Dealing with Missing Data
Multiple Imputation with Amelia II

15 March 2016

Why Imputation

- What would be standard ways to deal with missing values?

Why Imputation

- What would be standard ways to deal with missing values?

1. Listwise deletion (potential bias & inefficiency)

Why Imputation

- What would be standard ways to deal with missing values?

1. Listwise deletion (potential bias & inefficiency)
   - Best case: inefficiency (information is thrown away)

Why Imputation

- What would be standard ways to deal with missing values?

1. Listwise deletion (potential bias & inefficiency)
   - Best case: inefficiency (information is thrown away)
   - Worst case: estimates are severely biased

Why Imputation

- What would be standard ways to deal with missing values?

1. Listwise deletion (potential bias & inefficiency)
   - Best case: inefficiency (information is thrown away)
   - Worst case: estimates are severely biased
2. Omitting variables (bias)

Perfomance under MAR

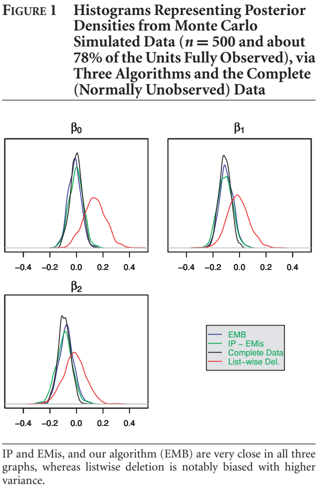

FIGURE 1 Histograms Representing Posterior Densities from Monte Carlo Simulated Data ($n = 500$ and about 78% of the Units Fully Observed), via Three Algorithms and the Complete (Normally Unobserved) Data

IP and EMis, and our algorithm (EMB) are very close in all three graphs, whereas listwise deletion is notably biased with higher variance.

Missingness Assumptions

Assumption	Acronym	Needed to Predict
Missing completely at random	MCAR	-
Missing at random	MAR	observed data (D_{obs})
Nonignorable/ not missing at random	NI or NMAR	observed and missing data (D_{obs}, D_{mis})

- D is a data matrix including y and all x
- M is a matrix with the same dimensions as D filled with 1's when an element in D is observed and 0 when missing
- $D = \{D_{obs}, D_{mis}\}$

Missing Completely at Random (MCAR)

If MCAR is violated, list-wise deletion yields biased estimates.

MCAR means M is independent of D. Neither observed nor missing values of D contain information about whether a value is missing or not.

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

- Example: ?

Missing Completely at Random (MCAR)

If MCAR is violated, list-wise deletion yields biased estimates.

MCAR means M is independent of D. Neither observed nor missing values of D contain information about whether a value is missing or not.

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

- Example: ?
- Data Collectors decide to delete values from their data based on coin flips

Missing at Random (MAR)

MAR means that whether data is missing or not is random after controlling for observed data. So in fact, it is not random but conditional on the observed data it is independent of unobserved data.

$$P(M|D) = P(M|D_{obs})$$

Independence of unobserved data seems like a strong assumption but the degree to which it is fulfilled is somewhat in your power and it is weaker than MCAR (the assumption needed for listwise-deletion)!

- Example: ?

Missing at Random (MAR)

MAR means that whether data is missing or not is random after controlling for observed data. So in fact, it is not random but conditional on the observed data it is independent of unobserved data.

$$P(M|D) = P(M|D_{obs})$$

Independence of unobserved data seems like a strong assumption but the degree to which it is fulfilled is somewhat in your power and it is weaker than MCAR (the assumption needed for listwise-deletion)!

- Example: ?
- I tend to show up in the office on sunny days and tend to stay at home on rainy ones.

Not Missing at Random (NMAR)

Whether a cell is missing depends on the value of the missing variable or some other unobserved variable.

$$P(M|D)$$

Under NMAR both list-wise deletion and multiple imputation are biased. MI aims to significantly increase efficiency while accepting some bias (bias-variance trade-off).

- Example: ?

Not Missing at Random (NMAR)

Whether a cell is missing depends on the value of the missing variable or some other unobserved variable.

$$P(M|D)$$

Under NMAR both list-wise deletion and multiple imputation are biased. MI aims to significantly increase efficiency while accepting some bias (bias-variance trade-off).

- Example: ?
- Rich and poor people tend not to report their income and we did not collect any other variables.

List-wise Deletion vs. MI

Missingness	List-wise Deletion	MI
MCAR	unbiased less efficient	unbiased more efficient
MAR	biased less efficient	unbiased more efficient
NMAR	biased less efficient	biased more efficient

MCAR can be rejected empirically while NMAR cannot. It is therefore possible to assess whether MI will outperform list-wise deletion.

What Amelia Does

- Imputes m values for each missing cell $\rightarrow$ m imputed data sets

Sequence of Analysis

- Impute 5 to 10 data sets using Amelia
  - Amelia assumes data are MAR conditional on the imputation model
  - Variables are assumed to be jointly multivariate normal (this was found to work well even for categorical data)
  - Observed values remain unchanged in all m imputed data sets
  - Missing values are estimated using the expectation maximization algorithm with bootstrapping(EMB)
  - The less certainty about a missing value, the more it varies across imputations
- Do your analysis on each of the imputed data sets
- Combine results

Amelia Assumptions

- MAR conditional on D_{obs}
- D is multivariate normal, $D \sim N(\mu, \Sigma)$
    - Each variable is a linear function of all others

Combining Quantities of Interest (1)

Point estimate: average over the j to m data sets.

$$\bar{q} = \frac{1}{m} \Sigma_{j=1}^{m} q_j$$

Standard error: Average of the estimated variances from within each complete data set, plus sample variance in point estimates across data sets, multiplied by a factor that corrects for bias.

$$\bar{s}^2 = \frac{1}{m} \Sigma_{j=1}^{m} s_j^2 + S_j^2 (1 + \frac{1}{m})$$

where s_j is the standard error estimated from data set j and $S_j^2 = \Sigma_{j=1}^{m} (q_j - \bar{q})/(m-1)$ is the sample variance in the point estimates across the data sets.

Combining Quantities of Interest (2)

- Simulate quantity of interest 1000 times for each model (m times)
- Stack the results into a single vector (length of 1000 * m)
- Extract the point estimate by taking the median value and the confidence interval with the corresponding percentiles.

Practical Points

- The imputation model should contain at least as much information as the analysis model
  - If it is in the analysis model it must be in the imputation (including DV!)
- Beware of non-linearities (except the standard logit and probit functional forms)
  - Quadratic terms in the analysis should go as quadratics into the imputation
  - Interactions in the analysis should go into the imputation

Practical Points 2

- Try meeting the multivariate normal assumption by applying variable transformations (unbounded and symmetric)
  - Skewed variables such as populations or budges can be logged
  - Take the square root of counts (stabilizes variance and makes them approx. symmetric)
  - Logistic transformation makes proportions unbounded and symmetric
  - Make ordinal variables approx. interval
- Whenever possible allow non-integer values to be imputed for ordinal variables including dichotomous ones
  - this carries more information (e.g.: male = 0.7 says it is likely that the respondent was a man)

Thank You!

When to Stick with Listwise Deletion

- The analysis model is conditional on X and the functional form is known to be correctly specified
- There is NMAR missingness and no variables in the data that could be used in the imputation to fix the problem
- Missingness in X is not a function of Y and unobserved omitted variables that affect Y do not exist
- The number of observations is very large making efficiency losses from listwise deletion negligible