

Workshop on Integrative Data Analysis

Using Imputation to Enhance Multiple Iterations of HINTS Data

Mandi Yu, Ph.D.
National Cancer Institute

HINTS Users Conference 2014



Outline

- Missing Data Principle
 - Missing data mechanisms
 - Basic principles for imputing missing data
- Applications
 - Impute missing data for income
 - Generate synthetic data for variables for which data are not collected at certain iterations - Use the internet to look for health information for self
- Concluding Remarks

Item Missing Data

ID	Design Variables	X	Y	Z
1	O	O	O	O
2	O	O	O	M
3	O	O	M	M
4	O	O	M	O
	O	O	O	O
⋮	O	O	O	O
	O	O	M	O
n	O	O	O	M

O: Observed
M: Missing

Complex Survey Design Variables:

- Stratum
- Cluster or Replicate Weights
- Survey Weights

Missing Data Mechanisms

- Missing Completely At Random (MCAR)
 - Missingness doesn't depend on anything.
 - Unbiased using deletion approach, but less efficient
- Missing At Random (MAR)
 - Missingness doesn't depend on the missing data, but can depend on the observed data
 - Unbiased if MAR is accounted for properly
- Not Missing At Random (NMAR)
 - Missingness depends on the missing data
 - Need to model missingness

How to determine the missing mechanism

- MCAR test for multivariate normal data
 - *Little R.J., JASA 1988*
- Unverifiable for MAR and NMAR because we don't know the missing data
 - Based on theoretical and/or substantive knowledge
- MAR works reasonably well for most applications

Imputation Under MAR

- Fill in missing values with plausible values using an imputation model
- Create a MAR situation by modeling covariates that are predictive to
 - Outcome that is subject to missing (M1)
 - Probability of missingness in outcome (M2)
- M1 is more useful than M2 in bias reduction
 - *Little, R.J. and Vartivarian, S., Survey Methodology, 2005)*

Single Imputation

- One imputed value for each missing observation
- Treat imputed values as if they were the observed, thus standard errors tend to be underestimated
- Remedy: Resampling procedures to incorporate imputation uncertainty in variance estimation
 - *Bootstrap* (Efron, *The Annals of Statistics* 1979)
 - *Jackknife* (Miller, *Biometrika* 1974)
 - Adjusted versions for simple random samples

Traditional Single Imputation Methods

- Mean substitution
 - Mean of observed data for all missing data
- Regression substitution
 - Predictive value from a regression model
- Stochastic regression substitution
 - Add random error to predictive value
- Hot deck (vs. Cold deck)

Multiple Imputation

- Generate imputed values
 - Frequentist approach: predict missing values, then add random error drawn from its residual distribution
 - Bayesian approach: randomly draw from the posterior predictive distribution of the variable
 - Estimate values explicitly or through MCMC
- Repeat multiple times
- Each imputed data is analyzed using standard statistical procedures separately
- Combine multiple estimates using a simple rule₉

Combining Rule

- *Rubin D., “Multiple Imputation for Nonresponse in Surveys”, Wiley & Sons 1987, Chapter 3*

$$\text{Point: } \bar{q} = \sum_{i=1}^M q_i / M$$

where, q_i is the estimate from analyzing each imputed data d_i ,
 $i=1,2,\dots,M$, and M is the number of imputation

$$\text{Variance: } T = \bar{u} + \left(1 + \frac{1}{M}\right) * b$$

where, \bar{u} is the average of within imputation variance
 b is the between imputation variance.

