

Session 4

Robust Small Area Estimation Based on a Survey Weighted
MCMC Solution for the General Linear Mixed Model

Hierarchical Bayes Small Area Estimation for Survey Data by EFGL: The Method of Estimating Function-Based Gaussian Likelihood

A. C. Singh, R. E. Folsom, Jr., and A. K. Vaish

RTI International

Abstract

In this paper a new approach representing a generalization of Fay-Herriot (1979) (FH) to unit-level nonlinear mixed models is presented which, like FH, employs data aggregation but through design-weighted estimating functions rather than estimators. Working with estimating functions (EFs) helps to alleviate the problems associated with FH because EFs, in general, can be better approximated by normality even for modest sample sizes, and can always be collapsed, if necessary, to improve the Gaussian approximation and the precision of variance estimates. Also, EFs can be based on unit-level covariate information, and can be specified at the lowest level of aggregation to avoid the problem of internal inconsistency. For hierarchical Bayes (HB) small area estimation, the proposed approach simply replaces the likelihood (typically computed under the assumption of ignorable design) with the estimating function based Gaussian likelihood which does not require ignorability of the design. The method is illustrated by means of a simple example of fitting a HB linear mixed model to data obtained from a nonignorable sample design. Both fixed and random parameters are estimated to construct small area estimates. Different scenarios for nonignorability are considered. MCMC is used for HB parameter estimation.

Key Words: Estimating functions; Pseudo Score Functions; Survey weighted HB; MCMC

1. INTRODUCTION

This research on small area estimation (SAE) was motivated by the problem of fitting generalized linear mixed models to survey data when unit-level covariate information is available. The problem arose in the context of the 1999 National Household Survey on Drug Abuse (NHSDA), see Folsom, Shah, and Vaish (1999). In the NHSDA, one of the outcome variables (y) of interest is past month marijuana use by persons aged 12-17. For this dichotomous variable, one can use as covariates person-level demographic variables,

census block-group-level demographic variables, census tract-level demographic and socioeconomic status variables, and inter-censal county-level variables including drug-related arrest, treatment and death rates. For estimating propensity of marijuana use at the state-level (treated as a small area), the following hierarchical Bayes (HB) model similar to the one considered by Folsom et al. may be formulated:

$$\begin{aligned}
y_{ijk} &= \mu_{ijk} + \varepsilon_{ijk}, y_{ijk} \sim \text{Bernoulli}(\mu_{ijk}) \\
g(\mu_{ijk}) &= x'_{ijk}\beta + \eta_i + v_{ij} \\
\eta_i &\sim_{iid} N(0, \sigma_\eta^2), v_{ij} \sim_{iid} N(0, \sigma_v^2) \\
\beta &\sim U(R^p), \sigma_\eta^2 \sim IG(\nu_0/2, \sigma_{\eta_0}^2/2), \sigma_v^2 \sim IG(\nu_1/2, \sigma_{v_0}^2/2)
\end{aligned} \tag{1.1}$$

where y_{ijk} denotes the observation on the k^{th} individual from the j^{th} cluster (such as a county) in the i^{th} stratum (such as a state), x_{ijk} is the corresponding individual-level covariate vector, and $g(\cdot)$ is the link function (such as the logit). The p -dimensional fixed parameter β has an improper uniform distribution on the p -dimensional space of real vectors, and the variance components $\sigma_\eta^2, \sigma_v^2$ have nearly flat inverse Gamma priors with very small location and scale parameters $\nu_0 > 0, \sigma_{\eta_0}^2 > 0, \nu_1 > 0, \sigma_{v_0}^2 > 0$. The model errors ε 's are independent of each other, and also independent of the random effects η 's and v 's.

In the context of survey data, the model (1.1) is a super-population model assumed to hold for the finite population U_N of size N . For U_N , let M be the number of strata ($i=1, \dots, M$), and N_i be the number of clusters in the i^{th} stratum ($j=1, \dots, N_i$), and N_{ij} be the number of individuals in the (i,j) th cluster ($k=1, \dots, N_{ij}$). The parameters η_1, \dots, η_M are the realized values from $N(0, \sigma_\eta^2)$. The (random) parameters of interest are the stratum means, μ_i , and the domain means, μ_d where the domain d may cut across strata. Thus,

$$\mu_i = \left(\sum_j \sum_k N_{ijk} \mu_{ijk} \right) / N_i, \quad \mu_d = \sum_i \gamma_{id} \mu_{id}, \tag{1.2}$$

where γ_{id} is the proportion of domain- d units in stratum- i , and μ_{id} is the mean of the domain- d units in stratum- i . Other parameters of interest may be the overall mean $\mu \left(= \sum_i \gamma_i \mu_i, \text{ where } \gamma_i = \sum_j N_{ij} / N \right)$, and the fixed parameters $\alpha, \beta, \sigma_\eta^2$ and σ_v^2 .

The observed data is a sample (s) of size n from the finite population U_N . If the sample design, $p(s)$, is

ignorable for the model (1.1), i.e., the model (1.1) also holds for the sample, s , then the usual HB estimation theory can be applied to s . However, if the design is nonignorable, use of the standard likelihood in the HB framework would lead to a biased posterior distribution, because the model (1.1) cannot be assumed to hold for the sampled data due to selection bias. This is discussed further in Section 2.

In Section 3, we consider existing solutions based on the seminal work of Fay-Herriot's (1979) aggregate-level model, and show how it takes account of the survey design. However, it does have some limitations which are also discussed. Section 4 provides motivation for the alternative proposed solution which is described in Section 5 in the context of a simple example of mixed linear models. The MCMC steps for the proposed HB-SAE method are described in section 6. Sections 7 and 8 describe the simulation experiment and results. The case of mixed nonlinear models is considered in Section 9 which also shows how the proposed method compares with the alternative method of Folsom et al. originally proposed for the NHSDA application. Finally we conclude the paper with some remarks in Section 10.

2. NONIGNORABILITY OF SAMPLE DESIGN

Consider a super-population model which is assumed to hold for the finite population U_N . For the sake of simplicity, we first consider a simple linear mixed model for the observations y_{ij} on the unit j in the i^{th} cluster, ($i = 1, \dots, M$; $j = 1, \dots, N_i$). We have

$$y_{ij} = x_{ij}'\beta + \eta_i + \varepsilon_{ij} \quad (2.1)$$

where $\varepsilon_{ij} \sim_{iid} N(0, \sigma_\varepsilon^2)$, $\eta_i \sim_{iid} N(0, \sigma_\eta^2)$, β is a p -vector of fixed effects, and x_{ij} is a p -vector of covariates associated with the unit j in the cluster i . Here η_i 's are random cluster effects.

We note that in practice it is almost impossible to include in the model all the factor effects (main and interaction) of design covariates such as cluster characteristics that are deemed to be related to the outcome variable y . This happens for several reasons: (i) the need for a parsimonious model, (ii) the need to avoid instability of parameter estimates, (iii) the model should correspond to the analyst's goals, and (iv) some covariates at lower levels are excluded due to unavailability of lower level population totals; these totals are needed in defining finite population parameters.

Since sample selection probabilities may depend on the outcome variables through design covariates, and

since all the factor effects due to design covariates may not be controlled in the model, it is difficult to assume that the design can be ignored for the model under consideration. This is why many survey samplers prefer to follow the conventional wisdom of playing it safe by taking the design into account. There are two main scenarios in small area modeling which make the design nonignorable.

Scenario I. Here small areas are, in fact, design strata, and the random effects η_i 's correspond to these strata.

Sampling within each stratum is informative in that the sample inclusion probability π_{ij} depends on ε_{ij} . Note that the factors corresponding to design covariates (x_2 , say), which are omitted from the model but are correlated with y_{ij} , become naturally part of ε_{ij} . This is easily seen from the following expression for the reduced model $y = E(y | x_1) + \varepsilon'$, $\varepsilon' = (E(y | x_1, x_2) - E(y | x_1) + \varepsilon)$ when the enlarged model is $y = E(y | x_1, x_2) + \varepsilon$.

Scenario II. Here, small areas are like domains, and the random effects η_i 's correspond to these domains.

Note that each domain may cut across design strata. In each stratum, sampling may be informative in that the sample inclusion probability of the $(i,j)^{\text{th}}$ unit in the h^{th} stratum, $\pi_{h(ij)}$ may depend on η_i or ε_{ij} or both. This is again for the reason that effects of design covariates which are not part of the model covariates x 's, become automatically part of the residual, $\eta_i + \varepsilon_{ij}$; here the residual has two components, η_i and ε_{ij} .

