Handling Missing Data

Tor Neilands, PhD
Estie Hudes, PhD, MPH
Center for AIDS Prevention Studies
Part 1: February 13, 2015

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 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

The Scary Box

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 probability-based 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: January 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.

Ad-hoc Approaches to Handling Missing Data

- Listwise deletion (a.k.a. complete-case analysis)
- 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

- 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)
- Consequences of listwise deletion of missing data:

	MCAR	MAR	NMAR
Biased Parameter Estimates	No	Yes	Yes
Biased Standard Errors	Yes	Yes	Yes
Reduced Power for Hypothesis Tests	Yes	Yes	Yes

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 does not use the same pairs of cases for each entry.
- 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.
- **If the amount of missing data is very small (e.g., 5% or less), then it may not matter much what method is used (Roth, 1994). Otherwise, these methods are not recommended.**

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
  - **Maximum likelihood estimation (FIML) – Part 1 (today)**
  - **Multiple imputation (MI) – Part 2 (next presentation)**
- Since MCAR missingness is a subsumed under MAR missingness, these methods will work for both MCAR and MAR missing data.
- 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*).

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 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 :

$$L(\theta) = \prod_{i=1}^{m} p(x_i, y_i \mid \theta) \prod_{i=m+1}^{n} g(y_i \mid \theta)$$

- The contribution to the likelihood from an observation with X missing is the marginal: $g(y_i \mid \theta) = \Sigma_x p(x, y_i \mid \theta)$
  - 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

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 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 *lEM* 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

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,  lnf, g, H)
      {
     lnf =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, .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  row of estimated se's  w/ 4
      decimals

end // exit Mata

log close

exit
```

Maximum Likelihood Demonstration (5)

Using Stata (Mata) for 2 x 2 Table

```
:    round(p,.0001)  // print out estimated probabilities w/ 4
   decimals

             1       2       3       4
    +--------------------------------+
  1 |  .1851   .2975   .1538   .3636  |
    +--------------------------------+

