Analyzing Group Randomized Trials (with a Few Clusters and Binary Outcomes)

Francis Huang, PhD

2021.06.04

@flhuang
https://faculty.missouri.edu/huangf
huangf@missouri.edu

Logistic regression[1] models have been the traditional 'go-to' models when analyzing data with binary outcomes

- Alternative models though may be able to provide more straightforward results
- An additional concern is the use of these models for the analysis of cluster randomized trials

Prevention Science
https://doi.org/10.1007/s11121-021-01228-5

Alternatives to Logistic Regression Models when Analyzing Cluster Randomized Trials with Binary Outcomes

Francis L. Huang[1]

Accepted: 21 March 2021

[1]Outside of education and psychology, probit models may be popular as well.

Some challenges with logistic regression

- Hard to interpret coefficients (e.g., logits, odds, odds ratios, probabilities)
  - Called 'unintelligible' and often a reason researchers present ORs but do not talk about the practical implications → often, readers do not know the difference! This though is not an issue of the model but a matter of understanding
- In an experiment, due to randomization, a comparison of treatment vs control outcomes usually will suffice
  - Covariates are added to improve model power
  - However, with logistic regression, adding variables to a model– even if uncorrelated with the treatment assignment variable– can change the treatment coefficient and does not necessarily improve power to detect effects!
- There is a difference between population average and cluster specific effects when dealing with multilevel logit models!

Interpretation case example (no need for a calculator!)

- On average, students have a 60% probability of passing an exam. An intervention is then tested: students are randomly assigned to either a treatment or control condition and then given the exam. The results of the experiment show that students in the treatment condition had double the odds of passing the exam compared to a student in the control group (i.e., OR = 2.0). What are the probabilities of passing for students in the treatment condition?
  - A. 70%
  - B. 75%
  - C. 85%
  - D. 95%
  - E. 120%

Interpretation case example: Solution

- On average, students have a 60% probability of passing an exam. An intervention is then tested: students are randomly assigned to either a treatment or control condition and then given the exam. The results of the experiment show that students in the treatment condition had double the odds of passing the exam compared to a student in the control group (i.e., OR = 2.0). What are the probabilities of passing for students in the treatment condition?
- Odds of passing = .60/.40 = 3 / 2 = 1.5
- Double the odds in treatment = 1.5 x 2 = 3.0
- Convert odds to probabilities:
  - odds / (1 + odds)
  - 3 / (1 + 3) = .75
- Another way of saying this is the likelihood of passing is (.75 / .60 = 1.25) is 25% higher in the treatment group ('risk' ratio → if the outcome is negative/detrimental)
- Or can say that treatment group had 15 percentage points higher of passing ('risk' difference or percentage difference)

Moving this to a clustered data setting: CS vs PA results

Example Showing Differences Between Cluster-Specific and Population Average (PA) Results

	Control		Treatment		
School	Prob	Odds	Prob	Odds	*OR*
A ($n = 200$)	.50	1.00	.67	2.00	2.00

Note. *OR* = Odds ratio. Prob = probability.

Moving this to a clustered data setting: CS vs PA results

Example Showing Differences Between Cluster-Specific and Population Average (PA) Results

	Control		Treatment		
School	Prob	Odds	Prob	Odds	*OR*
A ($n = 200$)	.50	1.00	.67	2.00	2.00
B ($n = 200$)	.70	2.33	.82	4.66	2.00
C ($n = 200$)	.30	0.43	.46	0.86	2.00

Note. *OR* = Odds ratio. Prob = probability.

Moving this to a clustered data setting: CS vs PA results

Example Showing Differences Between Cluster-Specific and Population Average (PA) Results

	Control		Treatment		
School	Prob	Odds	Prob	Odds	*OR*
A ($n = 200$)	.50	1.00	.67	2.00	2.00
B ($n = 200$)	.70	2.33	.82	4.66	2.00
C ($n = 200$)	.30	0.43	.46	0.86	2.00
(A) Average	**.50**		**.65**		

Note. *OR* = Odds ratio. Prob = probability.