Now, in Bayes or hierarchical Bayes estimation, we need specifications of the likelihood, $L(y | \beta, \eta, \sigma_\varepsilon^2)$ and of prior distributions. If $L(\cdot)$ is misspecified, the posterior distribution, $[\theta | y]$, is not correct for parameters of interest θ . (For instance, $\theta_i = A'_{xi}\beta + \eta_i$, is (approximately) the i^{th} area mean where $\sum_j \varepsilon_{ij} / N_i \approx 0$ for large N_i , and A_{xi} is the mean of x for the i^{th} area, i.e., $A_{xi} = T_{xi} / N_i$, $T_{xi} = \sum_j x_{ij}$.) Thus, any characteristic of $[\theta | y]$, in particular the posterior mean, could be (seriously) biased in that

$$E_{\theta|y} [\theta - E^*(\theta | y)] \neq 0 \quad (2.2)$$

where E^* denotes the posterior expectation based on the misspecified likelihood.

In the next section, we consider the existing solution of Fay and Herriot (1979, henceforth referred to as FH) in which the sampling design is taken into account by working with the aggregate-level data. Note that for aggregate statistics such as weighted sample totals or means, design-based variances and covariances can be estimated, and their distribution can be approximated as Gaussian. It is difficult in general to specify the

distribution of the unit-level data because there is not enough information about the distribution of the N-vector of sample inclusion indicators. In fact, typically, not even all the first order inclusion probabilities are known, let alone second or higher order inclusion probabilities. Some alternative approaches based on modeling of selection probabilities have been proposed by Pfeffermann and Sverchkov (1999). However, with the desirable goal of making minimum modeling assumptions for SAE, a way out might be to do efficient aggregation of data that incorporates unit-level information, and then use sampling weights as in FH, see Section 4. It may be remarked that unlike the census data which is based on nature's selection mechanism of the finite population, the sample from the finite population is based on man's selection mechanism, and hence the sampler knows very well what should not be assumed away. This is probably why the analysis of survey data becomes quite challenging, and thus distinguishes itself from the mainstream of statistics.

3. EXISTING SOLUTION: AGGREGATE LEVEL MODEL OF FH

The work of FH represents a milestone in the history of the development of SAE as it is the first method that takes design into account in small area modeling. The basic idea is to transform the unit-level data (y) to aggregate-level data (\tilde{y}) by using the direct small area estimates, $\hat{\theta}_{i,dir} \left(= \sum_{j=1}^{n_i} y_{ij} w_{ij} / w_{i+} \right)$ where w_{ij} s are the (calibrated) design weights, and $w_{i+} \left(= \sum_{j=1}^{n_i} w_{ij} \right)$ is typically equal to N_i due to weight calibration. Thus, in FH, the data is first condensed into M estimates which are modeled as follows. We will consider only Scenario I for the sake of simplicity. For $i = 1, \dots, M$; we specify the following

$$\begin{aligned} \text{Observation model: } & \hat{\theta}_{i,dir} = \theta_i + e_i, \text{ and} \\ \text{Link model: } & \theta_i = A'_{xi} \beta + \eta_i, \end{aligned} \tag{3.1}$$

where $e \sim N(0, \text{diag}(V))$, $\eta_i \sim_{iid} N(0, \sigma_\eta^2)$.

Here, V denotes the vector of design-based variance estimates that are regarded as known. In practice, they could be smoothed by suitable modeling; FH used generalized variance functions to smooth V , while Otto and Bell (1995) proposed a parameterization of $\text{Cov}(e)$ along with a suitable prior under a Bayesian framework. Even if variance estimates are not smoothed, one could still treat them as known and meet the goal of SAE modeling. The reason for this is that the main goal of SAE modeling is to see whether variances of SAEs after borrowing strength from other areas via modeling can be reduced appreciably in comparison to the variances of direct estimates. Note that under the assumption of unit-level model (2.1), there is another error term involving ε_{ij} in the link model (3.1) given by

$$\begin{aligned}\theta_i &= A'_{xi}\beta + \eta_i + \sum_{j=1}^{N_i} \varepsilon_{ij} / N_i, \\ &\approx A'_{xi}\beta + \eta_i\end{aligned}\tag{3.2}$$

where the term $\sum_j \varepsilon_{ij} / N_i \approx 0$ by SLLN, because N_i is expected to be very large in practice even though n_i may be small. Similarly, the Cov(e) in the observation model involves σ_ε^2 when the covariance is computed under both design and model randomizations, i.e., when the super-population expectation of the design-based covariance is taken. However, it is better to use just the design-based estimate of Cov(e) for several reasons: firstly, the actual computational form for Cov(e) under complex designs may be quite complex involving unknown second order inclusion probabilities, and so computation of its expectation may be prohibitive; secondly, even if the expectation involving σ_ε^2 is computable, one cannot produce good estimates of both σ_ε^2 and σ_η^2 from the aggregate-level data because it is hard to discriminate between them without unit-level data; and thirdly, the design-based estimate of Cov(e) has the desirable property of robustness to departures from the link model.

The Gaussian approximation of $\hat{\theta}_{i,dir} - \theta_i$ in the FH set-up is based on the Central Limit Theorem, and using this, FH proposed empirical Bayes estimators for θ_i s. However, if we were interested in HB estimation using the aggregate-level data, the unit-level likelihood $L(y|\cdot)$ can be replaced by the aggregate-level likelihood $L(\tilde{y}|\cdot)$, and one can then proceed as in Datta and Ghosh (1991).

Although the FH method represents a very important development in SAE methodology for survey data, it does suffer from a few limitations resulting mainly from aggregate-level modeling. Note that when the unit-level model is of interest, there is a loss of efficiency by using an aggregate-level model. This is analogous to the case of using the grouped data mle instead of the raw data mle in chi-square goodness-of-fit tests. While it is true that some loss of efficiency is inevitable when trying to take design into account, the issue under consideration is how to reduce this efficiency loss for unit-level models. Below we list some limitations of the FH approach.

(a) In the aggregate-level modeling approach of FH, unit-level covariate information is not exploited. The more unit-level information is used, the more efficient the resulting estimators are expected to be.

(b) The FH model is specific to the level of aggregation used. If we change the level of aggregation, we get a different model which is not internally consistent with the original model. Note that the exchangeability assumption about η_i 's is specific to the level of aggregation. This inconsistency problem becomes more acute when dealing with nonlinear models either in the mean function of the link model or in the dependent variable of the observation model. For example, with the logit link function, mean at a higher level is not sum of the means at lower levels that make up the higher level of aggregation. In practice, the additive property is clearly desirable. We run into similar problems if $\hat{\theta}_i$ is transformed through a nonlinear function such as $\log \hat{\theta}_i$. Here, an additional problem arises in the definition of $\log \hat{\theta}_i$ when $\hat{\theta}_i = 0$, see e.g. the report on SAIPE models by US Bureau of the Census (1998).

(c) In FH, the Gaussian approximation of $\hat{\theta}_i - \theta_i$ may not be reasonable for small to modest n_i 's. This may be more of a concern when dealing with discrete outcome variables.

(d) Finally, smoothed variance estimates V may not be a good approximation for very small n_i 's. Note that, if the direct small area estimates $\hat{\theta}_{i,dir}$ are unstable (this is precisely the reason why we are modeling to borrow strength), then the variance estimates V will, of course, be unstable.

4. MOTIVATION FOR THE ALTERNATIVE SOLUTION

In this paper we propose a generalization of FH to unit-level nonlinear mixed models such that unit-level covariate information is efficiently used as well as some form of data aggregation is used to account for the sample design. Recently, in an innovative attempt to account for the design, Prasad and Rao (1999) derived an aggregate-(or area-) level model for direct estimates from the unit-level model using survey weights, and obtained pseudo-optimal SAEs. It is pseudo in that the design was assumed to be ignorable, and so only the effect of unequal selection probabilities (i.e., sampling weights) was accounted for in the joint design-model variance. Moreover, for estimating variance components, in addition to assuming that the design was ignorable, the unequal weighting effect was also not accounted for. You and Rao (2003) used a similar framework for developing pseudo HB estimates. The above methods, however, are applicable to only linear models because the aggregate-level model for direct estimates is derived from the unit-level model. On the other hand, the method of Folsom et al. (1999) deals with unit-level mixed nonlinear models and develops a HB method using pseudo-likelihood involving survey weights and the corresponding survey weighted

estimating functions. However, the method assumes ignorability of the design, and the pseudo likelihood used for HB need not be a valid likelihood; see Section 7 for a brief discussion.

Our goal is to attempt to take full account of the survey design in unit-level modeling, and to develop methods that apply to both linear and nonlinear models. To this end, unlike FH we resort to data aggregation via survey-weighted estimating functions rather than through estimators. Use of survey weighted estimating functions has been implicitly invoked by survey statisticians for a long time in ratio and regression type estimators, see e.g., Fuller (1975), Cassel, Särndal, and Wretman (1976). The pioneering work of Binder (1983) explicitly introduced a general theoretical framework of survey weighted EFs for deriving estimators of super-population parameters, and their asymptotic properties under a given sample design. The optimality of survey-weighted EFs under joint design-model randomization was, however, established by Godambe and Thompson (1986) using the optimality framework of Godambe (1960). As an example, for the simple mixed linear model (2.1), the optimal EFs for β and η_i 's have heuristically appealing forms and are given by

$$\begin{aligned}\varphi_{\eta(i)} &= \sum_{j=1}^{n_i} (y_{ij} - x'_{ij}\beta - \eta_i) w_{ij}, \\ \varphi_{\beta} &= \sum_{i=1}^M \sum_{j=1}^{n_i} x_{ij} (y_{ij} - x'_{ij}\beta - \eta_i) w_{ij}\end{aligned}\tag{4.1}$$

where w_{ij} 's are inverse of the first order selection probabilities π_{ij} 's.

