A Simulation Study to Compare Unit-Level Linear Mixed Model Methods for Small Area Estimation for Survey Data

A. K. Vaish*, R. E. Folsom, and A. C. Singh
RTI International, Research Triangle Park, NC 27709
***avaish@rti.org**

KEY WORDS: Estimating Functions, Pseudo Score Functions, Survey Weighted Hierarchical Bayes, MCMC, and Pseudo Hierarchical Bayes

1. Introduction

In this paper we compare, via a simulation study, three small area estimation (SAE) methods for survey data. The first method, known as Estimating Function Gaussian Likelihood (EFGL), was developed by Singh, Folsom, and Vaish (2002a and 2002b); the second method, known as Pseudo Hierarchical Bayes (PHB), was developed by You and Rao (2003); and the third method is a hierarchical Bayes (HB) solution for the Fay and Herriot (F-H) (1979) model. For motivation, technical details, statistical properties, and performance of the three methods we refer to the above papers.

The present paper, in comparison to our previously published papers, makes the following contributions. Unlike the comparisons of the EFGL method with a HB version of the F-H solution and unweighted solution for nonignorable sample designs reported earlier, we present new simulation results with samples generated using ignorable and nonignorable sample designs. Only a brief description of the EFGL, PHB, and F-H methods is presented here while simulation study and results are presented in details.

2. Desirable Properties of Small Area Estimates (SAEs)

Most of the small area models for survey data assume that the sample design is ignorable i.e. the distribution of errors associated with the assumed finite population model remains the same for the sample. But it is often the case that the population model cannot be assumed to hold for the sampled data due to selection bias. If selection bias is present then standard model based solutions will lead to biased estimates for small areas such as posterior means and variances. Hence it is important for a small area estimator to be robust against nonignorable sample designs.

SAEs obtained from the EFGL and F-H models are robust against nonignorable sample design because sample design is fully accounted for in estimating the model parameters; whereas the PHB model assumes that the sample design is ignorable when estimating the variance components. Hence it produces biased SAE posterior variances when the design is nonignorable.

A natural way to overcome the presence of nonignorable sample designs in the joint design-model based estimation framework is to employ transformed or aggregate-level data such as the direct survey estimates for small areas. With aggregate-level data, one can take account of the sampling design in specifying the likelihood of model parameters by appealing to the central limit theorem for large samples; here it is assumed that variances of the direct estimates can be treated as known which, in practice, amounts to smoothing them via generalized variance functions. If the preferred model is unit-level which is often the case in practice, there is clearly a loss of efficiency by using aggregate-level data. The EFGL and PHB methods allow the use of unit-level data in the model whereas the F-H model works on aggregate-level data. Hence the F-H model produces less efficient SAEs as compared to the SAEs produced by EFGL and PHB methods.

The use of unit-level information also makes the SAEs internally consistent i.e. SAEs for higher levels (say states) are obtained by aggregating the SAEs at a lower level (say counties) by using appropriate population counts. The EFGL and PHB methods produce internally consistent SAEs whereas SAEs obtained by the F-H method are not internally consistent. For the F-H solution, if SAEs at the state-level are desired then data is aggregated at the state-level and the model is fitted to the state-level data. If county-level SAEs are desired then the data is aggregated at the county-level and a county-level model is fitted. In this case if the county-level SAEs are combined using appropriate county-level population counts to produce state-level SAEs then these state- level SAEs will be different from the ones produced by fitting the state-level model to the state-level data.

Often it is desirable to produce SAEs for binary outcome variables using nonlinear mixed models. The PHB approach adopted by You and Rao (2003) does not extend to nonlinear mixed models although the general pseudo hierarchical Bayes methodology is extendable to nonlinear mixed models, see Folsom, Shah, and Vaish (1999). The EFGL methodology is applicable to linear and nonlinear mixed models. In the interest of normality, there may be other concerns with the F-H methodology when formulating a model with transformed aggregate-level estimates e.g. when modeling a low prevalence outcome variable at a lower level of geography such as prevalence of heroin usage in a county, then many counties will be discarded from the model since log (0) is not finite.

It is also desirable that the SAEs converge to the true population parameters when the sample size increases. Such SAEs are called design-consistent (see Rao, 2003 for the definition of design-consistent). All three methods produce design-consistent SAEs. In the next section we give technical details of the three methods.

3. Brief Description of EFGL, PHB, and F-H Models

In this paper we consider a linear mixed model with one covariate. A simulation study for a nonlinear mixed model is in progress. First we give a brief description of the EFGL methodology then we describe PHB and a hierarchical Bayes version of the F-H methodology.

