The Impact of MNAR Dropout on Estimation of Latent Growth Curve Models with Binary Observed Variables

Jason T. Newsom
Portland State University

Brian T. Keller
University of Missouri

Nicholas A. Smith
Oregon Health and Science University

&

Mallory R. Kroeck
Portland State University

Paper Presented at the Modern Modeling Methods Conference, Storrs, CT, June 2023

Objective

- Examine the performance (parameter and SE bias) of different estimation approaches to latent growth curve models (LGCM) with binary observed variables when values are missing not at random (MNAR)—a common problem with longitudinal studies due to attrition.

Background

- Increasing interest in applying these methods to binary and ordinal data (e.g., Masyn, Petras, & Liu, 2014; Newsom, 2015; Mehta, Neale, & Flay, 2004)
- Far less is known about LGCMs with binary and ordinal observed variables

Background

- Diagonalized weighted least squares estimation with robust standard errors (WLSMV) commonly used for categorical variables and works reasonably well with modest sample sizes, e.g., N=200 (Flora & Curran, 2004; Maydeu-Olivares, 2001; Muthén, du Toit, & Spisic, 1997; Yang-Wallentin, Jöreskog, & Luo, 2010)

- Marginal maximum likelihood for categorical variables with robust standard errors (MLR), more commonly used with item response theory analysis programs/procedures, but less commonly for other models. Also works reasonable well with modest sample sizes (Bandalos, 2014; DeMars, 2012)

- Bayes estimation, used for categorical data as implemented in Mplus (Muthén & Asparouhov, 2012), uses a Markov chain Monte Carlo (MCMC) estimation process that estimates and models the underlying latent scores (Albert & Chib, 1993)

Background

- A recent simulation study compared WLSMV and MLR (Newsom & Smith, 2020) extended prior work (Finch, 2017; Muthén, 1996) found:
  - Both estimation approaches performed acceptably with at least five time points and sample size of 500 or seven time points and sample size of 200
  - Three time points and 100 cases are too few for accurate estimation with either estimation approach
  - MLR had superior convergence rates but both had similar parameter and SE bias and Type I errors

- However, this study only investigated performance with complete data, a rare circumstance in applied longitudinal studies

Background

- WLSMV is limited information method with hybrid approach to missing data, with some steps based on FIML and some based on pairwise deletion, and this may be less optimal when values are MAR or MNAR (Asparouhov & Muthén, 2010; 2021)

- MLR should be expected to behave similarly to FIML for continuous variables under various missing data mechanisms (e.g., Asparouhov & Muthén, 2021) [MAR assumption]

- Bayesian estimation takes into account parameters and auxiliary variables in the multivariate posterior distribution [MAR assumption]

Background

- Asparouhov and Muthén (2021) show biased estimates for WLSMV and unbiased estimates for Bayesian for a covariance between two variables under MAR
- For a growth model with MAR missingness, ML and Bayesian estimates had minimally biased estimates of growth parameter means and variances, whereas WLSMV estimates were biased for slope means and parameter variances (Asparouhov & Muthén, 2021)
- Another simulation comparing ML (non-robust) and Bayesian (informative and noninformative priors) with dropout pattern found ML outperformed Bayesian with noninformative priors (Kim, Huh, Zhou, & Mun, 2020).
  - Missing data mechanism unclear

Study Description

- Investigate performance of estimation approaches for binary variables when dropout patterns are not MAR
- Compare smaller and larger sample sizes
- Examine convergence issues and improper solutions, parameter and standard error (posterior standard deviation) bias under various missing data mechanisms

Method

- Binary indicators
- 5 time points
- 3 estimation approaches: WLSMV, MLR, Bayes
- 2 sample sizes (N=200, N=1000)
- 3 missing data mechanism conditions

Method

- Data were generated in SAS version 9.4 and the RandomMVBinary macro (Wicklin, 2013)
- 1,000 replications per cell
- Binary variables with nearly symmetric distributions at baseline (P = .45)
- Linear slope increment proportion of .025 per wave, a moderate effect, approximately equal to standardized slope value of .4
- Intercepts and slopes designed to have significant variance

Method

- Average proportion of cases with dropout equal to .2 per wave, based on similar rates for longitudinal studies, such as the Health and Retirement Study (Heeringa & Conner, 1995)
- Missing values were created to mimic three dropout patterns (missing on y_3, y_4, and y_5; missing on y_3 and y_4; or missing on y_5)

y_1	y_2	y_3	y_4	y_5
				.
			.	.
		.	.	.

