Research Article

Statistics in Medicine

Received 3 January 2012, Accepted 27 August 2012 Published online 13 September 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5624

Adjusting for confounding by neighborhood using generalized linear mixed models and complex survey data

Babette A. Brumback,[a*†] Hao W. Zheng[a] and Amy B. Dailey[b]

When investigating health disparities, it can be of interest to explore whether adjustment for socioeconomic factors at the neighborhood level can account for, or even reverse, an unadjusted difference. Recently, we proposed new methods to adjust the effect of an individual-level covariate for confounding by unmeasured neighborhood-level covariates using complex survey data and a generalization of conditional likelihood methods. Generalized linear mixed models (GLMMs) are a popular alternative to conditional likelihood methods in many circumstances. Therefore, in the present article, we propose and investigate a new adaptation of GLMMs for complex survey data that achieves the same goal of adjusting for confounding by unmeasured neighborhood-level covariates. With the new GLMM approach, one must correctly model the expectation of the unmeasured neighborhood-level effect as a function of the individual-level covariates. We demonstrate using simulations that even if that model is correct, census data on the individual-level covariates are sometimes required for consistent estimation of the effect of the individual-level covariate. We apply the new methods to investigate disparities in recency of dental cleaning, treated as an ordinal outcome, using data from the 2008 Florida Behavioral Risk Factor Surveillance System (BRFSS) survey. We operationalize neighborhood as zip code and merge the BRFSS data with census data on ZIP Code Tabulated Areas to incorporate census data on the individual-level covariates. We compare the new results to our previous analysis, which used conditional likelihood methods. We find that the results are qualitatively similar. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: confounding; generalized linear mixed models; pseudolikelihood; complex survey data; health disparities

1. Introduction

When investigating health disparities, it can be of interest to explore whether adjustment for unmeasured socioeconomic factors at the neighborhood level can account for, or even reverse, an unadjusted difference. Recently, we have been investigating racial and ethnic disparities in dental preventive care using complex survey data from the 2008 Florida Behavioral Risk Factor Surveillance System (BRFSS) survey [1,2]. We have needed to develop new statistical methods to simultaneously overcome two hurdles. First, accounting for neighborhood effects by including a dummy variable for each neighborhood into our model requires us to use either a conditional maximum likelihood approach or a generalized linear mixed model (GLMM) approach to estimation, because ordinary maximum likelihood theory is not applicable because of the large number of neighborhood parameters [3]. Second, we have needed to generalize these approaches for use with complex survey data. So far, we have generalized conditional likelihood methods for use with binary, ordinal, or multinomial outcomes and complex survey data [1, 2, 4], via a weighted composite conditional likelihood [5], based on ideas presented in [6–9]. We refer to these methods as conditional pseudolikelihood methods. The development treated each within-neighborhood pair of sampled observations as independent and used the inverse probability of selection for each pair as sampling weights. In the present paper, we show how to apply the same idea to develop GLMM methods for use with any type of outcome.

[a]*Department of Biostatistics, University of Florida, Gainesville, FL 32611, U.S.A.*
[b]*Department of Health Sciences, Gettysburg College, Gettysburg, PA 17325, U.S.A.*
**Correspondence to: Babette A. Brumback, Department of Biostatistics, University of Florida, Gainesville, FL 32611, U.S.A.*
[†]*E-mail: brumback@ufl.edu*

In our BRFSS example, the outcome is ordinal and represents how long it has been since an individual has had his or her teeth cleaned by a dentist or dental hygienist (within the past year, within the past 2 years, within the past 5 years, 5 or more years ago, or never). Our covariate of primary interest is race/ethnicity, which we categorized into white non-Hispanic, African American non-Hispanic, Hispanic, and other. Racial/ethnic or socioeconomic disparities have been observed across a spectrum of oral health outcomes, including presence of untreated dental caries [10, 11], other oral health problems (e.g., toothaches, tooth loss, or periodontal disease) [12–16], and self-rated or parent-rated dental health [14, 15, 17–19]. We operationalized an individual's neighborhood as the zip code and merged the BRFSS data with census data on ZIP Code Tabulated Areas (ZCTAs). We excluded from our analysis individuals without teeth, individuals with missing data on any of the covariates we used in the analysis, and individuals with zip codes not matching the ZCTAs. This resulted in a final sample of 8376 Florida residents. Using logistic regression software (SAS PROC SURVEYLOGISTIC) with a cumulative logit link [9] and accounting for the complex survey design of the BRFSS (disproportionate stratified sampling with poststratification), it is found that the crude odds ratios and 95% confidence intervals representing the association between race/ethnicity and more recent dental cleaning, with non-Hispanic whites as the reference group, were 0.55 (0.42, 0.73) for non-Hispanic African Americans, 0.81 (0.62, 1.07) for Hispanics, and 0.86 (0.57, 1.29) for other. Accounting for the simple demographic factors gender and age (categorized as 18–34, 35–54, 55–64, and 65+ years) as additive terms in the logistic regression model led to adjusted odds ratios of 0.60 (0.46, 0.79) for non-Hispanic African Americans, 0.95 (0.71, 1.26) for Hispanics, and 0.95 (0.63, 1.42) for other. Therefore, after accounting for demographic differences due to age and gender, we observe a disparity between non-Hispanic African American and white non-Hispanic individuals and no disparity between Hispanic and white non-Hispanic individuals. Brumback *et al.* [1] found, using new conditional pseudolikelihood methods, that further accounting for socioeconomic and unmeasured neighborhood factors removed the disparity between non-Hispanic African American and white non-Hispanic individuals and that it reversed the original disparity between Hispanic and white non-Hispanic individuals. In the present article, we will investigate whether our results are qualitatively similar when we apply our new GLMM methods.

