Notes 15: Heteroskedasticity and omitted variable bias.

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Foreword

You become an adult when you start making all your decisions, even if you are not sure, and paying the price for them. It is not different in research. You know the rules, you know how to do OLS right, you know how to interpret results under the world of assumptions 1, 2 and 3. You mature when you understand that in the real world things are never exactly like they are in the theory, and you have to make decisions, and sometimes risk being wrong. We will examine in the following classes all the ways in which our model can be wrong, how to watch out for problems, and what can be done sometimes. Welcome to adulthood!

2 Heteroskedasticity

When we were studying the regression method, we spoke about homoskedasticity and heteroskedasticity. You could suspect the problem by looking at the residuals. Do you remember how the residuals looked depending on the case?

You know informally what heteroskedasticity is, you even have an idea about what are the implications. First, what is heteroskedasticity?

- **Heteroskedasticity:** ____

  ____

You must reflect upon the meaning of this problem. Assumption 3 required that $Var(u|x_1, \ldots, x_k) = \sigma^2$. What does this mean? Think about our typical example

$$grade = \beta_0 + \beta_1 class + \beta_2 OH + \beta_3 sections + u.$$

So, let's think about one component of u. How about *responsibility*? Consider students that show up to 3 classes, no office hour, and 2 sections. How much do you expect their responsibility to vary? Probably not very much, these students are all not very responsible. How about the students who show up to every single class and section, and go to at least one office hour per week? How much do you expect their responsibility to vary? Not very much as well, these students are all very responsible. How about students that come to 80% of the classes, 80% of the sections, and one office hour per month? Now that is hard to predict, I can expect a great deal of variation. This group may include students that are responsible, but they are so smart that they don't really need to come to every class, or students that are less responsible, and come to more classes only to avoid having to study more. It's hard to say, there are probably many levels of responsibility that explain this behavior.

If we observed *responsibility*, and controlled for it, we would probably not see this pattern. Alas, heteroskedasticity is a problem that is related to the confounders. In other words, confounders can cause this kind of trouble. However, they can cause way more trouble than this, as we will study later.

- What happens if we have heteroskedasticity? ______________________________

  ____

It is important to notice that heteroskedasticity does not affect the fundamental properties of the OLS estimator. It still guesses right and close. In other words, it is still accurate and very precise. It is no longer the one that guesses the closest, the most precise of all, but this doesn't make it bad. However, we cannot use the old formula for the variance any longer. We need to recalculate it.

- Why do we need to recalculate the variance? Why did we need the variance in the first place? ____

  ____

All right, so here is the formula.

Theorem 4.

That looks weird, but it actually isn't. The interpretation remains exactly the same. The variance is affected by 4 factors: (1) the variance of the errors, (2) the sample size, (3) the variance of x_1, and (4) the level of multicollinearity. All the factors continue to affect the variance in the same way, except for (1). Now, the variances of the errors enter the variance equation as a weighted average of the variances of all the observations: $\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{r}_{1i}^2\sigma_i^2/n$. It seems fancy, but in terms of interpretation, it doesn't change much. If the variability of the unobservables is high, the OLS will be less precise, same as before, different formula.

So, you can still do everything we did before, all the hypothesis testing procedures, you just need to use the correct formula for the standard errors. Here is the formula:

where the $u_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i} - \cdots - \hat{\beta}_k x_{ki}$ are the residuals of the OLS regression. Of course, then

$$SE(\hat{\beta}_1) = \sqrt{\widehat{Var}(\hat{\beta}_1)}.$$

The formula for the variance estimator is very similar to the case with homoskedasticity. There, we substituted σ^2 by the average of the squared residuals. Here we do the same, except it is a weighted average. Although the formula is very natural, it turns out that proving that it worked took a long time. Researchers were very concerned with heteroskedasticity, and developed many estimators for the linear model in this case. Only in the 1980's the results that showed how to estimate the variance of the OLS appeared. Once they did, researchers went back to the OLS. The idea is that, sure, the OLS is not the best, but it behaves well, and we understand it. Other estimators that took care of heteroskedasticity (for example, GLS, FGLS, look them up) required that the researcher took some decisions, or ran preliminary steps, and they had problems of their own. Now that we have the right variance formula of the OLS, let's just keep using it.

So, what should you do? You should assume that there is always heteroskedasticity. Don't bother with assumption 3, simply use the OLS, and the variance formula for heteroskedasticity (which is known as the Eicker-White variance formula, by the way). If you have a large sample, over 1,000 degrees of freedom, most of the time you have no reason to worry. Besides, if your sample is not large enough, heteroskedasticity is probably just one of your many problems. So there, a problem with a relatively easy solution. It will only go downhill from here.

3 Omitted Variables

This is the fundamental failure of assumption 1, it is the essence of the problem of endogeneity. It is the most famous, and one of the most studied problems in econometrics. It has very bad consequences, and it is very hard to solve.

