© 2018 IJSRSET | Volume 4 | Issue 9 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section : Engineering and Technology

Logit Transformation on Unit Level Small Area Estimation with First Order Autoregressive Time Effects

Dian Surida[1], Anang Kurnia[1], Siti Muchlisoh[2]
*[1]Department of Statistics, Bogor Agricultural University, West Java, Indonesia
[2]Politeknik Statistika STIS, Jakarta,Indonesia

ABSTRACT

Indonesian labor force participation data are collected by Sakernas (National Labor Force Survey). The main purpose of Sakernas is to obtain information about the unemployment rate and its changes over time. Model-2 is the Rao-Yu unit model (Model-1) with modeling was done by logit transformation approach. The empirical study using Sakernas rotational panel data for 2011-2014 shows that the Empirical Best Linear Unbiased Prediction (EBLUP) of Model-2 gives a lower MSE compared to Model-1.

Keywords: EBLUP, Area Unit Level, Rao-Yu Model, Unemployment Rate, Logit Transformation

I. INTRODUCTION

Employment data are collected by the Statistics Indonesia (BPS) through Sakernas [2]. Sakernas data are collected quarterly under rotating panel survey. Based on Sakernas design, the number of samples was only adequate for estimating parameters at provincial and national levels beside information on employment at lower levels of the province (district/city) and changes that can be monitored in a relative time short is necessary for the line with the development of regional autonomy. The sample size at the regional level is usually very small so the statistics obtained will have a large variance. The estimation result cannot be done if the area is not selected to be an example in the survey. Therefore, a method of parameter estimation is developed that can overcome this. The method is known as the Small Area Estimation (SAE) method.

The Fay-Herriot model [4] is an SAE model based on area level estimation. This model can only be applied to cross-section data. One of SAE models for panel data is the Rao-Yu model as an extension of the basic Fay-Herriot model by adding a random area-time component which follows an autoregressive process order-1. However, the problem faced in applying the SAE level area is the assumption about the availability of variance of the sampling error, because in reality, the assumption is difficult to meet. To satisfy this assumption the variance of sampling errors is substituted by the estimated value. Therefore, to solve the problem, Muchlisoh [13] developed Rao-Yu model of area level to model Rao-Yu unit level.

Muchlisoh [13] developed the Rao-Yu model so it can be used in the unit level analysis. This model is used to estimate the unemployment rate using Sakernas data from 2011-2014. The results of the study indicate that the estimation using this model produce the estimated values that are commensurate with the Rao-Yu model and better than direct estimation. Therefore, This model can be used as an alternative in SAE when the variance of sampling errors is not available. However, the proportion of unemployment in this study is assumed to follow the Normal distribution. In fact, the Unemployed is a binary event with two possibilities, ie unemployed or works.

Binary responses with successful opportunities (unemployment) and upper limit (n of the labor force) generally follow the Binomial Distribution. Thus, the proportion of unemployment is assumed to follow the Binomial Distribution. Therefore, this study aims to improve the model developed by Muchlisoh [13] by assuming that the proportion of unemployment is assumed to follow the Binomial Distribution. So that the proportion of unemployment will be transformed by logit transformation.

II. RAO-YU MODEL AND ITS MODIFICATION

The Rao-Yu model has a sampling model for the direct survey estimation and a linking model for the small area parameters of interest [9]. The sampling model assumes that there exists a direct survey estimator $\hat{\theta}_{it}$ which is usually design-unbiased, for the small area parameter θ_{it}, such that

$$\hat{\theta}_{it} = \theta_{it} + \mathrm{e}_{it}, \qquad (2.1)$$

$i = 1, \dots, m$; $t = 1, \dots, T$. where m is the number of small areas, t is the number of times, and the e_{it} is the sampling error. It is customary to assume that error $\mathrm{e}_{it}'s$ are independently normal random variables $e_{it} \sim iidN(0, \sigma_{it}^2)$. The linking model for θ_{it} is given as

$$\theta_{it} = \boldsymbol{x}_{it}^T\boldsymbol{\beta} + v_i + u_{it}, \qquad (2.2)$$
$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, |\rho| < 1$$

