Bivariate Fay-Herriot Models with Application to Thai Socio-Economic Data

Annop Angkunsit and Jiraphan Suntornchost*

Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand
* Corresponding author. E-mail address: Jiraphan.S@chula.ac.th
Received: 6 March 2020; Revised: 15 June 2020; Accepted: 22 June 2020

Abstract

Welfare information is very important for policy makers and the government in order to improve the nation economic status. Most common welfare indicators widely used are expenditure and income. In practice, studying the two indicators separately could lead to different conclusions. Accordingly, to have precise viewpoints of the nation economic status, the two measurements should be simultaneously studied via a bivariate model. One of well-known models used in small area is the Fay-Herriot model. However, standard variance component estimation methods for the Fay-Herriot model frequently produce zero estimate of the strictly positive model variance. Therefore, Li and Lahiri proposed an adjusted method to prevent zero estimate of model variance for the univariate Fay-Herriot model. In this paper, we extend their technique to obtain an adjusted likelihood estimate for a bivariate Fay-Herriot model and apply the method to estimate income and expenditure in Thailand. In our study, simulation study is carried out to investigate the performance of our adjusted method comparing with the original profile likelihood method. The simulation results suggest that our adjusted profile likelihood estimates prevent zero estimates and outperform the profile likelihood estimates. Consequently, an empirical study is performed for the Thai income and expenditure welfare measurements using data from the 2017 Thailand Household Socio-Economic Survey (SES 2017) and the 2010 Thailand Population and Housing Census.

Keywords: Small area estimation, Bivariate Fay-Herriot model, Empirical best linear unbiased predictor, Adjusted maximum likelihood method, Income and Expenditure

Introduction

Household welfare is an important information of governments to measure the nation economic status and to make plan for the nation policy in order to improve the nation living standards. Two common welfare measurements widely used in many countries are income and expenditure. In some countries, particularly for developing countries, expenditure is often used as an indicator because expense data do not fluctuate much across time. Moreover, most of households are in the agriculture, which the spending pattern does not change much and most of the regular expenditures are food and necessities. While most of household income comes from agriculture, there is uncertainty in different years depending on climate and product price. In contrast, income is often used as an indicator in many developed countries because income data is more memorable than expenses. Most of their incomes come from regular salary and wages, while expenses have quite a lot of spending patterns.

However, both income and expenditure could give information on household welfare in different aspects. They could give different conclusions on welfare. Therefore, in order to efficiently measure household welfare, both income and expenditure should be considered. Due to this concern, many countries including Thailand conduct regular survey on both income and expenditure. The Thailand's National Statistical Office (NSO) conducts annual survey of household income and expenditure of Thai population called the Household Socio-Economic Survey (SES). The average household income and average household expenditure are computed by using the information of household collected from all districts and provinces in Thailand.

In general, direct survey estimates are used in presenting population estimates. The direct estimates are efficient if the sample size is sufficiently large. However, in some situations, we don't have good quality data for the direct estimation method to give reliable estimates. Therefore, alternative estimates have been proposed in literature such as the small area estimation (SAE) method. The basic concept of SAE method is to link the variables of interest with auxiliary information (e.g., Census and Administrative data) in a model to define the model-based estimator that "borrow strength" from the related area (Rao & Molina, 2015). One of widely used models is the Fay-Herriot model proposed by Fay and Herriot (1979) to improve direct estimates by incorporating sampling effect into models. For small area i ($i = 1, \dots, m$), let θ_i be the unobserved true area mean, and y_i be a direct estimate of the area mean. The model consists of two levels.

In level 1, called the sampling model, we assume that

$$y_i|\theta_i \sim N(\theta_i, D_i), \text{ independently for } i = 1, \dots, m,$$

where D_i $(i = 1, \dots, m)$ is a sequence of sampling variances assumed to be known. This level of the model accounts for sampling variability of the direct survey estimates y_i from the true population means θ_i.

In level 2, called the linking model, the true mean is linked with available auxiliary variables $(\boldsymbol{x})$. That is

$$\theta_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, A), \text{ independently for } i = 1, \dots, m,$$

where A is the regression variance and the auxiliary variables used in the model are usually from administrative records and census data.

The unknown parameter θ_i is commonly estimated by the empirical best linear unbiased predictor (EBLUP) estimate, denoted by $\hat{\theta}_i$. The EBLUP estimate is the weighted sum of the direct estimator y_i and the regression estimator $\boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. Specifically,

$$\hat{\theta}_i = \frac{\hat{A}}{\hat{A} + D_i} y_i + \frac{D_i}{\hat{A} + D_i} \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}, \quad (1)$$

where $\widehat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{m} \boldsymbol{x}_i \boldsymbol{x}_i' (\hat{A} + D_i)^{-1}\right)^{-1} \left(\sum_{i=1}^{m} \boldsymbol{x}_i y_i (\hat{A} + D_i)^{-1}\right)$ and $\hat{A}$ is an estimate of the regression variance A. We can see that the weights in (1) depend on the estimate $\hat{A}$. The precision in estimating the variance component A strongly influences the accuracy of the EBLUP estimates. Therefore, several methods in estimating $\hat{A}$ have been explored in literature such as the profile maximum likelihood method (Hartley & Rao, 1967). The profile maximum likelihood parameter estimation method has been widely used in many studies. However, the method can produce zero estimate of A in some situations. In such cases, the EBLUP estimate produces undesirable estimate because it ignores the direct estimator from survey data and reduces to the regression estimator. To prevent such situations, Li and Lahiri (2010) proposed an adjusted maximum likelihood method to avoid zero estimate of A in the EBLUP estimate for the univariate Fay-Herriot model.