We organize the paper as follows. Section 2 explains the details of estimation with the GLMM method. Section 3 presents a simulation study, and Section 4 applies the GLMM method to our health disparities question using the Florida BRFSS data. Section 5 concludes with a discussion.

2. Generalized linear mixed models with complex survey data

We are interested in estimating the parameter β of one of the following population-level models. Let $i = 1, \ldots, M$ index neighborhoods in the population and $j = 1, \ldots, N_i$ index all individuals in the population who belong to neighborhood i. We assume that the finite population has been sampled from a hypothetical infinite superpopulation as follows: the M neighborhoods are sampled independent and identically distributed (i.i.d.) from an infinite population of neighborhoods, and then the N_i individuals per neighborhood have been sampled i.i.d. from an infinite population of individuals within neighborhood. Then a simple model for estimating the effect of an individual-level covariate vector X_{ij} (note that this covariate vector does not include an intercept term) on a binary or continuous outcome Y_{ij} while accounting for confounding by a general neighborhood-level effect b_i is

$$E(Y_{ij}|X_i, b_i) = h(X_{ij}\beta + b_i), \tag{1}$$

where $X_i = \left(X_{i1}^T, \ldots, X_{iN_i}^T\right)$, h is a known inverse-link function, and $Y_{ij} \amalg Y_{il}|X_i, b_i$, for $j \neq l$. Note that X_{ij} does not include an intercept term because we do not require b_i to have mean zero. For ordinal or multinomial outcomes, as in our BRFSS example, Y_{ij} can equal one of the categories $1, \ldots, K$, and we require a more general model. One such model is the proportional odds model for clustered ordinal outcomes [9]. Let $V_{ijk} = 1$ if $Y_{ij} \leqslant k$ and 0 otherwise, for $k = 1, \ldots, K-1$. The model is

$$E(V_{ijk}|X_i, b_i) = \text{expit}(X_{ij}\beta + \alpha_k + b_i), \quad k = 1, \ldots, K-1, \tag{2}$$

where $\text{expit}(x) = \exp(x)/(1+\exp(x))$, the α_k are nondecreasing, and we assume $Y_{ij} \amalg Y_{il}|X_i, b_i$ for $j \neq l$. This model is very similar to model (1), and our parameter of main interest is again β, which now represents the log odds ratio for $Y_{ij} \leqslant k$ versus $Y_{ij} > k$, which the model assumes is constant

in k. Another model for ordinal or multinomial outcomes is the baseline category logit model [9]. The model is

$$\log(P(Y_{ij}=k|X_i,b_i))/\log(P(Y_{ij}=K|X_i,b_i))=X_{ij}\beta_k+\alpha_k+b_i,\quad k=1,\ldots,K-1, \tag{3}$$

where we assume $Y_{ij} \amalg Y_{il}|X_i,b_i$ for $j\neq l$. The parameters $\beta_k, k=1,\ldots,K-1$ represent log odds ratios for $Y_{ij}=k$ versus $Y_{ij}=K$, where K is the baseline, or reference, category.

Typical application of GLMMs assumes an ordinary cluster sampling design; that is, we have a simple random sample of m neighborhoods from the population of M neighborhoods, and our sample contains all N_i individuals within each sampled neighborhood i. GLMM estimation requires specification of a distribution for the $b_i, i=1,\ldots,m$. Typically, one treats the b_i as i.i.d. $N(\mu,\eta^2)$ random variables. However, when there is confounding by neighborhood, b_i is necessarily associated with X_i. We therefore need to assume a model that relates b_i to X_i. We let $b_i=\psi(X_i,\gamma)+\tau\delta_i$, where $\delta_i \amalg X_i$, the δ_i are i.i.d. $N(0,1)$, and $\psi(X_i,\gamma)$ is a parametric function of X_i, with parameter γ. Perhaps the most natural choice for the parametric function is

$$\psi(X_i,\gamma)=\gamma_0+\bar{X}_i\gamma_1, \tag{4}$$

where $\bar{X}_i=(1/N_i)\Sigma_{j=1}^{N_i}X_{ij}$. For the sake of identifiability, we will sometimes need to specify $\gamma_0=0$, depending on whether one of the α_k is set equal to 0. Neuhaus and Kalbfleisch [20] promoted the use of GLMM regression with model (4), and Neuhaus and McCulloch [21] dubbed the 'poor man's' alternative to conditional likelihood methods. When the inverse-link function h in model (1) is the identity function and $\text{var}(Y_{ij}|X_i,b_i)$ is constant in i and j, it so happens [22] that consistent estimation of β is achieved by setting $\psi(X_i,\gamma)$ equal to model (4), regardless of whether model (4) is the correct model for $\psi(X_i,\gamma)$. However, misspecifying $\psi(X_i,\gamma)$ can lead to an inconsistent estimator of β when h is the exponential function [22], or the expit function [23], and therefore also for models (2) and (3). In Section 3, we present a simulation study to document that the inconsistency can be substantial for very small neighborhoods.

Let θ represent the collection of parameters for a given model; for example, $\theta=(\alpha_1,\ldots,\alpha_{K-1},\beta,\tau,\gamma)$ for the proportional odds model (2). Let $L_{ij}(\theta;\delta)$ represent the likelihood of Y_{ij} given δ_i. Then the maximum likelihood estimator of θ optimizes

$$\log L(\theta)=\Sigma_{i=1}^{m}\log\int_{-\infty}^{+\infty}\left[\exp\left\{\Sigma_{j=1}^{N_i}\log L_{ij}(\theta;\delta_i)\right\}\right]\phi(\delta_i)\mathrm{d}\delta_i, \tag{5}$$

where $\phi(\cdot)$ is the standard normal pdf.

