From discrete conditions to continuous factors: Rethinking methodological simulations

Alexander M. Schoemann
Todd D. Little
Mijke Rhemtulla
Sunthud Pornprasertmanit

CENTER FOR RESEARCH METHODS & DATA ANALYSIS
College of Liberal Arts & Sciences
CRMDA.KU.EDU

Monte Carlo Simulations

- Monte Carlo simulations are a popular tool for methodologists with many uses
  - Determine the accuracy of new methods
  - Compare different methods
  - Perform power analyses

Monte Carlo Simulations

- General steps in a Monte Carlo Simulation
  1. Specify population parameters
  2. Create a sample of size N, based on population parameters
  3. Analyze sample data from step 2 with chosen statistical method(s).
  4. Repeat steps 2 and 3 for each of r replications.

The Typical Simulation Design

- Most simulations done involve a fixed set of conditions and a fully factorial design.
  - This can result in an extremely large number of simulation conditions.
  - "Crossing conditions defined by ICC, J , and n_j resulted in 4 X 6 X 3 = 72 conditions" Preacher, Zhang, & Zyphur (2011, p. 168)
  - Rhemtulla, Schoemann & Preacher (2011):

    9 X 9 X 6 X 4 = 1944 conditions

- Results from such a design are often interpreted via "eyeball"

The Typical Simulation Design

- Traditional designs require a trade off between study size and external validity.
  - More conditions = more external validity
  - More conditions = (much) larger design and more replications, greater difficulty interpreting results

The Typical Simulation Design

- Skrondal (2000) provided four recommendations to alleviate problems associated with simulation design
  - **Use of a meta-model**
  - Use of incomplete factorial designs
  - Use of common random numbers
  - **Use of fewer replications per condition**

Continuously Varying Factors

- Most factors in simulations are not categorical
  - e.g. sample size, parameter values
- Most simulation studies treat continuous factors as categorical.
  - This can bias results or hide important relationships
- What if factors in simulations were varied continuously?

Continuously Varying Factors

- With continuously varying factors, simulation parameters of interest (e.g., sample size, parameter values) are allowed to vary across a range of values.

Continuously Varying Factors

- Each replication is based on a population that is specified by a random draw from the range of population values.
  - A single (sample) dataset is generated and analyzed based on these parameters

- Results from the simulation are analyzed using a regression meta-model.