The Fay-Herriot model has been extended to multivariate models and studied by many authors. Fay (1987) and Datta, Fay, and Ghosh (1991) compared the precision of small area estimators obtained from univariate models for each response variable with the ones obtained by a multivariate model. Datta, Ghosh, Nangia, and Natarajan (1996) used also a multivariate Fay-Herriot model for obtaining hierarchical Bayes estimates of median income of four-person families for U.S. states. González-Menteiga, Lombardía, Molina, Morales, and Santamaría (2008) studied a class of multivariate Fay-Herriot model with a common random effect for all the components of the target vector. Benavent and Morales (2016) studied a class of multivariate Fay-Herriot models with one random effect per component of the target vector and allowing for different covariance patterns

between the components of the vector of random effects. However, based on our knowledge, there is no extension of adjusted maximum likelihood method available for bivariate Fay-Herriot models. Therefore, in order to simultaneously model income and expenditure, we will first extend the concept of adjusted maximum likelihood proposed by Li and Lahiri (2010) to obtain an adjusted maximum likelihood estimate for bivariate Fay-Herriot model. We then apply the bivariate Fay-Herriot model and the new obtained adjusted maximum likelihood estimates to produce EBLUP estimates of household income and expenditure in Thailand.

The remainder of this paper is divided into four sections as follows. First, we introduce the bivariate Fay-Herriot model, the adjusted maximum likelihood estimates for the EBLUP estimates, simulation setting and data description of the Thai socio-economic data. Second, we investigate the performance of our estimators via simulation experiments and discuss the application of our new estimates to household incomes and expenditures. Third, we discuss our results. Finally, we give conclusions and suggestions of future research.

Methods and Materials

In this section, we discuss methods and materials used in our study. We first describe the model used in our paper which is the bivariate Fay-Herriot model. We then extend the adjusted maximum likelihood method of Li and Lahiri (2010) to the bivariate Fay-Herriot model. Finally, we explain the setting of numerical simulations and data description of the socio-economic data.

Bivariate Fay-Herriot model

The structure of the bivariate Fay-Herriot model studied in this work is a special case of the multivariate Fay-Herriot model discussed in Benavent and Morales (2016). The model is described as follows supposing that the population is partitioned into m subpopulations. For domain i $(i = 1, \dots, m)$, let $\boldsymbol{\theta}_i = (\theta_{i1}, \theta_{i2})'$ be the vector of characteristics of interest and let $\boldsymbol{y}_i = (y_{i1}, y_{i2})'$ be a vector of direct estimators of $\boldsymbol{\theta}_i$. The model assumes that $\boldsymbol{\theta}_i$ is linearly related to the auxiliary variables $\boldsymbol{X}_i = \text{diag}(\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})$ with p explanatory variables $\boldsymbol{x}_{ij} = (x_{ij1}, \dots, x_{ijp})$, for $j = 1, 2$ and $i = 1, \dots, m$, through the model

$$\boldsymbol{\theta}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{v}_i, \qquad \boldsymbol{v}_i \sim N(\boldsymbol{0}, A\boldsymbol{I}_2) \quad i = 1, \dots, m, \tag{2}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$ is a vector of coefficients with $\boldsymbol{\beta}_j$, $j = 1, 2$ are column vectors of size p, and A is the variance of area random effect. The direct estimator $\boldsymbol{y}_i$ follows a sampling model

$$\boldsymbol{y}_i = \boldsymbol{\theta}_i + \boldsymbol{e}_i, \qquad \boldsymbol{e}_i \sim N(\boldsymbol{0}, \boldsymbol{D}_i) \quad i = 1, \dots, m, \tag{3}$$

where $\boldsymbol{D}_i$ is a 2×2 covariance matrix of sampling errors.

The bivariate Fay-Herriot model described by (2) and (3) can be rewritten as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{v} + \boldsymbol{e}, \qquad \boldsymbol{v} \sim N(\boldsymbol{0}, A\boldsymbol{I}_{2m}), \qquad \boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{D}), \tag{4}$$

where

$$\boldsymbol{y} = \text{col}_{1 \le i \le m}(\boldsymbol{y}_i), \qquad \boldsymbol{X} = \text{col}_{1 \le i \le m}(\boldsymbol{X}_i), \qquad \boldsymbol{v} = \text{col}_{1 \le i \le m}(\boldsymbol{v}_i), \qquad \boldsymbol{e} = \text{col}_{1 \le i \le m}(\boldsymbol{e}_i),$$

$\boldsymbol{D} = \text{diag}_{1 \le i \le m}(\boldsymbol{D}_i)$ and the random effects $\boldsymbol{v}$ are independent of the sampling errors $\boldsymbol{e}$.

Under model (4), the mean vector and the covariance matrix of $\boldsymbol{y}$ are

$$\text{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{\beta}, \ \text{Var}(\boldsymbol{y}) = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}(A) = A\boldsymbol{I}_{2m} + \boldsymbol{D}.$$

