Hierarchical Bayesian Modelling in Small Area for Estimating Binary Data

Hierarchical Bayesian Modelling in Small Area for Estimating Binary Data

A D Sari[1], F Yanuar[1*]

[1]Mathematics and Science Department, Andalas University, Padang, Indonesia

*corresponding author: ferrayanuar@sci.unand.ac.id

Abstract. Indonesian's data are obtained from BPS from census, but census are designed for large area. Now, local goverments need to have reliable and detailed information in small area. Direct estimation are unreliable to be applied in small area because produced high mean square error (MSE). To overcome this problem, we use the indirect estimation Small Area Estimation Hierarchical Bayesian (SAE HB) with Logit Normal as the model. From this study founded that HB produced a smaller MSE than direct estimation

1. Introduction

Statistics Indonesia (BPS) calculate about literacy rate, drop out children from school, etc periodly with cencus, which from its sampling design can provide direct estimation only on provincial level and district area. Along with establishment of autonomous regional policy, where regional governments had greater power to manage their own region, availability of data on lower level is necessary for regional government. Due to sampling design of cencus, accommodated only estimation on district level, the data will give high variance if used to estimate on lower sub-district level, although still unbiased. The high variance will result to broader confidence interval of estimation, which will make the estimation become unreliable [1].

One of method to obtain accurate estimator from inadequate sample size in small area is method of Small Area Estimation (SAE). Until now, the SAE method has been applied in various disciplines. The SAE method that is widely known and has been used in various subject which is Empirical Best Liniear Unbiased Predictor (EBLUP), Empirical Bayes (EB), Hierarchical Bayes (HB). EBLUP method can be applied for linear mixed models that are suitable to use if the response variable is a continuous variable. Some research that using EBLUP method are Krieg, Blaess and Smeets (2012), and Song (2011). On the other hand, EB and HB method can be generally applied because can be used for liniear mixed models, binary data and count data. The research using this method are Datta, Lahiri, and Maiti (2002), and Bukhari (2015). Rao (2003) explains some examples how to use these three methods in his book.

In this study, we used the binary data from R-software. Based on this information, it is known that the response variable in this study is binary data so the SAE method used is the HB method. The HB method is preferred because the EB method does not count the variance in hyperparameter estimation. The advantages of SAE HB are: (1) The model specifications are easy to use and can model different various (2) the inferential problem is relatively clearer and its computation is relatively easier by using the Markov Chain Monte Carlos technique (MCMC) [7].

Rifki Hamdani (2015) used HB SAE to estimate literacy rate on sub-district level in district of Donggala with Hierarchical Bayes method. In this research three methods were compared, the first is direct estimation, the second is HB spatial logit-normal SAE, and the third is HB non spatial logit-normal SAE. The conclusion is HB SAE the best method for estimating literacy rate in sub-district level either by including spatial or not. This study will compare the estimation of data for small areas using the small area estimation hierarchical bayes method and direct estimation.

2. Methods

2.1. Direct Estimation on Variable Binomial Response

The response variable y_{ij} is a binary variable count in i and j area where y_{ij} is 1 or 0. If the variable y_{ij} is assumed to have a Bernouli distribution with p_i as the parameter, so the density function of y_{ij} is:

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

or

$y_i|p_i \underset{\sim}{^{ind}} Binomial(n_i, p_i)$.

The proportion of p_i is:

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

If the sampling used the simple random sampling, then estimate the proportion of the i area is p_i, is derived through the method of Maximum Likelihood (ML), namely $\hat{p}_i = \sum_j \frac{y_{ij}}{n_i} = \frac{y_i}{n_i}$. ML estimation is unbiased estimator because the expectation velue of the estimator is the same as the parameter.

