Statistica Sinica **15**(2005), 1015-1032

QUANTIFYING PQL BIAS IN ESTIMATING CLUSTER-LEVEL COVARIATE EFFECTS IN GENERALIZED LINEAR MIXED MODELS FOR GROUP-RANDOMIZED TRIALS

Scarlett L. Bellamy[1], Yi Li[2], Xihong Lin[2] and Louise M. Ryan[2]

[1] *University of Pennsylvania and* [2] *Harvard School of Public Health*

Abstract: We derive the asymptotic bias and variance of the penalized quasilikelihood (PQL) estimator of the cluster-level covariate effect in generalized linear mixed models for group-randomized trials where the number of clusters n is small and the cluster size m is large. We show that the asymptotic bias is of order $\mathrm{O}_p(1/m)$ and the asymptotic variance is of order $O_p(1/n) + O_p\{1/(nm)\}$. The practical implication of our results is that the PQL method works well in settings involving small numbers of large clusters which are typical in grouped randomized trials. We illustrate the results using simulation studies.

Key words and phrases: Asymptotic bias, asymptotic variance, generalized linear mixed models, Penalized quasilikelihood.

1. Introduction

Group-randomized trials are becoming increasingly popular as a tool for evaluating the efficacy of behavioral health interventions, particularly those involving providers such as group therapists or primary care physicians, as well as other natural groupings like schools or neighborhood community centers. These trials are distinguished by randomizing intact units/groups of individuals (e.g., therapy groups, primary care practices, classrooms or neighborhoods) to various conditions (e.g., cognitive behavior therapy, depression specialist, method of teaching, or community-level classes). Such trials are also known as cluster-randomized trials in other literature (see, for example, Donner and Klar (2000)). Group-randomized trials are further distinguished by having a small number of clusters with a large number of observational units within each cluster. The goal of a typical behavioral intervention study is to compare the rates of various adverse health outcomes or risky lifestyle behaviors between intervention arms. The general goal is to make cluster-level inference of the efficacy of the intervention. These types of studies are often interested in modeling outcomes (e.g., number of risky 'events' in the past month, disease status yes/no, etc.) as a function of confounding and/or mediating variables, while adjusting for the potential correlation of responses from observations within the same cluster. From an analytic perspective there are several challenges to appropriately modeling such data, since it is important to account for the clustered nature of the responses, especially in settings when there are a few large clusters/groups and the standard asymptotic results are questionable.

Two popular approaches for modeling clustered data are via population-averaged (PA) models fitted by generalized estimating equations (GEEs), and subject-specific models fitted using generalized linear mixed models (GLMMs). The GEE approach is attractive in that it provides unbiased estimates while properly adjusting for possible misspecification of the correlation of responses from the same cluster, but these properties may be doubtful in settings involving small numbers of large clusters (see, for example, Bellamy et al. (2000) and Mancl and Leroux (1996)). Generalized linear mixed models account for the within-cluster correlation by introducing random effects in model specification. Inference in GLMMs is challenged by the required numerical integration in the likelihood function. In addition to numerical quadrature and Markov Chain Monte Carlo techniques for integration, the penalized quasilikelihood (PQL) approach (Breslow and Clayton (1993) and Green (1987)) provides an alternative easy approach to estimating covariate effects in GLMMs. However, various authors have noted that covariate effects based on the PQL may be asymptotically biased (see, for example, Neuhaus and Segal (1997), Breslow and Lin (1995) and Lin and Breslow (1996)). Breslow and Lin (1995) and Lin and Breslow (1996) derived asymptotic bias expressions for PQL regression coefficient and variance component estimators in settings where there are a large number of small clusters. Their results are not applicable to group-randomized trial settings where there are a small number of large clusters.

There have been related asymptotic explorations of the PQL estimators in the context of small area estimation. Jiang and Lahiri (2001) considered the asymptotic properties of estimating random effects in settings where the number of independent observations within clusters gets large, but here we are interested in the asymptotic bias associated with estimating fixed-effects. Jiang (1999) examined the asymptotic behavior of PQL-type estimators when both the numbers of clusters and observations within clusters increase. This result is not directly applicable to group-randomized trial settings where there is often a small number of clusters.

In previous work, we explored the bias associated with PQL methods in group-randomized trial settings via simulation studies and observed good empirical results in estimating covariate effects, especially for large cluster sizes (Bellamy et al. (2000)). Ten Have and Localio (1999) reported simulation results showing that PQL performed well with small numbers of large clusters, whereas numerical integration did poorly in settings where there are few clusters. Vonesh et al (2002) investigated the asymptotic distribution of the PQL estimator when the number of both the number of clusters and the cluster size go to infinity and showed the PQL estimator performed well in such settings.

In this paper we explore more rigorously those empirical findings by deriving expressions for the asymptotic bias and variance associated with estimating cluster-level covariate effects in a GLMM, using the PQL approach in settings which have a small number of large clusters such as group-randomized trials. The paper is organized as follows. In Section 2 we present the GLMM and the PQL method. In Section 3 we derive an expression for the asymptotic bias in estimating the cluster-level covariate effects via the PQL and the asymptotic variance of the PQL estimator. We consider in Section 4 the forms of the bias and variance expressions for common special cases. We present in Section 5 the results from a simulation study and close with a discussion in Section 6.

2. The generalized linear mixed model

2.1. Model formulation

Consider data from a study involving n clusters. For simplicity of presentation, we assume there are an equal number of m observations per cluster. Generalization of our results to unequal cluster size settings are straightforward and are presented at the end of Section 3. In grouped-randomized trials, n is often small and m is often large. For the jth observation ($j = 1, \ldots, m$) in cluster i ($i = 1, \cdots, n$), we observe a response y_{ij}, and a $p \times 1$ vector of cluster-specific covariates $\boldsymbol{x}_i$. Generalized linear mixed models (GLMMs) provide a broad class of random effects models to model such clustered data. Conditional on the cluster-specific unobserved random effects b_i, the outcomes y_{ij} are assumed to be independent and follow the exponential family

$$\ell_i(\boldsymbol{\beta}; b_i) = \sum_{j=1}^{m} \frac{a_{ij}}{\phi} \left\{ y_{ij}\eta_{ij} - c(\eta_{ij}) \right\} + k(y_{ij}, \phi), \tag{1}$$

where η_{ij} is a canonical parameter, a_{ij} is a known weight, ϕ is a scale parameter, and $c(\cdot)$ and $k(\cdot)$ are some known functions. The conditional mean of y_{ij} is $\mu_{ij} = E(y_{ij}|b_i) = c'(\eta_{ij})$ and is related to the covariate vector $\boldsymbol{x}_i$ and the random effect b_i through a generalized linear model (Breslow and Lin (1995))

$$g(\mu_{ij}) = \boldsymbol{x}_i\boldsymbol{\beta} + b_i, \tag{2}$$

where $g(\cdot)$ is a monotone link function and $b_i \sim N(0, \theta)$. We restrict attention in this paper to canonical link functions which satisfy $g(\mu_{ij}) = \eta_{ij}$, $\mathrm{var}(y_{ij}|b_i) =$

