APPLICATION OF BINOMIAL REGRESSION IN SMALL AREA ESTIMATION METHOD FOR ESTIMATING PROPORTION OF CULTURAL INDICATOR*

Yudistira[1], Anang Kurnia[2‡], and Agus Mohamad Soleh[3]

[1]Directorate General of Culture, Ministry of Education and Culture (Kemendikbud), Indonesia, aritsiduy.ui.0301@gmail.com
[2]Department of Statistics, Bogor Agricultural University (IPB), Indonesia, anangk@apps.ipb.ac.id
[3]Department of Statistics, Bogor Agricultural University (IPB), Indonesia, agusms@apps.ipb.ac.id
‡corresponding author

Indonesian Journal of Statistics and Its Applications (eISSN:2599-0802)
Vol 2 No 2 (2018), 56 - 62

Abstract

In sampling survey, it was necessary to have sufficient sample size in order to get accurate direct estimator about parameter, but there are many difficulties to fulfill them in practice. Small Area Estimation (SAE) is one of alternative methods to estimate parameter when sample size is not adequate. This method has been widely applied in such variation of model and many fields of research. Our research mainly focused on study how SAE method with binomial regression model is applied to obtained estimate proportion of cultural indicator, especially to estimate proportion of people who appreciate heritages and museums in each regency/city level in West Java Province. Data analysis approach used in our research with resurrected data and variables in order to be compared with previous research. The result later showed that binomial regression model could be used to estimate proportion of cultural indicator in Regency/City in Indonesia with better result than direct estimation method.

Keywords: binomial regression, cultural indicator, proportion estimation, small area estimation.

1. Introduction

In statistics, sampling survey is known as one of the sampling methods to provide information about population parameter that become interest of a research. In order to get good direct estimator about these parameter in particular domain, it was

* Received Jan 2018; Accepted Sep 2018; Published online on Oct 2018

necessary to have sufficient sample size. However in many practical surveys, there were lot of difficulties to ensure that sample obtained on its domain were adequate to obtain those estimators.

To overcome that problem, Small Area Estimation (SAE) is one of alternative methods to estimate parameter when sample size is not adequate. This method based on a linear model with the addition of a specific effects on desired level, and also provides researchers with a wide opportunity to add other closely-related explanatory variables (usually called auxiliary variables) that available in desired scope, to improving the accuracy of the direct estimators for that level (Rao & Molina 2015).
SAE method has been applied in such variation of models, with one of them is applied in binomial regression model. This model generally applied if the response variable is in form of binary numbers (0/1 or Yes/No), so that follows Bernoulli or Binomial distribution under certain proportion. Collett (2003) has a brief explanation about binomial logistic regression model, while Farrel et. al. (1997) also Jiang and Lahiri (2001) give some of theoretical view about the application of SAE Method on this model.

SAE method has also been applied in many fields of research, especially in economics and health. Nevertheless, there is one aspect in social research which also should be interesting, that is about culture. UNESCO (2009) stated that culture plays a key role in all societies around the world, influencing various facets of people’s lives, from leisure to professional activities. There are still few research in application SAE method on cultural field, such as Vidyattama et. al. (2015) that measured indigenous cultural participation in Australia from survey data.

Our research mainly focused on study how SAE method with binomial regression model is applied to obtained estimate proportion of cultural indicator. As research limitation, our research had aim to estimated proportion of people who appreciate heritages and museums in each regency/city level in West Java Province.

2. Methodology

Our research mainly used Data Analysis approaches. In that case, data analysis was the direct application of the method of estimating small areas in certain data to predict predetermined cultural indicators. The response variable was a proportion of people who appreciate heritages and museums in each regency/city in West Java Province.

Our research consisted of 1 response variable and 5 auxiliary variables. The response variable bas a proportion of people who appreciate heritages and museums in each Region/City level. Data source for this variable was from SUSENAS MSBP 2015 raw data, consisted of 18,103 individual samples that could be classify into 27 groups which indicated area level (Region/City). Furthermore, the auxiliary variables used in our research were variables which were theoretically related to the response variable. Data source for these variable were from administrative database and some publications from Statistics Indonesia which contained information about indicators that validated in area level. The variable information and source data used in our research data as showed in Table 1.

