Data Analysis

Sang-Hyop Lee
University of Hawaii at Manoa &
East-West Center

Seminar and Training Workshop on NTA
Organized by NUPRI and TDRI
December 16-25, 2010, Thailand

Data Sets for Statistical Analysis

- Cross section
- Time series
- Cross section time series; useful for aggregate cohort analysis
- Panel (longitudinal)
  - Repeated cross-section design: most common
  - Rotating panel design (Cote d'Ivore 1985 data)
  - Supplemental cross-section design (Kenya & Tanzania 1982/83 data, MFLS)
- Cross section with retrospective information
- Micro vs. Macro

Quality of Survey Data

- Constructing NTA requires individual or household micro survey data sets.
- A good survey data set has the properties of
  - Extent (richness): it has the variables of interest at a certain level of details.
  - Reliability: the variables are measured without error.
  - Validity: the data set is representative.

Data Problem (An example)

- FIES (64,433 household with 233,225 individuals)
  - Measured for only urban area (Valid?)
  - No single person household (Valid?)
  - No individual level income, only household level (Rich?)
  - No information of income for family owned business (Rich?)
  - Measured for up to 8 household members: discrepancy between the sum of individual and household income (Valid? Rich?)

Extent (Richness): Missing/Change of Variables

- Not measured in the data
  - 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

Reliability: Measurement Error

- Response error
  - Respondents do not know what is required
  - Incentive to understate/overstate
  - Recall bias: related with period of survey
  - Using wrong/different reporting units
- 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

Validity: Censoring

- Selection based on characteristics
- Top/Bottom coding
- Censoring due to the time of survey
  - Duration of unemployment (left and right censoring)
  - Completed years of schooling
- Attrition (Panel data)

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

Units, Real vs. Nominal

- Be careful about the reporting unit
  - Measurement 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

Solution for Missing Variables

- Ignore it; random non-response
- Find other source of data (FIES vs. LFS)
- Impute
  - Based on mean value
  - Based on the value of other peer group (based on characteristics)

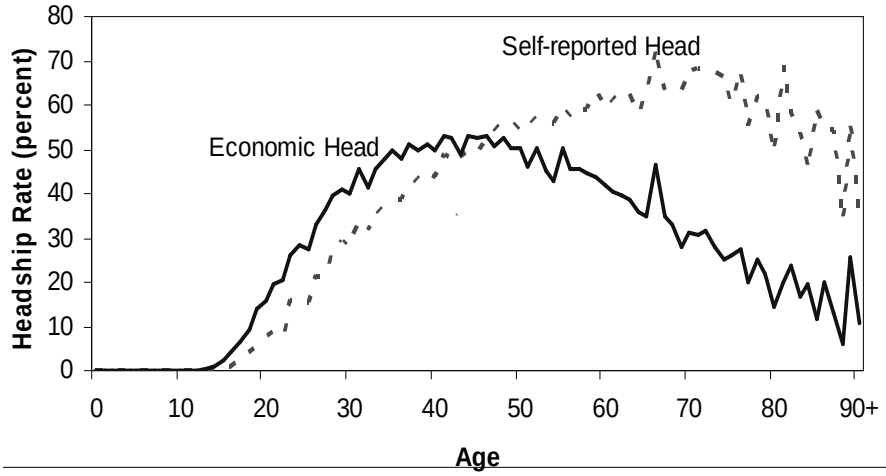
Analyzing Household: Measuring Private Consumption

- Underestimation: e.g. British FES
  - Using aggregate control mitigate the problem.
- Home produced items: both income and consumption.
- Allocation across individuals is difficult
- Estimating some profiles, such as health expenditure are also difficult in part due to various source of financing.

Analyzing Household: Income

- "All of the difficulties of measuring consumption apply with greater force to the measurement of income" (Deaton, p. 29).
  - Need detailed information on "transactions" (inflow and outflow): an enormous task
  - Incentive to understate: using aggregate control mitigate the problem if understatement is not systematic by age.
  - Some surveys did not attempt to collect information on asset income (e.g. NSS of India) → Exercise 1
- Allocating self-employment income across individuals is difficult.

Headship (Thailand, 1996)



Data Cleaning

- Case by case
- 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.

Weighting and Clustering

- 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]

Weighting and Clustering (cont'd)

- 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)

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.

Discussion

- Data type/quality varies across countries.
- Estimation method could vary across countries depending on data.
- However, some standard measure could be applied.
  - Definition → Specification → Estimation using weight → Smoothing → Macro control → Present your work!
  - If some component vary substantially by age, then it is estimated separately (education, health, etc)

The End
(Exercise Part 2)