Assessing imputation uncertainty NHES 2012

Recai M. Yucel
Nathaniel Schenker
Trivellore Raghunathan

January 25, 2018

Outline

Work in progress!

1. 2012 National Household Education Survey
2. Missing data due to item nonresponse
   - Rates of missingness
   - What impacts missingness?
3. Summary of NHES imputation routines
4. Assessing the imputation uncertainty using MI (with some empirical findings)

National Household Education Survey (NHES)

- The NHES consists of two topical surveys – the Early Childhood Program Participation (ECPP) Survey and the Parent and Family Involvement in Education (PFI) Survey
- The ECPP survey has a target population of children age 6 or younger who are not yet in kindergarten
- The PFI survey has a target population of children and youth age 20 or younger who are enrolled in kindergarten through 12th grade in a public or private school or who are being homeschooled for the equivalent grades
- NHES:2012 used an addressed-based sample covering the 50 states and DC, and proceeded as a two-stage, stratified sample. The first stage sampled the addresses, and the second stage selected the eligible child
- Around 73% unit response rates

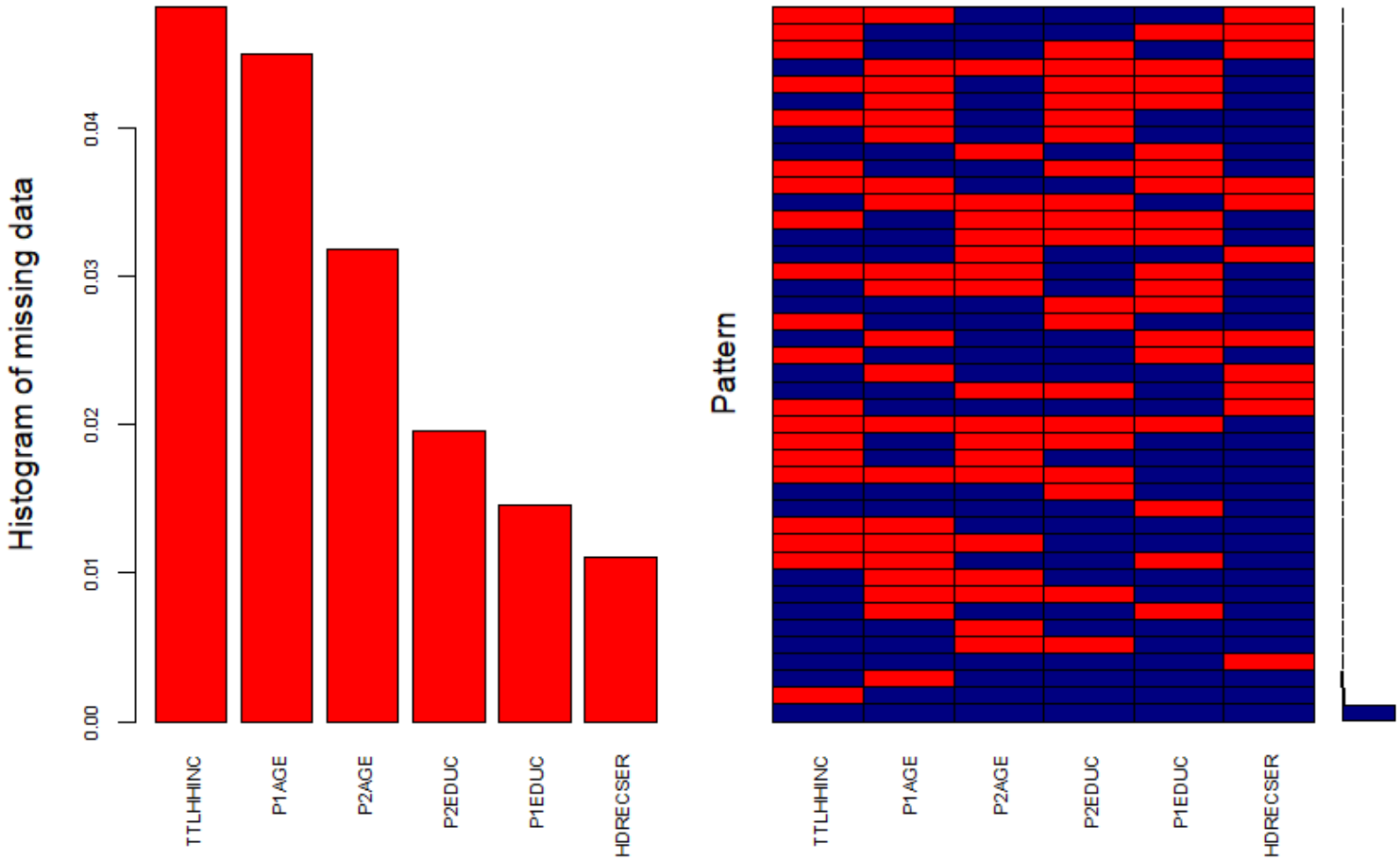
Missing data

- Similar to most surveys, NHES 2012 also has incompletely-observed survey items
- Median item response rates for both PFI (114 items for enrolled students, and 92 items for homeschooled) and ECPP (140 items) surveys were 96.4% and 97.9%, respectively