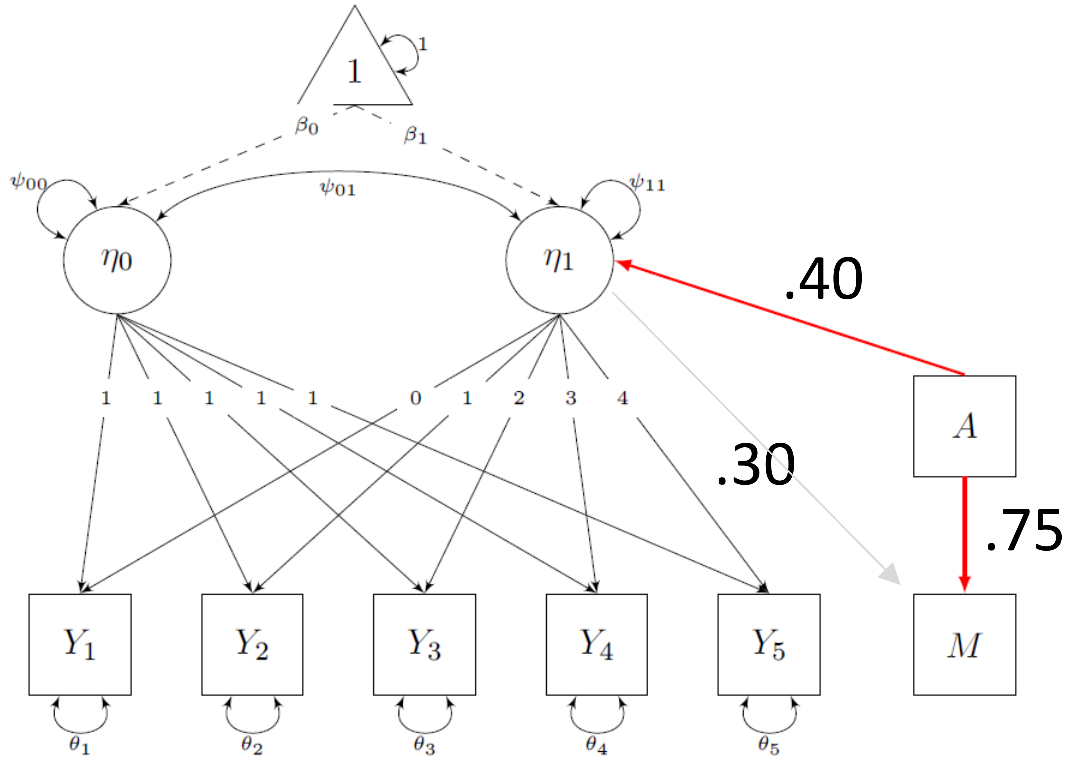
- Three patterns combined (i.e., we do not examine the differences among the three patterns)

Method

- Missing data mechanism conditions varied association of increment (slope) with probability of missingness/dropout [related to a random coefficient selection model process (Wu and Carroll, 1988)]:
  - "*Conditional MAR*" – auxiliary variable accounts for all of association with missingness
  - "*Partial MNAR*"' – auxiliary variable accounts for part of association with missingness
  - "*Pure MNAR*" – auxiliary variable accounts for none of association with missingness

