Small Area Prediction for a Unit-Level Lognormal Model

Emily Berg[1] **and Hukum Chandra**[2]
[1]National Agricultural Statistics Service, emily.berg@nass.usda.gov
[2]Indian Agricultural Research Service, hchandra@iasri.res.in

Abstract

Many variables of interest in business and agricultural surveys have skewed distributions. An example from the National Agricultural Statistics Service is the acres harvested for a particular crop. We investigate small area estimation methods for skewed data under the assumption that a lognormal model is a reasonable approximation for the distribution of the response given covariates. Empirical Bayes (EB) predictors and estimators of the mean squared error of the predictors are proposed. In simulation studies, the EB predictors are more efficient than a direct estimator and more efficient than a synthetic estimator.

1. Introduction

Small area estimation is a class of applications where domain sample sizes are too small to support reliable direct estimators. A common approach to small area estimation is to use model-based estimators instead of design-based estimators. Efficiency gains are realized if the models incorporate information about variability among the units in the population or the structure of the domain means.

1.1 Linear Unit-Level Model

Battese, Harter, and Fuller (1988) use a linear mixed model to predict the area planted to corn and soybeans in Iowa counties. Crop areas are obtained for a sample of segments in each county through farmer interviews. Covariates, obtained from satellite data, are the number of pixels classified as corn and soybeans. The covariates are available for all sampled segments, and the population mean of the covariates is known for each county.

In the Battese, Harter, and Fuller (BHF) model,

$$y_{ij} = \lambda_0 + \boldsymbol{x}'_{ij}\boldsymbol{\lambda}_1 + v_i + \epsilon_{ij}, \tag{1}$$

where y_{ij} is response for unit j in county i, $\boldsymbol{x}_{ij}$ is the corresponding vector of covariates, and $(v_i, \epsilon_{ij}) \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\sigma_v^2, \sigma_\epsilon^2))$. The quantity to predict is

$$\bar{y}_{N_i} = \lambda_0 + \bar{\boldsymbol{x}}'_{N_i}\boldsymbol{\lambda}_1 + v_i + \bar{\epsilon}_{N_i}, \tag{2}$$

where $\bar{\boldsymbol{x}}_{N_i}$ is the mean of $\boldsymbol{x}_{ij}$ for the population of segments in county i.

Three types of predictors have been used for the linear setting. First, a synthetic estimator (or indirect estimator) is obtained by replacing a nonsampled unit with an estimator of its expected value. Rao (2003, Chapter 4) discusses several synthetic estimators. A synthetic estimator for area i is of the form

$$\hat{y}_i^{syn} = N_i^{-1}\left\{\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} \hat{\lambda}_{0,ols} + \boldsymbol{x}'_{ij}\hat{\boldsymbol{\lambda}}_{1,ols}\right\}, \tag{3}$$

where $(\hat{\lambda}_{0,ols}, \hat{\boldsymbol{\lambda}}'_{1,ols})'$ is the OLS estimate of $(\lambda_0, \boldsymbol{\lambda}'_1)$, $j = 1, \ldots, n_i$ indexes the sampled units, and $j = n_i + 1, \ldots, N_i$ indexes the nonsampled units for area i. A second predictor for a linear model is a model based direct estimator (Chandra and Chambers, 2009). The model based direct estimator is of the form

$$\hat{y}_i^{MBDE} = N_i^{-1}\sum_{j=1}^{n_i} w_{ij} y_{ij}. \tag{4}$$

The model-based direct estimator is a weighted sum of the sampled units in area i. The weights are defined in such a way that the weighted sum of all the units in the sample is the BLUP of the population total. A third type of predictor, which is widely-used for small area estimation, is an EBLUP. An EBLUP for the population mean is

$$\hat{y}_i^{EBLUP} = N_i^{-1}\left\{\sum_{j=1}^{n_i} y_{ij} + \sum_{j=n_i+1}^{N_i} \hat{\lambda}_0 + \boldsymbol{x}'_{ij}\hat{\boldsymbol{\lambda}}_1 + \hat{\gamma}_i(\bar{y}_{si} - \hat{\lambda}_0 - \bar{\boldsymbol{x}}'_{si}\hat{\boldsymbol{\lambda}}_1)\right\}, \tag{5}$$