Missing data: example

- For this presentation, consider an analysis involving six variables: Age (783 out of 17563 respondents in PFI), Education (256 parent 1, 344 parent 2), Total Household Income (846 missing out of 17563) and indicator for receiving special health services
- 15663 cases from PFI module have complete data (1900 cases have at least one item missing) in this subset of variables

Example missing data pattern



Speculating factors causing missing data

- Some of the key factors influencing "missingness":
  - For missingness on income, parents' education level, grade level are key factors
  - For some other items subject to missingness, race and socio-economic factors also play a role
  - All estimated using design-based logistic regression on the relevant missingness indicator (R survey package by Lumley)

National Household Education Survey Imputation Routines

For various practical and operational reasons, missing values across the survey items were imputed using four successive imputation methods:

- Logic-based imputation
- Weighted random imputation
- Sequential hot deck imputation
- Manual imputation (mean/mode imputation if hot deck can not performed)

These routines were implemented in STATA for 2012 NHES, then SAS routines were developed for 2016 NHES. All imputation procedures are followed with a comprehensive post-imputation edits and imputation flags are added in the public datasets.

NHES Imputation Routines: Logic-based imputation

- In logic-based imputation, items for which a respondent is missing data are imputed using other data available for the same respondent.
- To impute a value to missing gate questions based on the presence of "yes" or valid data in follow-up items. Gate questions are defined as survey questions whose answers determine the subsequent routing of the respondent through the survey instrument.

NHES Imputation Routines: Weighted random imputation

- Imputation proceeds based on the empirical probability distribution of the variable
- For example, if 15% of the respondents report "high school diploma" on the item for highest education level attained, then "high school diploma" is imputed for a randomly selected 15% of the item nonrespondents

NHES Imputation Routines: Hot deck imputation (ctd.)

- Cross of boundary variables (which must be observed for all, missing ones are typically imputed using random imputation) are used to define imputation cells
- The algorithm samples from a pool of donor observations in these cells (same observation can not be used as imputation more than 5 times)
- The purpose of dividing the sample into imputation cells is to ensure that values are imputed from donor respondents that are sufficiently similar to each recipient respondent in terms of key "boundary" characteristics
- The variables were chosen because they are characteristics of households, respondents, or children that are likely to be associated with differences in item response propensities, such as parent(s) educational attainment; or are key variables in questionnaire paths and skip patterns, such as the child's grade and enrollment status

NHES Imputation Routines: Hot deck imputation (ctd.)

Donor rules are enforced to reduce the potential bias:

- an individual case may be used a maximum of five times as a donor for a particular variable. This is designed to reduce the likelihood that a single donor has a disproportionate effect on overall estimates
- Second, donors may have boundary variables that are imputed using weighted random imputation
- Donors are not eligible to impute a value for a specific variable if that variable was imputed, including logic-based imputation

NHES Imputation Routines: Manual imputation

Applied when no donors are available in hot deck imputation (not implemented for more than 10 cases per variable, on average)

- Collapsing boundary variables to produce more donors for imputation cells
- Reduced number of boundary variables
- Mean/mode imputation

  "Mean/mode imputation" refers to using the pre-imputation distribution of the item to assign an imputed value. For categorical variables, the modal value will be imputed. For continuous variables, the mean value will be imputed. This will either be the overall mean/mode, or that of a subgroup, depending on the variable.

Imputation Uncertainty

- Key problem with a single imputation (regardless of the underlying imputation methodology) is the underestimation of the uncertainty in the post imputation analyses unless care is taken to reflect the variation underlying the distribution of missing data (or uncertainty implied by the imputation process)
- As the surveys put forward by the federal agencies used by many entities, this is an important problem which has been extensively discussed in the missing-data literature

Incorporating imputation uncertainty

- **Resampling-based approaches** (Rao and Shao, 1992; Efron, 1994; Rao and Sitter, 1995, Kim and Fuller, 2004; Fuller and Kim, 2005)
- **Linearization approach** (Clayton et al. 1998; Shao and Steel, 1999; Robins and Wang, 200; Kim and Rao, 2009)
- **Multiple imputation (MI)** (Rubin, 1976, 1987 coined the term MI inference, initially he named it as repeated - imputation inference)

Incorporating imputation uncertainty using MI

- The key idea of MI is to generate multiple (say M) plausible versions of missing data, analyze each data by standard complete-data methods and then combine the results
- Consider for example, one wants to make inferences about a regression coefficient β
- We would obtain estimates of β across the imputed datasets: $\beta_1, \ldots, \beta_m$ along with its standard errors: $s_1, s_2, \ldots, s_m$
- To obtain an overall point estimate, we then simply average over the estimates from the separate imputed datasets:

$$\hat{\beta} = \sum_{m=1}^{M} \hat{\beta}_m$$

Incorporating imputation uncertainty using MI (ctd.)

