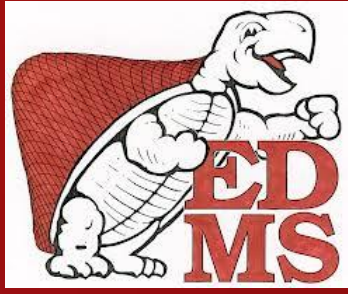


CLUSTERED DATA WITH A SMALL NUMBER OF CLUSTERS: COMPARING THE PERFORMANCE OF MODEL-BASED AND DESIGN-BASED APPROACHES

Dan McNeish and Jeffrey Harring

University of Maryland, College Park
Measurement, Statistics, and Evaluation

Background

- When data are collected, frequently observations are clustered within higher level units
- Common methods for handling clustered data require an “adequately large” number clusters
- This presentation aims to compare common methods when the number of clusters is small

Design Based vs. Model Based Methods

- Design Based – estimates take clustering into account by statistical corrections
  - Robust standard errors, Generalized Estimating Equations (GEE, Delta Estimator Method), or "Sandwich Estimators"

- Model Based – estimates take clustering into account by including the source of clustering into the model
  - Multi-level models or "HLM"

Minimum Sample Size, No Adjustments

- Design Based
  - With continuous outcomes, between 50 and 100 clusters may be necessary to obtain unbiased unadjusted GEE SE estimates (Morel, Bokossa, & Neerchal, 2003)
- Model Based
  - For continuous outcomes, between 20 and 30 clusters are suggested to obtain unbiased SE estimates for fixed effects and at about 100 clusters for variance component estimates (Mass & Hox, 2005)

Small Sample Size Adjustments

- Design Based
  - A few different methods have been proposed and implemented into software such as SAS (Proc Genmod, Proc Glimmix)
    (Hinkley,1977; Morel, Bokossa, and Neerchal, 2003)

- Model Based
  - Using a Kenward-Roger degree of freedom adjustment has been found to reduce bias in SE estimates with a small number of clusters
    (Bell et al, 2010)

Morel, Bokossa, Neerchal (2003)

- Applies a correction factor to the Huber-White estimator $[\tilde{V}(\hat{\beta})]$
- Correction factor is a function of the design effect (ϕ) and a multiplier (δ_n)that converges to zero as sample size increases
- $\tilde{V}(\hat{\beta}) + \delta_n \Phi [I_0(\hat{\beta})]^{-1}$
- [Sandwich estimates] + [multiplier*design effect *(model based SE)]

Kenward-Roger (1997)

- Formulae and derivation are lengthy and not the focus of this presentation
- Inflates the covariance matrix first to reduce the underestimation that occurs from a small number of clusters
- Then, the Satterthwaite degree of freedom approximation is applied to the inflated covariance matrix

Research Question

- If one has clustered data, as is common in social/behavioral sciences, how do different methods to account for clustering compare with a small number of clusters for the marginal model (fixed effects)?

Simulated Model

Generating Model:

$$y_{ij} = \beta_{0i} + \beta_{1i}X_{1i} + \beta_{2i}X_{2i} + \beta_{3i}X_{1i} * X_{2i} + r_{ij}$$
$$\beta_{0i} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{1j} * Z_{2j} + u_{0j}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \gamma_{13}Z_{1j} * Z_{2j} + u_{1j}$$
$$\beta_{2i} = \gamma_{20} + \gamma_{21}Z_{1j} + \gamma_{22}Z_{2j} + \gamma_{23}Z_{1j} * Z_{2j} + u_{2j}$$
$$\beta_{3i} = \gamma_{30} + \gamma_{31}Z_{1j} + \gamma_{32}Z_{2j} + \gamma_{33}Z_{1j} * Z_{2j} + u_{3j}$$

- Variables with 1 subscript are continuous and ~N(0,1)
- Variables with 2 subscript are binary with 50:50 prevalence
- Interactions are included at both L1 and L2
- Each coefficient has a random effect
- The fitted model had no misspecifications

Factors Manipulated	
*ICC (*4 levels) • 0.00 • 0.10 • 0.20 • 0.50 *Statistical Model (4 levels)* • Multi-level (HLM) • Multi-level (HLM) with Kenward-Roger • GEE Estimator • GEE Estimator with MBN adjustment	*Number of clusters (5 levels)* • 10-50, intervals of 10 *Cluster Size (4 levels)* • 5,15,30,50 • 1,000 replications were performed for each cell for 80,000 total replications • Each replicated dataset was analyzed using all 4 models

