Title: Bias Induced By Fitting GLMMs with Dichotomous Outcomes Using Penalized Quasi-Likelihood

Short Title: PQL Bias With Dichotomous Outcomes

Authors: Joshua Nugent, Bianca Doone, Ken Kleinman, (all: affiliation: Department of Biostatistics and Epidemiology, University of Massachusetts Amherst, Amherst, MA, USA)

Corresponding Author: Joshua Nugent, 415 Arnold House, School of Public Health and Health Sciences, University of Massachusetts, 715 North Pleasant Street, Amherst, MA 01003-9304

Telephone: 413-320-6667

jnugent@umass.edu

Grant Support: KEN - ANY OTHER REQUIRED ACKNOWLEDGMENTS?

Word Count: 2,671

Date: March 2019

Abstract: Generalized linear mixed models (GLMMs) are the most widely-used method for analyzing data from cluster-randomized trials (CRTs). Popular statistical software packages allow for different GLMM fitting algorithms, but one of these algorithms, penalized quasi-likelihood (PQL), has been shown to produce biased parameter estimates. We review the literature to assess how widely PQL may be used, and conduct a literature-informed simulation study to show the extent of the PQL bias in plausible CRT settings. We find that the algorithms employed are rarely reported in the literature, and that PQL bias is most extreme when the cluster size is small and the variability between clusters is large. Further, intraclass correlation coefficient (ICC) estimates from PQL-fitted models are also shown to vary by outcome prevalence and treatment effect. Alternatives to PQL estimation are demonstrated to be unbiased and feasible for most CRT data analysis needs.

Keywords: Cluster randomized trials, generalized linear mixed models, penalized quasi-likelihood, PQL

1 Background / Aims

Generalized linear mixed models (GLMMs) are the most commonly used correct method for analyzing data from cluster randomized trials (CRTs). GLMMs extend generalized linear models (GLMs) by including an additional random-effects term in the linear predictor. This term captures cluster variance - for example, the group-level differences between hospitals, classrooms, or individuals with repeated observations over time. In settings where interventions are applied at the cluster level, GLMMs can disaggregate treatment effects from any preexisting underlying variance between clusters. In the medical setting, CRTs with dichotomous outcomes are very common - for example, estimating the effect of a new infection control protocol on MRSA incidence, or the probability of a preterm birth for people enrolled in prenatal support groups - and GLMMs are a commonly-used tool for analysis.

The optimization problem of fitting a GLMM to data is a non-trivial task. Three common numerical methods for estimating the coefficients are *penalized quasi-likelihood* (PQL), *Gauss-Hermite quadrature* (GHQ), and *Laplace approximation*. Other methods, such as Newton quadrature, Monte Carlo integration, and Markov Chain Monte Carlo can be used as well [17], but since popular statistical software packages use PQL, GHQ, and Laplace as their standard GLMM fitting algorithms, we will focus on those in this paper. The full mathematical details of these three main methods have been elaborated in other sources [14][11], and an overview of the technical aspects of GLMMs and the algorithms is given in the Appendix, so we will focus here on the implications of using each method, rather than what's under the hood.

Penalized quasi-likelihood was popularized by Breslow and Clayton[2], though

similar methods were developed by others[16][4] around the same time. Though it is computationally efficient, especially for models with many random effects, PQL can induce bias in certain cases, in particular when the response variable distribution is far from normal[1][12][3][6], which we will investigate further in this paper. Additionally, PQL produces Wald-type test statistics, not true likelihoods, making it unsuitable for the likelihood ratio test used in nested model selection[17][11][9].

Gauss-Hermite quadrature is a more computationally demanding method, but shows no discernible bias, and can produce true likelihood statistics for model comparison. The accuracy with which it computes the model fit is a function of how many *quadrature points* it uses in estimating the model. All else being equal, the computation time is roughly proportional to $(N_q)^u$, where N_q is the number of quadrature points and u is the number of random effects at all levels of the model [13][11]. For a model with 4 random effects and 5 quadrature points, $(N_q)^u = 5^4 = 625$. Doubling the number of points to 10 changes that result to $10,000$, a factor of 16. For data sets with large numbers of random effects, this can limit the utility of GHQ.

