Common Data Cleaning Challenges and Approaches for the Fragile Families Challenge

Janet Xu and Matt Salganik

6/22/2018

Issue No. 1: Missing Data

As with many datasets, the Fragile Families and Child Well-being Study contains missing data. One thing to keep in mind when dealing with missing data is that the missing data codes contain information about the type of non-response – for example, -9 indicates that the respondent did not participate in that particular wave of the longitudinal study, while -6 means that the survey instrument skipped the question for the respondent because it does not apply or because it is already known based on prior information.

It is also useful to think about types of missing data and mechanisms that may generate them.

Some approaches to missing data include:

- Imputation:
    - Mean, median, or mode substitution
    - Hot-deck imputation , including Last Observation Carried Forward (LOCF)
    - Regression imputation
    - Linear interpolation
    - K Nearest Neighbors
    - Probabilistic PCA
- Listwise Deletion (but keep in mind that if you use a lot of variables, you might end up with zero observations)

(For the purposes of making a submission in the time we have, we would recommend mean or median imputation.)

Issue No. 2: Measures with unordered categories are coded as ordered categories

Some measures, such as race/ethnicity questions, might be stored as ordered categories even though they should be unordered. You should check the data type of the variable to make sure and recode it if necessary.

Issue No. 3: Some variables have little to no variation

Variables with little to no variation in the response may give you trouble when model-fitting. Remove these variables as necessary.