Note On Software

- Model estimates were obtained from SAS 9.2
  - HLM model- Proc Mixed
  - K-R model- Proc Mixed with DDFM=KR
  - GEE model – Proc Genmod with REPEATED
  - Adjusted GEE- Proc Glimmix with EMPIRICAL=MBN

Outcome Measure 1

- 95% Coverage Rate
  - Percentage of replications where the true value for the parameter was within the estimated confidence interval
- If standard errors are estimated accurately, this outcome measure should be close to 0.95
- Values less than 0.925 will be considered to have underestimated standard errors while values greater than 0.975 will be considered to have overestimated standard errors (Bradley, 1978)

Outcome Measure 2

- Statistical Power
  - Percentage of replications where 0 was not in the confidence interval for each parameter
- True values were non-zero so the null hypothesis of $\beta=0$ would be false
- The ability to reject this false null hypothesis is the definition of power

95% Coverage Rate

GEE								
NC	INT	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.75	0.83	0.79	0.68	0.76	0.85	0.64	0.76
20	0.87	0.89	0.88	0.78	0.89	0.90	0.82	0.85
30	0.90	0.90	0.91	0.80	0.91	0.91	0.87	0.88
40	0.91	0.91	0.92	0.79	0.92	0.92	0.88	0.89
50	0.92	0.91	0.92	0.77	0.93	0.92	0.89	0.90

MBN (2003)								
NC	INT	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.98	0.94	0.97	0.97	0.98	0.94	0.93	0.95
20	0.98	0.95	0.98	0.97	0.99	0.97	0.95	0.98
30	0.99	0.96	0.98	0.97	0.99	0.97	0.96	0.98
40	0.98	0.96	0.98	0.95	0.99	0.97	0.96	0.97
50	0.98	0.95	0.98	0.92	0.99	0.96	0.95	0.97

95% Coverage Rate

HLM								
NC	INT	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.97	0.96	0.97	0.92	0.94	0.94	0.91	0.92
20	0.96	0.95	0.95	0.91	0.95	0.94	0.93	0.92
30	0.95	0.95	0.95	0.89	0.95	0.94	0.93	0.92
40	0.95	0.95	0.95	0.87	0.95	0.94	0.91	0.91
50	0.95	0.95	0.95	0.84	0.95	0.94	0.90	0.92

Kenward-Roger								
NC	INT	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.96	0.96	0.95	0.94	0.95	0.94	0.94	0.93
20	0.95	0.95	0.95	0.92	0.95	0.94	0.95	0.93
30	0.95	0.95	0.95	0.90	0.95	0.94	0.94	0.92
40	0.95	0.95	0.95	0.87	0.95	0.94	0.92	0.92
50	0.95	0.95	0.95	0.84	0.95	0.94	0.91	0.92

95% Coverage Rate

GEE

NC	INT	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.75	0.83	0.79	0.68	0.76	0.85	0.64	0.76
20	0.87	0.89	0.88	0.78	0.89	0.90	0.82	0.85
30	0.90	0.90	0.91	0.80	0.91	0.91	0.87	0.88
40	0.91	0.91	0.92	0.79	0.92	0.92	0.88	0.89
50	0.92	0.91	0.92	0.77	0.93	0.92	0.89	0.90

HLM

NC	INT	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.97	0.96	0.97	0.92	0.94	0.94	0.91	0.92
20	0.96	0.95	0.95	0.91	0.95	0.94	0.93	0.92
30	0.95	0.95	0.95	0.89	0.95	0.94	0.93	0.92
40	0.95	0.95	0.95	0.87	0.95	0.94	0.91	0.91
50	0.95	0.95	0.95	0.84	0.95	0.94	0.90	0.92

95% Coverage Rate

MBN (2003)								
NC	INT	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.98	0.94	0.97	0.97	0.98	0.94	0.93	0.95
20	0.98	0.95	0.98	0.97	0.99	0.97	0.95	0.98
30	0.99	0.96	0.98	0.97	0.99	0.97	0.96	0.98
40	0.98	0.96	0.98	0.95	0.99	0.97	0.96	0.97
50	0.98	0.95	0.98	0.92	0.99	0.96	0.95	0.97

