Handling Missing Data

Tor Neilands, PhD

Estie Hudes, PhD, MPH

Center for AIDS Prevention Studies

Part 1: December 14, 2012

Contents

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 Binson's, Bill Woods', and Lance Pollack's work with flexible interviewing methods.
  - Choi study QDS example
- Asking longitudinal study participants if they anticipate barriers to returning for follow-up visits, then problem solving those issues. See: Leon, Demirtas, Hedeker, 2007, *Clinical Trials*

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.

Missing Data Mechanisms: Example 1

- Measuring systolic blood pressure (SBP) in January and February (Schafer and Graham, 2002, *Psychological Methods*, 7(2), 147-177)
  - MCAR: Data missing in February at random, unrelated to SBP level in January or February or any other variable in the study - missing cases are a random subset of the original sample's cases.
  - MAR: Data missing in February because the *January* measurement did not exceed 140 - cases are randomly missing data within the two groups: SBP > 140 and SBP <= 140.
  - NMAR: Data missing in February because the *February* SBP measurement did not exceed 140. (SBP taken, but not recorded if it is <= 140.) Cases' data are not missing at random.

Missing Data Mechanisms: Example 2

- Measuring Body Mass Index (BMI) of ambulance drivers in a longitudinal context (Heitjan, 1997, *AJPH*, 87(4), 548-550).
  - MCAR: Data missing at follow-up because participants were out on call at time of scheduled measurement, i.e., reason for data missingness is unrelated to BMI or other measured variables - missing cases are a random subset of the population of all cases.
  - MAR: Data missing at follow-up because of high BMI and embarrassment at initial visit, regardless of whether participant gained or lost weight since baseline, i.e., reason for data missingness in follow-up BMI is related to baseline BMI, a measured variable in the study.
  - NMAR: Data missing at follow-up because of weight gain since last visit (assuming weight gain is unrelated to other measured variables in the study, i.e., baseline BMI not available for some reason).

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.

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 where variables tend to be correlated with each other and with missingness (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.

Ignorability

- Ignorable data missingness - occurs when data are incomplete due to MCAR or MAR process
- If incomplete data arise from an MCAR or MAR data missingness mechanism, there is no need for the analyst to explicitly model the missing data mechanism (in the likelihood function), as long as the analyst uses software programs that take the missingness mechanism into account internally (several of these will be mentioned later)
- Even if data missingness is not fully MAR, methods that assume MAR usually (though not always) offer lower expected parameter estimate bias than methods that assume MCAR (Muthén, Kaplan, & Hollis, *Psychometrika*, 1987).

Ad-hoc Approaches to Handling Missing Data

- Listwise deletion (a.k.a. complete-case analysis)
  - Standard statistical programs typically delete the whole case from an analysis if one or more variables' values are missing and use only complete cases in analyses (listwise deletion)
- Pairwise deletion (a.k.a. available-case analysis)
- Dummy variable adjustment (Cohen & Cohen)
- Single imputation replacement with variable or participant means
- Regression
- Hot deck

Listwise Deletion of Missing Data

Consequences of listwise deletion of missing data:

- If missing data are due to MCAR:
  - Parameter estimates are 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
  - Robust to NMAR for predictor variables (all regression models) and robust to predictor variables OR outcome variable in logistic models (slopes only)

Pairwise Deletion of Missing Data

- Use pairs of available cases for computation of any sample moment.
  - For computation of means and variances, use all available data for each variable
  - For computation of covariances, use all available data on pairs of variables.
- Can lead to non-positive definite variance-covariance matrices because it uses different pairs of cases for each entry.
- Can lead to biased standard errors under MAR.

Dummy Variable Adjustment

Advocated by Cohen & Cohen (1985). Steps:

1. When *X* has missing values, create a dummy variable *D* to indicate complete case versus case with missing data.
2. When *X* is missing, fill in a constant *c*
3. Regress *Y* on *X* and *D* (and other non-missing predictors).

- Produces biased coefficient estimates (see Jones' 1996 *JASA* article)

Single Imputation Methods

- Mean substitution - by variable or by observation
- Regression imputation (i.e., replacement with conditional means)
- Hot deck: Pick "donor" cases within homogeneous strata of observed data to provide data for cases with unobserved values.
- These ad hoc approaches lead to biased parameter estimates (e.g., means, regression coefficients); variance and standard error estimates that are biased downwards.
  - One exception: Rubin (1987) provides a hot-deck based method of multiple imputation that may return unbiased parameter estimates under MAR.
- **Otherwise, these methods are not recommended.**

Summary: Ad hoc Methods (1)

- *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).

