Model Assisted Statistics and Applications 16 (2021) 241–250
DOI 10.3233/MAS-210538
IOS Press

Assessment of health and social security agency participants proportion using hierarchical bayesian small area estimation

Ferra Yanuar[a,*], Atika Defita Sari[a], Dodi Devianto[a] and Aidinil Zetra[b]
[a]*Department of Mathematics, Faculty of Mathematics and Natural Science, Andalas University, Kampus Limau Manis, Padang, Indonesia*
[b]*Department of Political Sciences, Faculty of Social and Political Sciences, Andalas University, Kampus Limau Manis, Padang, Indonesia*

Abstract. Data on the number of health insurance participants at the subdistrict level is crucial since it is strongly correlated with the availability of health service centers in the areas. This study's primary purpose is to predict the proportion of health and social security participants of a state-owned company named *Badan Penyelenggara Jaminan Sosial Kesehatan* (BPJS) in eleven subdistricts in Padang, Indonesia. The direct, ordinary least square, and hierarchical Bayesian for small area estimation (HB-SAE) methods were employed in obtaining the best estimator for the BPJS participants in these small areas. This study found that the HB-SAE method resulted in better estimation than two other methods since it has the smallest standard deviation value. The auxiliary variable age (percentage of individuals more than 50 years old) and the percentage of health complaints have a significant effect on the proportion of the number of BPJS participants based on the HB-SAE method.

Keywords: Hierarchical bayesian (hb), small area estimation (SAE), health and social security participants

1. Introduction

One indicator of Indonesia's medium-term goals is the proportion of the individuals registered in health and social security system namely BPJS Kesehatan (*Badan Penyelenggara Jaminan Sosial Kesehatan*, Health Social Security Agency), a state-owned company of aimed at providing universal and affordable health care to its citizens. Nationally, the number of the national health insurance participants in Indonesia has reached 221,580,743 people or 83.5% of all Indonesia's citizen as of May 2019. In Padang, the capital city of West Sumatra province, BPJS recorded that until December 1, 2019, the number of people registered in the National Health Insurance for the Indonesia Health Card (JKN-KIS) program had reached 1,463,097 people or 80.3% of all population in Padang. These information regarding the number of BPJS registrants/participants, provided by BPS (*Badan Pusat Statistik*, Central Bureau of Statistics), are only presented at the city/district level. There is no detail information available regarding the number of BPJS participants at the subdistrict level in Padang and this has become an obstacle for local governments in policy-making for regional autonomy implementation.

Information regarding the number of BPJS participants at the subdistrict level is vital because it corelates strongly to the availability of health service centers in the areas. Thus, the reliability of small-area estimation is critical in making proper decision or policies. In the estimation of a characteristic of such a small group, a direct estimate

*Corresponding author: Ferra Yanuar, Department of Mathematics, Faculty of Mathematics and Natural Science, Andalas University, Kampus Limau Manis, Padang, Indonesia. E-mail: ferrayanuar@sci.unand.ac.id.

ISSN 1574-1699/$35.00

based solely on data from the small group is likely to be unreliable, because only a small number of observations are available from the small group (Sugasawa & Kubokawa, 2020). The problem of small area estimation is how to generate a reliable estimate for a small group's characteristic, and small area estimation has been actively studied from both theoretical and practical aspects due to an increasing demand for reliable small area estimates from public and private sectors. A study by Yoshimori and Lahiri (2014) proposed a new adjustment factor that rectifies the problems associated with the existing adjusted likelihood methods. Sugasawa et al. (2018) developed an empirical Bayesian approach in which the Monte Carlo expectation-maximization algorithm for computing the maximum likelihood estimator. Diallo and Rao (2018) suggested relaxing the normality assumption of Molina et al. (2014) and derive the Empirical Bayesian estimator, assuming that the random errors in the nested-error model follow skew-normal distributions. Tsujino and Kubokawa (2019) suggested the nested error regression model with skew-normal distributions error terms. Sugasawa et al. (2019) proposed the unmatched sampling and linking models with an unknown link function modeled by a P-spline. They provided a hierarchical representation of the proposed model. Yanuar et al. (2019) employed empirical Bayesian estimator for small area estimation which use Poisson Gamma as prior distribution. Sugasawa and Kubokawa (2020) reviewed small area estimation techniques using mixed models, covering from fundamental to recently proposed advanced ones. Few types of research have been done in the application or employment of this method, such as an application on poverty indicators (Molina et al., 2014) and forestry estimation (Ver Planck et al., 2018).