where $\boldsymbol{x}_{it} = (x_{ij1}, \dots, x_{ijp})^T$ is area-time auxiliary variable size $p\ x$ 1which is assumed to be available in the i-th small areas and time t. $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$is a column vector of the p x 1 regression coefficient. Furthermore, v_i is an area-specific random effect with $v_i \sim iid\ N(0, \sigma_v^2)$ and u_{it} is a time-specific random effect with $\varepsilon_{it} \sim iid\ N(0, \sigma_\varepsilon^2)$, that is assumed to follow a first-order autoregressive process within each area. The constant ρ is the autoregressive coefficient. The combination of the sampling error model (2.1) and the linking model (2.2) will produce the following model:

$$\hat{\theta}_{it} = \boldsymbol{x}_{it}^T\boldsymbol{\beta} + b_i v_i + u_{it} + \mathrm{e}_{it}, \qquad (2.3)$$
$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, |\rho| < 1$$

The random effect $v_i, u_{it},$and e_{it} are assumed mutually independent. The condition $|\rho| < 1$ ensures stationarity of the series defined by (2.3) in order to obtain the first order autoregressive process.

Muchlisoh [13] modified the Rao-Yu model area level to make the Rao-Yu Model is suitable for analysis at the unit level when the sampling error not available. Let y_{itj} is *j*-th sample unit of *i*-th small area at time t. The Rao-Yu unit model is defined as,

$$y_{itj} = \boldsymbol{x}_{itj}^T\boldsymbol{\beta} + v_i + u_{it} + e_{itj}, \qquad (2.4)$$
$$u_{it} = \rho u_{i,t-1} + \varepsilon_{it}, |\rho| < 1,$$

Assume that unit specific auxiliary variable x_{itj} are available for each population element j in *i*-th small area at time t. where $i = 1,2, \dots, m$ is the index for a small area, $t = 1,2, \dots, T$ is the index for time and j = $j = 1,2, \dots, \eta_{it}$ with η_{it} is the number of sample units in a small area *i*-th at time t.

EBLUP for $\hat{\theta}_{it}$ when ρ is known $\hat{\theta}_{it(\rho)}$ obtained by subtituted $\widetilde{\boldsymbol{\beta}}$, $\widetilde{\boldsymbol{v}}_i$, and $\widetilde{\boldsymbol{u}}_{it}$ with $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{v}}_{i(\rho)}$, and $\widehat{\boldsymbol{u}}_{it(\rho)}$, respectively, that is

$$\hat{\theta}_{it(\rho)} = f_{it}\bar{y}_{it}^s + \left(\left(\overline{\boldsymbol{X}}_{it}^p\right)' - f_{it}(\overline{\boldsymbol{X}}_{it}^s)'\right)\widehat{\boldsymbol{\beta}}$$
$$+(1 - f_{it})\left(\widehat{\boldsymbol{v}}_{i(\rho)} + \widehat{\boldsymbol{u}}_{it(\rho)}\right) \qquad (2.5)$$

with

$$\widehat{\boldsymbol{v}}_i = \hat{\sigma}_v^2 \mathbf{1}_{\eta_i}'\left[\hat{\sigma}_v^2 \boldsymbol{J}_{\eta_i} + \hat{\sigma}_\varepsilon^2\left(\boldsymbol{I}_T \otimes \boldsymbol{J}_{\eta_{it}}\right) + \hat{\sigma}_e^2 \boldsymbol{I}_{\eta_i}\right]^{-1}$$
$$(\boldsymbol{y}_i^* - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})$$

and

$$\widehat{\boldsymbol{u}}_{it} = \hat{\sigma}_u^2 \boldsymbol{I}_T(\boldsymbol{I}_T \otimes \mathbf{1}_{\eta_{it}})'\left[\hat{\sigma}_v^2 \boldsymbol{J}_{\eta_i} + \hat{\sigma}_u^2(\boldsymbol{I}_T \otimes \boldsymbol{J}_{\eta_{it}})\right.$$
$$\left. + \hat{\sigma}_e^2 \boldsymbol{I}_{\eta_i}\right]^{-1}(\boldsymbol{y}_i^* - \boldsymbol{X}_i\widehat{\boldsymbol{\beta}})$$

$\eta_i = \sum_{t=1}^{T} \eta_{it}$. $\boldsymbol{J}_{\eta_i}$ and $\boldsymbol{J}_{\eta_{it}}$ are matrices of order η_i x η_i and η_{it} x η_{it} with elements 1, respectively. $\mathbf{1}_{\eta_i}$ is vectors size η_i with all the element 1 and $\mathbf{1}_{\eta_{it}}$ are vectors size η_{it} with all the element 1. $\boldsymbol{I}_{\eta_{it}}$ is identity matrix size η_{it} x η_{it}.