Kenward-Roger								
NC	INT	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.96	0.96	0.95	0.94	0.95	0.94	0.94	0.93
20	0.95	0.95	0.95	0.92	0.95	0.94	0.95	0.93
30	0.95	0.95	0.95	0.90	0.95	0.94	0.94	0.92
40	0.95	0.95	0.95	0.87	0.95	0.94	0.92	0.92
50	0.95	0.95	0.95	0.84	0.95	0.94	0.91	0.92

Standard Error Bias

- Unadjusted GEE SE estimates were consistently underestimated for continuous outcomes even with 50 clusters
  - Inflated Type-I error rates
- Using the MBN correction, SE were reasonably estimated and overestimation was actually more problematic (deflated Type-I error rate)
- HLM standard error estimates for L1 predictors and binary L2 predictors were surprisingly accurate
- Kenward-Roger adjustment seemed to perform well except for continuous level 2 predictors

Standard Error Bias

- The unadjusted design based and model based estimates seemed to have some issues with a small number of clusters
- The small sample size adjustments did a reasonable job although some estimates were still overestimated (MBN) or underestimated (K-R)
- While both the Kenward-Roger and MBN methods seem to do a reasonable job with standard error estimation, which method is more powerful to detect true differences may also be of interest
- Only the K-R and MBN methods will be compared since they best estimated the standard errors

The top number in each cell is power for Kenward-Roger

The bottom number in each cell is power for MBN

Power

NC	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.23	0.19	0.21	0.13	0.18	0.23	0.40
	0.14	0.10	0.09	0.04	0.18	0.18	0.28
20	0.43	0.34	0.37	0.29	0.33	0.45	0.61
	0.16	0.14	0.17	0.07	0.26	0.37	0.45
30	0.55	0.45	0.48	0.39	0.45	0.57	0.71
	0.19	0.19	0.23	0.13	0.34	0.49	0.55
40	0.63	0.52	0.55	0.47	0.57	0.63	0.76
	0.26	0.29	0.35	0.25	0.45	0.58	0.64
50	0.69	0.58	0.61	0.53	0.66	0.68	0.80
	0.34	0.40	0.44	0.36	0.54	0.65	0.71

