Handling X-side Missing Data with M*plus*

Tor Neilands

Center for AIDS Prevention Studies

November 19, 2010

Contents

- X-side and Y-side Missing Data
- Missing Data Mechanisms
- Approaches for Addressing Missing Data in Statistical Analyses
- Examples
  - Example 1: Linear Regression
  - Example 2: Logistic Regression with Clustered Data
  - Example 3: Multilevel Analysis
  - Example 4: MI for Clustered Categorical Data
- M*plus* Availability & Resources and Acknowledgements

Missing Data Overview

- Missing data are ubiquitous in applied quantitative studies
  - Don't know/don't remember/refused responses on cross-sectional surveys and self-administered paper surveys
  - Skip patterns
  - Interviewer error/A-CASI programming errors or omissions.
  - Longitudinal loss to follow-up

Preventing Missing Data

- Prevention is the best first step
  - A-CASI, CAPI, etc.
  - Rigorous retention protocols for participant tracking, etc.
  - Diane, Bill, and Lance's work with flexible interviewing methods.
  - Asking longitudinal study participants if they anticipate barriers to returning for follow-up visits, then problem solving those issues.

Missing Data Mechanisms

- What mechanisms lead to missing data?
- Rubin's taxonomy of missing data mechanisms
  - Rubin (1976), *Biometrika*
  - MCAR: Missing Completely at Random
  - MAR: Missing at Random
  - NMAR: Not Missing at Random
    - Also known as MNAR (Missing Not at Random).
  - Good articles that spell this out:
    - Schafer & Graham, 2002, *Psychological Methods*
    - Graham, 2009, *Annual Review of Psychology*

Missing at Random (MAR)

- Define *R* as an indicator of (non)missingness for variable *Y*. *R* = 1 if *Y* is observed; *R* = 0 if *Y* is missing.
- Denote Y*complete* as the complete data. Partition Y*complete* as
  - Y*complete* = (Y*observed*, Y*missing*)
- MAR occurs when the distribution of missingness does not depend on the values of Y that would have been observed had Y not been missing:
  - P(*R*|Y*complete*) = P(R|Y*observed*)

Missing Completely at Random (MCAR)

- Put another way, MAR allows the probabilities of missingness to depend on observed data, but not on missing data.
- MAR is a much less restrictive assumption than MCAR.
- MCAR is a special case of MAR where the distribution of missing data does not depend on Y*observed*, also:
  - $P(R|Y_{complete}) = P(R)$
- If incomplete data are MCAR, the cases with complete data are then a random subset of the original sample.

Not Missing at Random (NMAR)

- The probability that Y is missing is a function of Y itself.
- Missing data mechanism must be modeled to obtain good parameter estimates. Examples:
  - Heckman's selection model
  - Pattern mixture models
- Disadvantages of NMAR modeling: Requires high level of knowledge about missingness mechanism; results are often sensitive to the choice of NMAR model selected.

MCAR, MAR, NMAR Revisited

- From Schafer & Graham, 2002, p. 151: Another way to think about MAR, MCAR, and NMAR: If you have observed data X and incomplete data Y, and assuming independence of observations:
  - MCAR indicates that the probability of Y being missing for a participant does not depend her values on X or Y.
  - MAR indicates that the probability of Y being missing for the participant may depend on her X values but not her Y values.
  - NMAR indicates that the probability of Y being missing depends on the participant's actual Y values.

Listwise Deletion of Missing Data

- Standard statistical programs typically delete the whole case from an analysis if one or more variables' values are missing (listwise deletion). Consequences:
- If missing data are due to MCAR:
  - Parameter estimates unbiased but standard errors are enlarged and power for hypothesis testing is reduced
- If missing data are due to MAR:
  - Parameter estimates may be biased, standard errors enlarged, and power for hypothesis testing reduced
- If missing data are due to NMAR:
  - Parameter estimates may be biased standard errors enlarged, and power for hypothesis testing reduced

Occurrence of Missingness Types

- MCAR: Missing Completely at Random
  - A very stringent assumption unlikely to be met in practice
  - Example: computer failure loses some cases' data but not others
- MAR: Missing at Random
  - Much more likely to be met in practice, especially in social and behavioral research (Schafer & Graham, 2002, *Psychological Methods*)
- NMAR: Not Missing at Random
  - Unknown. MCAR vs. MAR can be formally tested via statistical tests, but MAR vs. NMAR cannot be tested.
  - Inclusion of measures during the study design phase that are likely to be correlated with subsequent data missingness can help to minimize NMAR missingness.
  - Some NMAR missingness may be inevitable, however.

