Approaches to imputing missing data in complex survey data

Christine Wells, Ph.D.

IDRE UCLA Statistical Consulting Group

July 27, 2018

Three types of missing data with item non-response

- Missing completely at random (MCAR)
  - Not related to observed values, unobserved values, or the value of the missing datum itself
- Missing at random (MAR)
  - Not related to the (unobserved) value of the datum, but related to the value of observed variable(s)
- Missing not at random (MNAR)
  - The value of the missing datum is the reason it is missing
- Each variable can have its own type of missing data mechanism; all three can be present in a given dataset
- Most imputation techniques only appropriate for MCAR and MAR data

Different approaches to imputing missing complex survey data

- Stata: multiple imputation (mi) (and possibly full information maximum likelihood (FIML))
- SAS: Four types of hotdeck imputation
  - Fully efficient fractional imputation (FEFI)
  - 2 stage FEFI
  - Fractional hotdeck
  - Hotdeck
- SUDAAN: Four methods
  - Cox-Iannacchione weighted sequential hotdeck (WSHD)
  - Cell mean imputation
  - Linear regression imputation
  - Logistic regression imputation

Handling imputation variation

- Stata
  - Multiple complete datasets
- SAS
  - Imputation-adjusted replicate weights (not with hotdeck)
    - BRR (Fay), Jackknife, Bootstrap
  - Multiple imputation (only with hotdeck)
- SUDAAN
  - Multiple versions of imputed variable (WSHD only)

Available methods with SAS's proc surveyimpute 1

- Hotdeck
  - Observed values from donor replace the missing values
  - Imputation-adjusted replicate weights cannot be created with this method, but multiple donors can be used, leading to multiple complete datasets
- Fractional hotdeck
  - Variation on hotdeck in which multiple donors are used
  - The sum of the fractional weights equals the weight for the non-respondent

Available methods with SAS's proc surveyimpute 2

- FEFI (default)
  - Variation on fractional hotdeck in which all observed values in an imputation cell are used as donors
- 2-stage FEFI
  - Particularly useful for continuous variables
  - The first stage is FEFI
  - The second stage uses imputation cells to determine imputed values
  - Imputation adjusted replicate weights are computed by repeating the first and second stage imputation in every replicate sample independently

General comments about SAS's proc surveyimpute

- None of the procedures are model-based
- Donor selection techniques include
  - Simple Random Sampling with or without replacement
  - Probability proportional to weight
  - Approximate Bayesian bootstrap
- All methods handle both continuous and binary variables
- Survey design elements can be incorporated into most methods
- All methods have a way to account for the imputation variance

Available methods with SUDAAN's proc impute 1

- Weighted Sequential Hotdeck (WSHD) (default)
  - For both continuous and binary variables
  - Uses imputation classes and multiple donors
  - Sampling weight is used to limit the number of times a donor is used
  - Currently the only method that allows for the creation of multiple versions of the same variable
- Cell mean imputation
  - For continuous variables only
  - Missing values replaced with mean of imputation class
  - Uses the same methodology as proc descript
  - Uses an explicit imputation model

Available methods with SUDAAN's proc impute 2

- Linear regression imputation
  - For continuous variables only
  - Fit a separate model for each continuous variable to be imputed
  - The same (complete) cases are used for each imputation model
  - The missing values are replaced with the predicted values
  - Uses an explicit imputation model
- Logistic regression imputation
  - For binary variables only
  - Similar to linear regression imputation
  - Predicted values are compared to a random number: 1 if x ge p; 0 otherwise
  - Uses an explicit imputation model

Pros of the mi approach

- Obviously accounts for the imputation variance
- Many researchers are familiar with it (at least with non-weighted data)
- Handles many types of outcomes (Stata)
- Can choose between multivariate normal (MVN) or imputation by chained equations (ICE) (Stata)
- Can use the multiply imputed datasets with other software packages

Cons of the mi approach

- No strong theoretical basis for ICE, but there is for MVN
- The imputation model may be different for different subpopulations
- The publicly-available dataset may not contain good predictors of missingness
- Multiple copies of a large dataset can create processing and/or storage problems

Pros of the hotdeck approach

- Does not require an explicit imputation model
- Only plausible values can replace missing values
- Preserves the distribution of the variable
- Minimal increase in the size of the dataset (just adding some variables)
- Lots of interest from big survey research organizations

Cons of the hotdeck approach

- No strong theoretical basis for hotdeck
- Not often used with non-weighted data
- May not have many (or any) donor cases for some subpopulations
- Can be problematic if the imputation variance is not taken into account

An example: Continuous NHANES 2015-2016 data

- dmqmiliz: Served active duty in US Armed Forces
  - binary
  - 3822 missing out of 9971 cases (38.33%)
- paq710: Hours watch TV or videos past 30 days
  - ordinal treated as continuous
  - 63 missing out of 9255 cases (including refused and don't know) (0.68%)

An example: Stata mi and analysis code

```
mi set flong
mi misstable summarize usmilitary paq710

gen descode = sdmvstra*10+sdmvpsu

mi register imputed usmilitary paq710
mi register regular riagendr ridageyr dmdfmsiz wtint2yr descode

mi impute chained (logit) ///
usmilitary (regress) ///
paq710 = riagendr ridageyr dmdfmsiz wtint2yr i.descode, ///
add(20) rseed(44587996)

mi svyset sdmvpsu [pw = wtint2yr], strata(sdmvstra)
mi estimate: svy: regress paq710 usmilitary riagendr ridageyr dmdfmsiz
```