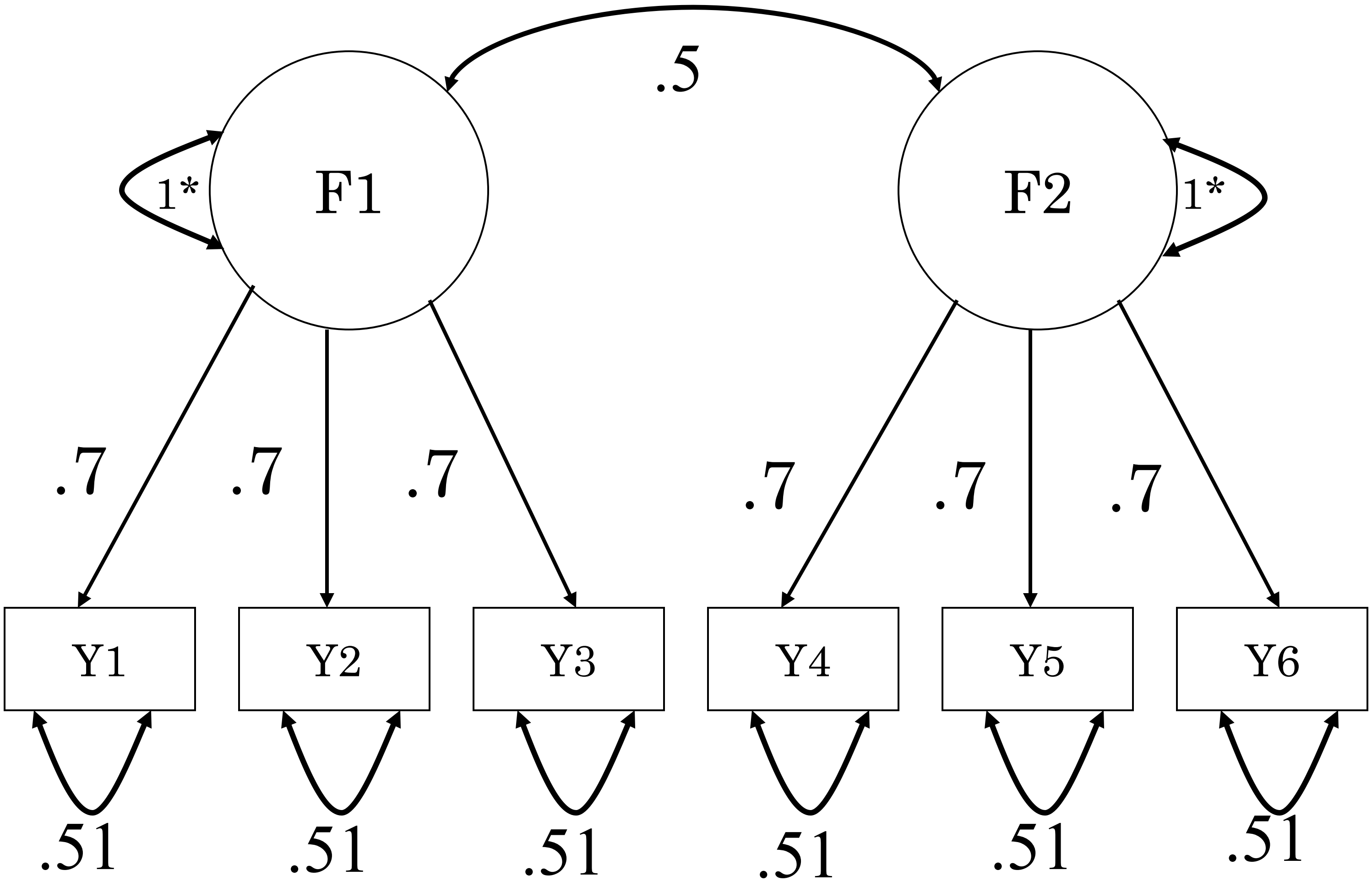
Example 1: Methodological Investigation

- A researcher is interested in studying the performance of full information maximum likelihood with missing data.
  - Traditional approach:
    - Select fixed values of the percent of missing data (e.g., 5%, 40%, 80%)
    - Generate 2000 replications in each condition
    - Analyze results using ANOVA/Present results in a large table

Example 1: Methodological Investigation

- A researcher is interested in studying the performance of full information maximum likelihood with missing data.
  - Continuous approach:
    - Specify a range of percent missing data (e.g., 1%-90%)
    - Generate 2000 replications with randomly varying percent missing data across replications
    - Analyze results using regression/Present results in figures

Example 1: Methodological Investigation



Example 1: Methodological Investigation

- Data were generated and analyzed with the simsem package (Pornprasertmanit, Miller, & Schoemann, 2012) in R.
  - R based SEM simulation utility (available on CRAN)
  - Advanced missing data simulation techniques
  - Built in functions to continuously vary simulation parameters

Example 1: Methodological Investigation

- Traditional approach results
  - Parameter bias

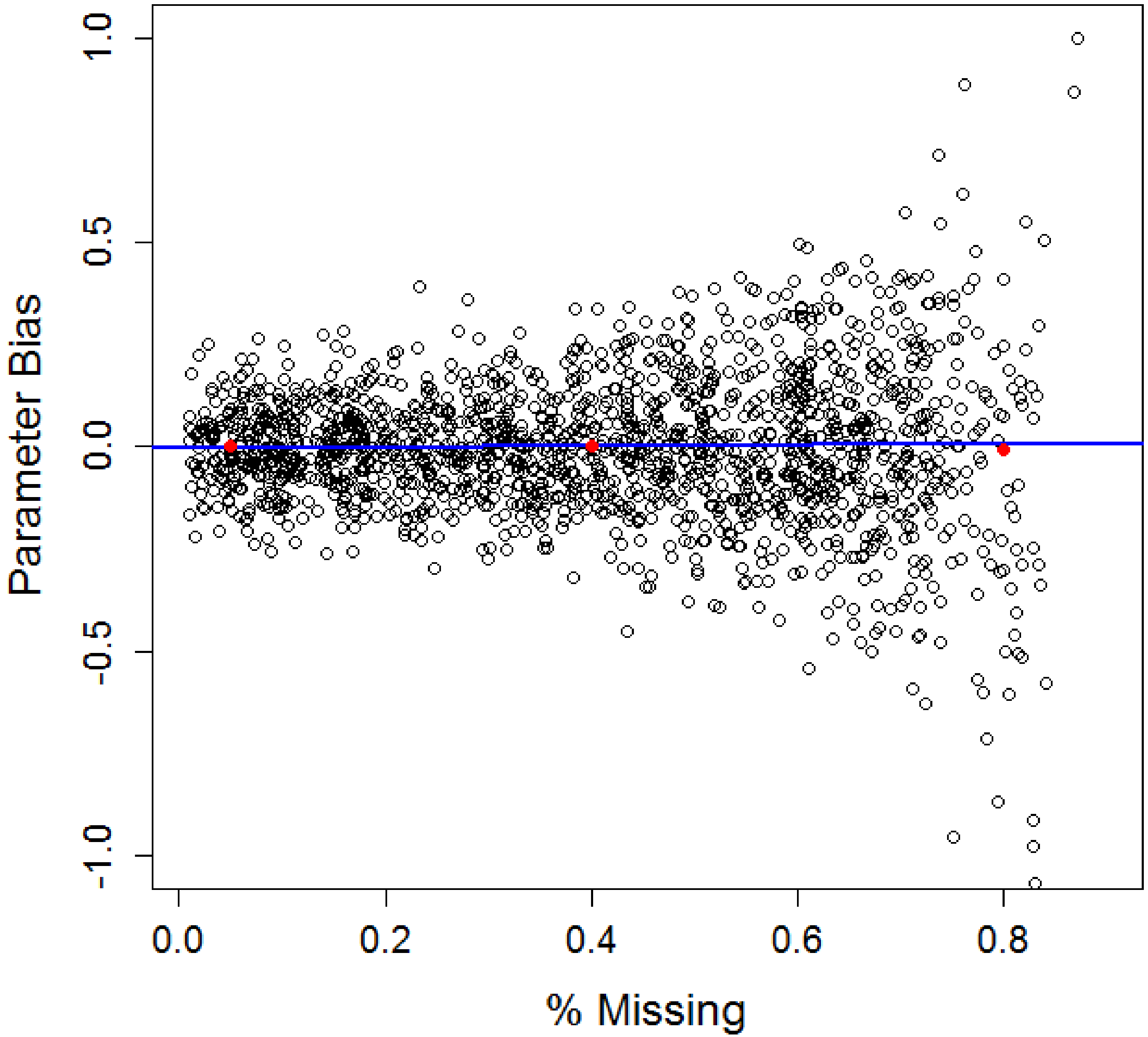
%Missing	Bias (PS 1,2)
.05	-.00004
.40	.00021
.80	-.00882
$R^2 = 0.0009$	

