1. Proxy Variables

Recall the Sisyphean task of satisfying MLR4. We can never hope to observe everything that affects y and is correlated with x, unless x is as good as randomly assigned (much more on this after the midterm). For example, in our oft-cited wage equation example,

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 ability_i + e_i \tag{1}$$

we are never going to be able to observe, let alone accurately measure, "true" ability. We would end up having to estimate

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + u_i \tag{2}$$

where $u_i = \beta_3 ability_i + v_i$. If ability is correlated with education or experience (as we suspect), then $\mathbb{E}[u_i|educ_i, exper_i] \neq 0$ and MLR 4 fails.

While we can't directly observe ability, we might think we can observe something that comes reasonably close, such as an IQ test or an SAT score. We call these **proxy variables**, since we might include them in our model as proxies for the variable we want. We could run something like

$$\log(wage_i) = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3^* IQ_i + u_i \tag{3}$$

where the asterisk on β_3 implies that we won't necessarily get the same effect β_3 we would if we could include *ability* itself.

If we think IQ scores are related to true ability, we could model

$$ability_i = \delta_0 + \delta_1 IQ_i + v_i \tag{4}$$

To see if including this proxy variable this helps us satisfy MLR 4, let's substitute (4) into (1)

$$\begin{aligned}
\log(wage_i) &= \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 ability_i + e_i \\
&= \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3(\delta_0 + \delta_1 IQ_i + v_i) + e_i && \text{by (4)} \\
&= \underbrace{\beta_0 + \beta_3\delta_0}_{\text{constant}} + \beta_1 educ_i + \beta_2 exper_i + \beta_3\delta_1 IQ_i + \underbrace{\beta_3 v_i + e_i}_{\text{new error term}} && \text{regrouping}
\end{aligned} \tag{5}$$

So we get an unbiased estimate of the effect ability (β_3) scaled by the effect of IQ on ability δ_1 (which is the best we can hope for with a proxy) *if* the new error term $\beta_3 v_i + e_i$ from (5) is uncorrelated with $educ$, $exper$, and IQ. Formally, MLR 4 now becomes $\mathbb{E}[\beta_3 v_i + e_i|educ, exper, IQ] = 0$. Since we have assumed 1 is the "true" model, we don't need to worry about the structural error e_i being correlated with anything. Likewise, v_i is independent of IQ_i by construction. So what MLR 4 boils down to with a proxy is

$$\beta_3 \mathbb{E}[v_i|educ_i, exper_i] = 0$$

If this holds, meaning the part of IQ that is uncorrelated with ability is also uncorrelated with education and experience, we have a good proxy for ability, and will be able to validly estimate (3). Is this a reasonable assumption though? Probably not. It's saying that whatever IQ measures miss about your true ability is uncorrelated with how much education and experience you end up getting. For example, if IQ tests are differentially accurate for people of different backgrounds,

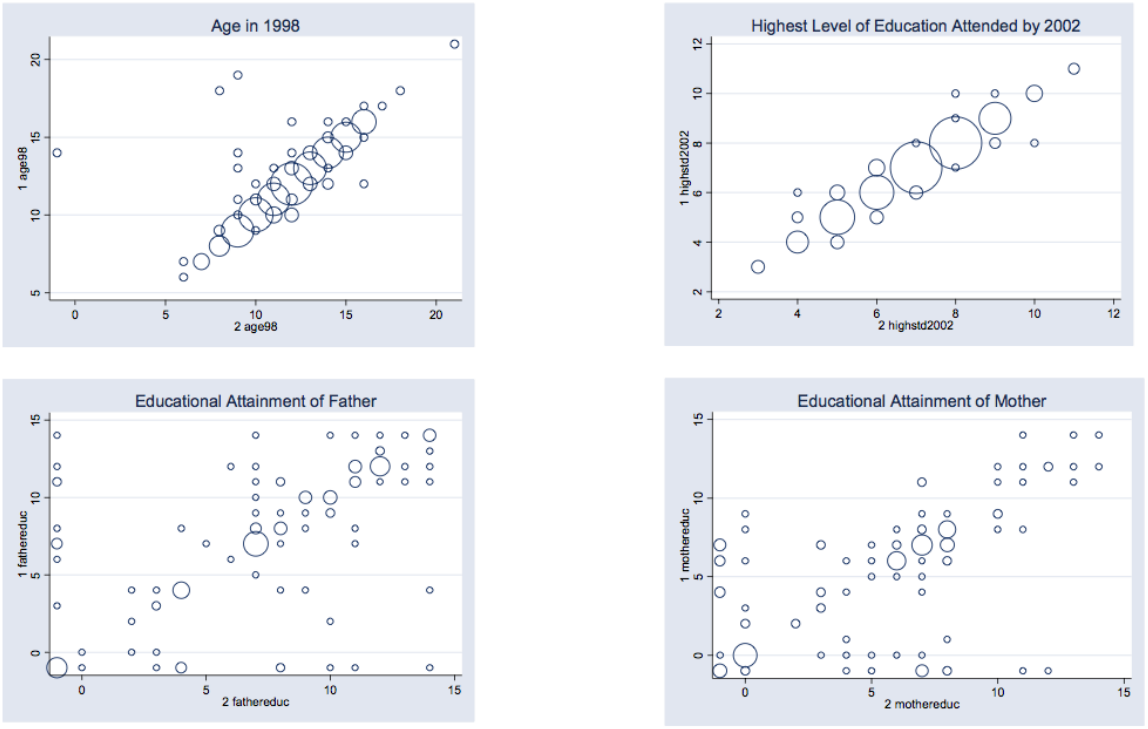
then they are a poor proxy for ability *and* will not give us unbiased estimates of the return to education or experience. The same might be true for using standardized test scores as a proxy for ability.[1]

