COMPARISON OF AGGREGATE VERSUS UNIT-LEVEL MODELS FOR SMALL-AREA ESTIMATION

Eric V. Slud, Census Bureau & Univ. of Maryland
Mathematics Department, University of Maryland, College Park MD 20742

Key words: EBLUP, generalized linear model, mean-squared error, mixed effects, simulation.

Abstract. This paper compares two methods of small-area estimation in a setting imitating the Census Bureau's county-level estimation of child poverty rates within the SAIPE (Small-Area Income and Poverty Estimates) program. The first method estimates a transformed Fay-Herriot (1979) regression model for log-rates of child poverty by county in terms of several county-level predictors, discarding data from sampled counties with 0 counts of child poor. The second method uses likelihood-based parameter estimates within a mixed-effect logistic regression model for poverty of individual CPS-sampled children. The Empirical BLUP small-area estimators from the Fay-Herriot model are compared, via simulation, with the analogous EBLUP estimators for the unit-level logistic model.

This paper reports on research undertaken by the author. Results and conclusions expressed have not been endorsed by the Census Bureau.

1. INTRODUCTION

As summarized by Bell (1997, 1998) and Citro and Kalton (1999), the Small Area Income and Poverty Estimates (SAIPE) program at the Census Bureau is congressionally mandated to estimate poverty rates among children at the state, county, and ultimately school district level. At the county level, which is all we consider here, the current methods rely on a mixed-effects linear model in terms of Census, Current Population Survey (CPS), and IRS predictor variables, for the logarithm of the observed number (of poor children) in counties for which CPS samples were taken and in which the sample contained a nonzero number of poor children. Sampled counties without poor children aged 5–17 in-sample are dropped from the analysis, a bothersome aspect of the current method highlighted in NAS reviews. It is desirable instead to model the essential discreteness of the response-counts by unit-level models.

The existing SAIPE methodology for small-area estimation is based upon the classic linear model of Fay & Herriot (1979) using aggregate-level data. Previous applications of unit-level generalized linear models in small area estimation include Malec et al. (1993) and Ghosh et al. (1998). R. Folsom and co-authors at Research Triangle Institute have used such methods for several years in connection with the National Household Survey on Drug Abuse. The primary approach to parameter estimation in these previous works is Gibbs sampling. A still-useful general review of small-area estimation methods is the paper of Ghosh & Rao (1994).

In this paper, we first present a model which can be used to simulate the county-level SAIPE data. We then describe two different small-domain working models for analyzing such data: (i) a mixed-effect linear-model fit to the logarithms of sampled counts, with zero-counts discarded, and (ii) a mixed-effect unit-level logistic regression model with county-level random effect, estimated by numerical maximization of the accurately approximated log-likelihood (Slud 2000). Both models are slightly misspecified: the quality of estimates they produce are compared here via simulation. A fuller discussion of the models and simulations presented can be found in the SAIPE technical report Slud (1999).

Acknowledgment. This work was supported by the Census Bureau's SAIPE program. I am grateful to Bill Bell for guidance and insights throughout.

2. AGGREGATE VS. UNIT MODELS

Suppose that for each county (PSU) in the nation, $i = 1, \ldots, m$, there is a population (e.g., the set of children 5–17) of size N_i which can be assumed known; a response variable Y_i^0 which is a count of population members in a desired response-category (e.g., poor child aged 5–17); and a vector X_i of explanatory variables such as "log of IRS poverty-rate", rate-variables related to Food Stamps and IRS exemptions, etc. The count Y_i^0 is not observable, but the corresponding count y_i^0 *is* for a random sample s_i of size n_i taken from each sampled PSU ($i \in s$). For many PSU's, $n_i = 0$; and for many sampled PSU's, the observed count y_i^0 will turn out to be 0. Assume that the PSU sizes N_i are always much larger than the sample size n_i . For simplicity,

assume that samples are drawn at random within each PSU. The parameters to be estimated are the ratios $\vartheta_i = Y_i^0/N_i$ for entire PSU's.