$\phi a_{ij}^{-1} v(\mu_{ij})$ and $g'(\mu_{ij}) = 1/v(\mu_{ij})$ (McCullagh and Nelder (1989)). The observed data likelihood is hence

$$L(\boldsymbol{\beta}, \theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\theta}} \int exp\left\{\ell_i(\boldsymbol{\beta}; b_i) - b_i^2/2\theta\right\} db_i. \tag{3}$$

2.2. PQL estimation of GLMM

Because the integrated likelihood (3) does not usually have a closed form expression, Breslow and Clayton (1993) proposed estimation of the regression coefficients $\boldsymbol{\beta}$ using the penalized quasilikelihood (PQL) method by applying the Laplace approximation to the integrated loglikelihood function. The PQL likelihood can be written as (Breslow and Lin (1995))

$$\ell_p(\boldsymbol{\beta}, \theta) = \sum_{i=1}^{n} \Big(\tilde{\ell}_i - \frac{\tilde{b}_i^2}{2\theta}\Big), \tag{4}$$

where $\tilde{b}_i$ satisfies $\tilde{b}_i = \theta\, \partial\ell_i(\boldsymbol{\beta}, b_i)/\partial b_i|_{b_i=\tilde{b}_i}$ and

$$\tilde{\ell}_i = \ell_i(\boldsymbol{\beta}, \tilde{b}_i) = \sum_{j=1}^{m} \frac{a_{ij}}{\phi}\{y_{ij}\ \tilde{\eta}_{ij} - c(\tilde{\eta}_{ij})\} + k(y_{ij}, \phi),$$

where $\tilde{\eta}_{ij} = \boldsymbol{x}_{ij}^T\boldsymbol{\beta} + \tilde{b}_i$.

In this paper, we assume θ is known and study the asymptotic bias and variance of the PQL estimator of the regression coefficients $\widehat{\boldsymbol{\beta}}$ in grouped randomized trial settings where the number of clusters n is small and fixed and the cluster size m is large and goes to infinity. We derive the asymptotic bias for the general GLMM (2) and then consider various special cases, including a random effects logistic regression model.

3. Asymptotic Bias and Variance of Estimated Cluster-Level PQL Covariate Effects

The PQL estimator of $\boldsymbol{\beta}$ is the solution of the estimating equations obtained from maximizing (4) with respect to $\boldsymbol{\beta}$ and simultaneously computing $\tilde{b}_i$ $(i = 1, \ldots, n)$ from

$$\sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{x}_i \frac{a_{ij}}{\phi}\{y_{ij} - \mu(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}} + \tilde{b}_i)\} = 0, \tag{5}$$

$$\sum_{j=1}^{m} \frac{a_{ij}}{\phi}\{y_{ij} - \mu(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}} + \tilde{b}_i)\} = \frac{\tilde{b}_i}{\theta}, \quad (i = 1, \ldots, n), \tag{6}$$

where $\mu(\cdot) = g^{-1}(\cdot)$.

Our goal is to derive the asymptotic bias of the PQL estimator when n is small and m goes to infinity. We proceed by deriving an asymptotic expansion of the PQL estimator $\widehat{\boldsymbol{\beta}}$, the solution to the estimating equations in (5) and (6), about its true value $\boldsymbol{\beta}$ for fixed n and large m. Since the second component of the estimating equations (6) involves solving n cluster-specific equations, standard M-estimation theory does not apply (see, for example, Van der Vaart (1998) or Akritas (1991)).

We first pre-multiply each of the n components of (6) by $\boldsymbol{x}_i$ then sum over the index i and apply (5). This gives the constraint

$$\sum_{i=1}^{n} \boldsymbol{x}_i \tilde{b}_i = 0, \tag{7}$$

where $\tilde{b}_i$ is the value of b_i that satisfies both (5) and (6), simultaneously.

Equation (6) can be written as

$$\bar{y}_{i.} - \mu(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} + \tilde{b}_i) = \frac{\tilde{b}_i \phi}{m \bar{a}_{i.} \theta}, \tag{8}$$

where $\bar{y}_{i.} = \sum_{j=1}^{m} a_{ij} y_{ij} / \sum_{j=1}^{m} a_{ij}$, the cluster-specific weighted average of the response variable, $\bar{a}_{i.} = \sum_{j=1}^{m} a_{ij}/m$. It follows from (6) that

$$\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} + \tilde{b}_i = g\Big(\bar{y}_{i.} - \frac{\tilde{b}_i \phi}{m \bar{a}_{i.} \theta}\Big). \tag{9}$$

Pre-multiplying this expression by $\boldsymbol{x}_i$, summing over the index i, and recalling the constraint (7), it follows that

$$\widehat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{n} \boldsymbol{x}_i \boldsymbol{x}_i^T\Big)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i g\Big(\bar{y}_{i.} - \frac{\tilde{b}_i \phi}{m \bar{a}_{i.} \theta}\Big). \tag{10}$$

Note that $\tilde{b}_i = \tilde{b}_i(\bar{y}_{i.}, \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}, \phi/m\bar{a}_{i.})$.

We derive the asymptotic bias of $\widehat{\boldsymbol{\beta}}$ using (10), assuming n is fixed and small while m goes to infinity. Our asymptotic calculations proceed by performing a series of Taylor expansions of $\bar{y}_{i.}$ about μ_i, $\tilde{b}_i$ about b_i, and $\widehat{\boldsymbol{\beta}}$ about $\boldsymbol{\beta}$. A detailed proof of our result is in the Appendix.