- A final variance estimate $Var(\hat{\beta})$ reflects variation within and between imputations:

$$Var(\hat{\beta}) = W + (1 + \frac{1}{M})B,$$

where $W = \frac{1}{M}\sum_{m=1}^{M} s_m^2$, and $B = \frac{1}{m-1}\sum_{m=1}^{M}(\hat{\beta}_m - \hat{\beta})^2$.

- B is essentially a key factor quantifying the variation in the missing data distribution, and ignored under single imputation procedures
- Kim et al. (2006) showed that for certain estimates, this variance can be biased and offered bias-adjustment

Multiple imputation under a hot deck algorithm

- One idea is to repeatedly execute the current hotdeck algorithm (hot deck MI)
- Missing values in income, education, age and indicator for receiving special health services were replaced by five donors selected randomly from plausible pool
- Not much difference is observed between SI and MI in terms of means (in fact, complete-case only analysis is also quite similar):

Table: Means and SEs of selected items from PFI

	Total Income	Special Health services	P2 Education
hotdeck SI	5.96 (0.022)	0.20 (0.0062)	3.61 (0.0261)
hotdeck MI	5.95 (0.0223)	0.20 (0.0066)	3.59 (0.0266)

Multiple imputation under a hot deck algorithm

Now consider some simple multivariate analyses:

- Model 1:

$$\text{logit}(P(\text{special health service})) = \beta_0 + \beta_1 Inc + \beta_2 Edu + \beta_3 Age$$

- Model 2:

$$Income = \beta_0 + \beta_1 Special.health.serv + \beta_2 Education + \beta_3 Age + \epsilon$$

Naïve comparison: MI versus SI (Model 1)

Table: Model 1 estimates– SI versus MI hotdeck

	$\hat{\beta}_0(SE)$	$\hat{\beta}_1(SE)$	$\hat{\beta}_2(SE)$	$\hat{\beta}_3(SE)$
hotdeck SI	-1.857 (0.102)	0.084 (0.016)	0.028 (0.0197)	-0.004 (0.0028)
hotdeck MI	-1.800 (0.103)	0.076 (0.017)	0.029 (0.0205)	-0.004 (0.0030)
r [1]	0.0101	0.0089	0.0465	0.0408

[1] estimated relative increase in the variances due to missing data (or due to imputation)

Naïve comparison: MI versus SI (Model 2)

Table: Model 2 estimates– SI versus MI hotdeck

	$\hat{\beta}_0(SE)$	$\hat{\beta}_1(SE)$	$\hat{\beta}_2(SE)$	$\hat{\beta}_3(SE)$
hotdeck SI	4.450 (0.034)	-0.087 (0.019)	0.52 (0.008)	-0.013 (0.001)
hotdeck MI	3.620 (0.131)	0.424 (0.095)	0.548 (0.018)	-0.01 (0.003)
r	0.02	0.02	0.06	0.07

Multiple imputation under parametric imputation model

- Assume a multivariate normal model as a rough approximation to the data-generation mechanism (variable-by-variable approach is better for surveys similar to this) (Schafer, 2016 : norm2 R package; Raghunathan et al (2016): IVEware; VanBuuren et al (2016): R package mice, White et al STATA package ice)
- More complex data structures: R packages pan (Schafer and Yucel, 2002); jomo (Carpenter et al 2011); shrimp (Yucel, Schenker and Raghunathan, 2017)
- Higher imputation-to-imputation variation leads to a bit larger SEs

MI under MVN

Table: Model 2 estimates– SI versus MI MVN

	$\hat{\beta}_0(SE)$	$\hat{\beta}_1(SE)$	$\hat{\beta}_2(SE)$	$\hat{\beta}_3(SE)$
hotdeck SI	4.450 (0.034)	-0.087 (0.019)	0.52 (0.008)	-0.013 (0.001)
MVN MI	3.630 (0.142)	0.422 (0.101)	0.551 (0.023)	-0.02 (0.003)
r	0.03	0.02	0.07	0.08

Notes

- This comparison is **naïve** in the sense that one can use a single imputation and still correct for the imputation uncertainty (see Kim's papers)
- However, public-use data files that include imputation need to make note of this; or MI versions should also be released as done in NHIS (Nat and Raghu's work) and NHANES (Schafer) with cautionary notes on combining inferences