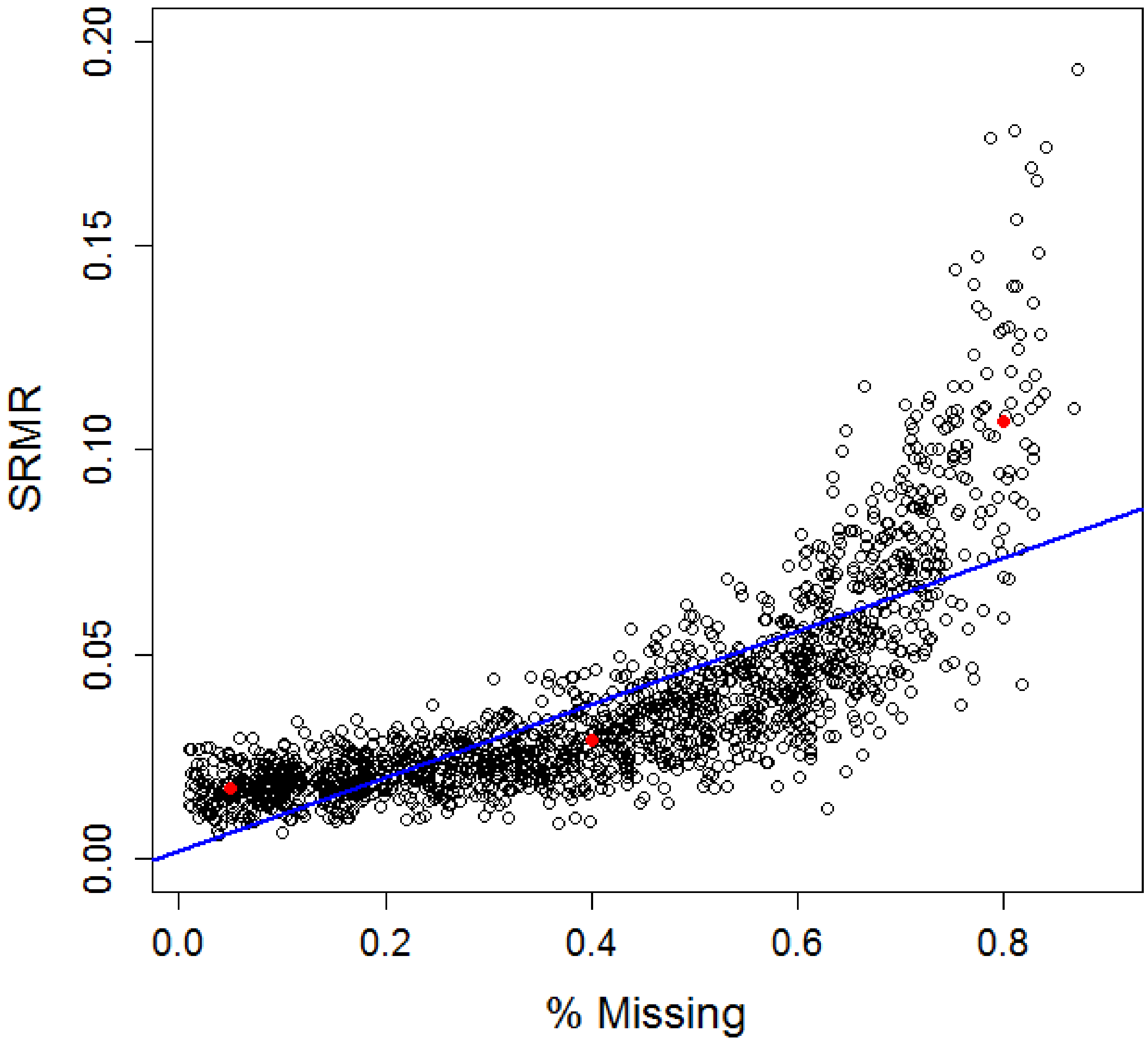
  - Model Fit

%Missing	$\chi2$	RMSEA	CFI	SRMR
.05	8.13	.012	.998	.017
.40	8.23	.013	.994	.029
.80	8.16	.014	.956	.107
R^2	0.00008	**0.002**	**0.19**	**0.86**

Example 1: Methodological Investigation

- Continuous approach results
  - Parameter bias
    - Bias (PS 1,2) = -0.0042 + 0.0151(%missing), R^2 = .00004
  - Model Fit
    - $\chi 2$ = 8.0184 + **0.9365**(%missing), R^2 = .002
    - RMSEA = 0.0115 + **0.0053** (%missing), R^2 = .005
    - CFI = 1.005 + **-0.0387** (%missing), R^2 = .120
    - SRMR = 0.001746 + **0.0899** (%missing), R^2 = .610


Parameter Bias
1.0
0.5
0.0
-0.5
-1.0
0.0
0.2
0.4
0.6
0.8
% Missing



SRMR
0.00
0.05
0.10
0.15
0.20
0.0
0.2
0.4
0.6
0.8
% Missing

Example 2: Power Analysis