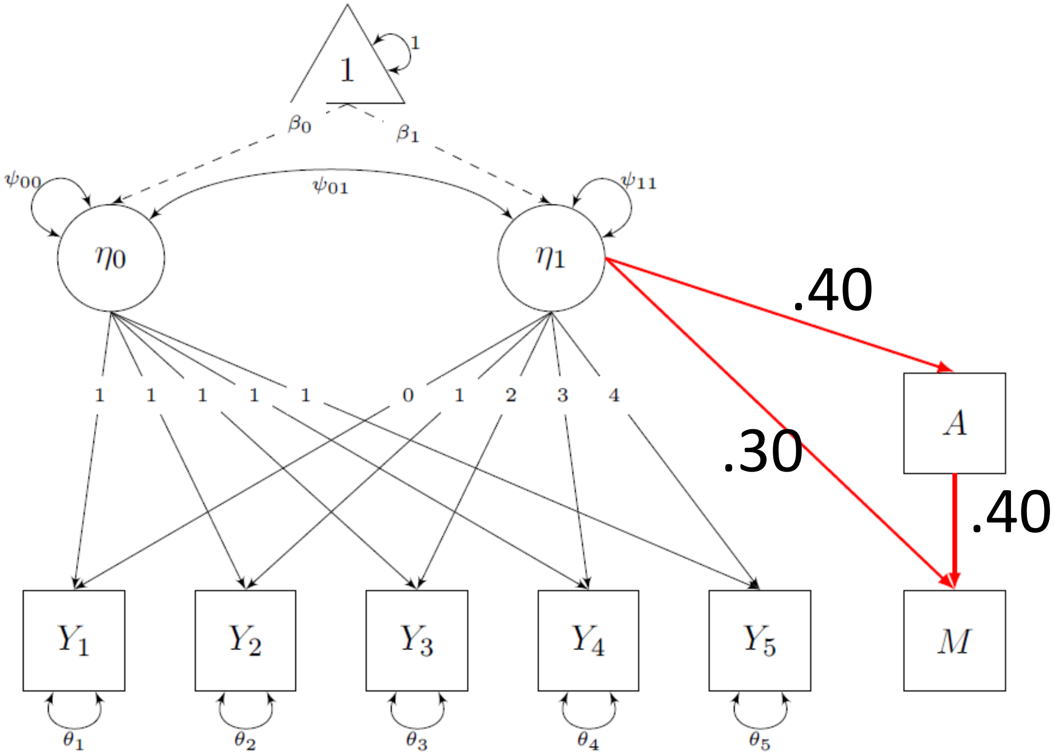
Method

"Conditional MAR"



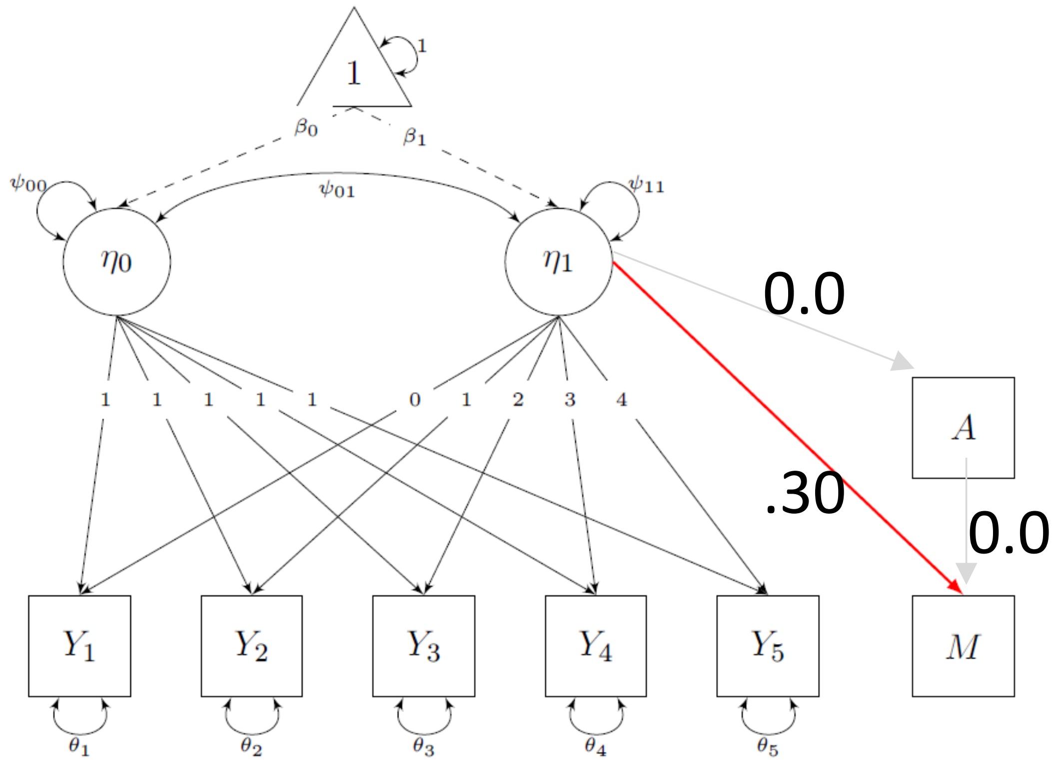
Method

"Partial MNAR"



