Notes 18: Measurement error

ECO 231W - Undergraduate Econometrics

Prof. Carolina Caetano

In the next two classes we will be talking about failures of the data (as opposed to failures of the model). Remember assumption 2:

Assumption 2: The data $\{(y_i, x_{1i}, \ldots, x_{ki}) : i = 1, \ldots, n\}$ was randomly collected from the population, and none of the regressors is constant, or has an exact linear relationship with the other regressors.

This assumption has three basic components. One of them is implied indirectly, but it turns out to be rather important.

1. ______________________________

2. ______________________________

3. ______________________________

Today we will be talking about the third component. What happens if we don't quite observe some of the components, but instead we observe a measure which may be mistaken?

1 Measurement error

We call measurement error all forms of mismeasurement in the variables in the model. The idea is that the true variable is, for example, x_1^*, but instead what we observe is a measurement x_1, which may be wrong. Observe the notation. Generally the true variable (sometimes called "latent variable") is denoted with a "*" and the observed variable is denoted without the star.

There are many reasons for measurement error. The main distinction is between classical and non-classical measurement error.

- **Classical measurement error:** ______________________________

  ____

  ____

The model with classical measurement error is often known as the ______________

____

In practice, classical measurement error happens when the mismeasurement is unintended, perhaps because of the instruments of measurement. Traditionally it had to do with measuring physical variables using tools which could err in a random way. For example, when measuring weight, a scale may err, even if just a little, because of variations in the environment on that particular day.

In Social Sciences we are not particularly concerned about this type of problem, because the level of precision at which we are making statements in not anywhere near this. However, we can still have classical measurement error for other reasons. One example is when manually transferring data. The person copying the data may make mistakes.

Another example is when we collect data directly from sources which may themselves make mistakes. How many times has it happened to you that someone asked you to check a form to see if all the information was correct? This happens often, for example, at the doctor's office, hospitals, the DMV, Social Security office, etc. How many times was there a wrong piece of information? I am sure it happened to all of you at least once. This unintentional type of error is also fairly random. Well, a researcher could have access to the data stored in that system before you had a chance to check and see that it was wrong. He would be running regressions with the wrong information here and there.

I bet you are expecting that I will say that this kind of error is of no consequence. Well, if that is the case then you are wrong! Unfortunately classical measurement error can cause a great deal of trouble, but not always.

1.1 Classical measurement error in the dependent variable

Suppose that

where e_0 is independent of $x_1^*, \ldots, x_k^*$. Let's suppose that only y has measurement error, so $x_1 = x_1^*, \ldots, x_k = x_k^*$. This way we can focus only on the problem of the dependent variable. The true model is

but unfortunately all you have is the wrong measurement y. Before we go about showing what happens in this case, let's think of an example to make things easier. Let's go back to the case of the effect of smoking on birth weight. Suppose that we may have mistakes in the measurement of birthweight. Then if we substitute the observed birthweight,

The model above is true. The question is whether the error term $u + e_0$ will cause a problem of endogeneity.

The first term may or not be zero because of endogeneity. This is a different issue. The second term, however, could give us problems. Observe:

So, the conclusion is

- ____

  ____
- Does it change the variance? ______________________________

  ____

Generally we would expect the variance to increase, but it is possible that it would decrease, depending on the correlation between u and e_0. If the measurement error is independent of everything in the model, including u, then the variance will increase. Depending on the magnitude of the errors, it could increase a great deal.

1.2 Classical measurement error in the independent variables

Suppose that

where e_1 is independent of $x_1^*, \ldots, x_k^*$. Let's suppose that only x_1 has measurement error, so $x_2 = x_2^*, \ldots, x_k = x_k^*$. This way we can focus only on the problem of one variable. The true model is

but unfortunately all we have is the wrong measurement x_1. So, let's substitute in the equation above

The question is whether the error term $u - \beta_1 e_1$ will cause a problem of endogeneity. Like before:

As before, the first term may or not be zero, but that is a different problem. The issue is that the second term is very likely not zero! Although $x_2 = x_2^*, \ldots, x_k = x_k^*$, the variable $x_1 = x_1^* + e_1$, so

The problem is that although e_1 is independent of $x_1^*, \ldots, x_k^*$, it most definitely is not independent of $x_1^* + e_1$! So, this is in essence a problem of endogeneity, and will cause bias in the coefficients.