Proposition 1. *Assume the number of clusters n is fixed and small and that the cluster size m goes to infinity. Suppose the variance component θ is known. Then, the asymptotic bias and variance of the PQL estimators of the regression* *coefficients* $\widehat{\boldsymbol{\beta}}$ *are*

$$\mathrm{E}(\widehat{\boldsymbol{\beta}})-\boldsymbol{\beta}=\frac{\phi}{m}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\sum_{i=1}^{n}\frac{1}{\bar{a}_{i.}}\boldsymbol{x}_i\boldsymbol{A}_i+\mathrm{O}_p\Big(\frac{1}{m^2}\Big), \tag{11}$$

$$\mathrm{cov}\Big(\widehat{\boldsymbol{\beta}}\Big)=\theta\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}+\frac{\phi}{m}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\boldsymbol{B}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}+\mathrm{O}_p\Big(\frac{1}{nm^2}\Big), \tag{12}$$

where

$$\boldsymbol{A}_i=E_{b_i}\Big[\frac{1}{2}v(\mu_i)g''(\mu_i)+\frac{1}{\theta}\Big\{\boldsymbol{x}_i^T\Big(\sum_{i'=1}^{n}\boldsymbol{x}_{i'}\boldsymbol{x}_{i'}^T\Big)^{-1}\boldsymbol{x}_i-1\Big\}g'(\mu_i)b_i\Big],$$

$$\boldsymbol{B}=\sum_{i=1}^{n}\frac{1}{\bar{a}_{i.}}\boldsymbol{x}_i\boldsymbol{x}_i^T\left\{E_{b_i}\left[v(\mu_i)\left\{g'(\mu_i)\right\}^2\right]-\frac{2}{\theta}E_{b_i}\left\{g'(\mu_i)b_i^2\right\}\right\}$$
$$+\frac{2}{\theta}E_{b_i}\Big\{\Big(\sum_{i=1}^{n}\boldsymbol{x}_ib_i\Big)\Big(\sum_{i=1}^{n}\frac{1}{\bar{a}_{i.}}\boldsymbol{x}_i\boldsymbol{x}_i^Tg'(\mu_i)\Big)\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i^Tb_i\Big)\Big\}.$$

This result indicates that PQL works well for estimating the cluster level covariate effect in group-randomized trials. It differs from that of Breslow and Lin (1995), who found the PQL estimator seriously biased for clustered binary data in conventional longitudinal/clustered data settings with a large number of clusters of small size.

Proposition 1 further shows that the variance of the PQL estimator of the cluster-level covariate effect, such as an intervention effect, is of order $O_p(1/n)+O_p\{1/(nm)\}$, which is of the same order as the maximum likelihood estimator in this small n, large m situation. This suggests that in group-randomized trials when n is fixed to be small, as expected, one often needs to have large cluster size m to achieve sufficient power to detect an intervention effect. Our simulation results in Section 4 further show that the finite sample variance and mean square error of the PQL estimator are smaller than that of the MLE. In view of the computational simplicity of PQL, this is encouraging for the use of PQL in analyzing data from grouped randomized trials.

Extension of Proposition 1 to unbalanced designs where the cluster sizes vary from cluster to cluster is straightforward. Calculations show that one simply needs to remove m and replace $\bar{a}_{i.}$ by to $\bar{a}_{i.}m_i$ in (11) and (12), and replace $O_p(m^{-2})$ with $O_p(\min(m_i)^{-2})$ and $O_p((nm^2)^{-1})$ in (12) with $O_p[\{n\min(m_i)^2\}^{-1}]$.

4. Special Cases

We examine the general bias and variance expressions in two common clustered data settings. We first take an identity link function, where expressions for the bias and variance associated with estimated cluster-level covariate effects have closed form solutions, and show that the same bias expressions can be obtained directly from our general bias and variance expressions. Second, we consider the clustered data setting with a binary outcome, where one is interested in the efficacy of, say, an intervention from a group-randomized trial.

4.1. Identity link function

Consider a random effects model with an identity link function $g(\mu) = \mu$. Assume $a_{ij} = 1$. The bias and variance expression for estimating cluster-level covariates can be obtained from (11) and (12), respectively, noting in this case that $g'(\mu_i) = 1$ and $g''(\mu_i) = 0$. Thus,

$$\begin{aligned}\mathrm{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= O_p(1/m^2),\\ \mathrm{cov}(\widehat{\boldsymbol{\beta}}) &= \Big(\theta + \frac{\phi E(v(\mu_i)}{m}\Big)\Big(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1} + O_p(1/nm^2).\end{aligned}$$

The above results apply to both normal and non-normal outcomes when an identity link is used. If one further assumes that the outcome y_{ij} is normally distributed as $y_{ij} = \boldsymbol{x}_i^T\boldsymbol{\beta} + b_i + e_{ij}$, where $b_i \sim N(0,\theta)$ and $e_{ij} \sim N(0,\phi)$, the above results become

$$\begin{aligned}\mathrm{E}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) &= O_p(1/m^2),\\ \mathrm{cov}(\widehat{\boldsymbol{\beta}}) &= \Big(\theta + \frac{\phi}{m}\Big)\Big(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1} + O_p(1/nm^2).\end{aligned}$$

One can easily show that the terms $O_p(1/m^2)$ and $O_p(1/nm^2)$ vanish. In other words, the above asymptotic bias and variance results are exact in the normal case.

4.2. Logit link function and 2 group comparison

Consider the case where one is interested in the efficacy of an intervention in a group-randomized trial involving a binary outcome. Let $g(\mu_k) = \mathrm{logit}(\mu_k)$, where μ_k is the probability of a positive response for a randomly selected individual in the kth group ($k = 1, 2$), conditional on the value of the individual's random effect. Note that while μ_k depends on b, we suppress this for notational simplicity.

Under this convention, $\mathrm{logit}(\mu_1) = \beta_0 + b$ and $\mathrm{logit}(\mu_2) = \beta_0 + \beta_1 + b$, where $b \sim \mathrm{Normal}(0, \theta)$. Take $\boldsymbol{x}_k = \begin{pmatrix}1\\0\end{pmatrix}$ for clusters in the control group ($k = 1$) and $\boldsymbol{x}_k = \begin{pmatrix}1\\1\end{pmatrix}$ for clusters in the treatment group ($k = 2$). Denote the number of clusters for the control and treatment groups as n_1 and n_2, respectively, with $n = n_1 + n_2$. Let $\gamma = n_2/n$ (the proportion of clusters in the treatment group or randomization fraction). Some calculations using (11) show that the asymptotic bias of the PQL estimator $\widehat{\boldsymbol{\beta}}$ is

$$\mathrm{E}(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta} = \frac{1}{m}\begin{pmatrix} S_1 \\ S_2 - S_1 \end{pmatrix} + \mathrm{O}_p(m^{-2}), \tag{13}$$

where for $j = 1, 2$,

$$S_j = \mathrm{E}_b\Big\{\frac{-(1-2\mu_j)}{2\mu_j(1-\mu_j)}\Big\} + \Big(\frac{1}{n_j} - 1\Big)E_b\Big\{\frac{b}{\mu_j(1-\mu_j)}\Big\}$$

The variance expression is complicated and is omitted here.

Equation (13) shows that the bias decreases as the cluster-size increases. For a fixed cluster-size, the magnitude depends on baseline event rate (reflected through S_1) as well as the true difference in the event rates for treated and control groups ($S_2 - S_1$). The randomization fraction γ, does not influence the magnitude of bias. The variance in this setting is also a function of the cluster-size and decreases as the cluster-size increases (for fixed n). Calculations using (12) show that the variance is influenced by the randomization fraction γ.

