Catalogue no. 11-522-XIE

Statistics Canada International Symposium Series - Proceedings

Symposium 2006 : Methodological Issues in Measuring Population Health

2006

Statistics Canada **Statistique Canada**

Canada

Proceedings of Statistics Canada Symposium 2006
Methodological Issues in Measuring Population Health

Mixed Linear Nonlinear Aggregate Level Models for Small Area Estimation from Surveys of Binary Counts

A.C. Singh[1] and F. Verret[2]

Abstract

We propose an aggregate level generalized linear model with additive random components (GLMARC) for binary count data from surveys. It has both linear (for random effects) and nonlinear (for fixed effects) parts in modeling the mean function and hence belongs to a class termed as mixed linear non-linear models. The model allows for linear mixed model (LMM)-type approach to small area estimation (SAE) somewhat similar to the well-known Fay-Herriot (1979) method and thus takes full account of the sampling design. Unlike the alternative hierarchical Bayes (HB) approach of You and Rao (2002), the proposed method gives rise to easily interpretable SAEs and frequentist diagnostics as well as self-benchmarking to reliable large area direct estimates. The usual LMM methodology is not appropriate for the problem with count data because of lack of range restrictions on the mean function and the possibility of unrealistic (e.g. zero in the context of SAE) estimates of the variance component as the model does not allow the random effect part of the conditional mean function to depend on the marginal mean. The proposed method is an improvement of the earlier method due to Vonesh and Carter (1992) which also uses mixed linear nonlinear models but the variance-mean relationship was not accounted for although typically done via range restrictions on the random effect. Also the implications of survey design were not considered as well as the estimation of random effects. In our application for SAE, however, it is important to obtain suitable estimates of both fixed and random effects. It may be noted that unlike the generalized linear mixed model (GLMM), GLMARC like LMM offers considerable simplicity in model fitting. This was made possible by replacing the original fixed and random effects of GLMM with a new set of parameters of GLMARC with quite a different interpretation as the random effect is no longer inside the nonlinear predictor function. However, this is of no consequence for SAE because the small area parameters correspond to the overall conditional means and not on individual model parameters. We propose a method of iterative BLUP for parameters estimation which allows for self-benchmarking after a suitable model enlargement. The problem of small areas with small or no sample sizes or zero direct estimates is addressed by collapsing domains only for the stage of parameter estimation. Application to the 2000-01 Canadian Community Health Survey for estimation of the proportion of daily smokers in subpopulations defined by provincial health regions by age-sex groups is presented as an illustration.

KEY WORDS: Domain collapsing; Range Restrictions; Self-benchmarking ;Variance-mean relationship

1. Introduction

The problem of aggregate level nonlinear mixed models for small area estimation considered in this paper arose in the context of producing estimates of number of daily smokers in health regions by age-sex groups for the province of Prince Edward Island based on the 2000-2001 Canadian Community Health Survey (CCHS). Denote by $t_{y,d}$ the direct estimate of the small area total $T_{y,d}$ of the outcome variable y for the domain d representing the subpopulation (r,a,g) where r is the r^{th} health region, a is the a^{th} age group (12-19, 20-29, 30-44, 45-64, and 65+) and g denotes the gender (male or female). In practice, for survey data, often a linear mixed model (LMM) at the aggregate level is used which can take full account of the underlying sampling design typically nonignorable for the model, see e.g. the well known method of Fay and Herriot (1979). Here, the term 'aggregate level' is used to signify that the covariates in the model are at the domain level. For example, under a subgroup common mean model, the mean $\mu_{y,d}$ for the d^{lt} domain is modeled as

$$\mu_{y,d} = \nu_{y,d} + \eta_d, \ \nu_{y,d} = A'_{x,d}\beta \qquad (1.1)$$

where $A_{x,d}$ is a D-vector of indicator covariates $1_{d,ag}, d = 1,...,D;$ $1_{d,ag} = 1$ if d belongs to the age-sex subgroup (a,g) and 0 otherwise, and the fixed parameter β is of dimension q, denoting the total number (10 in our example)

[1] Statistical Research and Innovation Division, Statistics Canada, Ottawa K1A 0T6
[2] Household Survey Methods Division, Statistics Canada, Ottawa K1A 0T6

of age-sex subgroups. The model error η_d is the random effect assumed to be $iid\ \ N(0, \sigma_\eta^2)$. One may also have the covariates such as average number of hospital admissions from domain d with asthma as the main diagnosis. Denoting the totals $N_d A_{x,d}$ *by* $T_{x,d}$, where N_d is the domain d subpopulation count, we can write the model for $t_{y,d}, d = 1, ..., D$, as

$$\begin{aligned} t_{y,d} &= N_d (A'_{x,d}\beta + \eta_d) + e_{y,d} \\ &= T'_{x,d}\beta + N_d \eta_d + e_{y,d} \end{aligned} \qquad (1.2)$$

where $e_{y,d}$ is the observation or sampling error with variance V_d assumed to be approximately known. It is assumed that the sample size n_d from domain d is not zero or very small such that $t_{y,d}$ is not zero. In practice, domains could be collapsed based on similarity with respect to $A_{x,d}$ to avoid this problem as well as for smoothing of the covariance matrix $\widetilde{V}$ of the collapsed vector t_y of dimension $\widetilde{D}$ in order to treat it as known, see Singh (2006) for more details.

Although there are several practical benefits in using LMM for the above problem such as having a semiparametric approach for BLUP with only first two moment assumptions (which allows for over dispersion effect in the interest of robustification), self-benchmarking of SAEs to reliable direct estimates for large areas in the interest of protection against possible model breakdowns, and a versatile set of diagnostics based on innovations from the use of Kalman Filter in deriving BLUPs (cf: Singh, 2006), there are a few major limitations. First, the model (1.1) for the small area means doesn't impose any range restriction on the parameters $\mu_{y,d}$ which is not reasonable for discrete data such as counts or proportions. In our example with proportions of individuals in a local health area diagnosed with diabetes, the parameter $\mu_{y,d}$ is necessarily between 0 and 1. For the above subgroup common mean model (1.1), $A'_{x,d}\beta$ lies in (0, 1), but with more general covariates it need not be, and so the BLUP (lying between the direct estimate $N_d^{-1} t_{y,d}$ and the indirect or synthetic estimate $A'_{x,d}\hat{\beta}$ at least for the case of diagonal V), need not be in the admissible range. This renders the optimality of BLUP tenuous in this LMM application. Moreover, the exchangeability assumption of common variance σ_η^2 for model errors η_d in (1.1) may be an over-simplifying approximation because for discrete data, the range of possible values of η_d depends on the marginal mean $v_{y,d}$ and so its variance should depend on the mean $v_{y,d}$, for example, η_d could be of the form $f(v_{y,d})\zeta_d$ for some smooth function $f(\cdot)$ and ζ_d being $iid\ \ N(0, \sigma_\xi^2)$. Without such a generalization of LMM, the usual REML-type estimate of σ_η^2 may turn out to be zero which is clearly inappropriate as the model (1.1) is known to be imperfect, i.e., $\sigma_\eta^2 > 0$.

