Statistics and Applications {**ISSN 2452-7395 (online)**}
Volume 16, No. 1, 2018 (New Series), pp 227-243

A Random Effect Model Approach to Survey Data Integration

Eunseon Gwak[1], **Jae Kwang Kim**[2] **and Youngwon Kim**[1]
[1]*Department of Statistics, Sookmyung Women's University, Seoul, Korea*
[2]*Department of Statistics, Iowa State University, Ames, Iowa, USA*

Received: December 27, 2017; Revised: Feb 02, 2018; Accepted: Feb 13, 2018

Abstract

Combining information from several surveys, or survey integration, is an important practical problem in survey sampling. When the samples are selected from similar but different populations, random effect models can be used to describe the sample observations and to borrow strength from multiple surveys. In this paper, we consider a prediction approach to survey integration assuming random effect models. The sampling designs are allowed to be informative. The model parameters are estimated using a version of EM algorithm accounting for the sampling design. The mean squared error estimation is also discussed. Two limited simulation studies are used to investigate the performance of the proposed method.

Key words: Best prediction; Small area estimation; Shrinkage method; Constrained EM algorithm.

1 Introduction

Combining information from several independent surveys is an important practical problem in survey sampling. By making the maximum use of information, we can minimize the cost associated with surveys and, at the same time, minimize the respondent burden (Bycroft, 2010). Such problem is called survey integration. Survey integration is an emerging area of research due to the increasing concerns of response burden (Citro, 2014) and availability of the auxiliary information in the era of big data (Tam and Clarke, 2015).

There are two approaches for survey integration. One is macro level approach, which combines the summary information from each source using general least squares estimation or other statistical tools. The US Consumer Expenditure Survey is one of the early applications of survey integration (Zieschang, 1990), where diary survey and quarterly interview survey are used to obtain improved estimates for diary survey item. Zieschang (1990) suggested the use of the generalized least squares adjustment algorithm to incorporate ancillary information to reduce the design variance of estimated survey total. Area level small area estimation (Fay and Herriot, 1979; Rao and Molina, 2015) is also an example of macro approach to survey integration.

The other approach is micro level approach, where the goal is to create a single data that contains all available information. Calibration weighting (Wu, 2004), synthetic data imputation (Kim and Rao, 2012) or statistical matching (Kim et al., 2016) can belong to the micro approach to survey integration. Statistical matching is a tool for integrating two or more sources where

Corresponding Author: Jae Kwang Kim
E-mail: jkim@iastate.edu

variables are never jointly observed. Statistical matching using file concatenation with adjusted weights and multiple imputations are discussed by Rubin (1986).

When the two survey data are obtained from the same target population, we may have systematic differences in two survey measurements, due to measurement errors in survey operation (Biemer et al., 2013). In this case, the survey integration can be developed using measurement error models. Kim et al. (2016) consider the case when the two data sources have a common measurement x but different measurements for y. They developed a statistical matching using fractional imputation based on measurement error models. Park et al. (2017) present an application of the measurement error model approach to survey data integration.

In this paper, we consider another important situation where the two surveys are obtained from two different populations but there are some similarities between the two populations. For example, if the two populations are different only in terms of survey years but otherwise the same, then we can expect that there exists some common structure shared in both populations. In this case, such shared features can be built into the model and we may wish to combine the information from two sources. If the two sub-populations are two mutually exclusive subset of the same finite population, we can expect that the two sub-populations share some common structures. Such information can be built into the model and survey integration method can be developed accordingly. Random effect model is a natural approach of building such information into the models.

In random effect models, we assume common slope and common model variances but we allow for differential intercept terms. The sample-specific effect is assumed to be random, that is $a_i \sim N(0, \sigma_a^2)$, and the parameter (σ_a^2) associated with the random effect model determines the level of homogeneity between the two populations. We first discuss parameter estimation for the random effect model under complex sampling. The sampling design is allowed to be informative in the sense of Sugden and Smith (1984). Once the model parameters are estimated, best prediction for the population mean of y can be developed in the form of synthetic data imputation. Such synthetic data imputation, or mass imputation, can be regarded as micro-level survey data integration.

This paper organized as follows. In Section 2, we introduce the basic setup for survey integration. In Section 3, we discuss the estimation of model parameters. We propose a method of moments estimation and a pretest method for estimation of model parameters. In Section 4, the best prediction for population mean of y and variance estimation are discussed. Results from two limited simulation studies are presented in Section 5. Concluding remarks are made in Section 6.

2 Basic Setup

Suppose that we have two surveys with the same measurement. For simplicity, we assume that two surveys are obtained from the same population in different time points, say $t = 1$ and $t = 2$. Furthermore, the two samples are independently selected. Let Y_1 be the study variable for population one ($t = 1$) and Y_2 be the study variable for population two ($t = 2$). We are interested in estimating the finite population means, denoted by $\mu_1 = N^{-1} \sum_{i=1}^{N} y_{1i}$ and $\mu_2 = N^{-1} \sum_{i=1}^{N} y_{2i}$.

We observe (x_i, y_{1i}) from A_1 selected from population one and observe (x_i, y_{2i}) from A_2 selected from population two. If the two populations share the same structure, for example,

$$y_{ij} = \beta_0 + \beta_1 x_i + e_{ij}, \quad j = 1, 2 \tag{2.1}$$

where $e_{ij} \sim (0, \sigma^2)$. Then there is no systematic difference between the two populations and one can simply combine the two sample and estimate the parameters. However, such assumption is probably unrealistic. Now, if the two populations are completely different, for example,

$$y_{ij} = \beta_{0j} + \beta_1 x_i + e_{ij}, \quad j = 1, 2 \tag{2.2}$$

where $e_{ij} \sim (0, \sigma_j^2)$. Then we cannot combine the two samples and we simply apply separate analysis from each sample.