We propose to use the above set of EFs as the starting point for Bayes or HB estimation, i.e., the likelihood would be defined by the distribution of these EFs. Clearly, EFs use unit-level information and they use it efficiently in view of their optimality properties. It is also known that EFs can be better approximated as Gaussian even for modest sample sizes (McCullagh, 1991) because by their very nature, they are simple sums of elementary zero functions, although the elementary functions could be complex by themselves. Moreover, EFs can be easily collapsed to improve the Gaussian approximation as well as the precision of variance estimates. Notice that the serious problem of internal inconsistency can be avoided by defining the EFs at the lowest level of aggregation. Thus, parameters at higher levels of aggregation can be obtained from the lowest level parameter estimates which serve as building blocks. It should also be noted that, typically in practice, the joint inclusion probabilities ($\pi_{i(jk)}$) of units j and k in stratum i are not available and therefore, survey weighted EFs can't be constructed if they involve cross-product terms, e.g., if they involve double sums within a stratum i . It is, therefore, desirable to specify the model (2.1) so that the error term ε_i 's are i.i.d. which, in turn, gives rise to single sums within strata for survey weighting.

Now, the vector ϕ of EFs (which involves data and parameters) serves as the condensed input data which after collapsing, if necessary, gives rise to an approximate Gaussian likelihood, $L(y^*|\beta, \eta, \cdot)$ where y^* denotes the implicit condensing of information in y via ϕ . Thus, for the unit-level HB analysis, the original likelihood $L(y|\cdot)$ (which would have been based on the ignorable design assumption) is replaced by the estimating function based Gaussian likelihood (EFGL), $L(y^*|\cdot)$ which does not assume ignorability of the design.

5. PROPOSED METHOD (EFGL)

We shall describe the proposed method of estimating function-based Gaussian likelihood (EFGL) in terms of the model (2.1). Suppose, the HB-framework at the census-level is defined as follows:

$$\begin{aligned} y_{ij} | \beta, \eta, \sigma_\varepsilon^2 &\sim N(x'_{ij}\beta + \eta_i, \sigma_\varepsilon^2) \\ \eta_i &\sim_{iid} N(0, \sigma_\eta^2), \quad \beta \sim U(R^p) \\ \sigma_\eta^2 &\sim IG(v_0/2, \sigma_{\eta_0}^2 | 2), \quad \sigma_\varepsilon^2 \sim U(0, \infty). \end{aligned} \quad (5.1)$$

Here an attempt is made to specify the priors to make them as noninformative as possible, and thus making the HB framework as objective as possible. Thus, the p-vector β of regression coefficients is assumed to have an improper uniform prior on the p-dimensional Euclidean space. However, this does not affect the propriety of the posterior of β . For variance component σ_η^2 , choice of the inverse Gamma as prior is computationally convenient because of its conjugate nature, and we can choose the shape parameter ($v_0/2$) and the scale parameter ($\sigma_{\eta_0}^2/2$) as very small positive numbers to make it nearly noninformative. The prior for σ_ε^2 , however, is improper like that of the mean parameter β , because in EFGL, as will be seen later, we introduce a separate EF, $\phi_{\sigma^2(\varepsilon)}$, for σ_ε^2 which treats σ_ε^2 as a mean parameter. It turns out as expected and as in the case of FH that σ_ε^2 is not functionally part of the V-C matrix Σ_ϕ of ϕ when a suitable design-based estimate of Σ_ϕ is substituted. So we need to add an extra EF if the estimation of σ_ε^2 is also of interest. It may be noted that there is quite a bit of flexibility in the EF framework in that all the pieces of information deemed important can be incorporated by augmenting the vector ϕ .

Now, the EFGL method will be defined for Scenario I in which small areas are strata. The EFs $\phi_{\eta(i)}$ and ϕ_β were defined earlier by (4.1). Further suppose,

$$\varphi_{\eta(i)} \sim_{\text{approx}} N(0, V_{\eta(i)}), \quad \varphi_{\beta} \sim_{\text{approx}} N(0, V_{\beta}) \text{ and } \text{Cov}(\varphi_{\beta}, \varphi_{\eta(i)}) = C_{\beta\eta(i)}. \quad (5.2)$$

Next define

$$\tilde{\varphi}_{\beta} = \varphi_{\beta} - \sum_i a_i \varphi_{\eta(i)}, \quad a_i = C_{\beta\eta(i)} / v_{\eta(i)}$$

which implies that $\tilde{\varphi}_{\beta}$ is uncorrelated with $\varphi_{\eta(i)}$'s. It should be remarked that if the model (2.1) has an intercept β_0 , then $\varphi_{\beta 0} = \sum_i \varphi_{\eta(i)}$ implying that $\tilde{\varphi}_{\beta 0} = 0$. We, therefore, drop one element from φ_{β} corresponding to β_0 . However, we shall continue to use φ_{β} to denote the reduced vector of dimension $p-1$.

Further, since

$$\text{Cov}(\tilde{\varphi}_{\beta}) \equiv \tilde{V}_{\beta} = V_{\beta} - C_{\beta\eta} V_{\eta}^{-1} C'_{\beta\eta}, \quad V_{\eta} = \text{diag}(V_{\eta(1)}, \dots, V_{\eta(M)}). \quad (5.3)$$

We have

$$\tilde{\varphi} = (\varphi_{\eta}, \tilde{\varphi}_{\beta})' \sim_{\text{approx}} N_{M+p-1}(0, \tilde{V}_{\varphi}), \quad \tilde{V}_{\varphi} = \text{blockdiag}(V_{\eta}, \tilde{V}_{\beta}) \quad (5.4)$$

and the EFG log-likelihood is given by

$$\ell(y^* | \beta, \eta) = \text{const} - \frac{1}{2} \left(\frac{\varphi_{\eta(1)}^2}{v_{\eta(1)}} + \dots + \frac{\varphi_{\eta(M)}^2}{v_{\eta(M)}} + \tilde{\varphi}_{\beta}' \tilde{V}_{\beta}^{-1} \tilde{\varphi}_{\beta} \right). \quad (5.5)$$

In the above EFG, the covariance matrix \tilde{V}_{φ} is design-based. This matrix may, in general, depend on unknown parameters which can be evaluated at their current values in the MCMC samples. It may be noted that there is, in fact, a second component involving σ_{ε}^2 when the V-C matrix of $\tilde{\varphi}$ is computed under joint design-model randomization. However, it is negligible in comparison to the first term, \tilde{V}_{φ} , under the usual assumption of $n_i \ll N_i$. It should also be emphasized that, in practice, some collapsing of $\varphi_{\eta(i)}$'s may often be required because the corresponding n_i 's (which are random under Scenario II) may be small. We may need this collapsing to improve the Gaussian approximation, as well as to improve the precision of the estimate \tilde{V}_{φ} . The effect of EF-collapsing on η_i -estimates is that all the prior estimates of θ_i 's ($\theta_i = A'_{xi} \beta + \eta_i$) that are part of a given collapsed EF, are shrunk toward the direct estimate of the corresponding collapsed small area. It is, therefore, important to choose EF-collapsing partners carefully so that they have similar η_i 's both in magnitude and sign. To this end, one can make a decision based on substantive considerations. However, in practice, as a yardstick one can use $\hat{\eta}_{i,HB}^{(0)}$ obtained under the ignorability assumption. Once it is decided

which η_i 's would be used in EF-collapsing, one can construct a new census EF under the assumption of common η_i 's for this set, and then employ survey weighting to get the appropriate collapsed EF.

If estimation of σ_ε^2 is also of interest, we add an extra EF as mentioned earlier. It is again motivated by census EF, and is given by

$$\varphi_{\sigma_\varepsilon^2} = \sum_i \sum_j \left((y_{ij} - x'_{ij}\beta - \eta_i)^2 - \sigma_\varepsilon^2 \right) w_{ij} \sim_{approx} N\left(0, V_{\sigma_\varepsilon^2}\right). \quad (5.6)$$

Note that in FH, although σ_ε^2 is not made explicitly part of the model, it could be done so by taking expectation of the design-based variance V . However, as mentioned earlier, using aggregate-level data $\hat{\theta}_{i,dir}$, it would be difficult to discriminate very well between the two variance components σ_η^2 and σ_ε^2 .