where $\hat{\gamma}_i = (\hat{\sigma}_v^2 + n_i^{-1}\hat{\sigma}_\epsilon^2)^{-1}\hat{\sigma}_v^2$, $(\bar{y}_{si}, \bar{\boldsymbol{x}}'_{si}) = n_i^{-1}\sum_{j=1}^{n_i}(y_{ij}, \boldsymbol{x}'_{ij})$, and $(\hat{\lambda}_0, \hat{\boldsymbol{\lambda}}'_1, \hat{\sigma}_v^2, \hat{\sigma}_\epsilon^2)$ is the vector of REML estimators. The EBLUP for the BHF model shrinks a direct estimator toward a synthetic estimator by a factor that depends on the relative magnitudes of estimates of σ_v^2 and σ_ϵ^2. The weight assigned to the direct estimator decreases as the ratio of the between-area variance component to the within-area variance component decreases or the sample size decreases. See Rao (2003) for a discussion of small area prediction based on mixed models.

1.2 Lognormal Unit-Level Model

We consider a situation where the distribution of the response variable has a positive support, the variance is a function of the mean, and relationships between the mean response and the covariates are nonlinear. Because the assumptions of the linear model with normal errors are violated, linear predictors are inefficient. We consider the specific situation where units in the population are assumed to have lognormal distributions. We write the loglinear mixed model for the variable of interest, y_{ij}, as

$$\log(y_{ij}) := l_{ij} = \beta_0 + \boldsymbol{z}_{ij}\boldsymbol{\beta}_1 + u_i + e_{ij}, \tag{6}$$

where $(u_i, e_{ij}) \sim \mathrm{N}(\mathbf{0}, \mathrm{diag}(\sigma_u^2, \sigma_e^2))$, and $\boldsymbol{z}_{ij}$ is a vector of appropriately transformed covariates. For example, $z_{ijk} = \log(x_{ijk})$ for $k = 1, \ldots, p$. Let the observations $\{(y_{ij}, \boldsymbol{z}_{ij}) : i = 1, \ldots, D; j \in s_i\}$ be available, where s_i denotes the set of j in the sample for area i, and $|\, s_i \,| = n_i$. Let U_i denote the set of N_i indexes in the population for area i, and let $\bar{s}_i$ denote the set of j in area i that are not in the sample. Assume that $\boldsymbol{z}_{ij}$ is available for the population of N_i values in area i, and let $\{y_{ij}; i = 1, \ldots, D, j \in s_i\} \cup \{\boldsymbol{z}_{ij} : i = 1, \ldots, D, j \in U_i\}$ be the available data. Let $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}'_1, \sigma_u^2, \sigma_e^2)'$ be the vector of model parameters, and let $\hat{\boldsymbol{\theta}} = (\hat{\beta}_0, \hat{\boldsymbol{\beta}}'_1, \hat{\sigma}_u^2, \hat{\sigma}_e^2)'$ be the REML estimator of $\boldsymbol{\theta}$. The quantity of interest is the area mean,

$$\bar{y}_{N_i} = \frac{1}{N_i}\sum_{j \in U_i} y_{ij}. \tag{7}$$

In this paper, we compare three predictors of $\bar{y}_{N_i}$.

The first predictor, based on an estimator of Karlberg (2000), is analogous to the synthetic estimator for the linear model. Karlberg (2000) considers a situation where the objective is to estimate a single finite population mean (equivalently, a total). He assumes that the units in the population are realizations from a lognormal distribution and that a covariate is observed for all units in the population. We adapt the Karlberg (2000) procedure to the small area context.

The second predictor is a model-based direct estimator developed in Chandra and Chambers (2011). The Chandra and Chambers (2011) estimator is a weighted sum of the sampled units, where the weights are defined to give the minimum mean squared error linear predictor of the population mean if the parameters of the lognormal distribution were known.

The third predictor is an empirical Bayes (EB) predictor. The empirical Bayes (EB) method is a general approach to small area estimation that is appropriate for a broad class of linear and nonlinear models. An EB predictor for squared error loss is an estimator of the conditional expectation of the small area parameter given the observed data and the underlying model parameters. For a linear mixed model with normal errors, an EBLUP is an EB predictor (See Rao, 2003 Section 9.1).