The bias of classical measurement error in the independent variables is quite special. Remember the bias formula:

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

where

- $\hat{\theta}_1 \approx \mathbb{E}[w|x_1 = 1, x_2, \ldots, x_k] - \mathbb{E}[w|x_1 = 0, x_2, \ldots, x_k]$
- $\hat{\beta}_w \approx \mathbb{E}[y|x_1, \ldots, x_k, w = 1] - E[y|x_1, \ldots, x_k, w = 0]$

In our case,

- What is w? ____
- What is $\hat{\beta}_w$? ____
- What is $\hat{\theta}_1$? This is harder. It's

$$\frac{Cov(x_1, e_1)}{Var(x_1)} = \frac{\sigma_{e_1}^2}{Var(x_1^*) + \sigma_{e_1}^2}.$$

Knowing where this comes from is not very easy. Those of you that like to know this should think about the univariate case. Look at the formal definition of $\hat{\theta}_1$, and remember that the coefficient of a regression of e_1 onto x_1 is approximately $Cov(x_1, e_1)/Var(x_1) = Cov(x_1^* + e_1, e_1)/Var(x_1^* + e_1) = \sigma_{e_1}^2/(Var(x_1^*) + \sigma_{e_1}^2)$.

So,

since $Var(x_1^*)/(Var(x_1^*) + \sigma_{e_1}^2) < 1$, the estimated $\hat{\beta}_1$ will always be of smaller magnitude than the true β_1. This phenomenon is called ______________________________,

____

which in other words means that when there is classical measurement error in the independent variables, the estimated coefficients are all attenuated, i.e. they are of smaller magnitude than the true values.

We could discuss what happens to the variance of the estimators. The takeaway is that it is hard to know. The denominator increases, because $Var(x_1) = Var(x_1^*) + \sigma_{e_1}^2$. However, $Var(u - \beta_1 e_1)$ could be smaller or larger than $Var(u) = \sigma^2$. Even if e_1 and u are independent, it is still hard to know if the variance increases or decreases.

1.3 Non-classical measurement error

All the other kinds of measurement error are called non-classical. Unfortunately, this kind of error is much more vicious than the classical kind, and very common in behavioral sciences. Among the several possible causes, here are some of the more common:

- Lying (or misreporting). Happens when either the participants in the survey or the data collectors willfully misrepresent the truth. This could be due to shame, guilt, or embarrassment, but it could also be due to political reasons, and self-protection, including privacy protection.

- Wrong recall. Often survey respondents find it hard to recall a particular piece of information. For example, if I asked you: how many classes did you attend this semester in this course? Many of you would not have an easy time remembering. Consider that this is actually an easy question. Many survey questions are much harder. Often, respondents faced with this kind of question tend to round their answers to whole numbers, or the nearest multiple of 5 or 10.

- Careless answers. This is often the case when the survey asks hard questions that require reflection, calculations, or introspection. For example, if you were asked how much do you think the average economics major makes 2 years after graduating, you would not know. In fact, although you have some knowledge about it, and perhaps even know the ballpark, you probably never thought about an actual number. In those circumstances, people tend to report rounded numbers, and they tend to take careless guesses which don't necessarily reflect their beliefs.

- Inadequate questions. This happens, for example, when the survey question can be understood in several ways. Also, it is common for researchers to use a question from a survey which is not exactly the variable they want, but rather an approximation. For example, someone may want to know how many hours people work at home in things like childcare, cooking, cleaning, paying bills, etc., but the survey may ask only about a subset of the activities. A very sneaky occurrence is when a series of the previous questions in the survey influence the respondent's answer to the question. For example, the options in the question may teach or remind the respondent about things he wouldn't otherwise think, let alone mention, on their own. Sometimes the very questions round the possible answers. For example: “Years of education” automatically forces the respondent to round up or down. However, the respondent could have dropped out any time during the year, and thus their true education could be something like 11.35 years.