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

So, the mean square error is same as the variety

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

2.2. Hierarchical Bayes Method with Logit-Normal Model

Small area estimation for each small area can calculate with Hierarchical Bayes Logit-Normal model. Rao (2003) defines the model as:

i. $y_i|p_i \sim ind\ Binomial\ (n_i, p_i)$

ii. $\theta_i = logit(p_i) = x_i^T\beta + v_i, v_i \underset{\sim}{^{iid}} N(0, \sigma_v^2)$

iii. β dan σ_v^2 are mutualy independent with $f(\beta) \propto 1$
$\frac{1}{\sigma_v^2} \sim gamma\ (a,b); a \geq 0, b \geq 0$

Thus, p_i is a parameter of the y_i variable that has a binomial distribution and it is the target to estimate. To connect p_i with the X_i variable we need the link fuction that matches with Generalized Linier Mixed Model. The suitable model is:

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

If v and y are vectors that contain the value v_i and y_i then y vector will take the distribution of binomial product:

$$f(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{v}) = f(y_1|\boldsymbol{\beta}, \boldsymbol{v}) f(y_2|\boldsymbol{\beta}, \boldsymbol{v}) \dots f(y_m|\boldsymbol{\beta}, \boldsymbol{v})$$
$$= \prod_{i=1}^{m} p_i{}^{y_i}(1-p_i)^{1-y_i}$$

The distribution for β and v will follow the Normal distribution:

$$f(\beta, v|\sigma_v^2) = \prod_{i=1}^{m} 1 \times (2\pi\sigma_v^2)^{-1}\exp(-\frac{1}{2\sigma_v^2}v_i^2)$$
$$\propto (\sigma_v^2)^{-m} exp\left(\sum_{i=1}^{m}\frac{1}{2\sigma_v^2}v_i^2\right)$$

If m indicates a small area, then the variance of the combined area will follow the inverse Gamma distribution:

$$f(\sigma_v^2) = \frac{b^a \exp(-\frac{b}{\sigma_v^2})}{\sigma_v^{2(a+1)}\Gamma(a)}$$

So, we get the new distribution from these variable is:

$$f(y, \beta, v, {\sigma_v}^2) \propto \prod_{i=1}^{m} {p_i}^{y_i}(1-p_1)^{y_i} \times ({\sigma_v}^2)^{-1} \exp(-\frac{1}{2{\sigma_v}^2} v_i^2) \times \frac{b^a \exp(-\frac{b}{{\sigma_v}^2})}{{\sigma_v}^{2(a+1)}\Gamma(a)}$$

$$f(p_1, \dots, p_m, \beta, {\sigma_v}^2 | y) \propto \prod_{i=1}^{m} f(p_i, \beta, {\sigma_v}^2) f(y|\beta, {\sigma_v}^2)$$

The function $f(p_1, \dots, p_m | y)$ is the marginal distribution from the $f(p_1, \dots, p_m, \beta, {\sigma_v}^2 | y)$ and the function is

$$f(p_1, \dots, p_m, \boldsymbol{\beta}, {\sigma_v}^2 | y) = \int_{\beta} \int_{{\sigma_v}^2} f(p_1, \dots, p_m, \boldsymbol{\beta}, {\sigma_v}^2 | y)\, d(\boldsymbol{\beta}) d({\sigma_v}^2)$$

To find the marginal function from $\boldsymbol{\beta}$, $\boldsymbol{v_i}$, and ${\sigma_v}^2$ we get from $f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2)$ function, they are:

$$f(\beta_0 | \boldsymbol{y}, \beta_1, \dots, \beta_k, \boldsymbol{v}, {\sigma_v}^2) = f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2) / \int f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2)\, d(\beta_0)$$

$$\vdots$$

$$f(\beta_u | \boldsymbol{y}, \beta_0, \dots, \beta_{u-1}, \beta_{u-1}, \dots, \beta_k\, \boldsymbol{v}, {\sigma_v}^2) = f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2) / \int f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2)\, d(\beta_u)$$

$$f(v_i | \boldsymbol{y}, \beta, v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_m\, {\sigma_v}^2) = f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2) / \int f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2)\, d(v_i)$$

$$f({\sigma_v}^2 | \boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}) = f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2) / \int f(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{v}, {\sigma_v}^2)\, d({\sigma_v}^2)$$

