AREC422 Notes

Youpei Yan

April 25, 2019

Data Issues

1. Proxy variable

Suppose that we miss a key variable due to data un-availability. To avoid OVB, we need to obtain a proxy variable for this omitted variable.
e.g., we use IQ score to proxy intelligence, use distance to a garbage incinerator to proxy the desire of good environment, use income for risk-averse level, etc.

Mathematically: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \beta_3 x_3^* + u$, where x_3^* is unobserved (but omitting it will cause OVB). We use x_3 as a proxy for x_3^*. $x_3^* = \delta_0 + \delta_3 x_3 + v_3$.

We run the regression $y \sim x_1 + x_2 + x_3$ in reality to get the unbiased estimators of β_1 and β_2.

Feature: 1. The proxy should have a positive relationship with the omitted variable. 2. The proxy should not introduce additional correlation with the error.
$corr(x_1, u) = corr(x_2, u) = corr(x_3^*, u) = corr(x_3, u) = corr(x_1, v_3) = corr(x_2, v_3) = corr(x_3, v_3) = 0$
$E(x_3^*|x_1, x_2, x_3) = E(x_3^*|x_3) = \delta_0 + \delta_3 x_3$.

With all these assumptions satisfied,
$y = (\beta_0 + \beta_3\delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \delta_3 x_3 + (u + \beta_3 v_3) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$

We run this regression and get the same old $\hat{\beta}_1$, $\hat{\beta}_2$ as desired, α_3 can reflect the impact of x_3^*.

2. Measurement Error

A. Measurement error in y

We observe y rather than y^* in our dataset, but GMAs are satisfied.
Let $e_0 = y - y^*$, then $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u + e_0$, usually, we simply assume that $corr(e_0, u) = 0$, so we just have a larger error term and larger error variance.
$Var(u + e_0) = \sigma_u^2 + \sigma_0^2 > sigma_u^2$.

We have larger variances of the OLS estimators, unprecised estimators but still unbiased.

B. Measurement error in the independent var. (could be troublesome)

Observe x_1 rather than x_1^*, $e_1 = x_1 - x_1^*$. $E(e_1 = 0)$, but $corr(u, x_1^*) = corr(u, x_1) = 0$.

Case 1: uncommon assumption: $Cov(x_1, e_1) = 0$.
Note that e_1 is a function of x_1, x_1^*, so $corr(x_1^*, e_1) \neq 0$.

$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$, $corr(x_1, (u - \beta_1 e_1)) = 0$, so $\hat{\beta}_1$ is consistent. $Var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_e^2$

We have larger variances of the OLS estimators, unprecised estimators but still unbiased.

Unfortunately, this is not a common assumption.

Case 2: classic assumption in econometrics:
2 unobserved variables are not correlated. $Corr(x_1^*, e_1) = 0$

In this case, $Cov(x_1, e_1) \neq 0$.
$Cov(x_1, e_1) = E(x_1 e_1) = E(x_1^* e_1) + E(e_1^2) = \sigma_e^2$
$Cov(x_1, (u - \beta_1 e_1)) = -Cov(x_1, e_1)\beta_1 = -\beta_1 \sigma_e^2 \neq 0$
$\hat{\beta}_1$ is inconsistent.

$$plim \hat{\beta}_1 = \beta_1 + \frac{Cov(x_1, (u - \beta_1 e_1))}{Var(x_1)} = \beta_1 + \frac{-\beta_1 \sigma_e^2}{Var(x_1^*) + Var(e)} = \beta_1 - \frac{\beta_1 \sigma_e^2}{\sigma_{x^*}^2 + \sigma_e^2}$$
$$\therefore plim \hat{\beta}_1 = \beta_1 \big(1 - \frac{\sigma_e^2}{\sigma_e^2 + \sigma_{x^*}^2}\big) = \beta_1 \frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2} < \beta_1$$

This is called the Attenuation Bias: the real impact is attenuated or under-estimated.

3. Missing Data

Missing at random: fine; nonrandom samples: violating GMA2. Are OLS estimators biased or inconsistent?

Case 1: Exogenous sample selection:
Sample selection is based on the independent variables.

If it is just missing indep. variables, we just have a smaller sample. OLS estimators are unbiased.

Case 2: Endogenous sample selection:
Sample selection is based on the dependent variable.

We'll get biased estimators.

e.g. profit $\sim$ production + pollution + other factors.
Firms with high pollution levels are observed in the data: unbiased estimator.
Firms with higher profit levels are observed in the data: biased estimator.