Small Area Estimation Combining Information from Several Sources

Jae-kwang Kim[*] Seunghwan Park[†] Seo-young Kim[‡]

Abstract
An area-level model approach to combining information from several sources is considered in the context of small area estimation. At each small area, several estimates are computed and linked through a system of structural error models. The best linear unbiased predictor of the small area parameters are computed by the general least squares method. Parameters in the structural error models are estimated using the theory of measurement error models. Estimation of mean squared errors is also discussed. The proposed method is applied to the real problem of labor force survey in Korea.

Key Words: Area-level model; Auxiliary information; Measurement error models; Structural error model; Survey integration.

1. Introduction

Combining information from different source is an important problem in statistics. In survey sampling, the source of information can come from a probability sampling with direct measurement, from another probability sampling with indirect measurement (such as self-reported health status), or from some auxiliary area-level information. We consider an area-level model approach to small area estimation when there are several source of auxiliary information. Pfeffermann (2002) and Rao (2003) provided thorough reviews of methods used in small area estimation. Elliot and Davis (2005) used dual-frame estimation methods to combine data from two surveys for estimating cancer risk factors in small areas. Ybarra and Lohr (2008) considered the small area estimation problem when the area-level auxiliary information has some measurement errors. Merkouris (2010) discussed the small area estimation by combining information from multiple surveys. Kim and Rao (2012) considered a design-based approach to combining information from two independent surveys.

To describe the setup, suppose that the finite population consists of H subpopulations, denoted by $U_1, \cdots, U_H$, and we are interested in estimating the subpopulation totals $X_h = \sum_{i \in U_h} x_i$ for area h. We assume that there is a survey that measures x_i from the sample but its sample size is not large enough to obtain estimates for X_h with reasonable accuracy. The main survey will be called survey A and let $\hat{X}_{h,a}$ denote a design-consistent estimator of X_h obtained from survey A. Often, we compute $\hat{X}_{h,a} = \sum_{i \in A_h} w_{ia} x_i$, where A_h is the set of sample A for subpopulation h and w_{ia} is the weight of unit i in sample A.

In addition to the main survey, suppose that there is another survey, called survey B, that measures an rough estimate for x_{it}. Let y_{1i} be the measurement taken from survey B. We may assume that y_{1i} is a rough measurement of x_i with some level of measurement error. Thus, we may assume

$$y_{1i} = \beta_0 + \beta_1 x_i + e_{1i} \tag{1}$$

[*] Department of Statistics, Iowa State University, Ames, Iowa, 50011, U.S.A.
[†] Department of Statistics, Seoul National University, Seoul, 151-747, Korea
[‡] Statistical Research Institute, Statistics Korea, Daejon, 302-847, Korea

for some (β_0, β_1), where e_{1i} is a random variable that is distributed with mean zero and variance σ_{e1}^2. The linear regression assumption or equal variance assumptions can be relaxed later. If $(\beta_0, \beta_1) = (0, 1)$, then model (1) means that there is no measurement bias. From survey B, we can obtain another estimator $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$, where w_{ib} is the weight of unit i in the t-th survey for sample B and B_h is the set of sample B for subpopulation h. Model (1) can be used to combine information from the two surveys.

Finally, another source of information can be the Census information. Census information does not suffer from coverage error or sampling error. But, it may have measurement errors and it does not provide updated information for each month or year. Let y_{2i} be the measurement for unit i from the Census. The subpopulation total $Y_{2h} = \sum_{i \in C_h} y_{2i}$ is available when C_h is the set of Census C for subpopulation h.

Table 1 summarize the major source of information that we can consider into small area estimation.

Table 1: Available information for small area estimation.

Data	Observation	Area level estimate	Discussion
Survey A	direct obs. (x_i)	$\hat{X}_h, \hat{V}(\hat{X}_h)$	Sampling error (large)
Survey B	aux. obs. (y_{1i})	$\hat{Y}_{1h}, \hat{V}(\hat{Y}_{1h})$	Measurement error Sampling error (small)
Census	aux. obs. (y_{2i})	Y_{2h}	Measurement error No updated information

In this paper, we consider an area-level model approach for small area estimation combining all available information. The proposed approach is based on the measurement error models, where the sampling errors of the direct estimators are treated as measurement errors, and all the other auxiliary information are combined through a set of linking models. The proposed approach is applied to the real problem of labor force survey in Korea.

The paper is organized as follows. In Section 2, the basic setup is introduced and the small area estimation problem is viewed as a measurement error model prediction problem. In Section 3, parameter estimation for the area level small area model is discussed. In Section 4, mean square estimation is briefly discussed. In Section 5, the proposed method is applied to the labor force survey data in Korea. Concluding remarks are made in Section 6.