A natural solution to the above problem is to consider generalized LMM (GLMM) for SAE instead of the LMM approach. The frequentist approaches for GLMM tend to focus on estimation of fixed first and second order parameters and not on random effects which are of course needed for SAE. A good frequentist solution to the problem of estimation of random effects for GLMM is in general quite difficult; see Jiang and Lahiri (2006) for a recent review. On the other hand, hierarchical Bayes (HB) methods with MCMC can be used for estimation of fixed and random parameters, see Rao (2003, Ch. 10). With survey data and nonignorable designs, the problem of SAE with unit level GLMM is even more difficult and is a topic of current research in the field. However, for aggregate level GLMM, an HB extension of Fay-Herriot methodology was proposed by You and Rao (2002). Although this is a useful extension, it does not seem possible to have the desirable feature of self-benchmarking in the HB framework. Moreover, the resulting SAEs have neither a simple practical interpretation nor the model diagnostics. It would therefore be useful to have LMM-type approach to aggregate level GLMM for SAE.

The purpose of this paper is to propose a new model based on GLM with additive random components (GLMARC) where the random part of the mean is linear and additive to the fixed nonlinear part of the mean function. The GLMARC models can be viewed as a special case of mixed linear nonlinear (MLNL) models. Earlier approaches to

MLNL models used the framework of small-sigma asymptotics (i.e. σ_η being small) which may not be realistic for SAE applications. Also, these earlier approaches were only concerned primarily with estimation of fixed parameters in the mean function and variance components and not with random effects.

In this paper, we propose a new MLNL model where both β and $\eta-$parameters are replaced to avoid the small sigma asymptotics assumption. This is reasonable for SAE because the original model parameters β and η are not of direct interest. In Section 2, we review the existing approaches to MLNL models based on small-sigma asymptotics as well as provide a motivation of the proposed method. In Section 3, the proposed method of GLMARC is described and a method of iterative BLUP is presented for parameter estimation. It is also shown how the covariates in GLMARC can be extended to allow for self-benchmarking analogous to the case of LMM. Innovation-based diagnostics are presented assuming that the direct estimates $t_{y,d}$ are approximately normal, after domain collapsing, if necessary. Section 4 presents an application of the proposed method to the data from 2000-01 CCHS. Finally, concluding remarks are given in Section 5.

2. Alternative MLNL models without Small Sigma Asymptotics

The mixed linear non-linear (MLNL) models were considered earlier by several authors where estimation of only fixed parameters in the mean function and variance components were of concern. Estimation of random effects was not of interest unlike the SAE problem considered here. For example, McCullagh and Nelder (1989, Ch 14) used first order Taylor linearization under small-sigma asymptotics (i.e., $\sigma_\eta^2 \approx 0$) to transform approximately the nonlinear predictor function of GLMM into an MLNL model form for simplified computations. However, they did caution about the caveat of small-sigma assumption which is not likely to be tenable, see also Drum and McCullagh (1993). Sutradhar and Rao (1996) used second order Taylor under small-sigma asymptotics to capture random interaction terms in cases where there are more than one main random factor. The main limitations of the above small-sigma based approaches are that the variance component in the context of SAE is generally not small, and that no range restrictions are placed on the random effect, and hence on the variance component which may lead to inadmissible error covariance structure in the sense of non-positive definiteness. Clearly an alternative framework for MLNL modeling without relying on small-sigma is needed. To this end, Vonesh and Carter (1992) proposed directly an MLNL model which in the case of binary data considered in this paper can be written as:

$$t_{y,d} = N_d[v_{y,d}(\alpha)+\zeta_d^*]+e_{y,d}, \qquad e_y \sim (0,V), \quad \zeta_d^* \sim^{iid} N(0,\sigma^2_{\zeta^*}) \tag{2.1}$$

and $$\text{logit } v_{y,d}(\alpha) = A'_{x,d}\alpha$$

following the notation for aggregate level LMM introduced in Section 1. Notice that the fixed regression parameters β are replaced by α signifying that in (2.1) we are modeling the marginal mean $v_{y,d}$ instead of the conditional mean $\mu_{y,d}$ given η_d which in the logit case is expressed as

$$\text{logit } \mu_{y,d} = A'_{x,d}\beta + \eta_d \tag{2.2}$$

Moreover, the new random effect ζ_d^* is now outside the nonlinear predictor function as an additive term. The model (2.1) is basically a new model deemed to be valid for the problem at hand instead of the model (2.2) which does not require small-sigma assumption. Moreover, no equivalence between old parameters (β,η) and new ones (α,ζ) is stipulated. However, a major limitation of the model (2.1) is that no range restrictions are placed on the random effect ζ_d^* and so the variance component $\sigma^2_{\zeta^*}$ does not account for the variance- mean relationship. The conditional mean given ζ_d^* under (2.1) is $\mu_{y,d} = (v_{y,d}(\alpha)+\zeta_d^*)$ which must lie between 0 and 1. So, the range restrictions on ζ_d^* take the form

$$v_{y,d}(\alpha) \le \zeta_d^* \le 1 - v_{y,d}(\alpha) \tag{2.3}$$

which implies that $\sigma^2_{\zeta^*}$ should depend on $v_{y,d}(\alpha)$ and not be constant for all d. For the GLMARC proposed in the next section, we take ζ_d^* to be of the form of $f(v_{y,d})\zeta_d$ where the functional form of $f(\cdot)$ is motivated from the