When the regression variance A is known, the true area mean is estimated by the best linear unbiased prediction (BLUP) proposed by Henderson (1975):

$$\tilde{\boldsymbol{\theta}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + A\boldsymbol{\Sigma}^{-1}(\boldsymbol{y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}}), \tag{5}$$

where $\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(A) = (\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}\boldsymbol{y}$. In practice, A is unknown but it can be estimated. One of widely used estimates is the profile maximum likelihood (PML) method. The profile maximum likelihood (PML) method maximizes the joint probability density function of the random vector $\boldsymbol{y}$. The joint probability density function of $\boldsymbol{y}$ is the profile maximum likelihood function:

$$L(A) = \frac{1}{(2\pi)^{n/2}}|\boldsymbol{\Sigma}|^{-1/2}\exp\left\{-\frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}\right\}, \tag{6}$$

where $\boldsymbol{P} = \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}^{-1}$. The corresponding profile log-likelihood function is

$$\ell(A) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y}.$$

Thus, the profile maximum likelihood estimator $\hat{A}$ is obtained by maximizing the profile log-likelihood function $\ell(A)$. Substituting $\hat{A}$ into (5), we obtain the Empirical BLUP or EBLUP ($\hat{\boldsymbol{\theta}}$) of $\boldsymbol{\theta}$. That is,

$$\hat{\boldsymbol{\theta}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} + \hat{A}\hat{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}), \tag{7}$$

where $\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}(\hat{A})$ and $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}(\hat{A})$. However, based on our investigation, the profile maximum likelihood method produces zero estimate of A which is also occurred in univariate model discussed in Li and Lahiri (2010). Therefore, in the next section, we extend the adjusted maximum likelihood method proposed by Li and Lahiri (2010) to obtain a nonzero estimate of the regression variance A for bivariate Fay-Herriot model.

Adjusted maximum likelihood method

To avoid the zero weight of the direct estimate in bivariate EBLUP, we apply an adjusted profile likelihood function of A in Li and Lahiri (2010) defined by

$$L_{\text{adj}}(A) = A \times L(A),$$

where $L(A)$ is the profile maximum likelihood function defined in (6). The corresponding adjusted profile log-likelihood function is

$$\ell_{\text{adj}}(A) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\boldsymbol{y}'\boldsymbol{P}\boldsymbol{y} + \log(A).$$

The adjusted profile maximum likelihood (APML) estimator $\hat{A}$ of A is obtained by maximizing the adjusted profile likelihood function $L_{\text{adj}}(A)$. The adjusted profile maximum likelihood estimator of A is strictly positive (see Li & Lahiri 2010) even for small m. Under the same regularity conditions given in Li and Lahiri (2010), we can show that the adjusted profile maximum likelihood estimator of A is a consistent estimate. Moreover, the bias and mean squares error of the adjusted profile likelihood are

$$E(\hat{A} - A) = \frac{\text{tr}(\boldsymbol{P} - \boldsymbol{\Sigma}^{-1}) + \frac{2}{A}}{\text{tr}(\boldsymbol{\Sigma}^{-2})} + o(m^{-1}), \text{ and } \quad E(\hat{A} - A)^2 = \frac{2}{\text{tr}(\boldsymbol{\Sigma}^{-2})} + o(m^{-1}),$$

respectively. The bias and mean squared error are equal to those of the original profile maximum likelihood estimator of A up to order $O(m^{-1})$.

Simulation setting

In this section, we describe a simulation study designed to analyze the behavior of the variance estimate $\hat{A}$ and EBLUP $\widehat{\boldsymbol{\theta}}$ based on bivariate Fay–Herriot model with different patterns of correlations among components of sampling errors. The simulation settings follow González–Menteiga et al. (2008); Li and Lahiri (2010).

In the simulation, we first simulate $\boldsymbol{\theta}_i$ $(i = 1, \dots, m)$ from (2). The matrix of covariates $\boldsymbol{X}_i = (\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2})'$ of two covariates are generated from a bivariate normal distribution with means $\mu_{x1} = \mu_{x2} = 10$, variances $\sigma_{x1}^2 = 1$ and $\sigma_{x2}^2 = 2$ and covariance $\sigma_{x12} = 1/\sqrt{2}$. This setting yields a correlation of 0.5. The regression coefficients are $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (1, 1)'$. The random effects $\boldsymbol{v}_i$ are generated from a normal distribution with mean zero and variance $A = 2$. Having obtained $\boldsymbol{\theta}_i$, we simulate $\boldsymbol{y}_i$ from (3) where sampling errors $\boldsymbol{e}_i$ are generated from a bivariate normal distribution with mean zero and covariance matrix $\boldsymbol{D}_i$. To study different situations of sampling errors, we let $\boldsymbol{D}_i = \left(D_{ijk}\right)_{j,k=1,2}$, where $D_{ijk} = r_{jk}\sqrt{w_i}$ and w_i are the heteroscedasticity weights. We assume $r_{11} = 1$, $r_{22} = 2$ and $r_{12} = r_{21} = \rho_e\sqrt{r_{11}r_{22}}$ with $\rho_e = 0.5$. Five scenarios are considered in this section based on heteroscedasticity and relation between regression variance and sampling variance.

Scenario 1: $w_i = 1$ representing homoscedastic model when sampling variances are smaller than regression variance.

Scenario 2: $w_i = 4$ representing homoscedastic model when sampling variances are the same as regression variance.

Scenario 3: $w_i = \max\limits_{1 \le j \le m}\left(\sqrt{x_{j1}^2 + x_{j2}^2}\right)$ representing homoscedastic model when sampling variances are larger than regression variance.

Scenario 4: $w_i = \sqrt{x_{i1}^2 + x_{i2}^2}$ representing heteroscedastic model when sampling variances vary according to regressors (González–Menteiga et al., 2008)

Scenario 5: $\boldsymbol{D}_i = \boldsymbol{L}_i\boldsymbol{L}_i^{'}$, where $\boldsymbol{L}_i = \left(L_{ijk}\right)_{j,k=1,2}$ with $L_{i11} = L_{i22} = \sqrt{\ell_n}$, $L_{i12} = 0$ and $L_{i21} = 0.5\sqrt{\ell_n}$. There are five groups G_t $(t = 1, \dots, 5)$, specifically, $\ell_n = 8.0$ if $n \in G_1$; $\ell_n = 4.0$ if $n \in G_2$; $\ell_n = 2.0$ if $n \in G_3$; $\ell_n = 1.0$ if $n \in G_4$; $\ell_n = 0.5$ if $n \in G_5$;. This case represents heteroscedastic model with different relations between sampling variances and regression variance (Li & Lahiri, 2010).

