Models for Simulation & Comparison of SAIPE Analyses

Eric V. Slud, Census Bureau & Univ. of Maryland

May 2, 2000

Abstract. Following Bell (1997), we first present a simplified ideal model which can be used to simulate underlying data for the small-area (county-level) estimation currently used in the SAIPE project. We then describe two different small-domain estimation methodologies which can be used on 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 fit to the sampled counts. The methods, both of which are based on slightly misspecified models, are compared via simulation. Initially, sampling weights are ignored, but then it is shown how they can be included in both the aggregated-linear and the unit-level models.

Key words: best linear unbiased predictor (BLUP), mean-squared errors, mixed effects linear model, mixed effect logistic regression, nonlinear regression, SAIPE, sampling weights, small area estimation, weighted census likelihood.

This preprint reports on research and analysis undertaken by Eric Slud, and is released to inform interested parties and to encourage discussion. Results and conclusions expressed are those of the author and have not been endorsed by the Census Bureau.

1 Introduction

As summarized by Bell (1997, 1998) the Small Area Income and Poverty Estimates (SAIPE) project at the Census Bureau has developed methods for estimating poverty and income statistics at the county and state level. At the county level, these methods rely on a mixed-effects linear model which is applied to the logarithms of the observed numbers (e.g. of poor children 5–17) in counties for which CPS samples were taken and in which the sample contained a nonzero number of poor children. Those sampled counties without poor children 5–17 in-sample are dropped from the analysis, a bothersome aspect of the methodology. It would be desirable instead to model the essential discreteness of the response-counts by some sort of unit-level model.

Based upon the methods proposed in Slud (1998, 2000) of estimating mixed-effect logistic regression models via mixed-nonlinear-regression software or by maximization of an approximate log-likelihood, we describe here a mixed-effect unit-level logistic regression model for SAIPE data. This model would make use of all of the SAIPE data. In order to compare this new method to the one which is now in use, we conduct a simulation based upon several idealized models which include the mixed linear model of Bell (1997) and the mixed logistic regression model which is probably the simplest available at the level of sampling-units.

2 Idealized Model for Simulated Data

Suppose that for each county (PSU) $i = 1, \ldots, m$ in the nation, there is a population size N_i which can (initially) be assumed known; a response variable Y_i^0 which is a count of individuals in a desired response-category (e.g., poor child aged 5–17); and an explanatory variable $X_i \equiv X_{1i}$ (such as "log of IRS poverty-rate"). Assume that the count Y_i^0 is never observable, but that the corresponding count y_i^0 is observable for a random sample of size n_i taken from the PSU. For many PSU's, n_i will be 0 (no sample drawn); and for many of the sampled PSU's, the observed counts y_i^0 will turn out to be 0. Assume that the PSU sizes N_i are always at least of order 1000 times larger than the sample size n_i . Initially, also assume

for simplicity that the samples are drawn at random, with identically equal weights.

The parameters which it is desired to estimate are the ratios

$$\vartheta_i \;=\; Y_i^0/N_i$$

for entire PSU's. For purposes of analysis, we assume as a first approximation that these quantities follow a possibly nonlinear regression model with a PSU-cluster (normally distributed) random effect and a normal sampling distribution with known variances:

$$\frac{Y_i^0}{N_i} \;=\; h(\gamma_0 + \gamma_1\, X_i + U_i)\,, \qquad U_i \;\sim\; \mathcal{N}\Big(0\,,\,\sigma^2\Big) \tag{1}$$

Here γ_0 and γ_1 are unknown constants; and h will be taken to be one of a few possible known (*exp, logistic,* or *identity*) functions. The variance of the cluster random effect U_i is not chosen to be a decreasing function of PSU size or sample size because the effect is shared by all individuals within the PSU. The unobservable Y_i^0 and the sampled quantity y_i^0 are assumed to be connected through the model

$$y_i \;\equiv\; h^{-1}\left(\frac{y_i^0}{n_i}\right) \;=\; h^{-1}\left(\frac{Y_i^0}{N_i}\right) \;+\; e_i \;\equiv\; Y_i \;+\; e_i \tag{2}$$

where

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

For the SAIPE models in current use, one often treats the variance v_e — which quantifies sampling error — as essentially known through generalized variance-function estimation. The model (2) expresses the error in estimating ϑ_i *on the measurement scale define by* h^{-1} due to sampling within the PSU. Presenting the errors on this measurement scale is convenient but slightly unnatural: the model (2) makes clear sense only if h is the identity function, so that

$$\frac{y_i^0}{n_i} \;\;=\;\; \frac{Y_i^0}{N_i} \;+\; e_i$$

But when this is true, we have by the delta method for large n_i and general h :

$$\sqrt{n_i}\,\Big(h^{-1}(\frac{y_i^0}{n_i}) - h^{-1}(\frac{Y_i^0}{N_i})\Big) \;\approx\; \sqrt{n_i}\,\Big(\frac{y_i^0}{n_i} - \frac{Y_i^0}{N_i}\Big)\cdot (h^{-1})'(\frac{Y_i^0}{N_i}) \;\approx\; \sqrt{n_i}\, e_i\, (h^{-1})'(\vartheta_i)$$

and this model differs from (2) in the dependence of the variance of the difference $h^{-1}(y_i^0/n_i) - h^{-1}(Y_i^0/N_i)$ on the true PSU parameter ϑ_i.

A particular way for models resembling (1), but with σ^2 replaced by a quantity depending upon i and X_i, to arise is by a *unit-level* model in which Y_i^0 is the sum of N_i independent indicators with identical probability $\pi_i = h(\gamma_0 + \gamma_1 X_i + U_i)$ of being 1. A natural model of this sort is a mixed-effect logistic regression, with

$$h(x) \;=\; e^x/(1+e^x)$$

3 Analysis Methods for Comparison

In the simulations which follow, the data $(n_i,\, y_i^0)$ will be analyzed in one of two possible ways. First, we imitate the linear-model analysis described by Bell (1997) for logarithms of (non-zero) counts. The 'model' is

$$y_i \;=\; \log\left(\frac{y_i^0}{n_i}\right) \;=\; \gamma_0 + \gamma_1 X_i + \frac{\delta_i}{\sqrt{n_i}} + U_i \tag{3}$$

where δ_i is taken to be normal with variance ω_*. Initially, ω_* will be taken to be known and, in our (unweighted-sampling) simulations, is set equal to the estimated variance of

$$\sqrt{N_i}\,\Big(\log(Y_i^0/N_i) - \log E(Y_i^0/N_i \,|\, U_i)\Big)$$

Within model (3), the cluster random effect $U_i \sim \mathcal{N}(0, \sigma^2)$ is independent of the random sampling error δ_i, and the unknown parameters γ_0, γ_1, σ^2 are estimated via maximum likelihood, for fixed $\omega_* = v_e$, omitting the data for sampled PSU's i having counts y_i^0 equal to 0. Later, within the description of simulations in Section 5 below, another estimation method is described for model (3), according to which σ^2 is assumed known and ω_* is estimated from simulated complete PSU data.

The second method of analysis which we consider in these comparisons is based upon the idea that, conditionally given each PSU cluster effect and sampling error, the individuals within each PSU respond independently with identical probability of a logistic-regression form

$$\pi_i \;=\; \frac{\exp(\gamma_0 + \gamma_1 X_i + U_i)}{1 + \exp(\gamma_0 + \gamma_1 X_i + U_i)} \tag{4}$$

This unit-level model corresponds to the aggregated model like (1) with $h(x) = e^x/(1+e^x)$ and (given X_i, U_i)

$$\frac{Y_i^0}{N_i} - \pi_i \;\sim\; \mathcal{N}(0,\, \frac{\pi_i(1-\pi_i)}{N_i})$$

Note that because N_i is so much larger than n_i it does no harm to treat

$$logit(Y_i^0/N_i) \;\equiv\; \log\left(\frac{Y_i^0/N_i}{1 - Y_i^0/N_i}\right) \;\approx\; \gamma_0 + \gamma_1 X_i + U_i$$

Through the unit-level model, we have also

$$y_i^0 \;\sim\; Binom(n_i,\, \pi_i) \tag{5}$$

which for moderate to large n_i implies approximately

$$\frac{y_i^0}{n_i} - \pi_i \;\sim\; \mathcal{N}(0,\, \frac{\pi_i(1-\pi_i)}{n_i})$$

and via the delta method,

$$\sqrt{n_i}\left(logit(y_i^0/n_i) - \gamma_0 - \gamma_1 X_i - U_i\right) \;\approx\; \mathcal{N}(0,\, ,\, \frac{\pi_i(1-\pi_i)}{(\pi_i(1-\pi_i))^2}\,)$$

or, with $h(x) \;=\; e^x/(1+e^x)$ and $h^{-1}(y) = logit(y) = \log(\frac{y}{1-y})$,

$$\sqrt{n_i}\left(h^{-1}(y_i^0/n_i) - h^{-1}(Y_i^0/N_i)\right) \;\sim\; \mathcal{N}(0,\, ,\, \frac{1}{\pi_i(1-\pi_i)}\,) \tag{6}$$

This development indicates the more general result that a unit-level model with $\mathcal{N}(0,\sigma^2)$ PSU-level random effects U_i

$$Y_i^0 \;\sim\; Binom\Big(N_i,\, h(\gamma_0+\gamma_1 X_i+U_i)\Big) \quad , \qquad y_i^0 \;\sim\; Binom\Big(n_i,\, h(\gamma_0+\gamma_1 X_i+U_i)\Big)$$

with moderate n_i *and much larger* N_i *(but no explicitly known sampling-error term* v_e*) is compatible with and well-approximated by an aggregated model (1) along with the model conditional upon* $\vartheta_i = Y_i^0/N_i$ *that*

$$h^{-1}\left(\frac{y_i^0}{n_i}\right) = h^{-1}(\vartheta_i) + e_i \ , \quad e_i \sim \mathcal{N}\Big(0\,,\, ((h^{-1})'(\vartheta_i))^2\,\frac{\vartheta_i(1-\vartheta_i)}{n_i}\Big) \qquad (7)$$

The approximate model (7) based on (aggregated) unit-level response differs from the model (2) primarily in allowing the 'sampling' variance term to vary with the PSU parameter $\vartheta_i = Y_i^0/N_i$.