2. Measurement error more generally

We can think of proxies more generally as examples of variables that are measured with error. Moreover, even things that we do observe are often going to be measured with error. Below are a series of figures with data from people who were surveyed twice within 1 week during the Kenya Life Panel Survey (KLPS).[2] For each of the four time-invariant (they should not change over time, at least not within a week) variables shown below, the first measurement is shown on the y axis and the second measurement of the same variable a week later is shown on the x axis.

If variables are measured perfectly then all observations should fall on the 45-degree line (as the prior measurement should equal the follow-up measurement). However, this is clearly not the case, especially when looking at parents' education. This is just an illustration to show how likely mismeasurement is to occur even in really simple contexts.

Figure 3: Reliability of survey data



Notes: These figures plot survey values against resurvey values. Points are weighted to denote number of observations included. A value of "-1" denotes a response of "don't know". Responses with impossible values are excluded.

[1] Hot takes on UCB abolishing the SAT and GRE available in office hours upon request.

[2] KLPS is a unique panel data set that has tracked a variety of outcomes over four rounds of data collection for individuals who participated in the randomized primary school deworming intervention in Miguel and Kremer (2004).

In this course, we will focus on **classical measurement error**, which essentially means variables are measured with random noise (more formal definition below). How badly does measurement error affect our estimates? It turns out to depend on whether it is the dependent or independent variable(s) that are measured with error – classical measurement error in y is not too bad but as we will show, even classical measurement error in x will lead to **attenuation bias**, or coefficients that are biased toward 0.

A. Measurement Error in the Dependent Variable

Suppose the true model is

$$y_i^* = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki} + u_i \tag{6}$$

but we measure y_i^* with error. We only observe $y_i = y_i^* + e_i$, where e_i is "white noise", or random error uncorrelated with y_i^*. When we run the regression using our observed variables we get:

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \ldots + \beta_k x_{ki} + v_i && (7)\\ y_i^* + e_i &= \beta_0 + \beta_1 x_i + \ldots + \beta_k x_{ki} + v_i && \text{subsituting for } y\\ u_i + e_i &= v_i && \text{substituting (6)} \end{aligned}$$

MLR 4 for 7 then becomes

$$\mathbb{E}[u_i + e_i | x_{1i} \ldots x_{ki}] = \mathbb{E}[u_i | x_{1i}, \ldots, x_{ki}] + \mathbb{E}[e_i | x_{1i}, \ldots, x_{ki}] = 0$$

So this holds when:

1. MLR 4 holds in the "true" (perfectly measured) model ($\mathbb{E}[u_i | x_{1i}, \ldots, x_{ki}] = 0$) **AND**
2. The measurement error in y_i is uncorrelated with the x_i's ($\mathbb{E}[e_i | x_{1i}, \ldots, x_{ki}] = 0$).

With classical measurement error or white noise, 2. holds and measurement error in y_i doesn't lead to bias. The only issue in this case is that our estimates will be less precise. This is because we have both u_i and e_i in the error term, which means it has a larger variance and our standard errors are proportional to $\sqrt{\sigma_u^2 + \sigma_e^2} > \sigma_u$.[3]

B. Measurement Error in Independent Variables

Let's focus on the SLR case for intuition. Suppose SLR 1-4 hold, but we mismeasure x_i^* in the model

$$y_i = \beta_0 + \beta_1 x_i^* + u_i. \tag{8}$$

Instead we observe $x_i = x_i^* + e_i$, so when we run the regression with x instead of x^*, we get

$$\begin{aligned} y_i &= \beta_0 + \beta_1 (x_i - e_i) + u_i \\ y_i &= \beta_0 + \beta_1 x_i + \underbrace{u_i - \beta_1 e_i}_{\text{new error term}} \end{aligned}$$

[3] With non-classical measurement error, 2. is unlikely to hold leading to bias in our estimates.