EFGL Methodology: The EFGL methodology is based on survey weighted Estimating Functions (EFs). To illustrate the EFGL methodology, we consider the framework (at the census level) for a linear mixed model with one covariate. Let
$y_{ij} = \beta_0 + \beta_1 x_{ij} + \eta_i + \varepsilon_{ij}$ where $\eta_i \sim N(0, \sigma_\eta^2)$,
$(\beta_0, \beta_1) \sim U(R^2), \sigma_\eta^2 \sim IG(\nu_0/2, \sigma_{\eta_0}^2/2)$,
$y_{ij} | (\beta_0, \beta_1, \eta_i, \sigma_\varepsilon^2) \sim N(\beta_0 + \beta_1 x_{ij} + \eta_i, \sigma_\varepsilon^2), \varepsilon_{ij} \sim N(0, \sigma^2)$, and
$i = 1, \ldots, M\,(strata)$ and $j = 1, \ldots, N_i$.

Note that the EFGL solution does not require an estimate of σ_ε^2. If an estimate for σ_ε^2 is desired, the prior for σ_ε^2 can be assumed to be improper like that of the mean parameters β's. In that case, we can introduce a separate EF, $\phi_{\sigma_\varepsilon^2}$, for σ_ε^2 which treats σ_ε^2 as a mean parameter. It turns out, as expected and as in the case of F-H, that σ_ε^2 is not a part of the variance-covariance matrix of ϕ's (defined below) when a suitable design-based estimate is substituted.

With a nonignorable sample design (small areas as strata), define survey weighted EFs for β_1 and $\eta_1, \ldots, \eta_M$ using the sample data as

$$\phi_{\eta_i} = \sum_{j=1}^{n_i} (y_{ij} - \beta_0 - \beta_1 x_{ij} - \eta_i) w_{ij}, \quad i = 1, \ldots, M \quad and$$

$$\phi_{\beta_1} = \sum_{i=1}^{M} \sum_{j=1}^{n_i} x_{ij} (y_{ij} - \beta_0 - \beta_1 x_{ij} - \eta_i) w_{ij}$$

where the w_{ij} are design weights. The EF for β_0 is simply the sum of ϕ_{η_i}'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, these 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 by a Gaussian distribution even for modest sample sizes (McCullagh 1991) because by their very nature, they are simple sums of elementary zero functions, although the elementary functions can be complex. Moreover, EFs can be easily collapsed to improve the Gaussian approximation as well as the precision of variance estimates.

Now, the vector of EFs serves as the condensed input data which after collapsing, if necessary, gives rise to an approximate Gaussian likelihood, $L(y^* | \beta, \eta)$ where y^* denotes the implicit condensing of information in y via EFs. 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, $L(y^*|\cdot)$, which does not assume ignorability of the design.

Let $\Phi = (\phi_{\eta_1}, \ldots, \phi_{\eta_M}, \phi_{\beta_1})$, and $V_\Phi = Cov(\Phi)$ where V_Φ is a design based variance-covariance matrix of Φ. Further suppose, $\Phi \sim N_{M+1}(0, V_\Phi)$. Due to this assumption, the proposed method is referred to as Estimating Function Based Gaussian Likelihood. The EF Gaussian log-likelihood is given by $l(y^* | \beta, \eta) = const - \frac{1}{2}(\Phi' V_\Phi^{-1} \Phi)$. It may be noted that there is, in fact, a second component involving σ_ε^2 when the variance-covariance matrix of the ϕ's is computed under joint design-model randomization. However, this term involving σ_ε^2 is negligible in comparison to the first term, V_Φ, under the usual assumption of $n_i << N_i$. It should also be emphasized that, in practice, some collapsing of the ϕ_{η_i}'s may often be required because the corresponding n_i's may be small. We may need this collapsing to improve the Gaussian approximation, as well as to improve the precision of the estimate V_Φ. With the specification of EFGL, estimation of parameters $[\eta = (\eta_1, \ldots, \eta_M), \beta = (\beta_0, \beta_1), \sigma_\eta^2]$ can proceed in the HB setup using MCMC steps. The next section gives details of the full conditional posterior distributions needed for MCMC.