Luckily for the data analyst, many CRTs have only one random effect, so computation time will increase linearly with the number of quadrature points, rather than as a power function. Furthermore, if many models need to be compared, using one quadrature point can give preliminary results rapidly. Then, after that model selection process, the number of quadrature points can be increased to make the final estimates as accurate as computationally possible. Empirical results suggest that 7 or fewer quadrature points often give suitably accurate estimates[10].

The Laplace method is equivalent to GHQ with a single quadrature point[7],

and as such it produces true likelihoods for model comparison. While its accuracy is typically lower than GHQ with $N_q \geq 2$, under certain circumstances it performs quite well[7].

For reasons of computational efficiency, PQL was a useful method for fitting GLMMs when it was developed, but with the advent of modern computers, more accurate methods such as GHQ have become an attractive alternative. For CRTs with binary outcomes, where the bias in PQL is the most extreme[9][5], and where the presence of only one random effect is typical, using GHQ is the best option: fast enough, and, more importantly, unbiased.

Most modern statistical software packages have functions to fit GLMMs with dichotomous outcomes, such as PROC GLIMMIX in SAS, meglm in Stata, and glmer (from the lme4 package, as well as others) in R. However, the default fitting algorithm in each of those functions varies. In SAS PROC GLIMMIX, the default is PQL, with Laplace or GHQ available if specified. In R, the glmer function default is Laplace, with GHQ available if specified; PQL is only available in R via the glmmPQL function in the MASS package. In Stata, meglm defaults to GHQ with 7 quadrature points, with Laplace available if specified.

Given that many data analysts may be unfamiliar with the fitting options, a function's default settings are influential in the final results. Below, we investigate how often functions and algorithms are reported in the literature and use simulations to describe the bias induced by PQL in a literature-informed, plausible CRT scenario.

2 Methods

We started by conducting a literature review among recent CRTs with dichotomous outcomes to determine a) common values for cluster size and number of clusters and b) what software, functions, and fitting algorithms were used to analyze the data, if reported. The review, searching for the phrase "cluster randomized trial" in the title or abstract of the article, spanned two databases over two timeframes. First, we searched the The New England Journal of Medicine, The British Medical Journal, The Journal of the American Medical Association, and The Lancet from January 1, 2014 through August 31, 2018, using the Web of Science database. Second, we did a broader search of all articles in the PubMed database published between March 1st, 2018 and August 31, 2018.

Having identified candidate articles, we filtered to completed CRTs with dichotomous outcomes. The mean number of observations per cluster and number of clusters for each study was recorded, as well as the software and functions/algorithms the authors used, if available.

The second phase of our work was a simulation study to investigate the bias of different GLMM fitting algorithms for dichotomous outcomes. To maximize the utility of the results, our simulations used a range of plausible cluster counts and cluster sizes drawn from the literature review.

Our data-generating mechanism for the simulations was a simple logistic-link GLMM with one fixed intercept, one treatment effect, and one random intercept, defined as:

$$\text{logit}[\pi(x_{ij}|u_j)] = \beta_0 + \beta_1 x_{ij} + u_j \tag{1}$$

with x_{ij} an indicator for treatment (1) or control (0) arm of the study for observation i in cluster j; $\pi(x_{ij}|u_j)$ the probability of the outcome for individual i in cluster j; e^{β_0} the baseline odds of the outcome across all clusters; e^{u_j} the odds ratio compared to baseline for the outcome, specific to observations in cluster j, with assumed distribution $u_j \sim N(0, \sigma_0^2)$; and β_1, our parameter of interest, the log odds ratio due to the treatment.

From that model, populations were generated with the following parameter values, informed by the literature review (see Table 2):

- Number of clusters $n \in \{10, 50, 100\}$
- Number of observations per cluster $p \in \{25, 100\}$
- β_0 values corresponding to a baseline prevalence of .01, .02, .03, .05, .1, and .2
- β_1 values corresponding to a treatment effect odds ratio of .5, .75, .9, 1.1, 1.33, 1.5, and 2
- σ_0^2 values of 1 (high cluster variability) and .1 (low cluster variability).