With the specification of EFGL, estimation of parameters $(\eta, \beta, \sigma_\eta^2, \sigma_\varepsilon^2)$ can proceed in the HB setup using MCMC steps. The next section gives details of full conditional posterior distributions needed for MCMC. Although so far, we have considered only Scenario I, the case of Scenario II is somewhat analogous. The main difference is that the V-C matrix of φ_η is no longer diagonal, and so the form of the EFGL is not as simple. However, full conditional posterior distributions (Section 6), can be derived easily by first orthogonalizing φ_β with respect to φ_η , and then for each i , orthogonalizing $\varphi_{\eta(i)}$ with all other $\varphi_{\eta(i')}, i' \neq i$.

6. MCMC FOR THE PROPOSED HB-SAE

For the Scenario I, the MCMC steps for finding full conditionals can be defined as follows. It is assumed that the regularity conditions for the convergence of the MCMC steps toward a stationary distribution hold.

Step I. $[\beta | y^*, \eta]$

We note that under the vague uniform prior for β , the posterior of β is simply proportional to the likelihood, and is given by

$$\log[\beta | \cdot] = const - \frac{1}{2} \left[\sum_{i=1}^M \varphi_{\eta(i)}^2 / v_{\eta(i)} + \tilde{\varphi}'_\beta \tilde{V}_\beta^{-1} \tilde{\varphi}_\beta \right]. \quad (6.1)$$

Since the kernel of the log-likelihood involves first and second powers of β , one can complete after some algebra the quadratic form in β . This implies that $[\beta | \cdot]$ is exact Gaussian with mean and V-C matrix given

respectively by the mode and curvature (at mode of the above kernel function if it depends on β). Thus,

$$[\beta | y^*, \eta] = N_{p-1} \left[\hat{\beta}_{\text{mode}}, \Sigma_{\psi(\beta)}^{-1} \right], \quad (6.2)$$

where $\hat{\beta}_{\text{mode}}$ solves the estimating equation $\psi_\beta = 0$,

$$\psi_\beta = (\partial/\partial\beta) \log L(y^* | \beta, \eta) = \left(\sum_{i=1}^M \varphi_{\eta(i)} x_{i+w} \right) / V_{\eta(i)} + (\tilde{X}' W \tilde{X}) \tilde{V}_\beta^{-1} \tilde{\varphi}_\beta. \quad (6.3)$$

where $\tilde{X}' W \tilde{X} \equiv X' W X - \sum_i a_i x'_{i+w}$, $x_{i+w} = \sum_i x_{ij} w_{ij}$, and $X' W X = \sum \sum x_{ij} x'_{ij} w_{ij}$. It is seen that similar to generalized least squares, $\hat{\beta}_{\text{mode}}$ can be obtained in a closed form. The V-C of ψ_β is easily obtained as

$$\Sigma_{\psi(\beta)} = -E \left[\frac{\partial \psi_\beta}{\partial \beta} \right] = \sum_{i=1}^M x_{i+w} x'_{i+w} / V_{\eta(i)} + (\tilde{X}' W \tilde{X}) \tilde{V}_\beta^{-1} (\tilde{X}' W \tilde{X}) \quad (6.4)$$

Step II. $[\eta_i | \eta_{i'}, \beta, y^*, \sigma_\eta^2, i' \neq i]$, $i = 1, \dots, M$

Since the posterior $[\eta_i | \cdot]$ is proportional to the product of the likelihood and the prior, we have

$$\log [\eta_i | \cdot] = \text{const.} - \frac{1}{2} \left[\sum_{i=1}^M \varphi_{\eta(i)}^2 / V_{\eta(i)} + \tilde{\varphi}_\beta' \tilde{V}_\beta^{-1} \tilde{\varphi}_\beta + \eta_i^2 / \sigma_\eta^2 \right]. \quad (6.5)$$

As in Step I, the kernel on the right hand side of (6.5) involves first and second powers of η_i , and one can complete the square in η_i . Therefore, $[\eta_i | \cdot]$ is also exact Gaussian with mean and variance given by the mode and curvature. That is,

$$[\eta_i | \cdot] = N \left[\hat{\eta}_{i, \text{mode}}, \sigma_{\psi(\eta(i))}^{-2} \right] \quad (6.6)$$

where

$$\begin{aligned} \hat{\eta}_{i, \text{mode}} \text{ solves } \psi_{\eta(i)} &= 0, \\ \psi_{\eta(i)} &= (\partial/\partial\eta(i)) \log [\eta_i | \cdot] = \varphi_{\eta(i)} w_{i+} / V_{\eta(i)} + \tilde{x}'_{i+w} \tilde{V}_\beta^{-1} \tilde{\varphi}_\beta - \eta_i / \sigma_\eta^2. \end{aligned} \quad (6.7)$$

where $\tilde{x}_{i+w} = x_{i+w} - a_i w_{i+}$. Again as in the case of β , $\hat{\eta}_{i, \text{mode}}$ has a closed form. The variance $\sigma_{\psi(\eta(i))}^2$ is obtained as

$$\sigma_{\psi(\eta(i))}^2 = -E \left(\left(\partial/\partial\eta(i) \right) \psi_{\eta(i)} \right) = w_{i+}^2 / V_{\eta(i)} + \tilde{x}'_{i+w} \tilde{V}_\beta^{-1} \tilde{x}_{i+w} + \sigma_\eta^{-2} \quad (6.8)$$

It is interesting to note that $\psi_{\eta(i)}$ and $\sigma_{\psi(\eta(i))}^2$ coincide with the usual BLUP theory when the design is ignorable and $w_{ij} = w$ (a constant). To see this, note that under the ignorality assumption,

$$a_i = C_{\beta\eta(i)} / V_{\eta(i)} = \text{Cov}\left(\sum_i \sum_j x_{ij} e_{ij} w_{ij}, \sum_j e_{ij} w_{ij} / \sigma_\varepsilon^2 \sum_j w_{ij}^2\right) = \sigma_\varepsilon^2 \sum_j x_{ij} w_{ij}^2 / \sigma_\varepsilon^2 \sum_j w_{ij}^2, \quad (6.9)$$

where $e_{ij} = y_{ij} - x'_{ij}\beta - \eta_i$. Thus, assuming $w_{ij} = w$, we have $a_i = \sum_j x_{ij} / n_i$ and $x_{i+w} - a_i w_{i+w} = 0$. Also,

$$w_{i+}^2 / v_{\eta(i)} = w_{i+}^2 / \sigma_\varepsilon^2 \sum_j w_{ij}^2 = n_i / \sigma_\varepsilon^2 \quad (6.10)$$

The reduced forms of $\psi_{\eta(i)}$ and $\sigma_{\psi(\eta(i))}^2$ are $\psi_{\eta(i)} = \sum_j e_{ij} / \sigma_\varepsilon^2 - \eta_i / \sigma_\eta^2$, and

$$\sigma_{\psi(\eta(i))}^2 = n_i / \sigma_\varepsilon^2 + 1 / \sigma_\eta^2 = \frac{\sigma_\eta^2 + \sigma_\varepsilon^2 / n_i}{\sigma_\eta^2 \sigma_\varepsilon^2 / n_i} \quad (6.11)$$

which implies that

$$\hat{\eta}_{i,\text{BLUP}} = \left(\frac{n_i}{\sigma_\varepsilon^2} + \frac{1}{\sigma_\eta^2} \right)^{-1} \frac{\sum_j (y_{ij} - x'_{ij}\beta)}{\sigma_\varepsilon^2} = \frac{\sigma_\eta^2}{\sigma_\eta^2 + \sigma_\varepsilon^2 / n_i} \sum_j (y_{ij} - x'_{ij}\beta) / n_i \quad (6.12a)$$

and

$$E\left(\hat{\eta}_{i,\text{EFGL}} - \eta_i\right)^2 = \sigma_{\psi(\eta(i))}^{-2} = \frac{\sigma_\eta^2 \sigma_\varepsilon^2 / n_i}{\sigma_\eta^2 + \sigma_\varepsilon^2 / n_i} = E\left(\hat{\eta}_{i,\text{BLUP}} - \eta_i\right)^2 \quad (6.12b)$$

Step III. $[\sigma_\eta^2 | \eta]$

In view of the conjugate nature of the prior, the conditional posterior also has the inverse Gamma distribution, and is given by

$$[\sigma_\eta^2 | \eta] = IG\left[(v_0 + M) / 2, \left(\sigma_{\eta_0}^2 + \sum_i^M \eta_i^2\right) / 2\right] \quad (6.13)$$

which implies that the conditional posterior mean of σ_η^2 ,

$$E[\sigma_\eta^2 | \eta] = \left(M \left(\sum \eta_i^2 / M\right) + (v_0 - 2) \sigma_{\eta_0}^2\right) / (M + v_0 - 2) \quad (6.14)$$

It follows that the unconditional posterior mean of σ_η^2 , i.e. $E[\sigma_\eta^2 | \bar{y}]$ is obtained by the average of MCMC realizations after convergence. This posterior mean is known to be approximately equal to the REML estimator for large M, see Kass and Steffey (1986), and Singh, Stukel, and Pfeffermann (1996).