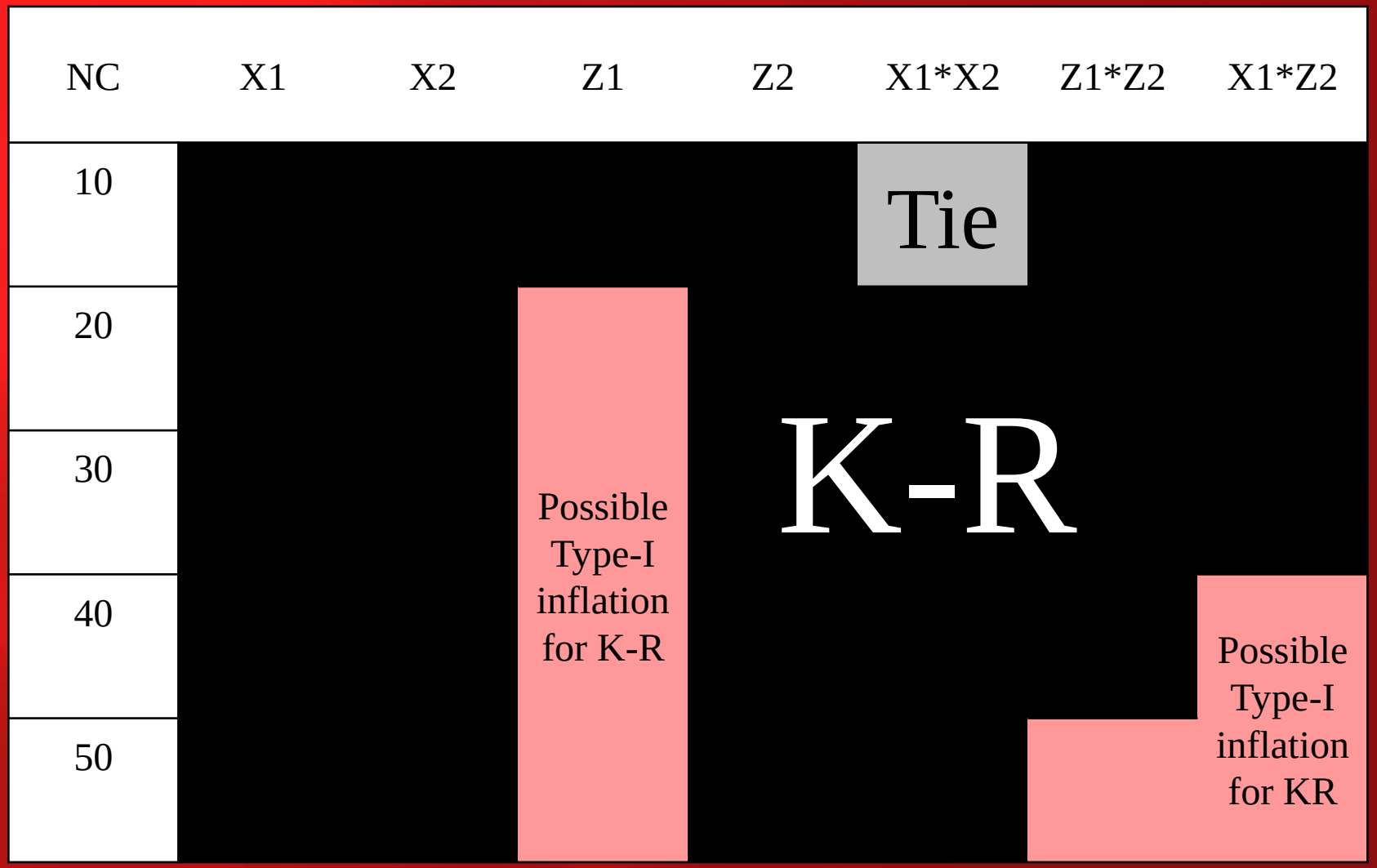
The top number in each cell is power for Kenward-Roger