first order Taylor expansion of GLMM under (2.2) around $\eta_d = 0$ as

$$\mu_{y,d} := \text{inv logit}\,(A'_{x,d}\beta + \eta_d) = \mu_{y,d}(\beta, \eta_d = 0) + \mu_{y,d}(\beta, \eta_d^*)(1 - \mu_{y,d}(\beta, \eta_d^*))\eta_d \qquad (2.4)$$

where η_d^* is somewhere between 0 and η_d. We need to express (2.4) in an approximately equivalent form involving the marginal mean $v_d(\alpha)$ and the new random effect ζ_d because η_d^* is not known. The expression (2.4) suggests that the working form of the function $f(v_{y,d})$ can be taken as $v_{y,d}(1 - v_{y,d})$ and then correct the variance component σ_η^2 by introducing the new random effect ζ_d with the new variance σ_ζ^2. In other words, for the model proposed in the next section, we define

$$\mu_{y,d} = v_{y,d}(\alpha) + v_{y,d}(\alpha)(1 - v_{y,d}(\alpha))\zeta_d \qquad (2.5)$$

where the random effect ζ_d must satisfy the range restrictions

$$-(1 - v_{y,d})^{-1} \le \zeta_d \le v_{y,d}^{-1} \qquad (2.6)$$

The above restrictions make the variance of ζ_d depend on d via $v_{y,d}$ which is rather awkward despite the factor $f(v_{y,d})$. A way out is to make the stronger assumption that $-1 < \zeta_d < 1$ for all d which in fact implies that (2.6) holds for all $v_{y,d}$'s and that $0 \le \sigma_\zeta^2 < 1$ (Note that in our experience the assumption $|\zeta_d| < 1$ is not serious in practice.) In view of the range restrictions, we will assume for simplicity that ζ_d's are approximately i.i.d. $N(0, \sigma_\zeta^2)$ because they are restricted to lie in (-1, 1).

We can justify the existence of above random effects in several ways. Typically with binary data the beta-binomial distribution is used for count data. So if we assume that $\mu_{y,d} \sim \text{Beta}(a_d, b_d)$ on the interval (0, 1) for some $a_d, b_d > 0$, we have

$$E(\mu_{y,d}) = \frac{a_d}{a_d + b_d}, \quad V(\mu_{y,d}) = \frac{a_d b_d}{(a_d + b_d)^2 (a_d + b_d + 1)} \qquad (2.7)$$

It follows that for general $f(v_{y,d})$,

$$a_d = v_{y,d}[v_{y,d}(1 - v_{y,d}) f^{-2}(v_{y,d})\sigma_\zeta^{-2} - 1] \qquad (2.8a)$$

and

$$b_d = (1 - v_{y,d})[v_{y,d}(1 - v_{y,d}) f^{-2}(v_{y,d})\sigma_\zeta^{-2} - 1] \qquad (2.8b)$$

Notice that with the choice of $f(v_{y,d})$ as $v_{y,d}(1 - v_{y,d})$, the condition $\sigma_\zeta^2 < 1$ is sufficient to ensure that both a_d and $b_d > 0$. This justifies the formulation (2.5). Alternatively if we assume that $[v_{y,d}(1 - v_{y,d})]^{-1}(\mu_{y,d} - v_{y,d})$ is truncated normal on (-1, 1) with mean 0 and variance σ_ζ^2, then we get ζ_d approximately $N(0, \sigma_\zeta^2)$. This is the assumption we make for the proposed model which is useful and convenient for model diagnostics under the added assumption of approximate normality of the observation error $e_{y,d}$. However, for the BLUP type estimation followed by their MSE estimation, the above normality assumptions are not necessary.

3. Iterative BLUP for GLMARC: The Proposed Method

3.1 The Model

The proposed method of SAE for binary count data is based on the following GLMARC model mentioned earlier in the previous section. For $d = 1, ..., D$,

$$t_{y,d} = N_d[v_d(\alpha) + v_d(\alpha)(1 - v_d(\alpha 1)\zeta_d] + e_{y,d} \qquad (3.1)$$

where $\text{logit}\, v_d(\alpha) = A'_{x,d}\alpha$, $e_y \sim N(0, \sigma_0^2 V)$, $\zeta_d \sim^{iid} N(0, \sigma_0^2 \tau_\zeta^2)$, $\sigma_\zeta^2 < 1$ and σ_0^2 is the over/under dispersion parameter representing adjustment to the covariance structure due to possibly omitted covariates or random effects. It is possible that for some domains, the realized sample sizes n_d are zero or small or the products $n_d N_d^{-1} t_{y,d}$ are

not large enough for approximate normality of $t_{y,d}$. To overcome this problem, we first collapse similar domains based on closeness in terms of covariates $A_{x,d}$'s so that the number of domains is reduced to $\widetilde{D}, \widetilde{D} < D$, and then estimate fixed model parameters α, σ_ζ^2 and σ_0^2. Once these model parameters are estimated, then the d-specific random effects corresponding to uncollapsed domains are estimated, thus preserving the identity of domain-specific small area parameters, see e.g. Singh (2006). The condition of $|\zeta_d| < 1$ to ensure that the small area parameters $\mu_{y,d}$'s satisfy range restrictions is not imposed initially during estimation via BLUP, but once $\hat{\zeta}_d$'s are obtained, they are truncated, if necessary, to satisfy this condition. Such truncation may not, in general, be needed in practice and is not expected to introduce any serious compromise in BLUP optimality. Moreover, the usual MSE estimates for BLUP can still be used in a conservative sense.

3.2 Parameter Estimation via iterative BLUP (IBLUP)

Given the over dispersion σ_0^2 , the proposed method of IBLUP can be used to estimate α , ζ_d and σ_ζ^2 . In fact, it turns out that, due to σ_0^2 being a multiplicative factor for V and σ_ζ^2, the estimates of α and ζ do not depend on σ_0^2, but their variances and the estimate of σ_ζ^2 do. If the model were LMM, both ($\sigma_0^2, \sigma_\zeta^2$) can be estimated using REML which can be simplified further by using the profile likelihood for estimation of σ_ζ^2. The reason for this is that REML of σ_0^2 can be obtained in a closed form as an average of standardized residuals which are themselves functions of σ_ζ^2, see e.g., Harvey (1989, Ch. 4) for an analogous approach in the Kalman Filter context. For GLMARC, we can use the above idea in an iterative manner. In the following, for simplicity, we set σ_0^2 = 1, and describe in steps how BLUP is modified to IBLUP for the GLMARC case. Here the fixed parameters α, σ_ζ^2 are first estimated. Given α, σ_ζ^2, BLUP estimates of $\zeta_d^{'s}$ are then obtained.