:    round(se_p,.0001)  // print out row of estimated se's w/ 4
   decimals

    +--------------------------------+
  1 |  .0311   .0361   .0297   .0384  |
    +--------------------------------+
:
: end // exit Mata
------------------------------------------------------------------------
```

ML for Regression Analyses

- Available in two main types of software programs:
  - Structural equation modeling (SEM) programs, which can fit models for continuous outcomes and mediating variables (if applicable) under the joint multivariate normality assumption.
    - Some SEM programs (e.g., Stata, M*plus*) feature robust standard errors for obtaining correct inferences with clustered data. Robust standard errors are also less vulnerable to violations of normality and constant variance assumptions.
    - Some SEM programs also allow for non-continuous mediators and outcomes.
- Mixed models programs, which are especially useful for longitudinal data sets where only dependent variables have missing data.
- See Appendix for a partial list of programs.

Example 1: 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. 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)
  - Participant 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 data from a planned missingness design using Stata and SPSS AMOS.
  - For illustrative purposes, venue clustering is not accounted for in this example.

Three Form Design (N=1217)

	Venue ID	Days smoked in past 30 (continuous)	Age in years (continuous)	Race (categorical)	Male gender (binary)	Sexual Orientation (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 yourself to be a smoker? (binary n=858)	How many people in your social network smoke? (continuous n=616)	Extraversion Index (continuous n=801)	Smokes within 30 min of waking (Binary Yes/No; can be used as an auxiliary variable)
Form X	Yes	no	Yes	Yes
Form Y	no	Yes	Yes	Yes
Form Z	Yes	Yes	no	Yes

Example 1: Tobacco and Bars Study (2)

- OLS regression is not possible because listwise deletion yields a data set with zero observations for the model.
  - This is an extreme form of the more typical problem where some subset of cases are complete, but a significant number of cases are incomplete.
- The analysis is straightforward using ML with -sem- in Stata, PROC CALIS in SAS, AMOS in SPSS, or a stand-alone SEM program with missing data ML (e.g., M*plus)*.
- To keep things simple, we did not include venue ID as a clustering variable in this example, but this can be done easily in Stata via specifying the -vce(cluster *clusterid*)- option and in M*plus*, as we demonstrate below in Example 2.
- Although we did not demonstrate it here, we could have used the smoking within 30 minutes of waking up variable as an *auxiliary variable* to add additional information to the analysis. An auxiliary variable is a variable that is (strongly) correlated with observed values for other variables in the analysis or is (strongly) correlated with missingness in variables that have missing data.
  - It is possible to include auxiliary variables in ML-based analyses
  - If you have more than a few auxiliary variables, it is more straightforward to include them via multiple imputation (MI). We cover MI in Part 2.
- See the appendix for more on auxiliary variables and for links on how to include auxiliary variables in Stata and SAS ML analyses.

Example 1: Results

- The results are based on maximum likelihood estimation (also known as full-information maximum likelihood [FIML] when the dependent variable is continuous & normality is assumed)
- There are no significant effects for age, race, male gender, and extraversion.
- There is a significant overall difference for LGBT status ($\chi^2[3] = 21.68$, p=.0001), with gay (*B* = -2.75, p=.005) participants having a lower mean number of smoking days relative to the straight participant reference group and bisexual (*B* = 3.05, p=.008) and other (*B* = 3.00, p=.016) having a higher mean number of smoking days relative to straight participants.
- Self-identification as a smoker is positively associated with the number of days smoked (*B* = 14.42, p<.001).
- More smoking occurring in one's social network is associated with more days smoked (*B* = .41, p=.006).

Example 2: 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 a general purpose logistic regression program, but we cannot do that because the listwise deletion of independent variables results in zero observations.
- Instead, we can use maximum likelihood estimation in M*plus* to fit the model. This time we'll illustrate including venue ID as a cluster variable to account for the clustering of participants in bars.
- We'll use the user-written Stata command file -runmplus-* to pass the data from Stata to M*plus* and display the M*plus* results within Stata. (You can use M*plus* directly without Stata if you want).
  - runmplus is written by Richard Jones and may be obtained from: https://sites.google.com/site/lvmworkshop/home/runmplus-stuff.
  - This site also features various utilities that work with Mplus and -runmplus-, including a handy Stata ado program, lli.ado, for comparing nested models using the robust likelihood ratio test.

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

Example 2: Results and Summary

- There is an overall effect for race (Wald chi-square with 4 DF = 11.86, p = .0184)
  - Latinos have a lower odds of daily smoking relative to Whites (OR = .70; p = .005)
  - Other race ethnic group members also have a lower odds of daily smoking relative to Whites (OR = .59; p = .033)
- There is an overall effect for sexual orientation (Wald chi-square with 3 DF = 15.68, p = .0013)
  - Bisexuals have a higher odds of daily smoking relative to heterosexuals (OR = 3.21; p = .009)
- Self-identified smokers have higher odds of daily smoking relative to self-identified non-smokers (OR = 21.36; p < .001).
- 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).
- Extraversion is positively associated with being a daily smoker (OR = 1.16; p = .049).

Example 3: Longitudinal Analysis

- Duncan, Moskowitz, Neilands, Dilworth, Hecht & Johnson developed a Mindfulness-Based Stress Reduction (MBSR) intervention to reduce symptoms experienced and bother/distress from taking antiretroviral HIV medications. [*J. Pain Symptom Mange*. 43(2), 161-171, 2012.]
- *N*=76 people living with HIV who were actively taking ART and reported distress from ART-related side effects were randomly assigned to an MBSR program or a wait-list control (WLC) standard care condition.
- Study retention was adequate, with 86% (*N*=65) completing three-month follow-up assessments and 93% (*N*=71) completing assessments at the six-month follow-up.