5. Simulation Study

We conducted a simulation study to examine the theoretical and empirical bias and variance of PQL, compared with the MLE, in estimating covariate effects at the cluster level under the logit link function for settings with small numbers of large clusters. We generated a common random effect for each of the $i = 1, \dots, n$ clusters (b_i) from a Normal$(0, \theta)$ distribution and, conditional on the random effect, a single Bernoulli random variable was generated from

$$y_{ij} = \mathrm{Bernoulli}(\mu_i), \quad \text{where} \quad \mu_i = \frac{\exp(\beta_0 + \beta_1 x_i + b_i)}{1 + \exp(\beta_0 + \beta_1 x_i + b_i)}. \tag{14}$$

In this model, x_i is a single cluster-level covariate assumed to follow a Normal$(1, 1)$ distribution. In our simulation study $\beta_0 = 1.5$, $\beta_1 = -1.2$ and $\theta = 0.5$.

For each simulated dataset, we estimated $\boldsymbol{\beta} = (\beta_0, \beta_1)^T$ using the PQL estimator via the SAS GLIMMIX macro, as well as the MLE using the SAS NLMIXED procedure. In both the PQL and NLMIXED approaches, we estimated $\boldsymbol{\beta}$ assuming θ was fixed and known ($\theta = 0.5$), as well as assuming θ was unknown. Whereas the PQL approach is based on the Laplace approximation to the loglikelihood function, NLMIXED approximates the marginal loglikelihood function (3) numerically (e.g., Gaussian quadrature) and finds MLEs for the parameters of interest via Newton Raphson. Variance estimates from the NLMIXED procedure are obtained from the appropriate function of the Hessian matrix at the final step of Newton Raphson. Simulation results are presented in the Tables 1 and 2 and Figures 1−3, the summaries are restricted to characterizing only the results for β_1.

Table 1. Simulation study results (from 500 simulated samples) summarizing estimated cluster-level covariate effects from the logistic-normal model (14) for θ known. (True population parameters: $\beta = -1.2$ and $\theta = 0.5$.)

	PQL			**MLE**		
	avg $\hat{\beta}(\text{SD}_{\hat{\beta}})$	avg $\text{SE}_{\hat{\beta}}$	avg MSE	avg $\hat{\beta}(\text{SD}_{\hat{\beta}})$	avg $\text{SE}_{\hat{\beta}}$	avg MSE
6 Clusters						
Cluster size=10	-1.351 (0.868)	0.673	1.526	-1.427 (0.915)	0.683	1.723
Cluster size=20	-1.228 (0.596)	0.520	0.710	-1.282 (0.622)	0.525	0.781
Cluster size=50	-1.212 (0.487)	0.451	0.474	-1.244 (0.502)	0.453	0.505
Cluster size=100	-1.199 (0.442)	0.404	0.390	-1.218 (0.451)	0.405	0.406
10 Clusters						
Cluster size=10	-1.223 (0.431)	0.439	0.372	-1.293 (0.456)	0.446	0.424
Cluster size=20	-1.203 (0.378)	0.359	0.286	-1.255 (0.396)	0.362	0.316
Cluster size=50	-1.194 (0.307)	0.299	0.188	-1.225 (0.317)	0.300	0.202
Cluster size=100	-1.189 (0.284)	0.279	0.161	-1.208 (0.290)	0.279	0.168
20 Clusters						
Cluster size=10	-1.137 (0.273)	0.280	0.153	-1.202 (0.289)	0.285	0.167
Cluster size=20	-1.159 (0.234)	0.233	0.111	-1.208 (0.245)	0.236	0.120
Cluster size=50	-1.192 (0.196)	0.201	0.076	-1.223 (0.202)	0.202	0.082
Cluster size=100	-1.185 (0.180)	0.184	0.065	-1.205 (0.184)	0.184	0.068
50 Clusters						
Cluster size=10	-1.139 (0.166)	0.170	0.059	-1.204 (0.176)	0.172	0.062
Cluster size=20	-1.168 (0.138)	0.143	0.039	-1.218 (0.145)	0.144	0.042
Cluster size=50	-1.170 (0.115)	0.122	0.027	-1.200 (0.119)	0.123	0.028
Cluster size=100	-1.172 (0.110)	0.113	0.025	-1.191 (0.113)	0.113	0.026

Table 1 has n =6, 10, 20 and 50, clusters of size m=10, 20, 50 and 100, and takes θ fixed and known, while Table 2 adopts the same settings when θ is unknown and estimated. Figure 1 compares the empirical biases of the PQL and MLE estimators of cluster-level covariate effects assuming θ is both fixed and known as well as for θ unknown; Figures 2 and 3 compare the empirical SEs and MSEs of $\widehat{\boldsymbol{\beta}}$ using the MLE and PQL methods, assuming θ is known and unknown.

We first examine the empirical bias properties of each estimation method. The results in Tables 1 and 2 and Figure 1 show that in practical group-randomized trial settings characterized by a small number clusters with moderate to large cluster sizes, the PQL estimate has small bias. In fact, PQL (REML and ML) generally performed better than the MLE in settings typical of group-randomized trials (e.g., 6 and 10 clusters in Figure 1), having the largest bias in those group-randomized settings where θ was fixed and known (top row of Figure 1; 6 and 10 clusters). Also, the empirical PQL bias (REML and ML) was similar (relatively and in absolute value) to its theoretical approximation (from (11)), especially in settings with few clusters. Figure 1 also suggests that the bias associated with PQL (ML and REML) cluster-level covariate effects is negligible in settings with few clusters, provided those clusters have large enough cluster sizes. Figure 1 and (11) also suggest that, theoretically, the PQL bias should approach the zero line for large cluster sizes but, practically, only the PQL bias seems to approximate zero line when there were only six clusters. It may be that the theoretical estimates are overly optimistic, except in the setting most typical of group-randomized trials (e.g., a small number of clusters). Finally, the MLE has the smallest bias in settings where there are large numbers of clusters (20 and 50), as expected.