SLR 4 is now

$$\mathbb{E}[u - \beta_1 e|x] = \mathbb{E}[u|x] - \beta_1 \mathbb{E}[e|x] = 0 \tag{9}$$

We don't need to worry much about the first part of (9) since it's innocuous to assume $\mathbb{E}[u|x] = \mathbb{E}[u|x^*] - \mathbb{E}[u|e] = 0$. This is because we've assumed SLR 4 holds for the model with x^*, and it's unlikely that the measurement error e is correlated with the structural error u.

It's the second part of (9) that we may need to worry about. We'll deal with two separate cases, though the distinction between them is subtle. The first is if $cov(e, x) = 0$ – i.e., the measurement error is uncorrelated with the measured x – and the second is if $cov(e, x^*) = 0$ – i.e., the measurement error is uncorrelated with the true x.[4] Note that these cases can't occur simultaneously. If $cov(e, x) = 0$, then $cov(e, x^* + e) = cov(e, x^*) + var(e) = 0$. But since $var(e) > 0$, $cov(e, x^*) \neq 0$. Of course, it may be that neither covariance is 0, but we won't explore the implications of this particular case.

What are the implications for our regression estimates under these two cases?

Case 1: $cov(e, x) = 0$

Recall from (9) SLR 4 requires e to be uncorrelated with x (under the palatable assumption that $\mathbb{E}[u|x] = 0$). This holds by assumption in this case. SLR 4 is satisfied because our error term is uncorrelated with the x we use in our regression. Like in the case for error in y, we end up with unbiased, albeit noisier (meaning with larger standard errors), estimates. This case is often seen as less likely to hold.

Case 2: $cov(e, x^*) = 0$

This is **classical measurement error**, which considered the likelier of the two cases and is the more problematic one in terms of regression estimates. We just argued that $cov(e, x^*) = 0$ implies $cov(e, x) \neq 0$, so this means that we can think of our measurement error as an omitted variable that is (negatively) correlated with our dependent variable. We therefore have

$y = \beta_0 + \beta_1 x_i - \beta_1 e_i + u_i$	the true population model
$y = \hat{\beta}_0 + \hat{\beta}_1 x_i$	what we estimate in a regression
$e_i = \delta_0 + \delta_1 x_i + v_i$	(because $cov(e, x) \neq 0$)

[4] If it helps, think about the where x is farm size in acres. If people with large *reported* farm sizes are just as likely to be making mistakes in their reported size as people with small reported farm sizes, then it has to be that people with truly small farms are making bigger mistakes than people with truly large farms in order for reported farm size to be similar for both groups. We have $x - x^* = e$, so if e doesn't vary with x it follows that x^* must move inversely with e. In the other case, if people are making the same types of mistakes regardless of *actual* farm size, the people with bigger reported farms are generally going to be those making bigger mistakes. Again with $x - x^* = e$, if e doesn't vary with x^* it follows that x must move together with e.

Recalling the OVB formula from earlier in the course we have

$$
\begin{aligned}
E(\hat{\beta}_1) =& \beta_1 + (-\beta_1)\delta_1 \\
E(\hat{\beta}_1) =& \beta_1 + (-\beta_1)\frac{Cov(x,e)}{Var(x)} \\
=& \beta_1 - \beta_1 \frac{Cov(x^* + e, e)}{Var(x^* + e)} \\
=& \beta_1 - \beta_1 \left(\frac{Cov(x^*, e) + Var(e)}{Var(x^*) + Var(e) + 2Cov(x^*, e)} \right) && \text{(expression for variance of a sum)} \\
=& \beta_1 - \beta_1 \frac{Var(e)}{Var(x^*) + Var(e)} && \text{(since } Cov(x^*, e) = 0\text{)} \\
=& \beta_1 \left(1 - \frac{Var(e)}{Var(x^*) + Var(e)} \right) \\
=& \beta_1 \left(\frac{Var(x^*)}{Var(x^*) + Var(e)} \right).
\end{aligned}
$$

Since $0 < \frac{Var(x^*)}{Var(x^*) + Var(e)} < 1$, we are scaling down the true β_1 towards zero. This is what we call **attenuation bias**.

C. Summary

Even if you didn't follow the math, what you should take away is the following.

- Classical measurement error in y will not produce biased estimates, but will reduce precision (i.e., increase standard errors).
- Classical measurement error in x will produced estimates that are *biased towards zero* (attenuated).
- A special case is when measurement error in x is independent of the measured value (even though it's correlated with the true value). In this case, we also get unbiased but noisier estimates (larger standard errors).
- We haven't covered what happens with multiple x variables if some are mismeasured. Analogous to the OVB case with MLR, generally all estimates will be biased if one variable is mismeasured (unless the other x variables are independent from the mismeasured x).
- Quite possible that the error is correlated with both measured and true x. Lots of other types of non-classical measurement error as well – all bets are off in these cases.