Missing Data Strategies for Multilevel Models

Stefany Coxe and Tyler Stout
Florida International University, Department of Psychology

Introduction

Data that exhibits clustering or multiple levels (e.g., individuals within groups, children within classrooms) is often analyzed using mixed or multilevel regression models [2, 5]

- Complicates both analysis and missing data handling

Methods exist to deal with missing data in mixed models [1,4]

- In practice, however, there is a trade-off between accuracy of estimation and usability

Method

Monte Carlo simulation in SAS 9.3
1000 replications per condition

Outcome: Continuous level 1 outcome (Y)
Predictors:
1 continuous level 1 predictor (X1) with effect size d = 0.5
1 continuous level 2 predictor (X2) with effect size d = 0.5
Cross-level interaction with effect size d = 0.15

Population model (random intercept mixed model [2, 5]):
Y = 100 + 3.02 (X1) + 3.02 (X2) + 1.51 (X1) (X2) + **G** + e
with **G** ~ N(0,15) and e ~ N(0,35)

30 clusters with 10 subjects each (N = 300)
Population intraclass correlation (ICC) = .30

Y value was missing at random (MAR)
Missingness was related to X1 value (monotonic increase):
•0% missing at lowest values of X1
•2 times the target missingness rate at highest values of X1
Target missingness rates of 1%, 5%, 15%, 25%

Analyses

Overall data analysis
Mixed model with SAS PROC MIXED in SAS [3,6]
Fixed effects of X1, X2, and X1*X2
Random intercept
Default degrees of freedom (containment)

Missing data handling:

Listwise deletion
•Complete case analysis
•Total sample size reduced from 300 to 225 to 297 (depending on missing data rate) with cluster sizes ranging from 7.5 to 9.9
•Generally not recommended for missing data; shows small bias with MAR and reduction in power

Maximum likelihood (ML)
•SAS "Method = ML"
•Recommended for mixed models when MAR holds

Expectation maximization (EM)

- Single EM imputation for missing values of Y
- EM imputation by cluster (cluster-specific mean)

Multiple imputation (MI)
•PROC MI with cluster as a categorical predictor
•EM estimates as starting values
•Markov chain Monte Carlo (MCMC) estimates for missing values
•5 imputations per replication
•PROC MIANALYZE used to pool fixed effects estimates
•Random effects pooled separately

Results

Listwise deletion

- Point estimates not severely impacted by missing data
- Standard errors inflated, especially for effects involving the variable related to missingness (X1)
- Little bias in random effects

Maximum likelihood

- Point estimates not severely impacted by missing data
- Standard errors inflated for 25% missing data rate, for effects involving the variable related to missingness (X1)

Expectation maximization

- Underestimation of fixed effects and s.e.s with missing > 15%
- Intercept variance overestimated; residual variance under
- Large overestimation of ICC

Multiple imputation

- 42 replications could not complete imputation step
- Underestimated interaction effect, all s.e.s with missing>15%
- Intercept variance overestimated

Conclusions

- All missing data approaches produced some bias in fixed effects estimates and/or variance components but the magnitude of bias varied

- Less sophisticated approaches (e.g., ML) are easy to use and show limited bias, but use caution with smaller samples and/or smaller effect sizes

Missing Data Strategies for Multilevel Models

Stefany Coxe and Tyler Stout
Florida International University, Department of Psychology

Fixed Effects (point estimates)

Percent missing		0%	1%	5%	15%	25%
Listwise deletion	Intercept	99.984	99.983	99.989	99.985	99.966
	X1	3.007	3.004	3.008	3.009	3.003
	X2	3.014	3.014	3.014	3.006	3.005
	X1*X2	1.496	1.497	1.493	1.487	1.491
Maximum likelihood	Intercept	99.984	99.983	99.989	99.984	99.965
	X1	3.007	3.004	3.008	3.009	3.003
	X2	3.014	3.014	3.014	3.006	3.004
	X1*X2	1.496	1.497	1.493	1.487	1.491
Expectation maximization	Intercept	99.984	99.968	99.912	99.763	99.597
	X1	3.007	2.973	2.855	2.560	2.256
	X2	3.014	3.007	2.979	2.895	2.817
	X1*X2	1.496	1.482	1.420	1.265	1.117
Multiple imputation	Intercept	99.984	99.983	99.991	99.996	99.986
	X1	3.007	3.003	3.008	3.010	3.001
	X2	3.014	3.000	2.973	2.860	3.732
	X1*X2	1.496	1.481	1.408	1.220	1.039