In this paper, hierarchical Bayesian models in small area estimation are employed for counts data. The SAE method using Bayesian specification are based on the Fay-Harriot model (Fay & Herriot, 1979), which considers a generalized linear Poisson model. You and Rao (2002) proposed a Normal-lognormal model within the class of the unmatched models. Recently, Nazir et al. (2016) has proposed a new improvement before hyperparameters of variance components and then consider a lognormal model. In this paper, a modification of the latter, a Binomial-log Normal model, is considered. Under appropriate conditions, any model could have some merits. Still, its appropriateness depends on various circumstances like the size of the areas, availability of suitable explanatory variables at the area level, the accuracy of sampling variance estimates, etc. The practical use of Bayesian hierarchical models has been boosted by software availability that implements MCMC simulations so that estimating the model can be straightforward and relatively easy.

This study aims to predict the proportion of the BPJS participants in the subdistrict level in Padang, Indonesia. The direct method, ordinary least square (OLS), and SAE-HB are employed to achieve the purpose. The three methods are compared to obtain the best estimator. The best approach is then implemented to predict the proportion in a subdistrict where the number of BPJS participants is unknown. The proportion of BPJS participants using the direct method cannot be predicted in the corresponding areas since not all subdistricts in Padang have complete information regarding the number of BPJS participants. Meanwhile, the OLS method is used for large data set and if the error term is not properly interpreted, the regression results are sensitive to functional form, which can lead to widely disparate conclusions depending on how the regression is initially set up. The SAE method with the HB approach is employed then. This method involves the auxiliary variables to forecast the proportion of the BPJS participants and its confidence interval. The confidence interval is essential for the BPJS party in allocating the benefit funds paid to customers in the future.

2. Materials and methods

2.1. Materials

This study uses data of 2018 BPJS participants in eleven subdistricts obtained from the Padang's BPJS branch. Data from ten subdistricts are analyzed to predict the model as presented in Table 1. One subdistrict that is Bungus Teluk Kabung, has no information regarding the number of BPJS participants. The number of individuals registered in BPJS for this subdistrict (Bungus Teluk Kabung), will then be estimated by the proposed model obtained later.

As can be seen in Table 1, the highest number of BPJS participants is in Koto Tangah subdistrict with 174,000 participants, with the number of individuals living in this subdistrict is 193,000. Meanwhile, the lowest of BPJS participants is in Padang Barat, with only 40,000 participants.

Table 1
The number of BPJS participants in each subdistrict in Padang (in thousand)

No	Subdistrict	y_i	N_i	X_1	X_2	$\hat{p}_i^{DM}$	Type
1	Koto Tangah	174	193	15.86	21.92	0.9106	1
2	Kuranji	125	150	15.65	12.27	0.8333	1
3	Lubuk Begalung	107	123	15.79	10.29	0.8699	1
4	Lubuk Kilangan	51	56	15.12	10.59	0.9107	0
5	Nanggalo	55	62	17.56	8.90	0.8871	0
6	Padang Barat	40	46	20.39	9.12	0.8695	0
7	Padang Selatan	55	60	17.57	8.81	0.9167	0
8	Padang Timur	78	80	17.63	18.81	0.9750	0
9	Padang Utara	55	71	14.61	11.06	0.7746	1
10	Pauh	55	74	13.37	5.24	0.7432	1
11	Bungus Teluk Kabung	Not Available	25	14.28	5.10	–	–

Many studies have been published regarding factors affecting individuals enrolling in health insurance. Mahumud et al. (2017) explored the socioeconomic, demographic, and behavioral characteristics of an individual that impacted health insurance expenditures in Bangladesh. Duku (2018) identified age, sex, educational level, marital status, health status and, travel time to the nearest health facility as determinants of enrolment in health insurance. Salari et al. (2019) considered the variables that affect the individual to register in health insurance: wealth, marital status and, age.

This study considered age (percentage of individual more than 50 years old), sex (percentage of female), marital status (percentage of married people), educational level (percentage of an individual had at least senior high school), and percentage of health complaints, to be auxiliary variables. These auxiliary variables are assumed could improve the prediction of parameters to be estimated in this model, that is the proportion of the number of BPJS participants in each subdistrict in Padang. In a preliminary analysis, it was obtained that only age (percentage of individuals more than 50 years old) and percentage of health complaints, have significant correlation on response. Therefore, only two auxiliary variables are then considered in the hypothesis model.

2.2. Methods

In this research, the problem is how to estimate model parameters whose response has Binomial distribution with parameter n and p, written as $Y \sim$ Binomial (n, p). The response variable x_k, $k = 1, \ldots, m$ is the binary response variable measured in the ith area. The data taken for this paper is, $x_k = 1$ if an individual in area i is registered as a BPJS participant, while $x_k = 0$ on the contrary. If there are n individuals in an area i, denoted as n_i, $i = 1, \ldots, n$ then Y is defined as $Y = \sum_{k=1}^{m} x_k$ as the number of certain individuals in an area i who are registered as BPJS participants.