With complex survey data, we may observe not only just a sample of m neighborhoods from the population of M neighborhoods but typically also just a sample n_i of the N_i individuals in neighborhood i; we also typically sample the individuals and sometimes the neighborhoods, with unequal probabilities. Let W_i be the inverse probability of selecting neighborhood i and $W_{j|i}$ be the inverse probability of selecting individual j from neighborhood i, given that neighborhood i has been selected. The GLLAMM software in Stata version 11.0 by Rabe-Hesketh and Skrondal [24] estimates θ by optimizing the log pseudolikelihood

$$\log\hat{L}(\theta)=\Sigma_{i=1}^{m}W_i\log\int_{-\infty}^{+\infty}\left[\exp\left\{\Sigma_{j=1}^{n_i}W_{j|i}\log L_{ij}(\theta;\delta_i)\right\}\right]\phi(\delta_i)\mathrm{d}\delta_i. \tag{6}$$

There are two problems with this approach. First, $\hat{L}(\theta)$ in (6) represents an inconsistent estimator of the limit of $L(\theta)$ in (5) when m tends to ∞ but the n_i do not tend towards N_i, because the weighted sum over j is exponentiated. Second, we typically do not know $\bar{X}_i$ based on the census of individuals $j=1,\ldots,N_i$ in neighborhood i. The best we can do is to replace it with the estimate $\bar{X}_i^w=\left(\Sigma_{j=1}^{n_i}X_iW_{j|i}\right)/\left(\Sigma_{j=1}^{n_i}W_{j|i}\right)$. However, especially for small n_i, this further biases $\hat{L}(\theta)$ as an estimate of $L(\theta)$. Brumback *et al.* [25] documented substantial bias in the resulting estimator of β with a simulation. Rabe-Hesketh and Skrondal [24] had also noted bias in some estimated components of θ in a simulation that did not include confounding by neighborhood.

Assuming that census data are available to bypass the second problem with a known $\bar{X}_i$, one could circumvent the first problem in the linear mixed model setting by using the method of Korn and Graubard [26], who developed a consistent estimator for θ of (1) with h as the identity function. However, their

estimator does not generalize for GLMMs, and it is also not easy to compute. Our solution is to optimize a weighted composite likelihood (see [5] for an overview of estimation based on composite likelihoods). Unlike the examples considered in [5], our weights pertain to consistency rather than precision; that is, because our weights are survey sampling weights (e.g., inverse probability of selection weights) rather than precision weights (e.g., inverse-variance weights), inclusion of our weights is necessary for consistency of the resulting estimator. We have already implemented this type of approach for conditional likelihood estimation with complex survey data [1, 2, 4], based on the idea of Graubard and Korn [6].

Our method is based on a composite likelihood for the population data composed of contributions from all possible within-cluster pairs, that is,

$$\log L^c(\theta) = \Sigma_{i=1}^{M} \Sigma_{j=1}^{N_i-1} \Sigma_{l=j+1}^{N_i} \log \int_{-\infty}^{+\infty} L_{ij}(\theta;\delta_i) L_{il}(\theta;\delta_i) \phi(\delta_i) \mathrm{d}\delta_i . \tag{7}$$

For complex survey data, let W_{ijl} be the inverse probability of selecting pair (ij, il) into the sample, assuming that every within-cluster pair of individuals in the population has a positive probability of being selected, which is known as a positivity assumption [27]. Then we can maximize a consistent estimate of $L^c(\theta)$, namely

$$\log \hat{L}^c(\theta) = \Sigma_{i=1}^{m} \Sigma_{j=1}^{n_i-1} \Sigma_{l=j+1}^{n_i} W_{ijl} \log \int_{-\infty}^{+\infty} L_{ij}(\theta;\delta_i) L_{il}(\theta;\delta_i) \phi(\delta_i) \mathrm{d}\delta_i , \tag{8}$$

which has the form of a weighted composite likelihood. Assuming that the $\bar{X}_i$ are known, maximizing (8) leads to a consistent estimator of θ, because the expectation of (8) with respect to the complex sampling design is equal to (7), and the estimator based on (7) is consistent. That is to say, Equation (8) is a Horvitz–Thompson estimator of Equation (7). However, when the $\bar{X}_i$ are estimated via $\bar{X}_i^w$, substantial bias can result. In Section 3, we provide a simulation study to document this.

Computing our estimator is relatively simple using the GLLAMM program [24] in Stata version 11.0. One needs only to form all within-neighborhood pairs of sampled observations and then treat each pair as its own neighborhood in the program. One sets the individual-level weights to one and the neighborhood-level weights to W_{ijl}. For our BRFSS example, our outcome is ordinal, and we provide SAS code for the data management and Stata code for the GLLAMM optimization in Appendix A. For our BRFSS example, the W_{ijl} are very large. This leads to excessively large values of the pseudolikelihood function during optimization. Because the GLLAMM program uses an absolute rather than a relative convergence criterion, we multiplied the W_{ijl} by a number $(1/100,000)$ small enough so that the values of the pseudolikelihood function were similar to its values without incorporating the weights. Otherwise, convergence would have taken so long as to be impractical.

2.1. *Estimating the sampling distribution*

In the population models (1)–(3), the M neighborhoods are sampled i.i.d., so that the $(Y_i, X_i, b_i), i = 1, \ldots, N_i$ are i.i.d. The first stage of a complex sampling design typically consists of primary sampling units (clusters) nested within primary strata $h = 1, \ldots, H$. Provided that our complex sampling design is such that either (a) each neighborhood is perfectly nested within a primary sampling unit or (b) each primary sampling unit is perfectly nested within a neighborhood and each neighborhood is perfectly nested within a primary stratum, then estimating the sampling distribution of our estimator of β is straightforward. In case (a), we let $c = 1, \ldots, C_h$ index the primary sampling units within primary stratum h. In case (b), we let $c = 1, \ldots, C_h$ index the neighborhoods nested within primary stratum h. Let (Y_c, X_c, b_c) represent the collection of sampled data pertaining to c. In either case (a) or case (b), the finite population sampling followed by the complex sampling design renders the clusters of data $(Y_c, X_c, b_c), c = 1, \ldots, C_h$ independent of one another.