It is clear from the foregoing discussion that an aggregated model (1) is likely to be well approximated by a unit-level Binomial model only if the h 'link' functions for the two models match. However, the most natural h to use is the *logistic*, and that is the one we use in (most of) the simulations reported below; while the h which has been used in the SAIPE county model (on the subset of the sampled data with non-zero response counts) is the exponential. Extra regression terms beyond the linear terms specified for these models may help in mitigating the effects of misspecifying the link. For this reason, we consider the effects in our simulations of incorporating a quadratic term in the explanatory variable.

As mentioned above, the methodology used so far in developing county-level small-area estimates of school-age poor children, as described in Bell (1997, 1998) and the 1999 Interim Report 3 (Citro and Kalton, editors) of the National Research Council, has been to fit a mixed-effect linear model like (3) on the sampled-county data for which at least one child in poverty was sampled. The model actually fitted was based on several covariates, not just the single one of model (3). For simplicity, the simulation will be based on data with a single covariate. The second method of statistical analysis studied here is based on approximate maximum likelihood estimation of parameters in the mixed-effect logistic model (1)-(2) with $h(x) = e^x/(1+e^x)$. As explained in the preprints Slud (1998, 2000), there is no generally available software for ML estimation in mixed logistic models with large numbers of PSU's which works sufficiently rapidly and reliably to be used in a comparative simulation. The approximate method devised in Slud (1998) was based on applying mixed-effect nonlinear regression software (function **nlme**) available in **Splus**, after transforming the PSU relative-frequency data by the variance-stabilizing transformation $\arcsin(\sqrt{x})$. An earlier version

of the work reported in Slud (2000) gave an approach to the approximate calculation of likelihoods within large mixed logistic models, which initially was intended as a way to calculate mixed-logistic deviances based on estimates obtained from the mixed nonlinear regression approach. However, simulation study of the behavior on moderate to large sample data of the nonlinear-regression estimation method showed that it can often get stuck at local optima on the boundary of the parameter space corresponding to 0 random-effect variances. An alternative approach based upon maximizing an accurately approximated log-likelihood function was shown in Slud (2000) to give generally good and numerically well-behaved ML estimators for the mixed logistic model which correct the evident tendency of the mixed nonlinear-regression estimators in the setting of the present simulation to underestimate the cluster variance. This alternative approximate ML estimator for the mixed-logistic model (1)-(2) is the one which is used in the simulations of later Sections.

4 Small-Area Estimates

It is desired to estimate the parameters $\vartheta_k = Y_k^0/N_k$, based on covariates X_k which are constant over the k'th PSU. The parameters $(\gamma_0, \gamma_1, \sigma^2)$ in model (1) are first to be estimated, either by a mixed-linear-regression methodology with $h(x) = e^x$ or an approximate ML methodology with $h(x) = (1+e^{-x})^{-1}$ ignoring the sampling errors e_k in (2). These estimators will then be substituted into modified 'EBLUP' small-area estimators (in the terminology of Prasad-Rao 1998 and Datta-Lahiri 1997). There are two separate cases: first, where the estimate of parameter ϑ_k in PSU k is based on no sampled data in the PSU, but only on the population-wide estimators $(\hat\gamma_0, \hat\gamma_1, \hat\sigma^2)$ and the known predictor variable X_k; and second, where ϑ_k in PSU k is estimated in terms of the population parameters and predictor plus an observed sample of n_k (with y_k^0 responses) in the k'th PSU. In each of the two cases, the sampling variability of the fixed-effect coefficient estimators $(\hat\gamma_0, \hat\gamma_1, \hat\sigma)$ can and should be taken into account. This is necessary in our context both in defining the small-area estimators themselves *and* in estimating mean-squared errors (MSE's). Many papers (see especially Prasad-Rao 1990 and Datta-Lahiri 1997) have underscored this in connection with MSE's, but these papers were dealing with cases where $h(x) = x$ in

the notation of model (1), so that the small-area estimators $\hat{\vartheta}_k$ would be linear in the observed responses y_k^0/n_k (and therefore also in the errors U_k and coefficients (γ_0, γ_1)) and the distribution (assumed normal, centered at the true values) of the fixed-effect coefficient estimators did not enter the EBLUP. Here, however, especially for nonlinear h, the estimators are based on nonlinear functions of observed response rates.

4.1 Modified EBLUP's & Approximate Distributions

The small-area estimator for ϑ_k should, if we knew the coefficients (γ_0, γ_1) in model (1) exactly, be based on the random quantity $\gamma_0 + \gamma_1 X_k + U_k$ and the conditional distribution of the PSU random effect U_k given the observed data. Hence, by analogy with BLUP's, our principle of estimation is that ϑ_k is to be estimated by the best available approximation or estimator for the conditional expectation

$$E\Big(h(\gamma_0 + \gamma_1 X_k + U_k) \,\Big|\, (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2), y_k^0\Big) = E\Big(E\Big(h(\gamma_0 + \gamma_1 X_k + U_k) \,\Big|\, (y_i^0, i \in s)\Big) \,\Big|\, (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2), y_k^0\Big) \quad (8)$$

In the setting where there is no sample in a PSU, the conditional expectation should instead be taken given only the parameter estimators $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2)$. The estimators so defined will be approximately unbiased by construction, and the corresponding (estimated) unconditional variances are the most natural candidates for estimating MSE.

Estimators based on (8) explicitly require some approximation to the joint distribution of U_k and $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2)$, and we adopt the following approximation, which can be tested in subsequent simulations:

> **Assumption A.** The k'th PSU random-effect U_k is approximately conditionally independent of the estimators $(\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2)$ and of $(y_i^0, i \in s\backslash\{k\})$, given y_k^0, and the parameter estimators are conditionally jointly normally distributed with means equal to some nonrandom values (the true values when model (1)-(2) holds) and variances the same for all k.

In fact, although the approximate normality does appear adequate for many real applications such as SAIPE, the displayed conditional independence assumption is suspect for the largest PSU's, which may be influential in the model-fitting. A correction to the assumption could be imagined, but in light of the simulation results to be reported below may not be necessary, since the modified EBLUP small-area estimators seem extremely accurate for large PSU's, and since the largest PSU's are by design chosen to be sampled.

We apply Assumption **A** separately for the two models we want to compare. Consider first the case of model (3) — motivated by (1)-(2) with $h(x) = e^x$ — and denote $y_i = \log(y_i^0/n_i)$. This model is a Fay-Herriot (1979) model as in Prasad-Rao (1990, Sec. 2.3), within which the conditional law of U_k given y_k is

$$\mathcal{L}\Big(U_k \,\Big|\, y_k\Big) \;=\; \mathcal{N}\Big(\, \frac{\sigma^2}{\sigma^2 + v_e/n_k}\,(y_k - \gamma_0 - \gamma_1 X_k)\,,\; \frac{\sigma^2 v_e/n_k}{\sigma^2 + v_e/n_k}\,\Big)$$

Then, by our approximate assumption **A**, we find that conditionally given the sampled value y_k and the parameter estimators,

$$\begin{aligned}\mathcal{L}\Big(\hat\gamma_0 + \hat\gamma_1 X_k + U_k \,\Big|\, (\hat\gamma_0,\, \hat\gamma_1,\, \hat\sigma^2),\, y_k\Big) \;\approx\; \mathcal{N}\Big(\hat\gamma_0 + \hat\gamma_1\, X_k + \\ \frac{\sigma^2}{\sigma^2 + v_e/n_k}\,(y_k - \hat\gamma_0 - \hat\gamma_1\, X_k)\,,\; \frac{\sigma^2 v_e}{n_k\sigma^2 + v_e}\,\Big)\end{aligned} \tag{9}$$

If the conditional law given only y_k^0 but *not* given the parameter estimators, were desired, then the estimators $\hat\gamma$ in the normal mean would be replaced by their true values, and the variance in the last normal distribution would increase by

$$a_k^2 \;=\; \begin{pmatrix} 1 \\ X_k \end{pmatrix}' \Sigma_\gamma \begin{pmatrix} 1 \\ X_k \end{pmatrix}$$

where Σ_γ denotes the (large-sample) variance-covariance matrix for the fixed-effect estimators $\hat\gamma_0,\ \hat\gamma_1$. In the case of the linear model for log-counts, the approximate conditional law given the parameter estimators is

$$\mathcal{L}\Big(\hat\gamma_0 + \hat\gamma_1 X_k + U_k \,\Big|\, (\hat\gamma_0,\, \hat\gamma_1,\, \hat\sigma^2)\Big) \;\approx\; \mathcal{N}\Big(\gamma_0 + \gamma_1 X_k\,,\, \sigma^2\Big) \tag{10}$$

Note that the conditional expectation which will be calculated in applying (8) in this setting has the form of the moment generating function for the normal distribution:

$$E(e^{\alpha+\beta Z}) \;=\; e^{\alpha+\beta^2\tau^2/2} \quad , \qquad Z \,\sim\, \mathcal{N}(0, \tau^2) \tag{11}$$

Next we apply assumption **A** in the model (4)-(5). In this setting, we have observations $y_k^0 \sim \text{Binom}(n_k,\, \pi_k)$ with $\pi_k = (1+\exp(-\gamma_0-\gamma_1 X_k-U_k))^{-1})$, and the conditional density of U_k given $(\hat\gamma_0,\, \hat\gamma_1,\, \hat\sigma^2)$ and $y_k^0 = m$ becomes approximately proportional to

$$\exp(m(\gamma_0 + \gamma_1 X_k + u) - u^2/(2\sigma^2)) \Big/ (1 + e^{\gamma_0+\gamma_1 X_k+u})^{n_k} \tag{12}$$

and again $\hat\gamma_0 + \hat\gamma_1 X_k$ is approximately independent of U_k with

$$\hat\gamma_0 + \hat\gamma_1\, X_k \;\approx\; \mathcal{N}\Big(\gamma_0 + \gamma_1\, X_k,\; a_k^2\Big) \tag{13}$$

The expectations which we will calculate in this model involve two functions which we now define:

$$g(x, b) \;=\; \int \frac{e^{x+bz}}{1+e^{x+bz}}\, \phi(z)\, dz \tag{14}$$

and

$$\Lambda(x, m, n, b) \;=\; \int \frac{e^{m(x+bz)}}{(1+e^{x+bz})^n}\, \phi(z)\, dz \tag{15}$$

where $\phi(\cdot)$ denotes the standard normal density. Note that by definition $g(x, b) = \Lambda(x, 1, 1, b)$, and the function (15) satisfies the recursion $\Lambda(x, m, n+1, b) + \Lambda(x, m+1, n+1, b) = \Lambda(x, m, n, b)$.