Moving this to a clustered data setting: CS vs PA results

Example Showing Differences Between Cluster-Specific and Population Average (PA) Results

	Control		Treatment		
School	Prob	Odds	Prob	Odds	*OR*
A ($n = 200$)	.50	1.00	.67	2.00	2.00
B ($n = 200$)	.70	2.33	.82	4.66	2.00
C ($n = 200$)	.30	0.43	.46	0.86	2.00
(A) Average	**.50**		**.65**		
Computed PA *OR* based on (A) averages		1.00		1.86	**1.86**

Note. *OR* = Odds ratio. Prob = probability.

PA results (with a logistic mixed model; not a linear mixed model) will be closer to zero compared to CS results. With interventions, our interest is generally on the PA results.

Viable alternatives (with experiments with random assignment) when outcomes are binary are:

- The linear probability model
- The log-binomial model
- The Poisson regression model

Often, in CRTs and experimental studies, the interest is in the treatment effect ("did it work?")

Have already shown that these can be effective in an experimental setting… but not in a clustered data setting with only a few clusters

The Journal of Experimental Education

ISSN: 0022-0973 (Print) 1940-0683 (Online) Journal homepage: https://www.tandfonline.com/loi/vjxe20

Alternatives to logistic regression models in experimental studies

Francis L. Huang

Why can a linear probability model be used?

- Several issues with LPMs (i.e., using OLS with a binary outcome)
  - Relationship modeled is linear when true relationship may be nonlinear
  - Violations of homoskedasticity
  - Result in 'out-of-bound' values (probabilities >1, < 0)

- However, with experiments, these are not necessarily issues:
  - Interest is in the outcome shift based on treatment assignment (can only assume a one or a zero)– will be an issue if the predictor of interest is continuous but not with an experiment (T = 1 vs T = 0)
  - Heteroskedasticity consistent standard error corrections have been around for a long time– can adjust for that
  - If the shift is 1 vs 0, not likely to happen (only moving 1 unit, not unless prevalence rates are extremely high or extremely low)
- Results in a percent change (e.g., 'risk' difference) instead of a logit or OR

What about a Poisson regression model?

- Traditional logistic model is: $\log(\pi/[1-\pi]) = \beta 0 + \beta 1 X 1$.
  - Model the log of the odds, exponentiating the coefficient gives the odds ratio (logit link and binomial distribution)
- A 'cousin' of the logistic regression model is the log-binomial model where: $\log(\pi) = \beta 0 + \beta 1 X 1$→ the outcome is log transformed but still uses a binomial distribution.
  - Exponentiating the coefficient gives an incident rate ratio (IRR) or risk ratio (RR) which is more readily understood as well compared to an OR
  - However, log binomial models are known to have convergence issues (so can't recommend this)
- One step further is the Poisson model where instead of the binomial distribution, the Poisson distribution is used instead:
  - Poisson regression is often used for count outcomes: in this case, a zero and one are counts with no chance of getting outcomes higher than one.
  - However, standard errors have to be adjusted: if not, standard errors will be too HIGH. Correction though is merely the robust standard error

A traditional and simple way of dealing with clustered data is to use 'cluster robust standard errors' (CRSEs) or 'sandwich' estimators

- Popularized by Liang and Zeger (1986):

$$var(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\Sigma} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \sum_{\boldsymbol{g}=1}^{\boldsymbol{G}} \boldsymbol{X}_{\boldsymbol{g}}' \widehat{\boldsymbol{\Omega}}_{\mathbf{g}} \boldsymbol{X}_{\boldsymbol{g}} (\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$\widehat{\boldsymbol{\Omega}}_{\mathbf{g}}$ are the squared residuals per cluster (i.e., $\boldsymbol{e}_{\boldsymbol{g}}\boldsymbol{e}_{\boldsymbol{g}}'$)

- Has been a traditional procedure in many fields to just cluster the standard errors
  - Available in many software packages: Stata, SAS
- However, CRSEs have the known limitation of still underestimating SEs if the number of clusters is still too low:
  - What is too low? < 100 (Maas & Hox, 2004); < 50 (Cameron & Miller, 2015.; Angrist & Pischke, 2008)
- Many CRTs though have < 30 groups! Then, we cannot use traditional CRSEs because of the risk of Type I errors!