Next, if σ_ε^2 also needs to be estimated, then logL gets modified due to inclusion of the EF $\varphi_{\sigma^2(\varepsilon)}$. It is easily

seen that in Step III, $[\eta_i | \cdot]$ can be obtained using the Metropolis-Hastings (MH)-step with a proposal distribution given by the earlier closed form of $[\eta_i | \cdot]$ where σ_ε^2 is not part of the likelihood.

Now, for estimating σ_ε^2 , we add a fourth step.

Step IV. $[\sigma_\varepsilon^2 | \cdot]$

It is similar to $[\beta | \cdot]$ because σ_ε^2 is treated like a mean parameter via EF. So,

$$[\sigma_\varepsilon^2 | \cdot] = \text{Const} \times N\left(\hat{\sigma}_{\varepsilon, \text{mode}}^2, V_{\sigma^2(\varepsilon)} / w_{++}^2\right) I_{\{\sigma_\varepsilon^2 > 0\}} \quad (6.15)$$

where $\hat{\sigma}_{\varepsilon, \text{mode}}^2 = \sum_i^M \sum_j^{n_i} (y_{ij} - x'_{ij}\beta - \eta_i)^2 w_{ij} / w_{++}$, and $w_{++} = \sum_i \sum_j w_{ij}$ is typically constant in practice due to weight calibration.

Before moving to the next section, we remark that in the HB framework, to get a reasonable shrinkage of the prior estimates of η_i toward the direct estimates, we need most of the η_i 's manifested in the sample. If the sampling design is such that this is not the case (e.g., if η_i 's correspond to random PSU effect), then we are faced with an undesirable scenario in which there is hardly any shrinkage of prior estimates of η_i 's. It is interesting to note an analogy of the above situation with the model-based estimation in survey sampling under the prediction approach, where the model-based predictor of the unobserved part of the population is simply given by the synthetic estimator.

7. SIMULATION EXPERIMENT

We design our study along the lines of Pfeffermann et al. (1998). Consider a universe of $i = 1, \dots, M$ strata (small areas) where $M = 100$ and let N_i denote the number of population members in stratum- i . In this simulation experiment, we set $N_i = N_0 (1 + \exp(u_i^*))$ where N_0 is a constant and u_i^* is obtained by truncating $u_i \sim N(0, 0.2)$ at $\pm\sqrt{0.2}$. For simplicity, we consider a single covariate super-population linear mixed model $y_{ij} = \beta_0 + x_{ij} \beta_1 + \eta_i + \varepsilon_{ij}$ where $\beta_0 = 0.5$, $\beta_1 = 1$, $\eta_i \sim N(0, 0.2)$, $\varepsilon_{ij} \sim N(0, 4)$, and $j = 1, \dots, N_i$. The covariate $x_{ij} = \nu_i + \delta_{ij}$ where $\nu_i \sim N(0, 0.1)$ and $\delta_{ij} \sim N(0, 1)$. We generate $K = 150$ population level data sets with common x_{ij} and N_i where N_i 's are generated using $N_0 = 3000$. Note that the substratum sizes vary over

the 150 populations. We selected two samples from each of these populations. The first sample was selected in such a way that the design was ignorable. The second sample was selected so that the design was nonignorable.

To select a sample with an ignorable design, we further stratify the stratum- i population into two substrata Ω_{i+} with $x_{ij} > 0$ and Ω_{i-} with $x_{ij} \leq 0$. To select a sample with nonignorable design, we stratify the stratum- i population into two substrata Ω_{i+} with $\varepsilon_{ij} > 0$ and Ω_{i-} with $\varepsilon_{ij} \leq 0$. Let N_{i+} , N_{i-} denote the sizes of these substrata and n_{i+} , n_{i-} denote the sizes of the simple random samples selected without replacement from these strata, respectively. Note that the substratum sizes vary across populations. Let $N = \sum_{i=1}^{100} N_i$ and $n = \sum_{i=1}^{100} n_i$ where $n_i = n_{i-} + n_{i+}$. For 150 populations, we generate the corresponding 150 samples. In our simulation experiment, $N = 628897$, $n_{i-} = 60$ and $n_{i+} = 20$ so that we have a sample of size 80 for each small area with a total sample of size 8000.

In our simulation study, we compare EFGL, FH, unweighted HB, and PHB (Pseudo-hierarchical Bayes method of You and Rao, 2003) solutions by comparing average posterior means and standard deviations of the parameters of interest. We also compare average 95% prediction interval coverage probabilities as well as the average lengths of 95% prediction intervals. These averages are taken over 150 replications corresponding to populations with varying η_i 's. The comparisons are made for samples generated under ignorable and nonignorable designs. For the FH method, we used a HB-version obtained from EFGL by transforming the unit-level auxiliary information to the aggregate-level, i.e., replacing x_{ij} with $\bar{X}_i = (\sum_{j=1}^{N_i} x_{ij}) \div N_i$. For the PHB method, we used version 2 of You and Rao (2003).

For each sample ($s = 1, \dots, 150$), using Gibbs sampling technique, we generate 10,000 MCMC samples for each of the model parameters, namely $\beta_0, \beta_1, \eta_1, \dots, \eta_M$, and σ_η^2 . These MCMC samples are tested for convergence criterion using CODA (Convergence Output Data Analysis software). First 1000 MCMC samples are deleted for “burn-in” period and from the rest of the 9000 MCMC samples we selected every ninth sample to minimize any auto-correlation among samples, yielding a final MCMC sample of size 1000.

Let $\theta_{sc} = (\beta_{0sc}, \beta_{1sc}, \eta_{isc}, \sigma_{\eta sc}^2)$ denote the parameter values from the c -th MCMC cycle corresponding to the s -

th sample. In Table 1, the average posterior mean of θ is defined as $(\sum_{s=1}^{150} \sum_{c=1}^{1000} \theta_{sc}) \div (1000 \times 150)$ and the average posterior standard deviation of each element of θ_{sc} is defined as the square root of $(\sum_{s=1}^{150} \sum_{c=1}^{1000} (\theta_{sc} - \bar{\theta}_s)^2) \div (1000 \times 150)$ where $\bar{\theta}_s = (\sum_{c=1}^{1000} \theta_{sc}) \div 1000$. Let $\Theta_{isc} = \beta_{0sc} + \bar{X}_i \beta_{1sc} + \eta_{isc}$ denotes the small area estimate from the s -th sample for the i -th area using the c -th MCMC cycle where $\bar{X}_i = (\sum_{j=1}^{N_i} x_{ij}) \div N_i$. Also, define $\Theta_{is}^* = \beta_0 + \bar{X}_i \beta_1 + \eta_{is}$ where η_{is} is the true value of η_i for the s -th population. Let L_{is} and U_{is} denote 2.5 and 97.5 percentiles of the posterior distribution of Θ_{is} obtained from 1000 MCMC samples of Θ_{isc} .

Define $\psi_{is} = \begin{cases} 1 & \text{if } \Theta_{is}^* \in [L_{is}, U_{is}] \\ 0 & \text{otherwise.} \end{cases}$

The coverage probability distribution characteristics given in Tables 2 are obtained from the distribution of 100 area- i specific values of $(\sum_{s=1}^{150} \psi_{is}) \div 150$.

8. SIMULATION RESULTS

Tables 1 and 2 summarize the simulation results obtained from the ignorable sample design, whereas Tables 3 and 4 present the corresponding results for the nonignorable samples. In Table 1, average posterior means and standard deviations for the EFGL method are compared with solutions from a HB version of the FH model, PHB and unweighted solutions for the ignorable sample design. Since the model holds in the sample, the unweighted solution is expected to be the most efficient solution. The average posterior means for all four methods are very close to each other. The average posterior standard deviation of β_1 for the FH model is approximately 13 times larger than the other methods. This is due to the fact that the FH solution uses aggregate-level covariate information. However, the average posterior standard deviations of β_0 and σ_η^2 for all the solutions are very close to each other.

In Table 2, 95% prediction interval coverage probabilities for the EFGL solution are compared with the FH, PHB, and unweighted HB solutions coverage probabilities. The coverage probabilities for all solutions are very close. However, the prediction intervals for the FH solution are 16% wider than the EFGL solution, which is expected, since the EFGL solution utilizes unit-level covariate information whereas the FH solution uses aggregate-level covariate information. The unweighted HB method, being the most efficient for the

ignorable sample design, results in prediction intervals that are approximately 10% shorter than the EFGL solution.

For the nonignorable sample design, Table 3, shows that the average posterior mean for β_0 from the unweighted solution is heavily biased (0.1043 vs 0.5) due to the fact that we over-sample the Ω_{i-} substrata. On the other hand, the average posterior means for the FH, EFGL and PHB solutions are very close to each other. The average posterior standard deviations of β_0 and σ_η^2 for all four solutions are also close to each other whereas the average posterior standard deviation for β_1 from the EFGL, PHB and unweighted solutions are more than 12 times smaller than the solution from the FH model.