Direct estimation for binary response

Direct estimator for parameter p which is representing the proportion of several BPJS participants is obtained by using the likelihood estimation method. The response Y is assumed to have Binomial distribution, which is obtained from x_k that has Bernoulli distribution which its probability distribution function as following (Sari & Yanuar, 2020)

$$f(x_k, p) = p^{x_k}(1-p)^{1-x_k}, x_k = 0, 1 \tag{1}$$

The likelihood function then can be constructed based on this pdf

$$L(x_k, p) = p^{\sum_{k=1}^{m} x_k}(1-p)^{n-\sum_{k=1}^{m} x_k}, \tag{2}$$

By maximizing Eq. (2) then we obtain the parameter estimated for p

$$\hat{p} = \frac{\sum_{k=1}^{m} x_k}{n} = \frac{y}{n} \tag{3}$$

In this study, the estimated proportion here refers to the proportion of BPJS participants in subdistrict i, $\hat{p}_i$ which is estimated by the ratio between the number of BPJS participants (Y_i) over the number of population at subdistrict (or area) i, denoted as (N_i), or formulated with $\hat{p}_i = \frac{Y_i}{N_i}$ for $i = 1, \ldots, N$. Meanwhile, the standard deviation for proportion is defined as the square root of the variance of $\hat{p}_i$, or

$$\text{Var}(\hat{p}_i) = \text{Var}\left(\frac{Y_i}{N_i}\right) = \frac{p_i(1-p_i)}{N_i} \tag{4}$$

Ordinary least square

Ordinary least squares (OLS) is a linear least squares method used to estimate unknown parameters in a linear regression model. The principle of least squares is used by OLS to select the parameters of a linear function of a set of explanatory variables: minimizing the sum of squares of the differences between the observed dependent variable (values of the variable being observed) in the given dataset and those predicted by the linear function of the independent variable (Higgins & Thomas, 2019).

Hierarchical Bayesian in SAE for binary response

In this present study, hierarchical Bayesian is constructed using the Binomial Logit-Normal model to linearized the correlation between response and its auxiliary variables. It is assumed that response $Y \sim$ Binomial (n, p) which the parameter p is affected by the auxiliary variables $X = (X_1, X_2, \ldots, X_p)$.

$$\log it(p_i) = \theta_i = X_i^T\beta + v_i + e_i, i = 1, 2, \ldots, n. \tag{5}$$

The random effect v_i is assumed Normal distribution $N(0, \sigma_v^2)$. Meanwhile, the sampling error (e_i) has Normal distribution as well, written as $e_i \sim N(0, \sigma_e^2)$. The coefficient regression, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$, and σ_v^2 are independent. Here, σ_v^2 is unknown and it's assumed the informative prior distribution for σ_v^2 is Gamma distribution, or $\sigma_v^2 \sim$ Gamma (a, b) with $a > 0, b \geqslant 0$, while σ_e^2 is assumed known. The prior distribution for β is Uniform distribution, $f(\beta) \propto 1$. Thus, parameter θ_i has normal distribution or $\theta_i \sim N(X_i^T\beta, \sigma_v^2)$ with its probability distribution function as following:

$$f(\theta|\beta, \sigma_v^2) = \frac{1}{\sqrt{\{2\pi\sigma_v^2\}}} \exp\left(-\frac{1}{2\sigma_v^2}(\theta - X_i^T\beta)^2\right). \tag{6}$$

Meanwhile, the likelihood function for θ is:

$$L(\theta|\beta, \sigma_v^2) \propto (\sigma_v^2)^{-\frac{m}{2}} \exp\left(\frac{1}{2\sigma_v^2}\sum_{i=1}^{m}(\theta - X_i^T\beta)^2\right). \tag{7}$$

The joint prior distribution for β and σ_v^2 conditional θ is proportional to (Yanuar et al., 2019)

$$f(\beta, \sigma_v^2|\theta) \propto f(\theta|\beta, \sigma_v^2)f(\beta, \sigma_v^2). \tag{8}$$

In this present study, the parameter p_i, $i = 1, 2, \ldots, m$ are also estimated. Therefore, the joint posterior distribution is

$$f(p_1, \ldots, p_m, \beta, \sigma_v^2|y) \propto \prod_{i=1}^{m} f(p_i, \beta, \sigma_v^2)f(y|p_i, \beta, \sigma_v^2) \tag{9}$$

The marginal posterior distribution for p_i is

$$f(p_1, \ldots, p_m|y) = \int_{\beta} \ldots \int_{\sigma^2_{\setminus nu}} f(p_1, \ldots, p_m, \beta, \sigma_v^2|y)d\beta d(\sigma_v^2) \tag{10}$$

