Forum Statistika dan Komputasi : *Indonesian Journal of Statistics* available online at:
ISSN : 0853-8115 *journal.ipb.ac.id/index.php/statistika*
Vol. 20 No.2, Oktober 2015, p: 48-52

SMALL AREA ESTIMATION OF LITERACY RATES ON SUB-DISTRICT LEVEL IN DISTRICT OF DONGGALA WITH HIERARCHICAL BAYES METHOD

Rifki Hamdani[1], Budi Susetyo[2], Indahwati[3]

[1]Master Student of Bogor Agricultural University
[2,3]Lecturer of Bogor Agricultural University
E-mail : rifkih@bps.go.id[1] , rifkipapaaim@gmail.com[1]

ABSTRACT

Literacy Rate (LR) is defined as percentage of population aged over 15 with ability to read and write. LR, as one of people welfare indicators, is a measurement of educational development. The indicator, as a measurement of government performance on education, can be measured if all variables related is available. Statistics Indonesia (BPS) each year calculated LR based on National Socio-Economic Survey (SUSENAS) with estimation available only on provincial level and district level. Along with establishment of autonomous regional policy, where regional government had greater power to manage its own region, availability of LR on lower levels to monitor educational development is necessary. Due to sampling design of SUSENAS, accommodated only estimation on district level, will give high variance if used to estimate on lower sub-district level, although still unbiased. Modelling LR was done with Logit-Normal approach, because LR data followed Binomial Distribution. Good estimators from inadequate sample size can be obtained with method of Small Area Estimation (SAE). Hierarchical Bayes (HB) method is one of SAE methods which are proven to give good estimate on binomial distributed data as LR. Estimation on sub-district level in District of Donggala with HB method gave better result compared to the direct estimation with lower Mean Square Error (MSE).

Key words : Small Area Estimation, Literacy Rate, Hierarchical Bayes, Logit-Normal Model

INTRODUCTION

Education is an important aspect on a country development, including Indonesia. Good education, with the right system and good quality will provide reliable human resource which is vital to national development. There are several measurements commonly used to measure effects of educational policies and programs. One of relevant educational measurements in a developing countries is Literacy Rates (LR). Literacy Rate (LR) is defined as percentage of population aged over 15 with ability to read and write. Sometimes, LR can be analyzed differently through its counterpart Illiteracy Rate (IR) with same interpretation. The importance of literacy rates is highlighted by UNESCO (2008), which state that literacy as one of Human Rights. Literacy is someone's main access of knowledge and information. Parents with literacy were able to educate their children so that they became literate too. Conclusively, literacy can give a capacity to a person to provide himself a better life through the age of information nowadays. Therefore, LR measurement has a strategic role on planning and evaluating a region development, especially on education.

Based on Regulation of the Minister of Home Affairs No. 54 on year 2010, secondary data related to indicators that measure performance of regional government (including LR) is provided by Statistics Indonesia (BPS). LR is estimated by BPS based on yearly National Socio-Economic Survey (SUSENAS). Sampling design of SUSENAS provides only estimate on national, provincial and district level due to its sampling size. On the other hand, along with establishment of autonomous regional policy, where regional government had greater power to manage its own region, there is a demand of availability of LR on lower levels to monitor educational development. Using SUSENAS data solely to estimate LR on sub-

district level directly will result to estimates with high variance, although still unbiased. Therefore, if our goal is to estimate LR on sub-district-level, another approach of estimation is necessary.

High variance of estimation can be solved by increasing sample size to improve reliability of the estimate. But this solution is not applicable because increasing sample size means increasing cost, man power, and duration of the survey. The solution, therefore, become ineffective. Another approach can be done via indirect estimation with method of Small Area Estimation (SAE). SAE models generally provide estimates on lower area level where direct estimation from sample is statistically inadequate. Some of SAE methods which are well-known and commonly used are Empirical Best Linear Unbiased Predictor (EBLUP), Empirical Bayes (EB) and Hierarchical Bayes (HB). Between the three, EB and HB are appropriate to be applied on Binomial distributed data. Calculation of LR is based on binary individual data where an individual aged over 15 year who can read and write is coded as 1, and the opposite as 0. The number of literate individual on an area therefore follows Binomial Distribution with LR estimate as its estimated probability (proportion).

Based on BPS data, LR in Province of Central Sulawesi on 2013 was 96.22. It means that 96.22 percent of aged over 15 population of Province of Central Sulawesi can read and write while the other 3.78 percent cannot. The rate was quite higher than national rate on 2013, which was 93.92 percent. District of Donggala on the other hand had the lowest LR in Province of Central Sulawesi (94.61%). Furthermore, to discover which sub-district has low LR is the purpose of this study. With estimate of LR on sub-district level, government can target policies or programs on education to increase literacy with more accurate, depends on condition of each sub-district. LR estimates on sub-district level in District of Donggala was obtained using Hierarchical Bayes and compared with Direct Estimation.