Validity of Multiple Imputation

- Depends on how imputation is carried out
- Adequately fit predictive model
- Model reflects MAR assumption
- “Congenial” to the analytic model: model assumptions are compatible
 - *Meng X.L., Statistical Science 1994*

Validity of Multiple Imputation cont'

- Assumptions are compatible
 - More relaxed assumptions → Less efficient
 - More strict assumptions → More efficient
- Assumption are not compatible
 - Omit an important term → biased (attenuation)
 - Include an unrelated term → unbiased but less efficient
- Imputation model need to be general to incorporate known and unknown statistical terms that may be included in an analytic model

Application 1

Imputing Item Missing Income Data HINTS4 Cycle 1

Item Missing Data at a Glance

Background Variables	Missing Obs (N)	Unwgted Missing Rate (%)
Age	68	1.7
Occupation	160	4.0
Marital Status	111	2.8
Education	85	2.1
Race/Ethnicity	220	5.6
Rent or Own Home	89	2.2
Comfortable Spk English	155	3.9
U.S. Born	47	1.2
Income	401	10.1

Weighted Sequential Hot Deck Imputation

- *Cox, B.G., ASA Proceedings 1980, Cox, B.G. and Folsom, ASA Proceedings 1981*
- Substitute the missing value using response from a donor who is similar to the recipient
- Means estimated using imputed data match weighted means using observed data in expectation
- *Not depends on the distribution of outcome, thus less sensitive to model failure*
- Variations of hot deck: random hot deck, predictive mean, predictive propensity, etc.

Select Covariates

- Select important covariates associated with
 - Income and/or Probability of missing income
 - Race/ethn, Education, Renter/Owner, Speak English, Nativity
 - Implemented in SUDAAN Hotdeck Procedure
- Records with missing data on covariates are not imputed
- Have the option for multiple imputation which is not ‘proper’ because the same donor pool is repeatedly used and variance is underestimated
- Modified by adding a Bayesian Bootstrap procedure before each imputation

Application 2

Data Enhancement: Synthetic Data for Variables not Collected in a Iteration of HINTS

Motivation

- Concern over response burden and data quality
- Not all items are asked in every iteration
- Limit the usefulness of the data for analyzing temporal trends
- A solution is to use imputation method to recover such 'missing' information based on reasonable model assumptions

Item Comparability

- Pattern of item comparability across all iterations
- Same inference population
- Survey designs vary slightly: sampling frame, design factors, and data collection mode

	Design Variables	X	Y	Z	T	U	V	W
Iteration 1 (2003)	O	O	O	O	M	O	M	M
Iteration 2 (2005)	O	O	O	M	O	M	O	M
Iteration 3 (2007)	O	O	M	O	O	M	M	O
Iteration 4 (2012)	O	O	O	O	O	O	O	O

↓ { { {
 4's 3's 2's

O: Observed; M: Not Asked and Treated as Missing

Example: Have you used the Internet to look for health or medical info. for self in the past 12 months?

- Not asked in 2007
- Large gap of 7 yrs btwn iteration 2 and 4
- Basic idea is to
 - Stack data from all iterations to create a concatenated data set
 - Turn to a typical missing data problem
 - Treat the 2007 data as missing
 - Fill in with multiple imputes

	Design Variables	Year	X	Y
Iteration 1 (2003)	0	2003	0	0
Iteration 2 (2005)	0	2005	0	0
Iteration 3 (2007)	0	2007	0	M
Iteration 4 (2012)	0	2012	0	0

Question: Have used the Internet to look for health or medical info. for self in the past 12 months?

- Missing At Random (MAR)?
- Missing data may depend on time, thus making MAR is a strong assumption
- In this illustration, we hope to use the correlations between X and Y , and Time and X to recover the not otherwise missing information on the correlation between Time and Y .

Distribution of Outcome

Year of Data Collection	N		% of Look for Health Info. for Self	
	Use internet	Missing	Point	S.E.
HINTS 1 (2003)	3,982	8	50.6	1.0
HINTS 2 (2005)	3,244	95	56.9	1.3
HINTS 3 (2008)	5,078	5,078	-	-
HINTS 4 (2011)	2,914	14	77.6	1.3

Note: Complex Survey Design Features are Incorporated in all Estimates.