Addressing Missing Data

- *Ad hoc* methods such as listwise deletion, pairwise deletion, or substitution of the variable's mean value usually assume MCAR and are not recommended. See Paul Allison's 2002 Sage publication for a readable treatment of the reasons why these methods don't usually work well.
- Listwise deletion may yield unbiased results in some circumstances, however:
  - Regression models where the probability of missing data on the independent variables does not depend on the value of the dependent variable (Allison, 2002, pp. 6-7).

Addressing Missing Data

- Regression models where the probability of missingness on Y depends on X values (covariate-dependent missingness. See: Little, 1995, *JASA*).
- In general, estimates of sample statistics such as means are more biased due to missing data than are regression parameter estimates (Little & Rubin, 2002: *Statistical Analysis with Missing Data*, Wiley, 2002).
- Remember, though, that efficiency in the regression analysis context is reduced due to missing data. You can lose a lot of statistical power, especially if there are many cases and missing data patterns, and the number of complete cases is a small fraction of the original number of cases.

Addressing Missing Data

- NMAR missingness can only be addressed through explicitly assuming a given missingness mechanism, which can lead to suboptimal results if the incorrect missingness mechanism is specified (Allison, 2002).
- There is even some evidence that methods that assume MAR missingness may outperform other approaches for NMAR situations (Muthen, Kaplan, & Hollis, 1987, *Psychometrika*).

Methods for MAR Missingness

- Ibrahim (*JASA*, 2005) reviewed four general approaches and found all to perform about equally well:
  - Inverse censoring weights
  - Fully Bayesian analysis
  - Multiple imputation (MI)
  - Direct maximum likelihood estimation (ML)
- A full treatment of each technique is beyond the scope of today's presentation. We will concentrate on how to employ M*plus* to address X-side missingness using direct maximum likelihood (ML) and multiple imputation (MI) under the assumption of MAR for incomplete data.

X-side and Y-side Missingness

- Some software programs implicitly incorporate direct maximum likelihood handling of an outcome variable Y. These are typically mixed models routines that can be employed to analyze longitudinal data with missing outcomes
  - PROCs MIXED, GLIMMIX, and NLMIXED in SAS
  - MIXED in SPSS
  - Stata -xt- commands (there are many) and -gllamm-
- However, these commands will drop the whole observation when one or more X values are missing.
- They cannot conveniently be used to handle cross-sectional missing data.

M*plus* for Addressing Missing Data

- Some of the most important developments in handling non-normal and incomplete data arose in the latent variable (structural equation modeling) field in the 1990s.
- For many years, EQS had the best non-normal data routines and AMOS had the most user-friendly implementation of direct ML suitable for use with cross-sectional and longitudinal X-side and Y-side missing data.
- In the late 1990s, Bengt and Linda Muthén developed M*plus*, a general latent variable modeling program that included and enhanced these approaches to include, among other things, the ability to handle categorical outcome variables.

M*plus* for Addressing Missing Data

- M*plus* is mainly known as a latent variable modeling program, but it can fit a wide class of regression models, including linear, binary and ordinal logistic, count data models (Poisson, negative binomial, hurdle, zero-inflated Poisson, zero-inflated negative binomial), and parametric and non-parametric (Cox) discrete and continuous time survival models.
- For clustered data structures, robust standard errors with or without weights and stratification are available for complex survey data. Multilevel modeling with two levels is available for model-based multilevel analyses.

How to transfer data to M*plus*: A very brief guide

- M*plus* only accepts delimited ASCII (raw text data as input)
  - Delimeter can be spaces, commas, etc.
- My approach: Use Stat/Transfer to save a space-delimeted text file with variable names in the first row and a specific value (e.g., -999) to denote missing values for all variables.
- Change the file extension from *.csv* to *.txt*.
- Open the text data file with Notepad and cut the first row containing the variable names into M*plus*. Delete the resulting blank line in the data file so that the first row of the data file starts with the first case (i.e., with actual data).
- Paste the variable names into M*plus* and remove the quotation marks. Note that M*plus* prefers variable names to be only 8 characters or less in length.

Example 1: Linear Regression

- Example from Paul Allison's monograph *Missing Data* (Sage Publications, 2002, p. 21)
- N = 1,302 US colleges and universities
- Outcome: gradrat (ratio of graduating seniors to number enrolled four years previously * 100)
- Explanatory variables:
  - CSAT – Combined mean scores on verbal and math SAT
  - LNEnroll – Natural log of number of freshmen
  - Private – 0 = public school; 1 = private school
  - Stufac – Ratio of students to faculty * 100
  - RMBrd – Total cost of room & board in thousands of dollars
- Auxiliary variable: ACT – Mean ACT score
  - Not in the regression model, but correlated with CSAT.
- Only Private has complete data for all 1,302 colleges.