An example: SAS hotdeck code - impute

```
proc surveyimpute data = nhanes_15_16 method = fefi (maxemiter = 300)
varmethod = jackknife;
weight wtint2yr;
strata sdmvstra;
cluster sdmvpsu;
class usmilitary paq710;
id seqn;
var usmilitary paq710;
output out = sas_2stage fractionalweights = frac_wts
outjkcoefs = sas_jkcoefs;
run;
```

An example: SAS hotdeck code - analysis

```
proc surveyreg data = sas_2stage varmethod = jackknife;
weight impwt;
repweights imprepwt: / jkcoefs = sas_jkcoefs;
model paq710 = usmilitary riagendr ridageyr dmdfmsiz;
run;
```

An example: SUDAAN hotdeck code - impute

```
proc impute data = nhanes_15_16 seed = 44587996 notsorted
method = wshd;
weight wtint2yr;
impvar usmilitary paq710;
impid seqn;
impname usmilitary = "usmilitary_ir" paq710 = "paq710_ir";
impby riagendr;
idvar seqn;
output / impute = default filename = wshd filetype = sas replace;
print / donorstat=default means=default;
run;
```

An example: SUDAAN hotdeck code - analysis

```
proc sort data = nhanes_15_16;
by seqn; run;

proc sort data = wshd;
by seqn; run;

data sudaan_merged;
merge nhanes_15_16 wshd;
by seqn; run;

proc sort data = sudaan_merged;
by sdmvstra sdmvpsu; run;

proc regress data = sudaan_merged filetype = sas design = wr;
weight wtint2yr;
nest sdmvstra sdmvpsu;
model pag710_jr = usmilitary_jr riagendr ridageyr dmdfmsiz; run;
```

Results - Coefficients

Term	Listwise	Stata	SAS	SUDAAN
Constant	1.612	1.966	1.968	1.976
usmilitary	0.416	0.466	0.559	0.445
riagendr	-0.018	-0.029	-0.017	-0.014
riageyr	0.020	0.013	0.013	0.013
dmdfmsize	-0.081	-0.067	-0.067	-0.066
Obs used	6135	9971	9971	9971
Population	244,344,506	316,481,044	316,481,044	316,481,044

Results - Standard errors

Term	Listwise	Stata	SAS	SUDAAN
Constant	0.156	0.117	0.117	0.112
usmilitary	0.119	0.112	0.109	0.088
riagendr	0.056	0.048	0.047	0.047
riageyr	0.002	0.001	0.001	0.001
dmdfmsize	0.019	0.017	0.017	0.016

These are not your only options

- R: many different packages
- Mplus: full information maximum likelihood (FIML)
- Stata: may be able to use the -sem- command and hence FIML
- IVEware: multiple imputation (mi model tied to analysis model)

Results - Coefficients

Term	Stata - FIML	Stata - mi	SAS	SUDAAN
Constant	1.946	1.966	1.968	1.976
usmilitary	0.409	0.466	0.559	0.445
riagendr	-0.025	-0.029	-0.017	-0.014
riageyr	0.014	0.013	0.013	0.013
dmdfmsize	-0.066	-0.067	-0.067	-0.066
Obs used	9971	9971	9971	9971
Population	316,481,044	316,481,044	316,481,044	316,481,044

Results - Standard errors

Term	Stata - FIML	Stata - mi	SAS	SUDAAN
Constant	0.121	0.117	0.117	0.112
usmilitary	0.115	0.112	0.109	0.088
riagendr	0.050	0.048	0.047	0.047
riageyr	0.001	0.001	0.001	0.001
dmdfmsize	0.017	0.017	0.017	0.016

Conclusions

- No clear consensus in the literature regarding the best way to handle missing data in complex survey datasets
- Better for determining associations between variables than precise parameter estimates
- Must be able to reasonably assume MCAR or MAR; not many options for MNAR data
- The quality of model-based imputations may depend on the quality of the variables in the dataset
- Lots of advances in this area, especially from the Census Bureau

References 1

Andridge, R. R. and Little, R. J. A. (2009). The Use of Sample Weights in Hot Deck Imputation. Journal of Official Statistics: 25(1): 21-36.

Andridge, R. R. and Little, R. J. A. (2010). A Review of Hot Deck Imputation for Survey Non-response. International Statistical Review: 78(1): 40-64.

Chen, Y. and Shao, J. (1999). Inference with Survey Data Imputed by Hot Deck When Imputed Values are Non-identifiable. Statistica Sinica 9, 361-384.

Cox, B. (1980). The Weighted Sequential Hot Deck Imputation Procedure

References 2

Heeringa, S. G., West, B. T. and Berglund, P. A. (2017). Applied Survey Data Analysis, Second Edition. New York: CRC Press.

Heeringa, S. G., West, B. T., Berglund, P. A., Mellipilan, E. R. and Portier, K. (2015). Attributable Fraction Estimation from Complex Sample Survey Data. Annals of Epidemiology. 25: 174-178.

Iannacchione, V. G. (1982). Weighted Sequential Hot Deck Imputation Macros. Seventh Annual SAS User's Group International Conference, San Francisco, CA, February, 1982.

Korn, E. L. and Graubard, B. I. (1999). Analysis of Health Surveys. New York: Wiley.

Contact information

Christine Wells, Ph.D.
IDRE UCLA Statistical Consulting Group
crwells@ucla.edu