1 Introduction

As researchers, we are typically interested in estimating the causal effect of some variable X on a dependent variable Y. So far in this class, we have used multiple regressions to study the relationship between different variables or phenomena using data. This section notes focus on the reliability of these regression models, and discuss the conditions under which the estimates from our regression models might fail to provide us with useful estimates of causal effects. We focus on two key concepts in this discussion: (i) **internal validity**, which relates to whether the effects estimated by a model based on a sample are valid for the population and the context of interest, and (ii) **external validity**, which relates to whether effects estimated by a model can be generalized to other populations and contexts.

2 Threats to Internal Validity

2.1 Problems

Let's assume that we are interested in the study of the causal impact of X on Y.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

This study is said to be internally valid if its estimated regression coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased and consistent for β_0 and β_1. We discuss here five cases in which the regression coefficients might be biased or inconsistent.

1. **Omitted Variables Bias** ($Z \Leftrightarrow X$ and $Z \Rightarrow Y$) occurs when there is some unobserved variable Z that is correlated with both our independent variable X and our dependent variable Y. Suppose we want to look at the effect of education spending on GDP at the country level. Here, GDP would be our Y variable and education spending our X variable. Suppose that there is a third variable Z that corresponds to health expenditures and that improving health may cause countries to both increase their income and spend more on education. This would mean that we would observe both large increases in education spending and large increases in income when there is improving health, even if there is no causal relationship between education spending and income.

2. **Wrong Functional Form**: If the specified functional form of the sample regression is different from the functional of the true population regression function, then our estimated model will be biased. A wrong functional form will most likely arise if the true population regression function is not linear. The misspecification can be interpreted as a form of omitted variable bias in which nonlinear terms of the true population regression function are the omitted variables in the sample regression. For example, in the example above, if it is the case that the quadratic polynomial of educational expenditures is an independent variable in the true population regression function, then this quadratic polynomial term would be a Z variable in the omitted variable framework presented above.

3. **Measurement Error**: We may be in a situation in which our X and Y variables are also measured imprecisely due to the variables being too abstract to measure (e.g., ability), or too sensitive to measure or ask about (e.g., income of individuals) or simply too hard to measure (e.g. yields of smallholder farmers). In these cases, the researcher might have to settle with less accurate measures of the variables needed. In our previous example, if education expenditures are measured imprecisely with errors, then the error term in our population regression will be correlated with the difference between the true education expenditures and these imprecise measured values of education expenditures. If this difference is uncorrelated with the true education expenditures and is correlated with the imprecise measured values, then our OLS estimator will be inconsistent and will tend towards 0, regardless of sample size. In general, (under the classic measurement error model), measurements errors in our X variable will produce estimates that are *biased towards zero* (attenuated) while measurement in Y will not produce biased estimates, but will reduce precision (i.e., increase standard errors). More details on this point in Section 2.3 of these notes.

4. **Sample Selection Bias** occurs when the sample for which we collect data might not be representative of the population we are interested in studying. For example, we might be interested in studying the relationship between GDP and education expenditures for all countries in the world but just have data on these two variables for a subset of these countries. We specifically analyse three cases of missing data.
   (a) If the reasons for data availability on income and education expenditures is unrelated to the levels of income or to education expenditures, then our estimated will be based on a reduced sample size but they will not be biased.
   (b) If the data is only missing for our X variable on education expenditure based on a threshold in this variable, then our estimate will not be biased for the set of countries which satisfies this threshold. For example, if we have education expenditure data for countries that spend at least 100 billion US dollars on education, then our estimates will be unbiased for this group of countries.
   (c) If the reason for our missing data in the X variable on education expenditure is related to our Y variable on income, then our estimate will be biased. For example, if it is the case that poorer countries, which have lower GDP levels, are also more likely to have missing data on education expenditures relative to richer countries with higher GDP levels, then this would be a threat to the internal validity of model.
5. **Simultaneous Causality Bias** ($Y \Rightarrow X$ and $X \Rightarrow Y$) occurs when the outcome variable causes selection into the treatment (X) and control, but treatment status also has effects on the outcome. For our purposes, this is a special case of reverse causality. In the example above, country income might lead to more investment in education, which might simultaneously increase country income. **Reverse causality** ($Y \Rightarrow X$), which means that the outcome variable (Y) actually affects the realization of X, is a special case of simultaneity bias. In the previous example, we might have reverse causality if countries with larger increases in GDP decide to increase their spending on education, because they can afford it. So it would not be that education causes growth, but that growth leads to more expenditure in education.

