Evaluation of Supplemental Samples in Longitudinal Research with Nonnormal Missing Data

Jessica A. M. Mazen[1] Xin Tong[1] Laura K. Taylor[2]

Department of Psychology
University of Virginia[1]

School of Psychology
Queen's University Belfast[2]

Modern Modeling Methods conference
May 2017

Longitudinal Research

- Popularity of longitudinal research is growing
- More attention paid to longitudinal theory, methodology, and research

Longitudinal Research

- Used in all areas of psychology to study a diverse set of topics (e.g., childhood abuse, mental illness, political violence)
- Popularity is not surprising, but longitudinal research is often encumbered with methodological challenges
- One such challenge is that missing data frequently arise

Missing Data

- Attrition rate - the percentage of participants from the initial wave that are missing at one or more time points
  - Permanent - a participant drops out of the study and does not return
  - Intermittent - a participant may not be available for one or more measurement occasions, but then returns at later waves of data collection

Missing Data

- Missing data mechanisms refer to the process that causes missing data (Little and Rubin, 2002)
  - Missing completely at random (MCAR) - missingness on Y is completely independent of other variables that influence Y
    - e.g., a student happens to be sick on the day of the math test
  - Missing at random (MAR) - missingness on Y is related to an observed variable (auxiliary variable) that affects Y
    - e.g., students with greater test anxiety tend to skip the test more than less anxious students, test anxiety is measured
  - Missing not at random (MNAR) - missingness on Y is related to an unobserved variable that influences Y
    - e.g., students with greater anxiety skip the test more, test anxiety is not measured

Missing Data

- 44% average attrition rate across 92 longitudinal studies in a recent meta-analysis examining personality traits (Roberts et al., 2006)
- 5% to 50% attrition rate across 25 population-based longitudinal studies of the elderly (Chatfield et al., 2005)
- Attrition may be especially problematic in longitudinal studies with at-risk populations
    - Attrition rates can be as high as 85% (Goemans, van Geel, and Vedder, 2015)

Current Strategies

- Deletion of cases (e.g., listwise or pairwise deletion)
  - Common approach to dealing with missing data (Jeličić et al., 2009)
- Modern missing data approaches
  - Full information maximum likelihood (FIML) estimation
  - Multiple imputation (MI)

Current Strategies

- Retention and tracking techniques (Ribisl et al., 1996)
  - e.g., increased financial incentive over time, driver's records, obtaining contact information of friends or family of participants
- Planned missing designs
  - Researchers intentionally collect incomplete data from participants
    - Missing items
    - Missing measures
    - Missing measurement occasions

Supplemental Sample Definition

- A set of new participants added to the original sample (after missing data appear) in the second or later measurement occasion

Supplemental Sample Approaches

- A set of new participants added to the original sample (after missing data appear) in the second or later measurement occasion
- Two approaches
  - Refreshment approach - researchers select additional participants using the same criteria as the initial participants (i.e., random selection from population of interest)
    - e.g., randomly select grade school children

Supplemental Samples Approaches

- A set of new participants added to the original sample (after missing data appear) in the second or later measurement occasion
- Two approaches
  - Refreshment approach - researchers select additional participants using the same criteria as the initial participants (i.e., random selection from population of interest)
    - e.g., randomly select grade school children
  - Replacement approach - researchers first identify auxiliary variables that explain the pattern of missingness in the data and then select new participants based on those attributes
    - e.g., researchers may over-select for children with high test anxiety

Supplemental Samples Use

- Supplemental samples are utilized to address attrition in many studies
  - Includes numerous large-scale studies
    - International Tobacco Control Policy Evaluation Project
    - Medicare Current Beneficiary Survey
    - International Alcohol Control (IAC) Study
    - Survey of Health, Ageing and Retirement in Europe
    - English Longitudinal Study of Ageing
  - Projects have generated over 2600 published articles
- Little research investigating supplemental samples -> little guidance for researchers