2.1. *Transformed Aggregate Models*

The transformed PSU- and sample- aggregated counts are often assumed to follow a linear regression model — a **Transformed Fay-Herriot (1979) Model** — with PSU or cluster random effect following a normal sampling distribution:

$$Y_i^0/N_i = h(\gamma_0 + \gamma_1' X_i + U_i), \quad U_i \sim \mathcal{N}(0, \sigma^2) \quad (1)$$

The unknown coefficients are a scalar intercept γ_0 and a vector γ_1 of the same dimension as X_i; and h is one of a few possible known (*exp, logistic,* or *identity*) functions. The cluster random effect U_i is shared by all individuals within the PSU. The unobservable total Y_i^0 and the sampled count y_i^0 (for $i \in s$) are connected through the model

$$y_i \equiv h^{-1}(y_i^0/n_i) = h^{-1}(Y_i/N_i) + e_i \equiv Y_i + e_i \quad (2)$$

on the measurement scale defined by h^{-1}, where

$$e_i \sim \mathcal{N}(0\,,\, v_e/n_i)$$

Either σ^2 or v_e is assumed known, as in Fay & Herriot (1979) for the case $h(x) = x$, and the unknown parameters (γ_0, γ_1 and σ^2 or v_e) are estimated by maximum likelihood (ML).

We imitate the SAIPE county-level analysis described in Bell (1997) by analyzing data assuming (1)–(2) with $h(x) \equiv \exp(x)$, using only sampled PSU's in which $y_i^0 > 0$. Early SAIPE models treated the sampling-error variance term v_e as essentially known through generalized variance-function estimation. ML estimates obtained with v_e known and σ unknown are labelled **linfitA**. However, the PSU variance is currently estimated from Census data in the 'bivariate' model of Bell (1997, 1998), motivating a second method of analysis, labelled **linfitB**, in which σ^2 but *not* v_e is taken as known.

In **linfitA** analyses, v_e is set equal within simulations to the estimated variance of

$$\sqrt{N_i}\left(\log(Y_i^0/N_i) - \log E(Y_i^0/N_i \mid U_i)\right)$$

In **linfitB** analyses within simulations, the quantity σ^2 is set equal to the residual mean-squared error in regressing $\log(\vartheta_i)$ on $1, X_i$ across PSU's i.

2.2. *Unit-level Models*

A model resembling (1)–(2), but with v_e depending upon i and ϑ_i, arises from a *unit-level* model

$$Y_i^0 \sim \text{Binom}(N_i, \pi_i)\,, \quad y_i^0 \sim \text{Binom}(n_i, \pi_i) \quad (3)$$

in which Y_i^0 is the sum of N_i independent indicators y_{ij}, $j = 1, \ldots, N_i$, with identical expectations

$$\pi_i = P(y_{ij} = 1) = h(\gamma_0 + \gamma_1' X_i + U_i) \quad (4)$$

and y_i^0 is the sum of y_{ij} for the n_i sampled units $j \in s_i$. A natural model of this sort is a mixed-effect logistic regression, with $h(x) = e^x/(1+e^x)$.

In the limit of large n_i and N_i, the Central Limit Theorem yields (for fixed covariates X_i and PSU random effects U_i), that $y_i^0/n_i \approx \mathcal{N}(\pi_i,\, \pi_i(1-\pi_i)/n_i)$, for π_i as in (4). Moreover, when $n_i \ll N_i$, the order of magnitude of $\vartheta_i - \pi_i = Y_i^0/N_i - \pi_i$ is much smaller than the order $1/\sqrt{n_i}$ of y_i^0/n_i, so that $y_i^0/n_i \approx \mathcal{N}(\vartheta_i, \vartheta_i(1-\vartheta_i)/n_i)$. Then the Delta Method shows under model (3)–(4) that

$$h^{-1}\left(y_i^0/n_i\right) = h^{-1}(\vartheta_i) + e_i$$

with the distribution of e_i conditional on X_i and U_i given approximately by

$$e_i \sim \mathcal{N}(\,0\,,\, \{(h^{-1})'(\vartheta_i)\}^2\, \vartheta_i(1-\vartheta_i)/n_i\,) \quad (5)$$