Using 3000 simulated populations for each combination of parameters, logistic-link GLMMs were fit via PQL, GHQ, and Laplace using SAS software version 9.4 (SAS Institute, Cary, NC). The distribution of $\hat{\beta}_1$ estimates from each method was compared to the true value from the data-generating mechanism and absolute bias was measured as the difference between the two. The standard errors of estimates and the estimated cluster variance ($\hat{\sigma}_0^2$) were also collected from the fitted models. Finally, the simulated populations were re-fit under the assumption that the outcome variable was normally

	Software cited (specific procedures or functions)				
	R	SAS	SPSS	Stata	Unreported
Top4 2014 - 2018 ($n = 42$)	11 (2)	13 (1)	2 (0)	18 (2)	4 (4)
PubMed Q2 2018 ($n = 43$)	3 (0)	11 (2)	8 (0)	14 (2)	10 (0)
Total ($n = 85$)	14 (2)	24 (3)	10 (0)	32 (4)	14 (0)

Table 1: Cited software used for data analysis. Numbers in parentheses indicate number of references to specific functions or fitting options within the software package. Some articles used multiple software packages, so the totals differ from n. While most articles report the software used, very few of them specifically reference the function or fitting algorithm.

	Median number of clusters (middle 50%)	Median observations per cluster (middle 50%)
Top4 2014 - 2018 ($n = 42$)	46 (24 - 116)	110 (40 - 487)
PubMed Q2 2018 ($n = 43$)	32 (15 - 70)	44 (14 - 205)
Total ($n = 85$)	40 (16 - 90)	66 (19 - 300)

Table 2: Common values for p and n from the literature review.

distributed, without the logit link, and the intraclass correlation coefficient (ICC) estimated from those models was collected.

3 Results

Results of the literature review are shown in Table 1. Of the entire group of 85 articles, only 9 identified the procedure (meglm or GLIMMIX, for example) used to fit their data. Among the 24 articles that identified SAS as one of the software packages, only 3 specified which SAS procedure was used and only one identified the model fitting algorithm.

Parameters for the simulations were chosen after observing the common values in Table 2. The number of clusters was typically below 100, as was the number of observations per cluster, though the latter showed significantly more dispersion.

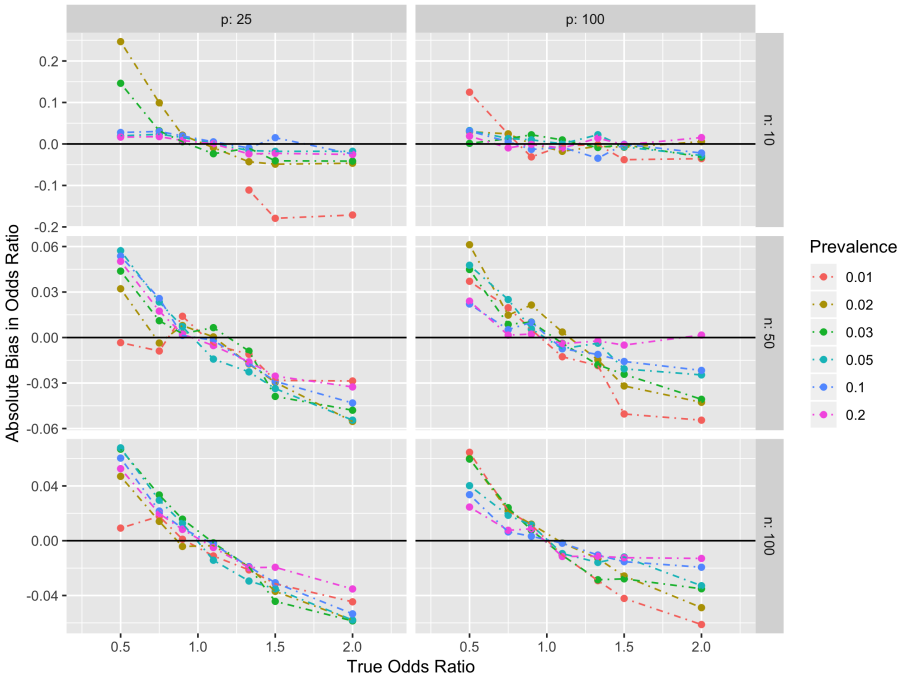

Figure 1: Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in PQL estimation, $\sigma_0^2 = 1$, cluster size p, number of clusters n. Simulation runs with less than 80% convergence omitted. Note larger scale in the first row.