We can therefore estimate the asymptotic sampling distribution of our estimator $\hat{\beta}$ of β by using the usual sandwich estimator of variance for complex survey data [27–29]. Let $U(\beta)$ denote the estimating equation for $\hat{\beta}$ corresponding to the derivative of the weighted composite loglikelihood at (8). In either case (a) or case (b), $U(\beta) = 0$ can be expressed as the sum $\Sigma_{h=1}^{H} \Sigma_{c=1}^{C_h} U_{hc}(\beta) = 0$, and its estimated variance is given by

$$\hat{\mathrm{var}}(\hat{\beta}) = \left[\nabla U(\hat{\beta})\right]^{-1} V(\hat{\beta}) \left[\nabla U(\hat{\beta})^T\right]^{-1}, \tag{9}$$

where $\nabla U(\beta)$ is the gradient of $U(\beta)$ with respect to β, and

$$V(\hat{\beta}) = \Sigma_{h=1}^{H} (C_h/(C_h - 1)) \Sigma_{c=1}^{C_h} \left(U_{hc}(\hat{\beta}) - U_{h.}(\hat{\beta})\right) \left(U_{hc}(\hat{\beta}) - U_{h.}(\hat{\beta})\right)^T, \tag{10}$$

where $U_{h.}(\hat{\beta}) = (1/C_h)\Sigma_{c=1}^{C_h} U_{hc}(\hat{\beta})$.

Our estimator of the asymptotic sampling distribution is design consistent; however, we need for $\Sigma_{h=1}^{H} C_h$ to be reasonably large for the estimator to perform well in practice. By the law of large numbers and the central limit theorem, $\hat{\beta}$ is approximately distributed as multivariate normal with mean β and variance $\hat{\text{var}}(\hat{\beta})$.

Conservative inference is simple using Stata's GLLAMM macro with the cluster and robust options; one specifies the cluster as the variable corresponding to c. Stata's GLLAMM macro does not currently accommodate stratification. Because ignoring primary stratification of the primary clusters can only lead to an increased estimated sampling variability, inference based on assuming all of the primary clusters belong to the same stratum (so that $H = 1$) is conservative. Furthermore, we have observed in practice that accounting for stratification does not typically reduce the estimated sampling variability of our population regression estimators that much. This is to be expected, because stratification is typically used to reduce variability of within-stratum estimators or to reduce the number of unproductive phone numbers when conducting the survey, rather than to reduce sampling variability of population regression estimators.

An alternative approach to estimating the sampling distribution would be to use the bootstrap, resampling the entities $c = 1, \ldots, C_h$ with replacement within their respective strata. For some sampling designs, the C_h are very small, for example, in the public use National Health Interview Survey (NHIS) file, all $C_h = 2$. Adaptations of the bootstrap are available [30, 31] for those complex survey data situations.

In the BRFSS example, the primary sampling units are the individuals, which are perfectly nested within the neighborhoods, that is, the zip codes. However, the zip codes are not nested within strata in the BRFSS example (there are 134 BRFSS strata in Florida, formed by crossing the 67 counties with two telephone density strata.) Therefore, we ignore stratification when estimating the sampling variability, which is conservative as we have just explained and which enables us to use the GLLAMM software for easy estimation and inference.

2.2. *Approximating the pairwise sampling probabilities*

A drawback of our method is that it requires accurate specification of the pairwise sampling probabilities, whereas standard methods for complex survey data only require accurate specification of the individual sampling probabilities. Public use survey datasets contain the individual sampling probabilities, but often there is no information on the pairwise sampling probabilities. Our BRFSS example is special in that the sampling design is relatively simple, which allows us to derive a reasonable method for approximating the pairwise sampling probabilities, as described in Section 4. In [4], we analyzed public use NHIS data and treated household as the neighborhood; this also allowed us to easily compute simple pairwise sampling probabilities, because all individuals in a household were sampled for the variables in our analysis. Therefore, the pairwise sampling probabilities equaled the individual sampling probability for individuals within the same household. In [25], we analyzed in-house NHIS data and treated the secondary sampling unit (SSU) of the survey as the neighborhood. We requested data that allowed us to compute w_i, the inverse probability of selecting the SSU, and $w_{j|i}$, the inverse probability of selecting an individual given that his or her SSU was selected. One could then adopt the theory that the individuals in the finite-population SSU were sampled i.i.d. from a superpopulation, so that one could approximate the inverse pairwise sampling probability for pair (ij, il) as $w_i w_{j|i} w_{l|i}$. One could, as we did in [25], ignore the individual-level poststratification weights, because it is difficult to translate them into accurate pairwise poststratification weights. However, see [6] for an approximate method.

3. Simulation study

First, we present a simulation to show that misspecifying $\psi(X_i, \gamma)$ using model (4) can lead to an inconsistent estimator of β when h is the expit function in model (1), even in the setting of ordinary cluster sampling. We simulated $m = 10,000$ neighborhoods, and we let $N_i = 2$ be constant in i. We simulated a variable u_i independently for each neighborhood, with an $N(0, 0.25^2)$ distribution. We then let

X_{ij} be independent normal random variables given u_i with mean u_i and variance 1. We then simulated $b_i = 5 * \max(X_{ij}, j = 1, 2)$ if $u_i > 0$, and $b_i = 5 * \min(X_{ij}, j = 1, 2)$ if $u_i < 0$. Therefore, $\psi(X_i, \gamma)$, which represents the conditional mean of b_i given X_i, is a complicated function that deviates substantially from model (4). Finally, we generated Y_{ij} according to model (1) with h as the expit function and $\beta = 0.5$. We then estimated β using SAS PROC GLIMMIX in version 9.2 using adaptive quadrature with five initial points and using model (1) in conjunction with model (4). We repeated the simulation five times; the average estimate of β was 0.231 with a range of 0.185–0.264, which was far away from the true value of 0.5. For comparison, we also estimated β using model (1) using PROC GLIMMIX and treating the covariate b_i as known and including it in the fixed effects part of the model (so that the true random effects variance should be zero). Correctly specifying $\psi(X_i, \gamma)$ for this model would require computing $P(u_i > 0|X_i)$ and $P(u_i < 0|X_i)$. The average estimate of β was 0.528 with a range of 0.428–0.583, which illustrates that the truth is recoverable from the data but that the use of model (4), that is, the poor man's approach [20, 21], can lead to substantial bias.