Covariates

Demographics

Age
 Gender
 Race/Ethnicity
 Interview in Spanish
 Marital Status
 Rural

Socioeconomics

Education
 Employment Status
 Household Income

Internet Use

Internet Use
 Use The Internet For Medication Purchase
 Use The Internet For Participating In Online Health Group
 Use The Internet For Communications With The Doctors
 Use The Internet To Look For Health and Medical Information For Others

Health Status

Overall Health Status
 BMI
 Ever Had Cancer
 Family Member Ever Had Cancer.

Health Care and Cancer Communication

Health Insurance Coverage
 Seek Health Care For Self
 Seek Cancer Information

Imputation Method

- Most covariates have missing data
 - Most ~ 1-5%, Max 14% for income
- Item missing data are simultaneously imputed together with the iteration missing data
- Sequential regression multivariate imputation (SRMI)
 - Raghunathan T.E., Lepkowski J.M. et al, Survey Methodology 2001
 - IVEware in SAS, MICE in R, ICE in Stata

Sequential regression multivariate imputation

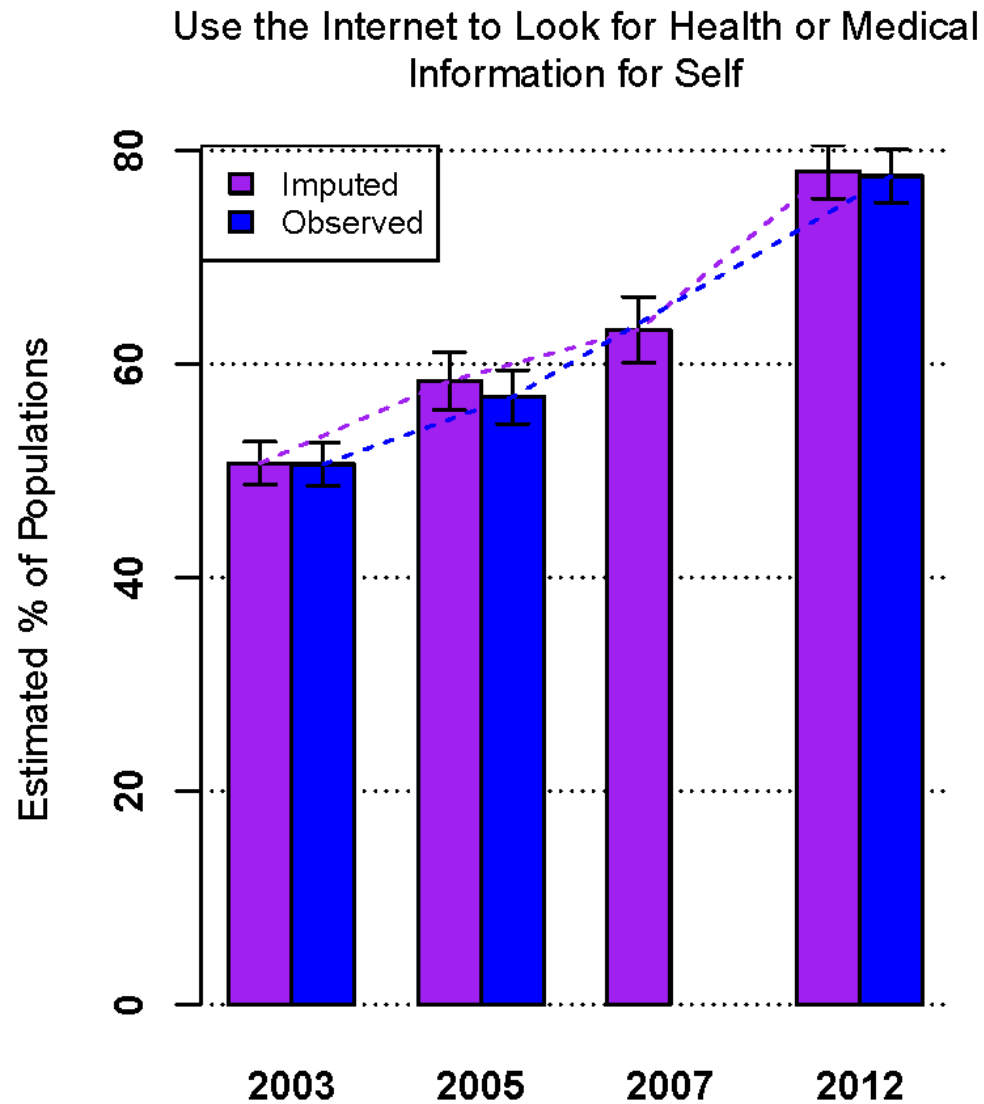
- Different conditional multivariate regression model depends on the type of outcome
 - Continuous: linear model based on normality
 - Binary: logistic model
 - Categorical: multinomial model
 - Count: Poisson model
- Handle skip pattern and higher order terms
 - For example, internet activity related items are only imputed for those who use the Internet
 - Interaction terms