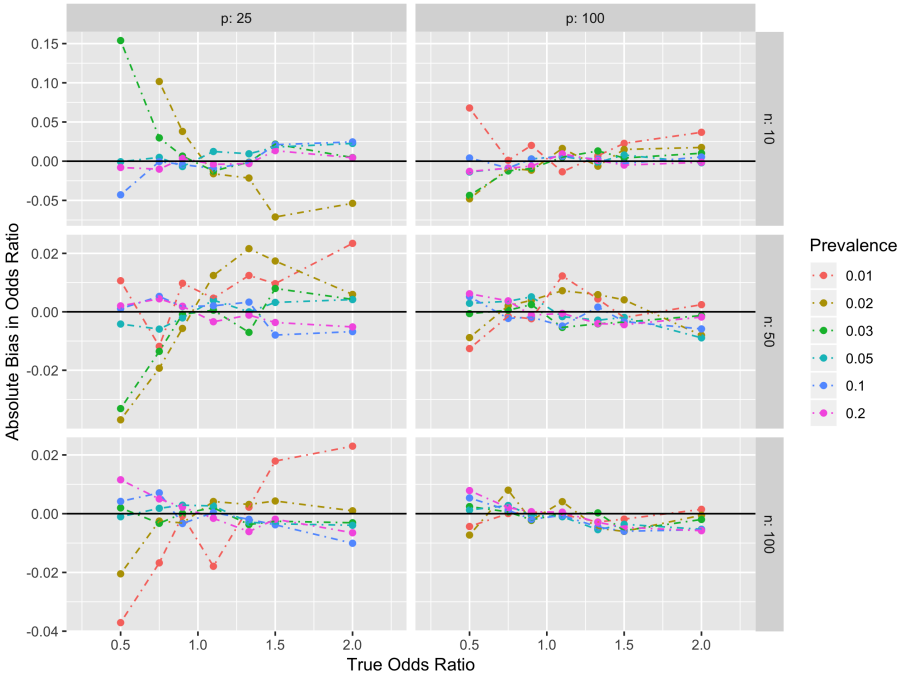

Figure 2: Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in PQL estimation, $\sigma_0^2 = 0.1$, cluster size p, number of clusters n. Simulation runs with less than 80% convergence omitted.