Our second simulation enlisted the simulation settings previously identified by Brumback *et al.* [25] to illustrate substantial bias with the usual GLLAMM approach that optimizes the log pseudolikelihood at (6), but here we show that the new GLLAMM approach that optimizes the weighted composite likelihood at (8) leads to a consistent estimator when the true $\bar{X}_i$ based on the census for neighborhood i is available. First, we simulated the population with $M = 1000$ and $N_i = 1000$ for each i. We let X_{ij} be independent (given u_i) Bernoulli random variables with probability expit(u_i), where u_i are i.i.d. $N(0, 1)$. We let $b_i = -5\bar{X}_i + \delta_i$, with δ_i i.i.d. N(0,1). Finally, we generated Y_{ij} according to model (1) with h the expit function and $\beta = 0.5$. Second, we sampled from the population. Individual observations were identified as concordant (C = 1) or discordant based on whether $Y_{ij} = X_{ij}$ or not. We included observations into the sample with independent probability 0.002 if $C = 1$ and 0.004 if $C = 0$. Within-neighborhood pair sampling weights were proportional to the product of the individual inverse probability weights. For example, if a pair included one concordant and one discordant observation, $W_{ijl} = 2$, whereas if the pair included two discordant observations, $W_{ijl} = 1$, or two concordant observations, $W_{ijl} = 4$. We repeated the simulation 100 times; the average estimate of β was 0.485 (truth = 0.5) with a range of 0.1–0.9 and a standard error of 0.017. The average estimate of the coefficient γ_1 of $\bar{X}_i$ was −5.06 (truth = −5) with a standard error of 0.053, that of the coefficient γ_0 was 0.031 (truth = 0) with a standard error of 0.022, and that of τ^2 was 0.981 (truth = 1) with a standard error of 0.014. This illustrates that the method generates unbiased estimates of all parameters of the model, as theory would indicate.

Finally, we repeated the second simulation but used the estimate $\bar{X}_i^w$ in place of $\bar{X}_i$ in the estimation. The average estimate of β was 0.151 (truth = 0.5) with a range of −0.29 to 0.66 and a standard error of 0.019. The average estimate of the coefficient γ_1 of $\bar{X}_i$ was −1.30 (truth = −5) with a standard error of 0.035, that of the coefficient γ_0 was −1.69 (truth = 0) with a standard error of 0.019, and that of τ^2 was 1.32 (truth = 1) with a standard error of 0.014. This illustrates that the method can generate biased estimates of all parameters in the model when census estimates of $\bar{X}_i$ are unavailable.

In summary, the first simulation highlights the importance of correctly modeling $\psi(X_i, \gamma)$. The second and third simulations show that even when $\psi(X_i, \gamma)$ is correctly modeled, census data on X_i from each sampled neighborhood may be necessary for consistent estimation.

4. Investigating dental health disparities using the BRFSS data

We next apply the new methodology to the 2008 Florida BRFSS survey data to investigate racial/ethnic disparities in oral health care. The Florida BRFSS uses disproportionate stratified sampling, in which only one person per household can be selected. In 2008, the BRFSS sampled 10,874 Floridians, of whom 8376 met our inclusion criteria described in Section 1. Each individual is assigned a sampling weight, representing the inverse probability of being selected into the sample multiplied by a poststratification adjustment, constructed so the joint distribution of race/ethnicity, gender, and age matches that of the most recent state census. For our analysis, we will need to estimate the inverse probability of selecting each possible pair of individuals within a given neighborhood. We are approximating this as the product of the two individual sampling weights. This approximation would be nearly exact if no poststratification adjustment had been made. However, if we assume that the inverse of the poststratification factor represents the conditional probability of responding to the survey given the survey design variables and race/ethnicity, gender, and age, and that the probability of one individual in a pair responding is independent of whether the other responded, then our approximation is valid. We also point out that the probability of pairs within the same household within a neighborhood being selected into the sample

is zero. Strictly speaking, this violates the 'positivity' [32] assumption, that is, that all pairs within a given neighborhood have a positive chance of selection into the sample. However, even if the BRFSS were to allow multiple individuals per household to be selected into the sample, such individuals would represent a negligible fraction of the sample. Thus for all practical purposes, the positivity assumption is satisfied, in that its violation in our context results in negligible bias.

In addition to the race/ethnicity covariate of primary interest, our analysis included the demographic covariates gender and age in the analysis, categorized as explained in Section 1, and also the socioeconomic variables education (high school or greater versus less than high school) and health insurance (covered versus uncovered). To estimate the census proportions for each zip code required by model (4), we merged the Florida BRFSS survey data with census data from Summary Forms 1 and 3 by matching BRFSS zip codes with US Census ZCTAs. Zip code proportions for gender and age were computed using data from adults 18 years and older from Summary Form 1. Zip code proportions for race/ethnicity were computed using data from all individuals from Summary Form 1. Zip code proportions for education were computed using data from the census's sample of individuals from Summary Form 3. For Summary Form 3, the census samples approximately 1 of every 6 individuals; although this does not provide exact census proportions, the sample sizes are large and the measurement error would not be substantial. Census data are not available on insurance status; instead, we used the zip code proportion of households earning less than $25,000. We did not include income as an individual-level covariate in our analysis, because of the large percentage of surveyed individuals who opt not to report household income.