2. Basic Theory

In this section, we first introduce the basic theory for combining the information for small area estimation. We first consider the simple case of combining two surveys. Assume that there are two surveys, survey A and survey B, obtained from separate probability sampling designs. The two surveys are not necessarily independent. From survey A, we obtain a design unbiased estimator $\hat{X}_h = \sum_{i \in A_h} w_{ia} x_i$ and its variance estimator $\hat{V}(\hat{X}_h)$. From survey B, we obtain a design-model unbiased estimator $\hat{Y}_{1h} = \sum_{i \in B_h} w_{ib} y_{1i}$ of $Y_{1h} = \sum_{i \in U_h} y_{1i}$. The sampling error of $(\hat{X}_h, \hat{Y}_{1h})$

can be expressed by the *sampling error model*

$$\begin{pmatrix} \hat{X}_h \\ \hat{Y}_{1h} \end{pmatrix} = \begin{pmatrix} X_h \\ Y_{1h} \end{pmatrix} + \begin{pmatrix} N_h a_h \\ N_h b_h \end{pmatrix} \tag{2}$$

and a_h and b_h represent the sampling errors associated with $\hat{X}_h/N_h$ and $\hat{Y}_{1h}/N_h$, respectively. Our parameter of interest is the population total X_h of x_i in area h.

If there is no way to obtain the observation pairs (x_i, y_{1i}) from the two surveys, then we derive the following area level model from (1):

$$Y_{1h} = N_h\beta_0 + \beta_1 X_h + \tilde{e}_{1h}, \tag{3}$$

where $(N_h, X_h, Y_{1h}, \tilde{e}_{1h}) = \sum_{i \in U_h}(1, x_i, y_{1i}, e_{1i})$. We can express (3) in term of population mean

$$\bar{Y}_{1h} = \beta_0 + \bar{X}_h\beta_1 + \bar{e}_{1h}, \tag{4}$$

where $(\bar{X}_h, \bar{Y}_{1h}, \bar{e}_{1h}) = N_h^{-1}\sum_{i \in U_h}(x_i, y_{1i}, e_{1i})$. If we use a nested error model

$$e_{1hi} = \epsilon_h + u_{hi} \tag{5}$$

where $\epsilon_h \sim (0, \sigma_e^2)$ and $u_{hi} \sim (0, \sigma_u^2)$, then $\bar{e}_{1h} \sim (0, \sigma_e^2 + \sigma_u^2/N_h)$. The nested error model is quite popular in small area estimation (e.g. Battese et al, 1988) and it assumes that $Cov(e_{1hi}, e_{1hj}) = \sigma_e^2$ for $i \neq j$. Because N_h is often quite large, we can safely assume that $\bar{e}_{1h} \sim (0, \sigma_e^2)$. The model (3) is called *structural error model* because it describes the structural relationship between the two latent variable Y_h and X_{1h}. The two models, (2) and (3), are often encountered in the measurement error model literature. Thus, the model for small area estimation can be viewed as a measurement error model, as suggested by Fuller (1991) who originally used the measurement error model approach in the unit-level modeling for small area estimation.

Now, if we define $(\bar{y}_{1h}, \bar{x}_h) = N_h^{-1}(\hat{Y}_{1h}, \hat{X}_h)$, combining (2) and (4), we have

$$\begin{pmatrix} \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_0 & \beta_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ \bar{X}_h \end{pmatrix} + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

which can also be written as

$$\begin{pmatrix} \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}. \tag{6}$$

Thus, when all the model parameters in (6) are known, the best estimator of $\bar{X}_h$ can be computed by

$$\hat{\bar{X}}_h = \left\{(\beta_1, 1)\, V_h^{-1}\, (\beta_1, 1)'\right\}^{-1} (\beta_1, 1)\, V_h^{-1}\, (\bar{y}_{1h} - \beta_0, \bar{x}_h)' \tag{7}$$

where V_h is the variance-covariance matrix of $(b_h + \bar{e}_{1h}, a_h)'$. The estimator in (7) can be called the Generalized Least Squares (GLS) estimator because it uses the technique of the generalized least squares method in the linear model theory. The GLS method is useful because it is optimal and it can incorporate additional source of information naturally. For example, if another estimator $\bar{y}_{2h}$ for $\bar{Y}_{2h}$ is also available and satisfies

$$\bar{Y}_{2h} = \gamma_0 + \gamma_1\bar{X}_h + \bar{e}_{2h}$$

and

$$\bar{y}_{2h} = \bar{Y}_{2h} + c_h,$$

then the extended GLS model is written as

$$\begin{pmatrix} \bar{y}_{2h} - \gamma_0 \\ \bar{y}_{1h} - \beta_0 \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} c_h + \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \tag{8}$$

and the GLS estimator can be obtained similarly.