Direct Estimation for Proportion

Response variable y_{ij} is binary response variable on area *i* where y_{ij} with value 1 or 0. On this study, y_{ij} represent individual ability to read and write y_{ij} is valued 1 if j^{th} individual on area *i* who are able to read and write and is valued 0 if j^{th} individual on area *i* who are unable to read and write. Variable y_{ij} thus assumed to follow Bernoulli distribution with parameter p_i , so the probability density function of y_{ij} is :

$$f(y_{ij}|p_i) = p_i^{y_{ij}}(1-p_i)$$

or can be written as,

$$y_i|p_i \overset{ind}{\sim} Binomial(n_i, p_i)$$

The parameter p_i is probability that an individual in area *i* is able to read and write. In a direct estimation the parameter is estimated as small area proportion.

$$p_i = \bar{Y}_i = \sum_j \frac{y_{ij}}{N_i}$$

On a simple random sampling design, estimate of proportion on area *i*, which is $\hat{p}_i$, is derived from maximum likelihood (ML) method resulting $\hat{p}_i = \sum_j \frac{y_{ij}}{n_i} = \frac{y_i}{n_i}$. Maximum Likelihood Estimator (MLE) is a unbiased estimator, characterized by its expected value is the estimated parameter.

$$E(\hat{p}_i) = E\left(\frac{y_i}{n_i}\right) = \frac{1}{n_i}E(y_i) = \frac{1}{n_i}n_i p_i = p_i$$

As an unbiased estimator MLE has mean square error (MSE) with the same value of its variance (Kismiantini 2010) as follows:

$$MSE(\hat{p}_i) = \widehat{Var}(\hat{p}_i) = \frac{\hat{p}_i(1-\hat{p}_1)}{n_i}$$

Hierarchical Bayes Model with Logit-Normal Model

Small area estimates for LR for every sub-district *i* are calculated using Logit-Normal Model with Hierarchical Bayes approach. Rao (2003) defined the model for the approach with assumptions:

i. $y_i|p_i \sim ind\ Binomial\ (n_i, p_i)$
ii. $\theta = logit(p_i) = \boldsymbol{X}_i^T\boldsymbol{\beta} + v_i, v_i \sim N_m(0, \sigma_v^2)$
iii. β and σ_v^2 are independent, $f(\beta) \propto 1$
$\frac{1}{\sigma_v^2} \sim gamma\ (a, b); a \geq 0, b > 0$

where :

a. Model (i) is sampling distribution of response variable proportion of literate population.
b. Model (ii) is relationship between response variable and explanatory variables with area based properties
c. Model (iii) is prior distribution of each parameter $\boldsymbol{\beta}$ and σ_v^2 on the model (ii)

Explanation for each variable on the models above are as follows:

- y_i is the number of individual aged over 15 on sub-district *i* who can read and write.

$$y_i = \sum_j y_{ij}$$

y_{ij} is binary variable on area *i* which valued 1 if j^{th} individual on area *i* who are able to read and write and is valued 0 if j^{th} individual on area *i* who are unable to read and write.

- p_i is a proportion of literate population which is a ratio between y_i with aged over 15 population in sub-district *i* (n_i).

$$p_i = \frac{\sum_j y_{ij}}{n_i} = \frac{y_i}{n_i}$$

n_i is sample size on area *i*.

Parameter p_i is a parameter for random variable y_i with binomial distribution. Parameter p_i is the target of the estimation in this study. Estimation of parameter is obtained from a model that connects parameter LR (p_i) and explanatory variable (x_i), through an appropriate *link function* in Generalized Linear Mixed Model (GLMM) principles. Link function for a Binomial distributed data is logit function as follows :

$$logit\ (p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$

- v_i is random area effect which is assumed to follow Normal Distribution with mean 0 and variance σ_v^2. On this case, prior conjugate distribution for σ_v^2 is Inverse Gamma distribution with hyper-parameter a and b. Value of hyper-parameter a and b on Invers Gamma distribution is conditioned close to 0 because the unavailability of prior information. This condition is called as low information prior (Zhou and You, in Bukhari, 2015).
- $f(\beta) \propto 1$ means that prior distribution for β in the HB model is flat prior. Flat prior is condition where the prior is equally likely/uniformly distributed. Flat prior is chosen because β only have value on certain range. The distribution is preferred because every value on the range has equally likely chance to be selected as supported likelihood in posterior form (Iriawan in Bukhari 2015).

To acquire posterior distribution from multidimensional integral in a closed form is very difficult and sometimes even impossible. Alternative solution that can be used is through calculation of posterior value with numeric approach (iteration). On of commonly used method to approach the solution numerically is Markov Chain Monte Carlo (MCMC). The principle of MCMC is to build a Markov chain of probability distribution which convergent to certain posterior distribution. Calculation of these posterior distributions resulted to samples of posterior values which are calculated to estimate parameter of posterior distribution.