The error-term e_i arising here from the unit-level model (3)–(4) has the same form as in the aggregate model (1)–(2) *except* that the analog of v_e in the latter is now PSU-dependent. Only in the very special case where $h(x) \propto \sin^2(x/2)$ does it turn out that e_i in (5) has (conditional) variance not depending on ϑ_i. For example, when $h(x) = e^x/(1+e^x)$ is the *logistic* (distribution) function, (5) yields conditional variance for e_i equal to $\{\vartheta_i(1-\vartheta_i)n_i\}^{-1}$.

The method of estimation used in the simulations below, labelled **glmfit**, is Maximum Likelihood (ML) based on the assumption of *logistic* h. Another estimation method, discussed in Slud (1998, 1999), is ML within a variance-stabilized mixed nonlinear regression model. However, that method is not treated here because it yields biased parameter estimates unless the sample-sizes n_i are very large.

An aggregated model (1)–(2) is likely to be well approximated by a unit-level Binomial model (3)–(4) only if the h functions for the two models match. Extra regression terms beyond the linear terms specified for these models do help in mitigating the effects of misspecifying h. For this reason, we consider the effect in our simulations of incorporating a quadratic explanatory variable.

3. EBLUP SMALL-AREA ESTIMATORS

The fractions $\vartheta_i = Y_i^0/N_i$, are to be estimated based on covariates X_i which are constant over the i'th PSU. The parameters (γ_0, γ_1) and σ^2 or v_e in model (1)–(2) are first estimated, either within a

mixed linear regression with $h(x) = e^x$ or via ML for (3)–(4) with *logistic* h. These estimators are substituted into modified 'EBLUP' small-area estimators (see Ghosh and Rao, 1994). There are two separate cases: first, where the estimator of parameter ϑ_i in PSU i is based on no sampled data in the PSU, but only on the predictor X_i and the estimators $(\hat\gamma_0, \hat\gamma_1)$ along with $\hat\sigma^2$ or $\hat v_e$; and second, where ϑ_i is estimated in terms of the parameters, predictor and observed sample of size n_i (with y_i^0 responses) in the i'th PSU (for $i \in s$). In both cases, sampling variability of the fixed-effect coefficient estimators $(\hat\gamma_0, \hat\gamma_1)$ should be taken into account, since the estimators are based on nonlinear functions of observed response rates.

For simplicity of notation from now on, define

$$\eta_i \;=\; \gamma_0 + \gamma_1' X_i \quad , \quad \hat\eta_i \;=\; \hat\gamma_0 + \hat\gamma_1' X_i$$

3.1. Modified EBLUP's

The small-area estimator for ϑ_i should, if we knew the coefficients (γ_0, γ_1) in model (1)–(2) or (3)–(4) exactly, be based on the random quantity $\eta_i + U_i$ and the conditional distribution of the PSU random effect U_i given the observed data. Recall that the conditional expectation for $\vartheta_i = Y_i^0/N_i$ given X_i, U_i is always $\pi_i = h(\eta_i + U_i)$. By analogy with BLUP's, our principle of estimation of ϑ_i is to estimate the conditional expectation

$$E(\vartheta_i \,|\, (\hat\gamma_0, \hat\gamma_1, \hat\sigma^2), y_i^0) \;=\; E(\pi_i \,|\, (\hat\gamma_0, \hat\gamma_1, \hat\sigma^2), y_i^0) \;=\; E\big(\, E(\,\pi_i \,|\, (y_k^0,\, k \in s)\,) \,|\, (\hat\gamma_0, \hat\gamma_1, \hat\sigma^2), y_i^0\big) \qquad (6)$$

The estimators defined via (6) are approximately unbiased. For **linfitB** analyses, $\hat\sigma^2$ in (6) is replaced by $\hat v_e$. However, for simplicity we continue to write formulas in terms of $\hat\sigma^2$. When there is no sample in a PSU, the expectation is conditioned given only the parameter estimators $(\hat\gamma_0, \hat\gamma_1, \hat\sigma^2, \hat v_e)$.