Note also that if $X \sim \mathcal{N}(\alpha, \beta^2)$, then

$$\begin{aligned} Eg(X, b) \;&=\; E \int \frac{e^{X+bu}}{1+e^{X+bu}}\, \phi(u)\, du \;=\; \int \frac{e^{\alpha+v\sqrt{b^2+\beta^2}}}{1+e^{\alpha+v\sqrt{b^2+\beta^2}}}\, \phi(v)\, dv \\ &=\; g(\alpha,\, \sqrt{b^2+\beta^2}) \end{aligned} \tag{16}$$

In the mixed logistic regression setting, the frequentist EBLUP small-area estimators are essentially like those of Booth and Hobert (1998), who also give an approach (slightly different from the one in Section 4.4 below) to the estimation of mean-squared errors. A Bayesian approach to small-area estimation for generalized-linear unit-level models is given by Ghosh et al. (1998).

4.2 Non-sampled PSU's

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

$$\hat{\vartheta}_k \;=\; \widehat{E}\,\Big(\, h(\gamma_0 + \gamma_1 X_k + U_k) \,\Big|\, (\hat{\gamma}_0, \hat{\gamma}_1, \hat{\sigma}^2)\,\Big)$$

where the expectation is taken over the errors U_k but *not* over the parameter estimators, and the estimator $\widehat{E}$ of the expectation will have parameter-estimators $(\hat{\gamma}_0,\, \hat{\gamma}_1,\, \hat{\sigma}^2)$ substituted and will be bias-corrected as far as possible. In the setting of the Fay-Herriot model(3) with $h(x) = e^x$, we obtain via assumption **A**, (10), and (11) after substituting estimators for the unknown parameters, that

$$\hat{\vartheta}_k \;=\; \exp\Big(\hat{\gamma}_0 \,+\, \hat{\gamma}_1\, X_k \,+\, \frac{1}{2}\,\hat{\sigma}^2 \,-\, \frac{1}{2}\,\hat{a}_k^2\Big) \tag{17}$$

Here and in what follows, we assume that the estimation-algorithm for fixed-effect coefficients automatically produces an estimator $\hat{\Sigma}_\gamma$ for the resulting covariance matrix of coefficient-estimators. For simplicity of notation from now on, we define

$$\eta_k \;=\; \gamma_0 \,+\, \gamma_1\, X_k \quad , \quad \hat{\eta}_k \;=\; \hat{\gamma}_0 \,+\, \hat{\gamma}_1\, X_k \quad , \qquad \hat{a}_k^2 \;=\; \begin{pmatrix} 1 \\ X_k \end{pmatrix}' \widehat{\Sigma}_\gamma \begin{pmatrix} 1 \\ X_k \end{pmatrix}$$

The term $\hat{a}_k^2/2$ in (17) is a bias-correction, since according to (11),

$$E\Big(\, \exp(\hat{\gamma}_0 + \hat{\gamma}_1\, X_k - \gamma_0 - \gamma_1\, X_k)\,\Big) \;\approx\; e^{a_k^2/2}$$

The bias-corrected small-area estimator (conditional on estimated parameters) for the model (4)-(5) becomes, after applying (8) along with (13) and (16), using the notation (14) and substituting parameter estimators,

$$\hat{\vartheta}_k \;=\; g\Big(\hat{\eta}_k\,,\, \sqrt{\hat{\sigma}^2 - \hat{a}_k^2}\Big) \;=\; g\Big(\hat{\gamma}_0 \,+\, \hat{\gamma}_1 X_k\,,\, \sqrt{\hat{\sigma}^2 - \hat{a}_k^2}\Big) \tag{18}$$

Of course, if the quantity $\hat{\sigma}^2 - \hat{a}_k^2$ in the square-roots were negative, it should be understood to be replaced by 0, and $g(\cdot, 0)$ is simply the logistic distribution function.

4.3 Sampled PSU's

In the case of a small-area estimator for a PSU k from which a sample has been drawn, we make direct use of (8), assumption **A**, and the conditional distributions described in Section 4.1. First, in the Fay-Herriot model (1) with $h(x) = e^x$, we apply the conditional distribution (9) along with (11) to obtain after substituting parameter estimators (and correcting for bias)

$$\hat{\vartheta}_k \;=\; \exp\Big(\hat{\eta}_k \;+\; \frac{\hat{\sigma}^2}{\hat{\sigma}^2 + v_e/n_k}\,(y_k - \hat{\eta}_k) \;+\; \frac{1}{2}\left[\frac{\hat{\sigma}^2\, v_e}{n_k\,\hat{\sigma}^2 + v_e} \;-\; \frac{(\hat{a}_k\, v_e/n_k)^2}{(\hat{\sigma}^2 + v_e/n_k)^2}\right]\Big) \tag{19}$$

Finally, using the mixed logistic regression model (4)-(5) and $h(x) = e^x/(1+e^x)$, we obtain for (8) via the approximate distributions in Section 4.1

$$E\Big(\,h(\hat{\gamma}_0 + \hat{\gamma}_1 X_k + U_k)\,\Big|\,(\hat{\gamma}_0,\,\hat{\gamma}_1,\,\hat{\sigma}^2),\,y_k^0\Big) \;\approx$$

$$\int \frac{\exp(y_k^0(\hat{\eta}_k + \sigma z))}{(1+e^{\hat{\eta}_k+\sigma z})^{n_k}}\; h(\hat{\eta}_k + \sigma z)\,\phi(z)\,dz \;\Big/\; \int \frac{\exp(y_k^0(\hat{\eta}_k + \sigma z))}{(1+e^{\hat{\eta}_k+\sigma z})^{n_k}}\;\phi(z)\,dz$$

which is equal by definition to

$$\Lambda(\,\hat{\eta},\,y_k^0+1,\,n_k+1,\,\sigma)\,\Big/\,\Lambda(\,\hat{\eta},\,y_k^0,\,n_k,\,\sigma)$$

However, it is not at all clear that a simple bias-correction would be available for this ratio. Thus we obtain, for the mixed logistic unit-level model

$$\hat{E}\Big(\,h(\hat{\gamma}_0 + \hat{\gamma}_1 X_k + U_k)\,\Big|\,(\hat{\gamma}_0,\,\hat{\gamma}_1,\,\hat{\sigma}^2),\,y_k^0\Big) \;=\; \hat{\vartheta}_k \;=\; \frac{\Lambda(\,\hat{\eta},\,y_k^0+1,\,n_k+1,\,\sigma)}{\Lambda(\,\hat{\eta},\,y_k^0,\,n_k,\,\sigma)} \tag{20}$$

4.4 Estimation of MSE's

Applications of small-area estimators $\hat{\vartheta}_k$ usually require standard errors as well as point estimates, and the ideas above lend themselves readily to the approximation and estimation of mean-squared errors. The key idea is that the proposed estimators $\hat{\vartheta}_k$ have the form

$$\hat{\vartheta}_k \;=\; t_s(y_k^0/n_k,\,\hat{\eta}_k,\,\hat{\sigma}^2,\,\hat{a}_k^2) \qquad \text{or} \qquad t_u(\hat{\eta}_k,\,\hat{\sigma}^2,\,\hat{a}_k^2)$$

respectively for sampled and non-sampled PSU's, where $t_s(\cdot)$ and $t_u(\cdot)$ are smooth, nonrandom, known functions given in formulas (17) and (19) for the aggregated linear model and (18) and (20) for the mixed logistic unit-level model. The estimators are constructed to be approximately unbiased, when parameters are replaced by their correct values and PSU sample averages are replaced by the PSU averages, which means in particular that for each model used we take

$$t_s(\frac{y_k^0}{n_k}, \eta_k, \sigma^2, 0) \approx E\left(h(\eta_k + U_k) \,\middle|\, y_k^0\right) \quad , \quad t_u(\eta_k, \sigma^2, 0) \approx E\left(h(\eta_k + U_k)\right)$$

where recall that

$$\vartheta_k = \frac{Y_k^0}{N_k} = h(\eta_k + U_k)$$

From this starting point, maintaining Assumption **A** and the requirement that sample sizes are large enough so that parameter estimators are approximately jointly normal with small variances, we ignore the randomness in $\hat{a}_k$ and apply the delta-method to estimate MSE's. We develop MSE's in detail for the case of non-sampled PSU's, with the convention that $t_{u,r,k}$ is the derivative of the function $t_u(\cdot)$ with respect to its r'th argument, evaluated at $(\eta_k, \sigma^2, 0)$:

$$E(\hat{\vartheta}_k - \vartheta_k)^2 \approx \mathrm{Var}\left(h(\eta_k + U_k)\right) +$$

$$E\left((\hat{\eta}_k - \eta_k)\, t_{u,1,k} + (\hat{\sigma}^2 - \sigma^2)\, t_{u,2,k} + a_k^2\, t_{u,3,k}\right)^2$$