It is not possible to get a close form from that final function with the form of a multi-dimensional integral in the above equations. One method used to solve this problem is Markov Chain Monte Carlos (MCMC) algorithm. The usually MCMC procedure is Gibbs Conditional. After simulations and iterations, the estimation of proportion of Hierarchical Bayes (p_i^{BB}) is

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

And the variation for Hierarchical Bayes estimatimation (p_i^{BB}) is

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

where
D = Iteration after burn in
d = Burn in period
k = The iteration

One measure of model compatibility that can be used in evaluating the compatibility of the Bayes model is D*eviance Information Criterion* (DIC). The smaller DIC value indicates the more suitbale model to use. According to Ntzoufras (2009), this criterion is defined as:

$$DIC = \overline{2D(\theta_c, c)} - D(\bar{\theta}_c, c) = D(\bar{\theta}_c, c) + 2p_c$$

2.3. Data Source

The data used in this study is data generated by R-software. Variables X_1 and X_2 are generated following the Normal distribution of 38 data for each variable. X_1 is generated by Normal distribution with an average of 10 and variant 5. X_2 is generated with Normal distribution with an average of 7 and variant 3. Furthermore, variable data is generated "n" with a Binomial distribution.

3. Result and discussion

3.1. Exploration p_i with Direct Estimation and Hierarchical Bayes Small Area Distribution

Descriptive statistics p_i of data generated through R-software, where y_i following Binomial distribution are presented in Table 1.

Table 1. Descriptive statistics p_i

Descriptive statistics p_i	
Mean	0,014166
Standard deviation	0,006324344
Maximum value	0,02439
Minimum value	0,0000
Total	38

From Table 1, it can be seen from the highest p_i value that is 0,02439 and the lowest is 0,000. The standard deviation is 0,006324344, this means that the value of p_i is not too diverse. Data from the table is processed using Gibbs Sampling algorithm by entering a variable predictor component (X_1, X_2) to obtain 3 models, there are:

Model A : $logit\ (p_i) = \beta_0 + \beta_1 X_1 + v_i;\ v_i \sim N(0, \sigma_v^2)$

Model B : $logit\ (p_i) = \beta_0 + \beta_1 X_2 + v_i;\ v_i \sim N(0, \sigma_v^2)$

Model C : $logit\ (p_i) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + v_i;\ v_i \sim N(0, \sigma_v^2)$

Tabel 2. Estimated results p_i HB SAE dan DIC

Model	Parameter	Mean	Std.Dev	Credible Interval 5%	Credible Interval 95%	Med	DIC
A	β_0	-5.99	0,7729	-7,582	-4,541	-5,959	87.905
	β_1	0,1709	0,07259	0,03299	0,3173	0,169	
B	β_0	-4,771	0,5808	-5,731	-3,829	-4,763	92.451
	β_1	0,06108	0,06676	-0,05011	0,01751	0,0126	
C	β_0	-6,129	0,8563	-7,564	-4,737	-6,115	89.819
	β_1	0,1654	0,0628	0,04021	0,2915	0,1648	
	β_2	0,02205	0,0045	-0,08264	0,1261	0,02237	

In Table 2 it can be seen that the model with all parameter that are significant at the 95% confidence level is model A. In Table 2 also presented the DIC value for each model. DIC value can be used as a measure of model compatibility, where the smaller DIC value of a model shows that the model is suitable for the data. The smallest DIC value is also obtained by model A. thus, the best model for estimating proportion of p_i with HB SAE method is A model, which is a model involving one predictor variable, namely X_1. The equation formed by the model is as follows:

$logit\ (p_i) = -5{,}99 + 0{,}1709 X_1$.

Futhermore, after obtaining parameters of p_i is carried out with the best HB SAE model as presented in Table 3 as below.

