Data Issues and Empirical Strategy

Sang-Hyop Lee

38th Summer Seminar: WS 1

Population, Development, and Policy

June 1, 2007

1

Type of Data I

- Cross section
- Time series
- Cross section time series; useful for aggregate cohort analysis
- Panel (longitudinal)
- Cross section with retrospective information

2

Type of data II: Good/Bad/Ugly

- Uneasy alliance (Zvi Griliches): "Badness" of the data provides econometricians with living.
- Criteria for good quality data
  - Extent (richness, quantity of variables)
  - Reliability (measured without error)
  - Validity (representative)

3

Nightmare

- FIES in 1996
  (64,433 household with 233,225 individuals)
  - Measured for only urban area
  - No single person household
  - No individual level income, only household level
  - No information of income for family owned business
  - Information of up to 8 household members (0.6% household has more than 8 members)

4

Measurement Error

- Response error (respondents do not know what is required; incentive to understate; recall bias)
- Reporting error → heaping or outliers
- Coding error
- Overestimate/Underestimate
  - Parents do not report their children until the children have name
  - Detect by checking survival rate of single age
- Discrepancy between aggregate value and individual value

5

Missing/Change of Variables

- Never been measured
  - Only measured for a certain group
  - Labor portion of self-employed income
- Change of variables over time
  - Institutional/policy change
  - New consumption items, new jobs, etc
- Change of survey instrument/collapsing

6

Censored Observation

- Top/Bottom coding
- Characteristics due to the time of survey
  - Seasonal fluctuation
  - Duration of unemployment
  - Completed years of schooling
- Attrition (Panel data)

7

Categorical/Qualitative Variables

- Converting categorical to single continuous variables
  - Grouped by age (population, public education consumption)
  - Income category (FPL)
- Inconsistency over time
  - Categorical → continuous, and vice versa

8

Units, Real vs. Nominal

- Careful about the reporting unit
  - Measuring units
  - Reporting period units (reference period, seasonal fluctuation, recall bias)
- Nominal vs. Real
  - Aggregation across items
  - Quality change (e.g. computer)
  - Where inflation is a substantial problem

9

Solution 1: Data Cleaning

- As clean as possible
  - Find out what data sets are available and choose the best one (template for workshop)
  - Detect outliers and examine them carefully
  - A serious examination is required when inflation matters to check whether actual estimation process generate a variable
  - Make variables consistent
  - Convert categorical variable to continuous variable, etc.

10

Solution 2: Missing Variables

- Ignore it; random non-response
- Find other source of data (FIES vs. LFS)
- Impute
  - Based on their characteristics or mean value
  - Modified zero order regressions (y on x)
    - Create dummy variable for missing variables of x (z)
    - Replace missing variable with 0 (x')
    - Regress y on x' and z, rather than y on x

11

LFS or FIES (or LSMS)?

- If FIES has information on individual earnings, no reason to use LFS.
- Otherwise, use LFS to construct individual earnings profile and use FIES to complement other source of labor income.

  - LFS may not be a representative sample
- Discrepancy between FIES, LFS, and NIPA
- How about using LSMS?

12

Estimation: Household vs. Individual

- Welfare measurement should be individual level
- But a lot of data are gathered from household
  - Allocating household consumption (income) to individual household members is the most critical part of estimation (June 4, by SH Lee)
  - Adjusting using aggregate (Macro) control (June 4, by A.Chawla)

13

Estimating Consumption

- Consists of Private and Public consumption
- *Private Consumption*: Value of the consumption good and services acquired and consumed by households
  - Education, health, and others
  - Rental value of owner occupied housing
  - Flow of services from durables
- *Public Consumption*: Value of the consumption goods and services provided by general government

14

Other Challenges; Consumption

- Consumption
  - Home produced items: both income and consumption.
  - Estimating private health expenditure age profiles: very complex in part due to various source of financing. Should depend on each county data (June 4, by SH Lee)
  - How to treat indirect taxes?
  - Should cost of childbearing be allocated to the mother or to the child?

15

Estimating Labor Income

- All compensation that is a return to work effort.
- Includes:
  - Wages and salaries (earnings)
  - Some portion of entrepreneurial or self-employment income (2/3)
  - Fringe benefits
  - Deferred compensation

16

Other Challenges; Labor Income

- Labor Income
  - How do we estimate and allocate self-employed income?
  - How do we estimate and incorporate deferred compensation? How about seniority-based wage systems?
  - In-kind fringe benefits

17

Weighting I

- Weight should be used in the summary of variables/direct tabulation/regression/smoothing.
- Frequency Weights; fw indicate replicated data. The weight tells the command how many observations each observation really represents.

  . tab edu [w=wgt] ⇔ tab edu [fw=wgt]
- Analytic Weights; aw are inversely proportional to the variance of an observation. It is appropriate when you are dealing with data containing averages.

  . su edu [w=wgt] ⇔ su edu [aw=wgt]

  . reg wage edu [w=wgt] ⇔ reg wage edu [aw=wgt]

18

Weighting II and Clustering

- Probability Weights; pw are the sample weight which is the inverse of the probability that this observation was sampled.

  . reg wage edu [pw=wgt] ⇔ reg wage edu [(a)w=wgt], robust

  . reg wage edu [pw=wgt], cluster(hhid)

  ⇔ reg wage edu [(a)w=wgt], cluster(hhid)

19

Smoothing

- Shows the pattern more clearly by reducing sampling variance
- Should not eliminate real features of the data
  - Avoid too much smoothing (e.g. old-age health expenditure.)
  - We don't want to smooth some profiles (e.g. education)
  - Basic components should be smoothed, but not aggregations
- Type of smoothing

  - "lowess" smoothing (Stata) does not incorporate sample weight

  - Friedman's super smoothing (R) does.

20

Discussion

- Estimation method could vary across countries depending on data.
- However, can apply some standard measure for all the countries and compare the results
  - Definition → Specification → Estimation using weight → Smoothing → Show your work!
  - If some component vary substantially by age, then it is better to estimate separately
    (education, health, etc)

21

The End

- Let's talk about measuring lifecycle deficit in detail (June 4, by SH Lee)
- For the time being, let's find relevant variables from your own data sets and clean them.

Mahalo!

22