# Handling Missing Data

Tor Neilands, PhD Estie Hudes, PhD, MPH Center for AIDS Prevention Studies Part 1: December 3, 2013

## Contents

- 1. Missing Data Overview
- 2. Preventing Missing Data
- 3. Missing Data Mechanisms
- 4. Handling Missing Data: Ad hoc methods
- 5. Handling Missing Data: Maximum Likelihood (ML)
- 6. Small ML demonstration with binary variables
- 7. Example 1: Linear regression via ML
- 8. Example 2: Linear regression via ML with an auxiliary variable and clustered data
- 9. Conclusions
- 10. Acknowledgements
- 11. References
- 12. Extra Example 3: Logistic regression via ML
- 13. Extra Example 4: Cox regression via ML
- 14. Appendix

# 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 and other forms of planned missingness
    - 3-form design; 2-method measurement design (Graham et al, *Psychological Methods*, 2006)
  - 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. (with lots of testing!)
  - Rigorous retention protocols for participant tracking, etc.
  - Diane Binson's, Bill Woods', and Lance Pollack'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. 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

## MCAR, MAR, NMAR

- From Schafer & Graham, 2002, p. 151: One 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.
  - See appendix for alternative definitions of these terms.

### Missing Data Mechanisms: Example

- 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.</li>
  - 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.</li>

# **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.

## 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
  - Allison (2002): Robust to parameter estimate bias under NMAR missing data for predictor variables (all regression models) and for 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 (i.e., non-invertible) matrices because it uses different pairs of cases for each entry.
- More fundamentally, in regression modeling with multiple X variables where the sample size fluctuates across different pairs of variables, it is difficult to know what N to specify for the analysis.
- Can lead to biased standard errors under MAR.

## Dummy Variable Adjustment

Advocated by Cohen & Cohen (1983). 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 at random 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.
  - Second exception: If the amount of missing data is very small (e.g., 5% or less), then it may not matter what method is used (Roth, 1994).

#### • Otherwise, these methods are not recommended.

### How should we handle missing data?

- It turns out that MCAR is a special case of MAR, so any method that capably addresses MAR missing data should also be able to address MCAR missing data.
- NMAR missingness can only be addressed through explicitly assuming a specific model for how the data became missing, which can lead to suboptimal results if an incorrect missingness model is specified (Allison, 2002).
- There is some evidence that methods that assume MAR missingness may outperform ad hoc approaches, yielding less biased parameter estimates, even when data are missing due to NMAR (Muthén, Kaplan, & Hollis, 1987, *Psychometrika*). Therefore it may be beneficial to use methods that assume MAR rather than MCAR missingness and there is probably generally little downside in doing so.
- Various NMAR models may be used to perform sensitivity analyses to evaluate parameter estimates under different missingness scenarios (see Appendix for a list of several popular NMAR models). NMAR sensitivity modeling is beyond the scope of today's presentation; we will focus on methods for handling MAR missing data.

# Methods for MAR Missingness

- Ibrahim (JASA, 2005) reviewed four general approaches for handling MAR missingness and found all to perform about equally well:
  - Inverse probability of censoring weights (IPCW)
  - Fully Bayesian analysis
  - Full Information Maximum likelihood estimation (FIML)
  - Multiple imputation (MI)
- 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) today in Part 1 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  $p(x, y | \theta)$  is the (joint) probability of observing (x, y) given a parameter  $\theta$ , 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 :  $L(\theta) = \prod_{i=1}^{m} p(x_i, y_i \mid \theta) \prod_{i=m+1}^{n} g(y_i \mid \theta)$
- The contribution to the likelihood from an observation
  - with X missing is the marginal:  $g(y_i|\theta) = \Sigma_x p(x,y_i|\theta)$ 
    - This likelihood may be maximized like any other likelihood function. Often labeled full-information ML (FIML) or direct ML.