Slud and Maiti (2006) construct an EB predictor for a small area mean under an assumption that the area-level direct estimators have lognormal distributions. They derive a closed form expression for the EB predictor and give an approximately unbiased MSE estimator. They use the EB predictor to obtain estimates of county-level rates of school-aged children in poverty using data from the US Census Bureau's "Small area income and poverty estimation" project. The EB predictor proposed in this paper differs from the Slud and Maiti (2006) predictor because we work with unit-level data instead of area-level data.

In Section 2, we discuss the estimator based on Karlberg (2000) and the Chandra and Chambers (2011) estimator in more detail. We also propose an empirical Bayes predictor for the lognormal model in Section 2. In Section 3, we compare the predictors defined in Section 2 through simulations. We conclude in Section 4 with a summary and a discussion of areas for future work.

2. Predictors for the Lognormal Model

2.1 A Type of Synthetic Estimator

Karlberg (2000) addresses a situation where the quantity of interest is a mean for a single area instead of many small area means. She derives a predictor under an assumption that the units in the population are realizations from the model (6) with $\sigma_u^2 = 0$. In the simulations in Karlberg (2000), the predictor that she suggests is more efficient than a regression estimator.

We modify the approach of Karlberg (2000) to define a type of synthetic estimator for the lognormal model. The resulting estimator of $\bar{y}_{N_i}$ is,

$$\hat{y}_{N_i}^{karlberg} = f_i \bar{y}_{n_i} + (1 - f_i) \left(\frac{1}{N_i - n_i} \right) (\sum_{j \in \bar{s}_i} \hat{y}_{ij}^{karlberg}), \tag{8}$$

where $\hat{y}_{ij}^{karlberg}$ is the estimator of $E[y_{ij} \,|\, \boldsymbol{\theta}, \boldsymbol{z}_{ij}]$ defined in (10) below. By properties of the normal moment generating function,

$$E[y_{ij} \,|\, \boldsymbol{\theta}, \boldsymbol{z}_{ij}] = \exp\{\beta_0 + \boldsymbol{z}_{ij}' \boldsymbol{\beta}_1 + 0.5(\sigma_u^2 + \sigma_e^2)\}. \tag{9}$$

Because $E[y_{ij} \,|\, \hat{\boldsymbol{\theta}}, \boldsymbol{z}_{ij}]$ is a nonlinear function of the REML estimator of $\boldsymbol{\theta}$, $E[y_{ij} \,|\, \hat{\boldsymbol{\theta}}, \boldsymbol{z}_{ij}]$ is a biased estimator of (9). To correct for the bias, we use the method of Karlberg (2000) and define the estimator,

$$\hat{y}_{ij}^{karlberg} \quad = \quad (\hat{c}_{ij}^{karlberg})^{-1} E[y_{ij} \,|\, \hat{\boldsymbol{\theta}}, \boldsymbol{z}_{ij}], \tag{10}$$

where

$$\hat{c}_{ij}^{karlberg} \quad = \quad \exp\left\{ 0.5((1, \boldsymbol{z}_{ij}') \hat{V}\{\hat{\boldsymbol{\beta}}\}(1, \boldsymbol{z}_{ij}')' + 0.25 \hat{V}\{\hat{\sigma}_u^2 + \hat{\sigma}_e^2\}) \right\}, \tag{11}$$

and $\hat{V}\{\hat{\boldsymbol{\beta}}\}$ and $\hat{V}\{\hat{\sigma}_u^2 + \hat{\sigma}_e^2\}$ are estimates of the variances of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\sigma}}$ obtained from the inverse information matrix. (See Rao, 2003, pg. 139.)