III. LOGIT TRANSFORMATION

One approach to modeling the variables of interest of value 0 and 1 is used the logit transformation [1]. Let n as the total individual, r is the number of successes. For the small area $(i = 1,2,\dots,m)$ with $p_i = {}^{r_i}/_{n_i}$ is the proportion of success, then the logit transform is defined as

$$y_i = logit(p_i) = ln\left[\frac{p_i}{1-p_i}\right] = ln\left[\frac{r_i}{n_i - r_i}\right] \qquad (3.1)$$

The logit transformation produces new variables with values that are in the interval $(-\infty,\infty)$. If p_i is 0 or 1, then the logit transformation will produce an infinite value. Therefore, Cox and Sneill [3] make the following modifications

$$y_i = ln\left[\frac{r_i + {}^1/_2}{n_i - r_i + {}^1/_2}\right] \qquad (3.2)$$

IV. EMPIRICAL STUDY

This section demonstrates the performance of the Rao-Yu unit model to estimate the unemployment rate for district level using West Java Sakernas panel rotation data for 2011-2014. The number of Sakernas samples of each quarter of the province of West Java is 400 census blocks or about 4000 households. The sample was allocated proportionally to 26 districts. The total samples for each district are contained of about 8-21 census blocks. The total samples of census blocks are separated into four sample packages (1,2,3,4). Every package, at each census block, consists of about 10 households. The auxiliary data come from The 2011 Village Potential Sensus. The auxiliary variables used are the ratio between the number of hotels to the population and the ratio of the number of industries to the population. Both variables are assumed to affect the employment absorption that will ultimately affect the unemployment rate.

Empirical study start from preparing the variable of interest y_{itj} and the auxiliary variables x_{itj} based on Sakernas panel rotation data 2011-2014. $M = 26$ is the index for a small area as many districts/cities as there are in West Java, $T = 13$ is the index for the time that is the number of quarters from the year 2011 quarter I to 2014 quarter I, and $\eta_{it} = 4$ is the number of census block packages (units) in the *i*-th small area at time t. So the number of units in this study is 1352 units. y_{itj} is the proportion of unemployement in *i*-th area *j*-th unit at time point t. y_{itj} is in the interval 0 and 1 so it needs to be transformed based on equation (3.2) to generate new variables that are in the interval $(-\infty,\infty)$. Direct estimation is done by simple random sampling with replacement then selected 4 samples of units in each district/city at any time so that y_{itj} is obtained. Then the average of direct estimates in *i*-th district/city and *t*-th time are computed, that is $\bar{y}_{it} = \frac{1}{n_{it}}\sum_{j=1}^{n_{it}} y_{itj}$. EBLUP for θ_{it} is calculated using equation (2.5). MSE estimation was treated by the bootstrap method with 150 replications (R=150). From each replication, the estimator of direct estimation and Rao-Yu unit model were computed. MSE estimated values by bootsrap method of any estimator, say $\hat{\theta}_{it}$, were computed as follows

$$MSE = \frac{1}{MT}\sum_{i=1}^{M}\sum_{t=1}^{T}\left(\frac{1}{R}\sum_{j=1}^{R}\left(\hat{\theta}_{itl} - \theta_{it}\right)^2\right)$$

In this study, the estimation is done by 3 models of prediction, i.e. direct estimation, Rao-Yu model unit that assumed y_{itj} is normally distributed (Model-1), and Rao-Yu model unit by performing logit transformation on y_{itj} (Model-2).

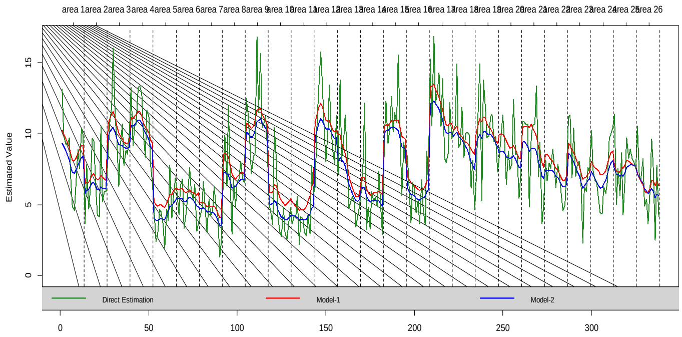

Figure 1. The EBLUP of the quarterly unemployment rate in the district/city level in West Java based on Sakernas rotation panel data from 2011 to 2014 using direct estimation (green), Model-1 (red) and Model-2 (blue).