- Given population parameters, what sample size will results in a given level of power (e.g., .80)?
  - Traditional approach
    - Specify model and one sample size
    - Generate 2000 replications at this sample size
    - Record power for parameters of interest (proportion of replications with significant parameters)
    - If power ≠ .80, choose different sample size and try again.

Example 2: Power Analysis

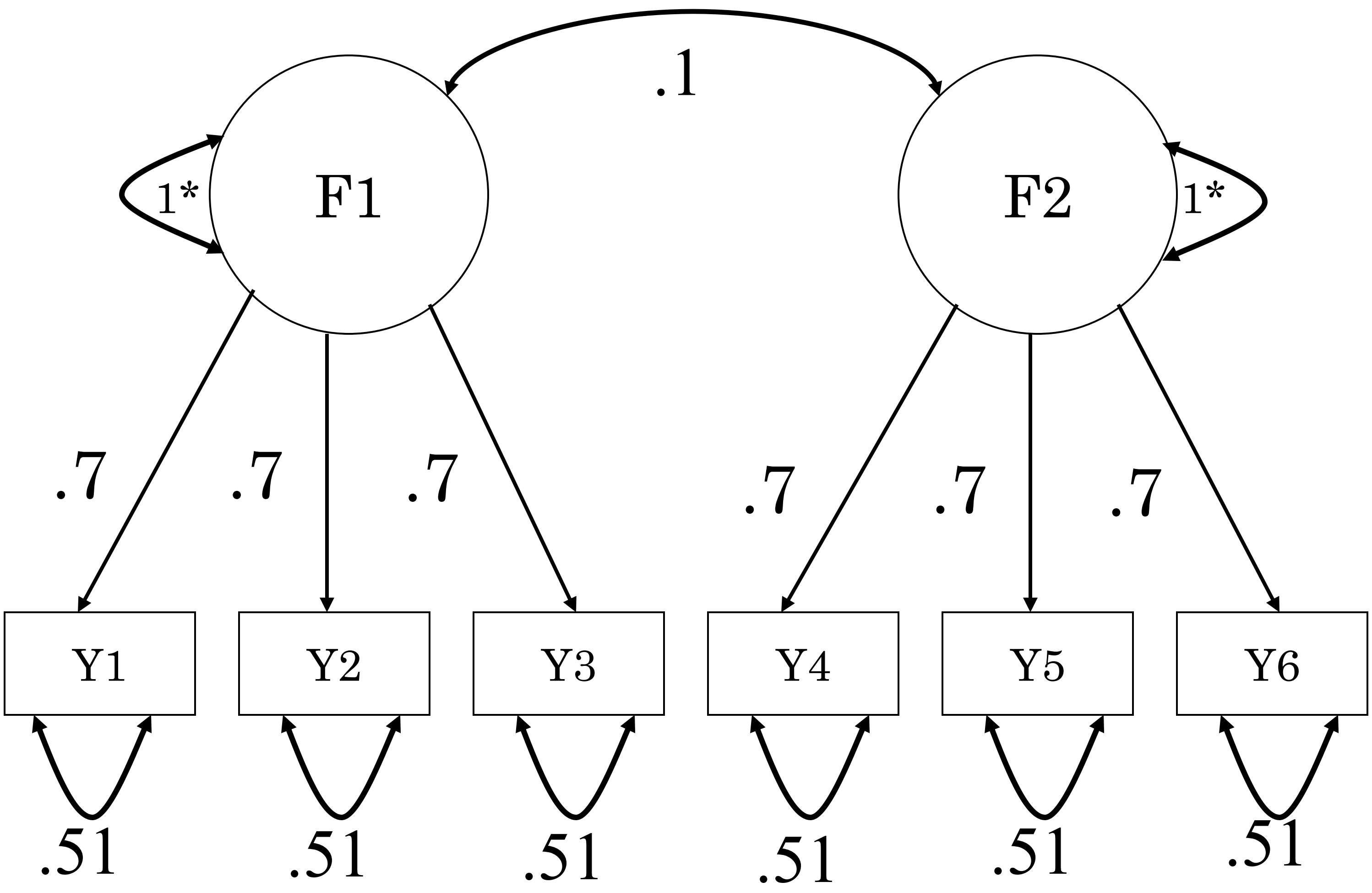
- Given population parameters, what sample size will result in a given level of power (e.g., .80)?
  - Continuous approach
    - Specify model and a range of sample sizes
    - Generate 2000+ replications varying sample size across replications
    - Record each parameter's significance for each replication (0 not sig., 1 sig.)