In all these five cases, the bias is due to the fact that the regressor X is correlated with the error term of the regression ϵ, violating the zero conditional mean assumption of linear regression models.

2.2 Solutions

Now that we have laid out the potential threats to internal validity, we turn to ways in which we can mitigate them:

1. **Omitted Variables Bias**: In order to correct an omitted variable bias, we should check whether or not the omitted variable is observed and whether there are adequate control variables for this omitted variable:
   (a) If the variable itself or adequate control exists, then we can directly add them to a regression. Here, we have to bear in mind the fact that adding variables to a regressions might make the coefficients of the model less precise, depending on the correlation between the existing regressors and the newly added regressors.
   (b) If the variable does not exist or no adequate control is available, then we can turn to panel data models (which can control for omitted variables that do not change over time) or to instrumental variables regressions or to randomized controlled trials.[1]
2. **Wrong Functional Form**: The solutions to account for non-linearities in our fonctional forms depend on the type of our dependent variables, and in particular whether they are discrete or not and whether they are binary or non-binary:
   (a) If our dependent variable is continuous (as is the case in the example above with GDP), then we can turn to polynomials, logarithmic transformations, and interactions terms as relevant.

[1] There will be more details in these alternative approaches in future handouts and sections.

As specified in previous handouts and section notes, the exact form that the added variables should take depends on the underlying known relationships between our dependent and our independent variables.

(b) If our dependent variable is discrete, then we can turn to linear probability models, or logit/probit models.[2]

3. **Measurement Error**: The most direct way to correct for measurement errors is simply to get accurate measures of our independent variable X whenever feasible. Otherwise, if the researcher possesses knowledge about the source and the nature of the measurement error, then they can model it to correct for the attenuation bias. Alternatively, the researcher can also turn to instrumental variables, which will be discussed with more details in later sections.

4. **Sample Selection Bias**: There are models that the researcher can use to model the selection process that led to the sample selection bias and to use this model to impute values for units with missing values. These alternative approaches are beyond the scope of this class.

5. **Simultaneous Causality Bias**: Randomized controlled trials or instrumental variables approaches, which will be covered in futures sections, are methods that the research can use to deal with simultaneous causality bias.

2.2.1 True/False Questions

Please respond with True or False to the following statements:

1. A larger sample size poses a threat to internal validity in linear regression models.

   False. A larger sample size typically improves the precision of parameter estimates and reduces the risk of sampling error. It does not pose a threat to internal validity.

2. Reverse causality is the opposite of simultaneous causality bias.

   False. Reverse causality is that the direction of causality runs from Y to X. It is a special case of simultaneous causality bias, which occurs when causality runs in both ways.

2.2.2 Long Question

Labor economists studying the determinants of women's earnings discovered a puzzling empirical result. Using randomly selected employed women, they regressed earnings on the women's number of children and a set of control variables (age, education, occupation, and so forth). They found that women with more children had higher wages, controlling for these other factors. Explain how sample selection might be the cause of this result. (Hint: Notice that women who do not work outside the home are missing from the sample.)

The key is that the selected sample contains only employed women. Consider two women, Beth and Julie. Beth has no children; Julie has one child. Beth and Julie are otherwise identical. Both can earn \$25,000 per year in the labor market. Each must compare the \$25,000 benefit to the costs of working. For Beth, the cost of working is forgone leisure. For Julie, it is forgone leisure and the costs (pecuniary and other) of child care. If Beth is just on the margin between working in the labor market or not, then Julie, who has a higher opportunity cost, will decide not to work in the labor market. Instead, Julie will work in "home production," caring for children, and so forth. Thus, on average, women with children who decide to work are women who earn higher wages in the labor market than.