The resulting MSE estimators, obtained by substituting parameter-estimators on the right-hand side after dropping the $t_{u,3,k}$ terms, are respectively

$$\widehat{\mathrm{MSE}}(\hat{\vartheta}_k) = e^{2\hat{\eta}+\hat{\sigma}^2}\left\{(e^{\hat{\sigma}^2} - 1) + \hat{a}_k^2 + \frac{1}{4}\widehat{\mathrm{Var}}(\hat{\sigma}^2)\right\} \tag{21}$$

in applying (17), where $t_u(\eta, \sigma^2, a^2) = \exp(\eta + (\sigma^2 - a^2)/2)$ and we have used the fact that $\hat{\eta}_k$ and $\hat{\sigma}^2$ are approximately uncorrelated in the linear model, along with (11) for $\alpha = 0$, $\beta = 2$, $Z = 2U_k$, and

$$\widehat{\mathrm{MSE}}(\hat{\vartheta}_k) = \widehat{\mathrm{Var}}\left(\frac{e^{\eta_k + U_k}}{1 + e^{\eta_k + U_k}}\right) + \hat{a}_k^2 \left(\frac{\partial g}{\partial \eta}(\hat{\eta}_k, \hat{\sigma})\right)^2$$

$$+ \frac{1}{\hat{\sigma}} \widehat{E}\left((\hat{\eta}_k - \eta_k)(\hat{\sigma}^2 - \sigma^2)\right) \left(\frac{\partial g}{\partial \eta}\frac{\partial g}{\partial \sigma}\right)_{(\hat{\eta}_k, \hat{\sigma})}$$

$$+ \frac{1}{4\,\hat{\sigma}^2} \widehat{\mathrm{Var}}(\hat{\sigma}^2) \left(\frac{\partial g}{\partial \sigma}(\hat{\eta}_k, \hat{\sigma})\right)^2 \qquad (22)$$

for the PSU estimator (18), for which $t_u(\eta, \sigma^2, a^2) = g(\eta, \sqrt{\sigma^2 - a^2})$. Within (22), we would approximate $g(\hat{\eta}_k, \hat{\sigma})$ by $\hat{\vartheta}_k$ or ϑ_k.

A bias-corrected form of (21) is given by

$$\widehat{\mathrm{MSE}}(\hat{\vartheta}_k) = \exp\left(2\hat{\eta} - 2\hat{a}_k^2\right) \cdot \Big(\exp(2\,\hat{\sigma}^2 - 2\,\widehat{\mathrm{Var}}(\hat{\sigma}^2)) - \exp(\hat{\sigma}^2 - \widehat{\mathrm{Var}}(\hat{\sigma}^2)/2) \cdot [1 - \hat{a}_k^2 - \frac{1}{4}\widehat{\mathrm{Var}}(\hat{\sigma}^2)]\Big) \qquad (23)$$

4.5 Robust Standard Errors

In practice, the models used to generate Small-Area Estimates cannot be expected to fit very closely. Therefore, in justifying the error structure of estimation formulas, any use of the assumption that models are properly specified is suspect. The parameters $(\gamma_0, \gamma_1, \sigma^2)$ which are being estimated will in any case have a meaningful interpretation as the parameters in the best-fitting model to the complete PSU data Y_i^0/N_i, but large-sample reasoning based upon likelihood methods can lead to incorrect conclusions when the models are actually misspecified (White 1982). However, even in such cases the parameter estimators, which can be viewed as 'M-estimators', will often have large-sample approximately normal distributions with variances which can be found via a famous 'sandwich formula' originally due to Huber and explained in the exposition of White (1982).

Although robust standard errors are desirable, we do not pursue methods to obtain them here, since the effect of the parameters Σ_γ and a_k on the performance of small-area estimators in the parameter- and sample-size ranges of the simulations below appear rather small.

5 Simulation Methodology & Results

In this section, we describe in detail a simulation experiment (without PSU or unit weighting) designed to compare small-area estimates based upon the SAIPE methodology for county estimates (involving mixed-effect linear models for log poverty rates, omitting sampled counties showing zero counts of poor school-age children) versus individual-level models with approximate maximum likelihood parameter estimates for a mixed logistic model.

We began by fixing once and for all the numbers n_i, a set of 1488 PSU sample-sizes corresponding to the (non-zero) numbers sampled by the Current Population Survey in counties over the 3-year period 1992-94 *(supplied by Matt Kramer in the file* **smpsiz.3yr***)*. The distribution of numbers sampled within PSU's is very skewed, ranging from 1 to 2226, with 34% of PSU's containing 10 or fewer sampled, and 57% containing 20 or fewer. The overall PSU sizes N_i play a direct role only in the estimation of $\omega_* = v_e$, but are regarded as fixed at some factor (initially 2000 — approximately the reciprocal sampling fraction of the CPS) multiplied by the numbers sampled within PSU.

The single predictor variable X_i which we use is meant to mimic the IRS-supplied estimate of log numbers of poor children in county *(column IRSPr5-17 from the predictor-variable file X93.SAIPE, also supplied by Matt Kramer)*. This predictor column has been simulated here, at most once for each Table based on 100 simulation iterations, as a column of independent $\mathcal{N}(7.33,\, 1.69)$ random variables. (The 7.33 is actually replaced by 0, since an intercept term is always fitted.) However, to prevent unrealistically large variation in response fractions for the PSU's with very large samples (those > 220), we made the predictors X_i for these PSU's much smaller: in Tables 1 to 9, all of which corresponded to the *same* column of X_i, we fixed the X_i values for these large-sample PSU's to be 0; in later Tables, we multiplied by 0.25 the simulated X_i values corresponding to the 30 indices i with $n_i > 220$.

In each simulation iteration, individual units j within PSU i have binary responses y_{ji} simulated as

$$y_{ji} \sim Binom\,(1,\, \pi_i)\ , \qquad \pi_i = h(\gamma_0 + \gamma_1 X_i + U_i) \tag{24}$$

where the function $h(x)$ (equal to $e^x/(1+e^x)$ unless indicated otherwise)

and parameters $(\gamma_0, \gamma_1, \sigma^2)$ are fixed separately within each simulation and $U_i \sim \mathcal{N}(0, \sigma^2)$. The initial choice of these parameters, was:

- To get 5 to 40 % in poverty, with overall rate of the order of 20%, we first take intercept $\gamma_0 = \log(0.2)$, or approximately -1.6.
- Then, with coefficient $\gamma_1 = 0.9$, the overall response rate is 0.22.

Initially, and throughout this Section, the simulations took all sampling weights identically equal to 1.

For each iteration of each simulation, estimators of $\gamma_0, \gamma_1, \sigma^2$ are calculated in the two ways described in the Analysis Methods section above, using **Splus** software, except that the mixed logistic estimates are themselves calculated in two ways: first via the nonlinear mixed-effect regression model described in Slud (1998)

$$\arcsin\left(\sqrt{\frac{y_i^0}{n_i}}\right) \;=\; \arcsin\left(\sqrt{\frac{e^{\eta_i+U_i}}{1+e^{\eta_i+U_i}}}\right) \;+\; \frac{\rho_i}{2\sqrt{n_i}} \tag{25}$$

where the variables ρ_i are standard normal and independent across PSU's, and then for comparison using the maximizer of the Laplace-method approximation to mixed logistic log-likelihood, as described in detail in Slud (2000). Next the small-area estimators $\hat{\vartheta}_i$ described in Section 4 are calculated. For each simulation iteration, the retained results include for each of the three sets of parameter estimators:

- the estimates for $(\gamma_0, \gamma_1, \sigma^2)$;
- empirical average biases and Mean-Squared Errors for small-area estimators by PSU, grouped into the 5 categories defined by number n_i sampled respectively in the intervals $(0, 10]$, $(10, 25]$, $(25, 75]$, $(75, 220]$, and $(220, 2500]$.

The respective numbers of PSU's with number sampled falling into these intervals are: 506, 448, 398, 106, and 30.

There is one further set of estimators which has been used to generate results in the aggregated linear model (3) to reflect the fact that SAIPE small-area estimates have on occasion proceeded from an assumption that σ^2 was

known and v_e estimated, by ML from the same Fay-Herriot model (3). The value $\sigma^2 = \sigma_*^2$ used in the simulations was estimated as the mean-squared residual error from the regression of $\log(Y_i^0/N_i)$ on 1, X_i. The estimators generated using the Fay-Herriot fit with σ^2 estimated and $v_e = \omega_*$ known are designated by column-headers **lmfitA**; those with σ_*^2 known and v_e estimated, by headers **lmfitB**. The column-headers for estimators within the mixed-logistic unit-level model are: **nlmfit** when parameters are estimated using the transformed nonlinear regression model (25), and **glmfit** when parameters are estimated by maximizing the approximate mixed-logistic log-likelihood of Slud (2000).

Throughout the simulations reported below, the estimated PSU response rates ϑ_k estimated from the nonlinear models were necessarily given as probabilities, and therefore lay between 0 and 1. However, the estimators derived from the linear (Fay-Herriot) models for log response rate did occasionally turn out to be greater than 1, either because of imprecision of the substituted parameter estimators or because of excessively large bias corrections. In those cases, it was important to replace those estimators by 1 to maintain a fair comparison between the Mean-squared Errors of the small-area-estimators derived from the linear and nonlinear models.

Throughout the simulations below, the quantities Λ appearing in formula (20) were calculated via a Laplace-method approximation in groups 3–5 and a very accurate series approximation of Crouch and Spiegelman in Groups 1–2. All of these approximate values were documented in Slud (2000) to be accurate to no less than one part in 10^5.