EXAMPLE 2: POWER ANALYSIS

- Given population parameters, what sample size will results in a given level of power (e.g., .80)?
  - Continuous approach
    - Use logistic regression to predict a parameter's significance (across all replications) from the sample size of each replication.
    - The predicted probability from the logistic regression at a given N is power for that parameter at that N

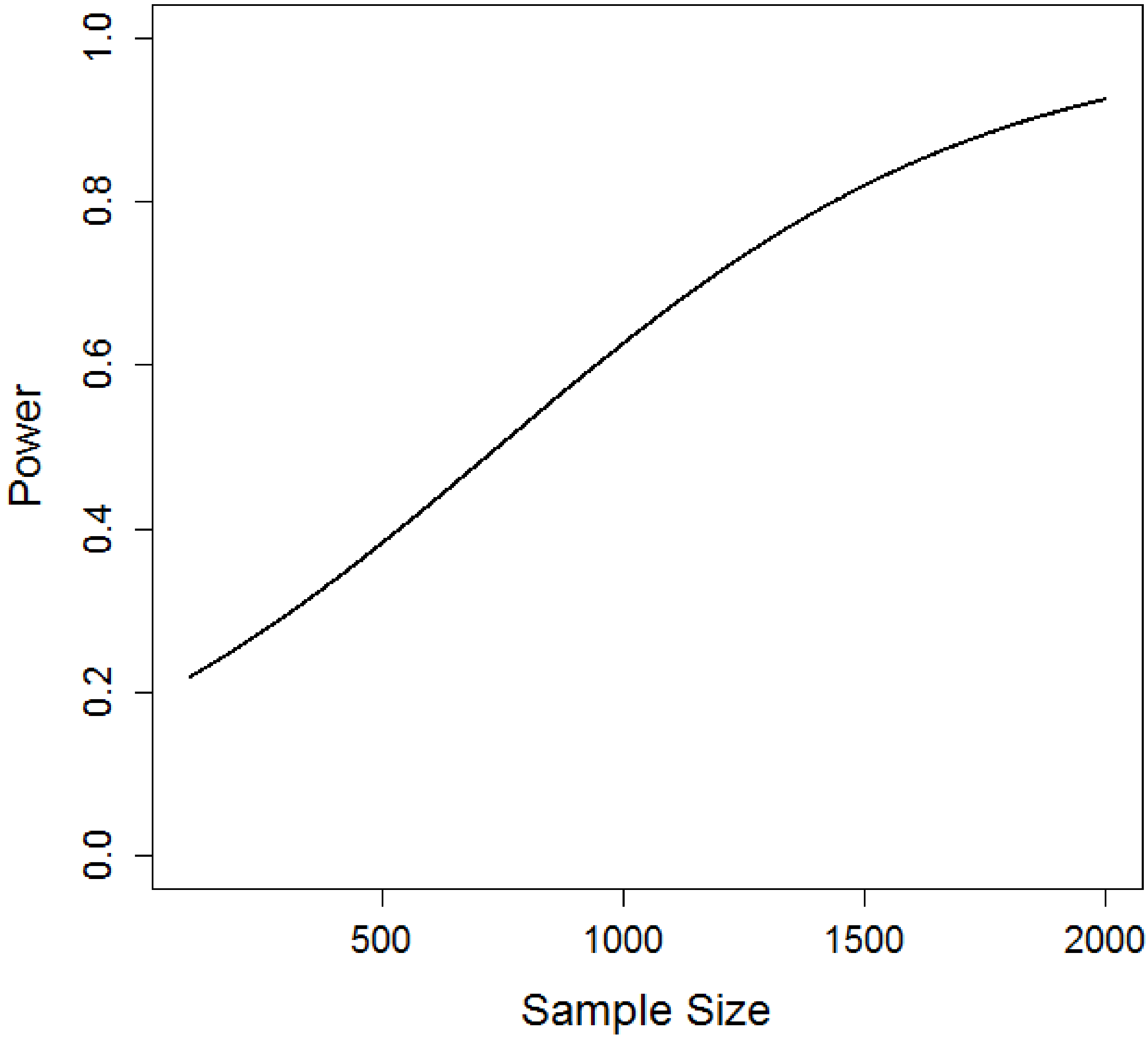
$$p = \frac{e^{B_0 + B_1 N}}{1 + e^{B_0 + B_1 N}}$$

