Available online at http://scik.org
Commun. Math. Biol. Neurosci. 2023, 2023:20
https://doi.org/10.28919/cmbn/7779
ISSN: 2052-2541

POISSON-LOGNORMAL MODEL WITH MEASUREMENT ERROR IN COVARIATE FOR SMALL AREA ESTIMATION OF COUNT DATA

FEVI NOVKANIZA[1,*], KHAIRIL ANWAR NOTODIPUTRO[2], KUSMAN SADIK[2], I WAYAN MANGKU[3]

[1]Department of Mathematics, Universitas Indonesia, Kampus FMIPA UI Depok 16424, Indonesia

[2]Department of Statistics, IPB University, Kampus IPB Dramaga 16680, Indonesia

[3]Department of Mathematics, IPB University, Kampus IPB Dramaga 16680, Indonesia

Abstract. Small Area Estimation (SAE) is a statistical technique to estimate parameters of subpopulation containing small-size of samples. SAE is an indirect estimation method by utilizing the strength of the neighbor area and data sources outside the area so that the sample becomes more effective and reduces the variance of parameter estimators. This paper deals with non-symmetrical count data in SAE which can be modelled based on Poisson-Lognormal distribution using hierarchical Bayesian (HB) approach and information on covariate contains measurement error. We develop the Poisson-Lognormal model for predicting small area counts with structural measurement error in the area-specific covariate. To get the HB Bayes estimator for Poisson-Lognormal model with measurement error in its covariates, the Metropolis-Hastings (MH) algorithm from the Markov Chain Monte Carlo (MCMC) technique is used. The HB estimator performance is studied through simulation and implementation of real data to predict the illiteracy rate at the sub-district level in Kepulauan Riau Province Indonesia based on The National Socio-Economic Survey (Susenas) data in March 2020.

Keywords: Bayes; count; hierarchical; lognormal; measurement error; Poisson.

2020 AMS Subject Classification: 62J12, 62C12,62F15.

*Corresponding author

E-mail address: fevi.novkaniza@sci.ui.ac.id

Received October 13, 2022

1. INTRODUCTION

Small Area Estimation (SAE) is concerned with the development of statistical procedures for estimating small areas (or domains) when direct survey estimates (i.e. based only on the area-specific sample data) are unreliable or even can not be calculated because of the limited sample size. In the context of small area estimation, direct estimators lead to unacceptably large standard errors for areas with unduly small sample sizes or no sample units may be selected from some small domains. Data obtained from sample surveys can be used to derive reliable direct estimates for large areas or domains, but sample sizes in small areas are rarely large enough for direct estimators to provide adequate precision for small areas.

Another method which can be used to obtain higher precision in SAE are developed by linking some information in particular area with some other areas through appropriate model and it is called indirect estimation. There are two connecting models in indirect estimation, implicit and explicit models. Traditional methods of indirect estimation are based on implicit models that provide a link to related small areas through supplementary data. The explicit small area models make specific allowance for between-area variation. [1] introduce mixed models involving random area-specific effects that account for between-area variation beyond that explained by covariates included in the model.

The SAE method was first used by [2] through a linear mixed model with area random effects for estimating per capita income in a small area in the United States. [2] uses the Empirical Best Linear Unbiased Prediction (EBLUP) which is an improvement on the Best Linear Unbiased Prediction (BLUP). The BLUP estimator was first introduced by [3] for linear mixed models. In this case, ”best” stands for minimum mean square error among all linear unbiased predictors, ”linear” means that the predictor is a linear combination of the response variable values and ”unbiased” means that the expected value of the prediction error is zero. The BLUP methods described in [3] assumed that the variances associated with random effects in the mixed model are known. In practice, variance components are unknown and have to be estimated from the data. The predictor obtained from the BLUP when unknown variance components are replaced by associated estimators is called the empirical best linear unbiased predictor (EBLUP) and is described in [5]. While BLUP was originally proposed by [3], the empirical version of

BLUP (EBLUP) is related to the classical shrinkage estimator studied by [4], who established analytically that EBLUP improves on the sample means when the number of small areas is larger than or equal to three.