Rao and Molina (2015) suggested employing Gibbs sampler to solve Eq. (10) and follow these steps:

$$f(p_i|\beta, \sigma_v^2, y) \propto h(p_i|y, \beta, \sigma_v^2)k(p_i),$$

where

$$h(p_i|y, \beta, \sigma_v^2) = \frac{\partial\theta_i}{\partial p_i} \exp\left(-\frac{2}{2\sigma_v^2}(\theta_i - X_i^T\beta)^2\right),$$

and

$$f(p_i) = p_i^{y_i}(1 - p_i)^{n-y_i}.$$

$$f(\beta|y, p, \sigma_v^2) \sim N_p\left(\beta^*, \sigma_v^2\left(\sum_i X_i X_i^T\right)^{-1}\right)$$

$$f(\sigma_v^2|y,p,\beta) \sim \textit{Gamma}\left(\frac{m}{2}+a, \frac{1}{2}\sum_i(\theta - X_i^T\beta)^2 + b\right).$$

The estimated value for each parameter is achieved by constructed marginal posterior distribution for the corresponding parameter, such as following:

The posterior distribution for the parameter p_i, $f(p_i|\beta,\sigma_v^2,y)$:

$$f(p|y) \propto f(p|y,\beta,\sigma_v^2,\sigma_e^2).$$

The posterior distribution for regression coefficients β:

$$f(\beta|y,p,\sigma_v,\sigma_e^2) = f(\beta|p,\sigma_v^2).$$

The posterior distribution for a random effect σ_v^2:

$$f(\sigma_v^2|y,p,\beta) = f(\sigma_v^2|p,\beta).$$

Since mean posterior and variance posterior for each parameter above could not be achieved analytically, a numerical approach using MCMC (Markov Chain Monte Carlo) is applied by generating random samples. Convergency tests of parameter estimated are based on trace plot, autocorrelation plot and Monte Carlo error (Yanuar, 2015).

For model selection and model comparison in Bayesian model, the most popular criteria is the Deviance Information Criteria (DIC) (Chan & Grant, 2016; Spiegelhalter, 2002). Assume that a model for observed data y postulates a density $p(y|\theta)$ (including covariates etc.). The deviance, $D(\theta) = -2\log\{p(y|\theta)\}$ is considered as a function of θ. Spiegelhalter et al. (Spiegelhalter et al., 2002, 2014) proposed this criterion based on the principle DIC. DIC is estimated by submission of goodness of fit and complexity. DIC is defined, analogously to AIC, as

$$\textit{DIC} = D(\theta) + 2p_c \tag{11}$$

with $p_c = E_{\theta|y}[-2\log\{p(y|\theta)\}] + 2\log[p\{y|\tilde{\theta}(y)\}]$ or p_c is equal to posterior mean deviance and deviance of posterior means. DIC can be easily monitored in BUGS.

3. Results and discussion

3.1. Model construction

In this section, we fit the data by employing direct, ordinary least square and HB-SAE estimation methods to construct the estimator for a proportion of the number of BPJS participants at the subdistrict level in Padang. The estimation results of the three methods are then compared to identify the suitable estimator to predict the proportion of binary response cases, i.e., as a participant or not, denoted by p.

Parameter estimated for p in direct method for each subdistrict is formulated as following:

$$\hat{p}_i^{DM} = \frac{y_i}{N_i}, i = 1,2,\ldots,10 \tag{12}$$

The estimated values for $\hat{p}_i^{DM}$ at each district are provided in Table 1 in the seventh column. The highest proportion of BPJS's participants is in Padang Timur ($\hat{p}_{TM} = 0.98$), and the lowest proportion is at Pauh ($\hat{p}_{pauh} = 0.74$).

We then implement OLS to construct the proposed model first, which is used to predict the proportion of BPJS's participants in Bungus Telung Kabung then. To do this analysis, we plot the trend between N_i and $\hat{p}_i$ for $i = 1,\ldots,10$ as provided in Fig. 1. The figure demonstrates that the pooled trend between N_i and $\hat{p}_i$ is in horizontal with a black solid line. The trend for two subclasses is also detected which each have a significant positive slope. The upper left (indicated with dash red line, denoted as Type = 0) trend corresponds to the sub-districts with less population, i.e., below 71 thousand inhabitants, with Padang Timur being an exception. The bottom right (indicated with dash blue line, denoted as Type = 1) corresponds to sub-districts with more population, except for Padang Utara and Pauh.

