Missing Data: Part 1
What to Do?

Carol B. Thompson
Johns Hopkins Biostatistics Center
SON Brown Bag – 3/20/13
Overview

• Missingness and impact on statistical analysis
• Missing data assumptions/mechanisms
• Conventional but not best options
• Preferred options
  – Maximum Likelihood
  – Multiple Imputation
  – Inverse Probability Weighting
Statistical Analysis & Missingness

Population to which inference is to be made

Sampling process

Sample used for inference (assume representative)

Inference process

Part of sample with missing data

Is sample with missing data still representative enough to make appropriate inferences to population of interest?????
Handling Missingness Impact

• How do cases with missing data compare to those without missing data? (*missingness mechanism*)

• What assumptions in the analysis can be made to either adjust for or account for effect of missing data?

• **GOAL** – to make estimates or inferences with missing data that can apply to population as if complete information were available
Analysis Goals When Data Are Missing

• Minimize bias (representative of Population)
• Maximize use of available information
• Good estimates of uncertainty (std errors and p-values)
• Method is easy to use
• NOT that imputed values are “close” to real values
Missingness Mechanisms

• Process by which observations become missing

• Can’t be known for sure; expressed only by observed data

• Mechanism types
  – Missing Completely at Random (MCAR)
  – Missing at Random (MAR)
  – Missing Not at Random (MNAR)
MCAR

• Missing Y are unrelated to value of Y or any variable to be analyzed with Y
• Can use complete data records as simple random subsample from full data set
• Missing Y may be related to some other X: missing age related to failure to report income
• Example - Lab sample dropped
• MCAR example? Subject falls under bus (what if psych trial?)
MCAR cont’d

• Violation: missing income related to age
  – Compare age for missing Y and non-missing Y
  – No difference does not indicate MCAR

• MCAR may occur because of research design
  – Two evaluations, one is very expensive
  – Everyone receives less expensive one
  – Random sample receives very expensive one

• MCAR is strongest assumption

• Unbiased estimates and don’t need to model missingness
MAR

- Weaker assumption than MCAR
- Missing Y are not related to value of Y, after controlling for other X in the analyses
- 2 subjects with same observed values have same statistical behavior on other variables, observed or not
- MAR - Ignorable
MAR - Ignorable

• Example:
  – Missing income depends on mental status
  – Within mental status category, missing income not related to value of income (i.e., when controlling for mental status).

• Not possible to test for MAR, because no info on missing Y
MAR – Ignorable cont’d

• Parameters governing missing data process are unrelated to parameters to be estimated
• Rare that above assumption is not satisfied
• No need to model missing data mechanism
• If assumption is not satisfied, ignorability methods work fine, but it might be better to model missing data mechanism
MAR – Non-Ignorable

• Missing Not at Random (MNAR)
• Even accounting for all available information, the reason for missing observations still depends on the unseen data
• Example – In clinical drug trials, subjects getting worse are more likely to drop out
• Must model missingness mechanism
  – Requires some prior knowledge to do this
  – Best to perform sensitivity analysis on likely models for missingness mechanism
Conventional But Not Best Options

• Listwise deletion
• Pairwise deletion
• Dummy variable adjustment
• Last observation carried forward (LOCF)
• Imputation
  – Simple mean
  – Regression mean
Listwise Deletion

• Use only complete cases

• Advantages
  – No restriction on type of analysis performed
  – No special methods required
  – If MCAR, can be considered simple random subsample
  – Unbiased estimates of standard errors and test statistics are appropriate
  – Std errors may be larger because fewer cases
Listwise Deletion cont’d

• Disadvantages
  – If MAR, estimates may be biased.
  – If multiple models, cases used per model may change → inconsistent inferences.