2.2 A Model-Based Direct Estimator

Chandra and Chambers (2011) derive an estimator of the form $\sum_{j \in s_i} \hat{w}_{ij} y_{ij}$, where $\hat{w}_{ij}$ is an estimator of the weight that gives the BLUP of the population mean if the parameters of the model (6) are known. To derive the predictor, Chandra and Chambers (2011), use the approximation,

$$E[y_{ij}] \quad \approx \quad \gamma_o + \gamma_1 \hat{y}_{ij}^{karlberg}, \tag{12}$$

and

$$C\{y_{ij}, y_{ik}\} \quad \approx \quad \hat{y}_{ij}^{karlberg} \hat{y}_{ik}^{karlberg} \left\{ \exp(\hat{\sigma}_u^2) - 1 + \exp(\hat{\sigma}_u^2)(\exp(\hat{\sigma}_e^2) - 1) I[j = k] \right\}, \tag{13}$$

where $\hat{y}_{ij}^{karlberg}$ is defined in (10), and the covariance between two units from different areas is zero. The approximations for the first and second moments in (12) and (13) follow from the moment generating function of a normal distribution. The vector form of (12) is

$$E[\boldsymbol{y}_N] \quad \approx \quad \hat{\boldsymbol{X}}_U \boldsymbol{\gamma},$$

and

$$C\{\boldsymbol{y}_N, \boldsymbol{y}_N'\} \quad \approx \quad \begin{pmatrix} \hat{\boldsymbol{V}}_{ss} & \hat{\boldsymbol{V}}_{ss} \\ \hat{\boldsymbol{V}}_{sr} & \hat{\boldsymbol{V}}_{rr} \end{pmatrix},$$

where $\boldsymbol{\gamma} = (\gamma_o, \gamma_1)'$,

$$\hat{\boldsymbol{X}}_U \quad = \quad (\hat{\boldsymbol{X}}_s', \hat{\boldsymbol{X}}_r')' = \left((\boldsymbol{1}_s', \boldsymbol{1}_r')', (\hat{\boldsymbol{y}}_s^{karlberg}, \hat{\boldsymbol{y}}_r^{karlberg})'\right),$$

$\boldsymbol{y}_N = (\boldsymbol{y}_s', \boldsymbol{y}_r')'$, $\boldsymbol{y}_s$ and $\boldsymbol{y}_r$ are the vectors of sampled and nonsampled units, respectively, and $\hat{\boldsymbol{y}}_s^{karlberg}$ and $\hat{\boldsymbol{y}}_r^{karlberg}$ are the vectors containing $\hat{y}_{ij}^{karlberg}$ for the sampled and nonsampled units. The elements of the covariance matrices $\hat{\boldsymbol{V}}_{ss}$ and $\hat{\boldsymbol{V}}_{sr}$ are defined in (13). If one treats the $\hat{y}_{ij}$, $\hat{\sigma}_u^2$, and $\hat{\sigma}_e^2$ in (12) and (13) as fixed values, then (12) is a linear model for the mean of y_{ij}, and the BLUP of $N^{-1}\sum_{i=1}^{D}\sum_{j=1}^{N_i} y_{ij}$ is $N^{-1}\hat{\boldsymbol{w}}'\boldsymbol{y}_s$, where

$$\hat{\boldsymbol{w}} = \boldsymbol{1}_s + \hat{\boldsymbol{H}}_s'(\hat{\boldsymbol{X}}_U'\boldsymbol{1}_U - \hat{\boldsymbol{X}}_s'\boldsymbol{1}_s) + (\boldsymbol{I}_s - \hat{\boldsymbol{H}}_s'\hat{\boldsymbol{X}}_s')\hat{\boldsymbol{V}}_{ss}^{-1}\hat{\boldsymbol{V}}_{sr}\boldsymbol{1}_r,$$

and

$$\hat{\boldsymbol{H}}_s = (\hat{\boldsymbol{X}}_s'\hat{\boldsymbol{V}}_{ss}^{-1}\hat{\boldsymbol{X}}_s)^{-1}\hat{\boldsymbol{X}}_s'\hat{\boldsymbol{V}}_{ss}^{-1}.$$

The model-based direct estimator defined in Chandra and Chambers (2011) is

$$\hat{y}_{N_i}^{TrMBDE} = N_i^{-1}\sum_{j \in s_i} \hat{w}_{ij} y_{ij}, \tag{14}$$

where $\hat{w}_{ij}$ is the element of $\hat{\boldsymbol{w}}$ associated with unit (i, j). The "Tr" in the label "TrMBDE" stands for "transformed" and is used to distinguish the estimator (14) from a model-based direct estimator for a linear model.