When we included zip code proportions for all categories of the individual-level covariates (e.g., three terms for the three dummy variables for the age covariate), the GLLAMM program failed to converge. This is probably due to the small remaining neighborhood-level variation. We therefore omitted the 'other' category for race/ethnicity (because of small variation about zero), and we selected one category per each of the other individual-level covariates based on a preliminary analysis using SAS PROC

Table I. Estimated odds ratios and 95% confidence intervals for more recent dental cleaning, adjusted for gender, age, education, health insurance, and neighborhood.

Covariate	OR	95% CI
Race/ethnicity		
White non-Hisp	1.0	
Afr Amer non-Hisp	0.65	(0.46, 0.93)
Hispanic	1.61	(1.12, 2.30)
Other	1.18	(0.69, 2.03)
Gender		
Male	1.0	
Female	1.31	(1.05, 1.62)
Age (years)		
18–34	1.0	
35–54	1.46	(1.14, 1.87)
55–64	1.88	(1.40, 2.52)
65+	1.75	(1.25, 2.45)
Education		
< High school	1.0	
⩾ High school	1.72	(1.15, 2.57)
Health insurance		
No insurance	1.0	
Insurance	3.22	(2.42, 4.28)
Neighborhood averages		
Proportion Afr Amer	7.61	(3.35, 17.29)
Proportion Hispanic	1.84	(0.71, 4.76)
Proportion female	0.10	(<0.001, 28.11)
Proportion 65+	3.72	(0.97, 14.25)
Proportion ⩾ high school	12.06	(1.05, 138.79)
Proportion earning less than $25,000	0.05	(0.01, 0.45)

SURVEYLOGISTIC to maximize the composite pseudolikelihood at (8) assuming $\tau^2 = 0$. Specifically, we used the paired dataset (see the construction of `ord.denpairs` in Appendix A) with the pairwise weights for each of the two observations within a pair (see Appendix A for the program). We selected the category with the smallest p-value; none of the covariates had two categories with p-values smaller than 0.10.

Table I presents the results of our final GLLAMM analysis. Our estimate of τ^2 was 9.69×10^{-14} with a standard error of 2.50×10^{-10}, effectively equal to zero. Therefore, the results of our GLLAMM analysis were exactly the same as the results of our preliminary analysis using PROC SURVEYLOGISTIC when we excluded the same zip code proportions from the latter. We observe that the disparity for non-Hispanic African Americans is barely significant (OR $= 0.65$, 95% CI $= (0.46, 0.93)$) and that the unadjusted disparity for Hispanic individuals has reversed (OR $= 1.61$, 95% CI $= (1.12, 2.30)$). These results are qualitatively similar to the results of Brumback *et al.* [1], except for the statistical significance of the disparity for non-Hispanic African Americans. The discrepancy may be due to our inclusion of fewer zip codes in this analysis as a result of the matching with ZCTAs, or it may be due to our operationalization of model (4) for the neighborhood random effect as a function of the census versions of the individual-level covariates.

It is noteworthy that when the variation of the outcome across neighborhoods is almost fully accounted for by the neighborhood census averages, our methods are exactly equivalent to the analytic approach using SAS PROC SURVEYLOGISTIC assuming $\tau^2 = 0$. The latter approach is easier to implement, simply because the GLLAMM program often takes a very long time to converge (our analysis took approximately 24 h). However, for our second simulation using the GLLAMM program (Section 3), we would have obtained different answers using SAS PROC SURVEYLOGISTIC, because $\tau^2 = 1$ in that example (based on 100 simulations, the average estimate of β was 0.42 with a standard error of 0.015; therefore, a t-test would reject the null hypothesis that $\beta = 0.5$, its true value.) In other words, for our simulation, the odds ratios conditional on neighborhood are unequal to the 'population-averaged' odds ratios unconditional on neighborhood. Therefore, in general, we recommend trying both approaches.

5. Discussion

We have presented a method based on GLMMs, which leads to a consistent estimator of the effect of individual-level covariates on an outcome when adjusted for unmeasured neighborhood-level confounding, using complex survey data. Our method differs from previous approaches in that we use a weighted composite likelihood [5] that effectively treats all within-neighborhood pairs of observations as though they were independent of one another and that we use pairwise weights equal to the inverse probability of selecting each pair into the sample. Unlike previous approaches [24, 25], our method is consistent even when the sampling is strongly biased with respect to the log odds ratio, exactly as in [25], and the sample sizes within neighborhood are small; we require, however, that the model $\psi(X_i, \gamma)$ for the association between the random effect and the individual-level covariates is correct and that census data are available on the individual-level covariates included in that model. Our simulation study demonstrated this consistency and also showed that when either of the two conditions is violated, the results are prone to substantial bias. In practice, it is most convenient to specify $\psi(X_i, \gamma)$ using the 'poor man's' model (4) of Neuhaus and Kalbfleisch [20]. This is the approach we took with the BRFSS example, except that we set some components of γ to zero because of the high dimension of X_i. The problem of model choice for $\psi(X_i, \gamma)$ represents an interesting topic for future research. Two obvious avenues to pursue are, first, a Wald test for nested GLMMs with complex survey data and, second, a change-in-estimate criterion for $\hat{\beta}$, in which successive components of $\psi(X_i, \gamma)$ are removed if they do not change the estimate of $\hat{\beta}$ by more than a specified amount (e.g., 10%). However, we suspect that when the sample sizes within neighborhoods are small and the number of components in X_i is large, there will be limited power available for model selection among alternatives for $\psi(X_i, \gamma)$. Therefore, the poor man's approach will probably remain popular, based on its familiarity and easy implementation.