Previous Research

- Taylor, Tong, and Maxwell (under review) systematically studied the effects of adding supplemental samples in growth curve modeling
- Compared refreshment and replacement approaches with MCAR and MAR data
- MCAR and MAR with refreshment approach
  - Bias similar to complete data analysis
  - Acceptable coverage rates
- MAR with replacement approach
  - Greater bias, increased as replacement sample increased
  - Unacceptable coverage rates

Previous Research

- Limitations:
  - Only focused on normally distributed data
  - Supplemental samples added at only one measurement occasion
  - Permanent attrition only
- Limit the applicability of findings to real-world studies

Current Study

- Extend previous findings by assessing effects of supplemental samples across a wide variety conditions
  - Nonnormal distributions
    - Practical data are more likely to be nonnormal in social and behavioral sciences (Micceri, 1989)
  - Permanent and intermittent attrition
  - Multiple measurement occasions

Model

- Growth curve model with time-invariant covariate
- A typical form of a linear growth curve models can be expressed as

$$y_i = \Lambda b_i + e_i,$$
$$b_i = \beta_0 + \beta_1 x_i + u_i,$$

y_i = Observations for individual i
Λ = Factor loading matrix determining the growth trajectories
b_i = Random effects
e_i = Intraindividual measurement errors
x_i = Covariate $\sim MVN(10, 1.5)$
β_0 = Regression coefficients = $(6, 0.3)$
β_1 = Regression coefficients = $(1, 0.1)$
u_i = Residuals $\sim MVN(0, 1)$

Conditions

- Original sample size
  - N = 50, 200, 500, 1000
- Number of measurement occasions
  - T = 4, 8
- Distribution of intraindividual measurement errors
  - normal distribution $N(0, \sigma_e^2)$
  - normal distribution $N(0, \sigma_e^2)$ with 2% outliers
  - normal distribution $N(0, \sigma_e^2)$ with 8% outliers
  - gamma distribution $\Gamma_{(1,1)}(0, \sigma_e^2)$
  - log-normal distribution $LN_{(0,1)}(0, \sigma_e^2)$
  - t distribution $t_{(5)}(0, \sigma_e^2)$

Conditions

- Variance of measurement errors
  - $\sigma_e^2 = 1, 3$
- Missing data pattern
  - MCAR, MAR
- Correlation between the auxiliary variable and latent slope
  - $r = .3, .8$
- Missing rate
  - MR = 3%, 5%, 8%, and 15%

Conditions

- Supplemental sample type
  - refreshment (RF) samples
  - replacement (RP) samples
- Size/timing of supplemental samples
  - 1 x number of missing observations at 2nd measurement occasion added at the 3rd measurement occasion (RF/RP (1))
  - (T-2) x number of missing observations at 2nd measurement occasion added at the 3rd measurement occasion (RF/RP(T-2))
  - 1 x number of missing observations at 2nd measurement occasion added at the 3rd measurement occasion and every subsequent measurement occasion (RF/RP(M))
- 7,152 conditions

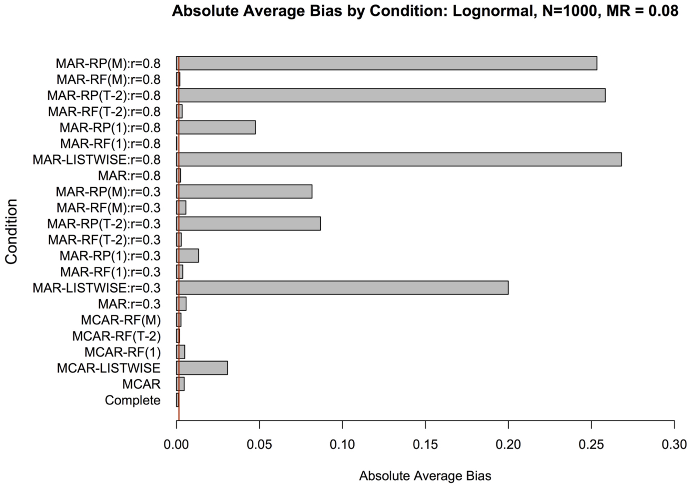
Analyses

- Two-stage robust procedure for structural equation modeling with missing data (Yuan and Zhang, 2012)
  - R package 'rsem'
  - Robust procedures are advantageous when analyzing data with missing values
    - Difficult to determine the distributional properties of the sample when missing values are present
    - Produce less biased parameter estimates and more reliable test statistics
- For comparison, we also applied listwise deletion and two-stage NML to analyze the data