Step 0: Computation of the initial estimate $\alpha^{(0)}$

Let $\sigma_\zeta^{2(0)} = 0$, and find a consistent estimate $\alpha^{(0)}$ using the standard iteratively reweighted least squares (IRLS) of GLM. More specifically, at iteration r , the mean function $v_d(\alpha)$ is linearized at $\alpha = \alpha_r$ as

$$v_d(\alpha) \approx v_d(\alpha_r) + v_d(\alpha_r)(1 - v_d(\alpha_r))A'_{x,d}(\alpha - \alpha_r) \tag{3.2}$$

which gives the approximate linear model for $t_{y,d}$ as

$$t_{y,d} \approx b_{d0(r)}N_d + b_{d1(r)}T'_{x,d}\alpha + e_{y,d} \tag{3.3}$$

where $b_{d0(r)} = v_d(\alpha_r) - b_{d1(r)}A'_{x,d}\alpha_r$, $b_{d1(r)} = v_d(\alpha_r)(1 - v_d(\alpha_r))$.

At $r = 0$, we do not need α_0 to estimate $v_d(\alpha)$. Instead, $v_d(\alpha_0)$ is estimated directly by the observed proportion with the continuity correction as $(n_d N_d^{-1} t_{y,d} + .5)/(n_d + 1)$, and $A'_{x,d}\alpha_0$ is easily computed as log of $v_d(\alpha_0)/(1 - v_d(\alpha_0))$. Now with the adjusted dependent variable $t_{y,d}$ - $b_{d0(r)}$, IRLS is used until convergence to obtain $\alpha^{(0)}$.

Step 1 Computation of $\sigma_\zeta^{2(1)}$

With $\alpha^{(0)}$ from step 0 , consider the linearized model

$$t_{y,d}^{*(0)} \equiv t_{y,d} - N_d b_{d0}(\alpha^{(0)}) \approx b_{d1}(\alpha^{(0)})T'_{x,d}\alpha + e_{y,d}^{*(0)} \tag{3.4}$$

where $e_{y,d}^{*(0)} = b_{d1}(\alpha^{(0)})N_d\zeta_d + e_{y,d}$, $vector\ e_y^{*(0)} \sim N(0, W(\alpha^{(0)}, \sigma_\zeta^2))$

$W(\alpha^{(0)},\sigma_\varsigma^2)=\Gamma(\alpha^{(0)},\sigma_\varsigma^2)+V,\quad \Gamma(\alpha^{(0)})=\text{diag}\left\{b_{d1}^2(\alpha^{(0)})N_d^2\right\}\sigma_\varsigma^2$

Now, use REML (see e.g., Rao (2003, Ch. 6) to obtain the estimate $\sigma_\zeta^{2(1)}$. To ensure $0<\sigma_\zeta^2<1, \sigma_\zeta^2$ can be first transformed via logit before the REML estimation.

Step II Computation of $\alpha^{(1)}$

With $\sigma_\zeta^{2(1)}$ from step 1, compute $\alpha^{(1)}$ by WLS from the above linearized model (3.4) except that $W(\alpha^{(0)},\sigma_\varsigma^2)$ is replaced by $W(\alpha^{(0)},\sigma_\zeta^{2(1)}) = \Gamma(\alpha^{(0)},\sigma_\zeta^{2(1)})+V$, $\Gamma(\alpha^{(0)},\sigma_\zeta^{2(1)})=\sigma_\zeta^{2(1)}\text{diag}\ \left\{N_d^2 b_{d1}^2(\alpha^{(0)})\right\}$.

The above steps I & II are iterated until convergence to obtain $\hat{\sigma}_\zeta^2$ and $\hat{\alpha}$.

Step III Estimation of $\zeta_d, d=1,...,D$

Given $\hat{\sigma}_\zeta^2$ and $\hat{\alpha}$, and the linearized model (3.4) with $W=\Gamma(\hat{\alpha},\hat{\sigma}_\zeta^2)+V = \hat{\sigma}_\zeta^2\,\text{diag}\,\{N_d^2 b_{d1}^2(\hat{\alpha})\}+V$, we have the EBLUP of ζ_d in the case of diagonal V as

$$N_d b_{d1}(\hat{\alpha})\hat{\zeta}_d=\Gamma_d W_d^{-1}(t_{y,d}-N_d v_{y,d}(\hat{\alpha})) \tag{3.5}$$

where Γ_d and W_d are evaluated at $\alpha=\hat{\alpha}$, $\sigma_\zeta^2=\hat{\sigma}_\zeta^2$. In general, for nondiagonal V, we have

$$\text{diag}\{N_d b_{d1}\}\hat{\zeta}=\Gamma W^{-1}(t_y\text{-diag}\{N_d\}v_y(\hat{\alpha})) \tag{3.6}$$

Now $\hat{\zeta}_d$ from (3.6) can also be expressed as

$$\begin{aligned}N_d b_{d1}(\hat{\alpha})\hat{\zeta}_d &= \Gamma_{dd}W_{dd.\bar{d}}^{-1}\left[(t_{y,d}-N_d v_{y,d}(\hat{\alpha}))-W_{d\bar{d}}W_{\bar{d}\bar{d}}^{-1}(t_{y,\bar{d}}-\text{diag}\{N_{d'},d'\neq d\}v_{\bar{d}}(\hat{\alpha}))\right]\\ &= \Gamma_{dd}W_{dd}^{-1}\left[W_{dd}W_{dd.\bar{d}}^{-1}(t_{y,d}-N_d v_{y,d}(\hat{\alpha}))-W_{d\bar{d}}W_{\bar{d}\bar{d}}^{-1}(t_{y,\bar{d}}-\text{diag}\{N_{d'},d'\neq d\}v_{y,\bar{d}}(\hat{\alpha}))\right]\end{aligned} \tag{3.7}$$

Where $\bar{d}$ is the set of all domains except the domain d, and $W_{dd}, W_{d\bar{d}}, W_{\bar{d}\bar{d}}$ are defined by partitioning the matrix W as

$$W=\begin{pmatrix}W_{dd} & W_{d\bar{d}}\\ W_{\bar{d}d} & W_{\bar{d}\bar{d}}\end{pmatrix},\quad W_{dd.\bar{d}}=W_{dd}-W_{d\bar{d}}W_{\bar{d}\bar{d}}^{-1}w_{\bar{d}d} \tag{3.8}$$

It follows that the SAE for $T_{y,d}$ is given by

$$t_{y,d}^{sae}=N_d[v_d(\hat{\alpha})+v_d(\hat{\alpha})(1-v_d(\hat{\alpha})\hat{\zeta}_d] \tag{3.9}$$

because $N_d b_{d0}(\hat{\alpha})+b_{d1}(\hat{\alpha})T_{x,d}'\hat{\alpha}=N_d v_d(\hat{\alpha})$.