The need to focus on methods to analyze cluster randomized trials (CRTs) with a limited number of clusters is of practical consideration…

- A review of 285 CRTs for health/medical interventions indicated that the median number of clusters was 21 (Ivers et al., 2011)
- Examples (from *Prevention Science*):
  - A first-aid training program for resident assistants was randomized between eight campuses with 566 participants (Thombs, Gonzalez, Osborn, Rossheim, & Suzuki, 2015).
  - A CRT tested a prevention program to reduce harmful legal product use and was randomized to 14 communities with 496 middle schoolers (Johnson et al., 2009).
  - To test an anti-bullying prevention program, Axford et al. (2020) had 3,214 student participants from 21 schools.
- **Of course running projects with more clusters is desirable but not always logistically possible! Have to balance this with the power requirements to find an effect!**

However, few cluster corrections have been around for nearly 20 years (Bell & McCaffrey, 2002; Mancl & DeRouen, 2001)!

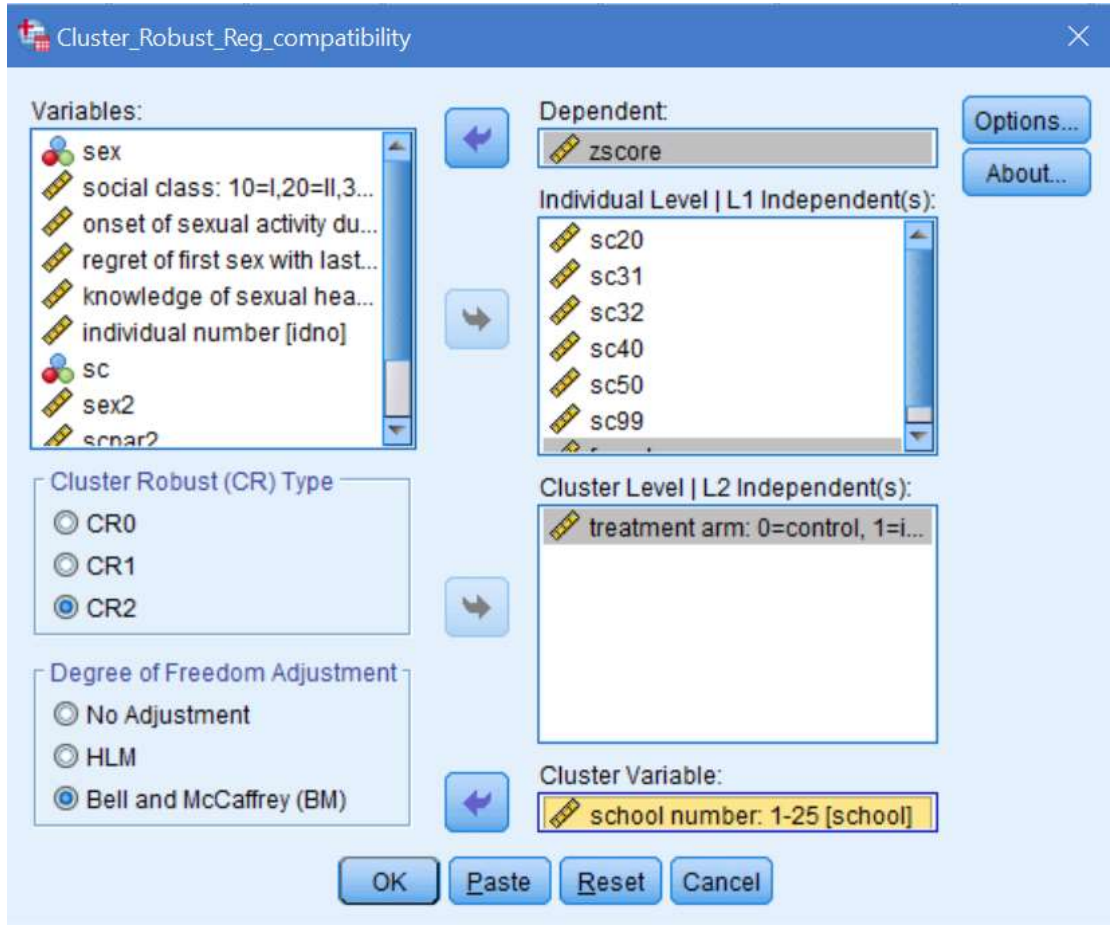
- In particular, I focus on the work of Bell and McCaffrey's (BM; 2002) biased-reduced linearization (BRL) estimator,
  - Has been the focus of more recent technically focused studies (e.g., Imbens & Kolesar, 2016) with "most papers … written by and for economists" (MacKinnon & Webb, 2017, p. 2) and statisticians with a focus on proofs and derivations
- Recent studies have referred to this as the CV2 (MacKinnon & Webb, 2017), CR2VE (Cameron & Miller, 2015), LZ2 (Imbens & Kolesar, 2016), or the CR2 estimator (Pustejovsky & Tipton, 2018).
- The suffix '2' is used with the correction since it is a generalization of the heteroskedasticity consistent standard error estimator (HC2) proposed by MacKinnon and White (1985) which also makes use of the hat matrix but extended for use in a clustered data setting.
- The use of the BRL estimator may not be readily accessible to applied researchers and is not available in some commonly-used statistical programs such as SPSS or Mplus. Even in a multilevel setting, this variant is not available in SAS or HLM
- In addition, BM propose a degrees of freedom (dof) adjustment which is important in settings with a few clusters

