Best Practices and Missing Data in SEM

Best Practices (Kline, Ch 13)

1. Specify the model before data are collected
2. Include all paths/associations that are substantively meaningful (no omitted paths and no omitted covariances)
3. Use psychometrically adequate measures (unreliability propagates across a model!)
4. Think about directionality and make no assumptions of causality without experimental data
5. Never add correlated residuals that are not theoretically justified and include those that are!
6. Remember parsimony!
7. Include enough indicators and remember rules of identification
8. Remember – the main goal of specification is to test a theory, not a model! And all of our models are wrong!

Best Practices (Kline, Ch 13)

8. Check the accuracy and distributions of your data, examine outliers, patterns of missing data
9. Consider potential nonlinearities
10. Do not ignore clustering (i.e., non-independence of data)
11. Watch out for empirical underidentification, non-sense values, large residuals, large residual correlations
12. Check the solution for admissibility and convergence, do the results "make sense?"
13. Carefully screen all of the output
14. Complex models require large samples
15. Report unstandardized and standardized estimates

Best Practices (Kline, Ch 13)

16. Consider fit statistics along with other information (correlation residuals, size of coefficients, variance predicted)
17. Make scales of variables meaningful, centering coefficients when necessary
18. Test invariance whenever examining group differences
19. Ignore a significant chi-square test
20. Use theory to guide model building, rather than relying solely on statistical criteria or rules of thumb
21. Always consider equivalent and near-equivalent models
22. Always report enough information for the reader to reproduce your results

Missing Data

Basics

- Definition: Data are missing on some variables for some observations.

- Problem: How to do statistical analysis when data are missing? Three goals:
  - Minimize bias
  - Maximize use of available information
  - Get good estimates of uncertainty

- Not a goal: imputed values "close" to real values.

Missing Data Mechanisms

- Missing complete at random (MCAR)
  - Probability of missing data is completely unsystematic.
  - Suppose some data are missing on Y. These data are said to be MCAR if the probability that Y is missing is unrelated to Y or other variables X.

$$P(Y\ is\ missing|X,Y) = P(Y\ is\ missing)$$

  - MCAR is the ideal(unrealistic) situation
  - If data are MCAR, complete data subsample is a random sample from original target sample → listwise deletion is appropriate.

MCAR Example

- Employees complete an IQ test during a job interview.
- A number of employees quit prior to the 6-month review
- Performance ratings are missing for no particular reason (e.g., maternity leave, spouse relocates, found higher paying job)

IQ	Performance (Hypothetical)	Performance (Observed)
78	9	
84	13	13
84	10	
85	8	8
87	7	7
91	7	7
92	9	9
94	9	9
94	11	11
96	7	
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	
113	12	12
115	14	14
118	16	16
134	12	

Missing Data Mechanisms

- Missing at random (MAR)
  - Systematic missingness, where missing data is related to other measured variables in the analysis.
  - Data on Y are MAR if the probability that Y is missing does not depend on the value of Y, after controlling for other variables X.

$$\mathrm{P}(Y\ is\ missing|X, Y) = \mathrm{P}(Y\ is\ missing|X)$$

  - Considerable weaker assumption than MCAR
  - This is the assumption which people will be working with most of the time.

MAR Example

- Prospective employees complete an IQ test during a job interview.
- The company use IQ as a selection measure and does not hire applicants in the lowest quartile.
- The missing performance scores depends on observed IQ scores. But after controlling for IQ, the missing performance scores does not depends on performance.

IQ	Performance (Hypothetical)	Performance (Observed)
78	9	
84	13	
84	10	
85	8	
87	7	
91	7	7
92	9	9
94	9	9
94	11	11
96	7	7
99	7	7
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

Missing Data Mechanisms

- Missing not at random (MNAR)
  - Probability of missing data on Y is related to the the would-be values of Y itself.
  - Cannot test between MNAR vs. MAR, often rely on substantive knowledge to the data.
  - NMAR is problematic and introduces bias when the would-be outcome scores determine missingness.
  - NMAR requires specialized analysis procedures (e.g., selection models, pattern mixture models).

