Multiple Imputation of Missing Data

Raymond Sin-Kwok Wong
University of California-Santa Barbara
September 8-12
Academia Sinica
Taipei, Taiwan
# The Problem

<table>
<thead>
<tr>
<th>SES</th>
<th>Age</th>
<th>Race</th>
<th>Education</th>
<th>Income</th>
</tr>
</thead>
<tbody>
<tr>
<td>?</td>
<td>36</td>
<td>3</td>
<td>12</td>
<td>?</td>
</tr>
<tr>
<td>60</td>
<td>45</td>
<td>2</td>
<td>15</td>
<td>18</td>
</tr>
<tr>
<td>30</td>
<td>25</td>
<td>?</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>25</td>
<td>57</td>
<td>4</td>
<td>17</td>
<td>24</td>
</tr>
<tr>
<td>40</td>
<td>?</td>
<td>3</td>
<td>13</td>
<td>17</td>
</tr>
<tr>
<td>30</td>
<td>29</td>
<td>?</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>50</td>
<td>47</td>
<td>1</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>30</td>
<td>42</td>
<td>3</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>40</td>
<td>?</td>
<td>2</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>20</td>
<td>65</td>
<td>2</td>
<td>?</td>
<td>10</td>
</tr>
<tr>
<td>45</td>
<td>58</td>
<td>3</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>35</td>
<td>62</td>
<td>5</td>
<td>12</td>
<td>22</td>
</tr>
<tr>
<td>?</td>
<td>45</td>
<td>?</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>80</td>
<td>38</td>
<td>2</td>
<td>12</td>
<td>?</td>
</tr>
<tr>
<td>?</td>
<td>70</td>
<td>4</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td>50</td>
<td>18</td>
<td>3</td>
<td>?</td>
<td>16</td>
</tr>
<tr>
<td>60</td>
<td>29</td>
<td>?</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>40</td>
<td>36</td>
<td>5</td>
<td>16</td>
<td>20</td>
</tr>
<tr>
<td>55</td>
<td>49</td>
<td>2</td>
<td>?</td>
<td>15</td>
</tr>
</tbody>
</table>
Missing Data

- Missing data frequently complicates data analysis for scientific investigations, but our statistical methods assume complete data.
- Determining the appropriate analytic approach in the presence of incomplete observations is a major question for data analysis.
- There are three types of concerns that typically arise with missing data:
  - (a) the loss of efficiency;
  - (b) complication in data handling and analysis; and
  - (c) bias due to differences between the observed and unobserved data.
Standard Methods to Handle Missing Data

- (1) casewise deletion: delete any case that has missing information on any variable (may lose up to 30%-50% of cases if the number of variables involved are many);
- (2) pairwise deletion: include any pair of observations for which data exist in both dimensions (the loss of cases may be small, but each regression model may be based on different samples);
- (3) mean-substitution: easy to use but does not take advantage of other information that might be available in the data (can be biased and inefficient);
- (4) regression-based substitution (unbiased but inefficient); and
- (5) do nothing is doing something (we assume MCAR! Is it a realistic assumption?)
## Missing Assumptions

<table>
<thead>
<tr>
<th>Assumptions</th>
<th>Acronym</th>
<th>You can predict with</th>
</tr>
</thead>
<tbody>
<tr>
<td>Missing Completely at Random</td>
<td>MCAR</td>
<td>------</td>
</tr>
<tr>
<td>Missing at Random</td>
<td>MAR</td>
<td>$D_{\text{obs}}$</td>
</tr>
<tr>
<td>Nonignorable</td>
<td>NI</td>
<td>$D_{\text{obs}}$ and $D_{\text{mis}}$</td>
</tr>
</tbody>
</table>
Missing Assumptions

- Missing values in processes that are MCAR cannot be predicted any better with information in $D$, observed or not. In other words, $M$ is independent of $D$.

$$P(M|D) = P(M)$$

- Example of MCAR: respondents decide whether to answer questions on the basis of coin flips.