2.3 An Empirical Bayes Predictor

The minimum mean squared error (MSE) predictor of $\bar{y}_{N_i}$ is $E[\bar{y}_{N_i} \,|\, (\boldsymbol{y}, \boldsymbol{z})]$, where $(\boldsymbol{y}, \boldsymbol{z}) = \{y_{ij}; i = 1, \ldots, D, j \in s_i\} \cup \{\boldsymbol{z}_{ij} : i = 1, \ldots, D, j \in U_i\}$. (See for example, Rao 2003, Chapter 9.) Under the assumptions of the lognormal model (6), a closed form expression for the minimum MSE predictor is

$$\bar{y}_{N_i}^{MMSE}(\boldsymbol{\theta}) = \frac{1}{N_i}[\sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} y_{ij}^{MMSE}(\boldsymbol{\theta})], \tag{15}$$

where

$$y_{ij}^{MMSE}(\boldsymbol{\theta}) = \exp\{\beta_0 + \boldsymbol{z}_{ij}'\boldsymbol{\beta}_1 + \gamma_i(\bar{l}_{is} - \beta_0 - \bar{\boldsymbol{z}}_{is}'\boldsymbol{\beta}_1) + 0.5\sigma_e^2(\gamma_i n_i^{-1} + 1)\}, \tag{16}$$

$(\bar{l}_{is}, \bar{\boldsymbol{z}}_{is}') = n_i^{-1}\sum_{j \in s_i}(l_{ij}, \boldsymbol{z}_{ij}')$, and $\gamma_i = \sigma_u^2(\sigma_u^2 + n_i^{-1}\sigma_e^2)^{-1}$. The form of the minimum MSE predictor in (15) follows from the moment generating function of the lognormal distribution and the property that

$$(u_i, e_{ij}) \,|\, (\boldsymbol{y}, \boldsymbol{z}) \sim \mathrm{N}\{[\gamma_i(\bar{l}_{is} - \beta_0 - \bar{\boldsymbol{z}}_{is}'\boldsymbol{\beta}_1), 0], \mathrm{diag}(\gamma_i n_i^{-1}\sigma_e^2, \sigma_e^2)\}$$

for $j \notin s_i$. A detailed derivation of (15) is given in the Appendix.

The minimum MSE predictor (15) is not possible to compute unless the true $\boldsymbol{\theta}$ is known. We replace the true $\boldsymbol{\theta}$ in (15) with the REML estimator to obtain the empirical Bayes (EB) predictor,

$$\hat{y}_{N_i}^{EB} = \bar{y}_{N_i}^{MMSE}(\hat{\boldsymbol{\theta}}) = \frac{1}{N_i}\left[\sum_{j\in s_i} y_{ij} + \sum_{j\in\bar{s}_i}\hat{y}_{ij}^{EB}\right], \tag{17}$$

where

$$\begin{aligned} \hat{y}_{ij}^{EB} &= y_{ij}^{MMSE}(\hat{\boldsymbol{\theta}}) \\ &= \exp\left\{\hat{\beta}_0 + \boldsymbol{z}'_{ij}\hat{\boldsymbol{\beta}}_1 + \hat{\gamma}_i(\bar{l}_{is} - \hat{\beta}_0 - \bar{\boldsymbol{z}}'_{is}\hat{\boldsymbol{\beta}}_1) + 0.5\hat{\sigma}_e^2(\hat{\gamma}_i n_i^{-1} + 1)\right\}. \end{aligned} \tag{18}$$

3. Simulations

We compare the EB predictors to the synthetic estimator and the direct estimator defined in Section 2 and evaluate the performance of the MSE estimator through simulations. The model for the simulations is (6) with a one-dimensional covariate z_{ij}, where $z_{ij} \sim N(\mu_z, \sigma_z^2)$. We generate data for 30 areas ($D = 30$) and set $(N_i, n_i) = (133, 5)$ for 15 of the areas and $(N_i, n_i) = (533, 20)$ for the other 15 areas so that $(N, n) = (9990, 375)$.

We pick the parameters in Table 1 so that the mean and variance of y_{ij} is approximately equal to the mean and variance of the number of chickens per segment in a 1960's United States Department of Agriculture area survey discussed in Fuller (1991). We alter the variance components to study the effects of varying the relative magnitudes of σ_u^2, σ_e^2, and σ_z^2 on the properties of the predictors. For the first two simulations, $\sigma_z = 1.58$ and the ratio $\sigma_u^2\sigma_e^{-2}$ is 0.51 or 0.15. For the second two simulations, we set $\sigma_z = 1.24$ and increase σ_u^2 and σ_e^2 so that $\sigma_u^2\sigma_e^{-2}$ is 0.45 or 0.15.