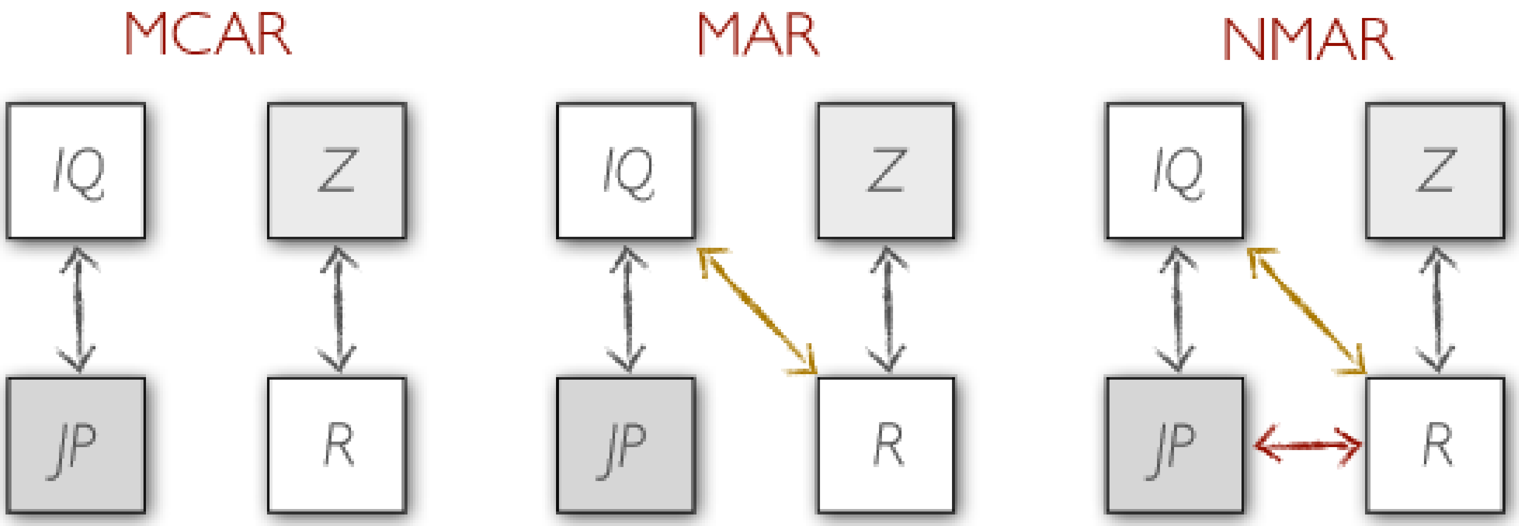
NMAR Example

- Employees complete an IQ test during a job interview
- The company terminates low-performing employees prior to their evaluation
- The would-be performance ratings determine missing data on the performance measure

IQ	Performance (Hypothetical)	Performance (Observed)
78	9	9
84	13	13
84	10	10
85	8	
87	7	
91	7	
92	9	9
94	9	9
94	11	11
96	7	
99	7	
105	10	10
105	11	11
106	15	15
108	10	10
112	10	10
113	12	12
115	14	14
118	16	16
134	12	12

Diagram of Mechanisms

- R is a binary missing data indicator for job performance ratings
- Z is a correlate or cause of missingness not in the data



Approaches for Handling Missing Data

- Conventional
  - Listwise deletion (complete case analysis)
    - If data are MCAR, does not introduce any bias in parameter estimates.
    - May delete a large proportion of cases, resulting in loss of statistical power.
    - Robust to NMAR for predictor variables in regression analysis
  - Pairwise deletion (available case analysis)
    - Approximately unbiased if MCAR
    - Uses all available information
    - Standard errors in correct

Approaches for Handling Missing Data

- Conventional
  - Dummy variable adjustment (Cohen & Cohen, 1985)
    - Produces biased coefficient estimates (Jones, 1996)
  - Imputation (any method that substitutes estimated values for missing values)
    - <u>Replacement with means</u>
    - <u>Regression (replace with conditional means):</u> use some predictors to predict the variables with missing, then use the regression model to generate predicted values for the cases with missing data.
    - <u>Hot deck</u>: Divide sample into homogeneous strata on observed variables. Within each stratum pick "donor" units with observed values to fill in missing values for other units.
    - Problems:
      - Often leads to biased parameter estimates.
      - Usually leads to smaller standard error estimates.

Approaches for Handling Missing Data

- Modern
  - Maximum likelihood
    - Factoring the likelihood for monotone missing data patterns.
    - EM algorithm.
    - Direct maximization of the likelihood (the focus for today).
  - Multiple imputation

Direct ML

- Also known as “raw” ML or “full information” ML.
- Directly maximizes the likelihood for the model of interest. Produces “consistent” estimates of the standard errors.

- Without missing data, the multivariate normal likelihood is

$$L(\theta) = \prod_i f(y_i | \mu(\theta), \sum(\theta))$$

Direct ML (cont.)

- With missing data, the likelihood becomes

$$L(\theta) = \prod_i f(y_i | \mu_i(\theta), \sum\nolimits_i(\theta))$$

- If data are missing for individual i, then y_i deletes the missing values, u_i deletes the corresponding means, and Σ_i deletes the corresponding rows and columns. This result follows from integrating the likelihood over the variables with missing data.

- This likelihood can be maximized by conventional methods, e.g., the Newton-Raphson algorithm.