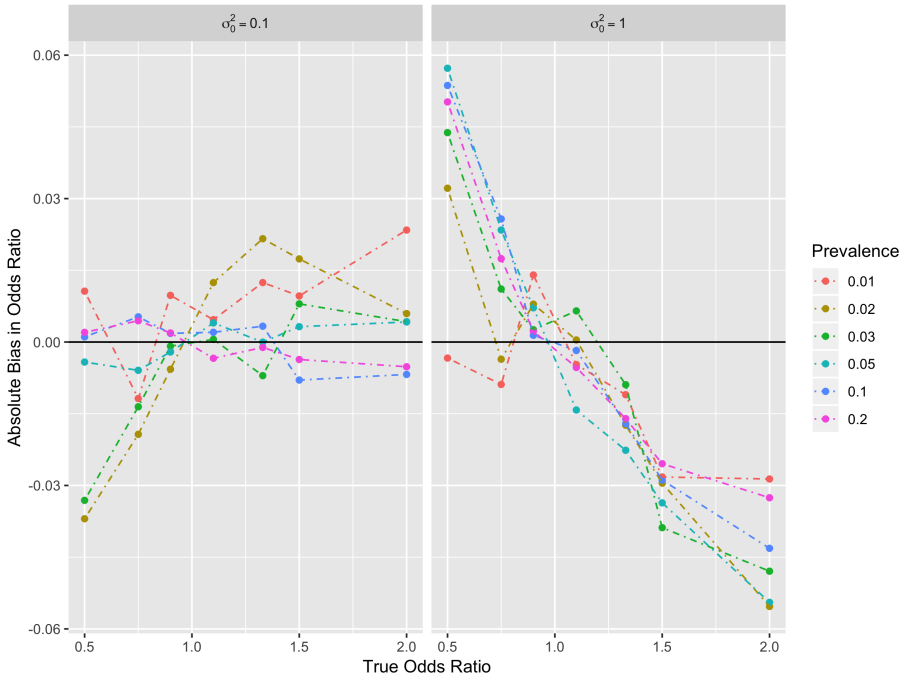

Figure 3: Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in PQL estimation for different σ_0^2, $p = 25$, $n = 50$.

The results of the bias investigation in PQL estimation are shown in Figures 1-3. Put simply, the results show bias towards the null: As the true odds ratio rises above 1, there is a negative bias, meaning the mean estimated odds ratio is closer to 1 than it should be. When the odds ratio is less than one, conversely, there a positive bias towards 1. Further, while the data are noisy, the bias is more pronounced for smaller cluster sizes and when the outcome's baseline prevalence is lowest. The bias occurs, to a differing extent, across all values of prevalence, treatment effect, cluster size, and number of clusters. Figure 3 demonstrates that the effect of high between-cluster variability is more pronounced bias.

Further, the standard errors of the estimate for β_1 are quite large in these simulations. For example, in the setting where $p = 25$, $n = 100$, and $\sigma_0^2 = 1$, it was not uncommon to have standard errors ten times larger than the bias. This means that even a small amount of bias on a large estimated value could pull a finding above the $p < .05$ threshold.

The fitted models using GHQ (a representative example is given in Figure 4) and Laplace approximation (Figure 5) did not show a clear bias; in each, PQL is shown for reference. In a small number of simulation runs, particularly with a small number of clusters and rare outcomes, the GHQ and Laplace algorithms became unstable, leading to estimates that tended toward infinity, making the mean bias comparison unhelpful - this can be seen in the outlier in Figure 5 when the prevalence is .01 and the treatment effect odds ratio is 2. On the other hand, when the number of clusters is small, the PQL algorithm was less likely to converge, implying that limited data is a challenge for all of the algorithms.

Given the bias it creates, why use PQL? Our simulations confirmed that the

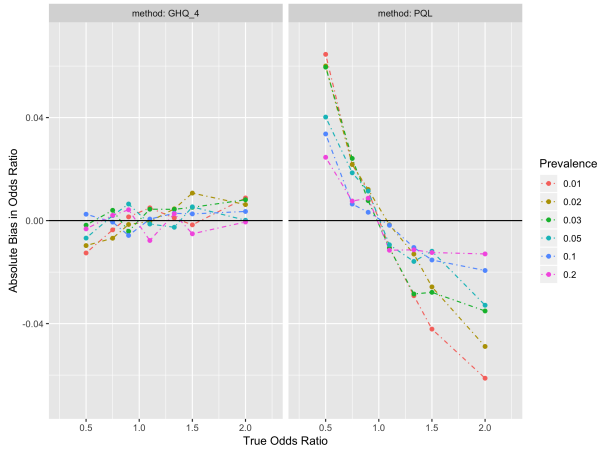

Figure 4: Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in GHQ (4 quadrature points) and PQL, $\sigma_0^2 = 1$, $p = 100$, $n = 100$.

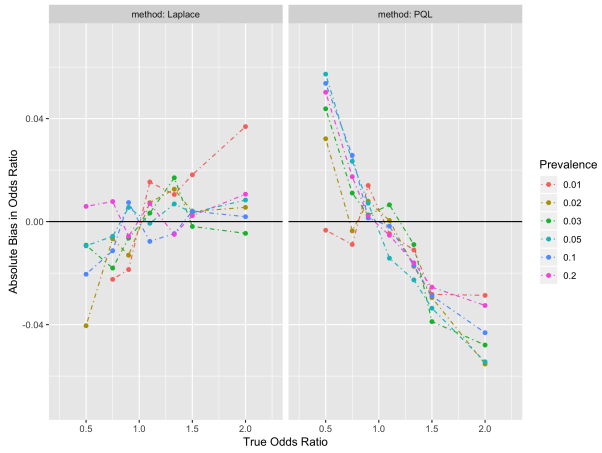

Figure 5: Odds ratio bias ($\exp\{\bar{\hat{\beta}}_1 - \beta_1\} - 1$) in Laplace and PQL, $\sigma_0^2 = 1$, $p = 25$, $n = 50$. Note that the lack on convergence in some of the Laplace models leads to the outlier at odds ratio 2.