The bottom number in each cell is power for MBN

Power

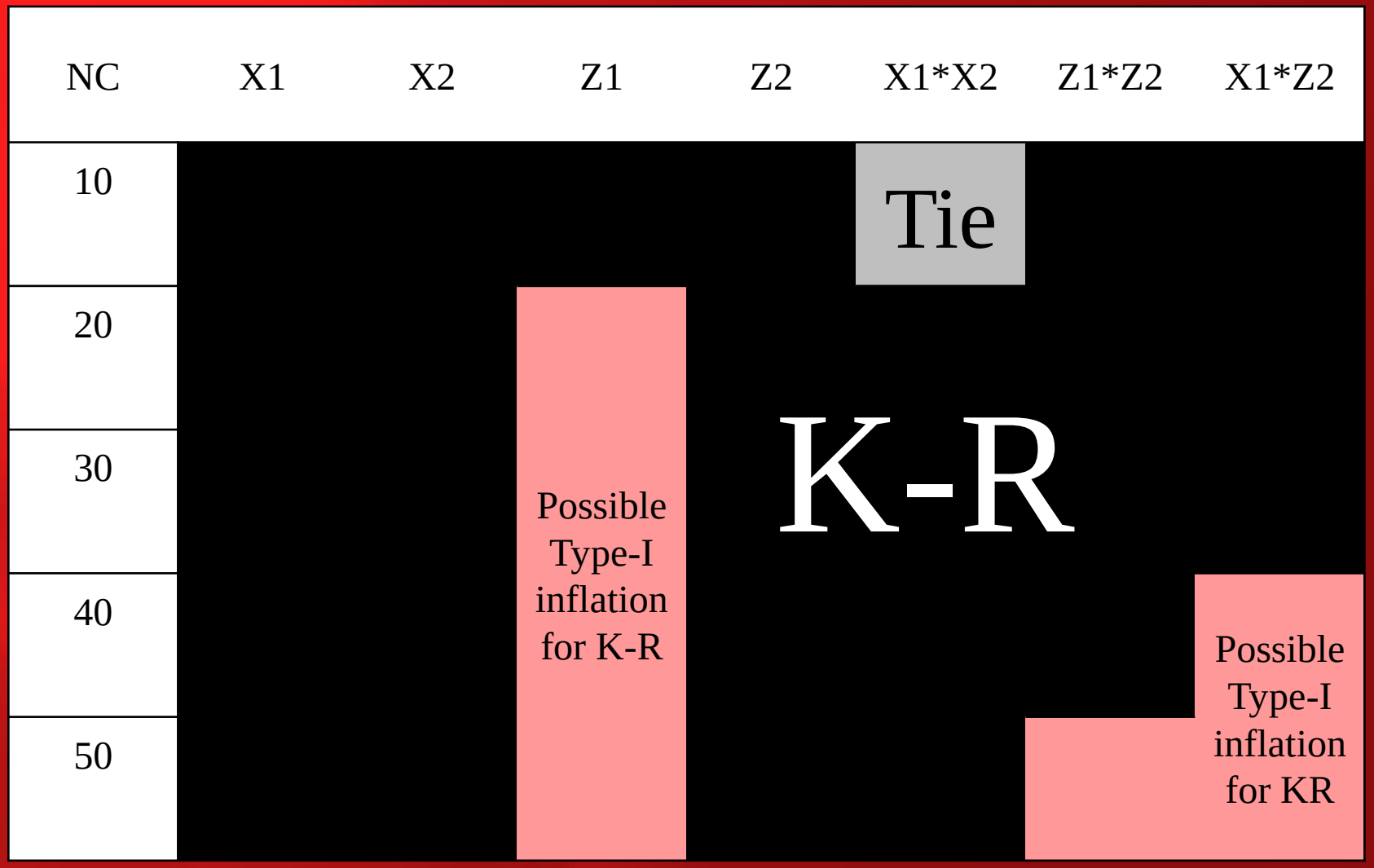
NC	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.23	0.19	0.21	0.13	0.18	0.23	0.40
	0.14	0.10	0.09	0.04	0.18	0.18	0.28
20	0.43	0.34	0.37	0.29	0.33	0.45	0.61
	0.16	0.14	0.17	0.07	0.26	0.37	0.45
30	0.55	0.45	0.48	0.39	0.45	0.57	0.71
	0.19	0.19	0.23	0.13	0.34	0.49	0.55
40	0.63	0.52	0.55	0.47	0.57	0.63	0.76
	0.26	0.29	0.35	0.25	0.45	0.58	0.64
50	0.69	0.58	0.61	0.53	0.66	0.68	0.80
	0.34	0.40	0.44	0.36	0.54	0.65	0.71

Power


NC
X1
X2
Z1
Z2
X1*X2
Z1*Z2
X1*Z2
10
20
30
40
50
Tie
K-R
Possible Type-I inflation for K-R
Possible Type-I inflation for KR

For small number of clusters, the model-based approach with Kenward-Roger adjustment was never less powerful than the design-based approach with MBN adjustment

Power



Maximum advantage of KR over MBN generally occurs at 30 clusters starts to decline thereafter

Power

NC	X1	X2	Z1	Z2	X1*X2	Z1*Z2	X1*Z2
10	0.09	0.09	0.12	0.09	0.00	0.05	0.12
20	0.27	0.20	0.20	0.22	0.07	0.08	0.16
30	0.36	**0.26**	**0.25**	**0.26**	0.11	**0.08**	**0.16**
40	**0.37**	0.23	0.20	0.22	**0.12**	0.05	0.12
50	0.35	0.18	0.17	0.17	0.12	0.03	0.09
Max	0.37	0.26	0.25	0.26	0.12	0.08	0.16

Conclusions

- Both the Kenward-Roger and MBN adjustments seem to improve standard error estimates with a small number of clusters with the MBN adjustment having a larger impact
- The unadjusted GEE approach did not estimate standard errors well for any estimate in any condition
- The Kenward-Roger adjustment never was less powerful than the MBN adjustment
- Even if variance components are not of interest, using a model based approach with the Kenward-Roger adjustment seemed to reasonably estimate SE and provide higher power than the GEE methods when number of clusters is small

Limitations & Future Directions

- Did not explore longitudinal data, which both methods are able to handle
- The delta parameter in the MBN adjustment is a function of a few parameters
  - SAS defaults were used but perhaps different values would perform better for small numbers of clusters.
- Comparisons were relative
  - One method outperforming another does not necessarily mean it works well in absolute terms