So, for V diagonal, $N_d^{-1}t_{y,d}^{\text{sae}}$ is necessarily between 0 and 1 as it is a convex combination of $N_d^{-1}t_{y,d}$ and $v_d(\hat{\alpha})$. However, for nondiagonal V, it need not be so, but is likely to be in the range (0, 1) because the first term in (3.7) is expected to be dominant. As mentioned earlier, we can truncate $\hat{\zeta}_d$ to ± 1 if it lies outside (-1, 1).

Finally, the MSE of t_y^{sae} can be easily obtained with second-order approximation as shown in Rao (2003, p. 155) because after linearization, the GLMARC model reduces to a Fay-Herriot Type model. Now so far, the overdispersion parameter σ_0^2 was set to 1 for simplicity. To estimate $(\sigma_0^2,\sigma_\zeta^2)$ jointly, we can first estimate σ_ζ^2 via REML on the profile likelihood after replacing σ_0^2 by its MLE given σ_ζ^2 from the REML likelihood. It turns out that $\hat{\sigma}_0^2(\sigma_\zeta^2)$ has a simple form as an average of squared standardized residuals; see e.g. Harvey (1989 Ch. 4) in the context of Kalman filtering for BLUP estimation.

3.3 Benchmarking of SAEs under GLMARC

For LMM, it is possible to get exact benchmarking by enlarging the model, see e.g. Singh (2006). It would be useful to review this before we consider benchmarking for GLMARC. Suppose there is only one global benchmark, i.e.,

the sum of SAEs for all domain totals should add up to the sum of the direct total estimates for all the domains in the benchmark subgroup, B(say). Here the benchmark subgroup B consists of all domains d ,$d = 1,...,D$. Now, for LMM, we consider the enlarged model

$$t_y = 1_B \beta_B + T(x)\beta + \text{diag}\{N_d\}\eta + e_y = T^*(x)\beta^* + T(c)\eta + e_y \qquad (3.10)$$

where 1_B is a D-vector with elements 1 or 0 depending on whether the d^{th} domain is in B or not, and $T^*(x)$ is the enlarged covariate matrix, β^* is the corresponding enlarged β -vector and $T(c)$ is simply $\text{diag}\{N_d\}$.

With one global benchmark, 1_B is just a vector of 1's. The covariate $V1_B$ involving the observation error covariance matrix is introduced with the extra regression parameter β_B as an artifact to induce benchmarking.

To see this, note that for LMM

$$t_y^{\text{sae}} = T^*(x)\hat{\beta}^* + \Gamma W^{-1}(t_y - T^*(x)\hat{\beta}^*) \qquad (3.11)$$

where $\Gamma = T(c)T'(c)\sigma_\eta^2 = \text{diag}\{N_d^2\}\sigma_\eta^2$, $W = \Gamma + V$, and $\hat{\beta}^*$ is the WLS estimate of β under (3.10).

Rewriting (3.11) as

$$\begin{aligned} t_y^{\text{sae}} &= t_y - (I - \Gamma W^{-1})(t_y - T^*(x)\hat{\beta}^*) \\ &= t_y - VW^{-1}(t_y - T^*(x)\hat{\beta}^*) \end{aligned} \qquad (3.12)$$

we have the desired benchmarking, i.e.,

$$1_B'(t_y - t_y^{\text{sce}}) = 1_B' VW^{-1}(t_y - T^*(x)\hat{\beta}^*) = 0 \qquad (3.13)$$

because, (3.13) is one of the WLS equations corresponding to the new covariate $V1_B$.

Now, for GLMARC, we introduce a new covariate vector $\text{diag}\{N_d^{-1}b_{d(1)}^{-1}(\tilde{\alpha}^*)\}V1_B$ and the corresponding regression parameter α_B, where $\tilde{\alpha}^*$ is the true unknown value of α^*, and consider the enlarged model

$$t_y = \text{diag}\{N_d\}v(\alpha^*) + \text{diag}\{N_d v_d(\alpha^*)(1 - v_d(\alpha^*))\}\zeta + e_y \qquad (3.14)$$

where $v(\alpha^*)$ =invlogit$(A^*(x)'\alpha^*)$, and $A^*(x)$ denotes the enlarged covariate matrix of domain averages including those for the new covariate vector. In estimating parameters via IBLUP, for the linearized model, we use a slightly modified version given below.

$$t_{y,d} \approx b_{d0}(\alpha^{*(r)})N_d + b_{d1}(\alpha^{*(r)})T_{x,d}'^*\alpha^* + b_{d1}(\alpha^{*(r)})N_d\zeta_d + e_d \qquad (3.15a)$$

where $b_{d0}(\alpha^{*(r)}) = v_d(\alpha^{*(r)}) - b_{d1}(\alpha^{*(r)})A_{x,d}'\alpha^{(r)}$ $-(d^{th}$element of diag $\{N_d^{-1}\}V1_B)$ (3.15b)

and $b_{d1}(\alpha^{*(r)})T_{x,d}'^*\alpha^* = b_{d1}(\alpha^{*(r)})T_{x,d}'\alpha + (d^{th}$ element of $V1_B)\alpha_B$ (3.15c)

Notice that in the last terms of (3.15b) and (3.15c), the product $b_{d1}(\alpha^{*(r)})b_{d1}^{-1}(\tilde{\alpha}^*)$ is replaced by 1. This is to get around the unusual specification of the new covariate involving the unknown α^*. This is reasonable because at convergence, the product will be 1 when $\tilde{\alpha}^*$ is replaced by $\hat{\alpha}^*$. This slight modification of b_{d0} and b_{d1} terms helps to speed up convergence of IBLUP.