- Interviewer effects. Some surveys are conducted in person. In such cases it is possible that the interviewers give different inflection to a question, or explain it in different ways. They can also show judgement in their face, or coax a “right” answer from the respondent. Also, sometimes the interviewer answers questions about the respondent, in which case their own biased opinion may be mistaken.

Non-classical measurement error can be prevented, in part, with very careful survey design. The problem is that people don’t usually lie randomly, or round randomly, make recall mistakes randomly, or answer hard questions randomly. People tend to do those things in a very selective way. For example, it is hard to imagine that someone would overreport how much they smoke. Whatever is the cause for the measurement error, this cause is often a confounder in the problem. So, in the problem of maternal smoking, if there is measurement error in smoking,

$$x_1 = x_1^* + e_1,$$

it is unlikely that the error is random. For starters, it will usually be negative. Moreover, who lies? Women that feel guilty, women that know better and feel shame. These are not just any type of women. So, the “omitted variable” e_1 is very related to certain mother characteristics which are confounders in this problem. Non-classical measurement error is, at its very essence, a problem of endogeneity.

The same argument goes for people that round. Who rounds how much they smoke? Who has trouble recalling how much they smoke? Who answers this carelessly? The answer to all those questions is not just any random woman. No, they are particular types of women. Whatever their common characteristic, it is likely a confounder in the problem.

So, when thinking about non-classical measurement error, consider the equations for classical measurement error, but then think that there is a true endogeneity problem brought by the “omitted variable” e_1, which is indeed correlated to the explanatory variables because of the factors that induce the error. The reasons for lying, not remembering, rounding, being careless, and misunderstanding a question are often very related to some important confounders of the problem.

1.4 Solving measurement error

Bad news. It is very, very hard to solve measurement error. Here are some recommendations:

- When you decide to use a particular data set, check the most important variables. How was the question phrased in the survey, how was the survey collected. Do you anticipate a measurement error problem? Sometimes it helps to plot the frequencies. For example, there are a disproportional number of women that smoke 5,10,15, and 20 cigarettes per day in comparison to those that smoke non-rounded numbers. It's hard to know if it is measurement error, because people could truly smoke proportions of a pack (so they are rounding consumption, not their responses), but this deserves investigation.

- For classical measurement error, the best solution is to find two measurements of the same variable. This is not very common in behavioral sciences data sets, but sometimes it is possible. Here is what you do. Suppose that there are two measurements of x_1^*:

  $$x_1 = x_1^* + e_1,$$
  $$x_1' = x_1^* + e_1',$$

  It is very important that the two measurements be independent, so that e_1 and e_1' are independent. Then, we can correct for the attenuation bias. Observe that

  $$Cov(x_1, x_1') = Var(x_1^*)$$

  and

  $$Var(x_1) = Var(x_1^*) + \sigma_{e1}^2$$

  so

  $$\frac{\widehat{Var}(x_1)}{\widehat{Cov}(x_1, x_1')}\hat{\beta}_1 \approx \frac{Var(x_1^*) + \sigma_{e1}^2}{Var(x_1^*)}\beta_1 \frac{Var(x_1^*)}{Var(x_1^*) + \sigma_{e1}^2} = \beta_1.$$

  Although this solution is extremely elegant, it is seldom the case that it can be implemented in behavioral data sets.

- Perhaps the best solution for measurement error is the use of instrumental variables. This is beyond the scope of this course, but you should know to look for this if you have the need. Search for "IV solution error-in-variables model." IV methods solve endogeneity problems, and since measurement error is, at the essence, a problem of endogeneity, it is natural that it should be solved the same way. However, don't be deceived. Instrumental variables can solve endogeneity (and therefore measurement error) problems, but they are not easy to find, and they have problems of their own.

- In your own work, when you cannot solve a problem of measurement error, make sure to acknowledge the possibility, explore its nature, and examine its likely consequences in your case. Do not waive hands just because others have also done so.