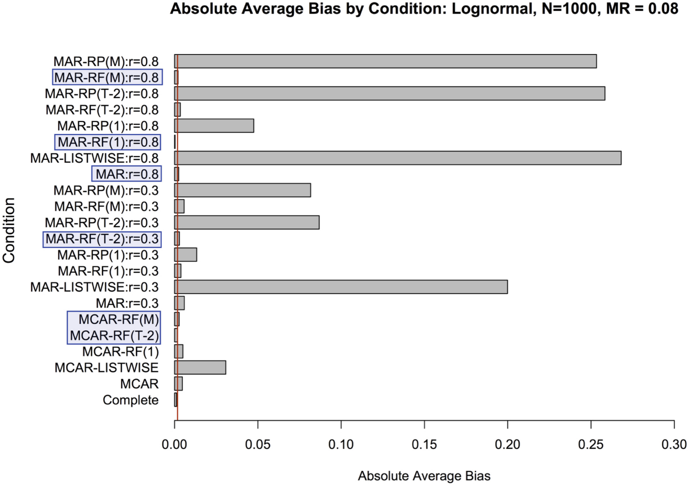
Evaluation Criterion

- Estimate of interest: population mean slope parameter
- Outcomes evaluated:
  - Absolute average bias - absolute value of bias (estimation minus the true parameter value) averaged across all replications
  - Relative efficiency - ratio of squared empirical standard error of complete data to incomplete data
  - Power - proportion of replications of which the 95% confidence interval does not contain zero
  - Average confidence interval width - upper confidence interval (CI) boundary minus lower CI boundary averaged across all replications

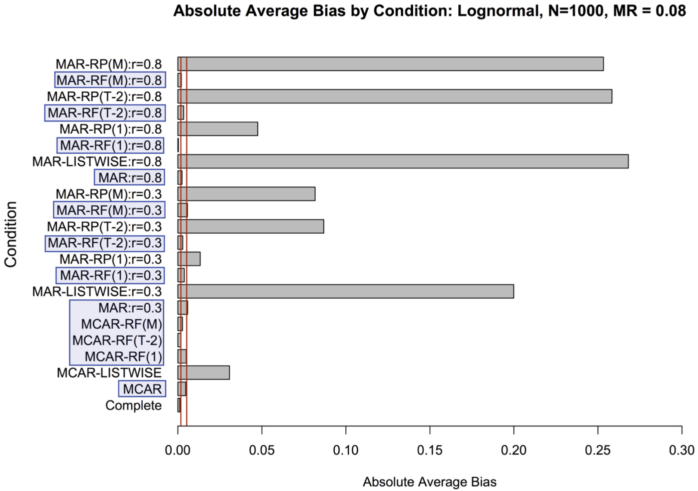
Bias


Absolute Average Bias by Condition: Lognormal, N=1000, MR = 0.08
Condition
MAR-RP(M):r=0.8
MAR-RF(M):r=0.8
MAR-RP(T-2):r=0.8
MAR-RF(T-2):r=0.8
MAR-RP(1):r=0.8
MAR-RF(1):r=0.8
MAR-LISTWISE:r=0.8
MAR:r=0.8
MAR-RP(M):r=0.3
MAR-RF(M):r=0.3
MAR-RP(T-2):r=0.3
MAR-RF(T-2):r=0.3
MAR-RP(1):r=0.3
MAR-RF(1):r=0.3
MAR-LISTWISE:r=0.3
MAR:r=0.3
MCAR-RF(M)
MCAR-RF(T-2)
MCAR-RF(1)
MCAR-LISTWISE
MCAR
Complete
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Absolute Average Bias