- MCAR rarely applies to non-experimental, survey-based data:
  - People in high income level are more likely to decline questions regarding their income level;
  - Independents are more likely to decline to answer a vote preference or partisan identification question
Missing Assumptions

- For MAR processes, the probability that a cell value is missing may depend on $D_{\text{obs}}$ but after controlling for $D_{\text{obs}}$, it is independent of $D_{\text{mis}}$. In other words, $M$ is independent of $D_{\text{mis}}$.

- Example of MAR:
  - If Democratic party identifiers are more likely to refuse the vote choice question, then the process is MAR so long as party identification is a question to which at least some people responded.
  - Data on attitudinal measures may be dependent on one’s educational level, but within each educational level, the data are missing randomly.

- MAR assumptions can be made to fit the data by including more variables in the imputation process to predict the pattern of missingness.
Missing Assumptions

- If the probability of a cell that is missing depends on the unobserved values of the missing response, the process is nonignorable (NI). In other words, M is not independent of D and \( P(M|D) \) does not simplify.
- Example of NI:
  - When high-income individuals are more likely to refuse to answer survey questions about income and when other variables in the data set cannot predict which respondents have high income.
- While it is possible to reject MCAR empirically in favor of MAR, the presence or absence of NI can never be demonstrated using only the observed data. In other words, while it is possible to verify whether one particular imputation method (MI) will outperform another method (listwise, pairwise, mean substitution, etc.), it is impossible to verify absolutely at any multiple imputation model (and, of course, any statistical model).
- The imputation model is only as good as its underlying assumptions and if the assumptions are wrong, the analysis can be wrong as well.
Disadvantages of Listwise Deletion

Disadvantages of Listwise Deletion (LD):
- LD may eliminate too many observations to apply any statistical model.
- Whenever it is possible to predict the probability of a cell in a data matrix that is missing (using $D_{obs}$ and $D_{mis}$), the MCAR assumption is violated and LD may generate biased parameter estimates. In other words, LD can result in different magnitudes or signs of causal or descriptive inference.
- LD always results in an efficiency loss.
- Even if we can assume LD does not generate any bias, some of the coefficients may change from statistical non-significance to statistical significance (especially for points that are one standard error farther away from the truth due to LD). King et. al. have shown that this problem is quite common in political science literature.
Steps in Multiple Imputation

Rubin (1996) described multiple imputation as a three-step process:

- (a) $m$ sets of plausible values for missing observations are created that reflect uncertainty about the nonresponse model. Each of these sets of plausible values can be used to “fill-in” the missing values and create a “completed” dataset.
- (b) each of these datasets can be analyzed using complete-data methods.
- (c) the results are combined, which allows the uncertainty regarding the imputation to be taken into account.
$Q$ quantity of interest

Overall point estimate $q^*$ of $Q$ is

$$q^* = \frac{1}{m} \sum_{j=1}^{m} q_j$$

Total Variance = Within Variance + Between Variance

$$SE(q)^2 = \frac{1}{m} \sum_{j=1}^{m} SE(q_j)^2 + S_q^2 \left(1 + \frac{1}{m}\right)$$

where $SE(q_j) = \text{standard error of } q_j \text{ from data set } j$ and

$$S_q^2 = \sum_{j=1}^{m} \frac{(q_j - q^*)}{(m-1)}$$
Multiple Imputations

- There are a wide variety of imputation models that have been used.
- When missingness is monotone, simple methods such as propensity methods (for continuous variables) (Rosenbaum and Rubin 1983), predictive mean matching (Little 1988), and discriminant analysis or logistic regression (for discrete variables).
- For more complicated missingness, Markov Chain Monte Carlo (MCMC) approaches have been suggested.
Analyzing and Combining Results

Combining Results

- Parameters are Student $t$ distributed
- Degrees of Freedom for each parameter calculated as follows:

$$\nu = (m - 1) \left[ 1 + \frac{W}{(1 + m^{-1})B} \right]^2$$

where $W$ is within-imputation variance and $B$ is between-imputation variance

- As the between-imputation variance increases, the degrees of freedom decrease
Multiple Imputations

- Both the predictive mean matching and MCMC approaches require assumptions of multivariate normality, but there is some evidence (from Schafter 1997) that inferences tend to be robust to minor departures from this assumption.
- Suggestions for what variables are included in the imputation model: variables that will be used in the analysis model, variables that are highly predictive of the variables in the analysis model, variables that are highly predictiveness of the missingness in the data, and variables that describe special features of the sample survey (probability surveys).
Notations

Consider general regression models with outcomes (denoted by $Y$, which may be scalar or vector valued) and a vector of predictors (denoted by $X$).