Now following the LMM argument under benchmarking, we have from (3.13) and (3.15a) that

$$1_B(t_y^* - t_y^{*\text{sae}}) = 0 \qquad (3.16)$$

where $t_y^* = t_y - \text{diag}\{b_{d0}(\hat{\alpha}^*)N_d\}1$ and $t_y^{*sae} = \text{diag}\{b_{d1}(\hat{\alpha}^*)\}T^*(x)\hat{\alpha}^* + \text{diag}\{b_{d1}(\hat{\alpha}^*)N_d\}\hat{\zeta}$

However,
$$\begin{aligned} t_y^* - t_y^{*\text{sae}} &= t_y - [v(\hat{\alpha}^*) + \text{diag}\{v_d(\hat{\alpha}^*)(1 - v_d(\hat{\alpha}^*)N_d\}\hat{\zeta}] \\ &= t_y - t_y^{\text{sae}} \end{aligned} \qquad (3.17)$$

which in view of (3.16) establishes the desired benchmarking under GLMARC.

3.4 Estimation and Benchmarking in the presence of Collapsed domains

Often in practice and indeed for the CCHS example considered in the next section, we need to collapse domains before we estimate the fixed parameters α and σ_ζ^2. Suppose, for illustration, only two domains d' and d'' are collapsed. Then the total number of domains D is reduced to $\tilde{D} = D - 1$, and in the model (3.1) d is replaced by $\tilde{d}$ varying from $1,...,\tilde{D}$ as well as V by $\tilde{V}$. Now, estimation of α and σ_ζ^2 is carried out as before. However, in estimating $\zeta_{d'}, \zeta_{d''}$, the residual for the collapsed domain $[(t_{y,d'} + t_{y,d''}) - (N_{d'} v_{d'}(\hat{\alpha}) + N_{d''} v_{d''}(\hat{\alpha})]$ is apportioned to d' and d'' according to relative variances of $b_{d'1}(\hat{\alpha}) N_{d'}, \zeta_{d'}$ and $b_{d''1}(\hat{\alpha}) N_{d''}, \zeta_{d''}$, somewhat similar to (3.5); see also Singh (2006).

The benchmarking property for the SAEs for collapsed domains goes through along the same lines. However, SAEs for domains d' and d'' within the collapsed subgroup are not readily available because the extra covariate vector involves the $\tilde{D} \times \tilde{D}$ matrix $\tilde{V}$ and not V. There doesn't seem to be an optimal way to resolve this problem, but a simple solution might be to allocate the total ($\tilde{d}^{th}$row of $V1_B$, *the symbol* $\sim$ denoting the reduced dimension $\tilde{D}$) in the case of LMM to individual domains d' and d'' proportionally to $N_{d'}$ and $N_{d''}$. This essentially modifies somewhat the covariate of the random effect term in the original model before collapsing. The case of GLMARC can be handled in a similar way.

4. Application of GLMARC to the CCHS Data

The proposed method was applied to the data of Cycle 1.1 of the Canadian Community Health Survey (CCHS) which was conducted in 2000-2001. The goal was to estimate the total number of daily smokers in 40 subpopulations or small areas of the province of Prince Edward Island (PEI). The subpopulations were defined by the four health regions (Queens, East Prince, West Prince and Kings) and by ten age-sex subgroups (12-19, 20-29, 30-44, 45-64 and 65 and over). To have sufficient degrees of freedom for modeling, it was decided to model data from all four Atlantic provinces together resulting in a total of 23 health regions and 230 small areas.

Benchmark subgroups were taken as the ten marginal age-sex subgroups and the four provincial subgroups of domains. This resulted in 13 non-redundant benchmark constraints. The model used for illustration was chosen as a subgroup common mean in which it is assumed that the mean number of daily smokers is common over domains with identical age-sex subgroup. The subgroup common mean model is a simple model often useful as a good starting point in the absence of other important predictor covariates. In future, it is planned to investigate the use of hospital admissions data and other administrative data for better model covariates. In the end, the models considered had 23 regression α^*-parameters, 10 for the age-sex subgroup common means, and the additional 13 for the benchmarks.

Some domain collapsing was also required to ensure that the input data of the models had the required properties. Certain rules of thumb were used to proceed such as no zero direct estimates, minimum effective sample size of at least 30 and effective sample size times the estimated probability of occurrence of at least one. Thus, the original domains were collapsed to a total of 217 collapsed domains for estimating fixed model parameters. Collapsing partners were chosen such that they are closest in Euclidean distance with respect to the domain-specific mean profiles of model covariates. Collapsing partners were restricted to lie within benchmark subgroups, but this restriction had to be relaxed occasionally to satisfy the three rules of thumb mentioned above. Keeping collapsing partners within benchmark subgroups gives self-benchmarking of the final small area estimates. However, a slight violation is not expected to seriously impact the benchmark constraints.

Both LMM and GLMARC models were applied to the CCHS data. Several diagnostics analogous to Kalman filter innovations were performed. This was done by ranking the domains in decreasing order of effective sample size and then treating the rank as a pseudo-time variable as described in Singh (2006). The innovations produced by both the models LMM and GLMARC successfully pass the Shapiro-Wilk normality test with respective p-values of 64.28%

and 78.13%. Figures 4.1(a,b) for GLMARC and 4.2 (a,b) for LMM show the scatter plots and the Q-Q plots of innovations and show no sign for particularly unusual pattern.