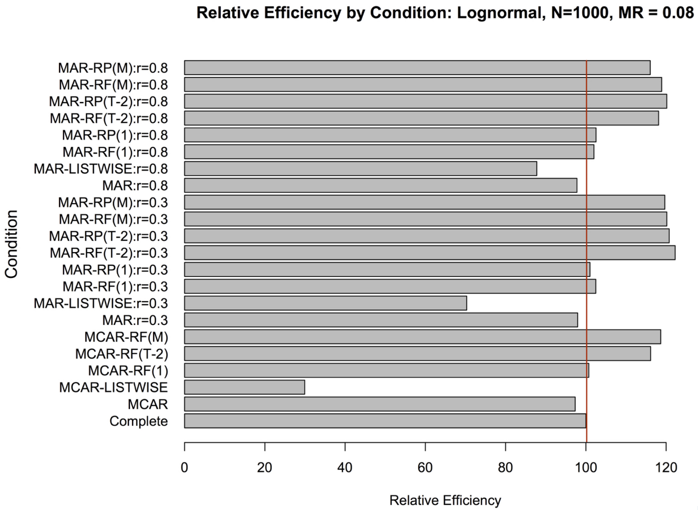
Bias


Absolute Average Bias by Condition: Lognormal, N=1000, MR = 0.08
Condition
MAR-RP(M):r=0.8
MAR-RF(M):r=0.8
MAR-RP(T-2):r=0.8
MAR-RF(T-2):r=0.8
MAR-RP(1):r=0.8
MAR-RF(1):r=0.8
MAR-LISTWISE:r=0.8
MAR:r=0.8
MAR-RP(M):r=0.3
MAR-RF(M):r=0.3
MAR-RP(T-2):r=0.3
MAR-RF(T-2):r=0.3
MAR-RP(1):r=0.3
MAR-RF(1):r=0.3
MAR-LISTWISE:r=0.3
MAR:r=0.3
MCAR-RF(M)
MCAR-RF(T-2)
MCAR-RF(1)
MCAR-LISTWISE
MCAR
Complete
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Absolute Average Bias

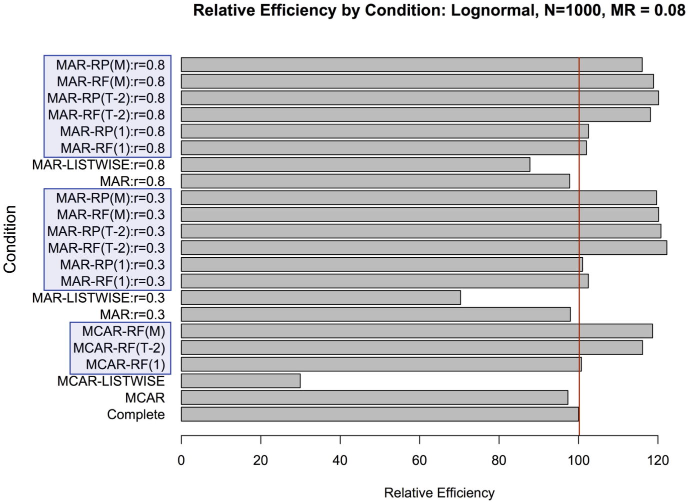
Bias


Absolute Average Bias by Condition: Lognormal, N=1000, MR = 0.08
Condition
MAR-RP(M):r=0.8
MAR-RF(M):r=0.8
MAR-RP(T-2):r=0.8
MAR-RF(T-2):r=0.8
MAR-RP(1):r=0.8
MAR-RF(1):r=0.8
MAR-LISTWISE:r=0.8
MAR:r=0.8
MAR-RP(M):r=0.3
MAR-RF(M):r=0.3
MAR-RP(T-2):r=0.3
MAR-RF(T-2):r=0.3
MAR-RP(1):r=0.3
MAR-RF(1):r=0.3
MAR-LISTWISE:r=0.3
MAR:r=0.3
MCAR-RF(M)
MCAR-RF(T-2)
MCAR-RF(1)
MCAR-LISTWISE
MCAR
Complete
0.00
0.05
0.10
0.15
0.20
0.25
0.30
Absolute Average Bias