Maximum Likelihood Demonstration (1) 2 x 2 Table with missing data\*

<u>Vo</u>	<u>te</u> (Y=V)				
<u>Sex</u> (X=S)	Yes No .		Y	Ν	
Male	28 45 <u>1</u> 0	(73)	<b>p</b> <sub>11</sub>	<b>p</b> <sub>12</sub>	
Female	22 52 15	(74)	р <sub>21</sub>	P <sub>22</sub>	
Total	50 97 <mark>25</mark>	(147)			1

Likelihood function: L(p<sub>11</sub>, p<sub>12</sub>, p<sub>21</sub>, p<sub>22</sub>) = (p<sub>11</sub>)<sup>28</sup>(p<sub>12</sub>)<sup>45</sup> (p<sub>21</sub>)<sup>22</sup> (p<sub>22</sub>)<sup>52</sup> (p<sub>11</sub>+p<sub>12</sub>)<sup>10</sup> (p<sub>21</sub>+p<sub>22</sub>)<sup>15</sup>

\* From Paul Allison, 2002, pp. 15-17

Maximum Likelihood Demonstration (2) 2 x 2 Table with missing data

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

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

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

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

#### Maximum Likelihood Demonstration (3) Using $\ell_{EM}$ for 2 x 2 Table

Output (partial)

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

\*\*\* (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

#### Maximum Likelihood Demonstration (4) Using Stata (Mata) for 2 x 2 Table

capture log close log using lem\_mata.log, replace

// 2 by 2 table, missing values on one margin. // same example as solved by LEM // the SV matrix (col vector) includes the // 4 (s,v) joint probabilities and 2 marginal (s,.) // p is the col vector of 4 joint probabilities Pr(S=s, V=v) // C is the row vector to multiply p; c is the element 1 // We impose the constraint C p = c, that is, the 4 probabilities add up to 1 // Cc is the row vector that is passed to Mata

mata: // start Mata

}

```
mata clear

SV = (28, 45, 22, 52, 10, 15)'

C = (1, 1, 1, 1)

c = (1)

Cc = (C,c)

void myfun(todo, p, SV, Inf, g, H)

{
```

```
Inf =SV[1]*log(p[1]) + SV[2]*log(p[2]) ///
+ SV[3]*log(p[3]) + SV[4]*log(p[4])
```

```
lnf = lnf + SV[5]*log(p[1]+p[2]) + /// SV[6]*log(p[3]+p[4])
```

S = optimize\_init()

optimize\_init\_evaluator(S, &myfun()) // optimize the liklihood function
 S
optimize\_init\_params(S, (.25, .25, .25)) // initial values of p
optimize\_init\_constraints(S, Cc) // constraints
optimize\_init\_argument(S, 1, SV)
optimize(S)
optimize\_result\_V\_oim(S)

p = optimize(S) // estimated probabilities varcov\_p=optimize\_result\_V\_oim(S) // var-cov matrix of estimates var\_p = diagonal(varcov\_p) // variances of estimates se\_p = sqrt(var\_p) // se's of estimates

round(p,.0001) // print out estimated probabilities w/ 4 decimals round(se\_p,.0001) // print out estimated se's w/ 4 decimals

end // exit Mata

log close

exit

#### Maximum Likelihood Demonstration (5) Using Stata (Mata) for 2 x 2 Table



### ML via SEM Programs

- Some of the most important developments in handling non-normal and incomplete data arose in the latent variable (structural equation modeling or SEM) field in the 1990s.
- For many years, the AMOS SEM program has had a user-friendly implementation of FIML missing data handling suitable for use with continuous cross-sectional and longitudinal exogenous (X-side) and endogenous (Y-side) missing data (Some commands in general purpose statistical software programs can handle longitudinal Y-side missing data via maximum likelihood. See the appendix for more regarding X-side and Y-side missingness and software programs).
- In the late 1990s, Bengt and Linda Muthén developed Mplus, a general latent variable modeling program that included FIML missing data handling and featured, among other things, the ability to model categorical and event history/survival outcome variables and hierarchically clustered (multilevel) data structures, with and without complete data, via ML.
  - Examples 3 and 4 demonstrate how to use Mplus to fit logistic regression and Cox regression models with incomplete covariate data.

#### Some Programs Supporting ML Analyses

- Commercial stand-alone SEM programs (e.g., Mplus, LISREL, EQS)
- Mx Freeware fits a wide variety of SEMs
- Loglinear & Event history analysis with Missing data
  - Freeware MS Windows program downloadable from the Internet (Jeroen Vermunt)
  - <u>http://members.home.nl/jeroenvermunt/</u>
  - Fits log-linear, logit, latent class, and event history models with categorical predictors.
- Availability in general purpose packages (ML for all):
  - SPSS AMOS: Continuous endogenous variables via ML; binary and censored endogenous (Y) variables via Bayesian estimation
  - SAS PROC CALIS: Continuous endogenous (Y) variables via ML
  - Stata's -sem- command: Continuous endogenous (Y) variables via ML, with robust standard error option to address non-normal and/or clustered data. These standard errors technically assume incomplete data arise from a mechanism in between MAR and MCAR (see <u>http://www.statmodel.com/discussion/messages/22/1047.html</u> for details) and may perform well in small to moderately-sized samples with nonnormality 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 <u>http://www.statmodel.com/download/webnotes/mc2.pdf</u>.)