- Primary Independent Variable: Intervention group (0 = control; 1 = MBSR)
- Dependent Variables: Sum of the number of symptoms reported (log-transformed) and average bother attributed to symptoms
- Analysis: Traditional general linear model (GLM)-based repeated measures ANOVA dropping cases with incomplete data and linear mixed models analysis using SAS PROC MIXED using all cases.

Example 3: Longitudinal Analysis Results

- If cases with incomplete data are excluded, the analysis *N* drops from 76 to 62, an 18% reduction.
- If cases with incomplete data are excluded, either automatically via the traditional GLM method or manually for the mixed models method, we conclude there is no group-by-time interaction.
- If cases with partial data are retained in the analysis using maximum likelihood estimation, there is a significant group-by-time interaction effect ($F[2,132]=3.87$, $p=.023$) and several significant follow-up findings showing MBSR participants had fewer reported symptoms and bother attributable to those symptoms (see Duncan et al for further details) at follow-up.
- The ML method is implemented automatically for continuous, normally distributed missing *dependent* variables in mixed models programs such as SAS PROC MIXED, SPSS MIXED, and Stata -mixed-.
- Several software programs support similar estimation commands for non-continuous data (e.g., binary, ordinal, counts). Examples include SAS PROC GLIMMIX and Stata's -meglm- family commands (e.g., -melogit-).
- Missing *independent* variables can still be a problem in longitudinal analyses. Possible remedies include recasting the analysis within the structural equation modeling framework as a latent growth curve model (LGCM) or using multiple imputation (MI), which is described in Part 2.

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 http://www.statisticalhorizons.com/ml-better-than-mi), 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 fully robust to violations of distributional assumptions (e.g., multivariate normality) and some of the usual regression 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 ML assume MAR (a less stringent assumption) and are readily available for some models/analysis scenarios.
- ML is 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 (Lightfoot, PI): Methods Core (Neilands, Hudes) supported time to prepare examples and slides. Also, NCI R01 CA141661 (Ling, PI) and U01 CA154240 (Ling, PI), P60 MD006902 (Bibbins-Domingo, PI), and R21 AT003102 (Johnson, PI) provided additional support (Neilands).
- Pamela Ling, MD: Tobacco and bars data; three-form design table
- Mallory Johnson, PhD: MBSR study data
- 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
- 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). Missing Data. Thousand Oaks, CA, Sage Publications.
- Buhi, E. R., et al. (2008). "Out of sight, not out of mind: Strategies for handling missing data." American Journal of Health Behavior, **32**(1): 83-92.
- Collins, L. M., et al. (2001). "A comparison of inclusive and restrictive strategies in modern missing data procedures." Psychological Methods, **6**(4): 330-351.
- Cohen J. and Cohen P. (1983) Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences, 2nd ed. LEA, NJ
- Davey, A. and Savla, J. (2010). Statistical Power Analysis with Missing Data: A Structural Equation Modeling Approach. New York: Routledge Academic.
- Dong, Y. and Peng CY. (2013).Principled missing data methods for researchers. *Methodology*, 2: 222. Open Access: http://www.springerplus.com/content/2/1/222. Accessed on November 26, 2013.
- Enders, C.K. (2010). Applied Missing Data Analysis. New York, NY: Guilford.
- Graham, J.W. (2012). Missing Data: Analysis and Design. New York: Springer.
- Graham, J. W. (2009). "Missing Data Analysis: Making It Work in the Real World." Annual Review of Psychology, **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." Multivariate Behavioral Research, **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." Clin Trials, **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." Psychometrika, **52**(3): 431-462.
- Roth, P. L. (1994). "Missing data: A conceptual review for applied psychologists." Personnel Psychology, **47**(3): 537-560.
- Rubin, D. B. (1976). "Inference and missing data (with discussion)." Biometrika, **63**: 581-592.
- Rubin, D. B. (1987). Multiple Imputation for Nonresponse in Surveys. New York, John Wiley and Sons.
- Schafer, J. L. and J. W. Graham (2002). "Missing data: Our view of the state of the art." Psychological Methods, **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.

Appendix (1)

Missing at Random (MAR)

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

  Y*complete* = (Y*observed*, Y*missing*)

- 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*|Y*complete*) = P(R|Y*observed*)

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 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.

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. Examples:
  - PROCs MIXED, GLIMMIX, and NLMIXED in SAS
  - MIXED in SPSS
  - Stata commands for longitudinal and clustered data which use ML estimation (there are many such as -mixed- and -melogit-) and user-written ML-based 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. For missing covariate data, consider alternatives:
  - Use structural equation modeling software
  - Use multiple imputation (MI) followed by standard analyses of the imputed data

Appendix (6): Software (Partial List)

- Commercial stand-alone SEM programs (e.g., M*plus*, LISREL, EQS)
- Mx - Freeware fits a wide variety of SEMs
- *ℓEM* Loglinear & Event history analysis with Missing data
  - Freeware MS Windows program downloadable from the Internet (Jeroen Vermunt)
  - http://members.home.nl/jeroenvermunt/
  - 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 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 .)

Appendix (7): Auxiliary Variables

- *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 methods are available for including auxiliary variables in Stata FIML analyses. These are illustrated in:
  - http://www.stata.com/meeting/new-orleans13/abstracts/materials/nola13-medeiros.pdf (Stata)
  - http://www.statisticalhorizons.com/wp-content/uploads/MissingDataByML.pdf (SAS)
- M*plus* has a convenience feature for including auxiliary variables in ML analyses involving continuous mediators and outcomes.
- If you have more than a few auxiliary variables, it can be more straightforward to use them in multiple imputation (MI) rather than in direct ML-based or FIML-based analyses. In MI, auxiliary variables are simply added to the imputation model along with the analysis variables.