Estimators based on (6) explicitly require some approximation to the joint distribution of U_i and $(\hat\gamma_0, \hat\gamma_1, \hat\sigma^2, \hat v_e)$, and we assume (**Assumption A**) that U_i is approximately conditionally independent given y_i^0 of the estimators $(\hat\gamma_0, \hat\gamma_1, \hat\sigma^2, \hat v_e)$ and of $(y_k^0,\, k \neq i)$, with conditional variance σ^2 for all i; and that the parameter estimators are jointly normally distributed, given each y_i, with means equal to the true values under model (1)–(2) or (3)–(4).

We apply Assumption **A** separately for the two models we compare. In the Fay-Herriot (1979) model (1)–(2), with $h(x) = e^x$ and $y_i = \log(y_i^0/n_i)$, the conditional law $\mathcal{L}(U_i \,|\, y_i)$ of U_i given y_i is

$$\mathcal{N}\Big(\, \frac{\sigma^2}{\sigma^2 + v_e/n_i}\,(y_i - \gamma_0 - \gamma_1' X_i)\,,\; \frac{\sigma^2 v_e/n_i}{\sigma^2 + v_e/n_i}\,\Big)$$

By Assumption **A**, conditionally given y_i and the parameter estimators, the law of $\hat\eta_i + U_i$ is

$$\mathcal{N}\Big(\hat\eta_i + \frac{\sigma^2}{\sigma^2 + v_e/n_i}\,(y_i - \hat\eta_i)\,,\; \frac{\sigma^2 v_e}{n_i\sigma^2 + v_e}\,\Big) \qquad (7)$$

The conditional law given only y_i^0 but *not* the parameter estimators is obtained from the normal distribution (7) by replacing $\hat\eta_i$ with η_i within the mean, and increasing the variance by

$$a_i^2 \;=\; \begin{pmatrix} 1 \\ X_i \end{pmatrix}' \Sigma_\gamma \begin{pmatrix} 1 \\ X_i \end{pmatrix}$$

where Σ_γ denotes the large-sample covariance matrix for the fixed-effect estimators $\hat\gamma_0$, $\hat\gamma_1$. In the linear model for log-counts, the conditional law of $\hat\eta_i + U_i$ given the parameter estimators is

$$\mathcal{L}(\hat\eta_i + U_i \,|\, \hat\gamma_0, \hat\gamma_1, \hat\sigma^2\,) \;\approx\; \mathcal{N}(\eta_i, \sigma^2\,) \qquad (8)$$

In the model (3)–(4), by Assumption **A** the conditional density of U_i at u, given $(\hat\gamma_0, \hat\gamma_1, \hat\sigma^2)$ and $y_i^0 = m$, is approximately proportional to

$$e^{m(\eta_i + u) - u^2/(2\sigma^2)}\,\big(1 + e^{\eta_i + u}\big)^{-n_i} \qquad (9)$$

and $\hat\eta_i$ is approximately independent of U_i with

$$\hat\eta_i \;\approx\; \mathcal{N}(\eta_i, a_i^2\,) \qquad (10)$$

Expectations in this model are given in terms of

$$\Lambda(x, m, n, b) \;=\; \int \frac{e^{m(x+bz)}}{(1 + e^{x+bz})^n}\,\phi(z)\,dz \qquad (11)$$

where $\phi(\cdot)$ denotes the standard normal density.

3.2. Non-sampled PSU's

For non-sampled PSU's, the estimator would be

$$\hat\vartheta_i \;=\; \widehat{E}\,\big(\,h(\gamma_0 + \gamma_1' X_i + U_i) \,|\, \hat\gamma_0, \hat\gamma_1, \hat\sigma^2\,\big)$$

where the estimated expectation $\widehat{E}$ will have estimators $(\hat\gamma_0, \hat\gamma_1, \hat\sigma^2)$ substituted and will be bias-corrected if possible. In the model (1)–(2) with $h(x) = e^x$, we obtain via Assumption **A** and (8), after substituting parameter estimators, that