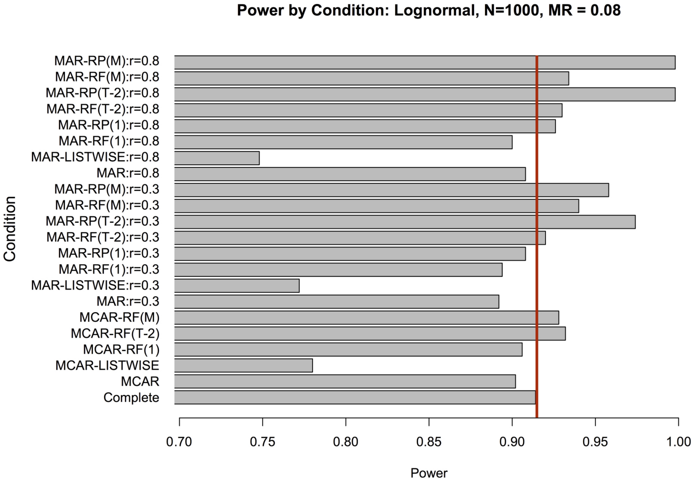
Relative Efficiency



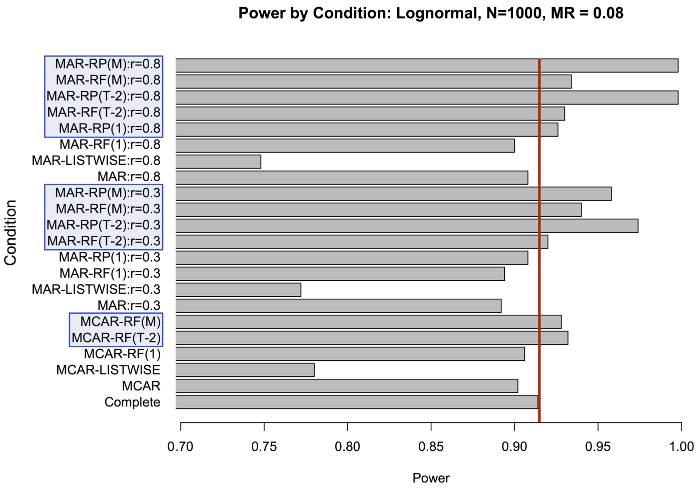
Relative Efficiency



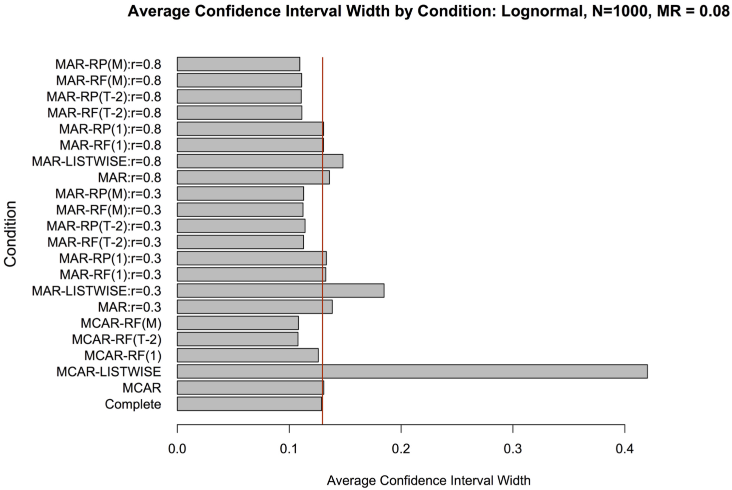
Power


Power by Condition: Lognormal, N=1000, MR = 0.08
MAR-RP(M):r=0.8
MAR-RF(M):r=0.8
MAR-RP(T-2):r=0.8
MAR-RF(T-2):r=0.8
MAR-RP(1):r=0.8
MAR-RF(1):r=0.8
MAR-LISTWISE:r=0.8
MAR:r=0.8
MAR-RP(M):r=0.3
MAR-RF(M):r=0.3
MAR-RP(T-2):r=0.3
MAR-RF(T-2):r=0.3
MAR-RP(1):r=0.3
MAR-RF(1):r=0.3
MAR-LISTWISE:r=0.3
MAR:r=0.3
MCAR-RF(M)
MCAR-RF(T-2)
MCAR-RF(1)
MCAR-LISTWISE
MCAR
Complete
Condition
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Power

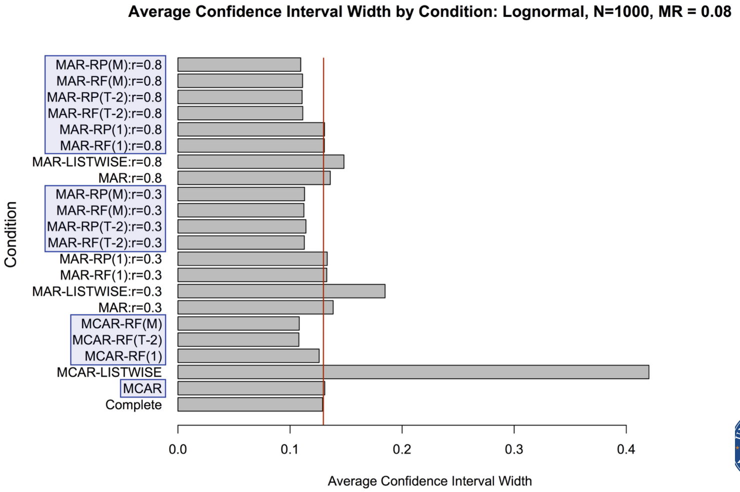
Power


Power by Condition: Lognormal, N=1000, MR = 0.08
Condition
MAR-RP(M):r=0.8
MAR-RF(M):r=0.8
MAR-RP(T-2):r=0.8
MAR-RF(T-2):r=0.8
MAR-RP(1):r=0.8
MAR-RF(1):r=0.8
MAR-LISTWISE:r=0.8
MAR:r=0.8
MAR-RP(M):r=0.3
MAR-RF(M):r=0.3
MAR-RP(T-2):r=0.3
MAR-RF(T-2):r=0.3
MAR-RP(1):r=0.3
MAR-RF(1):r=0.3
MAR-LISTWISE:r=0.3
MAR:r=0.3
MCAR-RF(M)
MCAR-RF(T-2)
MCAR-RF(1)
MCAR-LISTWISE
MCAR
Complete
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Power

Confidence Interval Width



Confidence Interval Width



General Discussion

- Bias
  - RF samples/ML estimation resulted in bias similar to complete data
  - RP samples led to biased estimates
    - Larger RP sample / higher missing rates equates to greater bias
- Relative Efficiency
  - Efficiency was greatest when supplemental samples were used
    - Increasing the size of supplemental sample resulted in higher efficiency
    - Differences between methods increased as missing rate increased

General Discussion

- Power
  - RP samples resulted in greater power than RF samples, and supplemental samples produced greater power than the ML estimation
- Average Confidence Interval Width
  - Interval widths similar to complete data for all supplemental sample/ ML methods
    - Increasing supplemental sample decreased interval width

Recommendations

- Replacement samples produce biased estimates and should not be used
- Refreshment samples can improve power and efficiency
  - Decision to use refreshment samples depends on many factors
    - Expected effect size, missing rate, cost/difficulty of obtaining supplemental sample

Thank You!

jm5ku@virginia.edu