Example 2: Power Analysis



Example 2: Power Analysis

- Results: What sample size results in power for the latent correlation of .80?
  - Continuous approach
    - 3000 replications, randomly varying N between 100-2000
    - logit(power) = $\beta_0 + \beta_1 N$
    - Power = .80 when N = 1436
  - Traditional approach: 3000 replications at n = 1436
    - Power = .810


1.0
0.8
0.6
0.4
0.2
0.0
Power
500
1000
1500
2000
Sample Size

Advantages of Continuously Varying Factors

- Graphical representation of results
  - Investigation of non-linear relationships
- More efficient use of resources
  - Continuously varying parameters allow for fewer replications over a greater range of conditions.
- Greater external validity
- Power analyses are easily specified.
  - Can vary multiple factors over replications (e.g., sample size and effect size)
  - Can easily determine minimum detectable effect size

Limitations

- Estimating empirical standard errors
  - Variability of parameter estimates across replications
  - Difficult to calculate when variability changes as a function of simulation parameters.
  - Possible solution: kernel ridge regression
- Software implementation
  - Currently only automated in simsem

QUESTIONS?

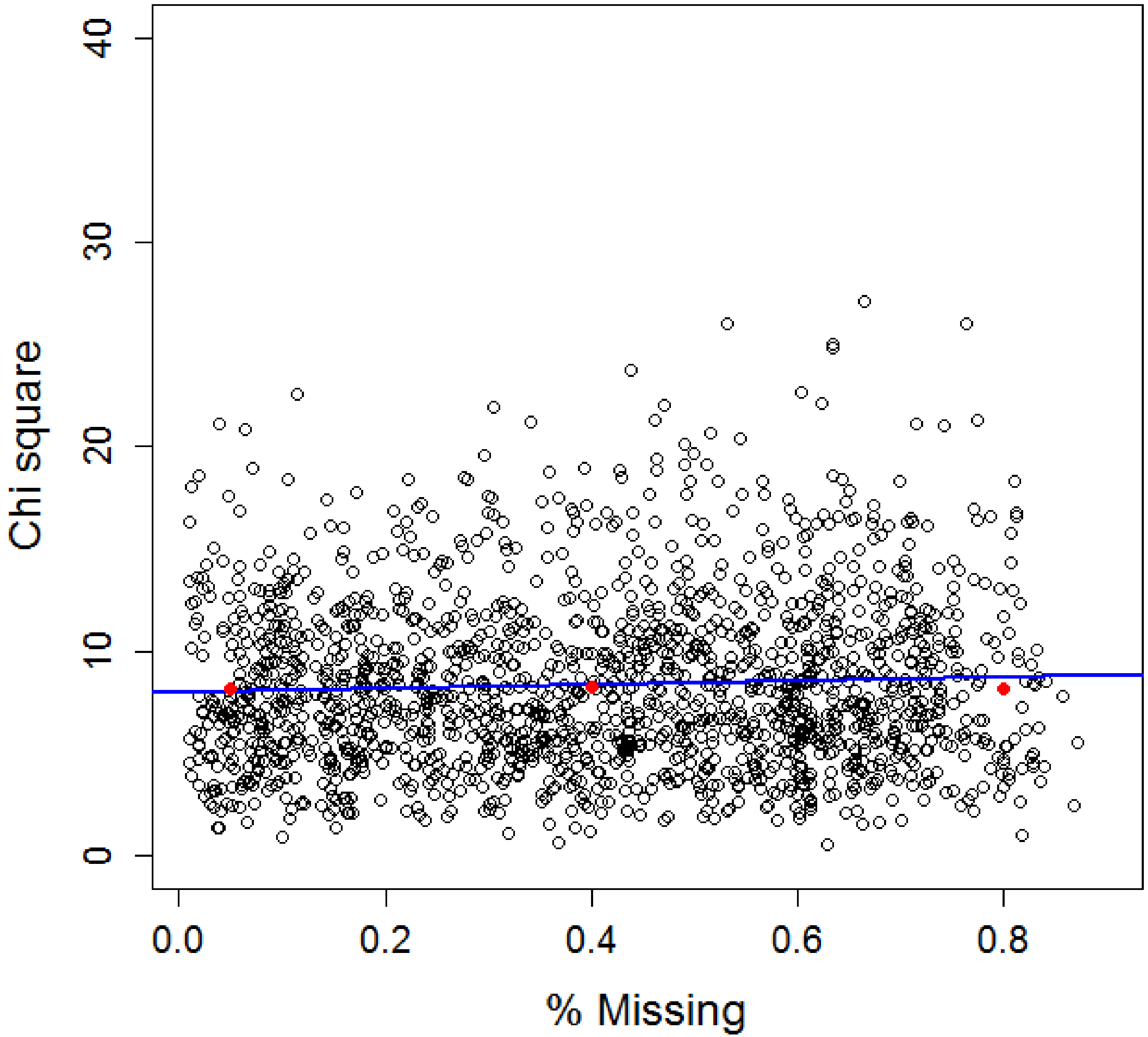
- Thanks to
  - Paul Johnson
  - Patrick Miller