Example 1: Linear Regression

- *N* = 455 cases with complete data. This would be the default available *N* for multiple linear regression analysis in Stata, SAS, SPSS, etc.
- And M*plus*, too: by default, M*plus* assumes that any x-side variable is fixed and drops cases with missing data for x-side variables. If, however, we are willing to convert x-side variables with missing data to random variables with distributions and assume multivariate normality of the underlying joint distribution, those variables can be brought into the likelihood and the cases with partial information will be included in the analysis. See http://www.statmodel.com/verhistory.shtml "Analysis Conditional on Covariates" for more information on the thinking behind this selection of this default behavior.

Example 1: Linear Regression

- See M*plus* model output files:
  - *Allison1.out:* Default listwise deletion of cases with missing X data (N = 455).
  - *Allison2.out:* Direct ML analysis (N = 1,302) with ML standard errors.
  - *Allison3.out:* Direct ML analysis (N = 1,302) with robust standard errors.
  - *Allison4.out:* Direct ML analysis (N = 1,302) with bootstrap standard errors and bias-corrected confidence intervals based on 5,000 bootstrap samples.
- Results: All regression coefficients are significantly different from zero across all analyses, except for student-faculty ratio:
  - Listwise analysis result: B = -.123, SE = .131, p = .347
  - Direct ML result: B = -.194, SE = .102, p = .058
  - Direct ML, robust SEs: B = -.194, SE = .106, p = .068
  - Direct ML, bootstrap SEs: B = -.194, SE = .123, p = .116
    - Bootstrap 95% bias-corrected CI: -.391, .099

Example 1: Linear Regression

- The number of M*plus* input and output files can grow rapidly for any given project. For instance, in the previous example we have 4 input + 4 output files = 8 M*plus* files.
- An alternative for Stata users is to use the runmplus utility that enables calls to M*plus* from within Stata. Input file syntax and data are passed to the M*plus* engine from Stata and results are passed back to Stata to display in the Stata Results window.

Example 1: Linear Regression

- Stata do file *usnews_Mplus_Models.do* example syntax block:

```
** --> Allison 2 (Direct ML estimation - N = 1,302) <-- **

runmplus gradrat csat stufac rmbrd private lenroll, estimator(ML) ///
model(gradrat ON csat stufac rmbrd private lenroll; csat; stufac; rmbrd; lenroll) ///
auxiliary are act;
```

Example 1: Linear Regression

- Features of the analysis:
  - Direct ML handling of incomplete X-side and Y-side data.
  - Easy incorporation of auxiliary variable(s) into the analysis.
    - The automatic auxiliary variable inclusion feature is available only for single-level analyses involving all continuous outcomes or mediators. You can still include auxiliary variables manually in other types of analyses, however.
- The LISTWISE = ON option is available in the M*plus* DATA command to explicitly request a listwise analysis. Useful for:
  - Verifying that your Mplus model specification is correct by comparing M*plus* results to those from the same model fitted in SAS, Stata, etc.
  - Use with care: LISTWISE = ON will delete cases with both Y-side and X-side missing data.

Example 1: Linear Regression

- Robust standard errors (M*plus* estimator MLR) and the bootstrap are available to address assumption violations due to non-normal or heteroskedastic outcome data.
- Mplus MLR estimator assumes MCAR missingness and finite fourth-order moments (i.e., kurtosis is non-zero); initial simulation studies show low SE bias for this estimator with MAR data. See http://www.statmodel.com/download/webnotes/mc2.pdf for more information about this estimator.

Example 2: Logistic Regression with Clustered Data

- Gay Couples Study (Colleen Hoff, PI)
- Study of 566 gay male couples from San Francisco
- Outcome: Any unprotected anal intercourse in the past three months with an outside partner of discordant or unknown serostatus (UAIOUTDU; 1 = yes, 0 = no).
- Explanatory variables:
  - Relationship length in years (couple-level)
  - Non-monogamous (couple-level; 1 = yes, 0 = no)
  - Age in years (individual-level)
  - CES-D depression (individual-level)
  - Sexual agreement investment (individual-level)
  - Couple serostatus (couple-level; 1 = -/-; 2 = +/-; 3 = +/+)
- Missing data:
  - 131 out of 1,132 men (11.6%) did not report having a sexual agreement and so were skipped out of the sexual agreement questions.