5.1 Simulation Results — Single Predictor

The Tables summarize the results of simulation experiments of 100 iterations each, with parameters specified as above. In all of these simulations, the reported **bias** and **MSE** numbers are empirical and do not yet reflect attempts to validate theoretical estimators of MSE. The simulations of this Section are all based upon unweighted analyses, and upon the same set of scalar predictors X_i, with two exceptions: Tables 5 and 7 are based on a different simulated predictor-column X_i .

Note that at present, no separate standard error estimates for parameter

estimators have been generated under the approximate mixed-logistic ML estimation method.

The first set of three simulations, with results in Tables 1 to 3, relate to $(\gamma_0, \gamma_1, \sigma) = (-1.6, 0.9, 0.3)$. Table 1 shows the behavior of estimators based on formulas (17) and (18). These should be interpreted as follows: imagine that, corresponding to each of the 1488 PSU's which were sampled in each simulation iteration, there is another PSU with the same size and predictor value X_i, from which no sample was drawn. The small-area estimators $\hat{\vartheta}_i$ would be those of (17) and (18), and Simulation 1 reflects their empirical performance. The second simulation shows the performance of modified-EBLUP estimators (19) and (20), and the third simulation is exactly like the second *except* that in the fitting of all models the predictor covariates X_i (plus intercept, with coefficients γ_1, γ_0) used in parameter estimation are augmented by X_i^2 (with a third fixed-effect coefficient γ_2).

Table 1 exhibits not only the groupwise averaged Mean-Squared Errors for the PSU response-rate estimators, but also the groupwise-average biases. However, these biases do not appear very informative, and we omit them from all later tables.

TABLE 1. SIMULATED PERFORMANCE OF SMALL-AREA ESTIMATORS FOR NON-SAMPLED PSU'S ($\sigma = 0.3$), VIA FORMULAS (17), (18). Average estimated ω^* in **lmfitA** is 9.916; average estimated σ_*^2 in **lmfitB** is 0.0802.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.67118	-1.59886	-1.84895	-1.83096
b1	0.95364	0.89725	0.63384	0.60788
sigP	0.25849	0.29991	0.15813	1.73853
Varb0	0.00044	NA	0.00028	0.00018
Varb1	0.00031	NA	0.00020	0.00012
Bias1	-0.00558	-0.00038	-0.00081	0.00382
MSE1	0.00231	0.00223	0.00389	0.00366
Bias2	-0.00599	0.00013	-0.00417	0.00104
MSE2	0.00217	0.00210	0.00333	0.00316
Bias3	-0.00521	0.00016	-0.00023	0.00420
MSE3	0.00237	0.00229	0.00403	0.00370
Bias4	-0.00475	0.00077	-0.00237	0.00244
MSE4	0.00257	0.00250	0.00375	0.00360
Bias5	-0.01154	-0.00053	-0.01388	-0.00650
MSE5	0.00195	0.00183	0.00201	0.00187

TABLE 2. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.3$), VIA FORMULAS (19), (20) WITH SINGLE PREDICTOR VARIABLE X_i. Average estimated ω^* in **lmfitA** is 9.983; average estimated σ_*^2 in **lmfitB** is 0.0800.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.67672	-1.60370	-1.85122	-1.83350
b1	0.95512	0.89790	0.63374	0.60749
sigP	0.25345	0.29784	0.15718	1.73772
Varb0	0.00044	NA	0.00028	0.00018
Varb1	0.00031	NA	0.00020	0.00012
MSE1	0.00232	0.00224	0.00378	0.00335
MSE2	0.00224	0.00216	0.00318	0.00259
MSE3	0.00142	0.00133	0.00334	0.00194
MSE4	0.00073	0.00070	0.00237	0.00085
MSE5	0.00030	0.00029	0.00061	0.00030

TABLE 3. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.3$), VIA FORMULAS (19), (20) WITH PARAMETERS FITTED USING PREDICTOR VARIABLES X_i, X_i^2. Average estimated ω^* in **lmfitA** is 9.984; average estimated σ_*^2 in **lmfitB** is 0.0624.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.65396	-1.60134	-1.81515	-1.80562
b1	0.98410	0.89819	0.64581	0.62563
b2	-0.02335	0.00235	-0.02596	-0.02237
sigP	0.24810	0.29768	0.14323	1.88163
Varb0	0.00051	NA	0.00037	0.00025
Varb1	0.00052	NA	0.00021	0.00013
Varb2	0.00014	NA	0.00006	0.00004
MSE1	0.00236	0.00229	0.00315	0.00295
MSE2	0.00221	0.00215	0.00276	0.00232
MSE3	0.00141	0.00133	0.00255	0.00173
MSE4	0.00077	0.00072	0.00200	0.00082
MSE5	0.00029	0.00029	0.00061	0.00029

The results in Tables 1-3 indicate very satisfactory performance by the estimators for the (properly specified) unit-level model, except that the nonlinear-regression seems to show persistent small biases in the population-wide parameter estimators in and underestimation of the PSU cluster variance component. These biases are not present in the approximate mixed-logistic ML methods, and appear to account for the slightly superior performance of its small-area estimates.

Not much emerges by examining the biases within groups of PSU's in the Tables, but the MSE's are very interesting and revealing. The MSE's for $\hat{\vartheta}_k$ in nonsampled PSU's are of the order of 0.002 across the board in the methods based on the unit-level model, but tend to be larger by a factor of roughly 1.5 (in PSU groups 1, 2, 3, and 4) in the linear-model methods, except for very large PSU's (Group 5) in which the linear models do about as well as the nonlniear. As would be expected, in Tables 2 and 3 corresponding to EBLUP's in sampled PSU's, the MSE's are much smaller in large PSU's, for all methods of estimation. However, the advantage of using the properly specified unit-level model over the analysis in **lmfitA** becomes especially pronounced in the larger PSU's (groups 3 to 5). In groups 1, 2 and 3 (PSU samples of 25 or less), the MSE's for **lmfitB** EBLUPs are

from 1.3 to 1.5 times worse than for the unit-level logistic model, although the nonlinear-model results and those of **lmfitB** are quite comparable in the larger-PSU groups (4 and 5). But very strikingly, when the quadratic predictor X_i^2 is used in the linear-model fitting, the MSE's for **lmfitB** (with the properly chosen value for σ^2 taken as known) are nearly as good as for the mixed-logistic analyses based on the single predictor X_i.

A technical aspect of the estimation methodology, evaluated and checked by further runs in the setting of Tables 1 to 3, was that the effect of omitting the terms a_k^2 in the EBLUP estimators of Table 2 is completely negligible to the accuracy shown: in only a few entries, this omission caused a change of 1 digit in the fifth decimal place. The same assertion was checked to hold also for all of the other EBLUP simulations reported in this sub-section.

The next three Tables are analogous to the first three, the only difference being that a smaller cluster standard deviation $\sigma = 0.2$ replaces the previous value of 0.3. One purpose of reducing this parameter was to mimic the realistic (for SAIPE) setting where the differences between (linear-model) small-area estimators for non-sampled PSU's and EBLUP estimators for sampled PSU's would be very small for small PSU's. That is, in the small PSU's it is found that the linear-model estimators for sampled PSU's place hardly any weight on the direct-sample estimators y_k within formula (19), and this is true also in the comparison of MSE's in the **lmfit** columns of Tables 4, 5.

TABLE 4. SIMULATED PERFORMANCE OF SMALL-AREA ESTIMATORS FOR NON-SAMPLED PSU'S ($\sigma = 0.2$), VIA FORMULAS (17), (18). Average estimated ω^* in **lmfitA** is 9.722; average estimated σ_*^2 in **lmfitB** is 0.0492.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.67454	-1.60020	-1.85158	-1.84394
b1	0.95638	0.89963	0.64481	0.62525
sigP	0.15086	0.19690	0.08275	1.87128
Varb0	0.00034	NA	0.00023	0.00016
Varb1	0.00025	NA	0.00018	0.00011
MSE1	0.00109	0.00101	0.00273	0.00252
MSE2	0.00104	0.00096	0.00219	0.00205
MSE3	0.00111	0.00103	0.00287	0.00255
MSE4	0.00119	0.00110	0.00245	0.00231
MSE5	0.00087	0.00077	0.00090	0.00083

Again the unit-level model estimators (**nlmfit** and **glmfit**) in the no-sample case (Table 4) uniformly outperform the linear-model estimators (with MSE's half as large) for all PSU groups other than 5. A similar comment holds for the EBLUP estimators (Table 5) in Groups 1 to 3, beyond which the advantage of **nlmfit** MSE's versus **lmfitB** is only 50% in Group 4 and disappears in Group 5.

TABLE 5. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.2$), VIA FORMULAS (19), (20) WITH SINGLE PREDICTOR VARIABLE X_i. Average estimated ω^* in **lmfitA** is 9.682; average estimated σ_*^2 in **lmfitB** is 0.0487.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.67123	-1.59672	-1.84753	-1.84029
b1	0.95647	0.89939	0.64163	0.62281
sigP	0.15017	0.19834	0.08190	1.87796
Varb0	0.00034	NA	0.00023	0.00016
Varb1	0.00025	NA	0.00018	0.00011
MSE1	0.00110	0.00103	0.00266	0.00237
MSE2	0.00103	0.00096	0.00212	0.00181
MSE3	0.00087	0.00077	0.00261	0.00159
MSE4	0.00060	0.00051	0.00210	0.00084
MSE5	0.00029	0.00024	0.00057	0.00027

Although we used formula (20) in estimating PSU response rate for the unit-level (nonlinear) models in PSU Groups 1 and 2, in these groups the nonlinear MSE entries are essentially identical in Tables 4 and 5. Note again that there are dramatic differences in Groups 3 to 5 for the **nlmfit** and **glmfit** MSE's respectively between the Table 4 and 5 entries.

