Violations of Gauss Markov Assumptions: Measurement Error

Econometric Methods, ECON 370

1 OLS under Measurement Error

We have found out that another source of endogeneity is derived from measurement error. This arises principally because we are not able to fully observe all variables all the time. We might attempt to create that variable through calculations using other variables. But these are ultimately imprecise measures of the variable we are trying to get at. As you should have realize by now, the structure of the problem created by measurement error is very similar to that obtained from omitted variable bias.

We will consider two types of measurement error, the first on the dependent variable, and the other from measurement error in independent variables.

1.1 Measurement Error in the Dependent Variable

Consider the case where the dependent variable is measured with error, y, which was essentially attempting to measure some true variable y^*. Let the error be $y - y^* = e$. Suppose what we were really interested in was,

$$y^* = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon$$

$$\Rightarrow y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + e + \epsilon$$

Notice that the error term is now $\epsilon + e$. This means that if we estimate the model, we would see the error in measurement appearing in the new error term. You can imagine that whether there is a violation of the assumption that the conditional expectation of the error should be zero is largely dependent on whether the error in measurement is correlated with the independent variables. Naturally by the OLS assumptions ϵ is uncorrelated with the covariates and consequently will not present any problems. What about e? It is natural to think of it as having on average zero mean, but even if it does not, all that is the estimate of the intercept β_0 which is hardly ever a concern. Typically, it will be assumed that the measurement error, e is uncorrelated with the covariates, which in turn implies that the OLS estimators will be unbiased and consistent. However, if you look carefully at the new error

term you would realize that the problem comes from inference. To see that,

$$var(\epsilon + e) = \sigma_\epsilon^2 + \sigma_e^2 > \sigma_\epsilon^2$$

assuming of course the error terms themselves are uncorrelated. **Ask yourself what it means if the they are?**. The last inequality means that the estimated variance is now larger than before, which means that all your inferences liable to type I error. There is nothing the researcher can do (we talked about this briefly prior), but to collect more data since more observations imply a better estimator of variance, and consequently reduces errors in inferences. However, if indeed e is correlated with the independent variables, the estimators would be biased and inconsistent just as in our earlier discussion of omitted variable bias. (Read the examples on pages 319 and 320 of your text).

1.2 Measurement Error in Independent Variable(s)

However, the real problem derived from measurement error is typically thought of as being derived from explanatory variables, and that the problem is more sinister here. Suppose what we're really interested in is the following,

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

which satisfies all the standard Gauss-Markov assumptions. However, if x_1^* is not observed, and is instead poorly measured by another variable, x_1 which we use. Suppose the measurement error is,

$$x_1 - x_1^* = e$$

where $e \in (-infty, infty)$. As usual, we will assume that e has mean of zero. Let's further assume that ϵ is uncorrelated with both x_1^* and x_1, that is $E(y|x_1^*, x_1) = E(y|x_1^*)$, and you should know that based on the law of iterated expectations, this simply means that x_1 does not provide any additional information once x_1^* is controlled for.

1. Suppose e is uncorrelated with the observed measure of x_1, that is $cov(x_1, e) = 0$, which would in turn imply that e must be correlated with x_1^*. This in turn implies that,

   $$y = \beta_0 + \beta_1 x_1 + (\epsilon - \beta_1 e)$$

   Since neither of the error elements are correlated with the observed independent variable, it then means that $E(\epsilon - \beta_1 e|x_1) = E(\epsilon - \beta_1 e) = 0$. Consequently, the OLS

estimators remain unbiased and consistent. Even though we assume the error terms are uncorrelated with each other, it is easy to see that,

$$var(\epsilon - \beta_1 e) = \sigma_\epsilon^2 + \beta_1^2 \sigma_e^2$$

And just like the case in the omission of a relevant variable, our variances are not efficient, and consequently our inferences are effected. Nonetheless, the OLS properties remains intact. In fact if β_1 is zero, our inference is totally unaffected. But then again, our model would be affected in so much as if the effect of x_1 is what we're interested in, than the coefficient being zero eliminates our need to study the model in the first place! Further note that this remains true even in the multiple variable case.