Example 2: Logistic Regression with Clustered Data

- Analysis: Logistic regression of UAIOUTDU onto the explanatory variables listed on the previous slide.
- Clustered data structure: 1,132 individual men nested within 566 couples.
- Requires logistic regression approach that allows for clustered data. Some possibilities:
  - Subject-specific coefficients:
    - Multilevel generalized linear mixed modeling (GLMM)
  - Population-average coefficients:
    - Logistic regression analysis with robust standard errors
    - GEE

Example 2: Logistic Regression with Clustered Data

- We will use the robust standard error approach. This is similar to using PROC SURVEYLOGISTIC in SAS or -logistic- with the -cluster- option in Stata. In M*plus*, we use the MLR estimator with the CLUSTER option and the COMPLEX analysis type.
- But… a default listwise deletion analysis would use *N* = 1,001, a 12% reduction from the original *N* of 1,132.

Example 2: Logistic Regression with Clustered Data

- How to handle the missing data?
  - Multiple imputation? This would require reshaping of long data into a wide format; imputing agreement investment scores; back-transposing to the long format, and analyzing. Tedious.
  - Direct ML in M*plus*: Run once with the original long data structure.

Example 2: Logistic Regression with Clustered Data

- Compare the default analysis results where cases with missing x-data are excluded to an analysis where they are included by specifying the x-variables with missing data as random variables.
  - Default analysis results (*HoffDemo1.out*)
  - Direct ML analysis results (*HoffDemo2.out*)

Example 2: Logistic Regression with Clustered Data

Effect	Listwise OR (p)	Direct ML OR (p)
HIV-Discordant Couples	2.16 (.004)	2.30 (.001)
HIV-Positive Couples	1.82 (.034)	1.43 (.168)
Relationship Length	1.03 (.125)	1.02 (.280)
Non-Monogamous	9.68 (<.001)	6.67 (<.001)
CES-D	1.00 (.971)	1.01 (.517)
Age	.981 (.097)	.981 (.067)
SAIS	.973 (.021)	.973 (.016)

Example 2: Logistic Regression with Clustered Data

- Features of the analysis
  - Binary outcome variable
  - Clustered data structure
  - Direct ML handling of incomplete X-side data

Example 3: Multilevel Analysis (Linear Mixed Model)

- Study of 99,966 HIV-positive patients from five geographic subregions (North America; East Africa; West Africa; South Africa; Asia) measured repeatedly following initiation of ART.
- Outcome: CD4 count
- Explanatory variables:
  - Sex
  - Age
  - Baseline CD4 (represented by three restricted cubic spline variables)
  - Baseline viral load stage (five stages via four dummy variables)
  - On AZT at baseline
  - On Efavirenz at baseline

Example 3: Multilevel Analysis

- Missing data:
  - Baseline viral load category: N = 24,486 (24.5%) spread across the five geographic subregions
  - Baseline Efavirenz: N = 3,055 (3.06%) in the S. Africa group only
  - Baseline AZT: N = 2,852 (2.85%) in the South Africa group only
- All other explanatory variables and the outcome were fully observed.
- Goal: Fit a linear mixed model with linear splines representing three time windows (1.5 mos-4 mos; 4 mos-12 mos; 12 mos-36 mos) with random slopes and intercepts with a completely unstructured covariance matrix among the random effects.
- Main effects only for all covariates but baseline CD4.
- Moreover, all mean and covariance parameters need to be group-specific. Time was centered at 4 months.

Example 3: Multilevel Analysis

- If we analyze these data with SAS PROC MIXED or Stata -xtmixed-, the analysis *N* is 72,731, a 27% loss of observations.
- Using direct maximum likelihood in M*plus*, the full *N* of 99,966 can be used.
- Sidebar: Convergence with the full model could not be obtained with SAS or Stata (even with multiply imputed data), but was possible with M*plus* in this particular scenario.
  - In other scenarios, SAS or Stata may converge, but M*plus* may not.

Example 3: Multilevel Analysis

- Refer to the handout *model_d_3spline_sqrt.out.*
- Features of the analysis:
  - Uses all available data
  - Separate variances and covariances among random effects are easily implemented within the multiple groups approach (55 random effects estimated).
  - Various model parameters are easily fixed or set to be equal. Non-focal covariate effects are set equal across splines and geographic regions via named parameters.
  - Model-based standard errors used here, but MCAR robust MLR standard errors are available if desired.
  - Post-estimation of spline knots at 4, 12, & 36 months is available via the MODEL CONSTRAINT command.