Fixed Effects (standard errors)

Percent missing		0%	1%	5%	15%	25%
Listwise deletion	Intercept	0.808	0.809	0.813	0.824	0.839
	X1	0.363	0.365	0.374	0.400	0.434
	X2	0.842	0.843	0.847	0.859	0.875
	X1*X2	0.380	0.383	0.392	0.419	0.456
Maximum likelihood	Intercept	0.808	0.782	0.785	0.796	0.810
	X1	0.363	0.364	0.372	0.398	0.431
	X2	0.842	0.814	0.818	0.830	0.845
	X1*X2	0.380	0.381	0.390	0.418	0.453
Expectation maximization	Intercept	0.808	0.809	0.814	0.826	0.848
	X1	0.363	0.361	0.355	0.339	0.320
	X2	0.842	0.843	0.848	0.861	0.884
	X1*X2	0.380	0.379	0.372	0.355	0.336
Multiple imputation	Intercept	0.808	0.811	0.823	0.855	0.897
	X1	0.363	0.365	0.375	0.402	0.438
	X2	0.842	0.845	0.857	0.889	0.930
	X1*X2	0.380	0.383	0.393	0.415	0.436

Random Effects (intercept variance)

Percent missing	0%	1%	5%	15%	25%
Listwise deletion	16.405	16.409	16.426	16.452	16.442
Maximum likelihood	16.405	15.104	15.110	15.104	15.056
Expectation maximization	16.405	16.494	16.832	17.763	19.295
Multiple imputation	16.405	16.483	16.864	17.882	19.273

Random Effects (residual variance)

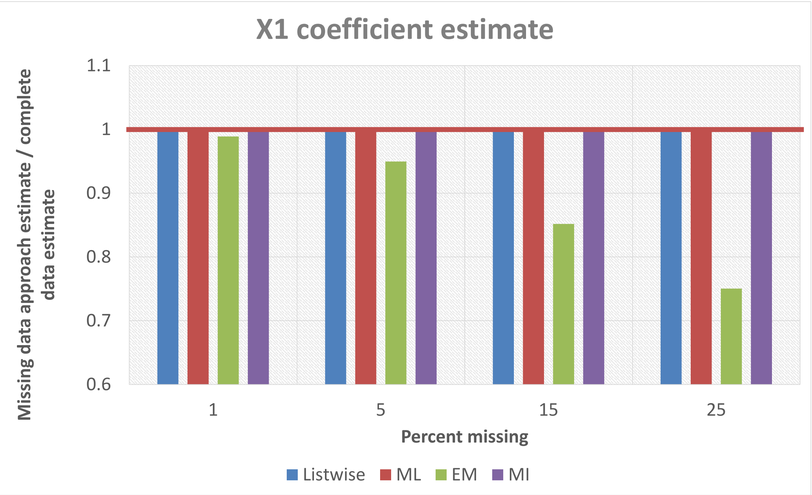
Percent missing	0%	1%	5%	15%	25%
Listwise deletion	35.189	35.187	35.226	35.254	35.185
Maximum likelihood	35.189	34.923	34.950	34.942	34.829
Expectation maximization	35.189	34.887	33.756	30.628	27.453
Multiple imputation	35.189	35.167	35.255	35.205	34.912

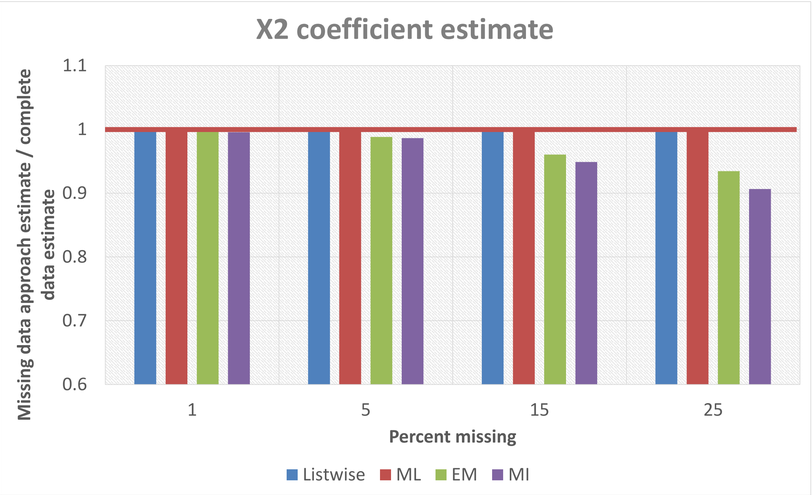
***Shaded cells are more than 10% different from complete data (0% missing) estimates**