Method

"Full MNAR"



Method

- Models were tested using Mplus version 8.9 (Muthén & Muthén, 1998-2017) using WLSMV, MLR, and Bayes estimation
- WLSMV estimation using theta parameterization with measurement intercepts equal to 0, and y variances equal to 1 (equivalent to the WLSMV1 delta specification in Newsom & Smith, 2020 with all thresholds = 0 and all scale factors = 1; based on Asparouhov & Muthén, 2002)
- MLR (probit link) used Monte Carlo integration with 8 integration points
- The Bayes estimation used the default prior, $IW(0,\text{-}m\ \text{-}1)$ for the growth parameters, a noninformative, uniform density, prior, shown to have minimal bias for factor models in smaller samples (Asparouhov & Muthén, 2021)

Method

- Because of limitations in the program for inclusion of auxiliary variables with binary indicators, the “extra-DV” approach was used (Graham, 2003)

Method

For this presentation, I will focus on the following results:

- Growth parameter mean relative bias
- SE relative bias for growth parameter mean estimates
- Growth parameter variance relative bias
- SE relative bias for growth parameter variance estimates

Samples with nonconvergence or other errors/warnings (improper solution warnings) were removed

Results

Convergence Failures and Improper Solutions

- WLSMV estimation was sensitive to the sample size, with a moderate number of samples with convergence failures or improper solutions (12%-16%) when N=200 but very few when N=1000 (<1%)
- MLR had low rates of convergence failures or improper solutions (4.2%-6.5%) in the Conditional MAR and Partial MNAR conditions, but high rates in the Pure MNAR condition for N=200 (47.6%) and modest rates for N=1000 (14.1%)
- Bayes estimation had no samples with warnings or errors for any condition

Results

Convergence Failures and Improper Solutions (out of 1,000 replications)

Conditional MAR									
	N	**WLSMV**			**MLR**			**Bayes**	
		Reps	Percent		Reps	Percent		Reps	Percent
	200	168	16.80%		56	5.60%		0	0.00%
	1000	4	0.40%		65	6.50%		0	0.00%
Partial MAR									
	200	129	12.90%		42	4.20%		0	0.00%
	1000	1	0.10%		65	6.50%		0	0.00%
Pure MAR									
	200	162	16.20%		476	47.60%		0	0.00%
	1000	3	0.30%		151	15.10%		0	0.00%

Results

Conditional MAR

- **Mean intercept** estimate bias was unacceptably high for both WLSMV (-22.76% for N = 200; -29.02% for N = 1000) and MLR (-11.87% for N = 200; -15.764% for N = 1000) but minimal bias for Bayesian (< 5%)
- **Mean slope** estimate bias was unacceptably high for WLSMV (-60.00% for N = 200; -63.63% for N = 1000) and MLR (-77.53% for N = 200; -78.42% for N = 1000) but modest or acceptable for Bayesian (-6.54% for N = 200; -2.87% for N = 1000)
- **Standard errors** for conditional MAR for both intercept and slope means tended to be slightly underestimated, had low bias for WLSMV and Bayes estimation but somewhat elevated for MLR estimation of intercepts for both sample size conditions (-12.94% for intercept) and slopes (-7.39%).