$$\hat\vartheta_i \;=\; \exp\big(\hat\gamma_0 + \hat\gamma_1' X_i + (\hat\sigma^2 - \hat a_i^2)/2\big) \qquad (12)$$

where the bias-correction term

$$\hat a_i^2 \;=\; \begin{pmatrix} 1 \\ X_i \end{pmatrix}' \hat\Sigma_\gamma \begin{pmatrix} 1 \\ X_i \end{pmatrix}$$

is defined in terms of a consistent estimator $\hat\Sigma_\gamma$, produced by each estimation-method, for the covariance matrix of estimators for fixed-effect coefficients.

The corresponding bias-corrected small-area estimator for the model (3)–(4), conditional on parameter estimates, is given in the notation (11) by

$$\hat{\vartheta}_i \;=\; \Lambda(\hat{\eta}_i,\, 1,\, 1,\, \sqrt{\hat{\sigma}^2 - \hat{a}_i^2}\,) \qquad (13)$$

where $\hat{\sigma}^2 - \hat{a}_i^2$ is replaced by 0 if it is negative.

3.3. Sampled PSU's

The small-area estimator for a sampled PSU i makes direct use of (6), Assumption **A**, and the conditional distributions of subsection 3.1. First, in the Fay-Herriot model (1)–(2) with $h(x) = e^x$, we find the expectation of the exponential of a variable with the distribution (7), with substituted parameter estimators and corrected for bias, as

$$\hat{\vartheta}_i \;=\; \exp\Big(\hat{\eta}_i \;+\; \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + v_e/n_i}\,(y_i - \hat{\eta}_i) \qquad (14)$$

$$+\; \frac{1}{2}\left[\frac{\hat{\sigma}^2\, v_e}{n_i\,\hat{\sigma}^2 + v_e} \;-\; \frac{(\hat{a}_i\, v_e/n_i)^2}{(\hat{\sigma}^2 + v_e/n_i)^2}\right]\Big)$$

In the mixed unit-level regression model (3)–(4) with $h(x) = e^x/(1+e^x)$, we obtain for (6) via the approximation of Assumption **A**

$$\int \frac{e^{y_i^0(\hat{\eta}_i + \hat{\sigma} z)}\phi(z)}{(1+e^{\hat{\eta}_i + \hat{\sigma} z})^{n_i}}\, h(\hat{\eta}_i + \hat{\sigma} z)dz \Big/ \int \frac{e^{y_i^0(\hat{\eta}_i + \hat{\sigma} z)}\phi(z)}{(1+e^{\hat{\eta}_i + \hat{\sigma} z})^{n_i}} dz$$

which is equal by definition to

$$\Lambda(\,\hat{\eta}_i,\, y_i^0 + 1,\, n_i + 1,\, \hat{\sigma}\,)\,/\Lambda(\,\hat{\eta}_i,\, y_i^0,\, n_i,\, \hat{\sigma}\,)$$

Since there is no simple bias-correction for this ratio, the small-area estimator for the unit-level model is

$$\begin{aligned} \hat{\vartheta}_i \;&=\; \widehat{E}(\,h(\hat{\eta}_i + U_i)\,|\,\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2, y_i^0\,) \\ &=\; \frac{\Lambda(\,\hat{\eta}_i,\, y_i^0 + 1,\, n_i + 1,\, \hat{\sigma})}{\Lambda(\,\hat{\eta}_i,\, y_i^0,\, n_i,\, \hat{\sigma})} \end{aligned} \qquad (15)$$

4. SIMULATION STUDY DESIGN

We now describe a simulation study designed to compare small-area estimates (without PSU or unit-level weighting) based upon the SAIPE aggregated mixed-effect log-linear models for county log child-poverty rates, omitting sampled counties with zero counts of poor school-age children, versus estimates from unit-level mixed logistic models.

We began by fixing the numbers n_i, a set of 1488 PSU sample-sizes corresponding to the (non-zero) numbers of school-age children sampled by the CPS in counties over the 3 years 1992-94. The distribution of numbers sampled within PSU's was very skewed, ranging from 1 to 2226, as Table 1 shows.