Different estimators in these five scenarios are compared using relative bias and mean square errors. The detailed steps of the simulation are as follows.

1. For each case of sampling covariance matrix, repeat $K = 10{,}000$ times $(k = 1, \dots, K)$
   (a). For each m = 5, 10, 20 and 50, generate $\left\{e_{ij}^{(k)}, u_{ij}^{(k)}, y_{ij}^{(k)}, \boldsymbol{x}_{ij}\right\}, i = 1, \dots, m, j = 1, 2$;
   (b). Calculate the variance estimator, $\hat{A}^{(k)}$ and EBLUP, $\widehat{\boldsymbol{\theta}}^{(k)}$ based on PML and APML methods;
2. Calculate the absolute bias of $\hat{A}$, $\frac{1}{K}\sum_{k=1}^{K}|\hat{A}^{(k)} - A|$, and mean squared error of $\hat{A}$, $\frac{1}{K}\sum_{k=1}^{K}\left(\hat{A}^{(k)} - A\right)^2$.
3. Calculate the average of absolute relative error ($\overline{\text{ARE}}$) and average of mean squared error ($\overline{\text{MSE}}$) of EBLUP as

$$\overline{\text{ARE}} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\left|\frac{\hat{\theta}_{ij}^{(k)} - \theta_{ij}^{(k)}}{\theta_{ij}^{(k)}}\right|, \qquad \overline{\text{MSE}} = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{K}\sum_{k=1}^{K}\left(\hat{\theta}_{ij}^{(k)} - \theta_{ij}^{(k)}\right)^2, \qquad j = 1,2.$$

Numerical results comparing the performance of the adjusted profile maximum likelihood estimates and the original profile maximum likelihood estimates are presented both for the variance component estimate $\hat{A}$ and the EBLUP estimate $\widehat{\boldsymbol{\theta}}$.

Data Description

The data used in this study is the average household income and average household expenditure data in Thailand from the Household Socio-Economic Survey 2017. The SES is conducted yearly by the National Statistical Office Thailand (NSO). The design sampling of SES is a stratified two-stage sampling. The SES is designed to produce estimates up to the provincial level. The total sample in SES 2017 is 43,210 households which are distributed in 5 regions including 77 provinces. Our study includes 76 provinces except Bangkok. Each province is divided into two parts according to the type of local administration area, namely, municipal area and non-municipal area.

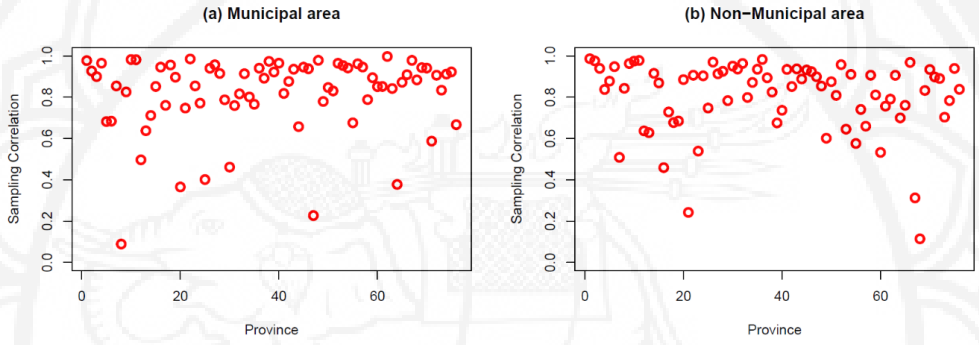

Figure 1 The sampling correlations of the average household income and average household expenditure

Figure 1 shows sample correlations of the average household income and average household expenditure in municipal and non-municipal areas. The correlations of two variables are generally close to 1. This suggests that the average household income and average household expenditure have high correlation. Thus, the bivariate model is more suitable than univariate model for this dataset.

Table 1 displays the means and standard deviations within group of the average household income and average household expenditure of SES 2017. The means of the average household incomes are higher than the means of the average household expenditures in all groups. For example, in municipal area of the central region, the mean of average household incomes, which is 31,229 Baht, is greater than the mean of average household expenditures, which is 23,147 Baht. The average household incomes and average household expenditures of municipal area are higher than the average household incomes and average household expenditures of non-municipal area in term of mean. For example, in central region, the average household income and average household expenditure of municipal area are 31,229 Baht and 23,147 Baht, respectively. The average household income and average household expenditure of non-municipal area are 27,230 Baht and 21,496 Baht, respectively.

Table 1 Sample size, mean and standard deviation of the average household incomes and average household expenditures of SES 2017

SES 2017	Region	Size	Mean		Standard Deviation	
			Municipal	Non-municipal	Municipal	Non-municipal
Average household incomes	Central	18	3.1229	2.7230	0.6693	0.6699
(Unit: 10,000 Baht)	East	7	2.9715	2.4834	0.3799	0.2609
	North	17	2.3638	1.7062	0.4822	0.2673
	Northeast	20	2.3727	1.8042	0.3414	0.3550
	South	14	2.9824	2.4504	0.7153	0.8221
	Total	**76**	**2.7158**	**2.1815**	**0.6321**	**0.6731**
Average household expenditures	Central	18	2.3147	2.1496	0.5168	0.5613
(Unit: 10,000 Baht)	East	7	2.2101	2.0382	0.1766	0.2323
	North	17	1.7734	1.4013	0.3014	0.2293
	Northeast	20	1.8855	1.5568	0.2631	0.2566
	South	14	2.3561	1.9602	0.4689	0.5073
	Total	**76**	**2.0787**	**1.7811**	**0.4455**	**0.4890**

For the corresponding area-specific explanatory variables, we use four explanatory variables selected from the AIC forward selection method. These variables are proportion of households that cement or brick dwellings (x_1); proportion of households that own land (x_2); proportion of households using gas for cooking (x_3); and average population per private household (x_4). These four variables are administrative data from the Population and Housing Census 2010.

