General Methods for Missing Data

John M. Abowd
March 2007

Outline

- General principles
- Missing at random
- Weighting procedures
- Imputation procedures
- Hot decks
- Introduction to model-based procedures

General Principles

- Most of today's lecture is taken from *Statistical Analysis with Missing Data*, 2nd edition, Roderick J. A. Little and Donald B. Rubin (New York: John Wiley & Sons, 2002).
- The basic insight is that missing data should be modeled using the same probability and statistical tools that are the basis of all data analysis.
- Missing data are not an anomaly to be swept under the carpet.
- They are an integral part of very analysis.

Missing Data Patterns

- Univariate non-response
- Multivariate non-response
- Monotone
- General
- File matching
- Latent factors, Bayesian parameters

Missing Data Mechanisms

- The complete data are defined as the matrix Y ($n \times K$).
- The pattern of missing data is summarized by a matrix of indicator variables M ($n \times K$).
- The data generating mechanism is summarized by the joint distribution of Y and M.

$$m_{ij} = \begin{cases} 0, \text{if } y_{ij} \text{ is observed} \\ 1, \text{if } y_{ij} \text{ is missing} \end{cases}$$

$$p\left(Y, M \middle| \theta, \phi\right)$$

Missing Completely at Random

- In this case the missing data mechanism does not depend upon the data *Y*.
- This case is called MCAR.

$$p(M|Y,\theta,\phi) = p(M|\phi)$$

Missing at Random

- Partition Y into observed and unobserved parts.
- Missing at random means that the distribution of M depends only on the observed parts of Y.
- Called MAR.

$$Y = \left(Y_{\text{obs}}, Y_{\text{mis}}\right)$$

$$p(M \mid Y, \theta, \phi) = p(M \mid Y_{\text{obs}}, \phi)$$

Not Missing at Random

- If the condition for MAR fails, then we say that the data are not missing at random, NMAR.
- Censoring and more elaborate behavioral models often fall into this category.

The Rubin and Little Taxonomy

- Analysis of the complete records only
- Weighting procedures
- Imputation-based procedures
- Model-based procedures

Analysis of Complete Records Only

- Assumes that the data are MCAR.
- Only appropriate for small amounts of missing data.
- Used to be common in economics, less so in sociology.
- Now very rare.

Weighting Procedures

- Modify the design weights to correct for missing records.
- Provide an item weight (e.g., earnings and income weights in the CPS) that corrects for missing data on that variable.
- See complete case and weighted complete case discussion in Rubin and Little.

Imputation-based Procedures

- Missing values are filled-in and the resulting “Completed” data are analyzed
  - Hot deck
  - Mean imputation
  - Regression imputation
- Some imputation procedures (e.g., Rubin’s multiple imputation) are really model-based procedures.

Imputation Based on Statistical Modeling

- Hot deck: use the data from related cases in the same survey to impute missing items (usually as a group)
- Cold deck: use a fixed probability model to impute the missing items
- Multiple imputation: use the posterior predictive distribution of the missing item, given all the other items, to impute the missing data

Current Population Survey

Census Bureau Imputation Procedures:

- Relational Imputation
- Longitudinal Edit
- Hot Deck Allocation Procedure

“Hot Deck” Allocation

Labor Force Status

- Employed
- Unemployed
- Not in the Labor Force

“Hot Deck” Allocation

	Black	Non-Black
Male		
16 – 24		
25+		ID #0062
Female		
16-24		
25+		

“Hot Deck” Allocation

	Black	Non-Black
Male		
16 – 24	ID #3502	ID #1241
25+	ID #8177	ID #0062
Female		
16-24	ID #9923	ID #5923
25+	ID #4396	ID #2271

CPS Example

- Effects of hot-deck imputation of labor force status.

Public Use Statistics

	Total AXLFSR	No change	Allocated
Total A_LFSR	220,284,576		
Working	131,704,236		
W/job,not at work	4,572,653		
Unemp,looking for work	7,967,976		
Unemp,on layoff	1,371,469		
Not in labor force	74,668,242		
Total A_AGE	220,284,576		
Average A_AGE	44.1		
Std Err A_AGE	0.15		
Total A_SEX	220,284,576		
Male	105,972,746		
Female	114,311,831		

Allocated v. Unallocated

	Total AXLFSR	No change	Allocated
Total A_LFSR	220,284,576	219,529,643	754,933
Working	131,704,236	131,294,888	409,348
W/job,not at work	4,572,653	4,564,589	8,063
Unemp,looking for work	7,967,976	7,919,562	48,414
Unemp,on layoff	1,371,469	1,367,766	3,703
Not in labor force	74,668,242	74,382,838	285,405
Total A_AGE	220,284,576	219,529,643	754,933
Average A_AGE	44.1	44.2	35.2
Std Err A_AGE	0.15	0.15	1.96
Total A_SEX	220,284,576	219,529,643	754,933
Male	105,972,746	105,603,454	369,292
Female	114,311,831	113,926,189	385,641

Model-based Procedures

- A probability model based on $p(Y, M)$ forms the basis for the analysis.
- This probability model is used as the basis for estimation of parameters or effects of interest.
- Some general-purpose model-based procedures are designed to be combined with likelihood functions that are not specified in advance.

Little and Rubin's Principles

- Imputations should be
  - Conditioned on observed variables
  - Multivariate
  - Draws from a predictive distribution
- Single imputation methods do not provide a means to correct standard errors for estimation error.