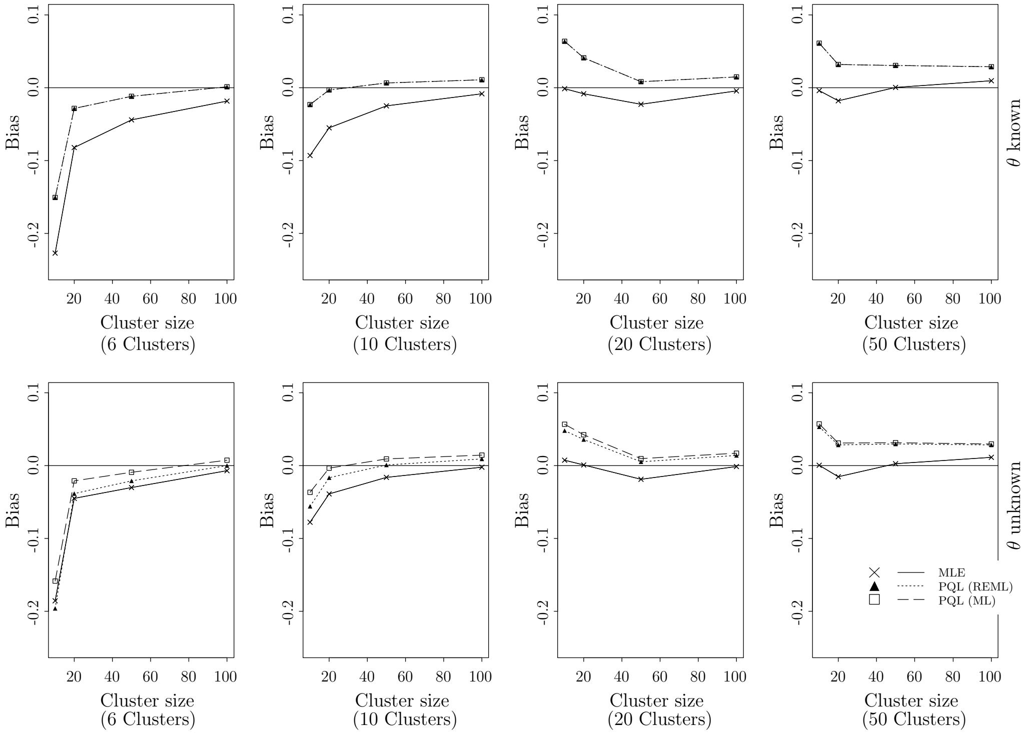

Figure 1. Average bias in estimating cluster-level covariate effects from the logistic mixed model (14) via penalized quasi-likelihood (PQL (REML) and PQL (ML)) and MLE expression in equation (11) (THEORETICAL) for θ known (top row) and θ unknown (bottom row).

Table 2. Simulation study results (from 500 simulated samples) summarizing estimated cluster-level covariate effects from the logistic-normal model (14) for unknown θ. (True population parameters: $\beta = -1.2$ and $\theta = 0.5$.)

	PQL (ML)			**PQL (REML)**			**MLE**		
	avg $\hat{\beta}(\mathrm{SD}_{\hat{\beta}})$	avg $\mathrm{SE}_{\hat{\beta}}$	avg MSE	avg $\hat{\beta}(\mathrm{SD}_{\hat{\beta}})$	avg $\mathrm{SE}_{\hat{\beta}}$	avg MSE	avg $\hat{\beta}(\mathrm{SD}_{\hat{\beta}})$	avg $\mathrm{SE}_{\hat{\beta}}$	avg MSE
6 Clusters									
Cluster size=10	-1.359 (0.890)	0.581	1.608	-1.396 (0.942)	0.686	1.813	-1.386 (0.895)	0.625	1.635
Cluster size=20	-1.221 (0.593)	0.414	0.702	-1.238 (0.605)	0.488	0.734	-1.245 (0.605)	0.432	0.897
Cluster size=50	-1.209 (0.492)	0.363	0.483	-1.221 (0.497)	0.437	0.494	-1.223 (0.501)	0.373	0.503
Cluster size=100	-1.193 (0.439)	0.325	0.386	-1.200 (0.442)	0.394	0.390	-1.207 (0.445)	0.331	0.396
10 Clusters									
Cluster size=10	-1.237 (0.449)	0.404	0.404	-1.256 (0.460)	0.446	0.425	-1.278 (0.464)	0.435	0.436
Cluster size=20	-1.204 (0.385)	0.316	0.296	-1.217 (0.390)	0.350	0.304	-1.239 (0.398)	0.333	0.318
Cluster size=50	-1.191 (0.308)	0.264	0.190	-1.199 (0.311)	0.294	0.193	-1.216 (0.315)	0.273	0.198
Cluster size=100	-1.186 (0.282)	0.245	0.160	-1.191 (0.284)	0.272	0.161	-1.202 (0.287)	0.249	0.164
20 Clusters									
Cluster size=10	-1.144 (0.284)	0.262	0.164	-1.152 (0.286)	0.275	0.166	-1.193 (0.297)	0.282	0.177
Cluster size=20	-1.158 (0.236)	0.215	0.113	-1.164 (0.237)	0.226	0.114	-1.199 (0.245)	0.227	0.120
Cluster size=50	-1.191 (0.196)	0.187	0.077	-1.195 (0.197)	0.196	0.078	-1.219 (0.202)	0.193	0.082
Cluster size=100	-1.183 (0.181)	0.170	0.066	-1.186 (0.181)	0.179	0.066	-1.202 (0.185)	0.173	0.068
50 Clusters									
Cluster size=10	-1.143 (0.170)	0.161	0.061	-1.147 (0.171)	0.164	0.061	-1.200 (0.181)	0.175	0.065
Cluster size=20	-1.169 (0.139)	0.136	0.039	-1.172 (0.139)	0.139	0.040	-1.216 (0.145)	0.144	0.042
Cluster size=50	-1.169 (0.115)	0.117	0.028	-1.170 (0.116)	0.120	0.028	-1.198 (0.119)	0.121	0.028
Cluster size=100	-1.170 (0.111)	0.108	0.025	-1.172 (0.111)	0.111	0.025	-1.189 (0.113)	0.111	0.026

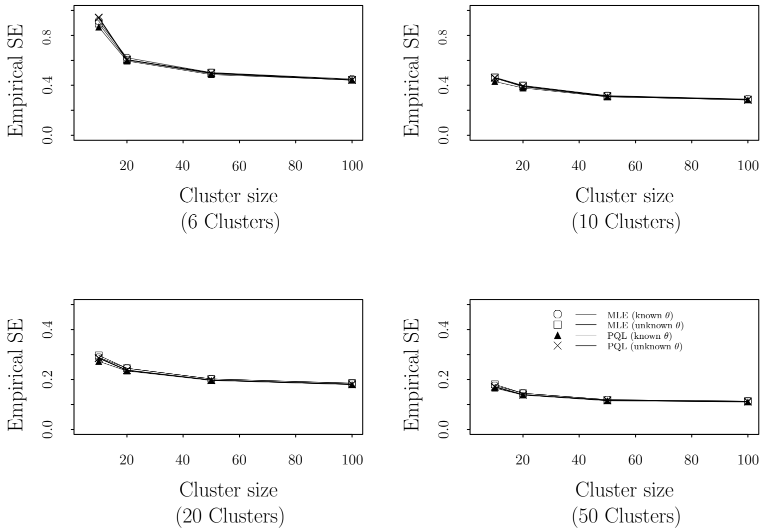

Figure 2. Comparison of empirical SEs for estimated cluster-level covariate effects from the logistic mixed model (14) via penalized quasi-likelihood (PQL) and the MLE assuming θ is known and unknown.

