Measurement Error

Madison Ashworth and Bailey Kirkland

University of Wyoming

December 12, 2019

Definition

- ”The difference between an observed variable and the variable that belongs in a multiple regression equation.”

Measurement Error Examples

- Reported annual income as a indicator of actual annual income
- Consumer price index
- Survey data

Why is Measurement Error Important?

- Source of endogeneity
- A lot of economic data relies on survey data, etc
- Can cause biased and inconsistent OLS estimates, leading to potentially false conclusions.

Measurement Error in the Dependent Variable

Let y^* represent the true variable that we are trying to represent, giving us the normal regression equation

$$y_i^* = \beta_1 + \beta_2 x_{1i} + ... + \beta_k x_{ki} + \epsilon_i$$

Now let y represent the observed value of y^*. Let us assume there is some form of measurement error, represented by

$$u_i = y_i - y_i^*$$

After substitution, the new regression equation is

$$y_i = \beta_1 + \beta_2 x_{1i} + ... + \beta_k x_{ki} + \epsilon_i + u_i$$

Results of Measurement Error in the Dependent Variable

If u_0 is mean zero, then

- OLS assumptions are satisfied

If u_0 is not mean zero, then

- β_0 will be biased

u_0 is uncorrelated with x_j, then

- Estimates are unbiased and consistent
- Hypothesis testing can be done
- Larger variance ($var(e + u_0)$)

If u_0 is correlated with x_j, then

- Estimates will be biased

Example: taken from Woolridge pg. 319:
Consider the regression

$$savings_i^* = \beta_1 + \beta_2 income_i + \beta_3 education_i + e_i$$

where $savings^*$ is the actual savings rate, and *savings* is the reported savings rate.
Our measurement error would be:

$$u_i = savings_i - savings_i^*$$

Measurement Error in the Independent Variable

Our true regression model is

$$y_i = \beta_1 + \beta_2 x_{i2}^* + \epsilon_i$$

The measurement error is:

$$u_{i2} = x_{i2} - x_{i2}^*$$

After substitution, our new regression model is

$$y_i = \beta_1 + \beta_2 x_{i2} + \epsilon_i - \beta_2 u_{i2}$$

Classical Errors-in-Variables (CEV) Assumption: $Cov(x_2, u_{i2}) \neq 0$

If we assume

$$Cov(x_2^*, u_{i2}) = 0,$$

this implies that

$$Cov(x_{i2}, u_{i2}) = E(x_2 u_{i2}) = E(x_2^* u_{i2}) + E(u_{i2}^2) = 0 + \sigma_{u_{i2}}^2 = \sigma_{u_{i2}}^2$$

The covariance of x_2 and our composite error term is

$$Cov(x_2, \epsilon - \beta_2 u_{i2}) = -\beta_2 \sigma_{u_{i2}}^2$$

Our Gauss-Newton assumptions are violated. We get biased and inconsistent estimators.

Inconsistency of the Estimator under the CEV Assumption

We can calculate the inconsistency by

$$
\begin{aligned}
plim(b_2) &= \beta_2 + \frac{Cov(x_2, \epsilon - \beta_2 u_{i2})}{Var(x_2)} \\
&= \beta_2 + \frac{Cov(x_2, \epsilon - \beta_2 u_{i2})}{Var(x_2^*) + Var(u_{i2})} \\
&= \beta_2 + \frac{-\beta_2 \sigma^2_{u_{i2}}}{\sigma^2_{x_2^*} + \sigma^2_{u_{i2}}} \\
&= \beta_2 \left(1 - \frac{\sigma^2_{u_{i2}}}{\sigma^2_{x_2^*} + \sigma^2_{u_{i2}}} \right) \\
&= \beta_2 \left(\frac{\sigma^2_{x_2^*}}{\sigma^2_{x_2^*} + \sigma^2_{u_{i2}}} \right)
\end{aligned}
$$

Impact of CEV Assumption

- The plim of b_2 is always biased towards 0.
  - Attenuation Bias
  - if β_2 is positive, OLS will underestimate b_2. If β_2 is negative, OLS will overestimate b_2.
- The magnitude of the bias depends on the ratio of the $Var(x_2^*)$ and $Var(u_{i2})$.
- For a multivariate regression all OLS estimators will be biased (except for in the rare case where x_2^* is uncorrelated with x_j. The direction and magnitude of the biases for the other estimates are difficult to derive.

Possible Fixes

- Instrumental Variable
- Two-Stage Least Square