Now, one can consider a compromise between (2.1) and (2.2), by considering a random effect model for the two populations.

$$y_{ij} = \alpha_j + \beta_0 + \beta_1 x_i + e_{ij}, \quad j = 1, 2 \tag{2.3}$$

where $\alpha_j \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. Note that $\sigma_a^2 = 0$ case reduces to model (2.1). Also, the case of $\sigma_a^2 \to \infty$ is essentially equal to model (2.2).

Thus, model (2.3) is a flexible model that we may borrow strength from the other survey. Note that we assume that the slopes are the same. If such assumption is not justifiable, we can use

$$y_{ij} = \alpha_j + \beta_j x_i + e_{ij}, \tag{2.4}$$

where $\alpha_j \sim N(\mu_a, \sigma_a^2)$, $\beta_j \sim N(\mu_b, \sigma_b^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. Such model can be called random coefficient model.

We first consider parameter estimation and prediction under model (2.3) where the samples are selected from complex sampling designs. Note that if x_i are available throughout the population, then we can predict the population mean of y_{ij} for given j by

$$\hat{\mu}_j^* = \frac{1}{N_j} \left\{ \sum_{i \in A_j} y_{ij} + \sum_{i \in A_j^c} \hat{y}_{ij}^* \right\}$$

where A_j is the set of sample indices for the j-th survey, N_j is the size of population j, and

$$\hat{y}_{ij}^* = \hat{\alpha}_j^* + (\hat{\beta}_0 + \hat{\beta}_1 x_i). \tag{2.5}$$

Here, $\hat{\alpha}_j^*$ is the best predictor of α_j which will be presented in section 4.

If x_i are available only in the two samples, under simple random sampling, we can still use

$$\hat{\mu}_1^* = n^{-1} \left\{ \sum_{i \in A_1} y_{1i} + \sum_{i \in A_2} \hat{y}_{1i}^* \right\}$$

where $n = n_1 + n_2$ and $\hat{y}_{1i}^*$ is obtained from (2.5). If the two surveys are obtained from complex sampling and w_{ij} is the sampling weight for unit i in survey j, then we can use

$$\hat{\mu}_1^* = P \frac{\sum_{i \in A_1} w_{1i} y_{1i}}{\sum_{i \in A_1} w_{1i}} + (1 - P) \frac{\sum_{i \in A_2} w_{2i} \hat{y}_{1i}^*}{\sum_{i \in A_2} w_{2i}}$$

as an estimator of $\mu_1 = E(Y_1)$ using the whole sample, where P is the proportion between 0 and 1. Optimal choice of P will be discussed in Section 4.

To compute the predicted values, parameters in model (2.3) need to be estimated. If both survey samples are obtained by simple random sampling, then the model parameters can be estimated by the maximum likelihood method using EM algorithm Dempster et al. (1977). Parameter estimation under complex survey sampling is discussed in the next section.

3 Parameter Estimation

3.1 ML Estimation using EM Algorithm

We now discuss parameter estimation for the random effects model (2.3) under complex sampling. Under the model (2.3) we are interested in estimating $\theta_1 = (\beta_0, \beta_1, \sigma_e^2)$ and $\theta_2 = \sigma_a^2$ consistently from the sample obtained from complex sampling design. The sampling design can be an element sampling, or any other complex sampling design where the first order and the second order inclusion probabilities are available. The sampling design can be informative when the sample distribution is different from that of the population.

Suppose that $I_i = 1$ if element i is sampled and $I_i = 0$ otherwise. For any measurable set B, the sampling design is called noninformative if

$$P(y_i \in B | x_i, I_i = 1) = P(y_i \in B | x_i) \tag{3.1}$$

The left side is the sample model and the right side is the population model. Equality (3.1) does not hold when the sampling design is informative. Kim et al. (2017) consider parameter estimation in the context of generalized linear mixed models under two-stage cluster sampling where the sampling distribution is informative. In this paper, we consider ML estimation using EM algorithm based on the proposed method by Kim et al. (2017). We will briefly introduce the the basic idea of Kim et al. (2017) and then apply it to our problem.

Suppose that survey units are obtained with two-stage cluster sampling. Let y_{ij} be the study variable for element i in cluster j. Let x_{ij} be the covariates for the regression model for y_{ij}. Assume that the elements in cluster j satisfy the following model

$$y_{ij} | \alpha_j \sim f_1(y_{ij} | x_{ij}, \alpha_j; \theta_1),\ i = 1, ..., M_j \tag{3.2}$$

for some θ_1 with

$$\alpha_j \sim f_2(\alpha_j; \theta_2),\ j = 1, ..., N \tag{3.3}$$

where α_j is the unobserved random effect associated with cluster j.

Assume that clusters are sampled with unequal selection probability and the first-order inclusion probability of cluster j are available. From the sampled cluster j, m_j elements are sampled by simple random sampling. Under this setup, model (3.2) can be written as

$$y_j \sim \tilde{f}_1(y_j | x_j, \alpha_j; \theta_1) = \prod_{i=1}^{m_j} f_1(y_{ij} | x_{ij}, \alpha_j; \theta_1) \tag{3.4}$$

where $y_j = (y_{1j}, \cdots, y_{n_j,j})$ and x_j is the covariate for cluster j.

Let $A^{(1)}$ be the set of indices for sampled clusters. Consider the following pseudo log-likelihood

$$l_p(\theta_1, \theta_2) = \sum_{j \in A^{(1)}} w_j l_j(\theta_1, \theta_2), \tag{3.5}$$

where w_j is the sampling weight associated with cluster j and

$$l_j(\theta_1, \theta_2) = \log \int \tilde{f}_1(y_j | x_j, \alpha_j; \theta_1) f_2(\alpha_j; \theta_2) d\alpha_j$$

is the likelihood function obtained from the marginal distribution of y_j on x_j.