Table 3. Result of p_i with HB SAE Method

p_i	Mean	Credible Interval 5%	Credible Interval 95%	Std. Dev
p_1	0.01922	0.01347	0.02594	0.003902
p_2	0.01969	0.01368	0.02674	0.004086
p_3	0.00869	0.005033	0.01335	0.0026
p_4	0.006674	0.003075	0.0116	0.002678
p_5	0.01622	0.01186	0.02102	0.002922
p_6	0.009283	0.005627	0.01369	0.002545
p_7	0.009167	0.005471	0.01362	0.002562
p_8	0.01591	0.01161	0.02069	0,002827
p_9	0.02835	0.01634	0.04458	0,008749
p_{10}	0.01512	0.01114	0.01945	0,002609
p_{11}	0.01297	0.009343	0.01694	0,00237

p_i	Mean	Credible Interval		Std. Dev
		5%	95%	
p_{12}	0.01286	0.009232	0.01688	0.002397
p_{13}	0.009653	0.00606	0.01415	0.002545
p_{14}	0.02389	0.01512	0.03484	0.006161
p_{15}	0,01769	0.01274	0,0235	0.003265
p_{16}	0,01702	0.01231	0.02232	0.003101
p_{17}	0,01387	0.01021	0.01794	0.002399
p_{18}	0,01728	0.0124	0.02277	0.003182
p_{19}	0,01282	0.009227	0.01675	0.002357
p_{20}	0,008325	0.005115	0.01337	0.00257
p_{21}	0.008325	0.004619	0.01311	0.002644
p_{22}	0.01209	0.008584	0.01616	0.002354
p_{23}	0.01364	0.009985	0.01775	0.002458
p_{24}	0.02341	0.01511	0.03403	0.005873
p_{25}	0.01452	0.01067	0.01869	0.002516
p_{26}	0.0106	0.007017	0.01475	0.002468
p_{27}	0.01198	0.008512	0.01597	0.002316
p_{28}	0.01154	0.007922	0.01565	0.002397
p_{29}	0.01511	0.01115	0.01948	0.002602
p_{30}	0.02114	0.01429	0.02942	0.004737
p_{31}	0.01473	0.01087	0.01915	0.002566
p_{32}	0.01585	0.01166	0.02045	0.002761
p_{33}	0.008834	0.005126	0.01338	0.002579
p_{34}	0.007628	0.003969	0.01237	0.002612
p_{35}	0.009484	0.00583	0.01385	0.002527
p_{36}	0.02072	0.01399	0.02849	0.004554
p_{37}	0.0126	0.009016	0.01653	0.002369
p_{38}	0.01275	0.009223	0.01677	0.002341

The next step is to compare the results of direct estimation (DE) and HB SAE as follows:

Table 4. The comparison results of Direct Estimation and HB SAE

p_i	Proportion p_i (DE)		Proportion p_i (HB SAE)	
	p	Sd	p	Sd
p_1	0,023255814	0,01613389	0.01922	0.003902
p_2	0,02173913	0,01507357	0.01969	0.004086
p_3	0,011627907	0,01147609	0.00869	0.0026
p_4	0	0	0.006674	0.002678
p_5	0,012195122	0,01204335	0.01622	0.002922
p_6	0,010869565	0,01071767	0.009283	0.002545
p_7	0,010752688	0,01060079	0.009167	0.002562
p_8	0,012048193	0,01189641	0.01591	0,002827
p_9	0,022727273	0,01576448	0.02835	0,008749
p_{10}	0,010752688	0,01060079	0.01512	0,002609
p_{11}	0,011363636	0,01121179	0.01297	0,00237
p_{12}	0,011764706	0,0116129	0.01286	0.002397
p_{13}	0,010869565	0,01071767	0.009653	0.002545
p_{14}	0,022727273	0,01576448	0.02389	0.006161
p_{15}	0,021978022	0,01524064	0,01769	0.003265
p_{16}	0,012048193	0,01189641	0,01702	0.003101