Parameter Configurations

Set	σ_z	$\sigma_u^2\sigma_e^{-2}$	σ_u	$E[Y]$	$V\{Y\}$
1	1.58	0.51	0.55	16	4493
2	1.58	0.16	0.35	16	4493
3	1.24	0.45	0.71	15.5	4006
4	1.24	0.15	0.46	15.5	5006

$(\mu_z, \beta_0, \beta_1) = (3.253, -1.62, 0.9)$

Table 1: Parameter configurations for simulations

We use a Monte Carlo (MC) sample size of 2000. For each MC sample, we generate a new set of z_{ij} from a normal distribution and select a stratified simple random sample where the areas are the strata. We compute the following predictors of $\bar{y}_{N_i}$:

1. Karlberg - the Karlberg (2000) estimator defined in (8)
2. TrMBDE - the Chandra and Chambers (2011) model-based direct estimator defined in (14)
3. EB - the empirical Bayes predictor defined in (17)

3.1 Empirical Properties of Small Area Predictors

We define the MC relative bias of predictor $\hat{y}_{N_i}$ for area i by,

$$\text{RB}_i = \frac{E_{MC}[\hat{y}_{N_i} - \bar{y}_{N_i}]}{E[\bar{y}_{N_i}]}, \tag{19}$$

where $E[\bar{y}_{N_i}]$ is given in the fifth column of Table 1 and $E_{MC}[\cdot]$ denotes the MC mean (the average of the 2000 samples). Table 2 contains the average MC relative biases of the alternative predictors, where the average is across areas with the same sample size. Estimates of the MC standard errors are in parentheses. For the two parameter sets with $\sigma_z = 1.58$, the MC relative biases of all predictors are small relative to the MC standard errors. The average MC relative biases of the TrMBDE and Karlberg predictors are essentially zero for all parameter configurations and sample sizes. For the two parameter sets with $\sigma_z = 1.24$, the EB predictor has a positive MC relative bias, and for fixed n_i, the average MC relative bias of the EB predictor is larger for $\sigma_u^2\sigma_e^{-2} = 0.15$ than for $\sigma_u^2\sigma_e^{-2} = 0.45$. For each parameter set with $\sigma_z = 1.24$, the average MC relative bias of the EB predictor is smaller for $n_i = 20$ than for $n_i = 5$. We conjecture that the average MC relative bias of the EB predictor increases as σ_z^2 decreases and (σ_u^2, σ_e^2) increase because the bias of the EB predictor increases as the variances of the REML estimators increase. For the parameters and sample sizes considered here, the average MC relative bias of the EB predictor is less than 3% of the average MC RMSE.

		Average Relative Biases (%) for $n_i = 5$			Average Relative Biases (%) for $n_i = 20$		
$\sigma_u^2\sigma_e^{-2}$	σ_z	TrMBDE	Karlberg	EB	TrMBDE	Karlberg	EB
0.51	1.58	-0.40	0.03	-0.17	-0.09	-0.22	0.42
		(0.33)	(0.36)	(0.24)	(0.17)	(0.33)	(0.14)
0.16	1.58	-0.06	-0.48	0.21	-0.29	-0.21	0.31
		(0.37)	(0.26)	(0.23)	(0.18)	(0.22)	(0.14)
0.45	1.24	0.64	-0.07	1.31	-0.43	0.01	0.44
		(0.55)	(0.48)	(0.33)	(0.27)	(0.46)	(0.19)
0.15	1.24	0.49	0.23	1.44	-0.09	0.03	0.87
		(0.61)	(0.33)	(0.29)	(0.30)	(0.29)	(0.19)

Table 2: Average MC relative biases (RB$_i$) of alternative predictors of $\bar{y}_{N_i}$. MC standard errors are in parentheses.