We then reran several candidate models using the OLS approach. We hypothesized several combination models,

Table 2
Candidate models based on OLS method

Variable	Linear model				Logit model		
	Lin1	Lin2	Lin3	Lin4	Logit1	Logit2	Logit3
Constanta	0.486**	0.848***	0.837***		−0.626		
X_1	0.018*	−0.002			0.068	0.033**	
X_2	0.007*	0.004			0.034*	0.045*	
N_i		0.001	0.001***	0.015***		−0.000	0.013***
Type		−0.145**	−0.163***	0.686***		−0.287	−0.852**
N_i*Type				−0.013***			
R^2_{adj}	46.51%	79.84%	81.12%	**98.30%**	39.22%	96.13%	78.65%

*$p < 0.1$; **$p < 0.05$; ***$p < 0.01$.

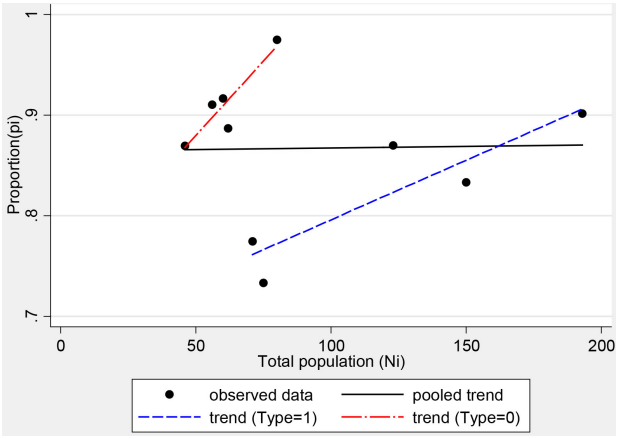

Fig. 1. The pooled trend between N_i and $\hat{p}_i$.

including explanatory variables (which have a significant correlation with the response), i.e., age (percentage of individuals more than 50 years old, X_i) and percentage of health complaints that ever had (X_2), Type of subdistricts (0 or 1), number of populations in each subdistrict N_i, and interaction between Type and N_i. The complete data for X_1 and X_2 are provided in Table 1. Here, we proposed four models with the proportion directly (named is as "Lin"), and three models with logit as the dependent variable (named it as "Logit"), as provided in Table 2.

As shown in Table 2, X_1 and X_2 are not significant determinants in any specification, with the exception of Logit1, where X_2 is significant at 10%. However, the adjusted R^2 (R^2_{adj}) in Logit1 is low, at 46.51% only. Removing X_1 and X_2 and replacing them with N_i and Type improves prediction significantly. Furthermore, including the interaction of N_i and Type performs even better. As a result, the best models are Lin4 since it has the highest value of adjusted R^2 (98.3%). Therefore, the proposed model based on OLS method obtained here is:

$$\hat{p}_i^{OLS} = 0.015 * N_i + 0.686 * \textit{Type} - 0.013 * N_i * \textit{Type} \tag{13}$$

The proportion for Bungus Teluk Kabung could be predicted by substituting $N_i = 25$ and Type $= 0$ and Type $=$ 1 to Eq. (13). Thus, there are two values here, for Type $= 0$, we obtain $\hat{p}_i = 0.375$, and for Type $= 1$ we obtain $\hat{p}_i = 0.736$. If we look at Fig. 1, the resemble most for Bungus Teluk Kabung is Type $= 1$. Thus, the estimated value for the proportion of BPJS participants in Bungus Teluk Kabung is 0.736.

We then employ the indirect estimation method or hierarchical Bayesian SAE for a binary response using the Binomial Logit-Normal link function. In this study, three hypothesis models are constructed to model the proportion of BPJS participants and several auxiliary variables. In SAE, the explanatory variables are known as auxiliary variables. Two auxiliary variables considered here are age (percentage of individuals more than 50 years old, X_1) and percentage of health complaints (X_2). Thus, we have three candidate models, as following,

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

Table 3
Candidate models based on HB-SAE

Model	Parameter	Mean	SD**	MC error	95% confidence interval		DIC
					Lower bound	Upper bound	
A	β_0	−1.5880	1.5860	0.0427	−3.4550	0.3224	60.7411
	β_1	0.2207*	0.0986	0.0027	0.1015	0.3377	
B	β_0	1.0850*	0.4290	0.0084	0.3997	1.5980	60.1072
	β_1	0.0719*	0.0342	0.0006	0.0324	0.1141	
C	β_0	−1.7610*	1.2520	0.0275	−3.3170	−0.2248	**57.5661**
	β_1	0.1902*	0.0807	0.0018	0.0610	0.2914	
	β_2	0.0521*	0.0265	0.0002	0.0209	0.0510	

*Significant at 0.05. **SD = standard deviation.