Summary: Ad hoc Methods (2)

  - Regression models where the probability of missingness on Y depends on X values (covariate-dependent missingness. See: Little, 1995, *JASA*).

- In general, when there are missing data, estimates of sample statistics such as means are more biased 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.

Summary: Ad hoc Methods (3)

- NMAR missingness can only be addressed through explicitly assuming a specific missingness mechanism, which can lead to suboptimal results if an 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 (Muthén, Kaplan, & Hollis, 1987, *Psychometrika*).
  - This suggests that it can be beneficial to use methods that assume MAR rather than MCAR missingness and there is likely little downside in doing so.

Methods for MAR Missingness

- Ibrahim (*JASA*, 2005) reviewed four general approaches for handling MAR missingness and found all to perform about equally well:
  - Inverse censoring weights
  - Fully Bayesian analysis
  - Multiple imputation (MI)
  - Full-information maximum likelihood estimation (FIML)
- A full treatment of each technique is beyond the scope of today's presentation. We will concentrate on how to employ Stata to address missingness using full-information maximum likelihood (FIML) and, in Part 2, multiple imputation (MI) under the MAR assumption.

Maximum Likelihood (1)

When there are no missing data:

- Uses the likelihood function to express the probability of the observed data, given the parameters, as a function of the unknown parameter values.
- Example: $L(\theta) = \prod_{i=1}^{n} p(x_i, y_i | \theta)$ where p(x,y|θ) is the (joint) probability of observing (x,y) given a parameter θ, for a sample of *n* independent observations. The likelihood function is the product of the separate contributions to the likelihood from each observation.
- MLEs are the values of the parameters which maximize the probability of the observed data (the likelihood).

Maximum Likelihood (2)

- Under ordinary conditions, ML estimates are:
  - consistent (approximately unbiased in large samples)
  - asymptotically efficient (have the smallest possible variance)
  - asymptotically normal (one can use normal theory to construct confidence intervals and *p*-values).
- The ML approach can be easily extended to MAR situations: $L(\theta) = \prod_{i=1}^{m} p(x_i, y_i \mid \theta) \prod_{j=m+1}^{n} g(y_j \mid \theta)$
- The contribution to the likelihood from an observation with X missing is the marginal: $g(y_j|\theta) = \Sigma_x p(x,y_j|\theta)$
- This likelihood may be maximized like any other likelihood function. Often labeled FIML or direct ML.

X-side and Y-side Missingness

- Some software programs implicitly incorporate FIML 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 (ML and REML), 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.

FIML via SEM Programs (1)

- 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, the AMOS SEM program has had a user-friendly implementation of FIML 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 FIML missing data handling and featured, among other things, the ability to handle categorical and event history/survival outcome variables and hierarchically clustered (multilevel) data structures, with and without complete data.

FIML Programs – Partial List (1)

- AMOS - Analysis of Moment Structures
  - Commercial program licensed as part of SPSS (CAPS has a 10-user license for this product)
  - Fits a wide variety of univariate and multivariate linear regression, ANOVA, ANCOVA, and structural equation (SEM) models
  - Fits ordered categorical and censored data models using Bayesian estimation
  - http://www.spss.com
  - No capabilities at present for analyzing multilevel data
- Mx - Freeware fits a wide variety of SEMs
  - http://views.vcu.edu/mx
- LISREL – LInear Structural RELations
  - Commercial program: http://www.ssicentral.com
  - Features capabilities for analyzing multilevel data

FIML Programs – Partial List (2)

- *lEM* Loglinear & Event history analysis w/ Missing data
  - Freeware DOS program downloadable from the Internet (Jeroen Vermunt)

    http://www.uvt.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html
  - Fits log-linear, logit, latent class, and event history models with categorical predictors.
- M*plus* – Fits a wide variety of models, including multilevel regressions and SEMs
  - http://www.statmodel.com
- SAS PROC CALIS and Stata -sem- routines now enable multiple linear regression, path analysis and SEM with missing data handling via FIML for models with continuous, normal mediators and outcomes.

