Notes 3: Confounders

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

1 Causal Effects

- Why do we care that the treatment and control are comparable? For that matter, why do we care to find the causal effect?

If we just compare observations that chose different treatment levels, the differences in outcome won't be due only to the differences in treatment. They will also be due to the differences in other variables. Therefore, this comparison does not tell us what will happen if we change the treatment, but leave these variables fixed.

For example, we like to know how much will the grade change if we increase class attendance. We want this because we can actually influence class attendance. The problem with comparing students that chose different numbers of classes to attend is that they also tend to have many differences, for example, they tend to have different grades in the pre-requisite courses, and so their grade differences are due not only to the differences in class attendance, they are also due to differences in the pre-requisite grades. However, if we change class attendance, we won't be changing the pre-requisite grades (nor any of the other factors), so we need to know what will happen if we change class attendance *ceteris paribus*.

The same way, we would like to know what will happen to the babies' birth weight if we make women smoke less in pregnancy. The problem is that women that chose to smoke different amounts are different not only in how much they smoked, but in many other ways. For example, they tend to have different income levels. Hence, the differences in birth weights are due not only to differences in smoking, but also to differences in income. However, if we change how much they smoke we won't be

changing their income (nor any of the other factors), so we need to know what will happen if we change the smoking amount *ceteris paribus*.

- Is there a solution?

  Yes, but it's not too simple. The idea is that if we observe the variables that are changing as well as the treatment (such as pre-requisite grades and income in the previous examples), we can control for them. Let me explain what this means in the examples. We will go slow.

  If we observe students' pre-requisite grades we can fix a grade level and then compare students with the same pre-requisite grade, but different number of classes attended. What I mean is that we get only the students with the same grade at the pre-requisite courses. We then see how many classes they attended, and their final grades in econometrics. We know that the differences in the final grade can be due to the differences in class attendance. They can also be due to several other factors (natural ability, patience, responsibility, etc.), but they definitely cannot be due to differences in pre-requisite grades, because we held those fixed. You see what we did there? We eliminated one layer of the problem. Now we can still have plenty of issues, but pre-requisite grades is not one of them. When we do this, we say that we are using pre-requisite grades as _____

  Now let's think of the smoking mothers. If we observe the mothers' income, we can fix an income level and get only the mothers with that income level. Then we see how much they smoked, and what was the birth weight. The differences in birthweight can be due to differences in smoking amounts. They can also be due to other factors, like how much the mother wanted the pregnancy, the food she ate, etc., but most definitely the difference is not due to income differences, because we held income fixed. Now we can have many problems, but income is not one of them, because we controlled for it.

- Fine, we solved the problem with one variable, but what about the rest?

  Ideally you will observe several of these factors in the data set. What you do is you control for them, which means that you hold them all fixed. Let's think of the examples.

  If we observe the students pre-requisite grades, their SAT scores, the number of other courses they are taking, and the amount of time they spend on social media, we can hold those fixed. This means that we fix the pre-requisites grades, the SAT scores,

the number of other courses they are taking and the amount of time they spend on social media, and now we look at the difference in the number of classes attended and their final grades. Now, the differences in the final grades may be due to the differences in the number of classes attended, and maybe there are other factors as well (patience, responsibility, love for the subject econometrics, work and internships), but we eliminated several factors from the problem. Pre-requisite grades, SAT scores, the number of other courses they are taking, and the amount of time they spend on social media are no longer part of the problem.

Let's look at the pregnant women. Suppose that we observe income, education, marital status, race, number of alcoholic drinks per week, number of prenatal visits. Then we can control for those factors. Consider only women with the same income, education, marital status, and race, who drank the same number of alcoholic drinks per week, and did the same number of prenatal visits. Now the differences in birth weight can be due to differences in smoking, and even to differences in other factors (nutrition, self control, how much they care about the pregnancy), but the differences in birth weight are most definitely not due to differences in the factors for which we controlled. They are no longer part of the problem.

NOTICE: In order to get sufficient observations with the same values of the controls, we will need a data set with a _____.

How many observations are enough? We will discuss this later in the course, but always the rule is: the more the better!

- How can we control for every possible variable? There must be millions!

  You are probably right. However, all is not lost. The beauty of controlling for things is that everything else leftover becomes less problematic. Let me explain in the examples.

  In the class attendance example, remember that we controlled for the pre-requisite grades, the SAT scores, the number of other courses students are taking, and the amount of time students spend on social media. Now we are concerned about the responsibility level. However, students with the same exact pre-requisite grade, SAT score, same number of other courses and the amount of time students spent on social media are likely to have similar responsibility levels as well. So, by controlling for several variables, we decreased the problems generated by other variables for which we did not control.

  In the smoking example, we controlled for income, education, marital status, race, number of alcoholic drinks per week and number of prenatal visits. We are still

concerned about nutrition. However, women with the exact same income, education, marital status, and race, who drank the same number of alcoholic drinks per week, and did the same number of prenatal visits are bound to have similar nutrition levels as well. By controlling for several things we decreased the problem caused by nutrition, for which we did not control.

Now, a little bit of language. Later in the course we will define these same terms formally in mathematical terms, but you can begin to use them in an intuitive level.

- If we did not control for a variable, but we controlled for so many other things that this variable was no longer a problem at all, we say that the controls _____

  _____

  In the class attendance example, we could say that pre-requisite grades, the SAT scores, the number of other courses students are taking, and the amount of time students spend on social media predicted the responsibility level. Now, this is probably not the case in this example, but if we had controlled for more things, perhaps we could indeed predict the responsibility.

- If a variable is predicted by the controls, we say that this variable is _____

  _____

  In the smoking example, if income, education, marital status, race, number of alcoholic drinks per week and number of prenatal visits predict the nutrition level, then we say that nutrition is redundant.

- Remember how we discussed confounders in the previous class? Now we will complete the definition of confounders. A variable is a **confounder** if it is:

  1. _____
  2. _____
  3. _____

For example: is the natural ability in econometrics a confounder in the problem of the effect of class attendance on the final grade?

1. Is it associated with the treatment?

   Yes. Students with higher natural ability grades may attend more classes because they are more pleasant to them. They may also attend less classes, because they can learn the material easily. Whatever is the predominant association, the point is that there is an association.

2. Is it associated with the outcome?

   Yes. Students with higher natural econometric ability tend to have higher grades in econometrics, for example because they may be able to wing a question in the exam better, or because they get more out of studying in comparison to the other students.

3. Is it not redundant?

   The question here is whether pre-requisite grades, SAT scores, the number of other courses students are taking, and the amount of time students spend on social media predict the natural ability in econometrics. Certainly students with the same pre-requisite grades, SAT scores, number of other courses amount of time students spent on social media will tend to have similar ability in econometrics, but the same ability? We don't know the answer, but we will know a lot more about this problem as the course progresses.

Hence, we are concerned only about those variables that are confounders. Even if a variable seems related to the problem, if it is redundant, and therefore not a confounder, it does not cause us trouble.

A good observational data set to answer our scientific question must contain:

1. _____

2. _____

3. _____

4. _____