Figure 1 is a graph of the comparison of the quarterly unemployment rate in the district/city level in West Java based on Sakernas rotation panel data for 2011-2014 using direct estimation, Model-1, and Model-2. Figure 1 shows that direct estimation yield more fluctuates and unstable estimated values compared to Model-1 and Model-2 EBLUP. This is because direct estimation is based solely on survey data and the component of variance is excluded from determining the estimated value of variables of interest.

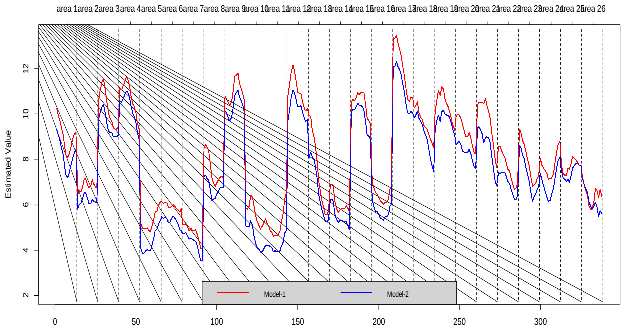

Figure 2. The EBLUP of the quarterly unemployment rate in the district/city level in West Java based on Sakernas rotation panel data from 2011 to 2014 using Model-1 (red) and Model-2 (blue).

Furthermore, Figure 2 is given to see more clearly the EBLUP comparison between Model-1 and Model-2. Figure 2 explains that Model-2 gives an always smaller EBLUP than Model-1. Determination of the best model that produces better-estimated value in estimating quarterly unemployment rate at district/city level in West Java based on Sakernas rotational panel data of the year 2011-2014 done by comparing MSE from each estimation. A model is better if the MSE is smaller than the MSE of other models. The statistical comparison of the predicted value of bootstrap MSE based on direct estimation, Model-1, and Model-2 are presented in Table 1.

Table 1. Statistical comparison of bootstrap MSE estimation to estimates quarterly unemployment rate at the district/city level in West Java based on Sakernas rotation panel data for 2011-2014 using Direct Estimation, Model-1, and Model-2.

MSE	Direct Estimation	Model-1	Model-2
Minimum	0.015	0.002	0.068
Mean	3.996	2.312	2.243
Median	2.737	1.229	1.739
Standard Deviation	4.447	2.928	1.924
Maximum	38.032	17.222	12.131

Based on Table 1, MSE on the direct estimation of 3.996 drops to 2.243 and 2.312 in the estimation method with Model-1 and Model-2, respectively. This suggests that the addition of random effects serves to calibrate the results of direct estimates based solely on survey data alone. As described above, the decrease in MSE is due to the decomposition of the component of variance present in the model in both Model-1 and Model-2. In addition, although the predicted value of MSE generated by Model-1 and Model-2 is not very different, the MSE Model-1 of 2.312 decreased to 2.243. This shows that the logit transformation performed on the variable of interest in Model-1 can improve the efficiency of the Model-1. For a clearer view of MSE comparisons based on direct estimation, Models 1, and Model 2 can be seen in Figure 3.

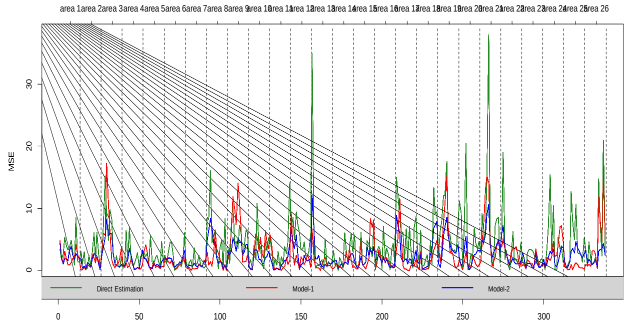

Figure 3. The comparison of MSE bootstrap estimation of the quarterly unemployment rate in the district/city level in West Java based on Sakernas rotation panel data for 2011-2014 using direct estimation (green), Model-1 (red) and Model-2 (blue).