With the BM adjustment, part of the "meat" matrix is

- $\widehat{\boldsymbol{\Omega}}_{\mathbf{g}} = \left[\boldsymbol{I}_{\boldsymbol{N}_{\boldsymbol{g}}} - \boldsymbol{H}_{\boldsymbol{gg}}\right]^{-\frac{\mathbf{1}}{\mathbf{2}}} \boldsymbol{e}_{\boldsymbol{g}} \boldsymbol{e}_{\boldsymbol{g}}' \left[\boldsymbol{I}_{\boldsymbol{N}_{\boldsymbol{g}}} - \boldsymbol{H}_{\boldsymbol{gg}}\right]^{-\frac{\mathbf{1}}{\mathbf{2}}}$ (see Imbens & Kolesar, 2016, p. 709) where $\boldsymbol{I}_{\boldsymbol{N}_{\boldsymbol{g}}}$ is an identity matrix for group g, $\boldsymbol{H}_{\boldsymbol{gg}}$ is the hat matrix for cluster g, and $\left[\boldsymbol{I}_{\boldsymbol{N}_{\boldsymbol{g}}} - \boldsymbol{H}_{\boldsymbol{gg}}\right]^{-\frac{\mathbf{1}}{\mathbf{2}}}$ is the inverse of the symmetric square root of the matrix $\left[\boldsymbol{I}_{\boldsymbol{N}_{\boldsymbol{g}}} - \boldsymbol{H}_{\boldsymbol{gg}}\right]$
- This is substituted in the sandwich formulation
- Using the inverse of the symmetric square root of the matrix has been shown to greatly reduce the bias even when the working covariance matrix does not follow $\sigma^2 I$ such as when the residuals within clusters are correlated with each other (i.e., the intraclass correlation coefficient or ICC ρ ≠ 0) (Bell & McCaffrey, 2002)