Table 4.1 gives the model estimates and R-square for both models. First, the random component variance estimate is smaller for the LMM model as expected because in GLMARC it appears as a multiplicative factor of a function known to be between 0 and 1. The overdispersion estimate being less than one for both models suggests some underdispersion. Finally, the R-square as a descriptive measure of model significance taking high values (over 90%) for both models shows no reason of concern. The comparison of CVs or RRMSEs (Table 4.2) begins to show some interesting differences between the two models. Out of 230 small areas, 76 of direct estimates were unpublishable using the criterion of CV being more than 33.3%. However, with LMM-based SAE, only 10 small areas remain to be unpublishable but none under GLMARC. The final results in terms of estimates and their precision are summarized in Table 4.3. The precision columns (Mod-CV, Mod-RRMSE and RRMSE) were calculated by dividing the standard error or the root mean squared error by the GLMM-ARC estimate for ease in comparability of the methods. The first 13 rows give benchmark estimates and their precision while the last 11 rows give estimates and their precision for 11 of the 40 target domains (ranked from smallest sample size to biggest and correspond to the deciles of the sample size distribution). All benchmarks are exactly met except for the New Brunswick benchmark because some collapsing partners were chosen from the province of Newfoundland and Labrador. Also notice that even if benchmarks are exactly satisfied, their precision may vary over direct, LMM and GLMM-ARC estimates because of the use of adjusted MSE due to over-dispersion. Finally we note that although in this example we didn't observe the problem of negative estimates or negligible variance component with LMM, the estimates under GLMARC generally performed well in comparison to those under LMM especially when the RRMSE under LMM tended to be on the high side. Moreover, in view of the theoretical desirable properties of GLMARC over LMM, it may be preferable in practice.

5. Concluding Remarks

In this paper, a model termed GLMARC was proposed for aggregate level small area modeling from survey data which uses LMM-type methods in the form of IBLUP for parameter estimation. The proposed method provides a simple alternative to HB approaches as well as self-benchmarking and easily interpretable model diagnostics. The method was illustrated for the CCHS data, and the idea of domain collapsing (Singh, 2006) was also illustrated to deal with the problem of domains with zero estimates or with small sample sizes. For the CCHS example, it was found under a simple subgroup common mean model that such problem domains, reasonable point estimates as well as the corresponding MSE estimates can be obtained using the proposed method which tend to be generally superior to those under LMM. This is an improvement over commonly used practice of keeping such problem domains outside the modeling process and later only synthetic estimates are reported for them. In the proposed approach, it was shown how the method of self-benchmarking applicable to LMM can be suitably generalized to the case of GLMARC, and this can be done even under domain collapsing. Note that self-benchmarking built in the modeling cannot be achieved if problem domains are not part of the modeling process. Finally we note that although with the simple subgroups common mean model, the proposed method showed that CVs of many estimates could be improved, it would be useful to find better predictor covariates so that model error variance σ_ξ^2 could be reduced further resulting in further gains in efficiency.

References

Drum, M.L., and McCullagh, P. (1993). REML estimation with exact covariance in the logistic mixed model. *Biometrics*, 677-689.

Fay, R.E., and Herriot, R.A. (1979). Estimates of Income for small places: an application of James-Stein procedures to Census Data. ASA., 74, 269-277.

Harvey, A.C. (1989). *Forecasting, Structural Time Series, and the Kalman Filter*. Cambridge University Press.

Jiang, J., and Lahiri, P. (2006). Mixed Model Prediction and small area estimation. Test, Vol. 15, 1-96.

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

Rao, J.N.K. (2003). Small Area Estimation. New York: John Wiley.

Singh, A.C. (2006). Some problems and proposed solutions in developing a small area estimation product for clients. ASA Proc. Surv. Res. Meth. Sec., 3673-3683.

Sutradhar, B.C., and Rao, R.P. (1996). On joint estimation of regression and overdispersion parameters in generalized linear models for longitudinal data. *Multivariate Analysis*, 56, 90-119.

Vonesh, E.G., and Carter, R.L. (1992). Mixed effects nonlinear regression for unbalanced repeated measures. *Biometrics*, 48, 1-17.

You, Y., and Rao, J.N.K. (2002). Small Area Estimation using unmatched sampling and linking models. Canadian Journal of Statistics, 30, 3-15.

Figure 4.1: Scatterplot and Q-Q plot of standardized innovations under the GLMARC model

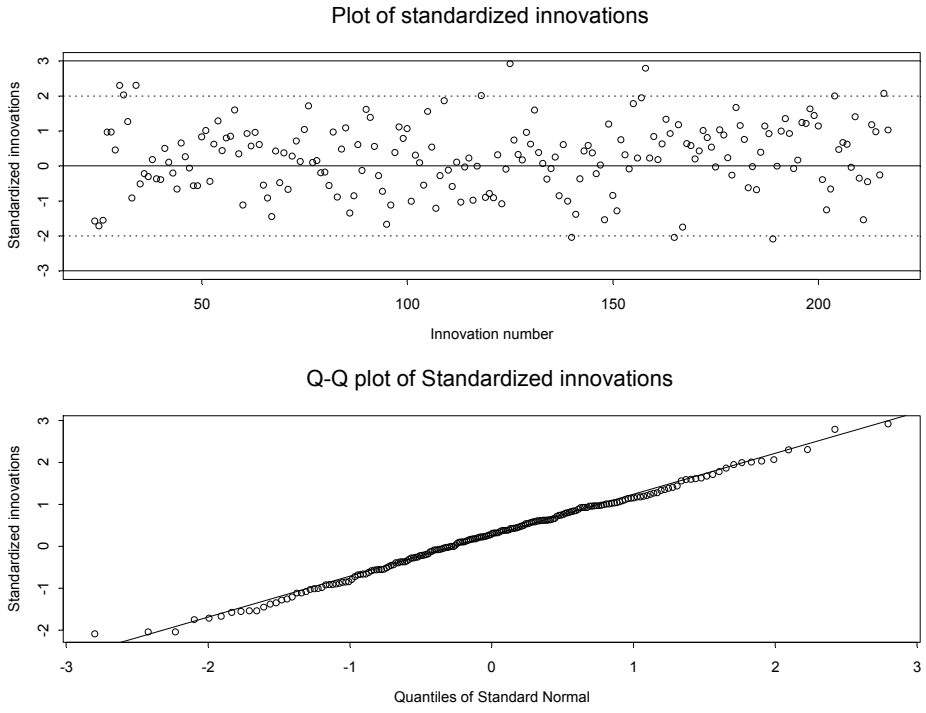

Figure 4.2: Scatterplot and Q-Q plot of standardized innovations under the LMM model