main advantage of using PQL over GHQ is speed; as expected, SAS's implementation of PQL wildly outperforms the other methods. The mean CPU time for SAS to fit a single large data set (500 clusters, 500 observations per cluster) on a modern laptop is 3.7 seconds for PQL, 11 seconds for Laplace, 18.4 seconds for GHQ with $N_q = 4$, and 27.1 seconds for GHQ with $N_q = 10$. Results are comparable for R's lme4 package, though the glmmPQL method from the MASS package, with a runtime of at 23.3 seconds, is significantly less efficient than in the SAS implementation. However, even for this large data set, none of the runtimes above are prohibitively long. In a situation where many models need to be tested, the Laplace method could be used to compare models, and then for the final analysis, a more accurate GHQ fit could be made with a large number of quadrature points.

For primarily historical reasons, analysts may be interested in the intraclass correlation coefficient (ICC), though it is not an actual parameter of the model. In normally distributed data, the ICC measures the proportion of total variance that is explained by the variance between groups. However, in the case of a non-normally distributed outcome variable, there has been considerable discussion about how to appropriately characterize and calculate the ICC[15][8]. The PQL-fitted models in our simulation generated two possible ICC estimates: first, a version that assumed a random intercept logistic model, implying an ICC of $\frac{\sigma_0^2}{\sigma_0^2+\frac{\pi^2}{3}}$, simply based on the estimated between-cluster variance, and second, by fitting a linear mixed model that assumed the outcome variable was normally distributed, leading to the typical ANOVA-based calculation $\frac{\sigma_0^2}{\sigma_0^2+\sigma_\epsilon^2}$, where σ_ϵ^2 represents the variance of the residuals. The results are shown in Figure 6 for the two values of σ_0^2 examined in out simulations. In both cases, the estimated ICC varies significantly by prevalence, treatment effect, and model class. Results when fitting the models

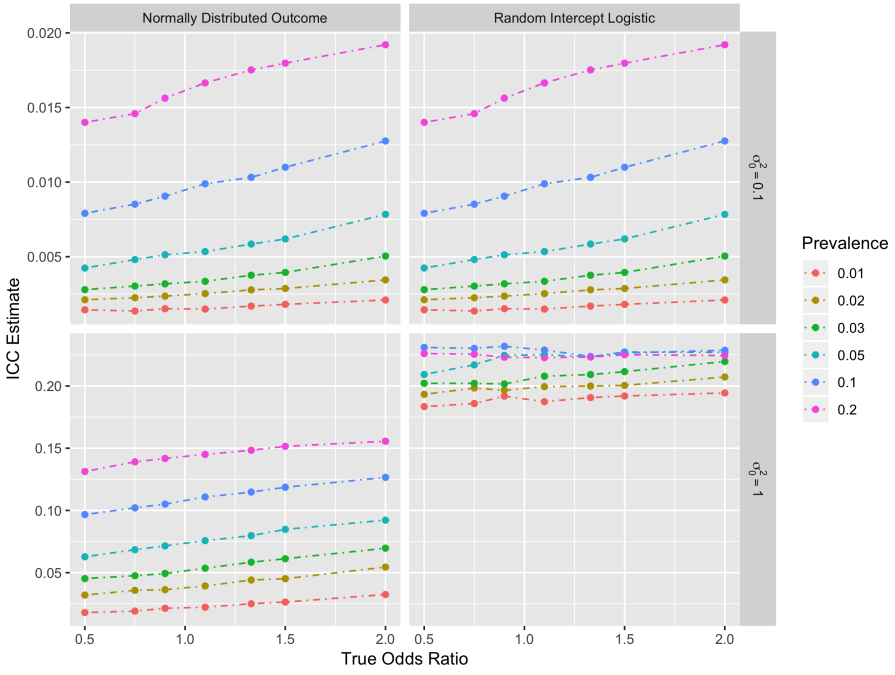

Figure 6: ICC estimates from models fit via PQL, $p = 100$, $n = 50$.