Table 6 shows that inclusion of a quadratic predictor X_i^2 makes the linear-model **limfitB** estimator almost competitive with the nonlinear-model estimators (although the **lmfitB** results shown here rely on the correctly known value σ_*^2). Results for (**lmfitA** are also much improved by the additional predictor, but still much worse than **lmfitB**.) But in Groups 1 to 3, it remains true that **lmfitB** has MSE's larger by at least half than either **nlmfit** or **glmfit**.

TABLE 6. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.2$), VIA FORMULAS (19), (20) WITH PARAMETERS FITTED USING PREDICTOR VARIABLES X_i, X_i^2. Average estimated ω^* in **lmfitA** is 9.662; average estimated σ_*^2 in **lmfitB** is 0.0309.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.65289	-1.59774	-1.81262	-1.81203
b1	0.98907	0.89913	0.65481	0.64262
b2	-0.02717	0.00019	-0.02890	-0.02560
sigP	0.15008	0.19832	0.07130	2.03317
Varb0	0.00039	NA	0.00030	0.00021
Varb1	0.00046	NA	0.00018	0.00011
Varb2	0.00012	NA	0.00005	0.00003
MSE1	0.00111	0.00104	0.00187	0.00183
MSE2	0.00104	0.00098	0.00162	0.00149
MSE3	0.00087	0.00078	0.00157	0.00123
MSE4	0.00063	0.00053	0.00160	0.00076
MSE5	0.00026	0.00023	0.00052	0.00024

After comparing the foregoing Tables, for unweighted sample-survey data simulated according to the mixed logistic model, we arrive at the tentative conclusion that the best available small-area estimator based upon the generalized-linear unit-level model would be the **glmfit** estimator, which uses direct-sample data in a PSU via the EBLUP formula (20). The best competitor using an aggregated log-linear model is the **lmfitB** estimator including a quadratic predictor, at least when the assumption of known σ_*^2 is approximately correct.

We proceed next to consider the performance of the same small-area estimators upon simulations of data from models other than the unit-level mixed-effect logistic. Tables 7 to 9 concern the model (24) with $h(x) = e^x$ with the particular parameter choices $\gamma_0 = -1.9$, $\gamma_1 = 0.3$, $\mathrm{var}(U_i) = (0.2)^2$. In view of the discussion surrounding equation (7), this model is designed to show the aggregated linear model (3) in its most favorable light.

Table 7 shows only a tiny advantage in MSE for the **lmfit**-based small-area estimators over those based on the **nlmfit**, **glmfit** mixed-effect logistic analyses. Table 8 does show an advantage in MSE for Groups 1 and 2 of the linear-model small-area estimators over those based on the (misspecified) mixed-effect logistic. In Table 8, where all models use only X_i as predictor,

the advantage (say of **lmfitB** over **glmfit**) is 5% in Group 1, 10% in Group 2, but disappears for PSU's with sample-size larger than 25.

TABLE 7. SIMULATED PERFORMANCE OF SMALL-AREA ESTIMATORS FOR NON-SAMPLED PSU'S ($\sigma = 0.2$, $h =$ *exp*), VIA FORMULAS (17), (18) WITH PARAMETERS FITTED USING PREDICTOR VARIABLE X_i. Average estimated ω^* in **lmfitA** is 6.380; average estimated σ_*^2 in **lmfitB** is 0.0402.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.78671	-1.72407	-1.92191	-1.92024
b1	0.39956	0.37167	0.28844	0.28705
sigP	0.18698	0.25215	0.19419	2.12572
Varb0	0.00034	NA	0.00021	0.00016
Varb1	0.00022	NA	0.00013	0.00011
MSE1	0.00150	0.00146	0.00145	0.00145
MSE2	0.00137	0.00132	0.00130	0.00130
MSE3	0.00138	0.00132	0.00135	0.00135
MSE4	0.00147	0.00143	0.00145	0.00145
MSE5	0.00096	0.00094	0.00093	0.00093

TABLE 8. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.2$, $h =$ *exp*), VIA FORMULAS (19), (20) WITH PARAMETERS FITTED USING PREDICTOR VARIABLE X_i. Average estimated ω^* in **lmfitA** is 6.351; average estimated σ_*^2 in **lmfitB** is 0.0403.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.78508	-1.72216	-1.91952	-1.91768
b1	0.40071	0.37215	0.28835	0.28672
sigP	0.18056	0.24719	0.18968	2.11283
Varb0	0.00034	NA	0.00020	0.00016
Varb1	0.00022	NA	0.00013	0.00010
MSE1	0.00153	0.00148	0.00143	0.00141
MSE2	0.00139	0.00135	0.00121	0.00119
MSE3	0.00106	0.00093	0.00108	0.00103
MSE4	0.00065	0.00055	0.00067	0.00060
MSE5	0.00025	0.00023	0.00024	0.00023

Remarkably, there are no other cases where the linear models perform better than the nonlinear: at least in this simulation, the misspecified linear-model fit due to the use of PSU-aggregated data which are discarded in those PSU's with zero response is no better than the fit due to an incorrectly specified

h function (logistic rather than exponential) in the unit-level model. In particular, when the quadratic predictor X_i^2 is included (Table 9), **lmfitB** is slightly better than **glmfit** in Group 2 and slightly worse in Group 3, but linear and nonlinear model fits are virtually equivalent from the vantage point of groupwise MSE.

TABLE 9. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.2$, $h = exp$), VIA FORMULAS (17), (18) WITH PARAMETERS FITTED USING PREDICTOR VARIABLES X_i, X_i^2. Average estimated ω^* in **lmfitA** is 6.380; average estimated σ_*^2 in **lmfitB** is 0.0404.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.78908	-1.73727	-1.92452	-1.92354
b1	0.39424	0.36276	0.28676	0.28499
b2	0.00480	0.01195	0.00403	0.00452
sigP	0.18184	0.24923	0.18994	2.11611
Varb0	0.00045	NA	0.00029	0.00023
Varb1	0.00026	NA	0.00013	0.00011
Varb2	0.00007	NA	0.00004	0.00003
MSE1	0.00149	0.00142	0.00143	0.00141
MSE2	0.00135	0.00128	0.00121	0.00118
MSE3	0.00103	0.00090	0.00104	0.00099
MSE4	0.00063	0.00054	0.00065	0.00060
MSE5	0.00024	0.00022	0.00023	0.00022

A tentative conclusion, well supported in all simulations done so far, is that using the **glmfit**-based small-area estimators in place of **lmfitB** can help considerably more than it is likely to hurt due to misspecification of the mixed logistic unit-level model. But other sorts of misspecifications remain to be explored in further simulations.

5.2 Displaying Results by Squared Relative Errors

Since results concerning mean-squared errors of small area estimators are often presented in terms of coefficients of variation (CV's), we calculated the results of the EBLUP simulations of Tables 2, 5, and 8 also by averaging squared relative errors

$$(\hat{\vartheta}_k - \vartheta_k)^2/\vartheta_k^2 \qquad \text{in place of} \qquad (\hat{\vartheta}_k - \vartheta_k)^2 \tag{26}$$

over PSU sample-size groups. Just for purposes of comparison with these respective Tables, we display the results here. As before, all simulations were done with 100 iterations.

TABLE 10. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.3$), VIA FORMULAS (19), (20) WITH SINGLE PREDICTOR VARIABLE X_i. Average estimated ω^* in **lmfitA** is 9.983; average estimated σ_*^2 in **lmfitB** is 0.0800. MSE's are now given as groupwise averages of squared relative errors (26).

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.6733	-1.60048	-1.84859	-1.83125
b1	0.95745	0.90037	0.63601	0.60898
sigP	0.25266	0.29790	0.15210	1.73669
MSE1	0.06302	0.07617	0.12743	0.18865
MSE2	0.06291	0.07664	0.12875	0.19667
MSE3	0.04435	0.05069	0.10435	0.12872
MSE4	0.02420	0.02531	0.05482	0.04550
MSE5	0.01120	0.01142	0.01880	0.01162

TABLE 11. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.2$), VIA FORMULAS (19), (20) WITH SINGLE PREDICTOR VARIABLE X_i. Average estimated ω^* in **lmfitA** is 9.682; average estimated σ_*^2 in **lmfitB** is 0.0487. MSE's are now given as groupwise averages of squared relative errors (26).

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.67123	-1.59672	-1.84753	-1.84029
b1	0.95647	0.89939	0.64163	0.62281
sigP	0.15017	0.19834	0.08190	1.87796
MSE1	0.03080	0.03134	0.07025	0.09566
MSE2	0.02981	0.03004	0.07171	0.10395
MSE3	0.02582	0.02540	0.06159	0.07606
MSE4	0.01625	0.01478	0.04195	0.03554
MSE5	0.00920	0.00865	0.01667	0.00949

The results are now different in detail than they were in Tables 2, 5, and 8. The groupwise averaged EBLUP MSE's are now not smaller in Group 2 than in Group 1, but are again monotonically decreasing as one moves from Group 2 (where PSU sample size was 11 to 25) to Groups 4 (where sample size

was 76 to 220) and 5. However, the general pattern of accuracy of the **nlmfit** and **glmfit** PSU response estimators is as before, with one glaring exception: as is probably due to the directions of bias in its parameter estimators the **nlmfit** MSE's (26) are now a bit better than those for **glmfit** in Groups 1 and 2, in all three Tables.

TABLE 12. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.2$, $h = exp$), VIA FORMULAS (19), (20) WITH PARAMETERS FITTED USING PREDICTOR VARIABLE X_i. Average estimated ω^* in **lmfitA** is 6.351; average estimated σ_*^2 in **lmfitB** is 0.0403. MSE's are now given as groupwise averages of squared relative errors (26).