For implementing Gibbs sampling, the full conditional posterior distributions are given below:

$$\left[\beta | \eta, data\right] \sim N\left[(X_1' V_\Phi^{-1} X_1)^{-1} X_1' V_\Phi^{-1} Y_1, (X_1' V_\Phi^{-1} X_1)^{-1}\right],$$

$$\left[\eta \mid \beta, \sigma_\eta^2, data\right] \sim$$

$$N\left[(X_2' V_\Phi^{-1} X_2 + \frac{I}{\sigma_\eta^2})^{-1} X_2' V_\Phi^{-1} Y_2, (X_2' V_\Phi^{-1} X_2 + \frac{I}{\sigma_\eta^2})^{-1}\right],$$

$$\left[\sigma_\eta^2 | \eta\right] \sim IG\left[\nu_1, \nu_2\right] \; where$$

$$\nu_1 = (\nu_0 + M)/2, \nu_2 = (\sigma_{\eta_0}^2 + \sum_{i=1}^{M} \eta_i^2)/2, \text{ and}$$

X_1, Y_1, X_2, and Y_2 are defined in Singh, Folsom, and Vaish (2002b). Using the full conditional posterior distributions given above, we can now use the Gibbs sampling technique and generate MCMC samples from the joint posterior distributions of the model parameters as described in the above mentioned paper.

PHB Methodology: You and Rao's PHB solution assumes the sampling mechanism is ignorable and begins by generating an MCMC sample of $(\sigma_\eta^2, \sigma_\varepsilon^2)$ pairs using the unweighted Gibbs algorithm given in the appendix section of You and Rao (2003). For each MCMC sample of $(\sigma_\eta^2, \sigma_\varepsilon^2)$, they then solve for $\beta_w = (\beta_{0w}, \beta_{1w})$ in the survey weighted conditional estimating equations:

$$\sum_{i=1}^{M} \sum_{j=1}^{n_i} w_{ij} x_{ij} [y_{ij} - \beta_{0w} - x_{ij}\beta_{1w} - \gamma_{iw}(\bar{y}_{iw} - \beta_{0w} - \bar{x}_{iw}\beta_{1w})] = 0$$

where $\gamma_{iw} = \sigma_\eta^2 / [\sigma_\eta^2 + (\delta_i \sigma_\varepsilon^2 / n_i)]$, δ_i is Kish's unequal weighting design effect, and $\bar{y}_{iw}$, $\bar{x}_{iw}$ are sample weighted averages of y_{ij}, x_{ij}, respectively. Once β_w is determined then a sample value of β is drawn from the conditional posterior distribution $[\beta \mid \beta_w, \sigma_\eta^2, \sigma_\varepsilon^2] \sim N(\beta_w, \Phi_w)$ where Φ_w is derived using the ignorable sample design assumption and is given in You and Rao (2003).

HB Version of the F-H Model: A HB version of the F-H model is fitted using the improper uniform prior on the fixed parameters and an inverse Gamma prior on σ_η^2 similar to the ones used in the EFGL and PHB methodologies. For this purpose, EFGL modeling was transformed to aggregate-level by replacing x_{ij} with $\bar{X}_i = (\sum_{j=1}^{N_i} x_{ij}) \div N_i$.

4. Simulation Experiment

We design our simulation study along the lines of Pfeffermann et al. (1998). Consider a universe of $i = 1, \cdots, M$ 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 $\beta_0 = 0.5$, $\beta_1 = 1$, $\eta_i \sim N(0, 0.2)$, $\varepsilon_{ij} \sim N(0,1)$, and $j = 1, \cdots, N_i$. The covariate $x_{ij} = \upsilon_i + \delta_{ij}$ where $\upsilon_i \sim N(0, 0.1)$ and $\delta_{ij} \sim N(0,1)$. We generate $K = 150$ population-level data sets with common x_{ij} and N_i where the N_i's are generated using N_0 =3000. We selected two samples from each of these populations. The first sample was selected in such a way that design was ignorable. The second sample was selected so that the design was nonignorable.

To select a sample with an ignorable design we stratify the stratum-i population into two substrata X_{i+} with $x_{ij} > 0$ and X_{i-} with $x_{ij} \le 0$. To select a sample with nonignorable design we grouped the stratum-i population into two substrata E_{i+} with $\varepsilon_{ij} > 0$ and E_{i-} with $\varepsilon_{ij} \le 0$.

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 $n_i = n_{i-} + n_{i+}$. From each population one sample was selected using the ignorable design and another was drawn with the nonignorable design. 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 and $n = 8000$.

In the simulation study we compare EFGL, F-H, unweighted and PHB solutions by comparing average posterior means and standard deviations of the parameters of interest. We also compare 95% prediction interval coverage probabilities as well as the average lengths of 95% prediction intervals. These comparisons are done for the ignorable as well as nonignorable samples.