Table 1: Size-categories of sample in PSU.

Group	Interval of n_i	# PSU's in Gp.
1	[1, 10]	506
2	[11, 25]	448
3	[26, 75]	398
4	[76, 220]	106
5	[221, 2500]	30

The total PSU sizes N_i played a direct role only in the estimation of v_e, and were fixed at a factor of 2000 (roughly the reciprocal sampling fraction of the CPS) multiplied by the sample sizes n_i.

Our single predictor variable X_i, resembling the IRS-estimated log number of poor children in county centered at 0, was simulated once for all simulations displayed, as a column of independent $\mathcal{N}(0,\, 1.69)$ random variables. However, to prevent unrealistically large variation in response fractions for PSU's with very large samples (those > 220), we fixed X_i for these large-sample PSU's to be 0.

Data for each simulation iteration were simulated according to model (3)–(4) with specified parameters $\gamma_0, \gamma_1, \sigma^2$ and $U_i \sim \mathcal{N}(0, \sigma^2)$; with binary response for individual j within PSU i of

$$y_{ji} \;\sim\; \mathit{Binom}\,(1,\, \pi_i)\,, \quad j = 1, \ldots, N_i$$

where π_i is as in (4); and with $y_i^0 \equiv \sum_{j=1}^{n_i} y_{ij}$. The function $h(x)$ (*logistic* unless indicated otherwise) and parameters $(\gamma_0,\, \gamma_1,\, \sigma^2)$ were fixed within each simulation. The initial choice of the fixed-effect coefficients was: $\gamma_0 = -1.6$, to get the average of ϑ_i values around 0.2, and $\gamma_1 = 0.9$ so that ϑ_i falls in the range $(0.05,\, 0.40)$ when $U_i = 0$.

For each simulation iteration, estimators of $\gamma_0,\, \gamma_1,\, \sigma^2$ were calculated in the three ways described in Section 2 (**linfitA**, **linfitB**, and **glmfit**), using specially coded **Splus** functions. The small-area estimators $\hat{\vartheta}_i$ were calculated as in Section 3, and the **linfit** estimated $\hat{\vartheta}_i$ values were replaced by 1 whenever greater than 1. For each simulation iteration, and each of the three sets of parameter estimators, the empirical Mean-Squared Errors (MSE's) for small-area estimators were averaged over each PSU Group, defined in Table 1 through the numbers n_i sampled within PSU.

5. RESULTS OF SIMULATION STUDY

Figures 1(a)–(d) display the groupwise average MSE results of 4 simulation experiments of 100 iterations each, with parameters shown in the graph headings. In all of these simulations, analyses were unweighted, and the plotted **MSE** numbers are empirical. Figure 1(a) shows the empirical behavior of

estimators based on formulas (12) and (13), which should be interpreted as estimates of ϑ_i as though no sample were drawn from the i'th PSU. An analogous simulation with $\sigma = 0.3$ in place of 0.2 (results not shown), yielded larger MSE's for all methods, but the comparison between methods was essentially the same as in Figure 1(a). Figure 1(b) shows the performance of modified-EBLUP estimators (14) and (15) for $\sigma = 0.2$, and the simulation of Figure 1(c) differs from that of 1(b) only in having augmented the set of predictor covariates 1, X_i with X_i^2 .

