Ordinal Variables and Missing Data

November 8, 2019

General

- This paper should be an intersection of ordinal variables and missing data. Jeff thinks there is room there for a contribution
- The general idea is that general treatments of missing data (listwise deletion, multiple imputation etc.) lead to different results – depending on the type of variable. How you treat `Don't Know` and `Refused` and how this treatment affects the results depends greatly on the type of variable that you use to impute the missing data. The idea is that you would need to approach things differently depending on whether you use a nominal, interval, or ordinal variable to fill in values for `Don't Know/Refused`. So I am developing a method to specifically treat `Don't Know/Refused` with ordinal variables that improves over current uses that are generic for all types of variables
- This method should be a customized form of multiple imputation that is closer to the data than general multiple imputation. Jeff and Skyler developed affinity scores and hot decking in their BJPS paper. They used the number of exact matches (in the form of other participants) to calculate the affinity score. Instead, I should use a weighted distance solution between ordinal variable categories. I would use the OP model from the blocking paper to weight the distances between the categories in matches (in the form of other participants). In other words, I would use the underlying ordered probit numbers to create the weights. So this would be a specific ordinal variable adjustment of the affinity score building
- Set up chapter structure

Code

- Currently `MCAR` with `prodNA`
  - `prodNA` adds a proportion of NAs
  - My code functions for one variable with NAs, for 10,000 iterations
    - I have saved results for `inc`, `age`, `Dem`, `Rep` for `hd.ord` and `hd.norm`
    - For all 4, `na.omit` performs the best. This isn't surprising, since deleting observations with missing values shouldn't be a problem with MCAR
  - My code does not function When run for several variables with NAs
    - It throws up a replacement error somewhere down the line when run for 10,000 iterations. It's never in the same place, so it must be something random
  - Always the same: # of NAs overall. Not always the same between some runs: # of NAs per column, # of rows with NAs
  - Find out why my code doesn't work for NAs in several variables
    - Fill in the `else` sections so that I don't save anything to dfs
- Test `mice amputate()`
  - It has `MCAR` and `MAR` options, which would be awesome
  - Testing is a bit annoying, since it doesn't work on just one variable
- Increase proportion of NAs

Theory

- Step by step fill in the sections on Missing Data, Deletion, and Imputation
- Rework the introduction
  - Current stuff in there is very broad and nowhere near detailed enough (taken from the Kerwin application)