Remark 1 *Note that model (6) can also be written as*

$$\begin{pmatrix} \beta_1^{-1}(\bar{y}_{1h} - \beta_0) \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} (b_h + \bar{e}_{1h})/\beta_1 \\ a_h \end{pmatrix}. \tag{9}$$

The GLS estimator obtained from (9), which should be the same as the GLS estimator obtained from (6), can be expressed as

$$\hat{\bar{X}}_h = \alpha_h \bar{x}_h + (1 - \alpha_h)\, \tilde{x}_h \tag{10}$$

where $\tilde{x}_h = \beta_1^{-1}(\bar{y}_{1h} - \beta_0)$ *and*

$$\begin{aligned} \alpha_h &= \frac{V(\tilde{x}_h) - Cov\,(\bar{x}_h, \tilde{x}_h)}{V(\bar{x}_h) + V(\tilde{x}_h) - 2Cov\,(\bar{x}_h, \tilde{x}_h)} \\ &= \frac{\sigma_{e,h}^2 + V(b_h) - \beta_1 C(a_h, b_h)}{\sigma_{e,h}^2 + V(b_h) + \beta_1^2 V(a_h) - 2\beta_1 C(a_h, b_h)}, \end{aligned}$$

The estimator $\tilde{x}_h$*, when computed with estimated parameter* $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$*, is called the synthetic estimator and the optimal estimator in (10) is often called the composite estimator. It can be shown that, ignoring the effect of estimating* β*, the variance of the composite estimator is equal to*

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \alpha_h V(\bar{x}_h) + (1 - \alpha_h)\, Cov\,(\bar{x}_h, \tilde{x}_h) \tag{11}$$

and, as $\alpha_h < 1$, the composite estimator is more efficient than the direct estimator.

3. Parameter estimation

Now, we discuss estimation of the model parameters in (4). If $\bar{X}_1, \cdots, \bar{X}_H$ were known, then the GLS estimator of (β_0, β_1) could be obtained by minimizing

$$Q(\beta_0, \beta_1) = \sum_{h=1}^{H} \begin{pmatrix} \bar{y}_{1h} - \beta_0 - \beta_1 \bar{X}_h \\ \bar{x}_h - \bar{X}_h \end{pmatrix}' \begin{pmatrix} \sigma_{e,h}^2 + V(b_h) & C(a_h, b_h) \\ C(a_h, b_h) & V(a_h) \end{pmatrix}^{-1} \begin{pmatrix} \bar{y}_{1h} - \beta_0 - \beta_1 \bar{X}_h \\ \bar{x}_h - \bar{X}_h \end{pmatrix}. \tag{12}$$

Because $\bar{X}_1, \cdots, \bar{X}_H$ are unknown, we minimize (12) for the choice of $\hat{\bar{X}}_h = \hat{\bar{X}}_h(\beta_0, \beta_1)$ in (7) or in (10). That is, we minimize

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} \begin{pmatrix} \bar{y}_{1h} - \beta_0 - \beta_1 \hat{\bar{X}}_h \\ \beta_1(\bar{x}_h - \hat{\bar{X}}_h) \end{pmatrix}' \left\{ V \begin{pmatrix} \bar{y}_{1h} - \beta_0 - \beta_1 \hat{\bar{X}}_h \\ \beta_1(\bar{x}_h - \hat{\bar{X}}_h) \end{pmatrix} \right\}^{-1} \begin{pmatrix} \bar{y}_{1h} - \beta_0 - \beta_1 \hat{\bar{X}}_h \\ \beta_1(\bar{x}_h - \hat{\bar{X}}_h) \end{pmatrix}. \tag{13}$$

After some algebra, it can be shown that (13) reduces to

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} \frac{(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)^2}{V(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h)}. \tag{14}$$

As we can write

$$\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1 = -a_h \beta_1 + b_h + \bar{e}_{1h},$$

we have

$$V\left(\bar{y}_{1h} - \beta_0 - \bar{x}_h \beta_1\right) = \sigma^2_{e,h} + (-\beta_1, 1)\, \Sigma_h \,(-\beta_1, 1)'\,. \tag{15}$$

where $\sigma^2_{e,h} = V(\bar{e}_{1h})$ and $\Sigma_h = V\left\{(a_h, b_h)'\right\}$. As we can obtain a consistent estimator of the variance-covariance matrix of (a_h, b_h), we can obtain $(\hat{\beta}_0, \hat{\beta}_1)$ minimizing $Q^*(\beta_0, \beta_1)$ in (14) if $\sigma^2_{e,h}$ is known. Thus, writing

