Data and Empirical Strategy Section

INT 93

Michael Topper

Progress Reports

Main Comments:

- Discretizing Variables
  - Many variables with Sex = 1 or Sex = 2. These need to be put into 2 categories: Male and Female.
- Inconsistent number of observations
  - The observations in the results tables should match the summary statistics
  - Read the codebooks—lots of missing codes that differ
  - Get one, consistent sample to speak about
- Names
  - Give informative names in your table. DO NOT use STATA names.
- Decimal Places
  - We do not need to know eight decimal places—keep to 3 max.
- Table formatting
  - Use the papers we've talked about in class as a template.

Data Section

Overview

- Sample sample sample
  - 1st and most important sentence
  - Remember, the main sample should generally not change
  - Should be a consistent set of individuals
- Do not get creative here
  - This section SHOULD be boring
  - Only for those most interested
- Candidate for writing 1st
  - Writing data section often helps you refine your sample

Components:

Components

1. Describe name, source, and period.
2. Describe central variables.
    - Panel/cross section/time series?
    - Extreme emphasis on important variables
        - Main outcomes
3. Discuss limitations
    - Missing variables/observations
    - Highlight critical limitations
    - Discuss minor limitations in footnote
4. Summary statistics
    - Table with means/standard deviations of important variables
    - "Descriptive" statistics
    - Important subgroups

Your Sample:

How your sample should "look"

- Consistent number of observations.
  - We want to know who the sample is, and we do not want this changing table-to-table
  - Describe who these are, and why you had to use these
  - Missing data problems?
- Do not switch between samples, unless you have a very clear reason to do so.
  - Switching samples confuses readers
- Look at your data and read your code books for missing codes!
  - Sometimes missings are coded differently (ie 999999, ., NA)

First Sentence:

The first sentence(s) is(are) critical for the audience to understand the sample. Every data section should (usually) begin the same.

Examples from papers:

- Our primary data source is the restricted-access 2000-2016 California Birth Statistical Master Files. These data cover the universe of California births during this period and come from birth certificate information that the parents and medical provider fill out at the time of birth. (Royer, Jacobson, Kogelnik 2020)
- Our main source of data are comprehensive administrative records covering 64,209 felony charges filed in Bexar County District Courts between 2005 and 2013 (Agan, Freedman, Owens, 2018)

Central Variables

When describing central variables, focus on the outcome variable(s), or important controls.

Keep in Mind..

- You likely manipulated your outcome
  - Explain how
  - Note if this changes interpretation (e.g. log(income))
  - Does this transformation cause issues? (e.g. missing data)
- Do not get too technical:
  - Bad: Using STATA, I subtracted `dispatch_time` from `call_entry_time` to create `time_to_dispatch` variable.
  - Good: Dispatch time is defined as the difference between dispatch time and the time of the call.

Limitations

While limitations are important to convey, do not undersell yourself!

Keep in Mind..

- Every study has limitations
  - Does later analysis address this limitation?
  - Is the limitation critical to your analysis?
- For certain data sets with limitations everyone knows of:
  - Use other literature as your guide. They had to deal with these problems first!
    - Remember to cite!

Summary Statistics

This can either be it's own subsection, or you can weave it into your writing. Both work well!

Keep in Mind..

- Used for the reader to envision the experimentees
- Summary statistics table should have:
  - Mean/SD of every important variable
    - Outcomes/controls etc.
- Note anything unusual in the table:
  - Large amount of one sex etc.
- Self contain your table and writing!
  - I should know what the table says by only reading the writing and vice-versa

Self-containing

Keep in Mind...

- You cannot simply write "summary statistics are shown in Table 1"
  - Elaborate
  - Make sure the reader never needs to look at the table
  - "The mean of BLANK is X, while the standard deviation is BLANK"
  - For some important variables, it may be great to write a sentence explaining why the numbers look the way they do.

Activity: Reading a Good Data Section

Exercise: Write Summary Statistics