In addition to EBLUP, empirical Bayes (EB) and hierarchical Bayes (HB) estimation and inference methods have been also applied to small area estimation. The BLUP/EBLUP method is applied to the linear mixed model which handles the problem of estimating small areas, but they are not suitable for handling binary or count data. The EB and HB methods are applicable more generally in the sense of handling models for binary and count data as well as normal linear mixed models. Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data. But in the HB approach, unknown model parameters (including variance components) are treated as random, with values drawn from specified prior distributions. Posterior distributions for the small area characteristics of interest are then obtained by integrating over these priors, with inferences based on these posterior distributions. As in [1], Bayesian specifications for small area models with count data are derived from basic area level models for SAE, e.g. the Fay-Harriot model [2] or a generalized linear Poisson model. Bayesian data analysis is based on the posterior distribution of relevant parameters given the data. For some simple posterior distributions it is possible to find the exact form of the posterior distribution and to find explicit forms for the posterior mean. The posterior distribution represents both prior information and observational data, each of which is represented by the prior distribution of its likelihood function. The likelihood function represents the data condition, while the prior distribution determines the researcher's subjectivity. The prior distribution specification in Bayes' inference is also very important because this prior distribution will affect the inference regarding its posterior distribution. However, it is commonly the case that for reasonably realistic model, it is not possible to obtain a closed form for the posterior distribution. In this situation, simulation methods must be made to to posterior sampling i.e. obtain samples from the posterior distribution which then can be summarised to get estimates of relevant quantities.

The properties of the small area estimators derived under the SAE model are based on the hypothesis that the auxiliary data or covariates are available for all the areas and that they are

measured without error. However, it is sometimes difficult to find error-free covariates. When the covariate used in the SAE model is measured with error, using a small area estimator such as the Fay-Herriot estimator while ignoring measurement error may be worse than simply using the direct estimator [7]. Survey data can be used as an up-to-date covariate but it may contain measurement error. As in [8], ignoring measurement errors can produce biased estimators and lead to incorrect conclusions in data analysis. Many statistical procedures have been developed for statistical inference in measurement error models [9]. The measurement error model is a combination of measurement errors and unobserved random variables where the observed value is known. Since the true random variables are unobserved, it is important to estimate the nature of the distribution of populations through the distribution of prior random variables that contain measurement errors. The problem of estimating the density function of a true or unobserved random variable using its observed sample (surrogate) is called deconvolution [10].

There are many parametric and nonparametric methods proposed to estimate the density function of the unobserved random variables due to measurement errors. In the additive measurement error model, [10] proposed the deconvolution kernel density estimator to recover the unknown density function from contaminated data, where the kernel idea and the Fourier inverse were employed in the construction of the estimator. In this article, we estimate the density function of unobserved covariate under the framework of measurement error model by Empirical Bayes Deconvolution (EBD) method proposed by [11, 12]. The basic idea of the EBD method is modeling prior density as a member of exponential family density. Empirical Bayes inference assumes an unknown prior density of the unobserved covariate produces an independent observation as surrogated with a known probability distribution mechanism.

The HB predictor for the area-level small area means based on the Fay-Herriot model with measurement error in covariates proposed by [15] is constructed for continuous responses. But for count data, the basic Poisson model is widely used with the assumption that the mean and variance of the count data are equal. However, this equal mean-variance relationship rarely occurs in observational data. In most cases, the observed variance is larger than the assumed variance, which is called overdispersion. The Poisson formulation has to be adjusted to accommodate the extra variety of sample data observations. So, we develop the Poisson-Lognormal

model using the HB approach in the context of small area estimation for handling overdispersion in nonsymmetrical count data. When area-level covariate contains measurement error is used in the Poisson-Lognormal model, it is necessary to utilizing information from covariates, even though it may contain measurement errors. Measurement error might either be introduced by the measuring technique which involves the subjective judgment by human action or due to a cheaper or more convenient substitution of the correct quantity. Treating imprecise data as the true value of covariate can result in some misleading outcomes. In the structural measurement error model, the true covariate is treated as a random variable and its surrogate distribution is assumed unknown. We construct the HB estimator of the Poisson-Lognormal model with measurement error in covariates using the HB framework by adding the distribution of unobserved covariate as the prior distribution. We used simulation studies to investigate the modeling aspects of the Poisson-Lognormal model with measurement error in covariate and used a real data example to illustrate the practical usefulness of this work.

The remainder of this study is organized as follows. In Section 2, we specify the EBD method for estimating the density function of unobserved covariate. Section 3 presents the Poisson-Lognormal model with measurement error in covariate using HB approach. Sections 4 and 5 present the results of the simulation and real data example. Conclusions follow in Section 6.