2.3 Measurement Error

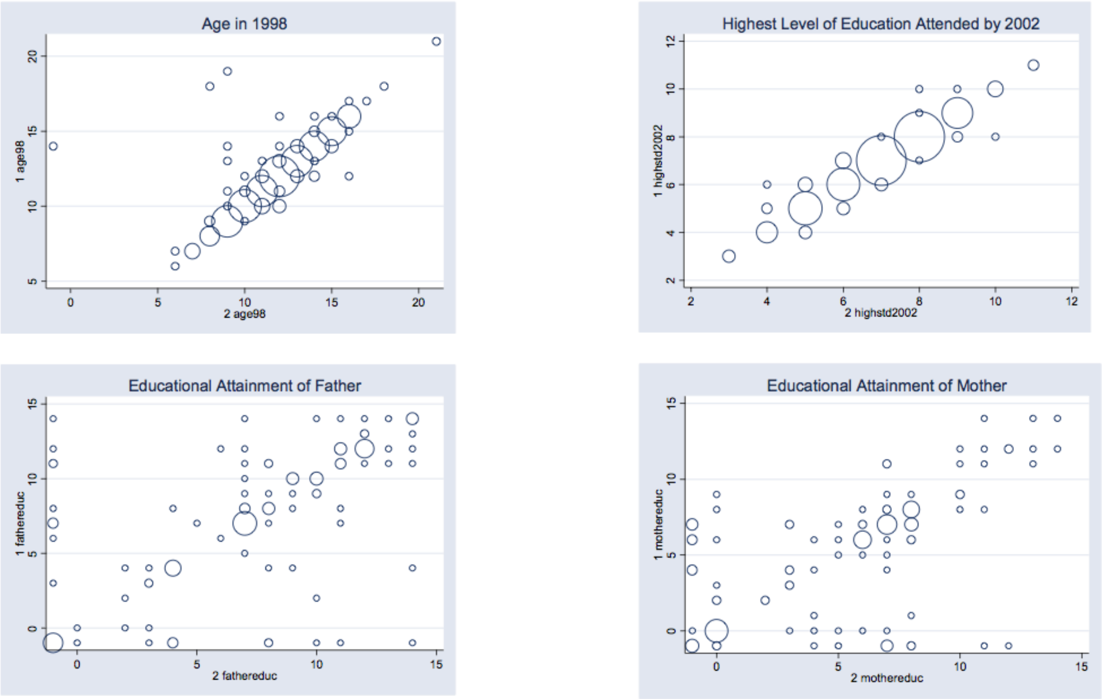
In this section, we zoom in on the issue of measurement error introduced above in Section 2.1 above.

[2] There will be more details in these alternative approaches in future handouts and sections.

In general, 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).[3] For each of the four time-invariant variables on age and education (they should not change over a short period of time, at least not within a week), 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.

Here, we 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.

2.3.1 Measurement Error in the Dependent Variable

Suppose the true model is the following:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_k X_{ki} + u_i \qquad (2)$$

[3] 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).

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. When we run the regression using our observed variables, we get:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 X_i + \ldots + \beta_k X_{ki} + v_i && && (3) \\ Y + e_i &= \beta_0 + \beta_1 X_i + \ldots + \beta_k X_{ki} + v_i && \text{subsituting } Y_i^* \text{ with } Y_i + e_i \\ u_i + e_i &= v_i && \text{substituting } Y_i \text{ with (2) in the previous line} \end{aligned}$$

The zero conditional mean assumption for (3) then becomes

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

This zero conditional mean assumption holds:

1. In the "true" (perfectly measured) model ($\mathbb{E}[u_i | X_{1i}, \ldots, X_{ki}] = 0$) **AND**
2. When 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 $\sqrt{\sigma_u^2 + \sigma_e^2} > \sigma_u$.[4]

2.3.2 Measurement Error in Independent Variables

Suppose that we mismeasure X_i in the model

$$Y_i = \beta_0 + \beta_1 X_i + u_i. \tag{4}$$

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

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 (X_i^* - e_i) + u_i \\ Y_i &= \beta_0 + \beta_1 X_i^* + \underbrace{u_i - \beta_1 e_i}_{\text{new error term}} \end{aligned}$$

Our zero conditional mean assumption is now:

$$\mathbb{E}[u - \beta_1 e | X^*] = \mathbb{E}[u | X^*] - \beta_1 \mathbb{E}[e | X^*] = 0 \tag{5}$$

We don't need to worry much about the first part of (5) since it's innocuous to assume $\mathbb{E}[u|X^*] = \mathbb{E}[\mathbb{E}[u|X, e]|X^*] = \mathbb{E}[0|X^*] = 0$. This is because we've assumed the zero conditional mean assumption holds for the model with X, and it is unlikely that the measurement error e is correlated with the structural error u.

It is the second part of (5) that we may need to worry about. We'll deal with two separate cases, though the distinction between them is subtle. The first case is if $Cov(e, X^*) = 0$ – i.e., the measurement error is uncorrelated with the measured X^* – and the second case is if $Cov(e, X) = 0$ – i.e., the measurement error is uncorrelated with the true X.[5] 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?

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

[5] If it helps, think about a case in which 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 the case that people with truly small farms are making bigger mistakes than people with truly large farms in order for the error in 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.