In our study, we apply model (4) with the two direct estimators of income and expenditure (y_1, y_2) and the four explanatory variables $x_1, x_2, x_3,$ and x_4. The variance component A is then estimated by the profile maximum likelihood (PML) and the adjusted profile maximum likelihood (APML) using the **optim** function in R (R Core Team, 2019). The study is divided into 10 small studies based on Region and municipality.

Results

In this section, we present simulation results and data application comparing our adjusted profile maximum likelihood estimate to the original profile maximum likelihood estimate.

Simulation Results

Tables 2 - 4 display percentage of zero estimates, absolute bias, and mean square error of estimates of $\hat{A}$, respectively. Tables 5 - 6 display the average of absolute relative errors and the average of mean squared errors of $\hat{\boldsymbol{\theta}}$.

Table 2 The percentage of zero estimates of A for $m = 5, 10, 20, 50$

Sample size	5		10		20		50	
Method	PML	APML	PML	APML	PML	APML	PML	APML
Scenario 1	24.28	0	2.81	0	0.05	0	0	0
Scenario 2	43.07	0	12.64	0	1.30	0	0	0
Scenario 3	59.16	0	31.98	0	11.17	0	0.84	0
Scenario 4	58.26	0	28.88	0	8.99	0	0.43	0
Scenario 5	73.89	0	14.87	0	1.14	0	0	0

From Table 2, we can see that the percentages of zero estimates of PMLs are very high in the cases of small sample sizes ($m = 5,10$). Results from scenarios 1 – 3 suggest that the percentages of zero estimates of PMLs are higher when sampling variances are large comparing to the regression variance. Considering heteroscedastic models in scenarios 4 and 5, we can see that percentages of zero estimates are very high particularly for small sample sizes. For all scenarios, the adjusted profile maximum likelihood method can prevent the zero estimate of A regardless of sample sizes and sampling variances.

Table 3 The absolute bias of different estimators of A for $m = 5, 10, 20, 50$

Sample size	5		10		20		50	
Method	PML	APML	PML	APML	PML	APML	PML	APML
Scenario 1	1.3479	1.0113	0.8984	0.7755	0.6031	0.5571	0.3724	0.3618
Scenario 2	1.5314	1.1109	1.1182	0.9199	0.7799	0.6915	0.4826	0.4621
Scenario 3	1.7126	1.5245	1.4252	1.2526	1.0898	0.9298	0.6965	0.6430
Scenario 4	1.7003	1.4747	1.3733	1.1754	1.0351	0.8837	0.6606	0.6137
Scenario 5	1.7575	1.2156	1.1075	0.9092	0.7211	0.6521	0.4292	0.4126

Table 4 The mean squared error of different estimators of A for $m = 5, 10, 20, 50$

Sample size	5		10		20		50	
Method	PML	APML	PML	APML	PML	APML	PML	APML
Scenario 1	2.2351	1.6023	1.1300	0.9401	0.5400	0.4844	0.2129	0.2054
Scenario 2	2.7990	2.4684	1.7063	1.4430	0.8964	0.7671	0.3569	0.3377
Scenario 3	3.4653	5.7724	2.6657	3.2083	1.6777	1.5364	0.7385	0.6680
Scenario 4	3.4135	5.3920	2.4847	2.7510	1.5267	1.3585	0.6658	0.6050
Scenario 5	3.4247	2.5392	1.6613	1.3011	0.7637	0.6605	0.2829	0.2682

From Tables 3 – 4, we can see that absolute bias and mean squared error decrease when sample size increases, or equivalently when sampling error covariance decreases. The absolute biases of the adjusted profile maximum likelihood method are less than those of profile maximum likelihood method for all cases of sampling error covariance matrix and for all cases of m. The mean squared errors of the adjusted profile maximum likelihood method is less than those of profile maximum likelihood method for all case of sampling error covariance and for all cases of m, except the case when $m = 5$ or 10 in scenarios 3 and 4.

Table 5 The average of absolute relative errors of different methods of EBLUPs $\widehat{\boldsymbol{\theta}}$ for $m = 5, 10, 20, 50$

	Parameter	θ_1				θ_2			
	Sample size	5	10	20	50	5	10	20	50
Scenario 1	PML	0.0415	0.0367	0.0343	0.0324	0.0543	0.0475	0.0439	0.0410
	APML	0.0389	0.0358	0.0340	0.0324	0.0520	0.0466	0.0436	0.0410
Scenario 2	PML	0.0542	0.0476	0.0435	0.0403	0.0695	0.0592	0.0531	0.0484
	APML	0.0515	0.0459	0.0429	0.0402	0.0678	0.0580	0.0526	0.0483
Scenario 3	PML	0.0682	0.0602	0.0540	0.0485	0.0878	0.0735	0.0635	0.0557
	APML	0.0668	0.0586	0.0528	0.0482	0.0876	0.0729	0.0628	0.0555
Scenario 4	PML	0.0671	0.0581	0.0524	0.0475	0.0864	0.0711	0.0619	0.0547
	APML	0.0656	0.0564	0.0512	0.0472	0.0861	0.0703	0.0612	0.0545
Scenario 5	PML	0.0562	0.0470	0.0423	0.0397	0.0595	0.0501	0.0454	0.0420
	APML	0.0511	0.0451	0.0423	0.0396	0.0550	0.0483	0.0450	0.0419