We have shown how to implement our method using the GLLAMM software in Stata version 11.0, with specific details presented in Appendix A. Implementation is relatively simple, but convergence of the GLLAMM procedure can sometimes be impractically slow. We discovered that the convergence criteria of the GLLAMM procedure are linked to absolute differences in the magnitude of the weighted pseudolikelihood, rather than to relative differences. This discovery allowed us to speed up convergence

by normalizing the weights by multiplying them by a small constant, chosen such that the weighted pseudolikelihood had similar magnitude to the unweighted pseudolikelihood. Difficulties with convergence may lead some analysts to prefer methods based on a conditional pseudolikelihood [1,2,4,6] or methods that produce 'population-averaged' estimates rather than 'neighborhood-specific' ones [33]. Nevertheless, the use of GLMMs remains popular in the social sciences and in social epidemiology [34–37], and we therefore hope it is helpful to provide a methodology that avoids the biases of previous approaches.

Appendix A

We used SAS PROC SQL in version 9.2 to create the paired dataset. Specifically,

```
*Make pairs for GLMM;
proc sql;
create table match as
select
one.id, one._finalwt,
one.zipcode,
one._finalwt*two._finalwt as weight_prod,
one.edu as edu_1, two.edu as edu_2,
one.ins as ins_1, two.ins as ins_2,
one.female as female_1, two.female as female_2,
one.age2 as age2_1, two.age2 as age2_2,
one.age3 as age3_1, two.age3 as age3_2,
one.age4 as age4_1, two.age4 as age4_2,
one.black as black_1, two.black as black_2,
one.hisp as hisp_1, two.hisp as hisp_2,
one.other as other_1, two.other as other_2,
one.plt25,
one.p25to50,
one.pblack,
one.phisp,
one.pfemale,
one.pgths,
one.denclean as denclean_1, two.denclean as denclean_2
from denclean one, denclean two
where (one.zipcode=two.zipcode and one.id > two.id);
quit;

data match;
  set match;
  pair_id=_n_;
run;

*Change to long format;
data ord.denpairs;
  set match;
  pairid=1;
  edu=edu_1;
  ins=ins_1;
  female=female_1;
age2=age2_1;
age3=age3_1;
age4=age4_1;
black=black_1;
hisp=hisp_1;
other=other_1;
```

```
denclean=denclean_1;
  output;
  pairid=2;
  edu=edu_2;
  ins=ins_2;
  female=female_2;
  age2=age2_2;
  age3=age3_2;
  age4=age4_2;
  black=black_2;
  hisp=hisp_2;
  other=other_2;
  denclean=denclean_2;
    output;
    drop edu_1 ins_1 female_1 age2_1 age3_1 age4_1 black_1 hisp_1
    other_1 edu_2 ins_2 female_2 age2_2 age3_2 age4_2
    black_2 hisp_2 other_2 denclean_1 denclean_2;
run;

*Preliminary analysis using population averaged approach;
proc surveylogistic data=ord.denpairs;
weight weight_prod;
cluster zipcode;
model denclean = black hisp other edu ins female age2 age3 age4
 plt25 pblack phisp pfemale pgths
page4 / link=clogit;
run;

*Export data to Stata;
PROC EXPORT DATA=ord.denpairs
     OUTFILE= "H:\NSF grant\GLMM\BRFSS Analysis\denpairs.dta"
     DBMS=DTA REPLACE;
RUN;
```

Next we used the GLLAMM program in Stata 11 to maximize the weighted composite likelihood and return 95% confidence intervals.

```
set memory 1G

set more off

log using "stata gllamm output.log", replace
use "denpairs.dta"

#delimit ;

gen pwt2=weight_prod/100000;
gen pwt1=1;

gen denreverse=1;
replace denreverse=2 if denclean==4;
replace denreverse=3 if denclean==3;
replace denreverse=4 if denclean==2;
replace denreverse=5 if denclean==1;

gllamm denreverse black hisp other edu ins female age2 age3 age4
  pblack phisp pgths plt25 page4 pfemale,
  i(pair_id) l(ologit) f(binom) cluster(zipcode) robust adapt nip(20)
```

```
  iterate(30) eform trace;
matrix a=e(b);

gllamm denreverse black hisp other edu ins female age2 age3 age4
  pblack phisp pgths plt25 page4 pfemale,
  i(pair_id) pweight(pwt) l(ologit) f(binom) cluster(zipcode) robust
  from(a) adapt nip(20) iterate(30) eform trace;

#delimit cr

log close
```

Note that, because of the way the GLLAMM program defines the ologit link, we needed to reverse the coding of our ordinal outcome to match the definition in our paper, which is more natural for our application.

Acknowledgements

We would like to thank Youjie Huang, Jamie Forrest, and Melissa Murray from the Florida BRFSS Office for their helpful support. We would also like to acknowledge the support of the National Science Foundation, the USDA/National Agricultural Statistics, the Department of Education/National Center for Educational Statistics, the Social Security Administration, and the Department of Agriculture/Economic Research Service through grant NSF SES-1115618.