Tables 1 and 2 also contain the average estimated mean square error (MSE) for each method of estimation. The PQL MSE is smaller than the MLE MSE in each simulation setting considered. In settings where θ was assumed to be known, the average MSE difference between PQL and MLE was larger in settings with few clusters and negligible in settings with many clusters. Regardless of the number of clusters, the MSE for each of the three estimation methods decreased as the cluster size increased. We observed similar results in settings where θ was estimated. Specifically, Table 2 suggests that the MSE associated with each of the three estimation methods are similar when θ is unknown and that, in general, PQL (ML) has modestly superior MSE properties when compared to the other methods.

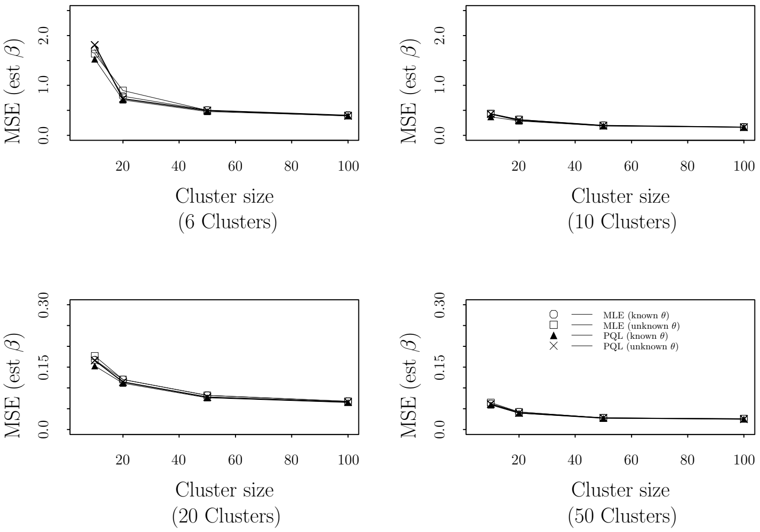

Figure 3. Comparison of MSEs for estimated cluster-level covariate effects from the logistic mixed model (14) via penalized quasi-likelihood (PQL) and the MLE assuming θ is known and unknown.

Figures 2 and 3 characterize the empirical standard errors and MSEs associated with estimating cluster-level covariate effects using PQL and MLE. The results also show that SE efficiency and MSE efficiency of all three methods was indistinguishable when θ was assumed fixed and known or when θ was estimated. However, the PQL estimates show a slight improvement in SEs and MSEs over the MLEs.

We also compared the computing resources required for the PQL approach versus the MLE using the SAS NLMIXED procedure to see if there were big differences. Although NLMIXED had comparable results (average parameter estimates, estimated standard errors, etc.), we found the procedure to be more computationally intensive, greatly increasing the total computing time for modeling the same datasets, when compared to fitting equivalent PQL models. For example, the CPU for modeling a single simulated dataset with 10 clusters and 100 responses per cluster was approximately five times greater using the NLMIXED procedure than the PQL approach (NLMIXED = 15.64 seconds and PQL = 3.47 seconds). This additional CPU time is likely due to the Newton Raphson algorithm, which involves inverting a matrix whose dimension depends on the number of independent responses within clusters. Because of this matrix inversion step, the method may be unstable if the matrix to be inverted is singular, or computationally intensive if the matrix is large. One may want to consider the trade-off in CPU time and bias in estimating covariate effects of interest associated with fitting the more computationally intense random effects models, using the NLMIXED procedure versus using the PQL approach.

6. Discussion

The PQL approach has been shown to produce biased estimates of covariate effects in clustered data settings with small numbers of observations per cluster or in settings with large variance components. Our early empirical results show that this bias may be small compared to MLEs in common group-randomized settings (small numbers of large clusters). In this paper, we have shown theoretically and in a simulation study that the bias in estimating cluster-level covariate effects using the PQL method is inversely proportional to the cluster-size. This result has important implications, especially in the context of community-based and/or group-randomized studies which typically have small numbers of large clusters. The results presented here suggest that the bias in estimating covariate effects at the cluster-level via PQL can be minimized to an extent in settings where the number of independent clusters may be small and/or fixed (e.g., number of census block-groups in a school district of interest) by sampling more subjects within clusters. We have also shown that the variance associated with these estimates is inversely proportional to the total number of clusters. Additionally, the variance is a function of both the between and within subject variability as well as the cluster size. We have found in our simulation study that the MSEs of the PQL estimates are slightly smaller than the MLE counterparts in practical group-randomized trial settings. We have presented main results both theoretically and from simulation studies for equal cluster-sizes, but have generalized our results unequal cluster size settings at the end of Section 3.

An alternative method is to estimate the regression coefficients using a two-stage method. Although the two-stage method is simple, it has major limitations: one would not be able to perform a two-stage analysis for cluster-level covariates, the main interest in the current paper; inference in two-stage analysis is difficult since one has to account for variability in estimation of both stages. Fitzmaurice, Laird and Ware (2004) note that their discussion of the two-stage random effects model formulation is for pedagogical purposes, and caution readers that such a formulation, although helpful conceptually, introduces extraneous and sometimes impractical model restrictions.

We have focused on studying the asymptotic bias of the regression coefficients of cluster-level covariates in group-randomized trial settings with a small number of clusters and large cluster sizes. We assumed the variance component θ known in our calculations. In practice the variance component is often unknown. The theoretical results for quantifying the bias of covariate effects when θ is unknown is beyond the scope of this paper, and is of significant interest for future research. Our simulation study results suggest that in typical group-randomized trial settings, the practical implication of assuming θ is known and fixed vs settings where θ is estimated are negligible, even in settings with small numbers of clusters, given a reasonable cluster size (e.g., cluster size ≥ 20). Although we have not presented theoretical arguments exploring a relaxation of this assumption, our simulation studies may give some practical insight on how the results might compare in settings where θ is unknown.

Another area of future research is to study the asymptotic bias of the regression coefficients of subject-level covariates in settings with a small number of clusters and large cluster sizes. This would require developing a different asymptotic analytic technique.

Acknowledgements

This research has been generously supported by the following National Institutes of Mental Health (MH) and National Cancer Institute (CA) research grants: Bellamy MH61892; Li CA95747; Lin CA76404; and Ryan CA48061. Additionally, the authors thank the Editor, an Associate Editor and two referees for their helpful and constructive suggestions that have improved the paper.