- **Endogeneity:** ____

  ____

When will this happen? There are many reasons why it will, but one reason stands out above the rest. Take our model

$$grade = \beta_0 + \beta_1 class + \beta_2 OH + \beta_3 sections + u.$$

- Do you believe that $\mathbb{E}[u|class, OH, sections] = 0$? ____________________

The discussion on heteroskedasticity evidenced the problem that was brought up by the responsibility of the student. We are interested in the causal effect of *class* on *grade*. We would like to control for your inherent quality as a student. Are you responsible? We can't influence that, I can't change who you are, but I can make the classes mandatory, I could force you to show up. Should I? Will it help? I can only find out if I know your responsibility, because this way I can control for it, I can hold it fixed, and see how much the extra class really affects the grade.

So, is *responsibility* a confounder? Yes, it is. It is related to *grade* and to *class*. Also, it is not redundant. How could it be? I can easily think of two students that come to all the sections, and no office hour per week, but have different responsibility. One is very smart. He doesn't know Stata, and he is lazy, so he knows that if he shows up to sections, he won't have to study at home. The other could be very responsible, but struggling, so he comes to Stata sections because he can't risk missing the material. So, can we predict the quality of this student? Not really, he could easily be the smart, lazy one, and he could be the responsible, struggling student.

Confounders are also known as "omitted variables." What we are studying is exactly the trouble that they cause. What is the relation between confounders and endogeneity? Well, *responsibility* is part of u. However, responsibility varies, even when we control for the observables. We need to know, when we control for *OH* and *sections*, whether *responsibility* varies with *class*. But of course! So, we got a student that comes to all sections, and no OH per week. Now, if he comes to every class, would you say he is the smart one, or the struggling one? Probably the struggling. He was struggling, and this is why he came to all sections. It would be natural for him to come to all classes. How about if he came to 4 classes? It was probably the smart and lazy one. He didn't know Stata, and thus came to sections, but he knows the material, he doesn't need to come to class very much. So, for few classes, the student is more likely to be the smart lazy one, and for many classes, the student is more likely to be the struggling and responsible one. I'd say that the responsibility probably varies with class, even controlling for *OH* and *sections*.

So, what is the problem? In theory, you already know. This is a failure of assumption 1, which will cause the OLS estimator to be biased. Intuitively you can already guess why. When we hold *OH* and *sections* fixed, and we vary *class*, the change in *grade* may not reflect the causal effect of classes, it may just reflect the change in the average responsibility of students. Now, let's try to give a formula to the omitted variable bias.

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

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 $\eta = w - d_0 - d_1 x_2 - \cdots - d_k x_k$, 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.)

This seems complicated, but it does make sense. You need to understand this. In fact, eventually you need to make these interpretations in your head, and fast. Think of the example. Remember the definition of $\hat{\beta}_1$, the coefficient of the regression. It is the average change in *grade* for an increase of one *class*, holding *OH* and *sections* constant. The change is due to the causal effect of *class*, β_1, and also because of the change in responsibility. To be precise, *responsibility* itself is not the trouble. The trouble is the part of *responsibility* that cannot be predicted by *OH* and *sections*, which is the residual η. So, when we vary one *class*, how much does η vary? $\hat{\theta}_1$. And how much does a change of 1 unit in η affects y? $\hat{\beta}_w$. So, the formula is pretty natural. What is nicest about it is the power it has to simplify the reasoning.

To see what the formula can do for us, let's try to guess what will be the bias. So, intuitively you know that even controlling for *OH* and *sections*, more responsible people go to more classes, and more responsible people do better in the grades. So, the bias is going to be positive, in the sense that the causal effect of classes is likely smaller than the OLS coefficient. Are you sure? Did you forget something in your reasoning? Sometimes things are not so simple.

Now, let's forget this intuition and just look at the formula. $\hat{\theta}_1$ should clearly be positive. After controlling for *OH* and *sections*, someone that goes to one extra class is still likely to be more responsible. So, $\hat{\theta}_1$ should be positive, there is little doubt about it. How about $\hat{\beta}_w$? That is tough. More responsibility could never hurt your grade, I am sure. What I mean is that the causal effect of responsibility in grad is bound to be positive. However, the bias is not about the causal effect, it's about the expected change (not *ceteris paribus*). Remember that, controlling for *OH* and *sections*, more responsibility could mean that you are a struggling student. So, it could very well be that a more responsible student that goes to the same office hours and sections as another, less responsible one, does so because he struggles, and perhaps his grade could be lower. Now, of course, he will also study more, and in the end perform better, so it's hard to know. So, what will be the net effect, the effect of studying more, or the effect of the struggle? It is unclear. Examining the bias formula allows us to think deeper, and see that, in this case, the bias could very well be negative.