Table 3 contains the average ratios of the MC MSE's of the alternative predictors to the MC MSE of the EB predictor. The relative MC MSE of predictor $\hat{y}_{N_i}$ for area i is

$$\text{RelMSE}_i = \frac{MSE_{MC}(\hat{y}_{N_i})}{MSE_{MC}(\hat{y}_{N_i}^{EB})}, \tag{20}$$

and Table 3 contains averages of RelMSE$_i$ across areas with the same sample size. For fixed (n_i, σ_z^2), the relative MSE of the model-based direct estimator (TrMBDE) is larger for, $\sigma_u^2\sigma_e^{-2} = 0.16$ or 0.15 than for $\sigma_u^2\sigma_e^{-2} = 0.51$ or 0.45. The relative MSE of the TrMBDE predictor is larger for $n_i = 5$ than for $n_i = 20$ for each parameter configuration. For a fixed parameter set, the relative MSE's of the Karlberg predictor increase as the sample size increases. For a fixed sample size and value of σ_z, the relative MSE of the Karlberg predictor increases as the ratio of σ_u^2 to σ_e^2 increases.

The relationships between the relative MSE's of the TrMBDE and Karlberg predictors to the values of the variance parameters and sample sizes are not surprising if we consider an analogy with the linear model of Section 1.1. For the BHF model of Section 1.1, the ratio of the MSE of a mixed-model predictor to the variance of the sample mean (a simple direct estimator) is approximately $\sigma_v^2(\sigma_v^2 + \sigma_\epsilon^2 n_i^{-1})^{-1}$, and the ratio of the MSE of a mixed-model predictor to the MSE of a synthetic estimator is approximately $1 - \sigma_v^2(\sigma_v^2 + \sigma_\epsilon^2 n_i^{-1})^{-1}$. Because the TrMBDE predictor is a version of a direct estimator, we expect the ratio of the MSE of the TrMBDE predictor to the MSE of the EB predictor to decrease as the sample size increases and the ratio $\sigma_v^2\sigma_\epsilon^{-2}$ increases. Because the Karlberg predictor is a type of synthetic estimator, we expect the opposite pattern in the relative MSE for the Karlberg predictor.

		Average RelMSE$_i$ for $n_i = 5$		Average RelMSE$_i$ for $n_i = 20$	
$\sigma_u^2\sigma_e^{-2}$	σ_z	TrMBDE	Karlberg	TrMBDE	Karlberg
0.51	1.58	1.809	2.215	1.478	6.052
0.16	1.58	2.781	1.317	1.766	2.499
0.45	1.24	2.747	2.163	2.115	5.930
0.15	1.24	4.481	1.301	2.653	2.533

Table 3: Average relative MC MSE's (RelMSE$_i$) of the alternative predictors of $\bar{y}_{N_i}$ to the EB predictor.

4. Concluding Remarks

We compared three predictors of a small area mean for a skewed response variable. The model-based predictor, TrMBDE (Chandra and Chambers, 2011), is a version of a direct estimator. The predictor based on Karlberg (2000) is a type of synthetic estimator (Rao, 2003, Chapter 4) because the predictor of a non-sampled unit is an estimator of $E[y_{ij} \mid \boldsymbol{z}_{ij}]$ and does not directly involve any of the sampled units. The EB predictor is an estimator of the minimum MSE predictor of the small area mean.

Because the Karlberg (2000) predictor is a type of synthetic estimator, the relative efficiency of the Karlberg (2000) predictor improves as n_i decreases and $\sigma_u^2\sigma_e^{-2}$ decreases. The efficiency of the Chandra and Chambers (2011) direct estimator improves as n_i increases and as $\sigma_u^2\sigma_e^{-2}$ increases. As discussed in Section 3, these patterns are expected by analogy with a linear model. Because the EB predictor is a nonlinear function of the REML estimate of $\boldsymbol{\theta}$, the EB predictor is biased. The MC bias of the EB predictor is typically less than 3% of the MC RMSE for the parameters and sample sizes considered in our simulation study. An examination of bias-corrected EB predictors is an area of current research.

We did not discuss MSE estimation in this paper. We obtained a closed-form MSE estimator that accounts for the estimation of the unknown parameters. In simulations not discussed here, the MC relative bias of the MSE estimator is less than 11%, and the empirical coverages of nominal 95% prediction intervals are between 94% and 96%. A more thorough evaluation of the MSE estimator is a topic for future research.