Appendix 1. Proof of Proposition 1

We can write $\tilde{b}_i$ as $\tilde{b}_i = h(\bar{y}_{i.}, \boldsymbol{x}_i^T\boldsymbol{\beta}, \phi, \theta, 1/m)$, where $h(\cdot)$ is the solution to (9). Then the PQL estimator $\widehat{\boldsymbol{\beta}}$ can be written as

$$\widehat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1} \sum_{i=1}^{n} \boldsymbol{x}_i g\left\{\bar{y}_{i.} - h(\bar{y}_{i.}, \boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}, \phi, \theta, 1/m)\frac{\phi}{m\bar{a}_{i.}\theta}\right\}.$$

We first expand $\bar{y}_{i.}$ in $h(\cdot)$ about μ_i as

$$\begin{aligned}\widehat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\Big[&g\Big\{\mu_i - h(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\frac{\phi}{m\bar{a}_{i.}\theta}\Big\}\\
&+(\bar{y}_{i.}-\mu_i)g'\Big(\mu_i - h(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\frac{\phi}{m\bar{a}_{i.}\theta}\Big)\\
&\times\Big(1-h'_{\mu_i}(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\frac{\phi}{m\bar{a}_{i.}\theta}\Big)\\
&+\frac{(\bar{y}_{i.}-\mu_i)^2}{2!}\Big\{g'\Big(\mu_i - h(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\frac{\phi}{m\bar{a}_{i.}\theta}\Big)\\
&\times\Big(-h''_{\mu_i\mu_i}(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\frac{\phi}{m\bar{a}_{i.}\theta}\Big)\\
&+\Big(1-h'_{\mu_i}(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\frac{\phi}{m\bar{a}_{i.}\theta}\Big)^2 g''\Big(\mu_i-h(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\frac{\phi}{m\bar{a}_{i.}\theta}\Big)\Big\}\Big]\\
&+\mathrm{O}_p(1/m^2),\end{aligned}$$

where $h'_{\mu_i}(\cdot) = \partial h(\cdot)/\partial\mu_i$ and $h"_{\mu_i\mu_i}(\cdot) = \partial h(\cdot)/\partial\mu_i\mu_i$. A further expansion of $g(\cdot)$ about $g(\mu_i)$ gives

$$\begin{aligned}\widehat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i\Big\{&g(\mu_i) - h(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\frac{\phi}{m\bar{a}_{i.}\theta}g'(\mu_i)\\
&+(\bar{y}_{i.}-\mu_i)\Big[g'(\mu_i) - \frac{\phi}{m\bar{a}_{i.}\theta}\Big\{h(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)g''(\mu_i)\\
&+g'(\mu_i)h'_{\mu_i}(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\Big\}\Big]\\
&+\frac{(\bar{y}_{i.}-\mu_i)^2}{2}\Big[g''(\mu_i) - \frac{\phi}{m\bar{a}_{i.}\theta}\Big\{h(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)g'''(\mu_i)\\
&+g'(\mu_i)h''_{\mu_i\mu_i}(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)+2g''(\mu_i)h'_{\mu_i}(\mu_i,\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}},\phi,\theta,1/m)\Big\}\Big]\Big\}\\
&+\mathrm{O}_p(1/m^2).\end{aligned}$$

If we expand $\widehat{\boldsymbol{\beta}}$ in $h(\cdot)$ about $\boldsymbol{\beta}$ and use the fact that $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}+b_i$, some calculations give

$$\widehat{\boldsymbol{\beta}} = \Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i(\boldsymbol{x}_i^T\boldsymbol{\beta}+d_i) + \mathrm{O}_p(1/m^2) + \mathrm{O}_p(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^2 - \frac{\phi}{m\theta}\boldsymbol{K}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta}),$$

where

$$\begin{aligned}d_i = \Big[&b_i - \frac{\phi}{m\bar{a}_{i.}\theta}g'(\mu_i)h_i(\cdot) + (\bar{y}_{i.}-\mu_i)\Big[g'(\mu_i) - \frac{\phi}{m\bar{a}_{i.}\theta}\{g''(\mu_i)h_i(\cdot)+g'(\mu_i)h'_{\mu_i}(\cdot)\}\Big]\\
&+\frac{1}{2}(\bar{y}_{i.}-\mu_i)^2\Big[g''(\mu_i) - \frac{\phi}{m\bar{a}_{i.}\theta}\{g'''(\mu_i)h_i(\cdot)+g'(\mu_i)h''_{\mu_i\mu_i}(\cdot)+2g''(\mu_i)h'_{\mu_i}(\cdot)\}\Big]\end{aligned}$$

and $(\cdot)$ denotes $(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \phi, \theta, 1/m)$, $\boldsymbol{K}$ is the $p \times p$ matrix

$$\boldsymbol{K} = \Big(\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1} \sum_{i=1}^{n} \frac{1}{\bar{a}_{i.}} \boldsymbol{x}_i \boldsymbol{c}_i^T,$$

$$\begin{aligned}\boldsymbol{c}_i = {} & g'(\mu_i)h'_{i\beta}(\cdot) + (\bar{y}_{i.} - \mu_i)\left\{g'(\mu_i)h''_{i\mu_i\beta}(\cdot) + g''(\mu_i)h'_{i\beta}(\cdot)\right\} \\ & + \frac{1}{2}(\bar{y}_{i.} - \mu_i)^2\left\{h'_{i\beta}(\cdot)g'''(\mu_i) + h'''_{i\mu_i\mu_i\beta}(\cdot)g'(\mu_i) + 2h''_{i\mu_i\beta}(\cdot)g''(\mu_i)\right\},\end{aligned}$$

and $h'_{i\beta}(\cdot)$ $h''_{i\mu_i\beta}(\cdot)$ and $h'''_{i\mu_i\mu_i\beta}(\cdot)$ denote the derivatives of $h_i(\cdot)$, $h'_{\mu_i}(\cdot)$ and $h''_{\mu_i\mu_i}(\cdot)$ with respect to $\boldsymbol{\beta}$. Using $g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} + b_i$ and collecting $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}$ terms,

$$\begin{aligned}\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} &= \Big(\boldsymbol{I} + \frac{\phi}{m\theta}\boldsymbol{K}\Big)^{-1}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i d_i + O_p(1/m^2) \\ &= \Big(\boldsymbol{I} - \frac{\phi}{m\theta}\boldsymbol{K}\Big)\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i d_i + O_p(1/m^2) \\ &= \Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i d_i\Big) \\ &\quad -\frac{\phi}{m\theta}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\Big(\sum_{i=1}^{n}\frac{1}{\bar{a}_{i.}}\boldsymbol{x}_i\boldsymbol{c}_i^T\Big)\Big(\sum_{i=1}^{n}\boldsymbol{x}_i\boldsymbol{x}_i^T\Big)^{-1}\Big(\sum_{i=1}^{n}\boldsymbol{x}_i d_i\Big) + O_p(1/m^2),\end{aligned}$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix.