Results

Conditional MAR

- Bias in **intercept variance** estimates and their **standard errors** was largely acceptable for WLSMV for both sample size conditions, high for MLR for both sample size conditions (> 26%), and high for Bayesian when N=200 (21.30%)
- Bias in **slope variance** estimates was worse for WLSMV (N=200, 15.36% and N=1000, 20.15%), MLR (>54%) for both sample size conditions, and Bayes (N=200, 75.61% and N=1000, 14.39%)

Results

MNAR Conditions

- Both the partial MNAR and the pure MNAR showed unacceptable underestimation of **intercept and slope means** for all three of the estimators (all bias estimates > 10% for intercepts and >31% for slopes)
- **Standard error** estimates for intercept and slope means continued to generally have acceptable or low levels of bias for the WLSMV and Bayesian estimation but were usually unacceptably high for MLR estimation
- **Intercept and slope variance** estimates were generally better but with unacceptably high levels of bias for most conditions except Bayesian estimation in the larger sample size condition (<1%-12.88%) which often had percent bias less than 5%. Parameter variance **standard errors** were generally only modestly biased although worse in the larger sample size condition.

Conditional MAR, N=200

WLSMV

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-28.931%	-3.420%	0.0263	0.931
means				
i	-22.759%	-2.900%	0.0184	0.929
s	-60.000%	-1.670%	0.0066	0.779
var				
i	2.660%	3.551%	0.4066	0.931
s	15.362%	-9.926%	0.0031	0.994

MLR

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-74.800%	0.525%	0.0218	0.717
means				
i	-11.872%	-12.935%	0.0159	0.960
s	-77.525%	-7.391%	0.0082	0.596
var				
i	-26.724%	7.091%	0.3924	0.662
s	-54.848%	-3.869%	0.0024	0.503

Bayes

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	43.280%	2.700%	0.0332	0.923
means				
i	4.286%	-5.803%	0.0177	0.955
s	6.535%	0.168%	0.0036	0.952
var				
i	21.302%	6.430%	0.6193	0.899
s	75.606%	1.713%	0.0066	0.840

Conditional MAR, N=1,000

WLSMV

n=1000				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-47.328%	2.806%	0.0084	0.789
means				
i	-29.015%	1.568%	0.0068	0.799
s	-63.627%	0.000%	0.0047	0.204
var				
i	-8.253%	3.010%	0.0838	0.876
s	-20.145%	6.557%	0.0008	0.837

MLR

n=1000				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-95.760%	-13.943%	0.0164	0.323
means				
i	-15.764%	-12.249%	0.0040	0.923
s	-78.416%	-8.374%	0.0067	0.042
var				
i	-35.688%	5.136%	0.3741	0.168
s	-87.576%	-41.176%	0.0034	0.113

Bayes

n=1000				
	% bias	SE Bias	MSE	coverage
cov				
I with s	0.800%	189.276%	0.0051	0.899
means				
i	-0.690%	2.653%	0.0036	0.945
s	-2.871%	4.981%	0.0070	0.929
var				
i	2.444%	8.484%	0.0750	0.915
s	14.394%	12.941%	0.0007	0.872

Partial MNAR, N=200

WLSMV

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-7.405%	0.533%	0.0286	0.956
means				
i	-22.562%	-3.753%	0.0184	0.936
s	-61.275%	-1.848%	0.0068	0.783
var				
i	6.251%	6.123%	0.4584	0.939
s	26.377%	-8.834%	0.0035	0.993

MLR

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-130.240%	-3.391%	0.0208	0.754
means				
i	-19.261%	-11.598%	0.0167	0.941
s	-77.426%	-8.547%	0.0083	0.618
var				
i	-26.878%	1.423%	0.3856	0.685
s	-52.121%	-11.677%	0.0023	0.576

Bayes

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	36.880%	5.639%	0.0342	0.913
means				
i	-10.542%	-5.107%	0.0176	0.952
s	-31.782%	-3.970%	0.0038	0.904
var				
i	19.747%	7.847%	0.6075	0.888
s	72.576%	-0.468%	0.0064	0.838

Partial MNAR, N=1,000

WLSMV

n=1000				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-31.603%	4.674%	0.0067	0.870
means				
i	-28.719%	1.042%	0.0067	0.816
s	-63.137%	-2.655%	0.0047	0.205
var				
i	-5.587%	3.951%	0.0776	0.893
s	-8.406%	5.578%	0.0007	0.907

