9.4 Properties of OLS under Measurement Error

1. Contrasting the case of omitted variable with measurement error: In the measurement error case, the variable that we do not observe has a well-defined, quantitative meaning (such as a marginal tax rate or annual income), but our recorded measures of it may contain error.

   In certain cases we could also calculate the size of the error asymptotically.

   - ▷ Thus econometrician has to use annual income of individuals to explain the dependent variable, but she does not have the data on annual incomes. She goes and asks people their annual incomes. Because people do not report their true income, the data collected is only an approximation of the actual annual incomes.

2. In measurement error problem, the mis-measured variable is of interest, whereas in the case of OVB, we might just be interested in its effect on the other key variables of interest.

9.4.1 Measurement Error in the Dependent Variable

1. Consider the following model in which we would ideally like to have data on y^*, but have an observable measure of y. For instance it is hard to obtain data on savings i.e y^* instead families reports a measure of savings y (self-reported measure):

$$y^* = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$$

2. Measurement error is defined as the difference between the observed value and the actual value

$$e_0 = y^* - y$$

If we substitute y^* in the equation with measurement error, we get a new composite error of the following form:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u + e_0$$

If e_0 has zero mean, then we get unbiased estimates. Even if e_0 has a non-zero mean, we do not worry about the bias as much.

3. Multiplicative measurement error: what if we have a measurement error of the following form: $y = y^* a_0$. In such a scenario, we take log on both sides and obtain $log(y) = log(y^*) + log(a_0)$, where $e_0 = log(a_0)$.

4. Conditional independence of the measurement error: If e_0 and x_j are statistically independent, then our estimates are consistent and unbiased. Our usual OLS inference procedures apply.

9.4.2 Measurement Error in the Explanatory Variable

1. Consider the simple regression model in which we have the desired independent variable which we want to explain the dependent variable:

$$y = \beta_0 + \beta_1 x_1^* + u$$

But instead of x_1^* we have a variable x_1 which is ridden with measurement error. We assume that the average measurement error is zero, which is not crucial. We assume more importantly that $E(y|x_1^*, x_1) = E(y|x_1^*)$.

2. Measurement error and independent variable: If we have the observed variable uncorrelated with the measurement error i.e.

$$Cov(x_1, e_1) = 0$$

we could simply plug in the observed variable and run the regression without affecting the properties of the OLS estimator. We estimate the following equation:

$$y = \beta_0 + \beta_1 x_1 + (u + \beta_1 e_1)$$

3. The estimates are consistent. Error variance is

$$Var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e_1}^2$$

4. If instead the covariance between the unobserved variable with the measurement error takes the following form:

$$Cov(x_1^*, e_1) = 0$$

then the assumption is refered to as the *Classical error in variable* assumption. However, if this holds, then the earlier assumption cannot be true i.e. $Cov(x_1, e_1) \neq 0$. Instead the correlation between the two is

$$Cov(x_1, e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1^2) = 0 + \sigma_{e_1}^2$$

5. This implies that our regression in point (2) does not satisfy the CLRM assumptions. The covariance between the x_1 and the composite error $u - \beta_1 e_1$ is

$$Cov(x_1, u - \beta_1 e_1) = -\beta_1 Cov(x_1, e_1) = -\beta_1 \sigma_{e_1}^2$$

Thus the OLS regression of y on x_1 gives a biased and inconsistent estimator.This implies the following:

$$plim(\hat{\beta}_1) = \beta_1 \overbrace{\left(\frac{\sigma_{x_1}^{2*}}{\sigma_{x_1}^{2*} + \sigma_{e_1}^2} \right)}^{<1}$$

This implies that for a positive β_1, under CEV assumptions, the estimated $\hat{\beta}_1$ will be smaller than its population counterpart, β_1. This is called the "attenuation bias" in OLS.

6. In a multiple regression model with measurement error for x_1^*, under the CEV assumptions, OLS will be biased and inconsistent. However, in this case all OLS estimatoes will be biased and not just the ones associated with x_1 - the measure of x_1^* in the sample for the following model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u - \beta_1 e_1$$

The attenuation bias can now be obtained as:

$$plim(\hat{\beta}_1) = \beta_1 \left(\frac{\sigma_{r_1}^{2*}}{\sigma_{r_1}^{2*} + \sigma_{e_1}^2} \right)$$

where r_1^* is the population error in the equation $x_1^* = \alpha_0 + \alpha_2 x_2 + \alpha_3 x_3 + r_1^*$.

We cannot say anything apriori and clearly about the biases in estimating β_2 and β_3 in this model.

7. Summing up: Generally the CEV assumption is still a strong assumption. It is more likely that the measurement error, e_1 is likely to be correlated with both x_1 and x_1^*.

   (a) IV estimation could provide us consistent estimators even in the presence of a general measurement error problem.