Aside from providing R syntax computing the CR2 and the BM dof adjustment…

- … we have developed a graphical interface for SPSS
- Took a while as SPSS handles matrices differently
- Thought this would be done before other papers but the review has taken longer…

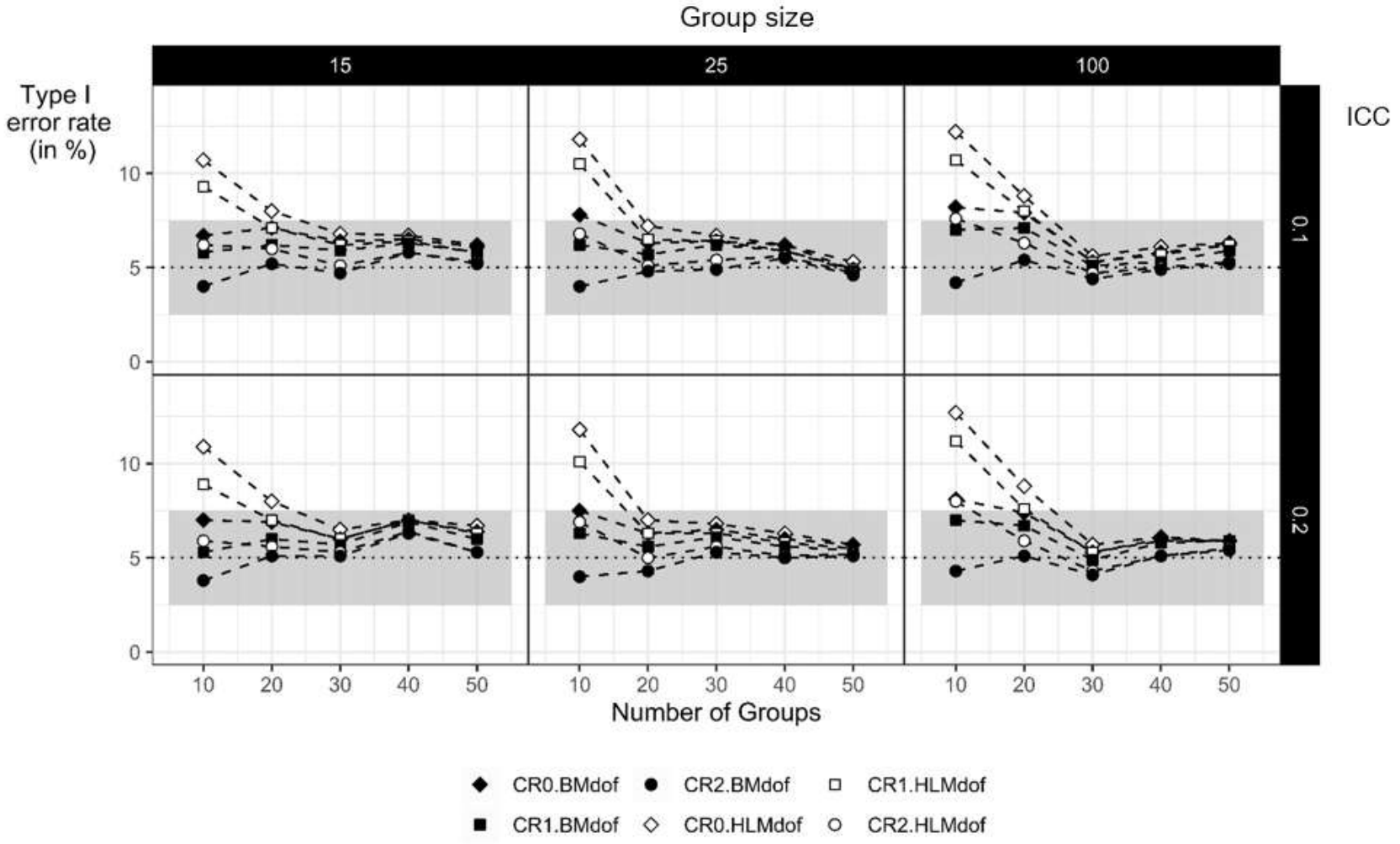


Using cluster robust standard errors when analyzing group randomized trials with few clusters. Huang, F., & Li, X. (Accepted, 2021.05.17). *Behavior Research Methods.*