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.78508	-1.72216	-1.91952	-1.91768
b1	0.40071	0.37215	0.28835	0.28672
sigP	0.18056	0.24719	0.18968	2.11283
MSE1	0.04369	0.04759	0.04478	0.04568
MSE2	0.04281	0.04697	0.04118	0.04186
MSE3	0.03324	0.03354	0.03358	0.03446
MSE4	0.02170	0.02027	0.02108	0.02074
MSE5	0.01129	0.01091	0.01067	0.01058

5.3 Simulations with More Than One Predictor

Several simulation experiments were also conducted with multiple predictor columns. First, we display tables 13 and 14 respectively for small-area estimators in nonsampled PSU's and for EBLUP's, based upon simulated data with 4 simulated independent predictor columns (the same for all 100 simulation iterations). All columns were simulated as $\mathcal{N}(0, 1.3^2)$ without downweighting in the largest PSU-sample group. In the simulation, the fixed-effect intercept was -1.5; the vector of fixed-effect coefficients was (0.5, 0.4, 0.2, 0.1); and the random PSU-effect standard deviation was 0.2.

TABLE 13. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR NON-SAMPLED PSU'S ($\sigma = 0.2$), VIA FORMULAS (17), (18) WITH FOUR UNCORRELATED PREDICTOR VARIABLES AND FIXED-EFFECT COEFFICIENTS (-1.5, 0.5, 0.4, 0.2, 0.1). Average ω^* in **lmfitA** is 6.500, σ_*^2 in **lmfitB** 0.0317.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.55696	-1.50065	-1.77326	-1.77025
b1	0.52600	0.49912	0.37963	0.37570
b2	0.41782	0.40066	0.30454	0.30138
b3	0.20943	0.19844	0.14788	0.14598
b4	0.10395	0.10067	0.07010	0.06989
sigP	0.13289	0.19739	0.15052	1.95765
Varb0	0.00029	NA	0.00019	0.00014
Varb1	0.00018	NA	0.00012	0.00008
Varb2	0.00017	NA	0.00012	0.00009
Varb3	0.00015	NA	0.00011	0.00008
Varb4	0.00015	NA	0.00012	0.00008
MSE1	0.00118	0.00113	0.00208	0.00203
MSE2	0.00118	0.00112	0.00191	0.00184
MSE3	0.00113	0.00107	0.00218	0.00211
MSE4	0.00112	0.00106	0.00248	0.00240
MSE5	0.00100	0.00097	0.00115	0.00114

These Tables show groupwise MSE's at least 70% larger for Groups 1–4 (corresponding to PSU sample-size of 75 or less) in the linear-model fits as opposed to the unit-level **nlmfit** and **glmfit**. This effect appears even stronger than in the single-predictor tables. The severe underestimation of PSU random-effect variance in **nlmfit** (which produced an estimate of 0.133 for the standard deviation which was actually 0.2) caused only slightly worse MSE performance for **nlmfit** versus **glmfit**. On the other hand, for the same reason, when squared **relative** errors were averaged groupwise (as in (26, not shown), the PSU response-rate estimators were a bit (5%-10%) better for **nlmfit** than for **glmfit** in all groups, but both gave consistently better relative-error MSE's by about 30% than the **lmfit** methods.

TABLE 14. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.2$), VIA FORMULAS (19), (20) WITH FOUR UNCORRELATED PREDICTOR VARIABLES AND FIXED-EFFECT COEFFICIENTS (-1.5, 0.5, 0.4, 0.2, 0.1). Average ω^* in **lmfitA** is 6.457, σ_*^2 in **lmfitB** 0.0318.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.55494	-1.49871	-1.77061	-1.76717
b1	0.52544	0.49919	0.37826	0.37403
b2	0.41856	0.40123	0.30253	0.29912
b3	0.20963	0.19906	0.14825	0.14593
b4	0.10668	0.10299	0.07161	0.07130
sigP	0.13384	0.19628	0.14996	1.93626
Varb0	0.00029	NA	0.00018	0.00013
Varb1	0.00017	NA	0.00011	0.00008
Varb2	0.00017	NA	0.00012	0.00009
Varb3	0.00015	NA	0.00011	0.00008
Varb4	0.00015	NA	0.00011	0.00008
MSE1	0.00123	0.00118	0.00206	0.00195
MSE2	0.00117	0.00112	0.00168	0.00151
MSE3	0.00094	0.00083	0.00178	0.00152
MSE4	0.00070	0.00059	0.00128	0.00085
MSE5	0.00036	0.00029	0.00041	0.00032

Next, we display results, respectively in Tables 15, 16, and 17 for small-area estimators in nonsampled PSU's and for EBLUP's, based upon simulated data with 3 Normal predictor columns simulated (and then fixed for all 100 simulation iterations) with means 0, variances 1.69, and all pairwise correlations 0.2, without downweighting in the largest PSU-sample group. The true fixed-effect intercept used in the simulation was -1.5, and the vector of fixed-effect coefficients was (0.5, 0.4, 0.2). The random PSU-effect standard deviation used was 0.15.

The pattern of greater accuracy (in the sense of smaller ordinary MSE) of the unit-level nonlinear models versus aggregated linear models continues to hold here. The underestimation of PSU random effect by **nlmfit** continues, perhaps with slightly greater severity than in the tables based on single-predictor data, but **glmfit** seems to be essentially free of bias in estimating parameters when (as here) the logistic mixed-effect unit-level model is properly specified. In the relative-error MSE counterpart of Table 15 (not

shown), due most likely to the smaller random-effect variance σ^2, the relative MSE's of the type (26) are somewhat lower for estimators based upon **glmfit** versus those based upon **nlmfit**. The analogous comparison for the case of sampled strata with EBLUP estimators is given in Table 16, which displays ordinary MSE's, and Table 17, which displays MSE's of the type (26). Again we find here that, except in Group 1 (the smallest PSU's, with samples of no more than 10), the relative-error MSE's (26) are smaller for method **glmfit** than for **nlmfit**.

TABLE 15. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR NON-SAMPLED PSU'S ($\sigma = 0.15$), VIA FORMULAS (17), (18) WITH THREE CORRELATED PREDICTOR VARIABLES AND FIXED-EFFECT COEFFICIENTS (-1.5, 0.5, 0.4, 0.2). Average estimated ω^* in **lmfitA** is 7.023; average estimated σ_*^2 in **lmfitB** is 0.0264.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.55515	-1.49918	-1.78838	-1.78489
b1	0.52329	0.49981	0.37192	0.36815
b2	0.41713	0.39973	0.29575	0.29264
b3	0.20792	0.19932	0.14794	0.14537
sigP	0.07234	0.14974	0.13107	1.98377
MSE1	0.00068	0.00063	0.00204	0.00195
MSE2	0.00070	0.00065	0.00260	0.00249
MSE3	0.00064	0.00060	0.00199	0.00189
MSE4	0.00058	0.00053	0.00083	0.00081
MSE5	0.00054	0.00050	0.00202	0.00190

TABLE 16. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.15$), VIA FORMULAS (19), (20) WITH THREE CORRELATED PREDICTOR VARIABLES AND FIXED-EFFECT COEFFICIENTS (-1.5, 0.5, 0.4, 0.2). Average estimated ω^* in **lmfitA** is 7.068; average estimated σ_*^2 in **lmfitB** is 0.0266.

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.55250	-1.49698	-1.78652	-1.78268
b1	0.52314	0.49985	0.37164	0.36757
b2	0.41767	0.40061	0.29542	0.29213
b3	0.20892	0.20026	0.15022	0.14692
sigP	0.06799	0.14386	0.12535	1.96749
MSE1	0.00071	0.00066	0.00205	0.00188
MSE2	0.00071	0.00066	0.00249	0.00223
MSE3	0.00061	0.00052	0.00180	0.00153
MSE4	0.00050	0.00038	0.00061	0.00052
MSE5	0.00036	0.00024	0.00092	0.00044

TABLE 17. SIMULATED PERFORMANCE OF SMALL-AREA EBLUP ESTIMATORS FOR SAMPLED PSU'S ($\sigma = 0.15$), VIA FORMULAS (19), (20) FOR ANOTHER SIMULATION WITH THE SAME PREDICTORS AS IN TABLE 16, BUT WITH MSE'S GIVEN AS GROUPWISE AVERAGES OF SQUARED RELATIVE ERRORS (26).

	nlmfit	glmfit	lmfitA	lmfitB
b0	-1.55250	-1.49698	-1.78652	-1.78268
b1	0.52314	0.49985	0.37164	0.36757
b2	0.41767	0.40061	0.29542	0.29213
b3	0.20892	0.20026	0.15022	0.14692
sigP	0.06799	0.14386	0.12535	1.96749
MSE1	0.01678	0.01693	0.02748	0.02953
MSE2	0.01601	0.01581	0.03029	0.03384
MSE3	0.01569	0.01433	0.02735	0.03056
MSE4	0.01516	0.01279	0.02145	0.02453
MSE5	0.01340	0.01031	0.01489	0.01609

6 Incorporating Sampling Weights

Since some PSU's and population subgroups are intentionally over-sampled in many sampling enterprises, such as SAIPE, it is important to be able to

incorporate the weighting scheme appropriately into model assumptions and estimators. We do this differently in the aggregated-linear and the unit-level logistic models. First, supposing that PSU's i are weighted by (slightly modified inverse inclusion probabilities) w_i, while the weights for selected individual units j within the i'th PSU are denoted $w_{j|i}$. We maintain the PSU-level aggregated model (1) as before, but it seems most natural to replace the unweighted sample average y_i^0/n_i in (2) by the weighted sample average

$$y_i^w \;=\; \sum_{j \in s_i} w_{j|i}\, y_{ji} \Big/ \sum_{j \in s_i} w_{j|i}$$

within PSU, where y_{ji} denotes the unit-level binary response. The resulting aggregated model, analogous to (2), is

$$y_i \equiv h^{-1}(y_i^w) \quad , \qquad y_i \;=\; Y_i \,+\, e_i \tag{27}$$

The method of parameter estimation is based upon the sample-weighted census likelihood idea of Binder (1983). Since PSU's are independent according to our superpopulation models (i.e., the cluster random effects apply only within PSU's), the weighted ML parameter estimators in the Fay-Herriot model are the maximizers over $(\gamma_0, \gamma_1, \sigma^2)$ of the Fay-Herriot model weighted log-likelihood

$$-\,\frac{1}{2}\,\sum_{i \in s}\, w_i \,\left\{ \log\left(\sigma^2 + \,\frac{v_e}{n_i}\right) \,+\, \frac{(y_i - \gamma_0 - \gamma_1 X_i)^2}{\sigma^2 + v_e/n_i} \right\}$$

