Lecture 11

Stochastic Regressors & Measurement Errors

There are random variables in the RHS.

- When those variables are correlated with the errors, there is a bias problem.
- There may also be a consistency problem.

For a two-variable model: $y_i = \beta_1 + \beta_2 x_i + u_i$.

According to OLS, $\hat{\beta}_2 = \dfrac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$.

- We also know that $\hat{\beta}_2 = \beta_2 + \dfrac{\sum(x_i - \bar{x})(u_i - \bar{u})}{\sum(x_i - \bar{x})^2}$.
    - Since $\bar{u}$ is expected to be zero, it is sometimes omitted.
    - Demonstration: We know that $\bar{y} = \beta_1 + \beta_2 \bar{x}$. We subtract it from the original model, then we have $y_i - \bar{y} = \beta_2(x_i - \bar{x}) + (u_i - \bar{u})$. Plugging it into the formula for $\hat{\beta}_2$, we obtain $\hat{\beta}_2 = \dfrac{\sum\left\{(x_i - \bar{x})\left[\beta_2(x_i - \bar{x}) + (u_i - \bar{u})\right]\right\}}{\sum(x_i - \bar{x})^2}$, which can be rewritten as $\hat{\beta}_2 = \beta_2 \dfrac{\sum(x_1 - \bar{x})^2}{\sum(x_1 - \bar{x})^2} + \dfrac{\sum(x_1 - \bar{x})(u_i - \bar{u})}{\sum(x_1 - \bar{x})^2}$.
- So, if $\mathrm{E}\left[\dfrac{\sum(x_i - \bar{x})(u_i - \bar{u})}{\sum(x_i - \bar{x})^2}\right] \neq 0$, $\hat{\beta}_2$ is <u>biased</u>, that is, $\mathrm{E}\left[\hat{\beta}_2\right] \neq \beta_2$.
- And if $N \to \infty : \dfrac{\sum(x_i - \bar{x})(u_i - \bar{u})}{\sum(x_i - \bar{x})^2} \nrightarrow 0$, $\hat{\beta}_2$ is <u>inconsistent</u>, that is, $\hat{\beta}_2 \nrightarrow \beta_2$ or $\operatorname{plim} \hat{\beta}_2 \neq \beta_2$.
    - If we multiply both the numerator and the denominator on the RHS by $\frac{1}{N}$, they become estimates of the covariance of x and u and variance of x, respectively.
    - As N approaches infinity, the estimates tend to the true values, that is, $\dfrac{1}{N}\sum(x_i - \bar{x})(u_i - \bar{u}) \to \mathrm{Cov}(x, u)$ and $\dfrac{1}{N}\sum(x_i - \bar{x})^2 \to \mathrm{Var}(x)$
        - $\mathrm{Var}(x)$ may also be written as σ_x^2.
    - If x and u **are not independent**, then $\dfrac{\mathrm{Cov}(x, u)}{\mathrm{Var}(x)} \neq 0$.

Measurement errors:

There are two kinds of measurement errors:

1. Error in the x's: bias, consistency problems.
2. Error in the y's: not a big problem.

Suppose the true relationship is given by $y_i = \beta_1 + \beta_2 z_i + v_i$.

We cannot, however, measure z_i directly; instead, we have a noisy measure $x_i = z_i + w_i$.

- x_i: measure.
- z_i: true value.
- w_i: measurement error.

Let us call $\text{Var}(z_i) = \sigma_z^2$ and $\text{Var}(w_i) = \sigma_w^2$, and assume that the measurement error and the true value are independent, $w_i \perp z_i$.

- That means that the measurement error is not a function of the measured variable's true value.

Thus, the model can be rewritten as $y_i = \beta_1 + \beta_2(x_i - w_i) + v_i$ or even $y_i = \beta_1 + \beta_2 x_i + (v_i - \beta_2 w_i)$.

- Let $u_i = v_i - \beta_2 w_i$.
- Then, $y_i = \beta_1 + \beta_2 x_i + u_i$.

Now, since $x_i = z_i + w_i$ and $u_i = v_i - \beta_2 w_i$ (both x_i and u_i depend on w_i), there is correlation between a RHS variable and the errors.

We know that $x = z + w$.

- Then $\text{Var}(x) = \text{Var}(z + w) = \text{Var}(z) + \text{Var}(w) + 2\text{Cov}(z, w)$.
- Since $z \perp w$, then $\text{Cov}(z, w) = 0$ and $\text{Var}(z + w) = \text{Var}(z) + \text{Var}(w) = \sigma_z^2 + \sigma_w^2$.

We also know that $u = v - \beta_2 w$.

- Then, $\text{Cov}(x, u) = \text{Cov}(z + w, v - \beta_2 w)$.
- Applying the properties of the covariance, we rewrite it as $\text{Cov}(x, u) = \text{Cov}(z, v - \beta_2 w) + \text{Cov}(w, v - \beta_2 w)$.
  - It can broken down again into $\text{Cov}(x, u) = \text{Cov}(z, v) - \beta_2 \text{Cov}(z, w) + \text{Cov}(w, v) - \beta_2 \text{Cov}(w, w)$.
  - Since $z \perp v$, then $\text{Cov}(z, v) = 0$.
  - Since $z \perp w$ (by assumption), then $\text{Cov}(z, w) = 0$.
  - Since $w \perp v$ (unrelated), then $\text{Cov}(w, v) = 0$.
- In the end, $\text{Cov}(x, u) = -\beta_2 \text{Cov}(w, w) = -\beta_2 \sigma_w^2$.

We have already seen that $\text{plim}\ \hat{\beta}_2 = \beta_2 + \dfrac{\text{Cov}(x, u)}{\text{Var}(x)}$.

- Thus, $\text{plim}\ \hat{\beta}_2 = \beta_2 - \beta_2 \dfrac{\sigma_w^2}{\sigma_z^2 + \sigma_w^2}$.
  - If there is no measurement error whatsoever, $\sigma_w^2 = 0 \to \text{plim}\ \hat{\beta}_2 = \beta_2$ (consistency).
  - As the variance of the measurement error grows, that is, $\sigma_w^2 \to \infty$, the estimated coefficient gets biased towards zero.

What if the measurement error is in the dependent variable (y)?

True relationship: $Q_i = \beta_1 + \beta_2 x_i + u_i$.

Measured variable: $y_i = Q_i + r_i$, where r_i is the measurement error.

- Then, $y_i - r_i = \beta_1 + \beta_2 x_i + u_i$.
- It can be rewritten as $y_i = \beta_1 + \beta_2 x_i + (u_i + r_i)$.
  - By assumption, x_i and u_i are not related ($x_i \perp u_i$).
  - And, according to the definition of the measured variable, x_i and r_i are correlated either ($x_i \perp r_i$).
- Then, the error might get a bit bigger, but there is no bias nor inconsistency.

Imperfect proxies:

> "When you use proxies for the real variables, you are going to introduce bias."

> "If you could show that there is a tight relationship between, say, house size and what you pay for it - and there probably is -, then your proxy is probably pretty *darn* good."
>
> "But it there is not a tight relationship between your proxy and the actual variable, you may be introducing quite a bit of bias."