From Table 4 (for the nonignorable sample design), we see that 95% coverage probabilities for the EFGL solution and FH solution are very close to each other whereas the coverage probabilities for the PHB solution are approaching 1 and the coverage probabilities for unweighted solution are close to 0. The unweighted method performed very poorly due to the heavily biased estimate of β_0 . It suggests that for our nonignorable samples, the PHB solution substantially overestimates the SAE posterior variances. The prediction intervals for the FH, PHB, and unweighted solutions are respectively 86%, 52%, and 32% wider than the EFGL solution. The inefficiency in the FH solution is expected for the reasons mentioned earlier, since the EFGL solution utilizes unit-level covariate information whereas the FH solution uses aggregate-level covariate information.

9. MIXED NONLINEAR MODELS: LOGISTIC CASE

The method of EFGL introduced in Section 5 for finding HB-SAE in the context of mixed linear models can be easily applied to mixed nonlinear models, the only difference being that full conditional posteriors of β and η have no longer analytic solutions. Therefore, as expected, the method gets more computer intensive. To illustrate the ideas, we consider a simpler version of the mixed logistic model (1.1) given by:

$$\begin{aligned}
 y_{ij} &= \mu_{ij} + \varepsilon_{ij}, \quad y_{ij} \sim \text{Bernoulli} \\
 \text{logit}(\mu_{ij}) &= x_{ij}'\beta + \eta_i \\
 \eta_i &\sim_{iid} N(0, \sigma_\eta^2), \quad \beta \sim U(R^p), \quad \sigma_\eta^2 \sim IG(v_0/2, \sigma_{\eta_0}^2/2).
 \end{aligned} \tag{7.1}$$

The EFs in this case remain similar to the linear case except that the elementary zero functions (or the residuals) $y_{ij} - \mu_{ij}$, are complex due to the nonlinear form of μ_{ij} 's. Observe that EFs continue to be simple linear functions of elementary zero functions, and hence they behave well in terms of Gaussian approximations. The EFs for the logistic case under Scenario I are given by

$$\begin{aligned}\varphi_{\eta(i)} &= \sum_{j=1}^{n_i} (y_{ij} - \mu_{ij}) w_{ij} \sim_{approx} N(0, V_{\eta(i)}) \\ \varphi_{\beta} &= \sum_i \sum_j x_{ij} (y_{ij} - \mu_{ij}) w_{ij} \sim_{approx} N(0, V_{\beta})\end{aligned}\quad (7.2)$$

We can orthogonalize φ_{β} with respect to $\varphi_{\eta(i)}$'s as before. Also with the intercept model, φ_{β} corresponding to the intercept should be dropped because of its linear dependence on $\varphi_{\eta(i)}$'s. Now, the likelihood, $L(y^* | \beta, \eta)$ can be approximately specified as before, but the MCMC steps are modified as follows:

Step I. $[\beta | y^*, \eta]$

Since the sample is typically very large, the full conditional posterior can be well approximated by

$$[\beta | y^*, \eta] \sim N(\hat{\beta}_{\text{mode}}, \Sigma_{\psi(\beta)}^{-1}) \quad (7.3)$$

where $\hat{\beta}_{\text{mode}}$ solves $\psi_{\beta} = 0$, $\Sigma_{\psi(\beta)} = -E((\partial/\partial\beta)\psi_{\beta})$,

$$\begin{aligned}\psi_{\beta} &= (\partial/\partial\beta) \log L(y^* | \beta, \eta) \\ &= \sum_{i=1}^M \varphi_{\eta(i)} \sum_{j=1}^{n_i} x_{ij} \mu_{ij} (1 - \mu_{ij}) w_{ij} / V_{\eta(i)} - \left(\sum_i \sum_j x_{ij} x'_{ij} \mu_{ij} (1 - \mu_{ij}) w_{ij} - \sum_i a_i \mu_i (1 - \mu_i) x_{ij} w_{ij} \right) \tilde{V}_{\beta}^{-1} \tilde{\varphi}_{\beta}\end{aligned}\quad (7.4)$$

Note that unlike the linear case, $\hat{\beta}_{\text{mode}}$ does not have an analytic form. Also note that instead of the approximate posterior (7.3), one can get realizations from an exact posterior by using the MH step within MCMC in which (7.3) can be used as a proposal.

Step II. $[\eta_i | \eta_{i'}, \beta, y^*, \sigma_{\eta}^2]$, $i = 1, \dots, M$.

As mentioned earlier, this again does not have an analytic solution. We could use MH with mle/prior for the proposed distribution. In other words, solve $\psi_{\eta(i)} - \sigma_{\eta}^{-2} \eta_i = 0$ to get $\hat{\eta}_{i, \text{mle-adj}}$, where $\psi_{\eta(i)} = (\partial/\partial\eta_i) \log L(y^* | \beta, \eta)$, and use $N(\hat{\eta}_{i, \text{mle-adj}}, (\sigma_{\psi(\eta(i))}^2 + \sigma_{\eta}^{-2})^{-1})$ as the proposal distribution where

$$\sigma_{\psi(\eta(i))}^2 = -E\left[\partial\psi_{\eta(i)}/\partial\eta_i\right].$$

Step III. $[\sigma_{\eta}^2 | \eta]$

We get the same result as in the linear case. Note that Step IV for $[\sigma_{\epsilon}^2 | \cdot]$ is not needed because σ_{ϵ}^2 is a known function of μ_{ij} in the logistic case.

We now consider the work of Folsom et al. (1999) mentioned earlier in Sections 1 and 2 which is related to the proposed EFGM method. For the logistic model, they constructed a pseudo log-likelihood (from the Bernoulli likelihood at the census-level) involving design weights. For this purpose, survey weights were scaled such that they sum to the effective sample size obtained by using the design effect within each area i. The design effect was, however, based only on the effect of unequal weighting under the working assumption of ignorability of the design. In other words, effects of stratification, clustering, and multistage designs were ignored.

Under their pseudo-likelihood approach, the score function for η_i involves $\varphi_{\eta(i)}$ multiplied by a scale adjustment for weights. This pseudo score function in conjunction with the prior information gives the appropriate prior-adjusted pseudo-mle for random effects. This prior-adjusted pseudo-mle along with its variance can be used for defining a Gaussian proposal distribution for the MH step in finding the full conditional posterior of η_i . In the case of β , the actual pseudo score function obtained from the pseudo likelihood was, however, not used, but a somewhat modified pseudo score function, namely ϕ_{β} obtained from the census likelihood was used as it has the appealing property of self-calibration or benchmarking explained later on. Note that the actual pseudo score function for β is not proportional to ϕ_{β} because of weight scaling. However, $\hat{\beta}_{\text{pseudo mle}}$ obtained by solving $\varphi_{\beta} = 0$, and the associated sandwich V-C matrix $(\partial\varphi_{\beta}/\partial\beta')^{-1} \Sigma_{\phi} (\partial\varphi_{\beta}'/\partial\beta)^{-1}$ used respectively as the mode and curvature of a Gaussian distribution is likely to be close to the conditional posterior based on the actual pseudo-score function for β . Here the V-C matrix Σ_{ϕ} is computed under the working assumption of ignorable designs, and thus reflects only unequal weighting effect. It may be noted that use of the sandwich V-C (and not the pseudo information) matrix is appropriate because the likelihood is pseudo.

For computing, $[\sigma_{\eta}^2 | \cdot]$, the distribution of Step III of EFGL was used. Thus, the above pseudo-likelihood approach has some similarity with the proposed EFGL. The main differences are that the likelihood is pseudo which need not be valid, and the working assumption of ignorable sample design may not be reasonable. In EFGL, the likelihood is based on EFs and approximated by a valid Gaussian likelihood where the covariance matrix takes full account of the design. However, in the NHSDA application, it was observed that the MCMC method for the pseudo-likelihood approach did converge and provided good results. Also, it can be shown that HB-SAE estimates based on the pseudo-likelihood approach have the desirable property of approximate self-calibration or benchmarking because SAEs obtained directly from pseudo score functions are very similar for large samples to the direct SAEs which are, of course, design-consistent. Thus, SAEs for big states will be approximately equivalent to the direct estimates. Also, aggregates of SAE estimates are nearly calibrated to the national direct estimates. By contrast, estimates resulting from the method of EFGL, although design-consistent, need to be modified to achieve benchmarking to direct estimates for areas with very large samples, see e.g., Singh and Folsom (2001).

10. CONCLUDING REMARKS

The method of EFGL was developed to exploit unit-level information, to take full account of the survey design, and to have a valid (approximate) likelihood for the HB-SAE methodology for generalized linear mixed models. It generalizes the aggregate-level model of FH (1979), and the pseudo-likelihood approach of Folsom et al. (1999). There are essentially two main ideas in EFGL, namely, the data aggregation via EFs and EF-collapsing. The main reason for EF-collapsing is to improve Gaussian approximation, and the secondary purpose is to improve the variance estimate's precision. In practice, it may be preferable to use separate modeling to make variance estimates more stable. However, even if variance estimates are not precise, it is often of interest, in practice, to see how much variance reduction can be realized through SAE modeling.

