are very few data sets that have simple questions like whether the woman was a smoker before the pregnancy, whether she planned the pregnancy, and when she found out when she was pregnant. The abundant information on the diseases can never make up for those three variables, condensed or not.