Under informative sampling, that is, when the equality (3.1) does not hold, the density function $\tilde{f}_1(y_j|x_j, \alpha_j; \theta_1)$ in (3.4) is unknown and we cannot obtain the predictive distribution

$$\alpha_j|(x_j, y_j; \theta) \sim \frac{\tilde{f}_1(y_j|x_j, \alpha_j; \theta_1) f_2(\alpha_j; \theta_2)}{\int \tilde{f}_1(y_j|x_j, \alpha_j; \theta_1) f_2(\alpha_j; \theta_2) d\alpha_j}. \tag{3.6}$$

Kim et al. (2017) proposed an alternative method that implements valid EM algorithm without calculating the sample likelihood in (3.4). The idea is to find $\hat{\alpha}_j = \hat{\alpha}_j(\theta)$ such that

$$\hat{\alpha}_j|\alpha_j \sim g_1(\hat{\alpha}_j|\alpha_j; \theta_1)$$

and use following approximate distribution instead of (3.6)

$$\alpha_j|(\hat{\alpha}_j; \theta) \sim \frac{g_1(\hat{\alpha}_i|\alpha_j; \theta_1) f_2(\alpha_j; \theta_2)}{\int g_1(\hat{\alpha}_j|\alpha_j; \theta_1) f_2(\alpha_j; \theta_2) d\alpha_j}. \tag{3.7}$$

By treating α_j as fixed, $\hat{\alpha}_i = \hat{\alpha}_i(\theta)$ be the solution to

$$\frac{\partial}{\partial \alpha_j} \hat{l}_{1j}(\theta_1, \alpha_j) = 0 \tag{3.8}$$

with

$$\hat{l}_{1j}(\theta_1; \alpha_j) = \sum_{i \in A_j^{(2)}} w_{i|j} log f_1(y_{ij}|x_{ij}, \alpha_j; \theta_1)$$

where $A_j^{(2)}$ is the index set of sample elements in cluster j and $w_{i|j}$ is the sampling weight associated with element i in cluster j. The EM algorithm can be implemented as

$$\hat{\theta}_1^{(t+1)} = \underset{\theta_1}{\text{argmax}} \sum_{j \in A^{(1)}} w_j E\{\hat{l}_{1j}(\theta_1, \alpha_j)|\hat{\alpha}_j, x_j; \hat{\theta}^{(t)}\} \tag{3.9}$$

$$\hat{\theta}_2^{(t+1)} = \underset{\theta_2}{\text{argmax}} \sum_{j \in A^{(1)}} w_j E\{\log f_2(\alpha_j; \theta_2)|\hat{\alpha}_j, x_j; \hat{\theta}^{(t)}\} \tag{3.10}$$

For the sampling distribution $g_1(\cdot)$, we can use

$$\hat{\alpha_j}(\hat{\theta}_1)|\alpha_j \sim N\big[\alpha_j, \hat{V}\big\{\hat{\alpha}_j(\hat{\theta}_1)\big\}\big] \tag{3.11}$$

where $\hat{V}\big\{\hat{\alpha}_j(\hat{\theta}_1)\big\}$ is the design unbiased estimator of the variance of $\hat{\alpha_j}(\hat{\theta}_1)$.

Note that under the model (2.3), we assume that two samples are independently selected from the same population in different time point. In the following, notation j and i denote the survey j and element i respectively. Under the model (2.3) by treating α_j as fixed, the score equation for $\theta_1 = (\beta_0, \beta_1, \sigma_e^2)$ and α_j is

$$\sum_j \sum_i w_{ij}\big\{(y_{ij} - \alpha_j - \beta_0 - \beta_1 x_i)\big\}(1, x_i) = (0, 0) \tag{3.12}$$

$$\sum_j \sum_i w_{ij}\big\{(y_{ij} - \alpha_j - \beta_0 - \beta_1 x_i)^2 - \sigma_e^2\big\} = 0 \tag{3.13}$$

$$\sum_i w_{ij}(y_{ij} - \alpha_j - \beta_0 - \beta_1 x_i) = 0 \tag{3.14}$$

where w_{ij} is the sampling weight associated with element i in survey j and for $\theta_2 = (\mu_a, \sigma_a^2)$ is

$$\sum_j (\alpha_j - \mu_a) = 0 \tag{3.15}$$

$$\sum_j \{(\alpha_j - \mu_a)^2 - \sigma_a^2\} = 0 \tag{3.16}$$

From (3.14), $\hat{\alpha}_j(\theta_1) = \bar{y}_j - \beta_0 - \beta_1 \bar{x}_j$, we have $\hat{\alpha}_j(\theta_1) = \alpha_j + \bar{e}_j$ where $\bar{e}_j = (\sum_{i \in A_j} w_{ij})^{-1}(\sum_{i \in A_j} w_{ij} e_{ij})$ and $e_{ij} = y_{ij} - \alpha_j - \beta_0 - \beta_1 x_i$. Since $\alpha_j \sim N(\mu_a, \sigma_a^2)$ with $\mu_a = 0$, we express the distribution of α_j given $\hat{\alpha}_j(\theta_1)$

$$\alpha_j | \hat{\alpha}_j(\theta_1) \sim N[c_j \hat{\alpha}_j(\theta_1), c_j \hat{V}_j(\beta)] \tag{3.17}$$