• Robust for violations of MAR
  – Assumes regression coefficients are same for all cases
  – If relationships vary across subpopulations, may weight coefficients toward one group or another (separate regressions or interactions)
Pairwise Deletion

• AKA: Available case analysis
• Works for analyses depending only on summary statistics
  – Means, std deviations, correlations
  – Not good for regression, factor analysis
Pairwise Deletion cont’d

- Ambiguities in implementation:
  - If MCAR, estimates are approximately unbiased for large samples
  - If MAR, but not “observed at random”, estimates and std errors may be seriously biased.
  - Need sample size to calculate std errors, but N from which pair of variables???
  - Not a good alternative to Listwise Deletion
Dummy Variable Adjustment

• Create variable D based on whether X’s value is missing or not (0,1)

• Adjust X:
  – Xc = X if not missing
  – Xc = c if missing (value of c doesn’t matter)

• Regress Y on Xc, D, other X
Dummy Variable Adjustment cont’d

• Advantages
  – Uses all available data
  – Choice of c only affects coefficient for D
  – Coefficient for D = predicted Y if missing X, controlling for other X
  – Coefficient for Xc = estimate of effect of X if not missing X
Dummy Variable Adjustment cont’d

• Disadvantages
  – Biased estimates even if MCAR
  – Impact of extra category in variable for missing value
    • Depends on distribution of missing responses across rest of categories
    • Depends on how probability of being missing depends on category
    • Dissimilar cases can be grouped together in “missing” category
    • Will likely lead to severe bias
LOCF

• Usually used in longitudinal studies
• Assumes that future observations on a subject are the same as past observations
  – Realistic assumption under treatment???
• Means and covariance structure are seriously distorted
Imputation

• Substitute with reasonable guess; analyze as if no missing data or “guesses”

• Mean imputation
  – Substitute with marginal mean (mean of all observations with data on variable)
  – Biased estimates of variances and covariances
  – AVOID
  – Substituting mean for items in scale to get total
    Consider # missing items – perhaps < 25%, and # cases with missing items
Imputation cont’d

• Conditional or Regression Mean Imputation
  – Regress missing X on all other X’s to generate predicted values of X
  – Use mean predicted value for each missing X
  – Can give unbiased estimates of mean, associations and regression coefficients in more situations than simple mean imputation
Imputation cont’d

– If missing is only on 1 X
  • If MCAR, coefficients are approximately unbiased in “large” samples
  • Can use with weighted least squares or generalized least squares

– These imputations underestimate standard errors and thus overestimate test statistics (more significance than warranted).
Imputation Example (Baraldi)

**Complete Data Set**

![Graph showing complete data set]

**List-wise Deletion**

![Graph showing list-wise deletion]

Mean – 76.35; Std Dev – 10.73

Mean – 81.8, Std Dev – 10.84
Imputation Example (Baraldi) cont’d

Mean Imputation

Regression Imputation

Mean - 81.80, Std Dev – 7.46

Mean – 76.12, Std Dev – 9.67
Conventional but not best options make things worse

- Introduce substantial bias
- Make analysis more sensitive to departures from MCAR
- Yield std error estimates that are incorrect (usually too low) → potential for more significance than warranted
- If # missing is small, may not affect much, but how small or what makes up missing may not be defensible.
Preferred Options

• Maximum Likelihood Estimation (MLE)
• Multiple Imputation (MI)
• Inverse Probability Weighting (IPW)
  – Weight observations inversely according to the estimated probability of not being missing
Preferred Options cont’d

• They make distributional assumptions about
  – The unseen data
  – Form of MV mechanism
• Based on well-defined statistical models
• Don’t replace MV directly
• Available info is combined with assumptions; doesn’t depend on observed data alone
• Analyses, inferences, conclusions are valid under these assumptions
MLE

• Identifies parameter values that have the highest probability of producing the observed data.

• Under a fairly wide range of conditions:
  – Approximately unbiased
  – Asymptotically normal – large samples
    • Can be used to construct CIs and p-values
MLE – General Missing Patterns

• **Expectation-maximization (EM)**
  – Basically reduces to regression imputation of missing values (E)
  – Computes means and covariance matrix with imputed values (M)
  – Cycles through E and M until estimates converge
  – Uses all variables as predictors

• **Advantages**
  • Easy to use; lots of commercial software

• **Disadvantage**
  • Linear modeling std errors + test statistics are incorrect
  • Biased estimates
MLE – General Missing Patterns cont’d

• Direct ML
  – Uses structural equation modeling (e.g., AMOS, MPlus)
  – Advantages
    • Estimates optimal for large samples
    • Estimates with correct std errors
  – Disadvantage
    • Requires specialized software
    • Some distribution assumptions may be unrealistic
Multiple Imputation (MI)

• MI has same optimal properties as ML, but removes some of ML limitations
• When used correctly, produces estimates:
  – Approximately unbiased, better as sample size increases
  – Asymptotically normal when data are MAR
    • Can construct CIs and p-values
MI cont’d