2. THE EMPIRICAL BAYES DECONVOLUTION METHOD

Density estimation is a research area in statistics and has been studied extensively in many fields. In density estimation we are interested in determining an unknown function f, given only random samples or observations distributed according to this function. The goal of density estimation is to infer the probability density function (pdf) from observations of a random variable. Density estimation approaches can be classified into two groups: parametric and non-parametric density estimation. Parametric methods make strict prior assumptions about the form of the underlying density function. In parametric density estimation approach, where one assumes that the observations come from a parametric family of densities that has to be specified in advance. Then, the task is to estimate the parameters that fit the data best. For instance, a parametric approach may assume the random variables have a Normal distribution.

Such assumptions simplify the problem since only the parameters of the chosen family distribution need to be determined. In a normal distribution case, the density estimation reduces to determining the mean μ and standard deviation σ of the sample points. But sometimes it is not possible to make such strict assumptions about the form of the underlying density function and non-parametric methods are more appropriate in these situations. These methods make few assumptions about the density function and allow data to drive estimation process more directly.

[11] proposed the Empirical Bayes Deconvolution (EBD) method for estimating the density function of an unobserved random variable based on empirical Bayes strategy. Empirical Bayes methods are procedures for statistical inference in which the prior distribution is estimated from the data. In the EBD method, an unknown prior distribution $g(x)$ has yielded (unobservable) parameters and each of X_i independently produces an observed random variable W_i according to a known probability mechanism. From [13, 14], in classical structural measurement error model, the true or unobserved random variable X_i and its observed variable (surrogate) W_i can be written as additive measurement error model:

$$W_i = X_i + e_i, i = 1, .., m \tag{1}$$

[8] assume that X_i's are independent and identically-distributed (i.i.d.) as X, the errors ei's are i.i.d. as e, and X and e are mutually independent. We apply the EBD methodology in structural measurement error problems for estimating the density function of a true or unobserved covariate X_i using observational sample W_i. The Bayes deconvolution problem is one of recovering $g(x)$ from observed sample W and g is restricted to be in a parametric exponential family.

$$X_i \overset{iid}{\sim} g(x), i = 1, ..., N \tag{2}$$

$$W_i \overset{iid}{\sim} p_i(W_i|X_i) \tag{3}$$

and the marginal density of W_i is

$$f(w) = \int_T p(w|x)g(x)dx \tag{4}$$

where the integral is taken over the X-space, T. [11] proposed a likelihood approach to Bayes deconvolution problem in (1) with the prior $g(x)$ modeled by an exponential family of densities

on the T, which is assumed to be a finite discrete support set

$$T = (x_{(1)}, ..., x_{(m)}) \tag{5}$$

and the support set of the observations W_i is also assumed finite and discrete $W = (w_{(1)}, ..., w_{(n)})$. The unknown prior density $g(x)$ is an m-vector $g = (g_1, g_2, .., g_m)^t$ specifying probability g_j on x_j,

$$g = g(\alpha) = \exp[Q\alpha - \phi(\alpha)]. \tag{6}$$

where α is a p-dimensional parameter vector while Q is a known $m \times p$ structure matrix with jth row denoted by Q_j^t. The jth component of $g(\alpha)$ is

$$g_j(\alpha) = \exp[Q_j^t - \phi(\alpha)], j = 1, .., m \tag{7}$$

where the function $\phi(\alpha)$ normalizes $g(\alpha)$ so that its components sum to 1:

$$\phi(\alpha) = \log \sum_{j=1}^{m} \exp(Q_j^t \alpha). \tag{8}$$

Let $p_{ij} = p_i(W = w_{(i)} | X_i = x_{(j)})$ and define the $n \times m$ matrix $P = (p_{ij})$, so the marginal density of W between $g(x)$ and $f(w)$ can be represented as matrix multiplication

$$f = Pg. \tag{9}$$

Define the count vector $y = (y_1, ..., y_n)^t$ with $y_i = -W_i = w_{(i)}$ for $i = 1, 2, ..., n$ as a sufficient statistics for g and $y \sim mult_n(N, f)$. The log-likelihood function $l(\alpha)$ for y the parameter vector $\alpha = (\alpha_1, ..., \alpha_p)^t$ is

$$l(\alpha) = \sum_{i=1}^{n} y_i \log f_i(\alpha), \tag{10}$$

where $f(\alpha) = Pg(\alpha)$. Define $B_i(\alpha) = \{b_{i1}(\alpha), ..., b_{im}(\alpha)\}^t$ with

$$b_{ij}(\alpha) = g_j(\alpha)\{p_{ij}/f_i - 1\}, \tag{11}$$

then the score function for α is

$$\dot{l}_i(\alpha) = (\frac{\partial l_i(\alpha)}{\partial \alpha_1}, ..., \frac{\partial l_i(\alpha)}{\partial \alpha_p})^t = Q^t B_+(\alpha) \tag{12}$$

which is needed for the maximum likelihood calculation. The maximum likelihood estimate $\hat{\alpha}$ for α is obtained by solving $\dot{l}(\alpha) = 0$.