with GHQ showed a similar pattern with the assumption of a normal distribution for the outcome, but more consistency when using the $\frac{\sigma_0^2}{\sigma_0^2 + \frac{\pi^2}{3}}$ formulation. This implies that fitting with GHQ gives more consistent estimates of $\hat{\sigma}_0^2$ than when fitting with PQL. [KEN - ADD SOME COMMENTARY HERE??]

4 Conclusions

For the vast majority of cluster randomized trials with dichotomous outcomes, GHQ is superior to PQL when fitting models. To fit a single data set with a small number of random effects, and given than most CRTs have only one

random effect, the speed of GHQ with 4-10 quadrature points is adequate and it produces no detectable bias. For data analysts who are experimenting with different nested models, using Laplace approximation during the model-selection stage can save time, and GHQ can be utilized for the fitting once the final model has been chosen. Laplace is also preferable to PQL in the model-selection process because PQL only provides quasi-likelihood, and hence it is not suited to nested model comparison with the likelihood ratio test.

The bias towards the null generated by PQL is more pronounced when clusters are small, between-cluster variance is high, and baseline incidence of an outcome is low. Given our simulations, we suspect that existing studies may have suffered this bias, though it is hard to be sure given that fitting methods are rarely reported in the literature. Statisticians should report methods/functions and the algorithm options in more detail, and take care when selecting procedures for fitting GLMMs, particularly in SAS, where PQL is the default option. Finally, ICC estimates generated by these algorithms may vary substantially by the method used to calculate them and by the baseline prevalence of the outcome, and should be approached with a degree of skepticism.

A GLMM Fitting

Mathematically, a GLMM can be modeled as

$$g(\mu_{it}) = \mathbf{x}_{it}^T\boldsymbol{\beta} + \mathbf{z}_{it}^T\mathbf{u}_i \tag{2}$$

with i a cluster indicator, t an observation indicator within cluster i, g the

GLM link function, $\boldsymbol{\beta}$ a vector of coefficients for covariate values $\mathbf{x}_{it}$, and $\mathbf{z}_{it}^T$ a vector of coefficients for random effects $\mathbf{u}_i$, assumed to be distributed as multivariate normal with mean 0 and covariance matrix $\boldsymbol{\Sigma}$. When the outcomes are dichotomous, the link function g is typically the logit, and the mean μ_{it} is the probability of the outcome given the covariate values and cluster membership.

To fit a GLMM with a vector $\mathbf{x}_{it}$ and corresponding outcome vector $\mathbf{y}$, it is necessary to integrate the random effects $\mathbf{u}_i$ out of the likelihood function[12]. That likelihood function, the probability mass function of y as a function of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$[1], is, in general,

$$\ell(\boldsymbol{\beta}, \boldsymbol{\Sigma}; \mathbf{y}) = f(\mathbf{y}; \boldsymbol{\beta}, \boldsymbol{\Sigma}) = \int f(\mathbf{y}|\mathbf{u}; \boldsymbol{\beta}) f(\mathbf{u}; \boldsymbol{\Sigma}) d\mathbf{u}. \tag{3}$$

For many link functions of interest, including the logit link function for dichotomous outcomes considered in this paper and other situations where the response variable is discrete, the integral above does not have a closed-form solution, in part because it involves integrating the product of discrete and continuous densities[9]. Numerical methods are required to approximate the integral in these circumstances.

PQL iteratively fits a linear mixed model[6] to the data, essentially approximating the discrete density using a Gaussian density[9]. Further details of PQL have been discussed above.