The MSE's for $\hat{\vartheta}_i$ in nonsampled PSU's (Fig. 1(a)) are of the order of 0.001 across the board in the **glmfit** method, but tend to be larger by a factor of 2 or more in PSU Groups 1 to 4 (PSU samples of 220 or less) for **linfitA** and **linfitB**. In Group 5, **linfitA** and **linfitB** respectively have MSE's larger than **glmfit** by 8% and 17%. The MSE's in sampled PSU's (Fig. 1(b)) are, for all methods of EBLUP estimation, only slightly better than in non-sampled PSU's for PSU Groups 1 and 2: although we used formula (15) in estimating PSU response rate for the unit-level (nonlinear) models in PSU Groups 1 and 2, in these groups the **glmfit** MSE entries are essentially identical in Figures 1(a) and 1(b). But there are clear differences in Groups 3 to 5 for the **glmfit** MSE's respectively between Figures 1(a) and 1(b). Figure 1(b) shows that in Groups 3 and 4, **linfitB** has MSE larger than **glmfit** by 100% and 60%, and in Group 5 **glmfit** is only about 10% better than **linfitB**. When the quadratic predictor X_i^2 is used in the linear-model fitting (Fig. 1(c)), the MSE's for **linfitB** (with the properly chosen value for σ^2 taken as known) improve considerably in Groups 1 to 4: they are larger than for **glmfit** by a factor of only about 1.5 for Groups 2, 3, and 4, and are just about the same as for **glmfit** in Group 5. Results for **linfitA** are also much improved by the additional predictor, but still much worse than **linfitB**. Again, analogous simulations with $\sigma = 0.3$ (not shown) replacing the value $\sigma = 0.2$ in Figures 1(b)-(c), yielded similar comparisons between methods.

We consider next the performance of the same small-area estimators upon simulated data from the unit-level model with $h(x) \equiv e^x$, which is designed to show the **linfit** analysis method in its most favorable light, and makes the **glmfit** analysis clearly mis-specified. Although Figure 1(d) displays the results only for EBLUP's (sampled PSU's), the **linfit** small-area estimators show only a very small advantage in MSE over those based on **glmfit**. The largest advantage in MSE for **linfit** versus **glmfit** appears in Figure 1(d) for Groups 1 and 2: **linfitB** has a 5% advantage over **glmfit** in Group 1 and 10% in Group 2, but none for PSU's with sample-size larger than 25. Remarkably, there are no other cases in our simulation where either **linfit** method outperformed **glmfit**, even in this setting with $h \equiv exp$.

6. CONCLUSIONS

The simulation results presented in the previous Section, together with the more complete results in Slud (1999), yield the following conclusions contrasting the small area estimators produced in the SAIPE context with the aggregate-level transformed Fay-Herriot model (1)-(2) versus those produced with the unit-level mixed model (3)-(4).

(1) **glmfit** gives uniformly smallest MSE. In the simulated-data comparisons, the loglinear-model **linfit** methods were implemented with σ^2 or v_e *accurately known*, which might tend to under-estimate the MSE's they would provide in practice.

(2) Even in the models with $h \equiv \exp$, where **linfit** methods should be at their best, unit-level logistic-model analyses are as good.

(3) Quadratic (and probably also interaction-) covariate terms help when the working model is misspecified, a *little* with **glmfit** and $h_0 \equiv \exp$ and a *lot* with **linfit** and $h_0 \equiv$ *logistic*.

(4) Simulations with multidimensional covariates X_i show still greater improvement of **glmfit** analysis methods over the **linfit** methods.

These observations together indicate that the **glmfit**-based small-area estimators using mixed logistic models in place of **linfit** (mixed log-linear aggregated model) are very promising in the SAIPE context as a way to overcome the disadvantage of discarding sampled 0-counts, and can help considerably more than they are likely to hurt due to misspecification of the mixed logistic unit-level model.

7. REFERENCES

Bell, W. (1997) Models for county and state poverty estimates. Preprint, Census Statistical Research Division.

Bell, W. (1998) Borrowing information over time in small area estimation: thoughts with reference to the American Community Survey. Presented at NAS workshop on ACS, Sept. 11, 1998.

Citro, C. and Kalton, G., eds. (1999) **Small-Area Estimates of School-Age Children in Poverty**, Interim Report 3 (National Research Council), Washington DC: Nat. Acad. Press.

Fay, R. and Herriot, R. (1979) Estimates of income for small places: an application of James-Stein procedures to census data. *JASA* **74**, 341-53.

Ghosh, M., Natarajan, K. Stroud, T. and Carlin, B. (1998) Generalized linear models for small-area estimation. *JASA* **93**, 273-82.