2. What economist are concerned with however isn't the first but the following. In fact it is so important we have a name for it, **Classical Errors-in-Variables (CEV)** problem, which assumes that the measurement error is uncorrelated with the unobserved variable x_1^*, consequently

$$cov(x_1^*, e) = 0 = E(x_1^* e) - E(x_1^*)E(e) \Rightarrow E(x_1^* e) = 0$$

Now lets express the measurement error as,

$$x_1 = x_1^* + e$$

Therefore if the first assumption holds,

$$cov(x_1, e) = E(x_1 e) - E(x_1)E(e) = E(x_1^* e + e^2) = E(e^2) = \sigma_e^2 \neq 0$$

i.e. the covariance between the covariate and the measurement error is the variance of the measurement error. What is the significance then? First note that the formula for β_1 is,

$$p\lim \beta_1 = \frac{cov(y, x_1)}{var(x_1)}$$

$$\Rightarrow p\lim \beta_1 = \frac{cov(\beta_0 + \beta_1 x_1 + \epsilon - \beta_1 e, x_1)}{var(x_1)}$$

$$\Rightarrow p\lim \beta_1 = \frac{\beta_1 var(x_1) - \beta_1 cov(x_1, e)}{var(x_1)}$$

$$\Rightarrow p\lim \beta_1 = \beta_1 \left(1 - \frac{cov(x_1, e)}{var(x_1)}\right)$$

$$\Rightarrow p\lim \beta_1 = \beta_1 \left(1 - \frac{\sigma_e^2}{var(x_1)}\right)$$

However, note that;

$$var(x_1) = var(x_1^* + e) = \sigma^2_{x_1^*} + \sigma^2_e$$

Therefore,

$$\Rightarrow p\lim \beta_1 = \beta_1 \left(1 - \frac{\sigma^2_e}{\sigma^2_{x_1^*} + \sigma^2_e}\right)$$

$$\Rightarrow p\lim \beta_1 = \beta_1 \left(\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*} + \sigma^2_e}\right)$$

The significance of the limiting value of the estimator is that $\frac{\sigma^2_{x_1^*}}{\sigma^2_{x_1^*}+\sigma^2_e}$ is always less than one, consequently, the OLS estimator of β_1 is always closer to 0, and that is why we call the bias an **attenuation bias**. Further this attenuation bias remains in the multiple variable case, and note that all parameters in the multiple variable case would be biased and inconsistent due this one single measurement error. Why? Recall that the OLS estimator in the multiple variable case has as its numerator the covariance between the error term from a regression with all the other covariates, and intuitively it introduces the measurement error everywhere since $cov(x_1, e) \neq 0$. The only time the other estimators are unbiased is when x_1^* is uncorrelated with the other variables, which is an unrealistic assumption.

3. The CEV assumption may not even hold themselves such that both x_1 and x_1^* are correlated with the error term, e. This then necessarily implies that all the estimators are biased and inconsistent. Fortunately, we still have another way out under certain assumptions, which we will cover shortly. The technique is known as Instrumental Variable, and is a Two Stage Estimation Technique.

2 Missing Data, Nonrandom Samples, and Outlying Observations

2.1 Missing Data

We have learned that both missing observations, and error in measurement of variables creates estimation problem under differing circumstances. What if have missing data points for observations that were examined during data collection. The answer is a simple one,

as long as the missing data are missing randomly, all we need to do is to drop the entire observation, since all it means that our sample size falls, and has no implications on the Gauss Markov assumption that we need for OLS, which is the the sample is random.

2.2 Nonrandom Samples

However the problem is more sinister when the missing data are deliberate in a sense. Consider an example such as a social mobility study where we wish to examine how income or educational attainment is transmitted between parents and children. Of course it is typical that not all respondents would recall or even know the education attainment of their parents. What if the lower educated individuals are more likely to not know the attainment of their parents. Then if we choose to drop those observations we would not have a representative sample, and further, the sample obtained would no longer be random, violating the Gauss Markov assumption that the sample be random. What happens then:

1. It turns out that when sample selection (as a result of our dropping of observations due to missing data) occurs on the basis on independent variables, we still get away with impunity in terms of estimator biasedness and consistency. This is commonly known as **Exogenous Sample Selection**, and means that there is sample selection based on the independent variables. An example is as follows, suppose we wish to examine how the various family structures affect the educational outcomes of children in those households, without recourse to the actual mechanism behind, such as whether it is because it is purely due to poor parenting skills or that their is a missing parent in single parent families, or the status of singleparenthood such as Divorced, Separated, Single Unmarried, Spouse Missing etc. Suppose the data is collected for only separated and divorced parent families, and intact families, excluding other family structures. Then we have selected the sample based on the independent variable of family structure. The reason why it will not affect our estimation or the use of OLS is that $E(Education|Parent'sEducation, ...)$ is the same for all family structures, consequently all it means is that we have a smaller sample size, and that we can't speak of the other family structures, but the parameter estimates remain consistent and unbiased. Of course there maybe a drop in variation in terms of the dependent variable, but it is to do with sample size.