Table 4
Estimated proportion based on HB-SAE method

Subdistrict	Proportion	SD*	95% confidence interval	
			Lower bound	Upper bound
Koto Tangah	0.9102	0.0186	0.8858	0.9234
Kuranji	0.8550	0.0198	0.8297	0.8682
Lubuk Begalung	0.8577	0.0194	0.8344	0.8695
Lubuk Kilangan	0.8508	0.0273	0.8206	0.8999
Nanggalo	0.8840	0.0237	0.8541	0.9126
Padang Barat	0.9205	0.0304	0.8809	0.9543
Padang Selatan	0.8881	0.0238	0.8581	0.9276
Padang Timur	0.9332	0.0177	0.9115	0.9561
Padang Utara	0.8178	0.0261	0.7806	0.8501
Pauh	0.7401	0.0423	0.6851	0.7932

*SD = standard deviation.

The best model is chosen based on the smallest DIC value (Chan & Grant, 2016). Table 3 provides the estimated results for the regression coefficient (β), standard deviation, and DIC values for all three models. It can also be seen that all model parameters have converged because the MC error value is less than 5% of the standard deviation. It was shown also that all the auxiliary variables in the two alternative models, i.e., model B and model C, are significant. The level of significance is known from the confidence interval that does not contain zero values. In model A, parameter β_0 is not different than zero, so that model A cannot be used. Based on the DIC value, model C has the smallest DIC value among others. So, it can be concluded that the best model for estimating proportions in this study is model C. Model C assumes that age (percentage of individual more than 50 years old, X_i) and percentage of health complaints (X_2) could improve the precision of estimating the proportion of BPJS participants in the subdistricts in Padang.

Based on model C, the proportion of BPJS's participants at each subdistrict based on HB-SAE method could be formulated as follows:

$$\log it(\hat{p}_i^{HB\text{-}SAE}) = -1.7610 + 0.1902X_1 + 0.0521X_2$$

or

$$\hat{p}_i^{HB\text{-}SAE} = \frac{\exp(-1.7610 + 0.1902X_1 + 0.0521X_2)}{1 + \exp(-1.7610 + 0.1902X_1 + 0.0521X_2)} \tag{14}$$

The proposed model in Eq. (14) then is used to estimate the mean of proportion, standard deviation of proportion, and 95% confidence interval for estimated mean of proportion for each district. The results of estimation are provided in Table 4.

The subsequent analysis is the convergency test for each parameter estimated. The indicator to convergency test is based on trace plot, density plot and ACF (autocorrelation plot). Figure 2 provides the results of the convergency test for estimated parameters in the proposed model.

Figure 2 provides the trace plot in the right side, density plot in the middle and ACF plot in the left side for each parameter estimated. Based on the trace plot for all four parameters, it can be seen that the parameter estimation algorithm has converged because there is no specific pattern in the plot. From the density plot, it can be seen

Table 5
Estimated proportion and standard deviation based on direct method, OLS and HB-SAE

Subdistrict	Estimated proportion (standard deviation)		
	Direct method	OLS	HB-SAE
Koto Tangah	0.9015 (0.0214)	0.9100 (0.0206)	0.9102 (**0.0185**)
Kuranji	0.8333 (0.0333)	0.8600 (0.0283)	0.8550 (**0.0197**)
Lubuk Begalung	0.8699 (0.0325)	0.8200 (0.0346)	0.8577 (**0.0194**)
Lubuk Kilangan	0.9107 (0.0399)	0.8200 (0.0513)	0.8508 (**0.0273**)
Nanggalo	0.8871 (0.0427)	0.9100 (0.0363)	0.8840 (**0.0236**)
Padang Barat	0.8695 (0.0533)	0.6700 (0.0693)	0.9205 (**0.0304**)
Padang Selatan	0.9167 (0.0373)	0.8800 (0.0419)	0.8881 (**0.0237**)
Padang Timur	0.9750 (0.0177)	1.1700 (NA)	0.9332 (**0.0176**)
Padang Utara	0.7746 (0.0563)	0.7600 (0.0507)	0.8178 (**0.0261**)
Pauh	0.7432 (0.0589)	0.7800 (0.0489)	0.7401 (**0.0423**)