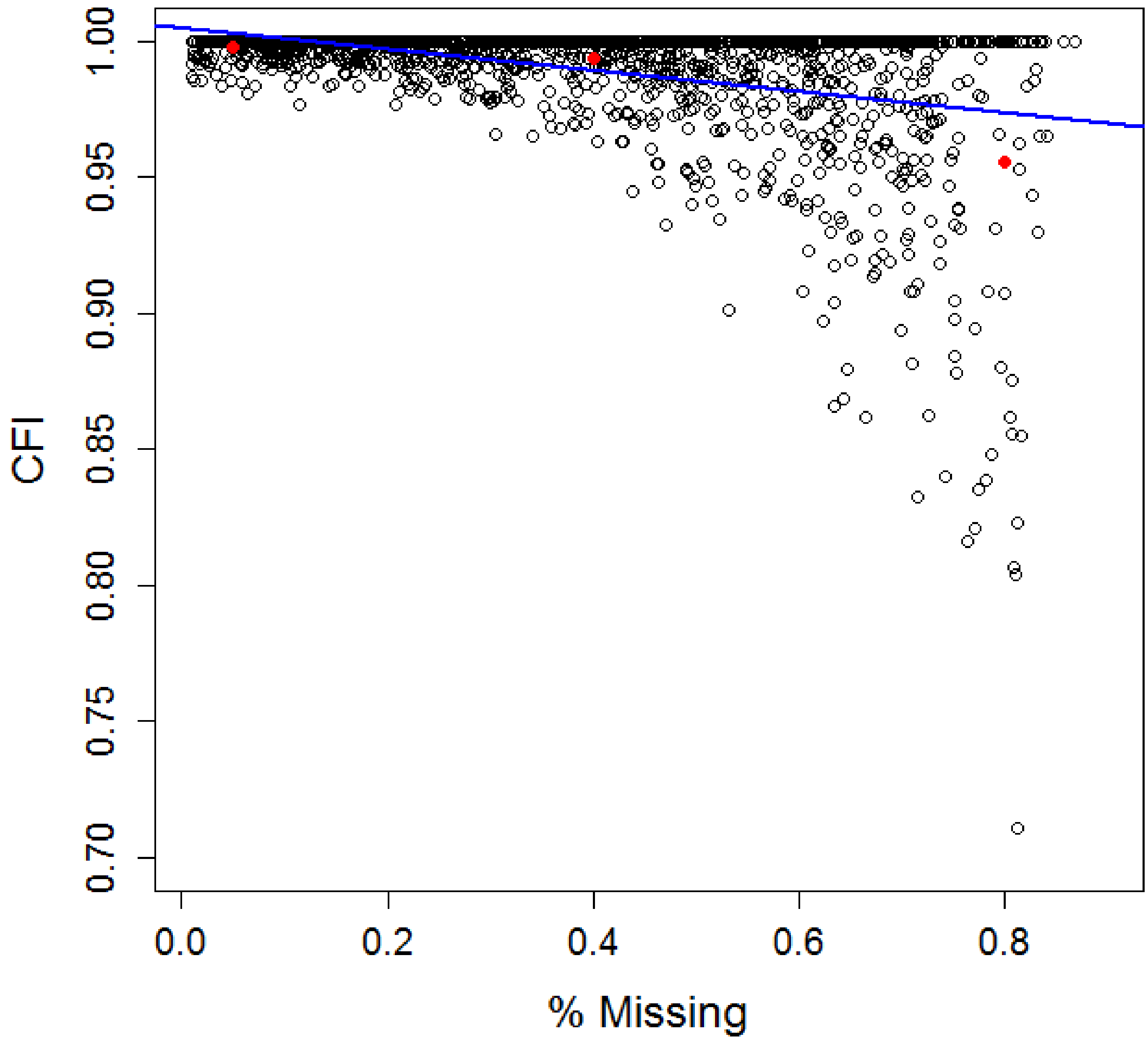
KU CENTER FOR RESEARCH METHODS & DATA ANALYSIS
College of Liberal Arts & Sciences
CRMDA.KU.EDU