Table 6 The average of mean squared errors of different methods of EBLUPs $\hat{\boldsymbol{\theta}}$ for $m = 5, 10, 20, 50$

	Parameter	θ_1				θ_2			
	Sample size	5	10	20	50	5	10	20	50
Scenario 1	PML	0.9089	0.7524	0.6719	0.6325	1.5475	1.2590	1.0986	1.0096
	APML	0.7963	0.7136	0.6631	0.6314	1.4168	1.2099	1.0866	1.0078
Scenario 2	PML	1.5478	1.2679	1.0841	0.9747	2.5426	1.9634	1.6139	1.4062
	APML	1.3973	1.1762	1.0515	0.9700	2.4215	1.8859	1.5856	1.4017
Scenario 3	PML	2.4549	2.0307	1.6688	1.4152	4.0761	3.0499	2.3151	1.8619
	APML	2.3601	1.9249	1.5966	1.3959	4.0582	3.0023	2.2729	1.8495
Scenario 4	PML	2.3821	1.9002	1.5764	1.3544	3.9473	2.8567	2.2027	1.7994
	APML	2.2796	1.7888	1.5084	1.3379	3.9193	2.7976	2.1606	1.7883
Scenario 5	PML	1.9295	1.3830	1.1270	1.0111	2.1285	1.5483	1.2570	1.1185
	APML	1.5742	1.2751	1.1046	1.0086	1.8003	1.4413	1.2344	1.1158

From Tables 5 – 6, we can see that average of absolute relative errors and the average of mean squared errors of the adjusted profile maximum likelihood method are less than that of the profile maximum likelihood method for all cases of sampling error covariances and for all cases of m. For example, considering the PML estimates for $m = 5$ in scenario 1, the average of absolute relative errors and the average of mean squared errors of EBLUP for average household income are 0.0415 and 0.9089, respectively. The average of absolute relative errors and the average of mean squared errors of EBLUP for average household expenditure are 0.0543 and 1.5475, respectively. The average of absolute relative errors and the average of mean squared errors decrease when sample size increases, or equivalently when sampling error covariances decrease.

Data analysis

In this section, we apply the adjusted profile maximum likelihood estimation method for bivariate Fay-Herriot model to study income and expenditure in Thailand. Table 7 displays the profile maximum likelihood estimate and adjusted profile maximum likelihood estimate of A.

Table 7 The estimates of A for different methods

Region		Municipal area		Non-municipal area	
	Sample size	PML	APML	PML	APML
Central	18	0.0075	0.0123	0	0.0042
East	7	0	0.0027	0	0.0015
North	17	0.0258	0.0343	0.0055	0.0079
Northeast	20	0.0124	0.0161	0.0144	0.0165
South	14	0.0082	0.0163	0	0.0229

From Table 7, we can see that the profile maximum likelihood estimates of regression variance A are zeros in some cases. This situation occurs particularly in the east regions (both municipal area and non-municipal area) when sample sizes are small. Moreover, profile maximum likelihood estimates are zeros for some cases with large sample size such as in non-municipal area of the central and the south regions. In all cases, our adjusted profile maximum likelihood estimates prevent zero estimates. For more illustrations, we demonstrate three examples of our results. First, Figure 2 presents the case where sample variances are relatively large and sample size is small which is the case when $m = 7$. In this case, the naive empirical maximum likelihood

estimate $\hat{A}$ is zero and it is not precise since sample size is small. The adjusted profile maximum likelihood estimate gives non-zero estimate. However, the estimate is also small since the sampling variances are large. The direct estimates are not reliable. Therefore, EBLUP estimates give small weight on direct estimates and large weight on the regression estimates. Second, Figure 3 presents the case where sample variances are smaller and sample size is medium $(m = 14)$. In this case, the naive profile maximum likelihood estimate gives zero estimate $\hat{A}$ while the adjusted profile maximum likelihood estimate gives a non-zero estimate. The weight of the direct estimate is higher than the previous case since sampling variances are smaller in this case. Therefore, we can see from the figure that the EBLUPs lie between the direct estimates and the regression estimates according to (7). Third, Figure 4 presents the case where sampling variances are smaller and sample size is larger $(m = 20)$ than in case two. In this case, the two estimates perform similarly and give non-zero weight on the direct estimates. The two EBLUP estimates lie between the direct estimates and the regression estimates.

