Eco311 Optional Reading: Identification

(Jing Li, Miami University)

1. Consider a simple regression

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

we already know that the OLS estimate $\hat{\beta}_1$ converges to β_1 if *exogeneity* holds

$$\hat{\beta}_1 = \frac{S_{xy}}{S_x^2} \to \frac{cov(x,y)}{var(x)} = \beta_1 + \frac{cov(x,u)}{var(x)} \tag{2}$$

$$\hat{\beta}_1 = \beta_1 \quad if\ cov(x,u) = 0\ (Exogeneity) \tag{3}$$

$$\hat{\beta}_1 \neq \beta_1 \quad if\ cov(x,u) \neq 0\ (Endogeneity) \tag{4}$$

2. Alternatively, we say that β_1 can be *identified* under the condition of exogeneity. By contrast, if exogeneity fails or *endogeneity* arises, β_1 *cannot* be identified since

$$\beta_1 = \hat{\beta}_1 - \frac{cov(x,u)}{var(x)} = \hat{\beta}_1 - \texttt{unknown omitted variable bias} \tag{5}$$

where $cov(x,u) \neq 0$ is unknown as we do not observe u.

3. There are three approaches to resolve the endogeneity issue (or achieve identification)

   (a) Matching

   - Let c be constant. Recall that $cov(x,c) = 0$. The key idea of matching is letting $u = c$, or in words, let other factors be held equal. Actually that is exactly the meaning of *ceteris paribus*.
   - In practice, we may used data that *match as pairs* and conduct a *matched pair analysis*. For instance, by using data of twins, we can ensure the innate ability, family background and etc are held constant. See Example C.3 of Wooldridge's book (on page 732 of seventh edition).
   - In some sense, the method of *Regression Discontinuity* is based on matching as well. Essentially that method compares observations just below and above some cut-off value, assuming those observations are otherwise comparable.

   (b) Randomization

- The main idea of randomization is letting x be *randomly assigned.* For instance, to identify the causal effect of going to selective schools on earnings, we may let high school students who have number 9 as the last digit of social security number go to selective schools. It follows that

  $$cov(x, u) = cov(SSN, u) = 0$$

  or randomization leads to exogeneity.
- One example is the paper titled "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination" in which the authors randomly assign names commonly used by white and black peoples to fake resumes. Then they are able to identify the causal effect of race on labor market outcome. Google "A lottery to lose, An enlightened scheme to benefit poor children seems to do the opposite" to see another example of randomization study.
- More realistically, randomization may be applied to variable z. As long as z and x are related, and y does not depend on z, then z can be used as an *instrumental variable* (IV)

  $$\hat{\beta}_1^{IV} = \frac{cov(y, z)}{cov(x, z)} = \beta_1 \quad if\ cov(z, u) = 0\ and\ cov(z, x) \neq 0 \tag{6}$$

  For instance, US military used draft *lottery* to determine the order of call to military service in Vietnam War. Google "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records".

(c) Controlling

- Controlling entails taking more factors out of the error terms, i.e., finding data for control variables and running a multiple regression with those controls.
- Panel data is very helpful since dummy variables can be used to control for factors that do not vary over time.
- In general, dummy variables can be used to control for factors that are constant in certain dimension. For instance, google "Estimating the Payoff to Attending a More Selective College: An Application of Selection on Observables and Unobservables", and see how the authors use dummy variables to control for unobserved ability, motivation, etc.