Table 1: Variable Information and Source Data.

Variable	Information	Data Source
Y	Number of people who appreciate heritages and museums in each Region/City level	SUSENAS MSBP 2015 (Statistics Indonesia)
X1	Ratio of heritages and museums by 100,000 residences in each Region/City level	Heritages and Museums Database (General Directorate of Culture, Indonesia)
X2	Score of Government affairs on cultural field in each Region/City level	Indonesia's Minister of Education and Culture, Act No. 61/2016
X3	Proportion of non-food expenditure in each Region/City level	Publication of Statistics Indonesia
X4	Proportion of people age 5 or over who doesn't have a formal educations in each Region/City level	Publication of Statistics Indonesia
X5	Proportion of people who have been gone out of town in last 6 months for Recreational purpose in each Region/City level	Publication of Statistics Indonesia

As the first step of data analysis in our research, we need to be checked for independence assumption in each of auxiliary variables. We examined this assumption from 2 aspects: significance for correlation and multicollinearity among those auxiliary variables. Result for examination of both aspects concluded that those auxiliary variables in data was still have dependence each other, because there were some significant values of correlation in variable X3, X4, and X5.

To overcome violation of independence assumption on this analysis, we were using Principal Component Analysis (PCA) and Factor Analysis (FA). PCA is used to determine number of principal component that could be form from origin variables, while FA is used to determine origin variables that included in each principal components. Result for this analysis showed that those auxiliary variables grouped into 3 principal components that explained about 83.81% variance of origin data. Information about those principal components and variables included is showed in Table 2. Furthermore, these value initial variables are converted into score of principal components.

Table 2: Variable of Principal Components and Relationship with Original Auxiliary Variables.

Variable	Information	Original Auxiliary Variables
PC1	Society's Social and Economy Condition	• Proportion of non-food expenditure in each Regency/City level • Proportion of people age 5 or over who doesn't have a formal educations in each Regency/City level • Proportion of people who have been gone out of town in last 6 months for Recreational purpose in each Regency/City level
PC2	Government's Cultural Concern	• Score of Government affairs on cultural field in each Regency/City level
PC3	Cultural Heritages Potential	• Ratio of heritages and museums by 100,000 residences in each Regency/City level

Our research used SAE Method based on Area Level Model for proportion, which is written in the form of the following equation (Rao & Molina 2015)

$$\bar{\pi}_i = f_i\bar{p}_i + (1 - f_i)\bar{\pi}_i^* \qquad (1)$$

where i indicated regency/city level in West Java Province ($i = 1, \dots, 27$). From this equation $\bar{p}_i$ was direct estimator of proportion, and f_i in our research formulated as standard sample fraction ($f_i = n_i/N_i$). Since in this case $n_i \ll N_i$ our research dominantly focused on estimate of $\bar{\pi}_i^*$. Moreover $\bar{\pi}_i^*$ was synthetic estimator of proportion based on desired model.

Since the response variable was about proportion, our research considered to used Binomial distribution approach in data analysis. Therefore $\bar{\pi}_i^*$ was synthetic estimator of proportion based on GLMM with logistic link function as written in following equation (Collett 2003)

$$\eta_i = log\ (\bar{\pi}_i^*) = \ln\left(\frac{\bar{\pi}_i^*}{1-\bar{\pi}_i^*}\right) = \beta_0 + \beta_1 PC1 + \beta_2 PC2 + \beta_3 PC3 + v_i \qquad (2)$$

where v_i denoted random effect on regency/city level in West Java Province ($i = 1, \dots, 27$). Then Empirical Bayes approach with assumption binomial distribution of response variable is used to find estimate of $\hat{\boldsymbol{\beta}}$, $\widehat{\sigma_v^2}$, $\widehat{v}_i$, and $\bar{\pi}_i^*$ in (2).