where $c_j = \frac{\sigma_a^2}{\sigma_a^2 + \hat{V}_j(\beta)}$, $\hat{\alpha}_j(\theta_1) = \bar{y}_j - \beta_0 - \beta_1 \bar{x}_j$ and $\hat{V}(\hat{\alpha}_j) = \hat{V}_j(\beta)$ evaluated at the current parameter value of θ. Therefore, the EM algorithm for estimating θ can be easily derived from (3.17). That is, the M-step for parameter can be written as

$$\sum_j \sum_i w_{ij}\{y_{ij} - c_j^{(t)} \hat{a}_j(\theta_1^{(t)}) - \beta_0^{(t+1)} - \beta_1^{(t)} x_i\} = 0 \tag{3.18}$$

$$\sum_j \sum_i w_{ij}\{y_{ij} - c_j^{(t)} \hat{a}_j(\theta_1^{(t)}) - \beta_0^{(t)} - \beta_1^{(t+1)} x_i\} x_i = 0 \tag{3.19}$$

$$\sum_j \sum_i w_{ij}\{(y_{ij} - c_j^{(t)} \hat{a}_j(\theta_1^{(t)}) - \beta_0^{(t+1)} - \beta_1^{(t+1)} x_i)^2 + c_j^{(t)} \hat{V}_j(\beta^{(t)}) - \sigma_e^{2(t+1)}\} = 0 \tag{3.20}$$

$$\hat{\mu}_a^{(t+1)} = (\sum_j)^{-1} \sum_j c_i^{(t)} \hat{a}_j(\theta^{(t)}) \tag{3.21}$$

$$\hat{\sigma}_a^{2(t+1)} = (\sum_j)^{-1} \sum_j [\{c_j^{(t)} \hat{\alpha}_j(\theta^{(t)})\}^2 + c_j^{(t)} \hat{V}_j(\beta^{(t)})] - \{\hat{\mu}_a^{(t+1)}\}^2 \tag{3.22}$$

where $c_j^{(t)} = \hat{\sigma}_a^{2(t)} / \{\hat{\sigma}_a^{2(t)} + \hat{V}_j(\beta^{(t)})\}$.

The ML estimation of σ_a^2 is obtained for the parameter space $\{\sigma_a^2; \sigma_a^2 > 0\}$. However, if we suspect that $\sigma_a^2 = 0$, then the ML estimator of σ_a^2 will be still positive and the resulting prediction is not necessarily efficient. We propose an alternative parameter estimation method that allows for $\hat{\sigma}_a^2 = 0$ for some cases. The idea is based on pretest estimation (Datta et al., 2011). More details will be presented in the next section.

3.2 Method of Moments Estimation and Constrained EM Algorithm

The ML estimation of σ_a^2 in Section 3.1 is obtained for the parameter space $\{\sigma_a^2; \sigma_a^2 > 0\}$, the ML estimator of σ_a^2 will be still positive and the resulting prediction is not necessarily efficient when the true population follows (2.3) with $\sigma_a^2 = 0$. That is, if the two samples are from the same distribution, ML estimation does not give $\hat{\sigma}_a^2 = 0$. Therefore the critical part is on how to estimate σ_a^2 when we suspect that $\sigma_a^2 = 0$.

We now introduce parameter estimation method for σ_a^2 that can allow for $\hat{\sigma}_a^2 = 0$ with positive probability. As discussed in Section 3.1, we can obtain the marginal distribution of $\hat{\alpha}_j(\theta_1)$ combining (3.11) with $\alpha_j \sim N(0, \sigma_a^2)$, as

$$\hat{\alpha}_j(\theta_1) \sim N(0, \sigma_a^2 + \hat{V}(\hat{\alpha}_j))$$

Since the two samples are independently selected, $Cov(\hat{\alpha}_1(\theta_1), \hat{\alpha}_2(\theta_1)) = 0$. Thus, the expectation of squares of the difference between $\hat{\alpha}_1(\theta_1)$ and $\hat{\alpha}_2(\theta_1)$ is

$$E\{(\hat{\alpha}_1(\theta_1) - \hat{\alpha}_2(\theta_1))^2\} = \hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2) + 2\sigma_a^2. \tag{3.23}$$

Now we define

$$T = \frac{\hat{\alpha}_1(\theta_1) - \hat{\alpha}_2(\theta_1)}{\{\hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2)\}^{1/2}},$$

then we can obtain

$$E\{(T^2 - 1)\} \doteq \frac{2\sigma_a^2}{\hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2)}$$

Thus, we can use

$$\hat{\sigma}_a^2 = \frac{1}{2}\{\hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2)\}(T^2 - 1) \tag{3.24}$$

as a method-of-moments estimator of σ_a^2.

By (3.24), if $|T| < 1$, then $\hat{\sigma}_a^2$ takes negative values. Thus, we will use

$$\hat{\sigma}_a^2 = max\{D, 0\}, \tag{3.25}$$

where $D = \frac{1}{2}\{\hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2)\}(T^2 - 1)$, to avoid the negative estimator of σ_a^2 .

A method-of-moments estimator of σ_a^2 in (3.24) is an approximately unbiased estimator for σ_a^2. This method can improve the precision of estimator for σ_a^2. Once estimator of σ_a^2 is estimated, then we can estimate other parameters $(\beta_0, \beta_1, \sigma_e^2)$ by EM algorithm. We propose a version of EM algorithm, so-called constrained EM algorithm, to estimate model parameters. To implement the constrained EM algorithm, we can consider the following steps:

1. Estimate σ_a^2 by a method-of-moments estimation given by (3.25).
2. Estimate parameters $(\beta_0, \beta_1, \sigma_e^2)$ by EM algorithm given by (3.18)-(3.20).
3. Update $\hat{\sigma}_a^2$ using (3.25) with estimated parameters $(\beta_0, \beta_1, \sigma_e^2)$ from Step 2.
4. Repeat Step 1 $\sim$ Step 3 until convergence.