$$Q^*(\beta_0, \beta_1) = \sum_{h=1}^{H} w_h(\beta_1) \left(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h\right)^2, \tag{16}$$

where

$$w_h(\beta_1) = \left\{\sigma^2_{e,h} + (-\beta_1, 1)\, \Sigma_h \,(-\beta_1, 1)'\right\}^{-1},$$

we have

$$\frac{\partial}{\partial \beta_0} Q^* = 0 \quad \Longleftrightarrow \quad \sum_{h=1}^{H} w_h(\beta_1) \left(\bar{y}_{1h} - \beta_0 - \beta_1 \bar{x}_h\right) = 0$$

and so

$$\hat{\beta}_0 = \bar{y}_w - \hat{\beta}_1 \bar{x}_w, \tag{17}$$

where

$$(\bar{x}_w, \bar{y}_w) = \left\{\sum_{h=1}^{H} w_h(\hat{\beta}_1)\right\}^{-1} \sum_{h=1}^{H} w_h(\hat{\beta}_1) \left(\bar{x}_h, \bar{y}_h\right).$$

Plugging (17) into (16), we have only to minimize

$$Q_1^*(\beta_1) = \sum_{h=1}^{H} w_h(\beta_1) \left\{\bar{y}_{1h} - \bar{y}_w - \beta_1(\bar{x}_h - \bar{x}_w)\right\}^2. \tag{18}$$

Thus, we need to find the solution to $\partial Q_1^* / \partial \beta_1 = 0$ where

$$\begin{aligned} \frac{\partial}{\partial \beta_1} Q_1^* \quad &= \quad \sum_{h=1}^{H} \left\{\frac{\partial}{\partial \beta_1} w_h\left(\beta_1\right)\right\} \left\{\bar{y}_{1h} - \bar{y}_w - \beta_1(\bar{x}_h - \bar{x}_w)\right\}^2 \\ &\quad -2 \sum_{h=1}^{H} w_h(\beta_1)(\bar{x}_h - \bar{x}_w) \left\{\bar{y}_{1h} - \bar{y}_w - \beta_1(\bar{x}_h - \bar{x}_w)\right\}. \end{aligned}$$

Using

$$\frac{\partial}{\partial \beta_1} w_h(\beta_1) = -2 \left\{w_h(\beta_1)\right\}^2 \left\{\beta_1 V(a_h) - C(a_h, b_h)\right\},$$

and

$$\left\{\bar{y}_{1h} - \bar{y}_w - \beta_1(\bar{x}_h - \bar{x}_w)\right\}^2 \xrightarrow{p} \sigma^2_{e,h} + (-\beta_1, 1)\, \Sigma_h \,(-\beta_1, 1)' = 1/w_h(\beta_1),$$

where $\overset{p}{\to}$ denotes the convergence in probability, the solution to $\partial Q_1^*/\partial \beta_1 = 0$ satisfies

$$\hat{\beta}_1 = \frac{\sum_{h=1}^H w_h(\hat{\beta}_1)\left\{(\bar{x}_h - \bar{x}_w)(\bar{y}_{1h} - \bar{y}_{1w}) - C(a_h, b_h)\right\}}{\sum_{h=1}^H w_h(\hat{\beta}_1)\left\{(\bar{x}_h - \bar{x}_w)^2 - V(a_h)\right\}}. \tag{19}$$

Note that the weight $w_h(\beta_1)$ depends on β_1. Thus, the solution (19) can be obtained by an iterative algorithm. Once $\hat{\beta}_1$ is computed by (19), then $\hat{\beta}_0$ is obtained by (17).

Now, we discuss estimation of model variance $\hat{\sigma}_{e,h}^2$. The simplest method is the method of moment (MOM). That is, we can use

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h\beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_{e,h}^2 \tag{20}$$

to obtain an unbiased estimator of $\sigma_{e,h}^2$. Under the nested error model in (5), we have $\sigma_{e,h}^2 = \sigma_e^2$ and

$$E\left\{(\bar{y}_{1h} - \beta_0 - \bar{x}_h\beta_1)^2 - \beta_1^2 V(a_h) + 2\beta_1 C(a_h, b_h) - V(b_h)\right\} = \sigma_e^2. \tag{21}$$

