Lecture Module 8

Agenda

1. Endogeneity
2. Instrumental Variables
3. Two-stage least squares
4. Panel Data: First Differencing
5. Panel Data: Fixed Effects
6. Panel Data: Difference-In-Difference

Endogeneity

Overview

- The most important assumption is $E(u|x) = 0$
- Biased estimates if $E(u|x) \neq 0$
- Generally, we call this a problem of "endogeneity"
- x is endogenous to unobserved factors which influence y.

Endogeneity

Examples

- Fast food availability and Obesity (Dunn, 2010)

$$ObesityRate = \beta_0 + \beta_1 McDensity + u$$

  - If u correlated with *McDensity*, then we have a problem
- Mental Health and Employment (Tefft , 2012)
  - Poor Mental Health ⇒ Poor Employment outcomes ⇒ Poor Mental Health....
  - Causality?
- Effect of CO2 on Temperature (Stock, Stern and Kaufmann, 2005)
  - More CO2 ⇒ Higher T ⇒ More CO2 ⇒ Higher T
  - Other causes for higher temperatures?
  - "Spurious" correlation

Instrumental Variable Estimation

- A common tool to correct for endogeneity is "Instrumental Variable Estimation"

$$wage = \beta_0 + \beta_1 educ + u$$

- The issue is that $E(u|educ) \neq 0$
- Need to find an "instrumental variable", z
- Need z that satisfies the following:

$$\begin{aligned} Cov(educ, z) &\neq 0 \\ Cov(u, z) &= 0 \end{aligned}$$

  - z should be correlated with *educ*
  - z should be uncorrelated with u

- z "instruments" for education
- Wish to "purge" *educ* of its endogeneity

Instrumental Variable Estimation

- Problem: Can't test $Cov(u, z) = 0$
- Two options:
  - Appeal to intuition
  - Regress on a proxy variable
- In the wage dataset, are there any $z's$ that are correlated with *educ*, but not *ability*?
  - meduc, feduc?
    - Probably not
    - Maybe try $\frac{meduc}{meduc+feduc}$?
  - sibs?
    - Potentially: might correlated with resources available for education, but not natural ability
  - urban, married?
    - Potentially, but probably not, as they are both choices and clearly endogenous to other factors.

Instrumental Variable Estimation - 2SLS

- Three-step procedure to estimate β_1 with endogenous *educ* from:

$$wage = \beta_0 + \beta_1 educ + \beta_2 iq + u$$

- Use *sibs*, *birthord*, and $\frac{meduc}{meduc+feduc}$ as instruments:

  1. Regress *educ* on *sibs*, *birthord*, $\frac{meduc}{meduc+feduc}$ and *iq*

$$educ = \pi_0 + \pi_1 sibs + \pi_2 birthord + \pi_3 \frac{meduc}{meduc+feduc} + \pi_4 iq + v$$

  2. Obtain predicted values of $\widehat{educ}$
  3. Regress

$$wage = \beta_0 + \beta_1 \widehat{educ} + \beta_2 iq + u$$

- If assumptions are correct and instruments are good, $\widehat{educ}$ not correlated with u, and $\hat{\beta}_1$ is not biased
- Can do all three steps at once using ivreg in stata.
- First stage regressions should have high full exclusion F-Stat (>10).

Endogeneity

Examples

- Fast food availability and Obesity (Dunn, 2010)

$$ObesityRate = \beta_0 + \beta_1 McDensity + u$$

  - Instrument: Distance from an interstate
  - Some effect of *McDensity* on Obesity rates.