3.3 Pretest Estimation

We now discuss another parameter estimation method using a pretest method. Pretest method for small area estimation has been discussed by Datta et al. (2011) and Molina et al. (2015). In Section 3.2, we proposed a Method of Moments (MOM) estimator for σ_a^2. The MOM estimator is approximately unbiased estimator for σ_a^2. Instead of using $|T| > 1$ one may also consider $|T| > \gamma$, where $\gamma > 1$, which gives more shrinkage toward zero. In this case, we can use

$$\hat{\sigma}_a^2(\gamma) = \frac{1}{2}\{\hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2)\}(T^2 - \gamma) \tag{3.26}$$

$$\hat{\sigma}_a^2(\gamma) = max\{D_\gamma, 0\} \tag{3.27}$$

where $D_\gamma = \frac{1}{2}\{\hat{V}(\hat{\alpha}_1) + \hat{V}(\hat{\alpha}_2)\}(T^2 - \gamma)$.

Since $\hat{\sigma}_a^2(\gamma)$ in (3.26) depends on unknown shrinkage parameter γ, we wish to find an optimal γ using survey sample data. The idea is based on cross validation. Cross validation is a model validation technique and mainly used in a prediction problem. Cross validation combines prediction errors to asses a performance of estimate of model prediction. Cross validation divide the sample into a training sample and a validation sample. The analysis conducted on the training sample and then validating analysis on the validation sample.

To find an optimal γ in the survey sample data using cross validation, we randomly divide survey sample data into 60% of data (called training set) and 40% of data (called validation set). If γ is given, we can compute a $\hat{\sigma}_a^2(\gamma)$ using (3.26)-(3.27), on the training set. Then the estimation for the other model parameters is implemented by the constrained EM algorithm. Using these estimated parameters, we can predict y_1 in the validation set of survey 1. Now, we define CV that is a sum of square of difference between observed value, y_{1i}, and its predicted value, $\hat{y}_{1i}$, in the validation set. That is

$$CV(\gamma) = \sum_i (y_{1i} - \hat{y}_{1i}(\gamma))^2 \tag{3.28}$$

Note that, the difference of y_{1i} and $\hat{y}_{1i}(\gamma)$ means the prediction error of $\hat{y}_{1i}(\gamma)$ given the choice of γ in the validation set. We will compute $CV(\gamma)$ for $\gamma \in [1, 4]$ in the simulation study. We define the γ^* as

$$\gamma^* = \arg\min_\gamma CV(\gamma) \tag{3.29}$$

where $CV(\gamma)$ is defined by (3.28). The optimal γ^* minimizes the mean squared prediction error of the validation set of survey 1. Once the optimal γ^* is determined by (3.29), we can use (3.27) to estimate σ_a^2 and then estimate the other parameters by the Constrained EM algorithm.

4 Best Prediction

Let Y_1 be the study variable for population one and Y_2 be the study variable for population two. We observe (x_i, y_{1i}) from A_1 selected from population one and observe (x_i, y_{2i}) from A_2 selected from population two. We now discuss the best prediction of y_{1i}. The model for the population elements can be described as

$$\begin{aligned} y_{1i} &= \alpha_1 + \beta_0 + \beta_1 x_i + e_{1i} && (4.1) \\ y_{2i} &= \alpha_2 + \beta_0 + \beta_1 x_i + e_{2i} && (4.2) \end{aligned}$$

where $\alpha_j \sim N(0, \sigma_a^2)$ and $e_{ij} \sim N(0, \sigma_e^2)$. The two error terms, e_{1i} and e_{2i} are independent because the error terms are defined for different units.

Using the parameter estimation method in Section 3, we can obtain consistent estimates for the model parameters, $\beta_0, \beta_1, \sigma_a^2$ and σ_e^2. Assuming that the parameters are all known, we can compute the best prediction of y_{1i} for $i \in A_2$. The best predictor of y_{1i} is given by

$$\hat{y}_{1i}^* = \hat{\alpha}_1^* + (\beta_0 + \beta_1 x_i) \tag{4.3}$$

where

$$\hat{\alpha}_1^* = c_1 \hat{\alpha}_1 \tag{4.4}$$

with

$$c_1 = \frac{\sigma_a^2}{\sigma_a^2 + \hat{V}(\hat{\alpha}_1)}$$

and

$$\hat{\alpha}_1 = (\sum_{i \in A_1} w_{1i})^{-1} \{ \sum_{i \in A_1} w_{1i} (y_{1i} - \beta_0 - \beta_1 x_i) \} = \bar{y}_1 - \beta_0 - \beta_1 \bar{x}_1.$$

Thus, using (4.3), we can compute

$$\begin{aligned} \hat{\mu}_1^* &= P \frac{\sum_{i \in A_1} w_{1i} y_{1i}}{\sum_{i \in A_1} w_{1i}} + (1 - P) \frac{\sum_{i \in A_2} w_{2i} \hat{y}_{1i}^*}{\sum_{i \in A_2} w_{2i}} \\ &:= P \hat{\mu}_{1,d} + (1 - P) \hat{\mu}_{1,s} && (4.5) \end{aligned}$$

as an estimator of $\mu_1 = E(Y_1)$ using the whole sample, where P is the proportion between 0 and 1. Here, the subscript in $\hat{\mu}_{1,d}$ indicates that this estimator is computed directly from the sample A_1 and the subscript in $\hat{\mu}_{1,s}$ indicates that this estimator is computed from synthetic sample data. Note that $\hat{\mu}_1^*$ is unbiased for any choice of $P \in (0, 1)$. A simple choice is $P = \sum_{i \in A_1} w_{1i} / \{ \sum_{i \in A_1} w_{1i} + \sum_{i \in A_2} w_{2i} \}$.