References

1. Brumback BA, Dailey AB, Zheng HW. Adjusting for confounding by neighborhood using a proportional odds model and complex survey data. *American Journal of Epidemiology* 2012; **175**(11):1133–1141.
2. Brumback BA, Cai Z, He Z, Zheng HW, Dailey AB. Conditional pseudolikelihood methods for clustered ordinal, multinomial, or count outcomes with complex survey data. *Statistics in Medicine*. DOI: 10.1002/sim.5625.
3. Neyman J, Scott EL. Consistent estimates based on partially consistent observations. *Econometrica* 1948; **16**:1–32.
4. Brumback BA, He Z. Adjusting for confounding by neighborhood using complex survey data. *Statistics in Medicine* 2011; **30**:965–972.
5. Varin C, Reid N, Firth D. An overview of composite likelihood methods. *Statistica Sinica* 2011; **21**:5–42.
6. Graubard BI, Korn E. Conditional logistic regression with survey data. *Statistics in Biopharmaceutical Research* 2011; **3**:398–408.
7. Liang K. Extended Mantel-Haenszel estimating procedure for multivariate logistic regression models. *Biometrics* 1987; **43**:289–299.
8. Breslow N, Day N, Halvorsen K, Prentice R, Sabai C. Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology* 1978; **108**:299–307.
9. Agresti A. *Categorical Data Analysis*, Second Edition. John Wiley & Sons: Hoboken, 2002.
10. Cheng NF, Han PZ, Gansky SA. Methods and software for estimating health disparities: the case of children's oral health. *American Journal of Epidemiology* 2008; **168**(8):906–914.
11. Tellez M, Sohn W, Burt BA, Ismail AI. Assessment of the relationship between neighborhood characteristics and dental caries severity among low-income African-Americans: a multilevel approach. *Journal of Public Health Dentistry* 2006; **66**(1):30–36.
12. Ahn S, Burdine JN, Smith ML, Ory MG, Phillips CD. Residential rurality and oral health disparities: influences of contextual and individual factors. *Journal of Primary Prevention* 2011; **32**(1):29–41.
13. Chattopadhyay A. Oral health disparities in the United States. *Dental Clinics of North America* 2008; **52**(2):297–318.
14. Sabbah W, Tsakos G, Chandola T, Sheiham A, Watt RG. Social gradients in oral and general health. *Journal of Dental Research* 2007; **86**(10):992–996.
15. Sabbah W, Tsakos G, Sheiham A, Watt RG. The effects of income and education on ethnic differences in oral health: a study in US adults. *Journal of Epidemiology and Community Health* 2009; **63**(7):516–520.
16. Sanders AE, Turrell G, Slade GD. Affluent neighborhoods reduce excess risk of tooth loss among the poor. *Journal of Dental Research* 2008; **87**(10):969–973.
17. Bramlett MD, Soobader M, Fisher-Owens SA, Weintraub JA, Gansky SA, Platt LJ, Newacheck PW. Assessing a multilevel model of young children's oral health with national survey data. *Community Dentistry and Oral Epidemiology* 2010; **38**(4):287–298.
18. Turrell G, Sanders AE, Slade GD, Spencer AJ, Marcenes W. The independent contribution of neighborhood disadvantage and individual-level socioeconomic position to self-reported oral health: a multilevel analysis. *Community Dentistry and Oral Epidemiology* 2007; **35**(3):195–206.
19. Wu B, Plassman BL, Liang J, Remle J, Remle RC, Bai L, Crout RJ. Differences in self-reported oral health among community-dwelling black, Hispanic, and white elders. *Journal of Aging and Health* 2011; **23**(2):267–288.

20. Neuhaus JM, Kalbfleisch JD. Between- and within-cluster covariate effects in the analysis of clustered data. *Biometrics* 1998; **54**:638–645.
21. Neuhaus JM, McCulloch CE. Separating between- and within-cluster covariate effects by using conditional and partitioning methods. *Journal of the Royal Statistical Society, Series B* 2006; **68**:859–872.
22. Goetgeluk S, Vansteelandt, S. Conditional generalized estimating equations for the analysis of clustered and longitudinal data. *Biometrics* 2008; **64**:772–780.
23. Brumback BA, Dailey AB, Brumback LC, Livingston MD, He Z. Adjusting for confounding by cluster using generalized linear models. *Statistics and Probability Letters* 2010; **80**:1650–1654.
24. Rabe-Hesketh S, Skrondal A. Multilevel modeling of complex survey data. *Journal of the Royal Statistical Society, Series A* 2006; **169**:805–827.
25. Brumback BA, Dailey AB, He Z, Brumback LC, Livingston MD. Efforts to adjust for confounding by neighborhood using complex survey data. *Statistics in Medicine* 2010; **29**(18):1890–1899.
26. Korn EL, Graubard BI. Estimating variance components by using survey data. *Journal of the Royal Statistical Society, Series B* 2003; **65**:175–190.
27. Binder DA. On the variances of asymptotically normal estimators from complex surveys. *International Statistical Review* 1983; **51**:279–292.
28. Skinner CJ, Holt D, Smith TMF. *Analysis of Complex Surveys*. John Wiley & Sons: Sussex, 1989.
29. Korn EL, Graubard BI. *Analysis of Health Surveys*. John Wiley & Sons: New York, 1999.
30. McCarthy PJ, Snowden CB. The bootstrap and finite population sampling. *Vital Health Stat 2* 1985; **95**:1–23.
31. Shao J. Impact of the bootstrap on sample surveys. *Statistical Science* 2003; **18**(2):191–198.
32. Cole SR, Hernan MA. Constructing inverse probability weights for marginal structural models. *American Journal of Epidemiology* 2008; **168**:656–664.
33. Zeger SL, Liang K-Y. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 1986; **42**:121–130.
34. Raudenbush SW, Bryk AS. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications: Thousand Oaks, CA, 2002.
35. Allison PD. *Fixed Effects Regression Models*. SAGE: Los Angeles, CA, 2009.
36. Dailey AB, Brumback BA, Livingston MD, Jones BA, Curbow BA, Xu X. Area-level socioeconomic position and repeat mammography screening use: results from 2005 National Health Interview Survey. *Cancer Epidemiology, Biomarkers, & Prevention* 2011; **20**(11):2331–2344.
37. Morenoff JD, House JS, Hansen BB, *et al.* Understanding social disparities in hypertension prevalence, awareness, treatment, and control: the role of neighborhood context. *Social Science and Medicine* 2007; **65**(9):1853–1866.