- Mental Health and Employment (Tefft , 2012)

$$Employment = \beta_0 + \beta_1 MentalHealth + u$$

  - Instrument: Latitude and Day of Year (SAD)
  - One additional day of poor mental health per month $\Rightarrow$ 5% increase in likelihood of unemployment
  - Unmarried, poor and uninsured are more adversely affected

Endogeneity

Examples

- Effect of CO2 on Temperature (Stock, Stern and Kaufmann, 2005)

$$\begin{aligned} Temp &= \beta_0 + \beta_1 CO2 + \beta_2 Erupt + u \\ CO2 &= \delta_0 + \delta_1 Temp + w \end{aligned}$$

  - *Erupt*: Volcanic Sulfates
  - Predictions are business as usual

Panel Data

- What is panel data?
  - Multiple agents surveyed over time
  - Ideal data, which allows for a more detailed accounting of unobserved variables
  - Also, can evaluate how the same sample of respondents changes over time.
- Panels can have two types:
  - Balanced - no attrition from the sample. Everybody is observed the same number of times.
  - Unbalanced - some individuals observed in the sample more than others. This requires careful analysis, beyond the scope of this class.
- Basic Panel Techniques for Regression
  - First-differences
  - Fixed effects

Panel Data

- Consider our basic wage and eduction regression in a panel setting:

$$wage_{it} = \beta_0 + \beta_1 educ_{it} + u_{it}$$

  - i is individual
  - t is the time period
  - With N individuals, T time periods, and a balanced panel, there are $N \cdot T$ observations.

- Do you think the error terms u_{it} are correlated over time, or uncorrelated?
- They are probably correlated by individuals. Consider natural $ability_i$ as an unobserved variable.

$$u_{it} = v_{it} + ability_i$$

- Since natural ability does not vary with t, we can do a few things to get rid of its effect on the returns to education.

Panel Data: First-differences

- The first technique that we will study is "first-differencing"
  - We estimate the model in changes *within the individual*, rather than levels.
  - Using "within" variation to estimate the model, though this will *not* be called a "within estimator".
  - Critically, anything that is time-invariant *within the individual* is eliminated.
- To see this, take differences of the wage equation between t and $t-1$

$$\Delta wage_{it} = \beta_1 \Delta educ_{it} + \Delta u_{it}$$

where

$$\Delta u_{it} = \Delta v_{it} + \Delta ability_i = \Delta v_{it}$$

- Since $ability_i$ is time invariant, differencing **removes** the bias.

Panel Data: First-differences

- What are we assuming about the effects of education using first-difference identification?
  - That the effects of the education "shock" are immediate.
- Suppose that *wage* and *educ* have the following one-to-one relationship, in order of time:

  wage: 5 5 5 6 6 6 7 7 7
  educ: 7 7 7 8 8 8 9 9 9

- In first differences,

  wage: 0 0 0 1 0 0 1 0 0
  educ: 0 0 0 1 0 0 1 0 0 → correlation=1

- However, if the effects on the wage are lagged one year,

  wage: 0 0 0 0 1 0 0 1 0
  educ: 0 0 0 1 0 0 1 0 0 → correlation=-0.286

- First-differencing evaluates the short-run effects of changes.
  - Informative, but can result in biases if the effects are lagged.

Panel Data: First-differences

- Let's now evaluate the effects of a gradual improvement in education:

  wage: 5 5 5 6 7 8 8 8 8
  educ: 6 6 6 7 8 9 9 9 9

- In first differences,

  wage: 0 0 0 1 1 1 0 0 0
  educ: 0 0 0 1 1 1 0 0 0 → correlation=1

- However, if the effects on the wage are lagged one year,

  wage: 0 0 0 0 1 1 1 0 0
  educ: 0 0 0 1 1 1 0 0 0 → correlation=0.5

- Lagged effects are less of an issue when the effects are gradual.

  Solutions:

  - Take "long differences"
  - Use a "within estimator" ie. Fixed effects.

Panel Data: Fixed effects

- The fixed effects, or "within estimator", starts with the same basic specification.

$$wage_{it} = \beta_0 + \beta_1 educ_{it} + u_{it} \tag{1}$$

where again we assume that:

$$u_{it} = v_{it} + ability_i$$

- Take the mean of the (1) equation for individual i:

$$\overline{wage}_i = \beta_0 + \beta_1 \overline{educ}_i + \overline{u}_i \tag{2}$$

- Then, for each individual i, subtract (2) from (1):

$$wage_{it} - \overline{wage}_i = \beta_1 \left(educ_{it} - \overline{educ}_i \right) + \left(u_{it} - \overline{u}_i \right)$$

where

$$u_{it} - \overline{u}_i = v_{it} - \overline{v}_i + ability_i - \overline{ability}_i = v_{it} - \overline{v}_i$$

- Each individual has been "de-meaned" for all variables, which eliminates the average effect of ability.