Ghosh, M., and Rao, J.N.K. (1994) Small Area Estimation: an appraisal. *Statist. Sci.* **9**, 55-93.

Malec, D., Sedransk, J. & Tompkins, L. (1993) Bayesian predictive inference for small areas for binary variables in the National Health Interview survey. In: *Case Studies in Bayesian Statistics*, **Lect. Notes in Stat. 83**, New York: Springer-Verlag.

Slud, E. (1998) Logistic regression with large cell-counts and multiple-level random effects. Unpublished paper.

Slud, E. (1999) Models for simulation and comparison of SAIPE analyses. Census Bureau preprint, posted at SAIPE web-site (see below).

Slud, E. (2000) Accurate calculation and maximization of log-likelihood for mixed logistic regression. Census Bureau preprint, posted at SAIPE web-site:

ftp.census.gov/hhes/www/saipe/tecrep.html

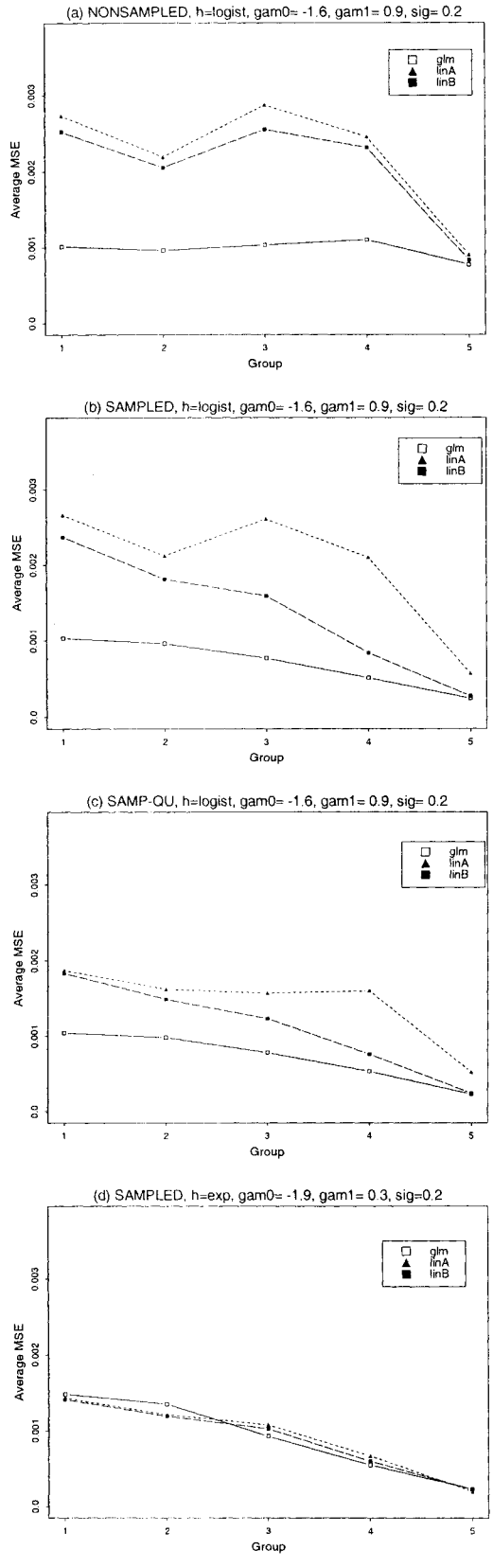

Figure 1, at right. Plots of groupwise Mean-Squared Errors, averaged over PSU Groups defined in Table 1 through similar sample sizes, of Small-Area Estimates $\hat{\vartheta}_i$ calculated by three different methods, based onsimulated data with $\sigma = 0.2$ and fixed-effect parameters as displayed in graph headings:

(a) MSE's in nonsampled PSU's, for $h \equiv$ *logist*;

(b) MSE's in sampled PSU's, for $h \equiv$ *logist*;

(c) MSE's in sampled PSU's, for $h \equiv$ *logist*, with an extra quadratic predictor; and

(d) MSE's in sampled PSU's, for $h \equiv$ *exp*.