Note that

$$\hat{\mu}_{1,d} = \beta_0 + \beta_1 \bar{x}_1 + \alpha_1 + \bar{e}_1 \tag{4.6}$$

where

$$(\bar{x}_1, \bar{e}_1) = \frac{\sum_{i \in A_1} w_{1i} (x_i, e_{1i})}{\sum_{i \in A_1} w_{1i}}.$$

Also,

$$\begin{aligned} \hat{\mu}_{1,s} &= \hat{\alpha}_1^* + \beta_0 + \beta_1 \bar{x}_2 \\ &= c_1 (\alpha_1 + \bar{e}_1) + \beta_0 + \beta_1 \bar{x}_2, && (4.7) \end{aligned}$$

where $\bar{x}_2 = (\sum_{i\in A_2} w_{2i})^{-1}\sum_{i\in A_2} w_{2i}x_i$. Therefore, combining (4.6) and (4.7) with (4.5), we obtain

$$\hat{\mu}_1^* \;=\; \beta_0 + \beta_1\bar{x}^* + \{P + (1-P)c_1\}(\alpha_1 + \bar{e}_1) \tag{4.8}$$

where $\bar{x}^* = P\bar{x}_1 + (1-P)\bar{x}_2$. Thus, we obtain

$$V\{\hat{\mu}_1^*\} = \beta_1^2 V(\bar{x}^*) + \{P + (1-P)c_1\}^2\{\sigma_a^2 + V(\bar{e}_1)\}. \tag{4.9}$$

We can use (4.9) to obtain the optimal choice of P that minimizes the variance (4.9) and also use it for variance estimation as follows

$$\hat{V}\{\hat{\mu}_1^*\} = \hat{\beta}_1^2\hat{V}(\bar{x}^*) + \{P^* + (1-P^*)\hat{c}_1\}^2\{\hat{\sigma}_a^2 + \hat{V}(\bar{e}_1)\}, \tag{4.10}$$

where $\hat{c}_1 = \frac{\hat{\sigma}_a^2}{\hat{\sigma}_a^2 + \hat{V}(\bar{e}_1)}$.

5 Simulation Study

To test our proposed methods, we performed two limited simulations. In this section, we present the results from two simulation studies. In the first simulation, we assume that the samples are selected by simple random sampling under the random effect model. In the second simulation, we assume that the samples are selected by two stage cluster sampling under the random effect model. In the simulation study, our goal is to obtain $\hat{\mu}_1$ from the two survey data. We will compare the proposed method and existing method in three cases, $\sigma_a^2 \in \{0, 0.2, 1\}$. Five methods are considered in the simulation study.

1. Separate: We compute a simple mean of y_{ij} in the survey 1 sample.

$$\hat{\mu}_1 = n_1^{-1}\sum_{i\in A_1} y_{1i}$$

2. Combined: We compute a simple mean of y_{ij} from the pooled sample data.

$$\hat{\mu}_1 = n^{-1}[\sum_{i\in A_1} y_{1i} + \sum_{i\in A_2} y_{2i}]$$

3. Best prediction using EM algorithm: We compute a simple mean of y_1 from the survey 1 sample and estimate a mean of y_1 in survey 2 sample.

$$\hat{\mu}_1 = n^{-1}[\sum_{i\in A_1} y_{1i} + \sum_{i\in A_2} \hat{y}_{1i}^*],$$

where $\hat{y}_{1i}^* = \hat{\alpha}_1^* + \hat{\beta}_0 + \hat{\beta}_1 x_i$. Here, $\hat{\alpha}_1^*$ can be obtained using (4.4) and model parameters$(\beta_0, \beta_1, \sigma_e^2, \sigma_a^2)$ are estimated by EM algorithm.

4. Best prediction using Method of Moments estimation: σ_a^2 is estimated by method of moments estimation described in (3.24)-(3.25) and then use the constrained EM algorithm method for other parameter estimation.

5. Best prediction using Pretest method: σ_a^2 is obtained from optimal γ^* in (3.29) which minimize a $CV(\gamma)$ in the validation set and then use the constrained EM algorithm for other parameter estimation.

Note that, when we set $\sigma_a^2 = 0$ in the simulation setup, then the population structure follows model (2.1). That is, there is no systematic difference between the two populations. We are mainly interested in investigating the performance of different estimation methods in this case.

5.1 Simulation One

For the simulation one, we generate two finite populations with size $N_1 = N_2 = 10,000$ from a random effect model and then select simple random samples with size $n_1 = n_2 = 100$ independently from each finite population. We repeat this Monte Carlo simulation independently $B = 2,000$ times.

The finite populations are generated from

$$y_{1i} = \alpha_1 + \beta x_i + e_{1i},$$
$$y_{2i} = \alpha_2 + \beta x_i + e_{2i}$$

where $\alpha_j \sim N(\mu_a, \sigma_a^2)$, $e_{ij} \sim N(0, \sigma_e^2)$, $x_i \sim N(1,1)$, $\mu_a = 1$, $\sigma_e = 0.1$, $\beta_1 = 1$. Note that, in this setup, $\beta = \beta_1$, $\mu_a = \beta_0$ in the model (2.3).

Table 1 presents the Monte Carlo biases, standard errors of the estimated parameters from ML estimation using EM algorithm(ML), Method of Moment estimation based on the Constrained EM algorithm(MOM) and Pretest method(Pretest) for different values of σ_a^2. The simulation result shows that all estimated parameters are nearly unbiased. Standard error(SE) of β_0 estimated by ML is bigger than MOM and Pretest when $\sigma_a^2 = 0$. When σ_a^2 increases, standard error(SE) of β_0 is almost the same in all methods.