Panel Data: Fixed effects

- Let's return to the simple numerical example from first-differences:

  wage: 5 5 5 6 6 6 7 7 7
  educ: 7 7 7 8 8 8 9 9 9

- Using the within transformation:

  wage: -1 -1 -1 0 0 0 1 1 1
  educ: -1 -1 -1 0 0 0 1 1 1 → correlation=1

- Now suppose that the wage effects are lagged one year:

  wage: 5 5 5 5 6 6 6 7 7
  educ: 7 7 7 8 8 8 9 9 9

- Using the within transformation:

  wage: -0.78 -0.78 -0.78 -0.78 0.22 0.22 0.22 1.22 1.22
  educ: -1.00 -1.00 -1.00 0.00 0.00 0.00 1.00 1.00 1.00

  → correlation=0.866

- De-meaning captures average effects, within each individual.
- The timing of the effects matter, but much less so than with FD.

Panel Data: Fixed effects

- The within estimator can be estimated directly using "fixed effects"

$$wage_{it} = \beta_0 + \beta_1 educ_{it} + \alpha_i + u_{it}$$

  - α_i is a "fixed effect" for individual i.

- The fixed effect α_i is actually short-hand for a dummy variable identifying individual i and an associated coefficient.

- Like a dummy variable, when including a constant β_0, we have to drop one of the α_i's.

  - Effects are measured relative to a base group.

- But α_i captures all average effects for each individual, including fully "absorbing" factors that are specific to i and invariant across time.

  - $Ability_i$ will be "absorbed" by α_i.

Panel Data: Fixed effects

- Fixed effects can represent multiple groups

$$wage_{it} = \beta_0 + \beta_1 educ_{it} + \alpha_i + \alpha_t + u_{it}$$

  - α_i is a "fixed effect" for individual i.
  - α_t is a "fixed effect" for year t.

- How might year effects be important in a wage on education regression?
  - Technical change. Higher technology requires more education, but also makes workers more productive thereby increasing wages.

- You can add fixed effects for any group, and interpret it as an estimator "within" those groups
  - In the above example, within individuals and years, the effect....

Difference in Difference

- "Diff-in-Diff" is a common panel technique used for identification.
- We have two periods, pre and post.
  - Post period identified by the dummy variable, $Post_t$.
- Two groups, Treatment and Control.
  - Treatment group identified by the dummy variable, $Treat_i$.
- Diff-in-Diff Specification on the wage:

$$wage_{it} = \beta_0 + \beta_1 Post_t + \beta_2 Treat_i + \delta Post_t \cdot Treat_i + u_{it}$$

- δ is called the "differences-in-differences" estimator.
- δ is also called the "average treatment effect".
- Where does the term "diff-in-diff" come from?

Difference in Difference

- The equation:

$$wage_{it} = \beta_0 + \beta_1 Post_t + \beta_2 Treat_i + \delta Post_t \cdot Treat_i + u_{it}$$

- "Diff-in-Diff" takes two differences:
  - Post vs. Pre
  - Treatment vs. Control
- Write out predictions, their difference, then the difference in differences:

	Before	After	After - Before
Control	β_0	$\beta_0 + \beta_1$	β_1
Treatment	$\beta_0 + \beta_2$	$\beta_0 + \beta_1 + \beta_2 + \delta$	$\beta_1 + \delta$
Control - Treatment	β_2	$\beta_2 + \delta$	δ

- Treatment group might naturally differ from control: β_2.
- Post might naturally differ from Pre: β_1.
- Getting rid of both yields δ, the average treatment effect.