For the simulations in Section 3, the model underlying the minimum MSE predictor is true, so it is not surprising that the EB predictor has a smaller MSE than the other predictors. An evaluation of the properties of the procedures when the model is misspecified and a study of the design properties of the model-based predictors are areas of current research.

In this study, we compared an EB predictor to a model-based direct estimator and a synthetic estimator. One can construct other shrinkage estimators for skewed data. An example is the area-level predictor of Slud and Maiti (2006) discussed in the Introduction. Current work includes a comparison of the EB predictors defined in Section 2.3 to other shrinkage estimators for skewed data.

In our analysis of the simulation study, we observed that the predictors, the mean squared errors of the predictors, and the mean squared error estimators are skewed and have large variances. This suggests that loss functions other than squared error loss may be more appropriate for highly skewed data. Future work may involve an evaluation of the proposed predictors for different loss functions or derivations of predictors that are optimal with respect to different loss functions.

We restricted our attention to a situation where the sample design within areas is simple random sampling. Modifications for complex sampling or nonresponse are potential areas for future work.

Appendix: Derivation of Minimum MSE Predictor

The minimum MSE predictor is

$$E[\bar{y}_{N_i} \,|\, (\boldsymbol{y}, \boldsymbol{z})] = \frac{1}{N_i}\left[\sum_{j \in s_i} y_{ij} + \sum_{j \in \bar{s}_i} E[y_{ij} \,|\, (\boldsymbol{y}, \boldsymbol{z})]\right]. \tag{21}$$

Under the model (6),

$$y_{ij} = \exp(\beta_0 + \boldsymbol{z}'_{ij}\boldsymbol{\beta}_1)\exp(u_i + e_{ij}),$$

and the conditional expectation of y_{ij} given the available data is

$$E_\theta[y_{ij} \,|\, (\boldsymbol{y}, \boldsymbol{z})] = \exp(\beta_0 + \boldsymbol{z}'_{ij}\boldsymbol{\beta}_1) E[\exp(u_i + e_{ij}) \,|\, (\boldsymbol{y}, \boldsymbol{x})]. \tag{22}$$

By properties of the normal distribution, for $j \in \bar{s}_i$,

$$(u_i, e_{ij}) \,|\, (\boldsymbol{y}, \boldsymbol{z}) \sim \mathrm{N}\{[\gamma_i(\bar{l}_{is} - \beta_0 - \bar{\boldsymbol{z}}'_{is}\boldsymbol{\beta}_1), 0], \mathrm{diag}(\gamma_i n_i^{-1}\sigma_e^2, \sigma_e^2)\}. \tag{23}$$

By (23) and the moment generating function of the lognormal distribution,

$$E[\exp(u_i + e_{ij}) \,|\, (\boldsymbol{y}, \boldsymbol{z})] = \exp\{\gamma_i(\bar{l}_{is} - \beta_0 - \bar{\boldsymbol{z}}'_{is}\boldsymbol{\beta}_1) + 0.5(\gamma_i n_i^{-1}\sigma_e^2 + \sigma_e^2)\} \tag{24}$$

for $j \in \bar{s}_i$. Together, (24), (22), and (21) justify (15).

Acknowledgements

We thank Ray L. Chambers and Wayne A. Fuller.

References

Battese, G.E., Harter, R.M., Fuller, W.A. (1988). An error component model for prediction of county crop areas using survey and satellite data. *Journal of the American Statistical Association*, 83, 28 - 36.

Chandra, H., and Chambers, R. (2009). Multipurpose weighting for small area estimation. *Journal of Official Statistics*, 25(3), 379-395.

Chandra, H. and Chambers, R. (2011). Small area estimation under transformation to linearity. *Survey Methodology*, 37, 39 - 51.

Fuller, W.A. (1991). Simple Estimators for the Mean of Skewed Populations. *Statistica Sinica*, 1, 137- 158.

Karlberg, F. (2000). Population total prediction under a lognormal superpopulation model. *Metron*, LVIII, 53-80.

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

Slud, E.V. and Maiti, T. (2006). MSE estimation in transformed Fay- Herriot models. *Journal of the Royal Statistical Society*, Series B, 68, 239- 257.