The idea of data aggregation in EFGL is somewhat similar to that of FH except it tries to take advantage of the unit-level information as much as possible. Since EFGL uses more information than FH, the resulting estimates are expected to be more efficient than those from FH. In particular, for the case of simple linear mixed models (2.1) with known variance components, it can be easily shown analytically that precision of the estimates of fixed effects (β) can be improved substantially in the case of unit-level models if the covariates (x_{ij}) have sufficient variability within areas. There is also some gain in efficiency of random effect

(η_i 's) estimates. However, if η_i 's are also defined as coefficients of suitable covariates (z_{ij} 's) as in the case of random regression coefficients, then high efficiency gains in estimating random effects can also be realized if there is sufficient variability in z_{ij} 's within areas.

We remark that the problem of HB-SAE arose in the context of NHSDA-SAE application where it was desired to fit a mixed logistic linear model. This was a daunting SAE application task with a very large data set and many covariates which was addressed by Folsom et al. (1999). Note that it was not possible to use any existing software for this task.

The ideas underlying the proposed method of EFGL are quite general, and the method is applicable to general nonlinear mixed models for survey data. However, it does have some limitations which the user should keep in mind: (i) Some loss of efficiency is inevitable due to data aggregation, and EF-collapsing. This is the price we pay for not having enough information about the likelihood of the sampled data, and by not being able to ignore the sample design. (ii) The EF-collapsing may be needed for the Gaussian approximation. In practice, it is better to avoid it if possible as it doesn't distinguish much between the areas involved in collapsing. At the design stage, one can take measures to ensure a sufficient number of observation in each small area in order to avoid EF-collapsing. It may be noted that one only needs a modest size of the realized sample in small areas for Gaussian approximation of EFs. However, SAEs are still needed for precise estimation.

Finally we mention an interesting problem (not on SAE though) considered by Pfeiffermann et al. (1998) on multi-level modeling (such as the mixed linear model (2.1)) for survey data for estimating fixed effects (β) and variance components ($\sigma_\eta^2, \sigma_\varepsilon^2$). Here we don't have the problem of small area estimation, and the random effects η_i are defined at the PSU-level which is lower than the area level. Under a frequentist approach, they proposed a probability-weighted iterative GLS for estimating all the fixed parameters which requires knowledge of both first-stage (π_i) and second stage (π_{ji}) selection probabilities separately, and a large number of PSUs as well as a large number of second stage units within each PSU to ensure consistency of the variance component estimates. In practice, since it is not realistic to assume large second stage sample sizes, the authors proposed scaling the weights as an option to reduce small sample bias. For a Bayesian approach as an alternative, if second order inclusion probabilities were known, it would be fairly straightforward to construct EFs for $\beta, \sigma_\eta^2, \sigma_\varepsilon^2$, and then the method of EFGL could be used to produce HB-SAE for these parameters. However, if only first order inclusion probabilities are known, as is often the case,

we need to modify the EFGM method. In its present form it doesn't seem applicable, because most PSUs need to be manifested in order to have a reasonable shrinkage as mentioned earlier in Section 6. A way to modify EFGM would be to include an additional EF of the form $(\sum_{i=1}^M 1_{i \in s} \eta_i^2 / \pi_i - \sum_{i=1}^M \eta_i^2)$ to account for the first stage of selection of PSU-level random effects in estimating σ_η^2 , and to allow for collapsing of PSUs, if necessary, for Gaussian approximation of EFs. Note that under the usual with-replacement assumption of PSUs, design-based variances of PSU-level EFs can be estimated within each design stratum provided there are at least two PSUs per stratum.

ACKNOWLEDGMENTS

This research was supported in part by grants from U.S. Census Bureau (through NSF), from NCHS of CDC to the University of North Carolina at Chapel Hill, and from the Behavioral Sciences Branch of CDC, Department of Health and Human Services) under Contract No. 200-98-0103. The first author's research was also supported in part by a grant from Natural Sciences and Engineering Research Council of Canada held at Carleton University, Ottawa under an adjunct research professorship. Earlier versions of this paper were presented at the International Conference on Small Area Estimation and Related Topics, Potomac, MD, April 11-14, 2001, and at the FCSM conference, Arlington, VA, Nov 14-16, 2001. The authors would like to thank Jon Rao, Bob Fay, Bill Bell, Don Malec, Malay Ghosh, and Nat Schenker for their helpful and encouraging comments.

REFERENCES

- Binder, D.A. (1983), "On the Variances of Asymptotically Normal Estimators from Complex Surveys," *International Statistical Review*, **51**, pp. 279-292.
- Cassel, C.M., Särndal, C.E., and Wretman, J.H (1976), "Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations," *Biometrika*, **63**, pp. 615-620.
- Datta, G. S. and Ghosh, M. (1991), "Bayesian Prediction in Linear Models: Applications to Small Area Estimation. *Annals of Statistics*," **19**, pp.1746-1770.
- Fay, R.E. and Herriot, R.A. (1979), "Estimates of Income for Small Places: An Application of James-Stein

Procedures to Census Data,” *Journal of the American Statistical Association*, **74**, pp. 269-277.

Folsom, R.E., Shah, B.V. and Vaish, A. (1999), “Substance Abuse in States: A Methodological Report on Model Based Estimates from the 1994-96 NHSDAs,” *Proceedings of the Survey Research Section, American Statistical Association*, pp. 371-375.

Fuller, W.A. (1975), “Regression Analysis for Sample Surveys,” *Sankhya*, Ser C, **37**, pp. 117-132.

Godambe, V.P. (1960), “An Optimum Property of Regular Maximum Likelihood Estimation,” *Annals of Mathematical Statistics*, **31**, pp. 1208-1212.

Godambe and Thompson, M.E, (1986), “Parameters of Super population and Survey Population, Their Relationship and Estimation,” *International Statistical Review*, **54**, pp. 127-38.

Kass, R. E. and Steffey, D. (1989), “Approximate Bayesian Inferences in Conditionally Independent Hierarchical Models (Parametric Empirical Bayes Models),” *Journal of the American Statistical Association*, **84**, pp. 717-726.

Kott, P. E. (1989), “Robust Small Area Estimation Using Random Effect Modeling,” *Survey Methodology*, **15**, pp. 3-12.

McCullagh, P. (1991). “Quasilikelihood and estimating functions”, In *Statistical Theory and Modelling: In honour of Sir David Cox, FRS*, ed. D.V. Hinkley, N. Reid, and E.J. Snell, London: Chapman and Hall, 265-286.

Otto, M.C., and Bell, W.R. (1995), “Sampling Error Modeling of Poverty and Income Statistics for States,” *Proceedings of the Government Statistics Section, American Statistical Association*, pp. 160-165.

Pfeffermann, D. and Sverchkov, M. (1999). Parametric and semi-parametric estimation of regression models fitted to survey data. *Sankhya: The Indian Journal of Statistics*, Ser. B, **61**, 166-186.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. (1998), “Weighting for Unequal Selection Probabilities in Multilevel Models,” *Journal of the Royal Statistical Society*, B, **60**, pp. 23-40

Prasad, N.G.N. and Rao, J.N.K. (1999), “On Robust Small Area Estimates Using a Simple Random Effects Model,” *Survey Methodology*, **25**, pp. 67-72.

Singh, A.C., Stukel, D.M. and Pfeffermann, D. (1998), “Bayesian versus Frequentist Measures of Error in Small Area Estimation,” *Journal of the Royal Statistical Society, B*, **60**, pp. 377-396.

Singh, A. C. and Folsom, R. E. (2001), “Benchmarking of Small Area Estimators in a Bayesian Framework,” *International Conference on Small Area Estimation and Related Topics*, Potomac, MD April 11-14.

U.S. Census Bureau (1998), “1993 Small Area Income and Poverty Estimates (SAIPE); Small Area Estimates of School-Age Children in Poverty,” *Interim Report 2*, National Academy Press.

You, Y., and Rao, J.N.K. (2003), “Pseudo Hierarchical Bayes Small Area Estimation Combining Unit Level Models and Survey Weights,” *Journal of Statistical Planning and Inference*, **11**, 197-208.