Beside analysis for original binomial distribution, our research considered alternative distribution in assumption for overcoming dispersion problem in modelling. Therefore synthetic estimator $\bar{\pi}_i^*$ in (2) was also estimated with assumption beta-binomial distribution of response variable (Faraway 2016).

To compare bias of estimation with direct estimation, some usual measurement for model comparison couldn't be used, because there was no actual value of proportion for each regency/city in West Java. The only reliable information as reference value was the estimated value of combined proportion for West Java Province from direct estimation on SUSENAS MSBP 2015. Therefore we proposed an alternative way to compared bias of estimation with following steps:

a) Each estimate value that obtained in regency/city were multiplied by number of residents in respective regency/city.
b) Summarized up result at point a), to obtained estimate number of people who answered "Yes" is combined for West Java Province.
c) The result on the point b) was divided by the total population of West Java Province as a whole, to obtained estimate of proportion for Province level.

If result in point c) was close enough to the reference value, then it could be said that the result had better in estimated proportion of people who appreciate heritages and museums in each regency/city in West Java Province.

3. Results and Discussion

Based on data explained in previous section, we conducted GLMM with logistic regression analysis based on initial assumption of binomial distribution approach (written as GLMM Binomial) and beta-binomial distribution approach (written as GLMM Beta-Binomial). Those models conducted with using R application, to estimated parameter β and σ_v^2, estimated random effects for each area as well as estimated proportion for each area. Results of those models showed in Table 3.

In addition, this result showed that estimated coefficient and standard error value between GLMM Binomial and GLMM Beta-Binomial was not much different. Moreover, dispersion parameter value between these two model was almost the same (GLMM Binomial = 0.210; GLMM Beta-Binomial = 0.156). So that, this result showed that GLMM Beta-Binomial still couldn’t able to address dispersion problem in GLMM Binomial.

Table 3: Estimated Coefficient, Standard Error, and Odds Ratio for Estimation of Proportion in GLMM Binomial and GLMM Beta-Binomial Analysis.

Variable	GLMM Binomial			GLMM Beta-Binomial		
	Coefficient[a]	Standard Error	Odds Ratio	Coefficient[a]	Standard Error	Odds Ratio
(Intercept)	-2.88936 *)	0.08695	0.05561	-2.88133 *)	0.09464	0.05606
PC1	0.15230	0.08720	1.16451	0.15266	0.09618	1.16493
PC2	-0.07356	0.08727	0.92908	-0.07638	0.09514	0.92646
PC3	0.29147 *)	0.08445	1.33839	0.29041 *)	0.09144	1.33698

[a] Values with *) showed that value was significant in level $\alpha = 5\%$

For estimated proportion $\hat{\pi}_i^*$ can be calculated with divided those values by number of samples in each area. Results of the model showed in Table 3 above. In addition, this result showed that only intercept and principal components RC3 were significantly influenced proportion of people whom appreciate heritages and museums in each Region/City level. However there was a possibility that this result didn’t reflect actual test, because it was still unknown whether standard error for each parameter has been in accordance with assumed response variable had a Binomial distribution.

To compared bias of estimation, we compared several proportion estimator, such as direct estimator and SAE Method estimator in Binomial logistic regression approach results, as well as actual proportion. All of those estimated are aggregated overall on West Java Province, which showed in Table 4.

Table 4: Comparison in Aggregate Proportion Estimates of West Java Province among Several Models in SAE Method a.

Method of Estimation	Proportion
Actual Value[a]	5.51%
Direct Estimate[b]	5.63%
GLMM Binomial	5.61%
GLMM Beta-Binomial	5.62%

[a] Results of processing raw data SUSENAS MSBP 2015

[b] Using Simple Random Sampling (SRS) Method

This result showed in general that both Binomial regression approach had less difference in estimated proportion to actual value than Direct Estimate Method. So that, it was indicated that these SAE Method in general could have better estimation in proportion than Direct Estimation Method.