One of the well-known MCMC procedure is Gibbs Conditionals. According to Rao (2003), the form of Gibbs Conditionals for Logit Normal model with explanatory variables on area level are as follows:

i. $[\beta|p,\sigma_v^2,y]\sim N_p[\beta^*,\sigma_v^2(\sum_{i=1}^m x_i x_i^T)^{-1}]$
ii. $[\sigma_v^2|\beta,p,y]\sim Gamma[\frac{m}{2}+a,\frac{1}{2}[(\theta - XT\beta T\theta - XT\beta + b]$
iii. $f(p_i|\beta,\sigma_v^2,y) \propto h(p_i|\beta,\sigma_v^2)k(p_i)$

Estimation of parameter β and σ_v^2 is generated directly from distribution (i) and (ii). Parameter β^* on (i) is acquired from:

$$\beta^* = \left(\sum_{i=1}^m x_i^T x_i\right)^{-1} (\sum_{i=1}^m x_i^T \theta_i)$$

Meanwhile, distribution (iii) can be presented as:

i. $h(p_i|\beta,\sigma_v^2) = \frac{\partial\theta_i}{\partial p_i}\exp\{-\frac{1}{2\sigma_v^2}[\theta_i - xiT\beta]2\}$
ii. $k(p_i) = p_i^{yi}(1-p_i)^{n-y_i}$

For not following certain distribution, proportion estimation of Hierarchical Bayes is estimated with Gibbs Sampling Metropolis-Hasting (M-H) simulation. Generating proportion estimate with M-H simulation is done simultaneously with generating estimate of β and σ_v^2 with Gibbs sampling algorithm. Steps of M-H algorithm are as follows:

1. Generating $\theta\sim ind\ N(X^T\beta,\sigma_v^2)$ then calculate $p_i^* = g^{-1}(\theta_i)$. Value β and σ_v^2 on each iteration is acquired from pervious *Gibbs sampling* process.
2. Calculate acceptance probability:

$$r\left(p_i^{(k)},p_i^*\right) = \min\left\{\frac{k(p_i^*)}{k\left(p_i^{(k)}\right)},1\right\}; k = 0,1,\dots,D$$

3. Generating *u* from *uniform* (0,1).
4. $p_i^{(k+1)} = p_i^*$ is preferred if $u \le r\left(p_i^{(k)},p_i^*\right)$.
5. Repeat step 1 to 4 until D samples is acquired.

From the proportion estimates posterior values can be calculated so that Hierarchical Bayes proportion estimate (p_i^{HB}) approaching:

$$p_i^{HB} \approx \frac{1}{D}\sum_{k=d+1}^{d+D} p_i^{(k)} = p_i^{(.)}$$

While, Variance for Hierarchical Bayes proportion estimate is calculated as:

$$V(p_i^{HB}|\hat{p}) = \frac{1}{D}\sum_{k=d+1}^{d+D} (p_i^{(k)} - p_i^{(.)})^2$$

where :

D = Number of iteration after burn-in period.

d = Burn-in period or period where Markov chain have not convergent yet.

k = current number iteration .

Research Variables

The data which are used in the study are individual data of National Socio-Economic Survey (SUSENAS) 2013 in District of Donggala. Those individual data on SUSENAS questionnaire are located in VSEN13.K block VC question 19. From the whether an individual is able to read and write can be obtained. Aggregation of these data resulted to indicator of literacy rate/proportion which is the response variable of the model.

As covariates, explanatory variables used to provide variance explained to the model and provide good small area estimates on sub-district level, are variables from Village Potency Enumeration (PODES) 2011. PODES 2011 is preferred as the nearest PODES and that the enumeration occurs before the SUSENAS, so then it can provide logical ground for causality.

Table 1. Variables in the model

No	Variable	Explanation	Type	Unit
1	Y	Number of literate individual	Numeric	-
2	x_1	Percentage of agricultural family	Numeric	Percent
3	x_2	Percentage of family using electricity	Numeric	Percent
4	x_3	Average of number of elementary school per 1000 population	Numeric	-

RESULT AND DISCUSSION

HB with logit-normal model was used to estimate proportion of literacy on sub-district i (p_i) by estimating β dan σ_v^2 first as parameter in the model through MCMC approach with Gibbs sampling algorithm and Metropolis Hasting algorithm. Iteration performed to obtain convergent value is as much as 100.000 times. The result showed after burning period iteration became stationary and convergent. The estimations are calculated from iteration 50.000 to 100.000 which are stationary, resulting to estimate model as follows:

$$logit(p_i) = 0.00324 + 0.01524x_{i1} + 0.01282x_{i2} + 0.00093x_{i3}$$

From the model above Comparison of proportion estimation of each sub-district in District of Donggala between HB method and direct estimate method is presented on Table 2. From the table can be seen that estimates with HB logit normal model and direct estimates for each sub district are not so different. The smallest difference was 0.016 on estimates of Sub-district of Banawa Selatan and the largest difference was 0.2837 on estimates of Sub-district of Pinembani. The small differences indicate that iteration of MCMC algorithm is quite convergent. But still the variation of the difference is quite high, which is possibly the result of not quite correct choice of covariates to explain in the model. Addition of covariates with higher correlation to the response variables will improve the accuracy of HB logit normal estimation.