# 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 modelbased, 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 bivariate linear regression analyses using the default missing data handling approach in Stata's regress- command
- Step 3: Perform multivariable linear regression analyses using the default listwise deletion approach in Stata's -regresscommand
- Step 4: Perform multivariable linear regression analyses using the default listwise deletion approach in Stata's -semcommand (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, not discussed today): Demonstrate how to perform bivariate 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)

- Bivariate results (pairwise 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.
- For simplicity, pairwise-based bivariate results are reported here. It is
  possible to obtain FIML-based bivariate results; see the do file for this
  example and our presentation from December 2012 to learn how to obtain
  FIML bivariate results using Stata.

### Example 1: Linear Regression Results (2)

- 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 (3)

- 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.

## Example 2: Tobacco and Bars Study (1)

- Dr. Pam Ling and her research group at the UCSF Center for Tobacco Control Research and Education (CTCRE) administered a brief survey to 1,217 young adult bar patrons in San Francisco. The design features clustered data from participants gathered within bars using a 3-form survey design with planned missingness and an auxiliary variable. Variables available for analysis include:
  - Number of days in past 30 the participant (PPT) smoked (the outcome; n = 1145)
  - Age in years from 18-26 (n = 1217)
  - Race (White, Latino, Black, A/PI, Other; n = 1207)
  - Male gender dummy variable (n = 1217)
  - Sexual Orientation (Gay, Straight, Bi, Other; *n* = 1212)
  - PPT considers self a smoker (0 = no; 1 = yes; n = 858)
  - Social network smoking: Sum of ordinal items asking how many friends, partying companions, and coworkers smoke (n = 616)
  - Extraversion index: Sum of ordinal outgoingness items (n = 801)
- This example demonstrates the FIML linear regression analysis of clustered data from a planned missingness design with an auxiliary variable, addict, which measures whether the PPT smokes within a half hour of getting up in the morning (n = 1207).

## Three Form Design (N=1217)

	Venue ID	Days smoked in past 30 (continuous)	Age in years (continuous)	Race (categorical)	Male gende (binary	e Sexual er Orientation y) (categorical)	
Form X	Yes	Yes	Yes	Yes	Yes	Yes	
Form Y	Yes	Yes	Yes	Yes	Yes	Yes	
Form Z	Yes	Yes	Yes	Yes	Yes	Yes	
	Do you consider vourself to be a		How many people in you	Extraversion r Index		Auxiliary: Smokes within	

	yourself to be a smoker? (binary n=858)	How many people in your social network smoke? (continuous n=616)	Index (continuous n=801)	Smokes within 30 min of waking (Binary Yes/No)
Form X	Yes	no	Yes	Yes
Form Y	no	Yes	Yes	Yes
Form Z	Yes	Yes	no	Yes

## Example 2: Tobacco and Bars Study (2)

- OLS regression using the standard approach is not possible because listwise deletion yields a data set with zero observations for the desired model.
- The analysis is straightforward using FIML with -sem- in Stata, PROC CALIS in SAS, AMOS in SPSS, or any specialized SEM program like Mplus.
- The Stata analysis addresses clustering due to participants being nested within recruitment sites (i.e., bars) through the use of robust standard errors.
- This example also includes an auxiliary variable, addict. *Auxiliary variables* are variables that are either (a) correlated with one or more of the observed variables in the analysis or (b) correlated with missingness on one or more variables that have missing data. These variables should only be included in the analysis if they are strongly correlated with observed values or missingness of the other variables already in the analysis (see Collins et al., A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 2001).
  - Several possible methods are available for including auxiliary variables in Stata FIML analyses. See <u>http://www.stata.com/meeting/new-</u> <u>orleans13/abstracts/materials/nola13-medeiros.pdf</u> for further details.

### Example 2: Linear Regression Results

- There are no significant effects for age, race, male gender, and extraversion.
- There is a significant overall difference for LGBT status, with gay
  participants having a lower mean number of smoking days
  relative to the straight participant reference group.
- Self-identification as a smoker is positively associated with the number of days smoked.
- More smoking occurring in one's social network is associated with more days smoked.