Table 2 presents Monte Carlo mean squared errors (MSE) of the point estimators for μ_1 for the three scenarios of σ_a^2. The MSE of the point estimators for μ_1 for ML method is bigger than those for MOM and Pretest method when $\sigma_a^2 = 0$. However, the differences are very small.

Since the ML estimation of σ_a^2 is obtained for the parameter space $\{\sigma_a^2; \sigma_a^2 > 0\}$, the ML estimation of σ_a^2 will be positive. When σ_a^2 is large, the MSE of point estimator for μ_1 for combined method is much bigger than those for MOM and Pretest method.

Table 3 shows that Monte Carlo means and variances of γ^* which is obtained from (3.29). Since γ is a unknown shrinkage parameter, we determined an optimal γ by pretest method based on cross validation. When $\sigma_a^2 = 0$, Monte Carlo Mean of γ^* is 1.311 and variance is 0.550. Note that, as described in Section 3.2, MOM estimator uses $\gamma = 1$ when we estimate σ_a^2.

5.2 Simulation Two

For the simulation Two, we generate two finite populations from a random effect model and then select a sample from each finite population using a two stage cluster sampling design. To implement the simulation we can consider the following steps:

1. The first stage: $n_j = 100$ clusters are sampled from each finite population with PPS.
2. The second stage: $m = 10$ elements are sampled from each cluster with SRS.

The finite population model as follows:

$$y_{kij} = \alpha_j + \beta x^*_{kij} + e_{kij}, \;\; i = 1, ...M_k, \;\; k = 1, ..., N, \;\; j = 1, 2$$

where $\alpha_j \sim N(\mu_a, \sigma_a^2)$, $e_{kij} \sim N(0, \sigma_e^2)$, $x^*_{kij} \sim N(1, 1)$, $\mu_a = 1$, $\sigma_e^2 = 0.02$, $\beta = 1$, $N = 10,000$. Note that, in this setup, $\beta = \beta_1$, $\mu_a = \beta_0$ in the model (2.3). Here, $x^*_{kij} = x_{kij} + r_k$ where $x_{kij} \sim N(1, 0.5)$ and $r_k \sim N(0, 0.5)$. The cluster size is $M_k = round(50\tilde{r}_k)$, with $\tilde{r}_k = exp(2.5 + r_k)/(1 + exp(2.5 + r_k))$. From each finite population, $n_j = 100$ clusters are sampled by the probability proportional to size(PPS) sampling with the measure of size M_k. That is, the first order inclusion probability is equal to $\pi_k^{(1)} = n_j M_k / \sum_{k=1}^{N_1} M_k$.

Once the clusters are sampled, $m = 10$ elements are sampled by simple random sampling in each cluster. We repeat this simulation independently with $B = 500$.

Table 4 present the Monte Carlo biases, standard errors of the estimated parameters from ML estimation using EM algorithm(ML), Method of Moment estimation based on the Constrained EM algorithm(MOM) and Pretest method(Pretest) for $\sigma_a^2 = 0, 0.2, 1.0$, respectively. Table 4 shows that all estimated parameter are nearly unbiased when $\sigma_a^2 = 0$. When $\sigma_a^2 = 0.2$, the estimator of β_0 has a positive bias 0.017 and estimator of σ_a^2 has a negative bias -0.002 in all methods. When $\sigma_a^2 = 1$, estimator of β_0 has a positive bias 0.039 in all methods. Table 4 show that Bias and Standard error(SE) of the parameter estimation methods has almost the same results for three methods.

Table 5 presents Monte Carlo mean squared errors (MSE) of the point estimators for μ_1 for the three scenarios for σ_a^2. The simulation results show that MOM and Pretest methods perform well in all cases.

Table 6 shows that Monte Carlo means and variances of γ^* which is obtained from (3.29). Optimal γ^* is determined by pretest method based on cross validation. As presented in table 6, Monte Carlo Mean of γ^* is 1.163 and variance is 0.336 when $\sigma_a^2 = 0$.

6 Concluding Remark

We have considered the problem of survey integration using random effect models in survey sampling. Parameter estimation for random effect model involves a version of EM algorithm adapted to survey sampling setup. Constrained EM algorithm is developed to obtain better prediction when there is some possibility that there is no systematic difference between the two populations. In the simulation study, the prediction based on constrained EM shows better performances.

The proposed method is developed only under the linear random effect model. Extension to generalized linear mixed models will be a topic of future study.

Acknowledgements

We thank an anonymous referee for very constructive comments. The research of the second author is supported by a grant from U.S. National Science Foundation (MMS- 1733572).

References

Biemer, P. P., Groves, R. M., Lyberg, L. E., Mathiowetz, N. A. and Sudman, S. (2013). *Measurement Errors in Surveys.* Wiley, Hoboken, New Jersey.

Bycroft, C. (2010). *Integrated household surveys: A survey vehicles approach.* Statistics New Zealand, Wellington.

Citro, C. F. (2014). From multiple modes for surveys to multiple data sources for estimates. *Survey Methodology*, **40**, 137–161.

Datta, G. S., Hall, P. and Mandal, A. (2011). Model selection by testing for the presence of small-area effects, and application to area-level data. *Journal of the American Statistical Association*, **106**, 362–374.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). *Maximum likelihood from incomplete data via the EM algorithm*, Vol. 39.

Fay and Herriot (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, **74**, 269–277.

Kim, J. K., Berg, E. and Park, T. (2016). Statistical matching using fractional imputation, *Survey Methodology*. **42**, 19–40.