On the other hand, MSE of the HB logit normal estimates of most sub-district are lower if compared to MSE of the direct estimates. It can be concluded that the precision of HB logit normal estimates are better. The estimates showed that in District of Donggala, highest LR estimate was 95.26 percent on sub-district Banawa and the lowest LR estimate was 74.20 percent on sub-district of Pinembani. Direct estimates provided the same result on sub-district with the highest and the lowest LR.

Table 2. Result of Estimation of Literacy Rate (LR) in District of Donggala

No	Sub-District	n_i	y_i	Direct Estimate		HB Estimate	
				LR	MSE	LR	MSE
1	Rio Pakava	108	103	0.95370	0.00041	0.88329	0.00070
2	Pinembani	24	11	0.45833	0.01034	0.74203	0.00484
3	Banawa	112	110	0.98214	0.00016	0.95264	0.00023
4	Banawa Selatan	127	109	0.85827	0.00096	0.87427	0.00028
5	Banawa Tengah	104	97	0.93269	0.00060	0.90241	0.00027
6	Labuan	61	60	0.98361	0.00026	0.89196	0.00031
7	Tanantovea	136	132	0.97059	0.00021	0.89926	0.00052
8	Sindue	155	147	0.94839	0.00032	0.89600	0.00027
9	Sindue Tobusabora	27	26	0.96296	0.00132	0.86099	0.00040
10	Sindue Tobata	53	49	0.92453	0.00132	0.87203	0.00037
11	Sirenja	126	123	0.97619	0.00018	0.90124	0.00037
12	Balaesang	75	71	0.94667	0.00067	0.86583	0.00036
13	Balaesang Tanjung	50	49	0.98000	0.00039	0.88175	0.00037
14	Damsol	164	158	0.96341	0.00021	0.91015	0.00026
15	Sojol	135	129	0.95556	0.00031	0.90311	0.00027
16	Sojol Utara	47	44	0.93617	0.00127	0.88663	0.00031

CONCLUSION AND SUGGESTION

Literacy Rate estimation in District of Donggala with HB Logit Normal model provided better result than direct estimation. It can be seen from the values of MSEs that 10 of 16 sub-district HB estimates resulted lower MSE if compared to the respective direct estimates. But if we look closely, the difference between HB logit normal estimates and the direct estimates are still quite high. The possible cause is that choice of covariates to be included in the HB logit normal model was not quite correct. Addition or replacement of covariates includd in the model, especially covariates that are highly correlated to the response variable is recommended as improvement of Small Area Estimation.

From those conclusion, an inquiry of possible covariates to improve HB estimation is recommended for further studies. More importantly, covariates that are highly correlated to the response variable are preferred. Another suggestion is to include spatial effect between sub-districts to the model to explain more variance within the model. Spatial effect were believed to have contributed on neighboring small area, in this case sub-district, response variables. Spatial correlation between neighboring areas will improve small area estimation if included in the model.

REFERENCES

[BPS] Badan Pusat Statistik. Angka Melek Huruf (AMH) dan Angka Buta Huruf (ABH). URL: http://sirusa.bps.go.id/index.php?r=indikator/view&id=7, accesed on 20 Maret 2015.

Bukhari AS. 2015. Pendugaan Area Kecil Komponen Indeks Pendidikan Dalam IPM di Kabupaten Indramayu dengan Metode Hierarchical Bayes Berbasis Spasial[Thesis]. Bandung(ID): Universitas Padjadjaran.

Kismiantini. 2010. Penerapan Metode Bayes Empirik pada Pendugaan Area Kecil untuk Kasus Biner (Studi tentang Proporsi Status Kepemilikan Kartu Sehat di Kota Yogyakarta). Seminar Nasional Penelitian, Pendidikan, dan Penerapan MIPA; Yogyakarta, Indonesia. Yogyakarta(ID): Universitas Negeri Yogyakarta.

Norlatifah. 2015. Pendugaan Area Kecil Terhadap Angka Melek Huruf di Kabupaten Kutai Kartanegara Dengan Metode Empirical Bayes Berbasis Model Beta-Binomial[Thesis]. Bandung(ID): Universitas Padjadjaran.