Eco311 Optional Reading: Measurement Error and IV estimation

(Jing Li, Miami University)

1. We already learn that a bias can arise due to an omitted variable w. One possible solution is trying to find data for the omitted variable, and then run a *multiple* regression that explicitly *controls* for that variable.

2. However, it is not uncommon that we can not find data or even agree on a definition for an omitted variable. For instance, we do not know how to quantify or measure "ability" or "motivation". Thus it's impossible to control for those two variables.

3. Instrumental Variable (IV) estimation is the method we can use to obtain an unbiased estimate of causal effect of x on y when the data of w is *unavailable*.

4. In this note, we show that a *different* bias can arise if data contain *measurement error*. IV estimation can also be used to resolve this bias.

5. Suppose the true relationship is

$$y = \beta x + u, \quad cov(x, u) = 0 \tag{1}$$

   Notice that we assume x is uncorrelated with the error term u, i.e., x is *exogenous*. That means the issue here is not about an omitted variable.

6. The issue is that we actually *observe* x^*, which differs from the true value x by an error (or noise). For instance, x is true family income, and x^* is reported income.

$$x^* = x + e, \quad cov(x, e) = 0, cov(e, u) = 0 \tag{2}$$

   where we assume the measurement error e is uncorrelated with x and u.

7. It follows that the OLS estimate of slope coefficient when regressing y onto x^* converges to

$$\hat{\beta} \to \frac{cov(x^*, y)}{var(x^*)} = \frac{cov(x + e, \beta x + u)}{var(x + e)} = \beta \frac{var(x)}{var(x + e)} = \beta \frac{var(x)}{var(x) + var(e)} \tag{3}$$

   Notice that $\frac{var(x)}{var(x)+var(e)} < 1$, and as $var(e)$ rises, $\hat{\beta} \to 0$. Therefore, the estimate $\hat{\beta}$ is biased toward 0. We call this *attenuation bias*. Do not confuse this bias with OVB.

8. Next we run a Monte Carlo simulation

```
> set.seed(12345); n = 1000; z = rnorm(n)
> x.true = runif(n)+z
> e = rnorm(n)
> x.observed = x.true + e
> u = rnorm(n)
> y = 3*x.true + u
> summary(lm(y~x.observed))$coef
             Estimate Std. Error   t value      Pr(>|t|)
(Intercept) 0.5840778 0.07771320  7.515811  1.255826e-13
x.observed  1.5464169 0.04962684 31.160897 1.962002e-149
```

   The true β is 3; the OLS estimate is 1.5464169. The estimate is closer to 0 than the true value, confirming the attenuation bias.

9. The instrumental variable is denoted by z, which should satisfy conditions that (i) z and e are uncorrelated; (ii) z and x are correlated; (iii) y does not depend on z directly. To use the instrumental variable, we first regress x^* onto z, and keep the fitted value $\hat{x}$. Next, we regress y onto that fitted value.

```
> xhat = fitted(lm(x.observed~z))
> summary(lm(y~xhat))$coef
               Estimate Std. Error   t value  Pr(>|t|)
(Intercept) -0.07043426 0.04518224 -1.558892 0.1193389
xhat         2.83460916 0.03851322 73.600941 0.0000000
```

   We see that the IV estimate 2.83460916 is close to the true value 3.

10. In general, when x is *endogenous*, one formula for the IV estimator is

$$\hat{\beta}^{iv} = \frac{S_{z,y}}{S_{z,x}} \to \frac{cov(z,y)}{cov(z,x)} = \frac{cov(z,\beta x+u)}{cov(z,x)} = \beta \tag{4}$$

    where we use the assumption (i) $cov(z,u)=0$, (ii) $cov(z,x)\neq 0$, and (iii) z is absent in the model for y. Those three *assumptions* are called *exogeneity*, *relevance*, and *exclusion*, respectively. Read chapter 15 of Wooldridge's book to learn more about IV estimation.