Model Validation

- Simulated data using data from 2003, 2005, and 2012
- Delete outcome data for 2005
- Build imputation model to impute 2005 data
- Compare to the original 2005 estimates
- Similar estimates suggests valid imputation models;
Large differences warrants further model improvements

Year	Actual Data (%)		Multiply Imputed Data (%)			
	Estimate	S.E.	Estimate	Within Var.	Btwn Var.	Combined S.E.
2003	50.6	1.0	50.7	0.9	0.0	1.0
2005	56.9	1.3	62.7	2.2	1.9	2.1
2012	77.6	1.3	78.0	1.7	0.0	1.3

Final Analysis Results



Thoughts on Future Survey Planning

	Design Variables	Year	X	Y
Iteration 1 (2003)	0	2003	0	0
Iteration 2 (2005)	0	2005	0	0
Iteration 3 (2007)	0	2007	0	0
	0		0	M
Iteration 4 (2012)	0	2012	0	0

A random subsample (10-20% of sample)

- This current approach is limited by the lack of information on the correlation between time and Y for 2007
- Improved by building a bridge btwn Time and Y by measuring Y on a random sample in 2007
- Some cost in precision, but large gain in data availability.

Concluding Remarks

- Imputation method is a flexible tool for enhancing the data usability by recovering missing information
- Key issue is to model reasons for missing data, if not ignorable.
- “General” model assumptions to protect against assumption failure
- Model diagnosis also apply to imputation to ensure an adequate fit

Questions and Suggestions

Mandi Yu
Surveillance Research Program
Division of Cancer Control and Population Sciences
National Cancer Institute
yum3@mail.nih.gov

More details on SRMI approach

SRMI

- Sort variables in increasing order by the amount of missing data
- Each variable is sequentially imputed
 - Step 1: impute 1st variable using fully observed sample
 - Step 2: impute 2nd variable on fully observed plus imputed sample in Step 1
 - Step 3: cycle through all variables with missing data completes the first iteration
 - Step 4: repeat Step 1-3, but conditional on all variables, each time use updated imputed values
- Typically 5-10 iterations are sufficient

SRMI: Simple Illustration

ID	Design Variables	X	Y	Z
1	0	0	0	0
2	0	0	0	M
3	0	0	M	0
4	0	0	M	M
5	0	0	0	M

For Iteration 1

1. Fit $Y \sim X$, e.g. $E(Y|X) = X\beta + \epsilon$, and generate imputed values (\hat{y}_3, \hat{y}_4) use this model
2. Fit $Z \sim X + \begin{pmatrix} Y \\ \hat{Y} \end{pmatrix}$, generate imputed values $(\hat{z}_2, \hat{z}_4, \hat{z}_5)$ use this 2nd model

For subsequent Iterations

1. Fit $Y \sim X + \begin{pmatrix} Z \\ \hat{Z} \end{pmatrix}$, and generate updated values (\hat{y}_3, \hat{y}_4)
2. Fit $Z \sim X + \begin{pmatrix} Y \\ \hat{Y} \end{pmatrix}$, and generate updated values $(\hat{z}_2, \hat{z}_4, \hat{z}_5)$