For a given subject, these quantities are either observed or missing.

- We denote $Y_{\text{obs}}$ as the observed component of outcome and $X_{\text{obs}}$ as the observed component of the predictors.
- Similarly, we denote $Y_{\text{mis}}$ and $X_{\text{mis}}$ as the unobserved component of the outcome and predictors, respectively.
- Finally, we also refer to $Z_{\text{mis}} = (Y_{\text{mis}}, X_{\text{mis}})$ and $Z_{\text{obs}} = (Y_{\text{obs}}, X_{\text{obs}})$. 
Imputation Modeling

- The goals are:
  - To reflect the uncertainty of the imputations properly
  - To preserve the important aspects of the distribution of data
  - To preserve important relationships between variables in the data

- The goals are not:
  - To predict the missing values with the highest accuracy
  - To describe the data in causal manner

- The imputation model is a predictive, not a causal, model and the only use of the model is to create imputations that reflect all available information of the missing values in the data.
Sources of information:


Available Softwares for Multiple Imputation

- (a) SOLAS version 3.0
- (b) SAS 8.2
- (c) Missing Data Library for S-Plus
- (d) Multiple Imputation by Chained Equations (MICE) [S-Plus and R]*
- (e) NORM*
- (f) AMELIA*
- (g) Others: SPSS, LISREL, EQS, HLM, IVEware

*Free, public domain programs
AMELIA
A Program for Missing Data
(Windows Version 2.1 7/15/2003 Gauss Version 2.1)

James Honaker, Anne Joseph, Gary King, Kenneth Scheve, Naunihal Singh
(c) Copyright 1998 - 2003, All rights reserved

This program is built on the Gauss version of Amelia. Both versions of the program and this paper are available at http://GKing.Harvard.Edu.

Press Any Key to Continue
Type the name of the input file and hit return.
Supported file types: Excel 2-5, 95-2000, Lotus WK1, Quattro WQ1, WB*, dBase2-5, Stata 4-7, Paradox 3-5, GenStat GSH, SAS PC 6.03-7.0, Minitab 8-13, Systat, MStat, Instat Epi-Info, SPSS/Win, Gauss Data/Matrix, Matlab, S*, ArcView/Info Shapefiles, MapInfo Interchange (MIF), Comma Delimited Text (CSV)

9/11/2003  Raymond Wong  20
MISUM & MIEST

- After multiple imputations, use MISUM and MIEST to obtain summary statistics and regression models by using the generated replicates.

MISUM
misum dataset varlist [weight] [if exp] [in range][, nsets(m) iout ]

Examples
Suppose you have created 5 multiply imputed datasets and named them mydata1, mdata2, etc. You want to calculate multiple imputation descriptive statistics for the variables income and gender.

. misum mydata income gender

You can show descriptive statistics for each imputed dataset.

. misum mydata income gender, iout

Or change the number of datasets used.

. misum mydata income gender, nsets(10)
miest dataset command varlist [weight] [if exp] [in range] [, nsets(m) iout ... ]

**Examples**

Suppose you have created 5 multiply imputed datasets and named them mydata1, mdata2, etc. You want to regress income on education using multiple imputation.

```
. miest mydata regress income education
```

You can add options to the stata command just as you would normally.

```
. miest mydata regress income education, robust
```

Or change the number of datasets used.

```
. miest mydata regress income education, robust nsets(10)
```

Or change the statistical model used.

```
. miest mydata logit turnout income education, nsets(10)
```

Or show the intermediate results.

```
. miest mydata logit turnout income education, nsets(10) iout