Simulating a CRT, the Type I error rates of the group treatment effect was investigated under several conditions…

- Type I error rates were acceptable with the CR2 estimator for all conditions using the BM dof adjustment
- Worked well even with just 10 clusters
- The traditional CRSEs (CR0 and CR1) did not have acceptable Type I error rates (as has already been shown)

Using Cluster Robust Standard Errors when Analyzing Group Randomized Trials with Few Clusters. Huang, F., & Li, X. (Accepted). BRM



So returning to our alternatives to logistic regression with clustered data…

- We have three viable methods (logistic, LPM, and Poisson)
- All models just use a standard GLM
- Need though to have the robust cluster adjustment, even with a limited number of clusters (the CR2 estimator)
- How do they compare?
  - Test using a simulation where we know the 'truth'!
    - Conditions include number of clusters, group size, ICC, prevalence rate
    - Analyze outcome using logistic regression, linear regression, modified Poisson
  - Investigate a level 2 treatment effect

- IF point estimates are biased, cannot recommend these
- IF standard errors are biased, cannot recommend these as well!

- Based on a simulation, the LPM and mod Pois models had comparable point estimates and standard errors (i.e., results were acceptable)

  The traditional CRSEs though had SEs that were too low

  CR2 did well

Table 2 Bias in point estimates and standard errors per simulation condition and model estimated

Intercept	ICC	NG	GS	Point estimates			Standard errors			
							CR0	CR2		
				LRM	LPM	Pois	LRM	LRM	LPM	Pois
1	.10	20	20	0.6	−0.9	0.1	−7.6	−1.8	−1.1	−1.1
			100	3.9	3.0	3.7	−10.9	−5.4	−4.9	−4.5
		50	20	0.0	−0.4	0.0	−1.9	0.3	0.3	0.2
			100	1.2	1.0	1.4	−2.7	−0.6	−0.1	0.1
	.20	20	20	5.5	2.4	3.7	−9.4	−3.7	−2.5	−2.0
			100	−2.2	−3.8	−2.3	−8.2	−2.6	−2.1	−2.3
		50	20	1.8	0.8	1.3	−2.3	−0.1	0.4	0.6
			100	7.2	6.0	6.3	−2.3	−0.2	−0.2	−0.4
2	.10	20	20	2.6	0.5	1.3	−6.7	−0.9	0.1	0.2
			100	−2.1	−2.6	−2.1	−9.6	−4.1	−3.0	−2.9
		50	20	6.0	5.1	5.5	−1.9	0.2	1.3	1.4
			100	3.1	2.3	2.4	−4.6	−2.6	−2.0	−1.9
	.20	20	20	3.2	−0.1	1.2	−12.3	−6.8	−4.0	−3.8
			100	8.0	5.1	6.1	−13.1	−7.8	−4.0	−3.3
		50	20	1.6	0.4	0.9	−3.2	−1.1	−0.7	−0.7
			100	1.2	−0.1	0.1	−2.1	0.0	1.1	1.3

Intercept (in logits) determines the prevalence rates. CR0 refers to cluster robust standard errors of Liang and Zeger (1986). CR2 refers to the bias reduced linearization standard error adjustment (Bell & McCaffrey, 2002).