simsem: http://github.com/simsem/simsem/wiki

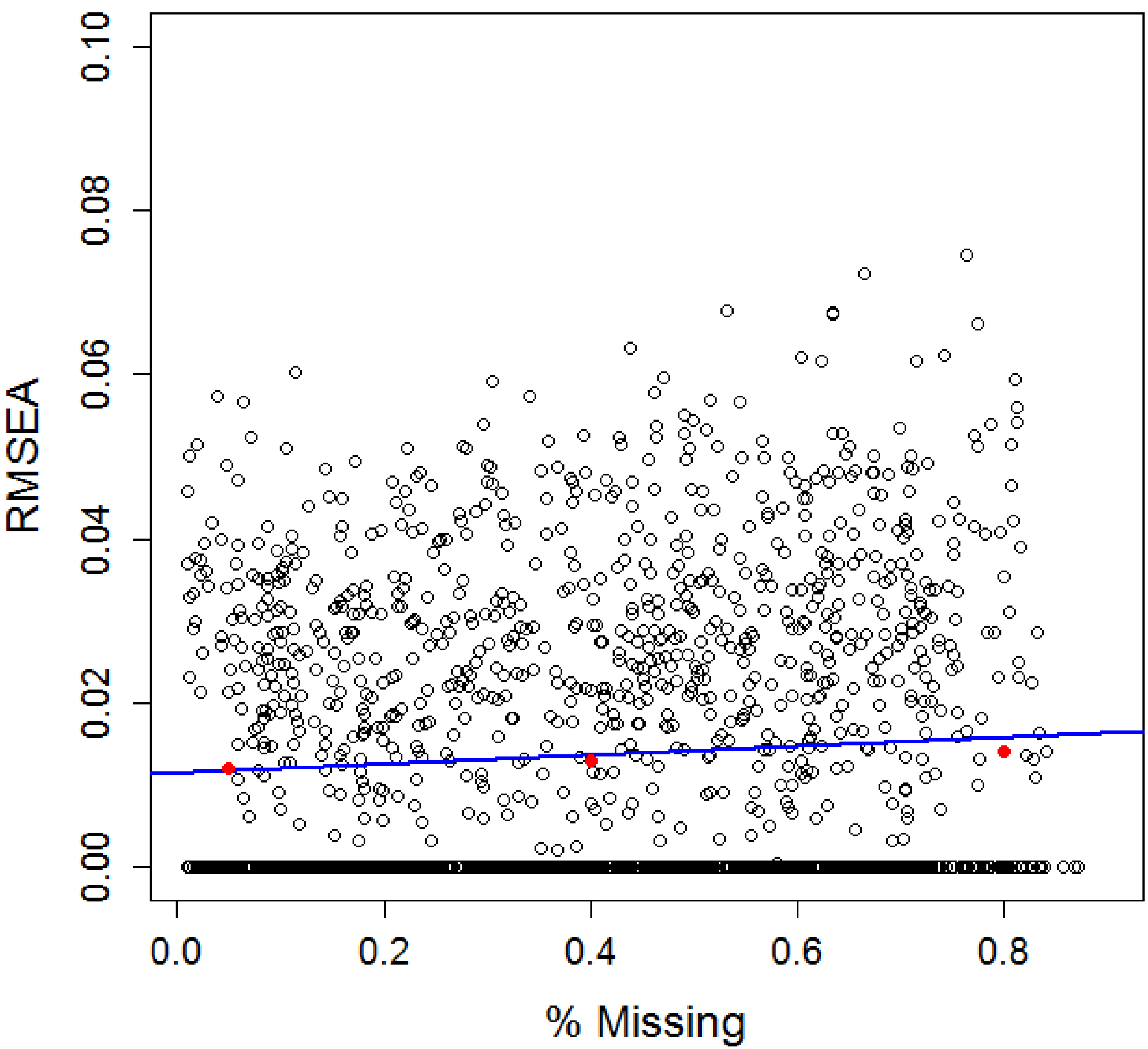
example code: http://github.com/simsem/simsem/wiki

email: schoemann@ku.edu


Chi square
0
10
20
30
40
0.0
0.2
0.4
0.6
0.8
% Missing



CFI
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.0
0.2
0.4
0.6
0.8
% Missing


RMSEA
0.00
0.02
0.04
0.06
0.08
0.10
0.0
0.2
0.4
0.6
0.8
% Missing