Gauss-Hermite quadrature approximates the integral of a function $f(\cdot)$ multiplied by a normal density function; note that it is very similar to the likelihood function presented earlier where $f(\mathbf{u}; \boldsymbol{\Sigma})$ was a multivariate normal

and $f(\mathbf{y}|\mathbf{u};\boldsymbol{\beta})$ the conditional likelihood. For univariate cases,

$$\int_{-\infty}^{\infty} f(u)exp(-u^2)du \approx \sum_{k=1}^{q} c_k f(s_k) \tag{4}$$

where c_k are weights, sometimes from a table, and s_k are the each of the quadrature points used to approximate the normal density. More quadrature points results in a more accurate approximation of the integral, but is more computationally intensive, though various GHQ subvariants have been developed that increase efficiency and reduce the number of quadrature points needed[11]. With GHQ, inversion of the Fisher information matrix can provide standard errors for the maximum likelihood estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$.

The Laplace method approximates the likelihood using a second-order Taylor expansion [10] and is is equivalent to GHQ with a single quadrature point[7]. Simulation studies have found Laplace approximations to exhibit mild bias in coefficient estimates, and significant bias in estimation of the variance components[11].

References

[1] A. Agresti. *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2013.

[2] N. E. Breslow and D. G. Clayton. "Approximate Inference in Generalized Linear Mixed Models". In: *Journal of the American Statistical Association* 88.421 (Mar. 1993), pp. 9–25.

[3] Norman E. Breslow and Xihong Lin. "Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion". In: *Biometrika* 82.1 (1995), pp. 81–91.

[4] B. Engel and A. Keen. "A simple approach for the analysis of generalizea linear mixed models". In: *Statistica Neerlandica* 48.1 (Mar. 1994), pp. 1–22.

[5] Xihong Lin. "Estimation using penalized quasilikelihood and quasi-pseudo-likelihood in Poisson mixed models". In: *Lifetime Data Analysis* 13.4 (Dec. 2007), pp. 533–544.

[6] Xihong Lin and Norman E. Breslow. "Bias Correction in Generalized Linear Mixed Models With Multiple Components of Dispersion". In: *Journal of the American Statistical Association* 91.435 (1996), pp. 1007–1016.

[7] Qing Liu and Donald A. Pierce. "A Note on Gauss-Hermite Quadrature". In: *Biometrika* 81.3 (1994), pp. 624–629.

[8] Nakagawa Shinichi, Johnson Paul C. D., and Schielzeth Holger. "The coefficient of determination R2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded". In: *Journal of The Royal Society Interface* 14.134 (Sept. 2017), p. 20170213.

[9] Edmond SW Ng et al. "Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood". In: *Statistical Modelling: An International Journal* 6.1 (Apr. 2006), pp. 23–42.

[10] José C. Pinheiro and Douglas M. Bates. "Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model". In: *Journal of Computational and Graphical Statistics* 4.1 (1995), pp. 12–35.

[11] José C. Pinheiro and Edward C. Chao. "Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models". In: *Journal of Computational and Graphical Statistics* 15.1 (2006), pp. 58–81.

[12] German Rodriguez and Noreen Goldman. "An Assessment of Estimation Procedures for Multilevel Models with Binary Responses". In: *Journal of the Royal Statistical Society. Series A (Statistics in Society)* 158.1 (1995), p. 73.

[13] StataCorp. *Stata 15 Base Reference Manual*. 2017.

[14] Russ Wolfinger and Michael O'connell. "Generalized linear mixed models a pseudo-likelihood approach". In: *Journal of Statistical Computation and Simulation* 48.3-4 (Dec. 1993), pp. 233–243.

[15] Sheng Wu, Catherine M. Crespi, and Weng Kee Wong. "Comparison of methods for estimating the intraclass correlation coefficient for binary responses in cancer prevention cluster randomized trials". In: *Contemporary Clinical Trials* 33.5 (Sept. 2012), pp. 869–880.

[16] Scott L. Zeger, Kung-Yee Liang, and Paul S. Albert. "Models for Longitudinal Data: A Generalized Estimating Equation Approach". In: *Biometrics* 44.4 (1988), pp. 1049–1060.

[17] Hui Zhang et al. “On Fitting Generalized Linear Mixed-effects Models for Binary Responses using Different Statistical Packages”. In: *Statistics in medicine* 30.20 (Sept. 2011), pp. 2562–2572.