Table 1: Average Posterior Mean and Standard Deviation for Model Parameters: Ignorable Sample Design

Parameter (True Value)	Average Posterior Mean				Average Posterior Standard Deviation			
	FH	EFGL	PHB	Unweighted	FH	EFGL	PHB	Unweighted
β_0 (0.5)	0.5009	0.5020	0.5020	0.5024	0.0473	0.0461	0.0482	0.0461
β_1 (1.0)	0.9946	0.9988	0.9989	0.9983	0.1650	0.0129	0.0131	0.0121
σ_η^2 (0.2)	0.1970	0.1974	0.1981	0.1981	0.0318	0.0309	0.0303	0.0303

Table 2: 95% Coverage Probability and Ratio of Prediction Interval (PI) Widths: Ignorable Sample Design

Percentiles and Means over Small Areas	Coverage Probability				Ratio of Average PI Widths		
	FH	EFGL	PHB	Unweighted	FH/EFGL	PHB/EFGL	Unweighted/EFGL
95%	0.973	0.970	0.973	0.980	1.19	1.03	1.00
75%	0.953	0.953	0.960	0.967	1.17	1.02	0.97
50%	0.940	0.940	0.953	0.953	1.16	1.01	0.91
Mean	0.942	0.941	0.950	0.950	1.16	1.01	0.89
25%	0.930	0.933	0.940	0.937	1.15	1.00	0.83
5%	0.913	0.907	0.913	0.920	1.14	1.00	0.75

Table 3: Average Posterior Mean and Standard Deviation for Model Parameters: Nonignorable Sample Design

Parameter (True Value)	Average Posterior Mean				Average Posterior Standard Deviation			
	FH	EFGL	PHB	Unweighted	FH	EFGL	PHB	Unweighted
$\beta_0(0.5)$	0.5043	0.5029	0.5029	0.1043	0.0472	0.0450	0.0459	0.0448
$\beta_1(1.0)$	1.0014	1.0004	1.0006	0.9999	0.1638	0.0131	0.0121	0.0103
$\sigma_\eta^2(0.2)$	0.1972	0.1977	0.1909	0.1909	0.0319	0.0294	0.0290	0.0290

Table 4: 95% Coverage Probability and Ratio of Prediction Interval (PI) Widths: Nonignorable Sample Design

Percentiles and Means over Small Areas	Coverage Probability				Ratio of Average PI Widths		
	FH	EFGL	PHB	Unweighted	FH/EFGL	PHB/EFGL	Unweighted/EFGL
95%	0.973	0.970	1.000	0.007	1.91	1.54	1.35
75%	0.953	0.953	1.000	0.000	1.88	1.53	1.33
50%	0.940	0.933	0.993	0.000	1.86	1.52	1.32
Mean	0.941	0.933	0.995	0.001	1.86	1.52	1.32
25%	0.927	0.913	0.993	0.000	1.84	1.50	1.31
5%	0.910	0.897	0.987	0.000	1.82	1.49	1.30

Discussion of “Estimating Function Based Approach to Hierarchical Bayes Small Area Estimation for Survey Data”

Phillip S. Kott

National Agricultural Statistics Service

Introduction

In their intriguing paper, Singh, Folsom, and Vaish develop an Estimating-Function Hierarchical Bayesian (EFHB) methodology to replace the standard Fay-Herriot (F-H) model for small-domain estimation. I will discuss two limitations of the F-H model overcome by their EFHB methodology and two other problems that are not. This leads to the obvious question: Why combine estimating functions and hierarchical Bayesian models in the way the authors choose?

The Fay-Herriot Model

Suppose we have M small domain totals (or means) satisfying the model:

$$Y_{i+} = \mathbf{X}_{i+}\boldsymbol{\beta} + \eta_i, \quad \eta_i \sim N(0, \sigma_\eta^2).$$

Suppose further that each of the component of the row vector \mathbf{X}_{i+} is known, but each of the Y_{i+} has a randomization-based estimator:

$$y_{i+(RB)} = Y_{i+} + d_i, \quad d_i \sim N(0, V_i) \text{ approximately.}$$

A better estimator for Y_{i+} is

$$y_{i+(\lambda)} = (1 - \lambda)y_{i+} + \lambda \mathbf{X}_{i+} \mathbf{b},$$

where \mathbf{b} is an unbiased estimator for $\boldsymbol{\beta}$, $\lambda = v_i / (v_i + s_\eta^2)$ when M is large, v_i is a randomization-based estimator for V_i , and s_η^2 is an estimator for σ_η^2 . A nice property of $y_{i+(\lambda)}$ is that as the sample size within domain i increases, so that V_i (and v_i) tends towards 0 under mild conditions, $y_{i+(\lambda)}$ approaches $y_{i+(RB)}$. Consequently, if $y_{i+(RB)}$ is *randomization consistent* (approaches Y_{i+} as the sample size within i grows arbitrarily large), then so is $y_{i+(\lambda)}$.

The Estimating-Function Hierarchical Bayesian Methodology

Let j be a unit within area i . The authors' EFHB technology lets us expand the F-H area-level model to the unit-level:

$$y_{ij} = \mathbf{x}_{ij}\boldsymbol{\beta} + \eta_i + \varepsilon_{ij}, \quad \text{where } E(\varepsilon_{ij}) = 0.$$

The model holds for the population, but not necessarily for the sample. In other words, the design may be informative. In this, the authors part company from most of the small-domain literature.

If $\mathbf{X}_{i+} = \sum_{j \in U(i)} \mathbf{x}_{ij}$ is known, and \mathbf{x}_{ij} is not constant within each area, then the EFHB technology produces a better estimator than $y_{i+(\lambda, F-H)}$ *under the model*. Moreover, as V_i (and v_i) approaches 0, $y_{i+(HB)}$ approaches Y_{i+}

In addition, the EFHB technology allows models of the form:

$$y_{ij} = \mu(\mathbf{X}_i\boldsymbol{\beta} + \eta_i) + \varepsilon_{ij},$$

where $\mu(\cdot)$ need not be the identity. This is particularly helpful when y_{ij} is a 0/1 variable. In that situation, $\mu(\cdot)$ can be logistic. Unfortunately, there is a limited ability to replace \mathbf{X}_{i+} with \mathbf{x}_{ij}

The EFHB methodology uses randomization-based estimators for V_i , but such estimators are notoriously error-prone when based on small samples. Collapsing domains won't help when the estimator, v_i , is zero but V_i is positive.

Another problem with the authors' EFHB methodology is that it is not *self-benchmarking*. A methodology having this property produces domain estimators satisfying

$$\sum_{i=1}^M y_{i+} = \sum y_{i+(RB)} = y_{++(RB)},$$

where $y_{++(RB)}$ is model free yet has a small variance. It should be noted that the standard F-H approach is likewise *not* self-benchmarking.

Why Estimating Functions?

Arguably the first model-assisted paper (Godambe 1955) requires the estimator to be randomization unbiased. The probability-weighted ratio (and regression) estimator can have good model-based properties but has a potential randomization bias. The correct way, in my view, to deal with the randomization bias of the certain probability-weighted estimators is to change the requirement from randomization unbiasedness to randomization consistency (Isaki and Fuller 1982), which assures that the estimator in question be close to what it estimates almost surely when the sample is large enough. The wrong way is to observe that an estimator like the probability-weighted ratio can be derived from the solution to an unbiased estimating equation (e.g., Godambe 1960, Godambe &

Thompson, 1986). This “wrong way” uncovered a technique that has found many practical uses, however. It is a useful technique built on a dubious theory.

Singh, Folsom, and Vaish use estimating functions (a mild generalization of estimating equations) to generate estimators that are randomization-consistent, but not self-benchmarking. With an alternative approach, You and Rao (2003) use estimating functions to produce estimators that are both randomization-consistent and self-benchmarking. They do this by modeling the sampling variance under an ignorable model. Unfortunately, they assume a linear $\mu(\cdot)$.

My Bottom Line

The less data we have the more we need models. Models with pre-determined functional forms have more power than semi-parametric models. Furthermore, hierarchical Bayesian models allow $\mu(\cdot)$ to be nonlinear.

Combining estimating functions Bayesian models appear to give us the best of both worlds, the robustness of estimating functions and the power of Bayes. The former’s reliance on the asymptotic normality of probability-weighted estimators, however, undercuts the advantage of a latter. We also need to ask whether sampling weights are needed because:

1. The model is correct in the population but not necessarily correct in the sample OR
2. The model may be wrong in both the sample and the population.

By positing the first, which is what the authors do, $E_{\text{model}}([y_{i+(RB)} - Y_i]^2 | \text{sample})$ cannot be estimated directly. Instead, one invokes the equality,

$$E_{\text{model}} \{E_{\text{rand}}([y_{i+(RB)} - Y_i]^2)\} = E_{\text{rand}} \{E_{\text{model}}([y_{i+(RB)} - Y_i]^2 | \text{sample})\},$$

and estimates the randomization variance for domain i , $V_i = E_{\text{rand}}([y_{i+(RB)} - Y_i]^2)$. This is often not a trivial thing to do well even with largish samples. In my view, it is much more sensible to accept the second position. Using sampling weights provides some asymptotic protection against the model being wrong in the population itself. Nevertheless, model-based parameters and predictors should be estimated as if the model were correct *and* the design noninformative. One can then estimate $E_{\text{model}}([y_{i+(RB)} - Y_i]^2 | \text{sample}) = E_{\text{model}}([y_{i+(RB)} - Y_i]^2)$ with relative ease, and the resulting estimator will usually have much more power than a randomization-based estimator for V_i . This is the approach effectively taken by You and Rao.

Singh, Folsom, and Vaish’s EFHB approach allows them to incorporate a nonlinear $\mu(\cdot)$ into their model. I wonder if that is enough to justify their having to rely on estimated randomization variances and put up with the inconvenience of domain estimators that are not self-benchmarking.