#### Example 2: Limitations and Further Thoughts

- The number of days smoked is not normally distributed. Parameter estimates should be somewhat robust to non-normality and robust variances help protect inferences from assumption violations. However, it could be beneficial to examine the robustness of the results via multiple imputation in which smokdays is imputed under less restrictive assumptions. We'll revisit this issue in Part 2.
  - Item-level missingness: With the three-form design, there should've been roughly 800 participants per survey form, but there are only 616 for the networking smoking variable. Why? Some respondents endorsed "not applicable" for one or more items. How to handle this type of situation?
    - Restrict the sample to only those who have valid responses for all three variables. Loses information and restricts the inference space.
    - Model network smoking as a latent factor representing the shared variance among the three network items. A more complicated proposition for binary or ordinal items and more complex to justify and report.
    - Compute the mean of the three items rather than the sum. Equivalent to mean substitution and therefore assumes MCAR missingness.
    - Impute responses at the item level via MI. May not be appealing if the answers for those questions should actually be "not applicable". Also, there are practical limitations on the number of items included in the same imputation run.
    - Two-part modeling at the item-level assuming separate "applicable" and N/A populations. Complex; probably not realistic for large numbers of items.

### **ML** Issues and Extensions

- Interactions between variables are handled seamlessly as part of the model.
- Availability of some regression-type model diagnostics may be more limited in the ML context (e.g., observation-level predicted value-by-residual scatterplots). Mplus features some case-deletion diagnostics (e.g., Cook's D) and, if available in software, robust standard errors may be compared with model-based standard errors as a crude gauge of how well normality and constant residual variance assumptions are met.
- What about generalized linear models for binary outcomes, count outcomes, and failure time (i.e., survival) outcomes? Earlier we demonstrated how LEM can be used to perform ML analyses for models involving exclusively categorical variables. Mplus features ML estimation for binary, ordinal, count, and nominal (i.e., multinomial) outcomes. Mplus allows the user to bring binary or continuous covariates with missing values into the models for these outcomes, switching covariates' status from fixed to random with a normal distribution. This is done by naming the variances and covariances of the covariates as explicit parameters to be estimated.
- Multilevel models with missing X-variables? Mplus can perform ML estimation for two-level models with most outcome variable types and bring covariates into the model as random variables as described above.

# Maximum Likelihood Summary (1)

- ML 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
  - Avoids a lot of the decision points involved in performing multiple imputation (see <a href="http://www.statisticalhorizons.com/ml-better-than-mi">http://www.statisticalhorizons.com/ml-better-than-mi</a>), including the complexities of dealing with situations where some cases' data need to be imputed, but others should have structurally missing data (e.g., number of pregnancies for males).

# Maximum Likelihood Summary (2)

#### • ML 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 ML 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) and some of the usual model diagnostic tools may not be as readily available as they are for standard regression methods.
  - 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 software programs with ML, consider multiple imputation, which we will discuss in Part 2.

# Acknowledgements

- NIMH P30 MH062246 (Morin, PI). Methods Core (Neilands, Hudes) supported time to prepare examples and slides. Also, NCI R01 CA141661 (Ling, PI) and U01 CA154240 (Ling, PI) and P60 MD006902 (Bibbins-Domingo, PI) provided additional support (Neilands).
- AIDS Foundation of Chicago: Gay harassment data
- Pamela Ling, MD: Tobacco and bars data; three-form design table
- Elvin Geng, MD: Africa mortality data
- David D. Burns, MD: Early and ongoing mentoring on AMOS and FIML
- Adam Carle, PhD: -runmplus- awareness
- Richard Jones, Sc.D.: -runmplus- programming and support
- Isabel Cannette, Stata Tech Support: -gsem- syntax to estimate models with missing data
- Steve Gregorich, PhD: Overall mentoring on missing data; review of previous slides
- Melissa Krone and Jesse Canchola: Contributions to early versions of this presentation in the early 2000s.