For each sample ($s = 1, \cdots, 150$), using the Gibbs sampling technique, we generated 10,000 MCMC samples for each of the model parameters, namely $\beta_0, \beta_1, \eta_1, \ldots, \eta_{\mathrm{M}}$, and σ_η^2. These MCMC samples were tested for convergence using CODA (Convergence Output Data Analysis software). The first 1000 MCMC samples were deleted as the "burn-in" period and from the rest of the 9000 MCMC samples we selected every ninth sample to minimize auto-correlation, yielding a final MCMC sample of size 1000.

Let $\theta_{sc} = (\beta_{0sc}, \beta_{1sc}, \sigma_{\eta sc}^2)$ denote the parameter values from the c-th MCMC cycle corresponding to the s-th sample. In Tables 1 and 3, 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. 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 the end points of the 95% prediction interval (PI) based on the posterior distribution of Θ_{is} obtained from 1000 MCMC samples of Θ_{isc}. Define

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

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

Note that for the PHB methodology, we used the PHB2 small area estimates and associated unconditional posterior variances as defined in You and Rao (2003).

5. 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 F-H 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 F-H model is approximately 13 times larger than the other methods. This is due to the fact that the F-H solution uses an aggregate-level model. 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 F-H, PHB, and unweighted solutions coverage probabilities. The coverage probabilities for all solutions are very close. However, the prediction intervals for the F-H solution are 16% wider than the EFGL solution, which is expected, since the EFGL solution utilizes unit-level covariate information whereas the F-H solution uses aggregate-level covariate information. The unweighted method, being the most efficient, 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 F-H, 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 F-H model.

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

6. Conclusion

Our results show that SAEs based on unit-level predictors may be considerably more efficient than F-H type estimators. For our nonignorable samples, the EFGL and F-H solutions achieved PI coverage probabilities close to the targeted 0.95 level.

In the nonignorable sample case, the F-H intervals were on average 86% wider than EFGL intervals. The nonignorable samples played havoc with the unweighted solution. The intercept estimate was severely biased and the PI coverage probabilities were close to zero. The PHB solution substantially overestimated the SAE posterior variances in the nonignorable case and yielded average PI coverage probabilities close to 1.

In the ignorable sample case, the EFGL estimates had PIs that were on average about 10% wider than the most efficient unweighted intervals. In conclusion, we have demonstrated that the EFGL solution extends the robustness of F-H against nonignorable designs to the more efficient unit-level model.

Acknowledgements

This research was partially supported by grants from the following sources: U.S. Census Bureau (through the National Science Foundation), Centers for Disease Control and Prevention to the University of North Carolina at Chapel Hill, and by the contract number 200-98-0103 T.O. #3 (awarded by the Behavioral Sciences Branch of the Centers for Disease Control and Prevention, Department of Health and Human Services).

References

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**, 269-277.

Folsom, R. E., Shah, B. V., and Vaish, A. K. (1999). Substance abuse in states: A methodological report on model based estimates from the 1994-96 NHSDAs. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 371-375.

McCullagh, P. (1991). Quasi-likelihood and Estimating Functions. *Statistical Theory and Modeling*: In honor of Sir David Cox, FRS, ed. D. V. Hinkley, N. Reid and E. J. Snell, Chapman and Hall, London, 265-286.

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**, 23-40.

Rao, J.N.K. (2003). Small Area Estimation. Wiley series in survey methodology, John Wiley & Sons, Inc.

Singh, A.C., Folsom, R.E., and Vaish, A.K. (2002a). Estimating function based approach to hierarchical Bayes small area estimation for survey data. *Proceedings of ICRASS 2002.* (in press)

Singh, A.C., Folsom, R.E., and Vaish, A.K. (2002b). A hierarchical Bayes generalization of the Fay-Herriot method to unit level nonlinear mixed models for small area estimation. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3258-3263.

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*, **111**, 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			
	Fay-Herriot	EFGL	PHB	Unweighted	Fay-Herriot	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 Predication Interval Widths: Ignorable Sample Design

Percentiles and Means over Small Areas	Coverage Probability				Ratio of Average PI Widths		
	Fay-Herriot	EFGL	PHB	Unweighted	Fay-Herriot/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			
	Fay-Herriot	EFGL	PHB	Unweighted	Fay-Herriot	EFGL	PHB	Unweighted
β_0 (0.5)	0.5043	0.5029	0.5029	0.1043	0.0472	0.0450	0.0459	0.0448
β_1 (1.0)	1.0014	1.0004	1.0006	0.9999	0.1638	0.0131	0.0121	0.0103
σ_η^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 Predication Interval Widths: Nonignorable Sample Design

Percentiles and Means over Small Areas	Coverage Probability				Ratio of Average PI Widths		
	Fay-Herriot	EFGL	PHB	Unweighted	Fay-Herriot/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