ICC intraclass correlation coefficient, *NG* number of groups, *GS* group size, *LRM* logistic regression model, *LPM* linear probability model, *Pois* modified Poisson, Two thousand replications per condition

- Because CR0 SEs were too low, Type I errors were higher and coverage rates were too low as well

Table 3 Coverage and type I error rates per simulation condition and model estimated

Intercept	ICC	NG	GS	Coverage rates				Type I error			
				CR0	CR2			CR0	CR2		
				LRM	LRM	LPM	Pois	LRM	LRM	LPM	Pois
1	.10	20	20	90.6	92.4	96.6	92.5	8.5	6.6	5.0	6.2
			100	92.5	93.7	95.3	94.0	9.1	7.3	5.1	7.0
		50	20	94.7	95.2	95.7	95.2	6.2	6.0	5.5	5.9
			100	94.2	94.7	95.1	94.8	6.4	5.7	5.1	5.6
	.20	20	20	91.3	92.8	94.9	93.5	7.9	6.3	4.9	5.9
			100	90.5	92.0	96.0	93.2	8.3	7.0	5.4	6.5
		50	20	93.1	93.7	94.5	94.4	5.1	4.9	4.3	4.3
			100	94.5	95.0	95.9	95.9	6.1	5.5	4.8	5.0
2	.10	20	20	92.4	93.7	95.2	94.4	7.3	5.8	4.5	5.4
			100	91.0	92.2	95.8	93.6	9.4	8.1	5.3	7.2
		50	20	93.8	94.0	95.0	94.3	5.9	5.4	4.6	5.0
			100	93.0	93.0	94.0	93.6	6.6	5.9	5.2	5.3
	.20	20	20	90.3	92.1	94.8	93.5	9.5	7.4	5.1	6.5
			100	90.4	91.5	94.9	93.4	10.4	8.8	5.6	6.9
		50	20	93.7	94.2	95.5	95.3	5.9	5.5	4.7	4.7
			100	92.6	93.2	94.8	94.5	5.4	4.9	4.1	4.1

Intercept (in logits) determines the prevalence rates. CR0 refers to cluster robust standard errors of Liang and Zeger (1986). CR2 refers to the bias reduced linearization standard error adjustment (Bell & McCaffrey, 2002).

ICC intraclass correlation coefficient, *NG* number of groups, *GS* group size, *LRM* logistic regression model, *LPM* linear probability model, *Pois* modified Poisson, Two thousand replications per condition

The question is: what would you like to interpret more?

- The logit and OR?
- The percentage point difference (ARD; average risk difference?)
- The risk ratio?

- Of course, using logits is fine and then converting to the different metrics is ok too– but if you can estimate the effect directly– why not?
- NOTE: For some, this might not be news as these methods have been around for a while– they are just not commonly used (nor combined in this manner)!
- The use of the modified Poisson is more novel though– prior studies (Zou & Donner) have recommended 50 clusters at least … but not anymore

- Can use these as supplementary analyses→ to help with communicating findings

Analyzing Group Randomized Trials (with a Few Clusters and Binary Outcomes)

Francis Huang, PhD

2021.06.04

@flhuang
https://faculty.missouri.edu/huangf
huangf@missouri.edu

Summer of R

R Workshop Series: August 16-20, 2021

This training is designed to provide the tools you need to start using the R statistical software – the basics and beyond!

Register now for one or more sessions:
http://moprevention.org/summer-of-r/

Slots are limited, so please register early for this annual event!

WHO ▶ Students, faculty, and other researchers in education, behavioral, and social sciences are all eligible to enroll.

WHAT ▶ Drs. Francis Huang, Wolfgang Wiedermann, Wes Bonifay, & Sonja Winter (University of Missouri) provide a five-part training in R.

PART I:	***Introduction to R and Basic Data Exploration*** (1 day)
PART II:	***Generalized Linear Modeling*** (1 day)
PART III:	***Multilevel Modeling*** (1 day)
PART IV:	***Psychometrics*** (1 day)
PART V:	***Introduction to R Shiny Applications*** (half day)

Missouri Prevention Science Institute
University of Missouri

WHEN ▶

PART I:	August 16, 2021
PART II:	August 17, 2021
PART III:	August 18, 2021
PART IV:	August 19, 2021
PART V:	August 20, 2021

Provided by the Missouri Prevention Science Institute
(http://moprevention.org)