MLR

n=1000				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-91.360%	-23.395%	0.0154	0.370
means				
i	-25.468%	-10.623%	0.0057	0.847
s	-79.604%	-8.673%	0.0068	0.035
var				
i	-35.657%	-1.161%	0.3733	0.176
s	-85.758%	-45.378%	0.0033	0.172

Bayes

n=1000				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-4.640%	11.872%	0.0052	0.891
means				
i	-15.025%	1.864%	0.0044	0.910
s	-39.604%	-2.703%	0.0021	0.583
var				
i	0.870%	8.923%	0.0731	0.904
s	12.879%	16.288%	0.0008	0.863

Full MNAR, N=200

WLSMV

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-19.008%	-1.300%	0.0267	0.934
means				
i	-23.547%	-2.650%	0.0187	0.930
s	-60.980%	-0.923%	0.0068	0.786
var				
i	3.713%	3.198%	0.4060	0.936
s	18.841%	-8.453%	0.0033	0.999

MLR

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	-24.320%	19.369%	0.0445	0.727
means				
i	-37.340%	-30.921%	0.0288	0.882
s	-72.079%	-2.694%	0.0088	0.662
var				
i	-9.377%	19.147%	0.5408	0.752
s	9.545%	36.230%	0.0067	0.634

Bayes

n=200				
	% bias	SE Bias	MSE	coverage
cov				
I with s	33.920%	0.291%	0.0312	0.915
means				
i	-23.103%	-5.666%	0.0192	0.938
s	-63.267%	-0.743%	0.0070	0.765
var				
i	20.543%	3.224%	0.5802	0.905
s	76.970%	1.490%	0.0071	0.852

Full MNAR, N=1,000

WLSMV

n=1000	% bias	SE Bias	MSE	coverage
cov				
I with s	-45.191%	4.348%	0.0083	0.810
means				
i	-28.916%	1.568%	0.0067	0.807
s	-63.627%	0.429%	0.0048	0.208
var				
i	-8.046%	3.796%	0.0841	0.870
s	-15.942%	6.478%	0.0007	0.877

MLR

n=1000	% bias	SE Bias	MSE	coverage
cov				
I with s	-45.040%	60.058%	0.0221	0.438
means				
i	-44.384%	35.653%	0.0191	0.584
s	-74.554%	33.815%	0.0069	0.172
var				
i	-21.184%	48.628%	0.2697	0.457
s	-3.939%	74.324%	0.0044	0.345

Bayes

n=1000	% bias	SE Bias	MSE	coverage
cov				
I with s	-18.240%	11.047%	0.0053	0.875
means				
i	-29.310%	1.368%	0.0070	0.820
s	-74.158%	2.193%	0.0061	0.100
var				
i	-0.901%	9.741%	0.0708	0.907
s	2.424%	15.625%	0.0007	0.869

Portland State UNIVERSITY

Discussion

- In general, accounting fully for all the association between the parameter estimates (MAR) and the probability of missingness somewhat improved estimates, but did not eliminate bias
- Larger sample sizes helped when fully accounting the association between parameter and missingness
- Accounting partially of the association with missingness somewhat improved estimates but rarely to acceptable levels of bias
- Standard error estimates were often relatively accurate
- Parameter variance was more poorly estimated than parameter means

Discussion

- Results indicated that Bayesian estimation of growth parameter means showed superior performance when compared to MLR and WLSMV estimation, which exhibited unacceptable bias in growth parameters even when MAR was conditionally met
- Although Bayesian estimation outperformed the other two estimation approaches, in general, slope variance estimates were unacceptable even in the conditional MAR condition and poor in the MNAR conditions

Discussion

- Bayesian estimation may potentially be improved with more informative, data dependent priors (McNeish, 2016), or small variance priors
- WLSMV estimation may be improved if used in combination with an inclusive multiple imputation approach (Asparouhov & Muthén, 2021)
- Results may differ with auxiliary predicted by the parameters (?)
- Additional conditions planned including incorporating additional comparisons (complete data, listwise deletion) and estimation conditions (multiple imputation, Bayes with more informative priors), zero slope
- Additional planned outcomes: coverage, MSE, Type I errors

Thank You

Questions?

Your comments and suggestions are greatly appreciated!

Thanks to Craig Enders for valuable comments at earlier stages of this project

Contact: newsomj@pdx.edu