Thus, similarly to Fuller (2009), the MOM estimator of σ_e^2 can be obtained by

$$\hat{\sigma}_e^2 = \sum_{h=1}^H \kappa_h \left\{\left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h\hat{\beta}_1\right)^2 - \left(-\hat{\beta}_1, 1\right)\Sigma_h\left(-\hat{\beta}_1, 1\right)\right\} \tag{22}$$

where

$$\kappa_h \propto \left\{\hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1\right)\Sigma_h\left(-\hat{\beta}_1, 1\right)\right\}^{-1}$$

and $\sum_{h=1}^H \kappa_h = 1$. Because κ_h depends on $\hat{\sigma}_e^2$, the solution (22) can be obtain iteratively, using $\hat{\sigma}_e^2 = 0$ as an initial value. Fay and Herriot (1979) used an alternative method which is based on the iterative solution to nonlinear equation:

$$\sum_{h=1}^H \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1\bar{x}_h\right)^2}{\sigma_e^2 + (-\hat{\beta}_1, 1)\Sigma(-\hat{\beta}_1, 1)'} = H - 2.$$

Writing the above equation as $g(\sigma_e^2) = H - 2$, a Newton-type method for $g(\theta) = 0$ with $\theta = \sigma_e^2$ can be obtained by

$$\theta^{(t+1)} = \theta^{(t)} + \frac{1}{g'(\theta^{(t)})}\left(H - 2 - g(\theta^{(t)})\right) \tag{23}$$

where

$$g'(\theta) = -\sum_{h=1}^H \frac{\left(\bar{y}_{1h} - \hat{\beta}_0 - \hat{\beta}_1\bar{x}_h\right)^2}{\left\{\theta + (-\hat{\beta}_1, 1)\Sigma(-\hat{\beta}_1, 1)'\right\}^2}.$$

Assuming $\sigma_{e,h}^2 \equiv \sigma_e^2$, we now describe the whole parameter estimation procedure as follows:

[Step 1] Compute the initial estimator of (β_0, β_1) by setting $\hat{\sigma}_e^2 = 0$ in (17) and (19).

[Step 2] Based on the current value of $(\hat{\beta}_0, \hat{\beta}_1)$, compute $\hat{\sigma}_e^2$ using the iterative algorithm in (23).

[Step 3] Use the current value of $\hat{\sigma}_e^2$, compute the updated estimator of (β_0, β_1) by (17) and (19).

[Step 4] Repeat [Step 2]-[Step 3] until convergence.

Remark 2 *If $\sigma_{e,h}^2 = \sigma_e^2$ is not true, we can consider some alternative model such as*

$$\bar{e}_h \sim \left(0, \bar{X}_h \sigma_e^2\right). \tag{24}$$

To check whether model (24) holds, one can compute

$$\nu_h = \left(\bar{y}_{1h} - \hat{\beta}_0 - \bar{x}_h \hat{\beta}_1\right)^2 - \hat{\beta}_1^2 V(a_h) + 2\hat{\beta}_1 \hat{C}(a_h, b_h) - V(b_h) \tag{25}$$

then plot ν_h and $\bar{x}_h$. If the plot shows a linear relationship, then (24) can be treated as a reasonable model. Under model (24), we can obtain σ_e^2 by a ratio method:

$$\hat{\sigma}_e^2 = \frac{\sum_{h=1}^H \kappa_h \nu_h}{\sum_{h=1}^H \kappa_h \hat{\bar{X}}_h} \tag{26}$$

where

$$\kappa_h \propto \left\{\hat{\bar{X}}_h \hat{\sigma}_e^2 + \left(-\hat{\beta}_1, 1\right) \Sigma_h \left(-\hat{\beta}_1, 1\right)\right\}^{-1}$$

with $\sum_{h=1}^H \kappa_h = 1$, $\hat{\bar{X}}_h$ is defined in (10), and ν_h is defined in (25). Because κ_h also depends on σ_e^2, the solution (26) can be obtained iteratively.

Remark 3 *We can also consider a transformation $\bar{x}_h^* = T(\bar{x}_h)$ and $\bar{y}_{1h}^* = T(\bar{y}_{1h})$ to improve the approximation to asymptotic normality. To check the departure from normality, plot $n_{ha}\bar{V}(\bar{x}_h)$ on $\bar{x}_h$. If the plot shows some structural relationship of $\bar{x}_h$ then the normality assumption can be doubted. Now, consider the following transformation*

$$T(x) = \log(x). \tag{27}$$

Note that the asymptotic variance of $\bar{x}_h^* = T(\bar{x}_h)$ is equal to

$$V\left(\bar{x}_h^*\right) \doteq \frac{1}{(\bar{x}_h)^2} V\left(\bar{x}_h\right).$$

Such transformation is a variance stabilizing transformation and is useful when we want to improve the approximation to normality.