## References

- Allison, P. (2002). <u>Missing Data.</u> Thousand Oaks, CA, Sage Publications.
- Buhi, E. R., et al. (2008). "Out of sight, not out of mind: Strategies for handling missing data." <u>American Journal of Health Behavior</u>, **32**(1): 83-92.
- Collins, L. M., et al. (2001). "A comparison of inclusive and restrictive strategies in modern missing data procedures." <u>Psychological Methods</u>, **6**(4): 330-351.
- Cohen J. and Cohen P. (1983) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2<sup>nd</sup> ed. LEA, NJ
- Davey, A. and Savla, J. (2010). <u>Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach</u>. New York: Routledge Academic.
- Dong, Y. and Peng CY. (2013).Principled missing data methods for researchers. *Methodology*, 2: 222. Open Access: <u>http://www.springerplus.com/content/2/1/222.</u> Accessed on November 26, 2013.
- Enders, C.K. (2010). <u>Applied Missing Data Analysis</u>. New York, NY: Guilford.
- Graham, J.W. (2012). <u>Missing Data: Analysis and Design</u>. New York: Springer.
- Graham, J. W. (2009). "Missing Data Analysis: Making It Work in the Real World." <u>Annual Review of Psychology</u>, 60: 549-576.
- Graham, J. W., et al. (1996). "Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures." <u>Multivariate Behavioral Research</u>, **31**(2): 197-218.
- Ibrahim, J. G., et al. (2005). "Missing Data Methods for Generalized Linear Models: A Review." Journal of the American Statistical Association, 100(469): 332-346.
- Jones, M. P. (1996). "Indicator and stratification methods for missing explanatory variables in multiple linear regression." Journal of the American Statistical Association, **91**: 222-230.
- Leon, A. C., et al. (2007). "Bias reduction with an adjustment for participants' intent to dropout of a randomized controlled clinical trial." <u>Clin Trials</u>, 4(5): 540-547.
- Little, R. J. A. and D. B. Rubin (2002). Statistical Analysis with Missing Data. New York, John Wiley and Sons.
- Muthén, B., et al. (1987). "On structural equation modeling with data that are not missing completely at random." <u>Psychometrika</u>, **52**(3): 431-462.
- Roth, P. L. (1994). "Missing data: A conceptual review for applied psychologists." <u>Personnel Psychology</u>, **47**(3): 537-560.
- Rubin, D. B. (1976). "Inference and missing data (with discussion)." <u>Biometrika</u>, **63**: 581-592.
- Rubin, D. B. (1987). <u>Multiple Imputation for Nonresponse in Surveys</u>. New York, John Wiley and Sons.
- Schafer, J. L. and J. W. Graham (2002). "Missing data: Our view of the state of the art." <u>Psychological Methods</u>, **7**(2): 147-177.
- Yuan, K.-H. and P. M. Bentler (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. Sociological Methodology 2000. M. E. Sobel. Washington, DC, American Sociological Association: 165-200.

#### Example 3: ML Logistic Regression with Bar Data

- Revisiting the tobacco and bars data set, what if we wanted to know what the associations of the previously studied explanatory variables with daily smoking (yes/no) are? (variable: smkdaily)
- Ordinarily one would fit a logistic regression model using the Stata -logistic- command, but that is not possible in this example because the listwise deletion intersection of explanatory variables yields zero observations.
- Instead, we can use maximum likelihood estimation in Mplus to fit the model.
- We'll use the user-written Stata command file -runmplus-\* to pass the data from Stata to Mplus and display the Mplus results within Stata.
  - runmplus is written by Richard Jones and may be obtained from: <u>https://sites.google.com/site/lvmworkshop/home/runmplus-stuff</u>.
  - This site also features various utilities that work with Mplus and -runmplus-, including a handy Stata ado program, Ili.ado, for comparing nested models using the robust likelihood ratio test.

\* We appreciate Dr. Adam Carle recommending -runmplus- to us.

#### Example 3: Results and Summary

#### Results

- There is an overall effect for race (Wald chi-square = 13.06, p = .011)
  - Latinos have a lower odds of smoking relative to Whites (OR = .70; p = .003)
  - Other race ethnic group members also have a lower odds of smoking relative to Whites (OR = .57; p = .024)
- There is an overall effect for sexual orientation (Wald chi-square (3) = 16.24, p = .001)
  - Bisexuals have a higher odds of smoking relative to heterosexuals (OR = 3.41; p = .006)
- Self-identified smokers have higher odds of daily smoking relative to self-identified nonsmokers (OR = 21.69; p < .001).</li>
- For every one-unit increase in tobacco exposure through one's social network, the odds
  of reporting daily smoking increase by 19% (OR = 1.21; p < .001).</li>
- Extraversion is positively associated with being a daily smoker (OR = 1.16; p = .040).
- Features of the Analysis
  - Maximum likelihood handling of missing data with a binary outcome.
  - Robust standard errors address clustering of participants within bars.
  - Auxiliary variable contributes information through explaining the outcome and by being correlated with other explanatory variables.