Case 1: $Cov(e, X^*) = 0$

The zero conditional mean assumption in (5) requires e to be uncorrelated with X^* (under the palatable assumption that $\mathbb{E}[u|X^*] = 0$). This holds by assumption in this case. The zero conditional mean assumption is satisfied because our error term is uncorrelated with the X^* we use in our regression. Like in the case for measurement errors 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 the **classical measurement error**, which consider 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:

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_i^* - \beta_1 e_i + u_i && \text{the true population model} \\ Y &= \delta_0 + \delta_1 X_i^* + \tilde{u}_i && \text{what we estimate in a regression} \\ cov(e, X^*) &\neq 0 \end{aligned}$$

Recalling the OVB formula from earlier in the course we have

$$\begin{aligned} \hat{\delta_1} \xrightarrow[n\to\infty]{p} \delta_1 =& \beta_1 + (-\beta_1)\rho_{X^*e}\sqrt{\frac{Var(e)}{Var(X^*)}} \\ =& \beta_1 + (-\beta_1)\frac{Cov(X^*, e)}{Var(X^*)} \\ =& \beta_1 - \beta_1 \frac{Cov(X+e, e)}{Var(X+e)} \\ =& \beta_1 - \beta_1 \left(\frac{Cov(X,e) + Var(e)}{Var(X) + Var(e) + 2Cov(X,e)}\right) && \text{(expression for variance of a sum)} \\ =& \beta_1 - \beta_1 \frac{Var(e)}{Var(X) + Var(e)} && \text{(since } Cov(X,e) = 0\text{)} \\ =& \beta_1 \left(1 - \frac{Var(e)}{Var(X) + Var(e)}\right) \\ =& \beta_1 \left(\frac{Var(X)}{Var(X) + Var(e)}\right). \end{aligned}$$

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

2.3.3 True / False Questions

1. Classical measurement errors in independent variables always lead to underestimated regression coefficients.

   False. Classical measurement errors in independent variables can also lead to overestimated regression coefficients if the coefficient is negative.

2. Measurement errors in dependent variables affect the precision (standard errors) of the estimated regression coefficients.

   True. See notes above.

3. If both the independent variable and dependent variable have classical measurement errors, they cancel each other out, resulting in unbiased estimates.

False. Measurement errors in both variables can compound the bias and lead to more substantial estimation errors.

4. Measurement errors in the independent variable impact the slope of the regression line but not the intercept.

   False. Measurement errors in the independent variable can affect both the slope and intercept of the regression line.

3 Threats to External Validity

In this section, we discuss the potential threats to the external validity of regression models. When it comes to external validity, the two factors that matter are the population and the context.

1. **Populations**: If there are significant differences between the population studied and the populations we (or policymakers) are interested in, then the results from our regression models from the population studied might not be generalizable to the population of interest, i.e. our results might not apply to the population of interest.
2. **Contexts**: Similarly, if there are significant differences between the context studied and the context we (or policymakers) are interested in, then the results from our regression models from the context studied might not be generalizable to the context of interest. Differences in contexts include differences in institutions, regulatory frameworks, or physical environments.

3.1 True/False Questions

1. As long as an econometric study is internally valid, it will be externally valid.

   False. A study can be internally valid without being externally valid. The factors that are considered for external and internal validities are different from one another.

2. Convenience sampling (where you sample the data that is most easy to access) is a potential threat to the external validity of a linear regression study.

   True. Convenience sampling may introduce a threat to external validity because the sample may not be representative of the broader population, affecting generalizability.

3. Selection bias is not a potential threat to external validity. It is only a threat to internal validity.

   False. Selection bias, often associated with non-random samples or self-selection, can threaten external validity when the sample does not represent the population of interest.

4. External validity considers whether the findings of a study have been validated and certified by an impartial third party audit.

   False. External validity simply involves assessing the applicability of study findings to other time periods, locations, or contexts, addressing the broader generalizability of results. it does not require the oversight of audit entities.

3.2 Long Questions

1. Suppose you have just read a careful statistical study of the effect of advertising on the demand for cigarettes. Using data from New York during the 1970s, the study concluded that advertising on buses and subways was more effective than print advertising. Use the concept of external validity to determine if these results are likely to apply to Boston in the 1970s, Los Angeles in the 1970s, and New York in 2018.
   Potential threats to external validity arise from differences between the population and setting studied and the population and setting of interest. The statistical results based on New York in the 1970's are likely to apply to Boston in the 1970's but not to Los Angeles in the 1970's. In 1970, New York and Boston had large and widely used public transportation systems. Attitudes

about smoking were roughly the same in New York and Boston in the 1970s. In contrast, Los Angeles had a considerably smaller public transportation system in 1970. Most residents of Los Angeles relied on their cars to commute to work, school, and so forth. The results from New York in the 1970's are unlikely to apply to New York in 2018. Attitudes towards smoking changed significantly from 1970 to 2018.