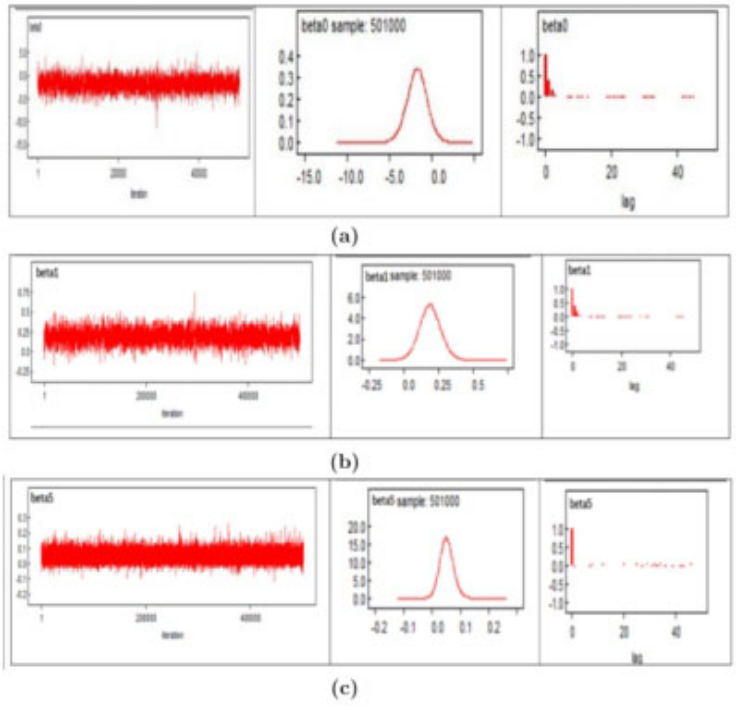

Fig. 2. Trace plot, density plot, ACF plot for (a) β_0, (b) β_1, (c) β_2, (d) σ_v^2.

that the distribution of the sample has approached the normal distribution. This indicates that the convergence of the algorithm has been achieved. Meanwhile, from the autocorrelation plot, it can be seen that the autocorrelation values in the first lag approach one and then the autocorrelation values in the next lag continue to decrease to 0. This indicates that there is a weak correlation in the chain. This weak correlation suggests that the errors are uncorrelated and the algorithm is already in the target distribution area.

3.2. Model selection

In this section, the results based on all estimation methods are compared to identify the best approach in estimating the proportion of BPJS's participants in each subdistrict in Padang. The best one should have the lowest value of standard deviation. Table 5 provides the comparison result of estimated proportions based on all three methods for each subdistrict. The estimation is based on the direct method using Eq. (12), while Eq. (13) is used to construct the estimated proportion based on OLS model, and the estimated proportion based on HB-SAE is by using Eq. (14).

Table 5 informs us that the standard deviation obtained from the SAE-HB method results in smaller values than other methods, indicated in bold. It could be concluded that the proportion of BPJS's participants obtained based on SAE-HB using model C results in a better model than other models. Thus, the final model for the proportion of BPJS's participants in Padang is formulated as:

$$\hat{p}_i^{HB\text{-}SAE} = \frac{\exp(-1.7610 + 0.1902X_1 + 0.0521X_2)}{1 + \exp(-1.7610 + 0.1902X_1 + 0.0521X_2)} \tag{15}$$

3.3. Predicting the proportion of BPJS participants in Bungus Teluk Kabung

After obtaining the best method with its proposed model to estimate the proportion of BPJS participants in each subdistrict in Padang. It has been informed that one subdistrict, that is Bungus Teluk Kabung has no complete information. The available data from this subdistrict is age (percentage of individual more than 50 years old, X_1) and percentage of health complaints (X_2). We then substitute the value of $X_1 = 14.28$ and $X_2 = 5.10$ to the Eq. (15) as the proposed model based on the best method. The estimated proportion of BPJS's participants in Bungus Teluk Kabung is 0.7722 with a standard deviation of 0.04311. It is informed that in average, for every 100 individuals living in Bungus Teluk Kabung, there are 77 individuals that have registered BPJS participants. Moreover, the estimated value for 95% confidence interval for the proportion is between 0.7066 to 0.8378. These values inform us that there are at least 70 participants out of every 100 individuals and at most 84 participants from every 100 individuals will pay a premium to the BPJS party. Reciprocally, the BPJS party should allocate the benefit funds paid in the future to at least 70 participants out of every 100 individuals and at most 84 participants from every 100 individuals.

This study identifies that the proportion of BPJS participants at Pauh and Bungus Teluk Kabung are less than 80%, i.e., the proportion at Pauh is 0.7432, and at Bungus Teluk Kabung is 0.7722. These results inform us that need to promote counseling on the importance of joining the BPJS program or health service quality should be improved in these subdistricts so that more individuals enroll in health insurance. This activity can be arranged based on the conditions of each sub-district then.

4. Conclusions

We have demonstrated the comparison result between classics method (direct and OLS method) and HB-SAE in estimating the proportion of BPJS participant in eleven districts in Padang, Indonesia. In this study, we assume that response is binomial distributed. We have proved that if we applied the direct and OLS method here, both methods resulted unsatisfactory estimated values. The direct method tends to result higher values of standard deviation and this method cannot result the prediction model. While OLS cannot estimate all values for proportion and its standard deviation.