Remark 1 *It seems a natural idea to try to re-allocate the weight w_i of sampled PSU's with 0 responses artificially to increase the weight of sampled PSU's with similar characteristics and sample sizes and low ($< 10\%$) rates of poor children. But the SAIPE researchers have apparently never tried this.*

By contrast, the unit-level binomial-logistic model with PSU random effect, does not change at all due to sample weighting, but the parameter estimators do. As indicated in the previous Section, we have two different methods for estimating parameters in the mixed logistic regression unit-level model, and we now describe the sample-weighted versions of these methods. The first method, described in Slud (1998), is to transform the mixed-logistic model to give (25). The underlying idea was the delta method (assuming both

N_i and n_i were large), and we now apply the same idea with y_i^0/n_i replaced by y_i^w (which we assume is self-weighting, i.e., consistently estimates $\vartheta_i = Y_i^0/N_i$). For

$$\vartheta_i \;=\; h(\eta_i + U_i) \;=\; \frac{e^{\eta_i + U_i}}{1 + e^{\eta_i + U_i}}$$

this yields

$$\left[\frac{\sum_{j \in s_i} w_{j|i}}{\sqrt{\sum_{j \in s_i} w_{j|i}^2}}\right] \, (y_i^w - h(\eta_i + U_i)) \;\approx\; \mathcal{N}\Big(0,\, h(\eta_i + U_i)(1 - h(\eta_i + U_i))\Big)$$

and

$$\arcsin\sqrt{y_i^w} \;\approx\; \arcsin\sqrt{h(\eta_i + U_i)} \;+\; \left[\frac{\sqrt{\sum_{j \in s_i} w_{j|i}^2}}{\sum_{j \in s_i} w_{j|i}}\right] \frac{\rho_i}{2} \qquad (28)$$

with standard normal errors ρ_i. As before, parameter estimation would proceed by maximizing the log-likelihood within this nonlinear regression model, but now the i'th sampled PSU would receive the weight w_i.

However, in the mixed nonlinear regression analysis (**nlmfit** columns), existing software does not easily accommodate maximization of a stratum-weighted sum of log-likelihoods over PSU's, so that the modified log-likelihood which we actually maximize in sample-weighted simulations below is:

$$\sum_{i \in s} \log \int \exp\Big(-\frac{w_i \, (\sum_{j \in s_i} w_{j|i})^2}{4 \sum_{j \in s_i} w_{j|i}^2} \, (\arcsin\sqrt{y_i^w} - \arcsin\sqrt{\eta_i + \sigma u})^2\Big) \, \phi(u) \, du$$

in place of the census-weighted log-likelihood expression, which would remove the PSU weights w_i from the exponents in the integrand and instead position them to multiply the logarithms of the integrals.

The second method used in estimating mixed logistic model parameters was described in Slud (1999, 2000). It involves an approximation to the log-likelihood which can be directly weighted both within and across PSU's. Apart from special modifications made in PSU's for which the response rate is either 0 or 1, the approximation used in the unweighted case for the i'th PSU log-likelihood is a constant plus

$$y_i^0 \, \log\frac{y_i^0}{n_i} \;+\; (n_i - y_i^0) \, \log\left(1 - \frac{y_i^0}{n_i}\right) \;-\; \frac{1}{2}\Big\{\log\Big(1 \;+\; y_i^0 \, (1 \;-\; \frac{y_i^0}{n_i}) \, \sigma^2\Big) +$$

$$+ \frac{y_i^0 \, (n_i - y_i^0) \, (logit(y_i^0/n_i) - \eta_i)^2}{n_i \, + \, y_i^0 \, (n_i - y_i^0) \, \sigma^2} \Big\} \tag{29}$$

Denote the total weight within the i'th PSU by

$$n_i^* \; \equiv \; \sum_{j \in s_i} w_{j|i}$$

Then the weighted log-likelihood contribution corresponding to (29), which would be weighted by a further factor w_i and summed over sampled PSU's i, is

$$\log \int \prod_{j \in s_i} \Big(h(\eta_i + \sigma u)^{y_{ji}} \, (1 - h(\eta_i + \sigma u))^{1 - y_{ji}} \Big)^{w_{j|i}} \, \phi(u) \, du$$

which (for PSU's such that $y_i^0 \neq 0, \, n_i$) is approximately

$$n_i^* \, (y_i^w \, \log y_i^w \, + \, (1 - y_i^w) \, \log(1 - y_i^w)) \, - \, \frac{1}{2} \Big\{ \log(1 \, + \, n_i^* \, y_i^w \, (1 - y_i^w)) \, +$$

$$+ \frac{n_i^* \, y_i^w \, (1 - y_i^w) \, (logit(y_i^w) - \eta_i)^2}{1 \, + \, n_i^* \, y_i^w \, (1 - y_i^w) \, \sigma^2} \Big\}$$

So the effect of including sampling weights in the approximate maximum likelihood analysis for mixed-effect logistic data is essentially that the response-rates y_i^0/n_i should be replaced by y_i^w and the sample-size by $\sum_{j \in s_i} w_{j|i}$ within the i'th PSU, and the log-likelihoods across PSU's should be combined combined using weights w_i.

The EBLUP's corresponding to the various estimation methods of this report require only simple modifications to take account of sample weights, after population-wide parameter estimators have been calculated:

- For **nlmfit** and **glmfit**, the EBLUP's are estimated, in terms of the plug-in parameter estimators $\hat{\gamma}_0, \, \hat{\gamma}_1, \, \hat{\sigma}^2$, using formula (20).
- For the Fay-Herriot **lmfit** methods, the only change required in formula (19) is to replace $y_k = \log(y_i^0/n_i)$ by $y_k = \log(y_i^w)$.

7 Summary & Directions for Further Research

This report compares two different types of estimation methodology for small area estimation in a two-level modeling framework allowing for linear or generalized-linear dependence of within-PSU response rates on predictor variables, and also allowing for a normally distributed random PSU effect. Since measured counts in a sampling framework like that of the SAIPE program (Citro et al. 1999) necessarily have many PSU's showing counts of 0, the Fay-Herriot mixed-linear-model methodology currently employed on log-counts works with data from which such 0-count sampled PSU's have been deleted (essentially, treated as though they were not sampled). We have here explored a small-area estimation methodology based upon a generalized-linear unit-level model with PSU random effects. Simulations clearly show that when a mixed-effect logistic model is correctly specified, small-area EBLUP estimators based upon the approximate maximum likelihood parameter estimates outperform — sometimes substantially — the Fay-Herriot methodology which uses only PSU's with non-zero counts. This is so even when the Fay-Herriot model is fitted with a known PSU variance component. At least in the case of a single strong predictor variable, the advantage of the generalized-linear unit-level methodology over the Fay-Herriot log-linear methodology is much lessened, but still present, when the latter model is allowed to include a quadratic predictor term. In settings where there are multiple predictors, whether or not these are highly correlated, our simulations suggest that the advantage for unit-level generalized-linear models becomes more pronounced than in the single-predictor case, and it is very unclear whether the inclusion of interaction and quadratic terms in the Fay-Herriot estimation approach could mitigate this disadvantage. Further research is needed on these issues.

Although the effect of incorporating sample weights has not been explored directly in the simulation studies reported here, both the Fay-Herriot SAIPE analysis method and the method based on logistic unit-level models are implemented to take account of sample weights, as described in Section 6, and sample weighting is not expected to change any of our conclusions materially.

One theoretical element of Small-Area Estimation which still requires further research is the prediction from data of the MSE's of small-area GLM-based EBLUP estimators. Although some ideas for this have been advanced

in the present report, the necessity of including some corrections for misspecified models has made this a more substantial problem, and the development of MSE formulas has seemed less important in the present study, where the main issue has been the comparison of the two available small-area estimation methodologies.

It remains in future research also to explore the performance of the mixed-logistic (or other generalized-linear unit-level mixed-effect) analysis and small-area estimation methodology on the real SAIPE data.

8 References

Bell, W. (1997) Models for county and state poverty estimates. Preprint, Census Bureau, Statistical Research Division, Nov. 17, 1997.

Bell, W. (1998) Borrowing information over time in small area estimation: thoughts with reference to the American Community Survey. Census Bureau preprint, Statistical Research Division. Presented at NAS workshop on ACS, September 11, 1998.

Binder, D. (1983) On the variances of asymptotically normal estimators from complex surveys. *Int. Statist. Rev.* **51**, 279-92.

Booth, J. and Hobert, J. (1998) Standard errors of prediction in generalized linear mixed models. *Jour. Amer. Statist. Assoc.* **93**, 262-72.

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

Datta, G. and Lahiri, P. (1997) A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. Preprint.

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

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

Prasad, N. and Rao, J.N.K. (1990) The estimation of mean squared errors of small area estimators. *Jour. Amer. Statist. Assoc.* **85**, 163-71.

Prasad, N. and Rao, J.N.K. (1998) On robust small area estimation using a simple random effects model. Preprint.

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

Slud, E. (2000) Accurate calculation and maximization of log-likelihood for mixed logistic regression. Preprint.

White, H. (1982) ML in misspecified models. *Econometrica*