• Advantages
  – Can be used with virtually any kind of data,
  – any kind of model,
  – and with unmodified conventional software

• Disadvantages
  – Can be cumbersome to implement,
  – Is easy to do wrong,
  – Produces different estimates every time it is used (hopefully small differences)
Single Random Imputation

• Replaces missing value with random draw from a distribution based on observed data (hot-deck)
• Does not produce a unique set of numbers (random variation in process)
• Without random component, produces underestimates of variances for variables with missing data.
MI Process

• Repeat the random imputation process more than once (5 times is generally enough)
• Each imputation process represents random sample from distribution of plausible values for missing values
• Important for imputation processes to be independent – large number of iterations between each saved data set
• Analyze data set from each imputation process as if no missing data
MI Process – Pooling Estimates

• Calculate mean of estimates
• Calculate mean of squared std errors
• Calculate variance of estimates
• Calculate square root of mean of variances plus variance of estimates
• Can be used with any parameter
MI Example (Howell- Part 2)

Regression coefficients from five imputed data sets

<table>
<thead>
<tr>
<th>Data set</th>
<th>Estimated parameter</th>
<th>$b_0$</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$b_4$</th>
<th>$b_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Coefficient</td>
<td>-11.535</td>
<td>-2.780</td>
<td>1.029</td>
<td>-0.031</td>
<td>-0.359</td>
<td>0.572</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>43.204</td>
<td>3.323</td>
<td>0.013</td>
<td>0.013</td>
<td>0.013</td>
<td>0.012</td>
</tr>
<tr>
<td>2</td>
<td>Coefficient</td>
<td>-11.501</td>
<td>-4.149</td>
<td>1.040</td>
<td>-0.093</td>
<td>-0.583</td>
<td>0.876</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>40.488</td>
<td>2.680</td>
<td>0.010</td>
<td>0.009</td>
<td>0.009</td>
<td>0.007</td>
</tr>
<tr>
<td>3</td>
<td>Coefficient</td>
<td>-10.141</td>
<td>-5.038</td>
<td>0.766</td>
<td>0.123</td>
<td>-0.252</td>
<td>0.625</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>42.055</td>
<td>3.301</td>
<td>0.010</td>
<td>0.010</td>
<td>0.010</td>
<td>0.009</td>
</tr>
<tr>
<td>4</td>
<td>Coefficient</td>
<td>-11.533</td>
<td>-6.920</td>
<td>0.870</td>
<td>0.084</td>
<td>-0.458</td>
<td>0.815</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>28.751</td>
<td>1.796</td>
<td>0.081</td>
<td>0.007</td>
<td>0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>5</td>
<td>Coefficient</td>
<td>-14.586</td>
<td>-1.115</td>
<td>0.718</td>
<td>0.050</td>
<td>-0.373</td>
<td>0.814</td>
</tr>
<tr>
<td></td>
<td>Variance</td>
<td>32.856</td>
<td>2.362</td>
<td>0.009</td>
<td>0.009</td>
<td>0.009</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Mean $b_i$          | -11.859  | -4.000  | 0.885   | 0.027   | -0.405  | 0.740   |
Mean Var.($\bar{W}$)| 37.471   | 2.692   | 0.025   | 0.010   | 0.010   | 0.009   |
Var. of $b_i$ (B)   | 2.682    | 4.859   | 0.022   | 0.008   | 0.015   | 0.018   |
$\overline{T}$      | 40.69    | 8.523   | 0.051   | 0.020   | 0.028   | 0.031   |
$\sqrt{T}$          | 6.379    | 2.919   | 0.226   | 0.141   | 0.167   | 0.176   |
t                | -1.859   | -1.370  | 3.916*  | 0.191   | 2.425*  | 4.204*  |

* $p < .05$  “Var.” refers to the squared standard error of the coefficient.
Additional Rules of Thumb (Allison)

• Dependent Variable (DV) should always be included in imputation regression analysis

• Impute missing values on DV if:
  – There are auxiliary variables strongly correlated with DV.

• Don’t impute DV if:
  – No missing predictor data or auxiliary variables
  – No auxiliary variables and missing predictor data
References

- [http://www.missingdata.org.uk](http://www.missingdata.org.uk)