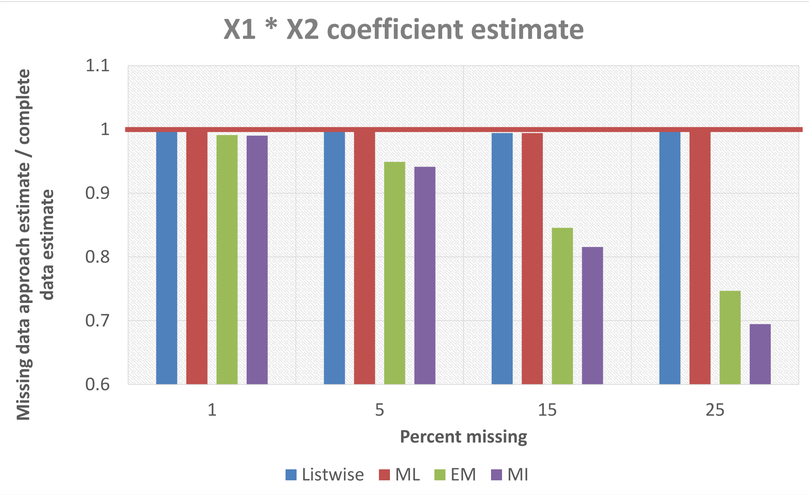
Missing Data Strategies for Multilevel Models

Stefany Coxe and Tyler Stout
Florida International University, Department of Psychology

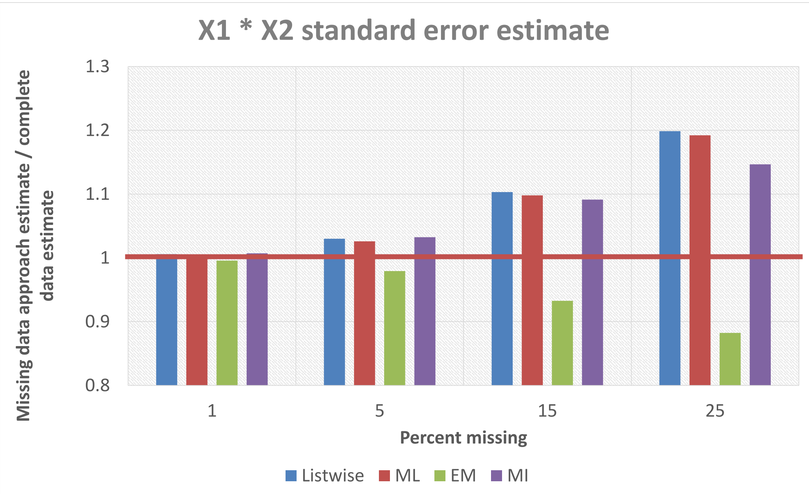
Fixed effects: Underestimation increases with more missingness, especially for EM and MI and for effects involving X1







Standard errors: LD, ML, MI overestimate and EM underestimates s.e.s, especially with more missingness and for effects involving X1