In Figure 3, direct estimation gives the larger MSE value compared to Model-1 and Model-2. Overall, there was a significant reduction of MSE values when using Model-1 and Model-2. While to see more clearly the MSE bootstrap comparison of Model-1 and Model-2 can be seen in Figure 4. Based on Figure 4, the MSE with the bootstrap method in Model-2 tends to have a smaller value than Model-1. This indicates that EBLUP Model-2 is better than Model-1 in estimates the parameters.

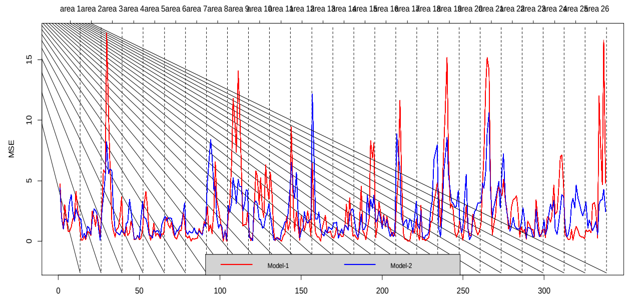

Figure 4. The comparison of MSE estimation of the quarterly unemployment rate in the district/city level in West Java based on Sakernas rotation panel data for 2011-2014 using Model-1 (red) and Model-2 (blue).

V. CONCLUSION

The estimation of the quarterly unemployment rate in the district/city level in West Java based on Sakernas rotational panel data for 2011-2014 using Rao-Yu unit model with transformation logit in the variable of interest gives a more stable estimated value compared to Model-1. The MSE estimation by the bootstrap method in Model-2 tends to have a smaller value than Model-1. This indicates that EBLUP Model-2 is better than Model-1 in estimates the parameters.

VI. ACKNOWLEDGEMENTS

The author would like to thank to the Indonesia Endowment Fund for Education (LPDP) who has provided “Indonesian Education Scholarship 2016” in the Department of Statistics FMIPA Bogor Agricultural University Indonesia.

VII. REFERENCES

[1]. Baum CF. 2008. Modelling Proportion. The Stata Journal. 2:299-303

[2]. [BPS] Badan Pusat Statistik. 2014. Pedoman pengawas survei angkatan kerja nasional (Sakernas) triwulanan. Jakarta (ID): Badan Pusat Statistik Republik Indonesia.

[3]. Cox DR, Snell EJ. 2002. Analysis of Binary Data Second Edition. USA: Chapman and Hall.

[4]. Fay RE and Herriot RA. 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.

[5]. Pfeffermann D, Terryn B, Fernando M. 2008. Small Area Estimation under a Two Part Random Effects Model to Estimation if Literacy in Developing Countries. Survey Methodology 34: 67-72.

[6]. Friedman EM, Jang D, Williams TV. 2002. Combined Estimates From Four Quarterly Survey Data Sets. Proceedings of the Section on Survey Research Methods. American Statistical Association. 1064-1069.

[7]. Ghosh, M. dan Rao, J.N.K. 1994. Small area estimation: an appraisal. Statistical Sciences 9, 55-93.

[8]. Hidiroglou M. 2007. Small-area estimation: theory and practice. Section on Survey Research Methods, 3445–3456.

[9]. J.N.K. Rao and M.Yu, Small area estimation combining time series and cross-sectional data, The Canadian Journal of Statistics, 22 (1994), 511-528.

[10]. Kurnia A, Notodiputro KA. 2007. Generalized additive mixed model for small area estimation. 2nd. International Conference on Mathematical Sciences 2007 (IcoMS-2007). Universiti Teknologi Malaysia.

[11]. Levy, P.S. dan Lemeshow, S. 1999. Sampling of Populations, Methods and Applications. New York: John Wiley and Sons, Inc.

[12]. Longford NT. 2007. On standard errors of model-based small-area estimators. Survey Methodology. 33: 69-79.

[13]. Muchlisoh S. 2017. Small Area Estimation of Unemployement Rate Based on Unit Level Model with First Order Autoregressive Time Effects. Journal of Applied Probability and Statistics 12(2) :65-77

[14]. Rao JNK. 2003. Small Area Estimation. New York: John Wiley and Sons, Inc.