This study found that the hierarchical Bayesian in small area estimation method yielded better estimation values than other methods. It is concluded that the small area estimation proposed in this study is suitable to be implemented in the case of binary data. The proposed model informed that percentage of individuals more than 50 years old and percentage of health complaints have a significant effect of improving the precision of mean response. Pauh and Bungus Teluk Kabung have lower proportion of BPJS participants than other districts, i.e., less than 80%.

Acknowledgments

This research was support by grant from DRPM, the Deputy for Strengthening Research and Development of the Ministry of Research and Technology/National Research and Innovation Agency of Indonesia, in accordance with Contract Number 163/SP2H/AMD/LT/DRPM/2020.

Conflict of interest

The authors declare no conflict of interest.

References

Chan, J.C.C., & Grant, A.L. (2016). On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics*, *14*(4), 772-802.

Diallo, M.S., & Rao, J.N.K. (2018). Small area estimation of complex parameters under unit-level models with skew-normal errors. *Scandinavian Journal of Statistics*, *45*(4), 1092-1116.

Duku, S.K.O. (2018). Differences in the determinants of health insurance enrolment among working-age adults in two regions in Ghana. *BMC Health Services Research*, *18*(1), 384.

Fay, R.E., & Herriot, R.A. (1979). Estimates of income for small places: An application of james-stein procedures to census data. *Journal of the American Statistical Association*, *74*(366), 269.

Higgins, J.P.T., & Thomas, J. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (Second edition). Wiley-Blackwell.

Mahumud, R.A., Sarker, A.R., Sultana, M., Islam, Z., Khan, J., & Morton, A. (2017). Distribution and determinants of out-of-pocket healthcare expenditures in bangladesh. *Journal of Preventive Medicine and Public Health*, *50*(2), 91-99.

Molina, I., Nandram, B., & Rao, J.N.K. (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical Bayes approach. *The Annals of Applied Statistics*, *8*(2), 852-885.

Nazir, N., Mir, S.A., & Bhat, M.I.J. (2016). Hierarchical bayes small area estimation under a unit level model with applications in agriculture. *Pakistan Journal of Statistics and Operation Research*, *12*(3), 491.

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

Salari, P., Akweongo, P., Aikins, M., & Tediosi, F. (2019). Determinants of health insurance enrolment in Ghana: Evidence from three national household surveys. *Health Policy and Planning*, *34*(8), 582-594.

Sari, A.D., & Yanuar, F. (2020). Hierarchical bayesian modelling in small area for estimating binary data. *Journal of Physics: Conference Series*, *1554*, 012049.

Spiegelhalter, D.J. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*(4), 583-639.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*(3), 485-493.

Spiegelhalter, D.J., Nicola, G.B., & Carlin, B. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B*, *64*(4), 583-639.

Sugasawa, S., & Kubokawa, T. (2020). Small area estimation with mixed models: A review. *Japanese Journal of Statistics and Data Science*, *3*(2), 693-720.

Sugasawa, S., Kubokawa, T., & Rao, J.N.K. (2018). Small area estimation via unmatched sampling and linking models. *TEST*, *27*(2), 407-427.

Sugasawa, S., Kubokawa, T., & Rao, J.N.K. (2019). Hierarchical Bayes small-area estimation with an unknown link function. *Scandinavian Journal of Statistics*, *46*(3), 885-897.

Tsujino, T., & Kubokawa, T. (2019). Empirical Bayes methods in nested error regression models with skew-normal errors. *Japanese Journal of Statistics and Data Science*, *2*(2), 375-403.

Ver Planck, N.R., Finley, A.O., Kershaw, J.A., Weiskittel, A.R., & Kress, M.C. (2018). Hierarchical Bayesian models for small area estimation of forest variables using LiDAR. *Remote Sensing of Environment*, *204*, 287-295.

Yanuar, F. (2015). The Use of Uniformative and Informative Prior Distribution in Bayesian SEM. *Global Journal of Pure and Applied Mathematics*, *11*(5), 3259-3264.

Yanuar, F., Eka Putri, N.C., & Yozza, H. (2019). Poisson gamma model in empirical Bayes of small area estimation (SAE). *Journal of Physics: Conference Series*, *1317*, 012006.

Yanuar, F., Zetra, A., Muharisa, C., Devianto, D., Putri, A.R., & Asdi, Y. (2019). Bayesian quantile regression method to construct the low birth weight model. *Journal of Physics: Conference Series*, *1245*, 012044.

Yoshimori, M., & Lahiri, P. (2014). A new adjusted maximum likelihood method for the Fay-Herriot small area model. *Journal of Multivariate Analysis*, *124*, 281-294.

You, Y., & Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *Canadian Journal of Statistics*, *30*(1), 3-15.