Once the GLS estimator $\hat{\bar{X}}_h^*$ of $\bar{X}_h^*$ is obtained, then we need to apply the inverse transformation to obtain the best estimator of $\bar{X}_h = T^{-1}(\bar{X}_h^*) := Q(\bar{X}_h^*)$. Simply applying the inverse transformation will lead to biased estimation. To correct for the bias, we can use a second-order Taylor linearization. Using a Taylor expansion, we have

$$Q\left(\hat{\bar{X}}_h^*\right) \doteq Q\left(\bar{X}_h^*\right) + Q'\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right) + \frac{1}{2}Q''\left(\bar{X}_h^*\right)\left(\hat{\bar{X}}_h^* - \bar{X}_h\right)^2$$

and so, if we use $Q\left(\hat{\bar{X}}_h^*\right)$ as an estimator for $\bar{X}_h = Q(\bar{X}_h^*)$, we have, ignoring the smaller order terms,

$$E\left\{Q\left(\hat{\bar{X}}_h^*\right)\right\} = \bar{X}_h + \frac{1}{2}Q''\left(\bar{X}_h^*\right) V\left(\hat{\bar{X}}_h^*\right) = \bar{X}_h + \frac{1}{2}\bar{X}_h V\left(\hat{\bar{X}}_h^*\right).$$

For the transformation in (27), we have $Q(\bar{X}_h^*) = \exp(\bar{X}_h^*)$ *and so* $Q^{''}\left(\bar{X}_h^*\right) = \bar{X}_h$*. Thus,* $\hat{\bar{X}}_h = Q(\hat{\bar{X}}_h^*)$*, we have*

$$E\left(\hat{\bar{X}}_h\right) \quad \cong \quad \bar{X}_h + \frac{1}{2}\bar{X}_h V\left(\hat{\bar{X}}_h^*\right).$$

and the bias-corrected estimator of $\bar{X}_h$ *is*

$$\hat{\bar{X}}_{h,bc} = \frac{\hat{\bar{X}}_h}{1 + 0.5V(\hat{\bar{X}}_h^*)} \tag{28}$$

and $V(\hat{\bar{X}}_h^*)$ *is computed by the MSE estimation method which will be discussed in Section 4.*

4. MSE Estimation

We now discuss mean squared error (MSE) estimation of the GLS estimator $\hat{\bar{X}}_h$ which is given by (10). Note that the GLS estimator is a function of (β_0, β_1) and σ_e^2. If the model parameters are known, then the MSE of $\hat{\bar{X}}_h$ is equal to $M_{h1} = \alpha_h V(\bar{x}_h)$, as discussed in Remark 1. That is, writing $\theta = (\beta_0, \beta_1, \sigma_e^2)$ and $\hat{\bar{X}}_h = \hat{\bar{X}}_h(\theta)$, the actual prediction for $\bar{X}_h$ is computed by $\hat{\bar{X}}_{eh} = \hat{\bar{X}}_h(\hat{\theta})$. To account for the effect of estimating the model parameters, we first note the following decomposition of $MSE(\hat{\bar{X}}_h^*)$:

$$\begin{aligned} MSE(\hat{\bar{X}}_{eh}) &= MSE(\hat{\bar{X}}_h) + E\left\{(\hat{\bar{X}}_{eh} - \hat{\bar{X}}_h)^2\right\} \\ &=: M_{h1} + M_{h2} \end{aligned}$$

We consider a jackknife approach to estimate the MSE. The following steps can be used for the jackknife computation.

Step 1. Calculate the k-th replicate $\hat{\theta}^{(-k)}$ of $\hat{\theta}$ by deleting the k-th area data set $(\bar{x}_k, \bar{y}_{1k})$ from the full data set $\{(\bar{x}_h, \bar{y}_{1h})\,; h = 1, 2, \cdots, H\}$. This calculation is done for each k to get H replicates of θ: $\left\{\hat{\theta}^{(-k)}; k = 1, \cdots, H\right\}$ which, in turn, provide H replicates of $\hat{\bar{X}}_h$: $\left\{\hat{\bar{X}}_h^{(-k)}; k = 1, 2, \cdots, H\right\}$, where $\hat{\bar{X}}_h^{(-k)} = \hat{\bar{X}}_h(\hat{\theta}^{(-k)})$.