FIML in Stata's -sem- command: Robust SE

- Stata's -sem- command has the -robust- option to generate robust "sandwich" standard errors
- Stata's -sem- command also has the -cluster- option to generate robust "sandwich" standard errors for multilevel data (one level of clustering)
  - Resulting robust standard errors technically assume incomplete data arise from a mechanism in between MAR and MCAR (see http://www.statmodel.com/discussion/messages/22/1047.html for details) and may perform well in small to moderately-sized samples with non-normality and missing data (Yuan & Bentler, 2000, *Sociological Methodology*, 30(1), 165-200). Initial simulation studies show low SE bias for this estimator with MAR data. (See http://www.statmodel.com/download/webnotes/mc2.pdf). If you are concerned about MAR vs. MCAR, the bootstrap is another option.

Maximum Likelihood Example (1)

2 x 2 Table with missing data

Sex (X=S)	Vote (Y=V) Yes	No	.		Y	N	
Male	28	45	10	(73)	p_{11}	p_{12}	
Female	22	52	15	(74)	p_{21}	p_{22}	
Total	50	97	25	(147)			1

Likelihood function: $L(p_{11}, p_{12}, p_{21}, p_{22}) = (p_{11})^{28}(p_{12})^{45}$
$(p_{21})^{22}(p_{22})^{52}(p_{11}+p_{12})^{10}(p_{21}+p_{22})^{15}$

Maximum Likelihood Example (2)

2 x 2 Table with missing data

$$p_{11} = \left(\frac{28}{73}\right)\left(\frac{73+10}{172}\right) = 0.1851$$

$$p_{12} = \left(\frac{45}{73}\right)\left(\frac{73+10}{172}\right) = 0.2975$$

$$p_{21} = \left(\frac{22}{74}\right)\left(\frac{74+15}{172}\right) = 0.1538$$

$$p_{22} = \left(\frac{52}{74}\right)\left(\frac{74+15}{172}\right) = 0.3636$$

Maximum Likelihood Example (3)

Using ℓ_{EM} for 2 x 2 Table

Input (partial)

```
* R = response (NM) indicator
* S = sex; V = vote;

man 2        * 2 manifest variables
res 1        * 1 response indicator
dim 2 2 2    * with two levels
lab R S V    * and label R
sub SV S     * defines these two subgroups
mod SV       * model for complete
dat [28 45 22 52   * subgroup SV
     10 15]        * subgroup S
```

Output (partial)

***** (CONDITIONAL) PROBABILITIES *****

* P(SV) * complete data only

1 1	0.1851 (0.0311)	0.1905 (0.0324)
1 2	0.2975 (0.0361)	0.3061 (0.0380)
2 1	0.1538 (0.0297)	0.1497 (0.0294)
2 2	0.3636 (0.0384)	0.3537 (0.0394)

* P(R) *

1	0.8547
2	0.1453

Example 1: FIML Linear Regression

- The AIDS Foundation of Chicago administered a questionnaire to 570 HIV-positive men. Variables available for analysis include:
- Gay harassment scale score (the outcome; *n* = 551)
- Race (White, Black, Hispanic, Other; *n* = 569)
- Sexual Orientation (Gay, Straight, Bi, Other; *n* = 548)
- Age in years (*n* = 570)
- Visited doctor in last six months? (yes; no; *n* = 450)
- Months living with HIV (*n* = 559)
- HIV stigma scale score (*n* = 552)
- Internalized heterosexism scale score (*n* = 481)
- Disclosure items: 5-point Likert (none, a few, half, most, all)
  - Close friends know HIV status (*dss1*; *n* = 557)
  - Family members know HIV status (*dss2*; *n* = 552)
- HIV treatment beliefs scale (BMQ concerns; *n* = 556)
- Social support scale (*n* = 562)

Example 1: Analysis Approach

- Research question: What are the associations of age, doctor visit, race, and sexual orientation with experiences of gay harassment?
- If there were no missing data, how would we proceed?
  - We have a continuous outcome, gay harassment for all analyses considered here.
  - Continuous explanatory variable (age): Pearson or Spearman correlation
  - Binary explanatory variable (doctor visit): t-test or analogous two-group non-parametric test
  - Multi-category explanatory variable (race, sexual orientation): OLS regression; ANOVA
  - Multivariable analyses involving all of these plus other control variables: OLS regression/general linear modeling (GLM) framework
- FIML analyses: Because the FIML approach is model-based, uses all information in the likelihood, and is based on first- and second-order moments (i.e, means, variances, and covariances), the analyses are cast in the covariance matrix and multiple regression framework.