College Example

1994 U.S. News Guide to Best Colleges

- 1302 four-year colleges in U.S.
- Goal: estimate a regression model predicting graduation rate (no. graduating/no. enrolled 4 years earlier X 100)
- 98 colleges have missing data on graduation rate

Independent variables:

- 1st year enrollment (logged, 5 case missing)
- Room & Board Fees (40% missing)
- Student/Faculty Ratio (2 cases missing)
- Private=1, Public=0
- Mean Combined SAT Score (40% missing)
- Auxiliary variable: Mean ACT scores (45% missing)

FIML with Mplus

```
DATA:
FILE IS college.csv;

VARIABLE:
    NAMES ARE gradrat lenroll rmbrd private stufac csat act;
    USEVARIABLE ARE gradrat lenroll rmbrd private stufac csat;
    MISSING ARE ALL (9999);

MODEL:
    gradrat ON lenroll rmbrd private stufac csat;

    !making exongenous variables to be endogenous variables
    lenroll;
    rmbrd;
    private;
    stufac;
    csat;

OUTPUT:
```

FIML with Lavaan

```
data.college <- read.table ("C:\\Users\\yuyuhsiao\\Dropbox\\UNM\\Courses\\607\\(14) Best Practice
head(data.college)


model.college <- "
gradrat ~ lenroll + rmbrd + private + stufac + csat
"
fit.college <- sem(model=model.college, data=data.college, missing="fiml", fixed.x=FALSE)
summary(fit.college)
varTable(fit.college)
```

Output: Mplus

SUMMARY OF ANALYSIS

Number of groups	1
Number of observations	1302
Number of dependent variables	1
Number of independent variables	5
Number of continuous latent variables	0

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value
GRADRAT ON				
LENROLL	2.166	0.605	3.582	0.000
RMBRD	2.364	0.578	4.091	0.000
PRIVATE	13.023	1.321	9.861	0.000
STUFAC	-0.194	0.102	-1.893	0.058
CSAT	0.066	0.005	13.357	0.000

Output: Lavaan

```
Optimization method                               NLMINB
Number of free parameters                             27

Number of observations                              1302
Number of missing patterns                            14
```

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  gradrat ~
    lenroll           2.166    0.605    3.582    0.000
    rmbrd             2.364    0.578    4.091    0.000
    private          13.023    1.321    9.861    0.000
    stufac           -0.194    0.102   -1.893    0.058
    csat              0.066    0.005   13.357    0.000
```

SEM with Auxiliary Variable

- Including auxiliary variables can potentially reduce biases and standard errors without directly influencing parameter estimates.

- A good auxiliary variable should be highly correlated with variables with missingness in the model.

Adding Auxiliary Variables in Mplus

```
VARIABLE:
    NAMES ARE gradrat lenroll rmbrd private stufac csat act;
    USEVARIABLE ARE gradrat lenroll rmbrd private stufac csat;
    MISSING ARE ALL (9999);
    AUXILIARY = act (M);
```

Adding Auxiliary Variables in Lavaan

- Method 1:

```
model.college2 <- "
gradrat ~ lenroll + rmbrd + private + stufac + csat
act~gradrat+lenroll + rmbrd + private + stufac + csat
"
fit.college2 <- sem(model=model.college2, data=data.college, missing="fiml", fixed.x=FALSE)
summary(fit.college2)
```

- Method 2:
  - Packages ("semTools")

```
model.college <- "
gradrat ~ lenroll + rmbrd + private + stufac + csat
"
fit.college3 <- sem.auxiliary(model=model.college, data=data.college, aux = "act", missing="fiml", fixed.x=FALSE)
summary(fit.college3)
```

Outputs

- Mplus

```
MODEL RESULTS

                                                    Two-Tailed
                    Estimate       S.E.  Est./S.E.    P-Value

 GRADRAT  ON
    LENROLL            2.083      0.598      3.483      0.000
    RMBRD              2.404      0.568      4.235      0.000
    PRIVATE           12.914      1.298      9.952      0.000
    STUFAC            -0.181      0.101     -1.788      0.074
    CSAT               0.067      0.005     13.797      0.000
```

- Lavaan