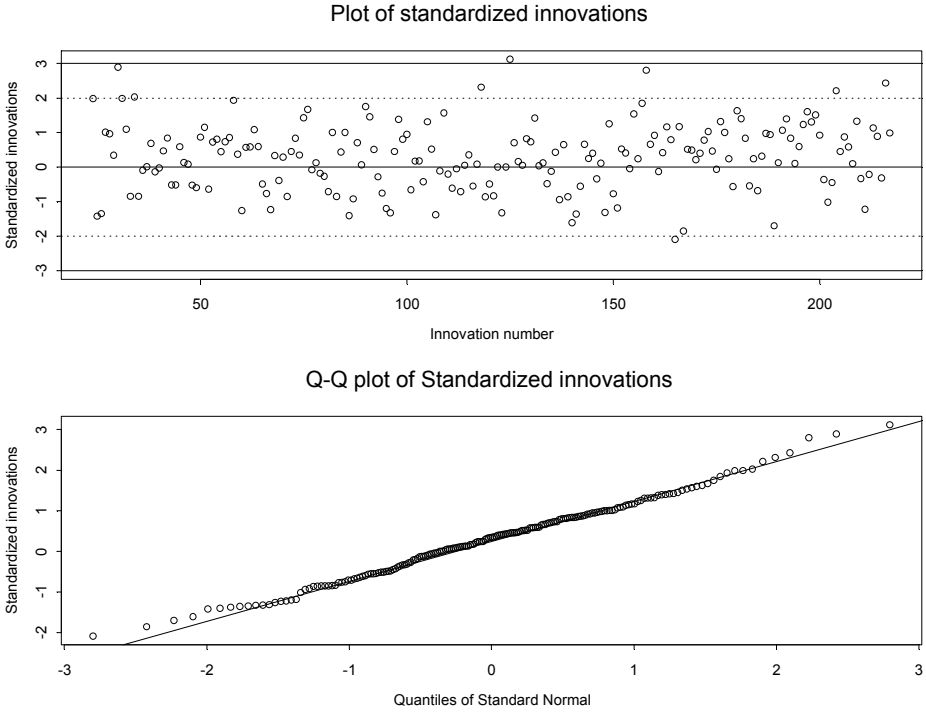

Table 4.1: Parameter estimates and R-square of the GLMARC and LMM models

Parameter					LMM	GLMARC
	Benchmark	*Age*	*Gender*	*Geography*		
$\hat{\beta}$	•	20-29	Male		-0.0030	-0.0240
	•	30-44	Male		-0.0015	-0.0164
	•	45-64	Male		-0.0023	-0.0203
	•	65+	Male		-0.0009	-0.0076
	•	12-19	Female		0.0001	0.0018
	•	20-29	Female		-0.0013	-0.0125
	•	30-44	Female		-0.0025	-0.0228
	•	45-64	Female		-0.0017	-0.0154
	•	65+	Female		0.0001	0.0063
	•			PEI	0.0011	0.0067
	•			Nova Scotia	-0.0006	-0.0031
	•			New Brunswick	-0.0004	-0.0021
	•			Atlantic	0.0029	0.0235
		12-19	Male		0.0853	-2.3091
		20-29	Male		0.3776	-0.6017
		30-44	Male		0.2865	-0.9313
		45-64	Male		0.2471	-1.1227
		65+	Male		0.1017	-2.1653
		12-19	Female		0.0804	-2.3570
		20-29	Female		0.2122	-1.3821
		30-44	Female		0.2771	-0.9438
		45-64	Female		0.1997	-1.4150
		65+	Female		0.0709	-2.5236
$\hat{\sigma}_{\eta}^2$					0.00123	0.06053
$\hat{\sigma}_{0}^2$					0.95139	0.91428
R^2					0.9619	0.9865

Table 4.2: CV categories of direct and SAE estimates under GLMARC and LMM models

CLASS	Mod-CV of the direct	Mod-RRMSE of LMM	RRMSE of GLMARC
[0]	2	0	0
(0-16.5%)	32	130	119
[16.5%-33.3%)	120	90	111
[33.3%-50%)	58	10	0
[50%-100%)	14	0	0
[100%-)	4	0	0

Table 4.3 Small area Estimates and their precision under GLMARC and LMM

Geography	Age	Gender	Sample size	Direct Estimate	LMM Estimate	GLMARC Estimate	Mod-CV Direct	Mod-RRMSE LMM	RRMSE GLMARC
---	20-29	MALE	881	52496.97	52496.97	52496.97	5.85	5.71	5.63
---	30-44	MALE	2153	90252.64	90252.64	90252.64	3.70	3.61	3.56
---	45-64	MALE	2520	74592.38	74592.38	74592.38	4.10	4.00	3.95
---	65+	MALE	1247	16720.60	16720.60	16720.60	8.99	8.77	8.65
---	12-19	FEMALE	1256	15952.30	15952.30	15952.30	8.94	8.72	8.61
---	20-29	FEMALE	1141	39406.69	39406.69	39406.69	5.92	5.77	5.69
---	30-44	FEMALE	2605	82692.72	82692.72	82692.72	3.95	3.85	3.80
---	45-64	FEMALE	2830	67787.47	67787.47	67787.47	4.52	4.41	4.35
---	65+	FEMALE	1996	17095.44	17095.44	17095.44	8.42	8.21	8.10
PEI	---	---	3651	27491.75	27491.75	27491.75	4.70	4.58	4.52
NS	---	---	5319	184684.46	184684.46	184684.46	3.33	3.24	3.20
NB	---	---	4996	147603.97	147733.11	147735.09	3.17	3.09	3.05
ATLANT.	---	---	17836	474686.16	474686.16	474686.16	1.88	1.83	1.81
1140	12-19	MALE	25	67.41	86.71	87.61	46.21	35.34	24.51
1140	20-29	FEMALE	30	100.36	211.42	205.95	24.57	17.58	19.63
1110	20-29	MALE	51	625.73	556.08	573.79	22.44	10.74	13.52
1120	12-19	MALE	59	159.32	172.29	178.17	42.19	28.75	22.11
1110	20-29	FEMALE	66	388.55	303.10	295.83	28.91	16.46	18.04
1140	65+	FEMALE	79	123.87	102.39	88.01	38.85	28.15	19.40
1120	30-44	MALE	92	1288.21	1213.87	1230.43	20.24	10.11	12.76
1120	65+	FEMALE	115	230.57	222.06	208.74	36.78	27.08	20.49
1110	65+	FEMALE	123	67.61	80.80	95.85	28.78	25.28	21.32
1110	45-64	FEMALE	159	518.33	525.00	516.05	19.55	13.35	13.85
1130	45-64	FEMALE	218	1592.19	1764.76	1734.63	15.48	11.83	12.39