p_i	Proportion p_i (DE)		Proportion p_i (HB SAE)	
	p	Sd	p	Sd
p_{17}	0,011494253	0,01134242	0,01387	0.002399
p_{18}	0,010638298	0,01048638	0,01728	0.003182
p_{19}	0,010526316	0,01037439	0,01282	0.002357
p_{20}	0,01087	0,010718	0,008325	0.00257
p_{21}	0,011765	0,011613	0.008325	0.002644
p_{22}	0,012195	0,012043	0.01209	0.002354
p_{23}	0,010526	0,010374	0.01364	0.002458
p_{24}	0,023529	0,016325	0.02341	0.005873
p_{25}	0,011494	0,011342	0.01452	0.002516
p_{26}	0,011236	0,011084	0.0106	0.002468
p_{27}	0,011111	0,010959	0.01198	0.002316
p_{28}	0,0125	0,012348	0.01154	0.002397
p_{29}	0,021053	0,014593	0.01511	0.002602
p_{30}	0,02381	0,016521	0.02114	0.004737
p_{31}	0,024096	0,016721	0.01473	0.002566
p_{32}	0,02439	0,016926	0.01585	0.002761
p_{33}	0,011628	0,011476	0.008834	0.002579
p_{34}	0	0	0.007628	0.002612
p_{35}	0,010753	0,010601	0.009484	0.002527
p_{36}	0,024096	0,016721	0.02072	0.004554
p_{37}	0,0125	0,012348	0.0126	0.002369
p_{38}	0,011364	0,011212	0.01275	0.002341

Based on the Table 4 above, it can be seen that the estimated proportion value between DE and HB SAE shows a similarity. However, the estimation of HB produces a debiation value smaller than estimated by DE. So, the indirect estimation is better than the direct estimation method.

4. Conclusion and Suggestion

The study concluded that the estimation of the proportion of p_i with HB method better than DE method. This is because estimation with HB SAE produces a smaller standard deviation. In this study, we only compare the standard deviation between HB SAE and DE, in the next study it can be compared DIC value between estimation with HB SAE and DE. After that, in this study only two predictor variable were used and formed three models, we were hoped that in the next study more predictor variable would be used to obtain more model

Reference

[1] Satriya A M 2015 Small Area Estimation Pengeluaran Perkapita di Kabupaten Bangkalan dengan Metode Hierarchical Bayes *Jurnal Statistika* **3** Universitas Muhammadiyah Semarang

[2] Krieg S, Blaess V, dan Smeets M 2012 Small Area Estimation of Turnover of The Structural Business Survey, Discussion Paper, Statistics Netherlands, The Hague

[3] Song S 2011 Small Area Estimation of Unployment: From Feasibility to Implementation. Paper presented at the New Zealand Association of Economists Conference, Wellington

[4] Datta G S, Lahiri P, and Maiti T 2002 Empirical Bayes Estimation of Median Income of Four Person Families by State Using Time Series and Cross-Sectional Data *Journal of Statistical Planning and Inference* **102**

[5] Bukhari A S 2015 Pendugaan Area Kecil Komponen Indeks Pendidikan Dalam di Kabupaten Indramayu dengan Metode Hoerarchical Bayes Berbasis Spasial. Tesis. Unversitas Padjadjaran, Bandung.

[6] Rao J N K 2003 Small Area Estimation. New York: John Wiley and Sons.
[7] Hajarisman N 2013 Pemodelan Area Kecil untuk Menduga Angka Kematian Bayi Melalui Pendekatan Model Regresi Poisson Bayes Berhirarki Dua- Level, Disertasi, Institut Pertanian Bogor, Bogor
[8] Hamdani R 2015 Small Area Estimation of Literacy Rates on Sub-District Level in District of Donggala with Hierarchical Bayes Method *Indonesian Journal of Statistics* **20** Institut Pertanian Bogor.
[9] Ntzoufras I 2009 Bayesian Modeling Using Winbugs New Jersey: John Wiley and Sons