Step 2. Calculate the estimator of M_{h2} as

$$\hat{M}_{2h} = \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\bar{X}}_h^{(-k)} - \hat{\bar{X}}_h\right)^2. \tag{29}$$

Step 3. Calculate the estimator of M_{h1} as

$$\hat{M}_{1h} = \hat{\alpha}_h^{(JK)}\hat{V}(\bar{x}_h) + \left(1 - \hat{\alpha}_h^{(JK)}\right)\hat{C}\left(\bar{x}_h, \bar{y}_{1h}\right) \tag{30}$$

where $\hat{\alpha}_h^{(JK)}$ is a bias-corrected estimator of α_h given by

$$\hat{\alpha}_h^{(JK)} = \hat{\alpha}_h - \frac{H-1}{H}\sum_{k=1}^{H}\left(\hat{\alpha}_h^{(-k)} - \hat{\alpha}_h\right),$$

$$\hat{\alpha}_h = \frac{\hat{\sigma}_e^2 + V(b_h) - \hat{\beta}_1 C(a_h, b_h)}{\hat{\sigma}_e^2 + V(b_h) + \hat{\beta}_1^2 V(a_h) - 2\hat{\beta}_1 C(a_h, b_h)},$$

and

$$\hat{\alpha}_h^{(-k)} = \frac{\hat{\sigma}_e^{(-k)2} + V(b_h) - \hat{\beta}_1^{(-k)} C(a_h, b_h)}{\hat{\sigma}_e^{(-k)2} + V(b_h) + (\hat{\beta}_1^{(-k)})^2 V(a_h) - 2\hat{\beta}_1^{(-k)} C(a_h, b_h)}.$$

Remark 4 *For the transformation in (27), we use the bias-corrected estimator in (28) and its MSE estimation method needs to be changed. Using $\hat{\bar{X}}_{eh,bc}$ to denote the bias-corrected estimator in (28) evaluated at $\hat{\theta}$, we can have the*

$$\begin{aligned} MSE(\hat{\bar{X}}_{eh,bc}) &= MSE(\hat{\bar{X}}_{eh}) \\ &= MSE\left\{Q\left(\hat{\bar{X}}_{eh}^*\right)\right\} \\ &\cong \left\{Q'\left(\bar{X}_h^*\right)\right\}^2 \cdot MSE\left(\hat{\bar{X}}_{eh}^*\right) \\ &= \bar{X}_h^2 \cdot MSE\left(\hat{\bar{X}}_{eh}^*\right). \end{aligned}$$

where the first equality follows that $\hat{\bar{X}}_{h,bc} - \hat{\bar{X}}_h$ is of order $O_p(n_h^{-1})$. The MSE of $\hat{\bar{X}}_h^*$, the EGLS estimator of $\bar{X}_h^*$ after transformation, is computed by (29) and (30). Once $MSE\left(\hat{\bar{X}}_{eh}^*\right)$ is estimated, we should multiply it by $\hat{\bar{X}}_h^2$ to obtain the MSE estimator of the back-transformed EGLS estimator $\hat{\bar{X}}_{eh,bc}$.

5. Application to Korean Labor Force Survey

We now consider an application of the proposed method to the labor force surveys in Korea. In Korea, two different labor force surveys are used to obtain the information about employment. One is the Korean Labor Force (KLF) survey and the other is the Local Area labor force (LALF) survey. The KLF survey has about 7K sample households but LALF has about 200K sample households. Because LALF is a large-scale survey employing a lot of part time interviewers, there are certain level of measurement errors in the LALF survey. We assume that the KLF has no measurement error, although it has significant sampling errors in small area levels. The KLF sample is a second-phase sample from the LALF sample. Thus, the sampling errors for two survey estimates are correlated. Let $\bar{X}_h$ be the (true) unemployment rate for area h. The small area level we considered is called *Gu*. The number of "Gu" in Korea is 229.

We observe $\bar{x}_h$ from KLF survey and $\bar{y}_{1h}$ from the LALF survey. To construct linking models, we first partition the population into two regions, urban region and rural region, based on the proportion of the households working on agricultural practice. Within each region, we build models separately (same model but allows for different parameter) and estimate the model parameters separately. The structural model is

$$\bar{Y}_h = \beta_1 \bar{X}_h + e_h \tag{31}$$

with $e_h \sim (0, \sigma_e^2)$. Here, we set $\beta_0 = 0$ to guarantee that the GLS estimator of $\bar{X}_h$ is nonnegative. The sampling error model remains the same. In this case, β_1 can be estimated by

$$\hat{\beta}_1 = \frac{\sum_{h=1}^H w_h(\hat{\beta}_1)\left\{\bar{x}_h \bar{y}_{1h} - C(a_h, b_h)\right\}}{\sum_{h=1}^H w_h(\hat{\beta}_1)\left\{\bar{x}_h^2 - V(a_h)\right\}}. \tag{32}$$

The model variance is estimated by the method of moment technique in (22) with $\hat{\beta}_0 = 0$. The GLS estimator can be computed by (10) with $\tilde{x}_h = \hat{\beta}_1^{-1}\bar{y}_{1h}$.