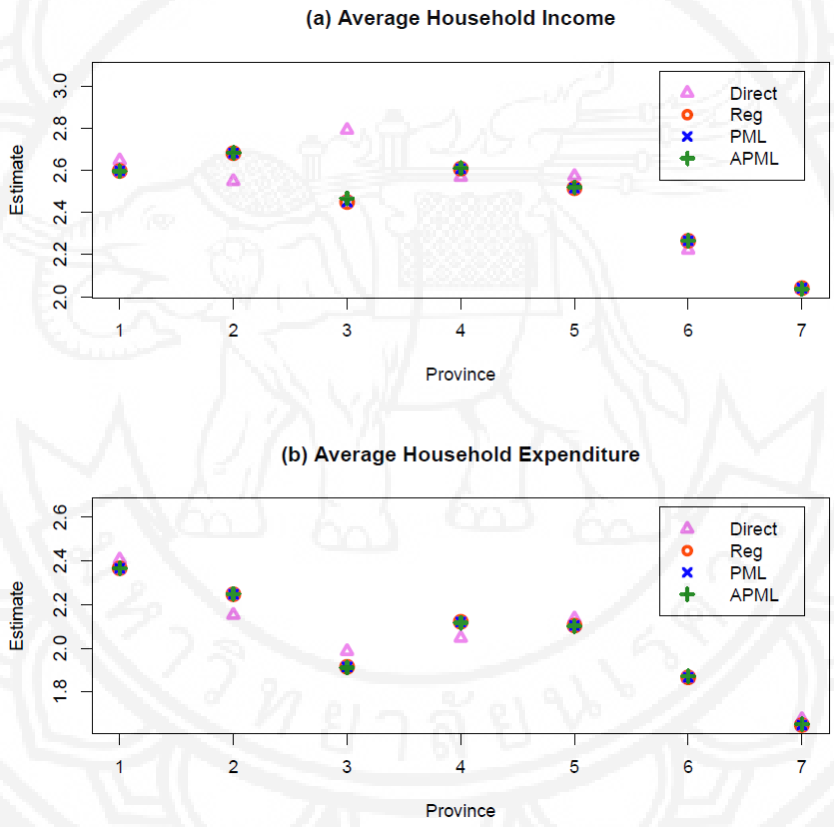

Figure 2 The estimates of the average incomes (a) and average expenditures (b) in non-municipal area of east region

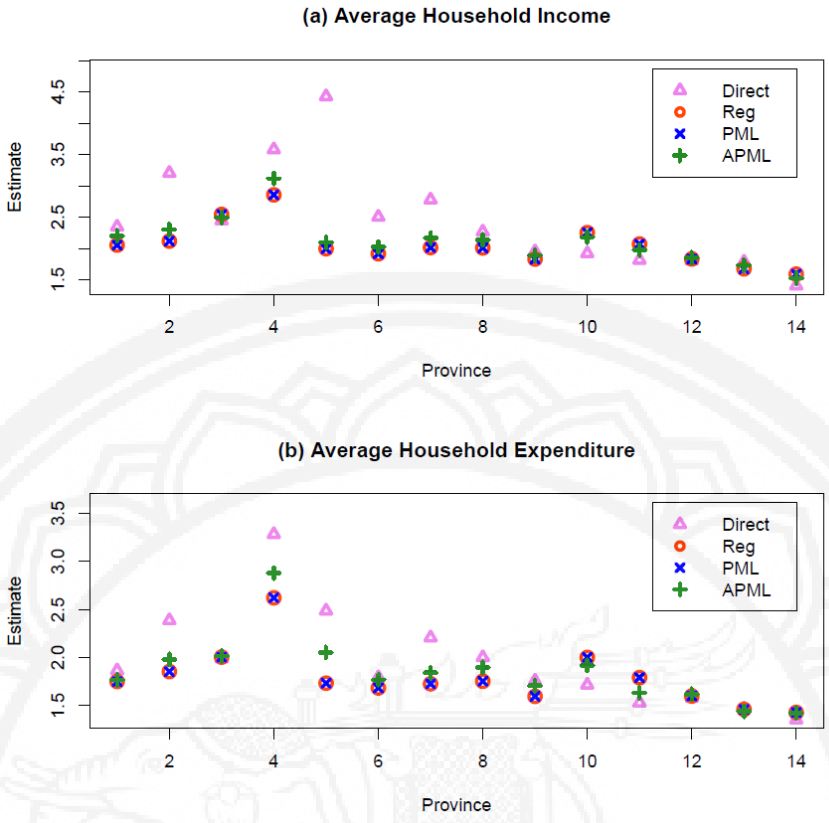

Figure 3 The estimates of the average incomes (a) and average expenditures (b) in non-municipal area of south region

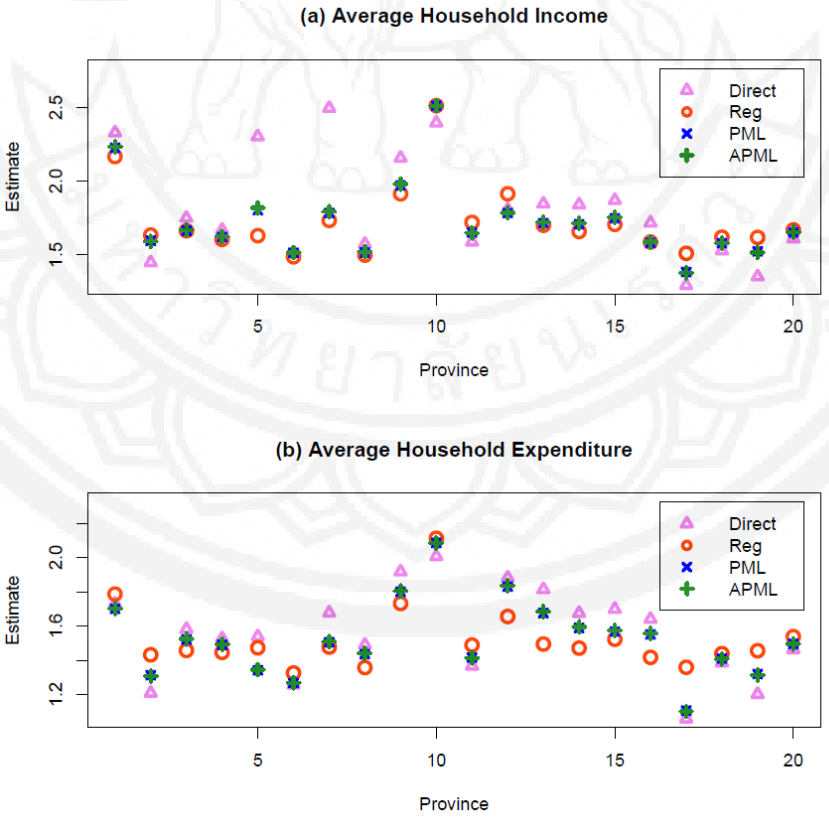

Figure 4 The estimates of the average incomes (a) and average expenditures (b) in non-municipal area of northeast region

Table 7 and Figures 2 – 4 suggest that our adjusted profile maximum likelihood estimates improve the naive profile maximum likelihood estimates.