Example 4: MI for Clustered Categorical Data

- The range of model types that can be fitted in M*plus* using direct ML or Bayesian estimation is very broad.
- However, sometimes you may still want to generate multiple imputations. Examples:
  - To obtain regression diagnostics, which are more readily available in programs like SAS and Stata
  - To conveniently obtain specific tests or assumptions (e.g., score test for proportional odds for ordinal logistic regression; Hosmer-Lemeshow lack-of-fit test for logistic regression, etc.).

Example 4: MI for Clustered Categorical Data

- SAS and Stata have well-developed suites for generating multiple imputations under the assumption of joint multivariate normality
  - This approach is robust to non-normality if the subsequent data analyses are carried out using approaches that address the non-normality (Schafer, 1997, p. 147-148).
  - Though it is not large, there is still some bias in this approach, however (Horton, Lipsitz, Parzen, American Statistician, 2003).
- Multiple imputations via chained equations (MICE) provides an alternative approach for imputing categorical data.
- Stata has a MICE implementation via the user-written program –ice- for imputing IID (independent and identically distributed; i.e., unclustered data)

Example 4: MI for Clustered Categorical Data

- For repeated measures data with a small number of fixed time points, it is possible to transpose or reshape a multiple record long data file into a multiple variable wide data file with one record per subject, generate the imputations, and then retransform the imputed data back into the long file format used by mixed models and GEE software routines.
- Handling clustered data with unequally spaced repeated measurements or non-longitudinal clustered data structures is more difficult. For instance, one could impute within each cluster and include dummy variables indicating cluster membership in the imputation model (Graham, 2009, *Annual Review of Psychology*).

Example 4: MI for Clustered Categorical Data

- M*plus* 6 allows for imputation of clustered continuous and ordered categorical data assuming random intercepts for clusters and an unrestricted covariance structure among the variables.
- It is also possible to impute data under a more complex random effects structure (e.g.,. random intercepts and random slopes), but the analyst must specify a specific model under which the data must be imputed. If an ordered categorical variable is represented in the model as a series of dummy variables, this can be problematic because M*plus* may assign 1s to multiple dummy variables during the imputation process.

Example 4: MI for Clustered Categorical Data

- Impute missing values for Efavirenz (binary), AZT (binary), and viral load WHO status (ordinal) among South African participants in the previous example (N = 65,401).
  - As noted previously, only Efavirenz, AZT, and baseline VL category had missing values
    - 41,185 (63%) had complete data
    - 21,137 (32%) were missing baseline VL but nothing else
    - 2,547 (3.9%) were missing Efavirenz and AZT, but nothing else
    - 305 (<1%) were missing Efavirenz, AZT, and baseline VL
    - 202 (<1%) were missing Efavirenz only
    - 24 (<1%) were missing Efavirenz, AZT, baseline VL and the outcome
    - 1 (<1%) was missing baseline VL and Efavirenz, but not AZT
- Use the M*plus* basic random intercepts multiple imputation scheme for missing clustered ordered categorical data to generate 11 imputed data files
- See the output file *sqrtcd4_group4_imputations_basic_output.pdf*
- Features of the analysis
  - Multiple imputation of clustered categorical data

M*plus* Availability & Resources

- For CAPS investigators and staff, two M*plus* licenses are accessible via the Citrix terminal server system.
- For all others or remote use without an internet connection, licenses are available for purchase at http://www.statmodel.com.
- Online user's guide available at this same URL, along with many examples and a discussion forum.
- Sample programs and Monte Carlo counterparts come with the program.
- Extra manuals from previous versions are in the CAPS reading room.
- For Stata users, there is a user-written ado program that enables Stata do files to pass M*plus* syntax to the M*plus* engine from Stata. http://sites.google.com/site/lvmworkshop/home/runmplus-stuff

Acknowledgements

- Slide Reviewers:
  - Dee Chakravarty, MS
  - Steve Gregorich, PhD
  - Estie Hudes, PhD, MPH
- Stata-to-M*plus* utility:
  - Adam Carle, PhD, James M. Anderson Center for Health Systems Excellence, Cincinnati, OH
  - Richard Jones, ScD, Institute for Aging Research, Hebrew SeniorLife, Boston, MA
- Data Sharing:
  - Colleen Hoff, PhD, PI: Gay Couples Study
  - Jeff Martin, MD, MPH & Elvin Geng, MD, MPH: NA-ACCORD CD4 data