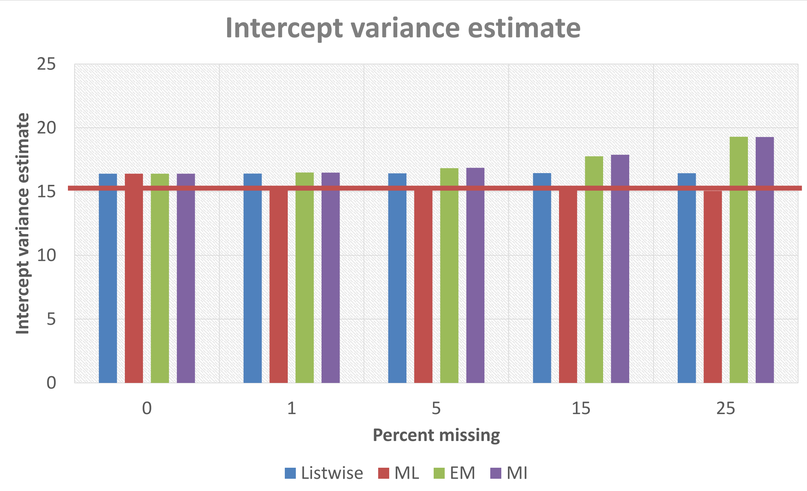


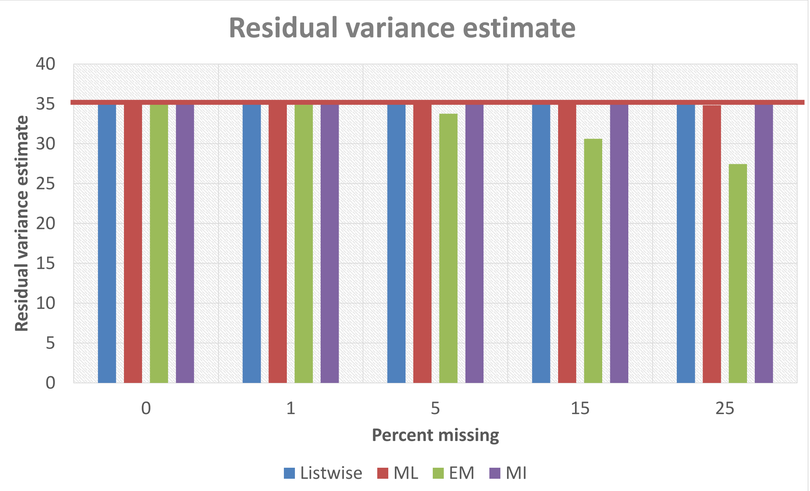
LD = listwise deletion; ML = maximum likelihood; EM = expectation maximization; MI = multiple imputation

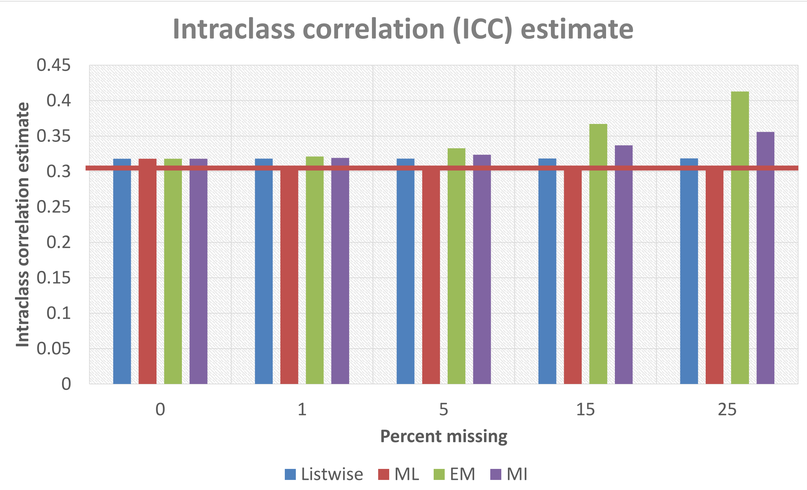
Missing Data Strategies for Multilevel Models

Stefany Coxe and Tyler Stout
Florida International University, Department of Psychology

Random Effects: intercept variance underestimated by ML, overestimated by LD, EM, MI; residual variance underestimated by EM; ICC overestimated by EM and MI







References

1. Enders, C. K. (2010). *Applied missing data analysis*. Guilford Press.
2. Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.
3. Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberger, O. (2006). SAS for mixed models. SAS Institute. *Inc., Cary, NC*, 814.
4. Mistler, S. A. (2013, April). A SAS macro for applying multiple imputation to multilevel data. In *Proceedings of the SAS Global Forum*.
5. Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1). Sage.
6. Singer, J. D. (1998). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *Journal of educational and behavioral statistics*, *23*(4), 323-355.

LD = listwise deletion; ML = maximum likelihood; EM = expectation maximization; MI = multiple imputation