Kim, J. K., Park, S. and Lee, Y. (2017). Statistical inference using generalized linear mixed models under informative cluster sampling. *Canadian Journal of Statistics*, **45**, 479–497.

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

Molina, I., Rao, J. N. K. and Datta, G. S. (2015). Small area estimation under a fay-herriot model with preliminary testing for the presence of random area effects. *Survey Methodology*, **41**, 1–19.

Park, S., Kim, J. K. and Stukel, D. (2017). A measurement error model for survey data integration: combining information from two surveys. *Metron*, **75**, 345–357.

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation*. Wiley, Wiley series in survey methodology. 2nd ed. Hoboken, New Jersey.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, **4**, 87–94.

Sugden, R. A. and Smith, T. M. F. (1984). Ignorable and informative designs in survey sampling inference. *Biometrika*, **71**, 495–506.

Tam, S.-M. and Clarke, F. (2015). Big data, official statistics and some experience of the australian bureau of statistics. *International Statistical Review*, **83**, 436–448.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *The Canadian Journal of Statistics*, **32**, 15–26.

Zieschang (1990). Sample weighting methods and estimation of totals in the consumer expenditure survey. *Journal of the American Statistical Association*, **85**, 986–1001.

Table 1: **Monte Carlo biases, standard errors of the proposed methods for simulation study one**

		Parameter	**ML**	**MOM**	**Pretest**
$\sigma_a^2 = 0$	Bias	β_0	0.000	0.000	0.000
		β_1	0.000	0.000	0.000
		σ_e^2	0.000	0.000	0.000
		σ_a^2	0.000	0.000	0.000
	S.E	β_0	0.008	0.003	0.002
		β_1	0.006	0.005	0.005
		σ_e^2	0.000	0.000	0.000
		σ_a^2	0.000	0.000	0.000
$\sigma_a^2 = 0.2$	Bias	β_0	-0.006	-0.006	-0.006
		β_1	0.000	0.000	0.000
		σ_e^2	0.000	0.001	0.001
		σ_a^2	-0.001	-0.001	-0.001
	S.E	β_0	0.316	0.312	0.311
		β_1	0.006	0.025	0.026
		σ_e^2	0.001	0.015	0.016
		σ_a^2	0.278	0.278	0.278
$\sigma_a^2 = 1$	Bias	β_0	-0.006	-0.006	-0.006
		β_1	0.000	0.000	0.000
		σ_e^2	0.000	0.001	0.001
		σ_a^2	-0.001	-0.001	-0.001
	S.E	β_0	0.316	0.312	0.311
		β_1	0.006	0.025	0.026
		σ_e^2	0.001	0.015	0.016
		σ_a^2	0.278	0.278	0.278

Table 2: **Monte Carlo mean squared errors(MSE) of point estimator for μ_1**

Parameter	**Method**	$\sigma_a^2 = 0$	$\sigma_a^2 = 0.2$	$\sigma_a^2 = 1$
μ_1	Separate	0.01078	0.01036	0.01036
	Combined	0.00521	0.10449	0.50095
	ML	0.00529	0.00555	0.00555
	MOM	0.00525	0.00570	0.00570
	Pretest	0.00526	0.00570	0.00570

Table 3: **Monte Carlo Means and Variances of γ^* for Pretest method**

	Mean	**Variance**
$\sigma_a^2 = 0$	1.311	0.550
$\sigma_a^2 = 0.2$	2.311	2.161
$\sigma_a^2 = 1$	2.386	2.196

Table 4: **Monte Carlo biases, standard errors of the proposed methods for simulation study two**

		Parameter	**ML**	**MOM**	**Pretest**
$\sigma_a^2 = 0$	Bias	β_0	0.000	0.000	0.000
		β_1	0.000	0.000	0.000
		σ_e^2	0.000	0.000	0.000
		σ_a^2	0.000	0.000	0.000
	S.E	β_0	0.003	0.003	0.003
		β_1	0.002	0.001	0.001
		σ_e^2	0.000	0.000	0.000
		σ_a^2	0.000	0.000	0.000
$\sigma_a^2 = 0.2$	Bias	β_0	0.017	0.017	0.017
		β_1	0.000	0.000	0.000
		σ_e^2	0.000	0.000	0.000
		σ_a^2	-0.002	-0.002	-0.002
	S.E	β_0	0.315	0.315	0.315
		β_1	0.002	0.002	0.002
		σ_e^2	0.000	0.000	0.000
		σ_a^2	0.268	0.268	0.268
$\sigma_a^2 = 1$	Bias	β_0	0.039	0.039	0.039
		β_1	0.000	0.000	0.000
		σ_e^2	0.000	0.000	0.000
		σ_a^2	-0.013	-0.013	-0.013
	S.E	β_0	0.704	0.704	0.704
		β_1	0.002	0.002	0.002
		σ_e^2	0.000	0.000	0.000
		σ_a^2	1.340	1.340	1.340

Table 5: **Monte Carlo mean squared errors(MSE) of point estimator for** μ_1

Parameter	**Method**	$\sigma_a^2 = 0$	$\sigma_a^2 = 0.2$	$\sigma_a^2 = 1$
μ_1	Separate	0.001038	0.00102	0.00102
	Combine	0.000526	0.09933	0.49407
	ML	0.000540	0.00050	0.00050
	MOM	0.000538	0.00050	0.00050
	Pretest	0.000537	0.00050	0.00050

Table 6: **Monte Carlo Means and Variances of γ^* for Pretest method**

	Mean	**Variance**
$\sigma_a^2 = 0$	1.163	0.336
$\sigma_a^2 = 0.2$	2.394	2.161
$\sigma_a^2 = 1$	2.436	2.197