For the rest of this paper, we apply our adjusted profile maximum likelihood estimates of the regression variance for bivariate Fay–Herriot model to produce EBLUP estimates of the average household incomes and average household expenditures. Table 8 displays aggregated mean and standard deviation of the EBLUP estimates using the adjusted profile maximum likelihood estimate of the regression variance. The statistics are presented at Region × Municipality levels. From Table 8, we see that the means of average household incomes are generally higher than the means of average household expenditures. For example, in municipal area of the central region, the mean of average household incomes, which is 27,665 Baht, is greater than the mean of average household expenditures, which is 22,053 Baht. The average household income and average household expenditure of municipal area are higher than the corresponding average household incomes and average household expenditures of non–municipal area in term of mean. For example, the average household income and average household expenditure of municipal area in east region are 28,493 Baht and 21,980 Baht, respectively. The corresponding average household income and average household expenditure of non–municipal area of the east region are 24,534 Baht and 20,385 Baht, respectively.

Table 8 Sample size, mean and standard deviation of the EBLUP of average household incomes and average household expenditures

SES 2017	Regions	Size	Mean		Standard Deviation	
			Municipal	Non–municipal	Municipal	Non–municipal
Average household incomes (Unit: 10,000 Baht)	Central	18	2.7665	2.5160	0.4774	0.5112
	East	7	2.8493	2.4534	0.2042	0.2279
	North	17	2.2072	1.6450	0.3337	0.2037
	Northeast	20	2.2570	1.7297	0.2820	0.2602
	South	14	2.7241	2.1259	0.4238	0.3766
	Total	**76**	**2.5071**	**2.0366**	**0.4505**	**0.4967**
Average household expenditures (Unit: 10,000 Baht)	Central	18	2.2053	2.0534	0.4130	0.4590
	East	7	2.1980	2.0385	0.1926	0.2432
	North	17	1.7096	1.3541	0.2126	0.1829
	Northeast	20	1.8378	1.5236	0.2395	0.2242
	South	14	2.2567	1.8548	0.3477	0.3547
	Total	**76**	**2.0065**	**1.7196**	**0.3727**	**0.4175**

Discussion

In this study, we have applied the bivariate Fay–Herriot model to study income and expenditure at provincial level of Thailand. The bivariate model was used due to the correlation found in the survey presented in Figure 1. In our analysis, we have developed an extension of variance estimation by an adjusted profile maximum likelihood estimate to the bivariate Fay–Herriot model.

Conclusion and Suggestions

The simulation results suggest that our adjusted profile maximum likelihood estimate prevents zero estimate of regression variance A. Consequently, using the obtained adjusted profile maximum likelihood estimate, we can obtain better EBLUP estimates of the population mean(θ). Further investigation on real data was also performed using the Thai Household Socio-Economic data. The results showed that our adjusted profile maximum likelihood estimate outperforms the naive profile maximum likelihood estimates. Several extensions of our study can be considered. For example, investigating the performance of the adjusted maximum likelihood method to other forms of likelihood function such as residual likelihood. Alternatively, we can consider an extension of the method to more general multivariate models.

Acknowledgments

The first author would like to thank fund from the Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, and the His Royal Highness Crown Prince Maha Vajiralongkorn Scholarship from the Graduate School, Chulalongkorn University to commemorate the 72nd anniversary of his Majesty King Bhumibol Aduladej is gratefully acknowledge.

References

Benavent, R., & Morales, D. (2016). Multivariate Fay-Herriot models for small area estimation. *Computational Statistics and Data Analysis, 94,* 372-390. https://doi.org/10.1016/j.csda.2015.07.013

Datta, G. S., Fay, R. E., & Ghosh, M. (1991). Hierarchical and empirical Bayes multivariate analysis in small are estimation. In *Proceedings of Bureau of the Census 1991 Annual Research Conference, 17-20 March 1991* (pp. 63-79). Washington DC, U.S.: Bureau of the Census.

Datta, G. S., Ghosh, M., Nangia, N., & Natarajan, K. (1996). Estimation of median income of four-person families: a bayesian approach. *Journal of the American Statistical Association, 91*(436), 1423-1431.

Fay, R. E. (1987). Application of multivariate regression of small domain estimation. In R. Platek, J. N. K. Rao, C. E. Särndal, M. P. Singh (Eds.), *Small Area Statistics,* (pp. 91-102). New York: Wiley.

Fay, R. E., & Herriot, R. A. (1979). Estimates of income for small places: An application of James-Stein procedure to census data. *Journal of the American Statistical Association, 74*(366), 269- 277. https://doi.org/10.1080/01621459.1979.10482505

González-Menteiga, W., Lombardía, M. J., Molina, I., Morales, D., & Santamaría, L. (2008). Analytic and bootstrap approximations of prediction errors under a multivariate Fay-Herriot model. *Computational Statistics and Data Analysis, 52*(12), 5242-5252. https://doi.org/10.1016/j.csda.2008.04.031

Hartley, H. O., & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, *54*(1/2), 93-108. https://doi.org/10.2307/2333854

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under selection model. *Biometrics*, *31*(2), 423-427. https://doi.org/10.2307/2529430

Li, H., & Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis, 101*(4), 882-892. https://doi.org/10.1016/j.jmva.2009.10.009

R Core Team. (2019). *R: A language and environment for statistical computing [Computer software manual].* Retrieved from https://www.R-project.org/

Rao, J. N. K., & Molina, I. (2015). *Small area estimation* (2nd ed.). New York: John Wiley and Sons.