Example 1: Linear Regression

- Step 1: Describe the data, including amounts and patterns of missing data
- Step 2: Perform a few bivariable linear regression analyses using the default listwise deletion approach in Stata's -regress- command to illustrate the approach
- Step 3: Perform multivariable linear regression analyses using the default listwise deletion approach in Stata's -regress- command
- Step 4: Perform multivariable linear regression analyses using the default listwise deletion approach in Stata's -sem- command (this is to show how to fit a regression model using -sem- and to demonstrate that the results will be highly similar to what was obtained in Step 3 using -regress-)
- Step 5: Reprise the analysis from Step 4 using FIML via -sem-
- Step 6 (optional): Demonstrate how to perform bivariable FIML analyses via -sem- (oddly, this is a bit more tricky than multivariable analyses)
  - In a real application, you would most likely generate a FIML-based covariance/correlation matrix for bivariate analyses and then perform multivariable regressions for multivariable analyses

Example 1: Linear Regression Results (1)

- Bivariable results (listwise deletion):
  - Age (*n* = 551): Negatively associated with harassment.
  - Six-month doctor visit (*n* = 435): Not associated with gay harassment.
  - Race (*n* = 550): Overall difference in means with Blacks and Hispanics reporting less gay harassment than Whites
  - Sexual orientation (*n* = 540): Overall difference in means with straight-identified persons reporting less gay-harassment than gay-identified individuals.

Example 1: Linear Regression Results (2)

- Bivariable results (FIML using -sem-; *n* = 570):
  - Age: Negatively associated with harassment.
  - Six-month doctor visit: Not associated with gay harassment.
  - Race: Black race negatively associated with gay harassment.
  - Sexual orientation: Straight sexual orientation negatively associated with gay harassment.

Example 1: Linear Regression Results (3)

- Multivariable results (listwise deletion; *n* = 340):
  - Age: Negatively associated with harassment.
  - Six-month doctor visit: Not associated with gay harassment.
  - Race: No overall mean difference; Blacks still report less gay harassment, but Hispanic comparison with Whites is now non-significant.
  - Sexual orientation: No overall mean difference between groups and no paired differences are significant.

Example 1: Linear Regression Results (4)

- Multivariable results (FIML using -sem-; *n* = 570):
  - Age: Negatively associated with harassment.
  - Six-month doctor visit: Not associated with gay harassment.
  - Race: Marginally-significant overall difference in means with Blacks and Hispanics reporting less gay harassment than Whites.
  - Sexual orientation: Overall difference in means with straight-identified person reporting less gay-harassment than gay-identified individuals.

Maximum Likelihood Summary (1)

- FIML advantages:
  - Provides a single, deterministic set of results appropriate under the MAR assumption with a single reportable *N*.
  - Well-accepted method for handling missing values (e.g., in grant proposals and manuscripts); simple to describe
  - Generally fast and convenient

Maximum Likelihood Summary (2)

- FIML disadvantages:
  - Only available for some models via standard software (would need to program other models), though the number of models and programs supporting those models continues to grow
  - Because FIML uses full information and estimates means, variances, and covariances for all variables simultaneously, more care must be taken to ensure convergence, especially when there are large numbers of variables and relatively few numbers of cases.
  - Parametric: may not be robust to violations of distributional assumptions (e.g., multivariate normality)
    - However, robust standard errors seem to work pretty well for inferential purposes (the bootstrap is an alternative).

Part 1 Conclusions

- Planning ahead can minimize cross-sectional non-response and longitudinal loss to follow-up.
- Use of ad hoc methods, while convenient, assume incomplete data arise from an MCAR mechanism (a fairly strict assumption) and can lead to biased results.
- Maximum likelihood methods such as FIML assume MAR (a less stringent assumption) and are readily available for some models/analysis scenarios.
- FIML/direct ML are most convenient for models that are supported by software and when parametric assumptions are met or not too badly violated.
- For scenarios not supported by FIML software programs, consider multiple imputation, which we will discuss in Part 2.