2. However, the case such as that in the social mobility study is alittle complicated, since

the selection into the sample is not based on parental educational attainment. If the selection into sample based on the other variables are independent of the error term in the regression, there is no problem still, just as in the first case. For example, the case might have been selected based on the provincial weight, and that this variable is uncorrelated with the error, OLS remains unbiased and consistent.

3. The problem arises when the selection is based on the dependent variable . Consider the social mobility example again; suppose the data was selected based on the attainment levels of children, where we only select individuals with high school education or above. Then OLS is no longer valid. This kind of selection is known as sample selection based on dependent variable, and is also referred to as **Endogenous Sample Selection**. Why is that so? Consider the conditional expectation;

$$E(Educ|Prt'sEdu, ...) \neq E(Educ|Prt'sEduc, ..., Educ \geq HighSchool)$$

This problem is a problem with the design of the sample collection, and must be addressed there. **Read your text on Stratified Sampling on page 327 for your own knowledge of some alternative sampling schemes besides random samples**

2.3 Outliers and Influential Observations

It is quite inevitable that we may sometimes be faced with using small samples, and in the use of which we have to be concerned with lost of degrees of freedom. That is however not the sole concern, because in such small samples, we would also have to examine the degree of dispersion of the observations since it is in small samples that **outliers** can exercise the greatest effect on our estimation. An outlier is an observation that should we drop from our analysis, would totally change our estimate of interest. For example in an examination of how foreign direct investment might affect a nation's GDP among developing countries, suppose you included Taiwan, and Hong Kong, then the latter two would be outliers driving the estimation, and may not accurately give you the effect based on the majority of the observations. Note finally that when outliers occur in large data sets, we refer to them as influential observations. How could there be extreme effects in large sample? Consider a census data, where suppose you're interested in examining the educational attainments effect on income. Suppose on the average, the general population's income is $100,000.

However, the individuals of power in this census, are all well educated, and have extremely high incomes, say $1,000,000. Then their educational attainment would drive the value of an additional year of education sky high, even though it is not true.

The best way to anticipate problems arising from outliers to do a plot of the observations with the dependent variable on the vertical axis, and the independent variable on the horizontal axis. You can think of the outliers being on the far tail of the same distribution, or you could think of the outliers being in truth from a totally other distribution, in which case, in should not be included. In the FDI example, for intents an purposes, it is perhaps more useful to think about it as the latter, since if we think of Hong Kong and Taiwan, they are culturally different, and if each culture should be of a differing distribution, then we should drop them.

Of course sometimes the outliers are created by human error during the coding of the data, in which case, the best way to find out is to examine the summary statistics and the extreme values to see if there were codes that were not noted in the description of the data set.

If we cannot think of the outliers as being from a totally different distribution, it would serve us well to show results with and without the outliers, since all observations contain important information.

A good technique to reduce the effect of outliers is to use the logarithmic functional form, since as we had noted before, it shrinks the variation more thereby reducing the effect the outlier may have on our estimates.

Another method which we have not dealt with, but of which is not difficult to learn is the use of the estimation technique known as **Least Absolute Deviations (LAD)**. The idea behind OLS is that by definition, it is a conditional mean, conditioning on all independent variables. Consequently, it awards the same weight to all observations, consequently allowing all observations to affect it. Another way to think about it is that OLS gives greater weight to large residuals, precisely where the outliers are typically found (since OLS minimizes the sum of squared errors, and recalling that a quadratic function has a convex shape). However, the Least Absolute Deviation technique is a conditional **median** technique, and we know that the median is not susceptible to outlier effects. However, the technique has its own drawbacks:

1. Their is no close form solution to the technique and to solve, we have to write a program

that solves for the parameters iteratively. It cannot be a "hill climbing "algorithm since the absolute value is not a concave function.

2. The variances are valid asymptotically, while OLS is not. So you cannot use LAD in small samples.

3. The mean and median need not correspond, consequently, the estimates from the two methods may be very different when they do not correspond.

You can read more about this technique on page 332 and 333 of your text, and in more detail in more advanced books.