#### Example 4: Cox Regression via ML

- Cohort study conducted by Dr. Elvin Geng featuring N = 33,947 research participants with HIV from Kenya, Tanzania, and Uganda.
- Outcome: Time to death. Implies Cox proportional hazards model.
- 1,082 cases excluded due to having zero observation time.
- Predictors (listwise *n* = 26,883; 82% of the sample of 32,865)
  - Country: (1 = Kenya; 2 = Tanzania; 3 = Uganda). n = 32,865
  - Sex (0 = female; 1 = male). n = 32,865
  - Age at study entry. n = 32,463 (402 missing)
  - Already in care (0 = no; 1 = yes). n = 32,865
  - Pre-therapy CD4 T-cell count modeled by three restricted cubic spline variables. n = 29,347 (3,518 missing)
  - Tuberculosis infection at study start (0 = no; 1 = yes). n = 30,292 (2,573 missing)
- Case weight included to improve estimates based on random sampling and subsequent re-contacting of participants who were originally lost to followup.
- Clustering present due to participants being nested within clinics.
- Use Mplus (via -runmplus- in Stata) to perform the Cox regression analysis using direct ML.

#### Example 4: Cox ML Regression Results

#### Results

- Listwise:
  - There is more than a two-fold hazard of death being in country 2 (Tanzania) versus the reference country (Kenya; HR = 2.40; p < .001).</li>
  - There is a lower hazard of death being in country 3 (Uganda) versus the reference country (Kenya; HR = 0.62; p < .001).</li>
  - Males have a higher hazard of death relative to females (HR = 1.43, p = .009).
  - The hazard of death increases 16.5% for every 10 year increase in age (HR = 1.165, p <.001).</li>
  - The linear component of CD4 is negatively associated with hazard of death (HR = .880, p < .001), though the cubic terms are also significant, so the association is likely non-linear.
  - Being in care is strongly negatively associated with the hazard of death (HR = 0.38, p < .001).</li>
  - Having TB means a higher hazard of death (HR = 1.31, p = .001).
- ML: Similar to listwise, although country 3 (Uganda) relative to country 1 (Kenya) is no longer significant (HR = .84; p = .566).
- Features of the analysis
  - Continuous failure time outcome variable
  - Missing values on predictors handled seamlessly by direct ML estimation under the MAR assumption, avoiding the complexities inherent in imputing survival data via multiple imputation
  - Clustering due to clinic handled through robust standard errors
  - Case weights incorporated into the analysis

## Appendix (1)

#### Missing at Random (MAR)

• Denote Y*complete* as the complete data. Partition Y*complete* as:

Ycomplete = (Yobserved, Ymissing)

- Define R as an indicator of (non)missingness for variable Y.
   R = 1 if Y is observed; R = 0 if Y is missing.
- MAR holds 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*|Ycomplete) = P(R|Yobserved)

## Appendix (2)

#### 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 Yobserved, also:
   P(R|Ycomplete) = P(R)

• If incomplete data are MCAR, the cases with complete data are then a random subset of the original sample.

## Appendix (3)

#### 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
  - Weighted multiple imputation
- Disadvantages of NMAR modeling: Requires high level of knowledge about missingness mechanism; results are often sensitive to the choice of NMAR model selected (Allison, 2002)

## Appendix (4)

#### Ignorability

- Ignorable data missingness occurs when data are incomplete due to MCAR or MAR processes (Allison, 2002)
- 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 methods (and software that implemens those methods) that take the missingness mechanism into account
- 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).

## Appendix (5)

#### A Few Words About X-side and Y-side Missingness

- Some software programs implicitly incorporate ML handling of an outcome variable Y under the MAR assumption. 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 which use ML estimation (there are many) and user-written MLbased analysis commands (e.g., -gllamm-)
- However, these commands will drop the observation row when one or more X values in that row are missing.
- These commands are very useful for analyzing longitudinal data with no missing covariates (e.g., complete baseline covariate data).
- They cannot conveniently be used to handle cross-sectional missing data or longitudinal data with missing covariates.