Now examine $h_i(\cdot) = h(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \theta, 1/m)$ and $h_{i\beta}(\cdot) = \partial h(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \theta, 1/m)/\partial\boldsymbol{\beta}$. We show in Appendix 2 that $h_i(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \theta, 1/m) = b_i + O_p(1/m)$, and in Appendix 3 that $h_{i\beta}(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \theta, 1/m) = -\boldsymbol{x}_i^T + O_p(1/m)$. Replacing $h_i(\cdot)$ by b_i in d_i and $h'_{i\beta}(\cdot)$ by $-\boldsymbol{x}_i^T$ in $\boldsymbol{c}_i^T$, noticing $E\{g'(\mu_i)b_{i'}\} = E\{g'(\mu_i)b_i\}$ if $i = i'$ and 0 otherwise, and keeping the terms of order $O_p(1/m)$, some calculations give the bias and variance expressions in (11) and (12).

Appendix 2. Proof of $h_i(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \phi, \theta, 1/m) = b_i + O_p(1/m)$

By definition, $h(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \theta, 1/m)$ is the solution for $\tilde{b}_i$ of the equation,

$$\begin{aligned}\mu_i - g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i) &= \frac{\tilde{b}_i\phi}{m\bar{a}_{i.}\theta} \\ \mu_i - \mu(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i) &= \frac{\tilde{b}_i\phi}{m\bar{a}_{i.}\theta},\end{aligned} \tag{15}$$

where $\mu(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i) = g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i)$. Expanding the left side of the above equation in a first order Taylor series about b_i, we get

$$\mu_i - \mu(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i) \approx \mu_i - \Big\{\mu(\boldsymbol{x}_i^T\boldsymbol{\beta} + b_i) + (\tilde{b}_i - b_i)\mu'(\boldsymbol{x}_i^T\boldsymbol{\beta} + b_i)\Big\} = \frac{\tilde{b}_i\phi}{m\bar{a}_{i.}\theta}$$

$$\mu_i - \mu_i - (\tilde{b}_i - b_i)\mu'(\boldsymbol{x}_i^T\boldsymbol{\beta} + b_i) = \frac{\tilde{b}_i\phi}{m\bar{a}_{i.}\theta}.$$

Solving for $\tilde{b}_i$,

$$\tilde{b}_i = b_i\Big(\frac{\mu'(\boldsymbol{x}_i^T\boldsymbol{\beta} + b_i)}{\mu'(\boldsymbol{x}_i^T\boldsymbol{\beta} + b_i) + \frac{\phi}{m\bar{a}_{i.}\theta}}\Big).$$

For large m, one can easily show $\tilde{b}_i = b_i + O_p(1/m)$.

Appendix 3. Proof of $h'_{i\beta}(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \phi, \theta, 1/m) = -\boldsymbol{x}_i^T + O_p(1/m)$.

By definition,

$$\mu_i - g^{-1}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i) = \mu_i - \mu(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i) = \frac{\tilde{b}_i\phi}{m\bar{a}_{i.}\theta}. \tag{16}$$

Taking derivatives of (16) wrt $\boldsymbol{\beta}$,

$$-\Big(\boldsymbol{x}_i^T + h'_\beta(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \theta, 1/m)\Big)\mu'_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i) = \frac{h'_\beta(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \phi, \theta, 1/m)}{m\theta}$$

$$h'_{i\beta}(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \prime, \phi, \theta, 1/m) = -\boldsymbol{x}_i^T\Big(\frac{\mu'_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i)}{\frac{\phi}{m\bar{a}_{i.}\theta} + \mu'_{\boldsymbol{\beta}}(\boldsymbol{x}_i^T\boldsymbol{\beta} + \tilde{b}_i)}\Big).$$

As m gets large, $h'_{i\beta}(\mu_i, \boldsymbol{x}_i^T\boldsymbol{\beta}, \phi, \theta, 1/m) = -\boldsymbol{x}_i^T + O_p(1/m)$.

References

Akritas, M. G. (1991). Robust M estimation in the two-sample problem. *J. Amer. Statist. Assoc.* **86**, 201-204.

Bellamy, S. L., Gibberd, R., Hancock, L., Howley, P., Kennedy, B., Klar, N., Lipsitz, S. and Ryan, L. (2000). Analysis of dichotomous outcome data for community intervention studies. *Statist. Meth. Medical Res.* **9**, 135-159.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *J. Amer. Statist. Assoc.* **88**, 9-25.

Breslow, N. E. and Lin, X. (1995). Bias correction in the generalized linear mixed model with a single component of dispersion. *Biometrika* **82**, 81-91.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research.* Arnold, London.

Fitzmaurice, G. M., Laird, N. M. and Ware, J. H. (2004). *Applied Longitudinal Analysis.* John Wiley, New York.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Internat. Statist. Rev.* **55**, 245-259.

Jiang, J. (1999). Conditional inference about generalized linear mixed models. *Ann. Statist.* **27**, 1974-2007.

Jiang J. and Lahiri P. (2001). Empirical best prediction for small area inference with binary data, *Ann. Inst. Statist. Math.* **53**, 217-243.

Lin, X. and Breslow, N. E. (1996) Bias correction in generalized linear mixed models with multiple components of dispersion *J. Amer. Statist. Assoc.* **91**, 1007-1016.

Mancl, L. A. and Leroux, B. G. (1996). Efficiency of regression estimates for clustered data. *Biometrics* **52**, 500-511.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* 2nd Edition. Chapman and Hall, London.

Neuhaus, J. M. and Segal, M. R. (1997). An assessment of approximate maximum likelihood estimators in generalized linear mixed models. *Modeling Longitudinal and Spatially Correlated Data. Methods, Applications and Future Directions.* Springer.

Ten Have, T. and Localio, R. (1999). Empirical bayes estimation of random effects parameters in mixed effects logistic regression models. *Biometrics* **55**, 1022-1029.

Van Der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.

Vonesh, E. F., Wang, H., Nie L. and Majumdar, D. (2002). Conditional second-order generalized estimating equations for generalized linear and nonlinear mixed-effects models. *J. Amer. Statist. Assoc.* **97**, 271-283.

Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104, U.S.A.

E-mail: sbellamy@cceb.upenn.edu

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

Dana-Farber Cancer Institute, Boston, MA 02115 U.S.A.

E-mail: yili@hsph.harvard.edu

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

E-mail: xlin@hsph.harvard.edu

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, U.S.A.

E-mail: lryan@hsph.harvard.edu

(Received December 2003; accepted January 2005)