```
Regressions:
                   Estimate  Std.Err  z-value  P(>|z|)
  gradrat ~
    lenroll           2.083    0.598    3.483    0.000
    rmbrd             2.404    0.568    4.235    0.000
    private          12.914    1.298    9.952    0.000
    stufac           -0.181    0.101   -1.788    0.074
    csat              0.067    0.005   13.797    0.000
```

Mplus vs. Lavaan in FIML

- Lavaan can do FIML for a very wide class of linear models in SEM.
- Mplus can do the same, but it can also handle missing data for
  - Logistic regression.
  - Poisson and negative binomial regression.
  - Models in which data are not missing at random.

Limitations of Maximum Likelihood

- Requires estimation of a model for the joint distribution of all the variables.
  - May be hard to comp up with an appropriate model for the different kinds of variables (e.g., discrete vs. continuous).
  - Results may not be robust to model choice.

Multiple Imputation

- Why multiple imputation?
  - Single imputation not fully efficient because of random variation.
  - Standard errors biased.
- Do it multiple times
  - Generate multiple imputed dataset. Use the same model to fit each imputed dataset.
  - Average the parameter estimates.
  - Variability among the estimates provides information for correcting the standard errors.

Combining the Imputations

- Parameter estimate is just the mean of the multiple estimates.

- Standard error is calculated by the following steps:
  1. Square the estimated standard errors and average them across the replications.
  2. Calculate the variance of the parameter estimates across the replications.
  3. Add the results of 1 and 2 and take the square root.

Formula for Standard Error

$$\sqrt{\frac{1}{M}\sum_{k=1}^{M} s_k^2 + (1+\frac{1}{M})(\frac{1}{M-1})\sum_{k=1}^{M}(b_k - \bar{b})^2}$$

- b_k is the parameter estimate
- s_k is the standard error of b_k
- M is the number of replications

- This formula is used with generally every application of multiple imputation.

MI with Mplus

- Step1: generating imputed dataset

```
DATA IMPUTATION:
    IMPUTE = gradrat lenroll rmbrd private stufac csat;
    NDATASETS = 20;
    SAVE = missimp*.dat;

ANALYSIS: TYPE=BASIC;
          BSEED=2019;
```

- 20 imputed datasets ("missimp1.dat", "missimp2.dat", ..., "missimp20.dat") were created.
- A file named "missimplist.dat" which contains the names of the imputed dataset was also created.

MI with Mplus (cont.)

- Step2: analyzed imputed dataset and aggregate the results

```
DATA:
FILE IS missimplist.dat;
TYPE=IMPUTATION;

VARIABLE:
    NAMES ARE gradrat lenroll rmbrd private stufac csat act;
    USEVARIABLE ARE gradrat lenroll rmbrd private stufac csat;
    MISSING ARE ALL (9999);
MODEL:
    gradrat ON lenroll rmbrd private stufac csat;
```

MI with Lavaan

- Packages: “Amelia”, “mice”

```
model.college2 <- "
gradrat ~ lenroll + rmbrd + private + stufac + csat
act~gradrat+lenroll + rmbrd + private + stufac + csat
"

MIresults2 <- runMI(model=model.college2, data=data.college, m=20, miPackage="mice", fun="sem", seed=2019)
summary(MIresults2)
```

Outputs

Why the results are different?

- Mplus

MODEL RESULTS

	Estimate	S.E.	Est./S.E.	Two-Tailed P-Value	Rate of Missing
GRADRAT ON					
LENROLL	2.228	0.579	3.850	0.000	0.147
RMBRD	2.404	0.555	4.336	0.000	0.481
PRIVATE	13.192	1.253	10.528	0.000	0.170
STUFAC	-0.202	0.107	-1.894	0.058	0.397
CSAT	0.065	0.005	12.376	0.000	0.491

- Lavaan

Regressions:

	Estimate	Std.Err	z-value	P(>|z|)
gradrat ~				
lenroll	2.083	0.598	3.483	0.000
rmbrd	2.404	0.568	4.235	0.000
private	12.914	1.298	9.952	0.000
stufac	-0.181	0.101	-1.788	0.074
csat	0.067	0.005	13.797	0.000

Multiple Imputation

- Pros:
  - Properties similar to ML.
  - Can be used with any kind of data or model.
  - Analysis can be done with conventional software.

- Cons:
  - Get a different result every time you use it.
    - Increase imputation time
    - Select seed (does not guarantee results are effecient)
  - The imputation model must be consistent with the analysis model (not a problem for ML).

Side by Side Comparisons

- E.g., gradart ON(~) lenroll

	Coefficient	Standard Error
Listwise	2.417	0.953
FIML	2.166	0.605
FIML with Auxiliary variable	2.083	0.598
MI (Mplus)	2.228	0.579
MI(Lavaan)	2.051	0.830

Class Practice (NLSYMISS)

N=581 Children, Variables are:

DV:

- ANTI antisocial behavior, measured with a scale ranging from 0 to 6.

IV:

- SELF self-esteem, measured with a scale ranging from 6 to 24.
- POV poverty status of family, coded 1 for in poverty, otherwise 0.
- BLACK 1 if child is black, otherwise 0
- HISPANIC 1 if child is Hispanic, otherwise 0
- CHILDAGE child's age in 1990
- DIVORCE 1 if mother was divorced in 1990, otherwise 0
- GENDER 1 if female, 0 if male
- MOMAGE mother's age at birth of child
- MOMWORK 1 if mother was employed in 1990, otherwise 0

Goal: Run the Regression with…

- Listwise deletion on the predictor
- FIML
- Multiple Imputation with 20 imputed dataset