In addition to the two surveys, we can also use the Census information. The GLS model incorporating the three information can be expressed as

$$\begin{pmatrix} \bar{Y}_{2h} \\ \bar{y}_{1h} \\ \bar{x}_h \end{pmatrix} = \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix} \bar{X}_h + \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix}$$

where $\bar{Y}_{2h}$ is the census result for area h. Because Census estimate does not suffer from sampling error, we have only only model error e_{2h} which represents the error when we model $E(\bar{Y}_{h2}) = \gamma_1 \bar{X}_h$. The model parameters can be obtained using the method in Section 3 with $\Sigma = \text{diag}(0, V(a_h))$. The GLS estimator of $\bar{X}_h$ can be obtained easily. The MSE part can be computed by using the fact that

$$V\left(\hat{\bar{X}}_h - \bar{X}_h\right) = \left[\begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix}' \left\{ V \begin{pmatrix} \bar{e}_{2h} \\ b_h + \bar{e}_{1h} \\ a_h \end{pmatrix} \right\} \begin{pmatrix} \gamma_1 \\ \beta_1 \\ 1 \end{pmatrix}\right]^{-1} := M_{h1}$$

and applying the jackknife method for bias correction.

Table 2: Summary of the MSE performance of the small area estimates

MSE	1st Q	Median	3rd Q	Mean
KLF	0.0000630	0.0001210	0.0002395	0.0002476
LALF	0.0001123	0.0001330	0.0001695	0.0001482
GLS 1	0.0000444	0.0000738	0.0001210	0.0000893
GLS 2	0.0000405	0.0000543	0.0000721	0.0000575

Table 2 presents the performance the small area estimates in terms of the MSE estimates. We considered four different estimators of $\bar{Y}_h$. KLF represents the result derived using data only Korea Labor Force survey, LALF represents the result using only Local Area Labor Force survey, GLS 1 represents the result for combining both of survey KLF and LALF, and GLS 2 represents the result for combining KLF, LALF and the census data. Table 2 shows that the GLS 2 method provides the smallest mean squared errors.

6. Concluding Remark

Small area estimation problem is treated as a measurement error model prediction problem. The sampling errors of the direct estimators are treated as measurement errors and the structural error model can be used to link the other auxiliary information to the direct estimator. The measurement error model approach is particularly useful when there are several auxiliary information in area-levels. The resulting estimator is optimal in the sense of minimizing the mean squared errors among the class of unbiased estimators that are linear in the available information.

In the example of the Korean labor survey application, two sample estimates and the Census information are used to compute the GLS estimates for small area parameters and the two sample estimates are correlated due to the two-phase sampling structure. We simply used linear regression models for the linking models, mainly for the sake of computational simplicity. Instead of the linear model, one could consider a generalized linear model to improve model prediction power. Such extension would involve the theory for nonlinear measurement error models. Further investigation on this extension will be a topic of future research.

REFERENCES

Elliot, M.R. and Davis, W.W. (2005), "Obtaining cancer risk factor prevalence estimates in small areas: Combining data from two surveys," *Applied Statistics*, 54, 595–609.

Fay, R.E. and Herriot, R.A. (1979), "Estimation of income from small places: an application of James-Stein procedures to census data," *Journal of the American Statistical Association*, 74, 269–277.

Fuller, W. A. (1991), "Small area estimation as a measurement error problem," in *Economic Models, Estimation, and Socioeconomic Systems: Essays in Honor of Karl A. Fox*, eds. Tij K. Kaul and Jati K. Sengupta, Elsevier Science Publishers, pp. 333–352.

Fuller, W.A. (1987), *Measurement error models*, New York: Wiley.

Fuller, W.A. (2009), *Sampling Statistics*, John Wiley & Sons, Inc., Hoboken, NJ.

Kim, J.K. and Rao, J.N.K. (2012). "Combining data from two independent surveys: a model-assisted approach," *Biometrika* **99**, 85–100.

Merkouris, T. (2010), "Combining information from multiple surveys by using regression for efficient small domain estimation," *Journal of the Royal Statistical Society*, Ser. B, 68, 509–521.

Pfeffermann, D. (2002), "Small area estimation - New developments and directions," *International Statistical Review*, 70, 125–144.

Prasad, N.G.N. and Rao, J.N.K. (1990), "The estimation of the mean squared error of small-area estimators," *Journal of the American Statistical Association*, 85, 163–171.

Rao, J.N.K. (2003), *Small Area Estimation*, John Wiley & Sons, Inc., Hoboken, NJ.

Ybarra, L.M.R. and Lohr, S.L. (2008), "Small area estimation when auxiliary information is measured with error," *Biometrika*, 95, 919–931.