Figure 1 as showed below was a map of estimated proportion $\bar{\pi}_i^*$ in each Region/City level in West Java Province, with darker colors showed larger estimated proportion. In this map, it showed that most of city area level had tendency of larger proportion of people whom appreciate heritages and museums, such as Bandung City (Capital of West Java Province) and Depok City (as part of The Megapolitan of Jabodetabek, next to Jakarta as Capital City of Indonesia). Cirebon City had the largest estimated proportion because that area was became one of central government in the Islamic empire era of such as Keraton Kasepuhan, Kacirebonan, and Surosowan.

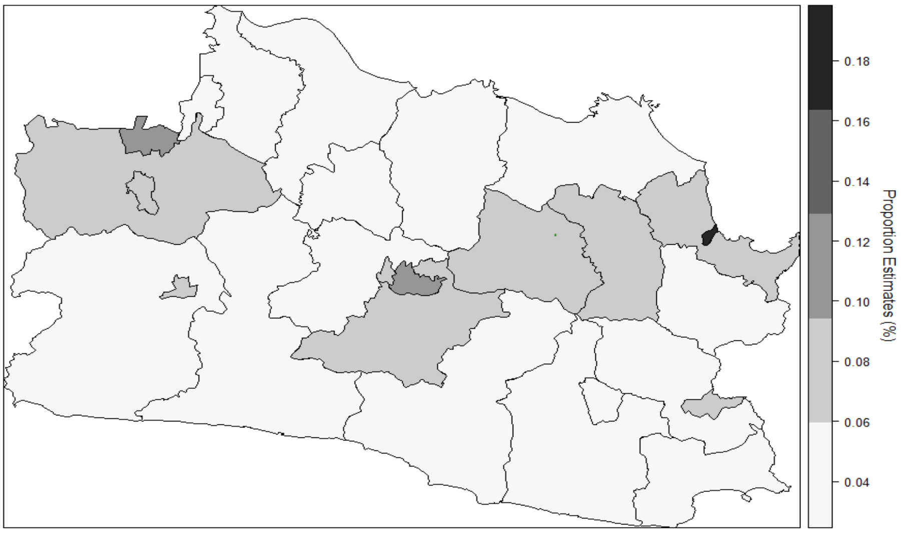

Figure 1: Spatial Map of Estimated Proportion of People who Appreciate Heritages and Museums in West Java Province.

4. Conclusion

According to all of these results, we could say that binomial regression in SAE Method could be used to estimate cultural indicator in Regency/City in Indonesia, at least with better result than direct estimation methods in term of parameter bias with actual value. Based on 2 approaches of binomial regression model that applied in our research, it showed that SAE Method with GLMM Binomial was rather more accurate to estimate proportion, while in opposite GLMM Beta-Binomial still couldn't able to improve accuracy of proportion estimate.

There is one thing to note that all of these results only applied on our research's case, especially for estimate proportion of people who appreciate heritages and museums in each regency/city in West Java Province. Advanced research about this topic such as using data simulation or alternative modelling is recommended to check consistency of these results.

References

Collet, D. (2003). *Modelling Binary Data*. Chapman & Hall.

Faraway, J. J. (2016). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models* (Vol. 124). CRC press.

Farrell, P. J., MacGibbon, B., & Tomberlin, T. J. (1997). Empirical Bayes small-area estimation using logistic regression models and summary statistics. *Journal of Business & Economic Statistics*, ***15***(1), 101-108.

Jiang, J., & Lahiri, P. (2001). Empirical best prediction for small area inference with binary data. *Annals of the Institute of Statistical Mathematics*, ***53***(2), 217-243.

Rao, J. N., & Molina, I. (2015). *Small area estimation*. John Wiley & Sons.

UNESCO. (2009). The 2009 UNESCO framework for cultural statistics (FCS).

Vidyattama, Y., Tanton, R., & Biddle, N. (2015). Estimating small-area Indigenous cultural participation from synthetic survey data. *Environment and Planning A*, ***47***(5), 1211-1228.