Exploring Data in R

Jason Thomas

November 7th, 2018

Bloomberg Philanthropies

DATA FOR HEALTH INITIATIVE

CDC Foundation
Together our impact is greater

Vital Strategies

The Ohio State University
INSTITUTE FOR POPULATION RESEARCH

In collaboration with the WHO VA Reference Group

Motivation

- No data are prefect, as there are many potential sources of problems
  - design and implementation of the the questionnaire (XLS From)
  - interviewer error; respondent does not understand the question or gives problematic answer
  - mistakes during data entry
- VA algorithms are only as good as the inputs, so it is best to perform data quality checks and ensure the values are sensible

Goals for today

- Learn about tools in `R` for exploring data.
- Develop a strategy for cleaning data
- Get practice with checking for errors and "fixing" them
- Introduction to `CrossVA`

Exploring Data (cont)

When exploring data . . .

- determine the valid range (and type) of values for each variable and compare to the observed range of responses
  - natural boundaries (ages & times must be positive)
  - "choices" worksheet on the XLM Form
  - conditional responses: *the infant had trouble breathing for 2 years* (given the age, the duration must be less than 1 year)
- look at the distribution of all responses and check for outliers or extreme values
- think of common practices that may compromise data quality
  - abbreviations or alternative respresentations (two, 2, TWO, Two)
  - typos (027 instead of 1027; too vs. two)
  - codes for missing data (e.g., 9999)
- sanity checks: given the question being asked, does the value "make sense"?

Exploring Data (cont)

There are some basic commands in `R` that will give us an initial picture of our data. . .

- `summary()` – useful for *continous* variables (many values)
- `table()` – useful for *categorical* variables (only a few values)
  - also helpful when you need to consider two variables at the same time (e.g., conditional relationships: men should not experience problems with giving birth)
- Visual displays
  - `hist()` – (histogram) looking at the distribution
  - `plot(x, y)` – scatterplot for visualizing the relationship between to variables
- We will see many examples of these commands when walking through the `R` script for today.

Cleaning Data

- While exploring the variables in our data, we often see changes that need to be made.
  - Always make changes to a copy of the variable (or a copy of the entire data sets) **NEVER** change the original data.
  - Creating Yes/No, 0/1, or True/False indicators that flag a problem is also a good practice.
- This process invloves 3 steps. . .
  1. Find the "problem" (e.g., negative values for ages, people who were sick for 8,000 years, women with 14 children who are 20 years old)
  2. Create an *index* that identifies the cases with the particular problem (or potential problem).
  3. Create a copy of the variable and use the *index* to assign new values

R Package: CrossVA

- Install `CrossVA()`
- Read an ODK Briefcase export into R
- Example run of CrossVA()
- Brief look under the hood