3. THE POISSON-LOGNORMAL MODEL WITH MEASUREMENT ERROR IN COVARIATE

When the response variables are discrete such as binary, count, and multi-category, the Poisson regression model is a widely used model for count data [20]. But the assumptions of the Poisson regression model turn out to be unrealistic. More specifically, the Poisson model involves the assumption that the mean is equal to the variance and, therefore, the model cannot account for overdispersion. For this reason, alternative distributions are used which imply different mean and variance that handle overdispersion, such as negative binomial distribution. An alternative for Poisson regression is to assume that the logarithmic parameter of the Poisson distribution follows a normal distribution with unknown variance, and unknown mean that depends linearly on a vector of covariates. Unfortunately, this model is not trivial to analyze because the likelihood function is not available in closed form. This model has been suggested before as a plausible competitor to Poisson-Gamma, and Poisson-Inverse Gaussian [21]. [21, 22, 23] has explored the relationships between count response and error-free covariates in the Poisson-Lognormal model using HB approach.

[16] discussed the use of generalized linear mixed models for small area estimation under this situation, and provided hierarchical Bayesian methods. Within the HB framework, alternative model specifications can be considered when the variable of interest is a count or a proportion [17, 18]. According to [19], HB models have many advantages: (i) their specification allows taking into account the different sources of variation, and (ii) inferences are clear-cut and computationally feasible in most cases by using standard MCMC techniques. [19] proposed six HB area-level models for producing small area estimation for count data. They choose the HB approach since inference on non-standard specifications may not be feasible using classical estimation methods. One of the HB models which can be used in SAE is the Poisson-Lognormal model. The Poisson-Lognormal specification does convert a standard generalized linear mixed model for count variates into a proper HB form. The Lognormal stage depends on two sources of random variability, the sampling error and the random area effect. [1] represent has the relationship of count response and error-free covariates in the Poisson-Lognormal model as HB SAE model.

When covariate contains measurement error, the representation of Poisson-Lognormal model as HB model needs additional stage for considering the covariate measurement error. We consider the situation in which covariate data are available from a survey which are measured with error. We assume that the true or unobserved covariate X_i has an unknown probability distribution with distribution of its surrogate W_i is assumed known. As in [15], we adopt a Bayesian approach and rewrite the Poisson-Lognormal model with measurement error in covariate as a hierarchical model. We applied the EBD method in Section 2 to estimate the density function of unobserved covariate and use it as the prior distribution in HB stages. The simulation study and a real data application are conducted to investigate some related issues to the use of covariates measured with error in the Poisson-Lognormal model will be discussed. For simplification, we focus on one covariate measured with error. The Poisson-Lognormal model with measurement errorin its covariate can be written as the following HB stages model:

(1) Stage 1: *Sampling model*

As in the first stage of the HB Poisson-Lognormal model, the count response y_i is written as:

$$y_i|\lambda_i \overset{ind}{\sim} Poisson(\lambda_i), i = 1,..,m$$

(2) Stage 2: *The linking Model*:

$$log(\lambda_i) = x_i^t\beta + v_i$$

with area random effect $v_i \overset{ind}{\sim} N(0,\sigma_v^2)$ and

$$log(\lambda_i)|X_i,\beta,\sigma_v^2 \overset{ind}{\sim} N(x_i^t\beta,\sigma_v^2)$$

(3) Stage 3: *Density estimation of true covariate X using the EBD method*

First, we assume the observed measurement W_i as a Poisson random variable whose mean and variance equal to unobserved covariate X_i. The observed sample $W = \{W_1,...,W_m\}$ is the realization of unobserved covariate X_i. Denote the vector probability function of X as $f = (f_1,...,f_m)$. The vector probability f is the prior density of W and is modeled as a family of exponential family distribution. From the EBD method proposed by [11] we get the density estimation $f(x)$ which will use as prior density for X.

(4) Stage 4: *Prior Distribution*

Prior distribution of parameters of the Poisson-Lognormal model with measurement error in covariate X as follows:

$$f(x,\beta,\sigma_v^2) \propto f(x)f(\beta)f(\sigma_v^2)$$

with $X_i \sim f(x)$ from EBD method, $f(\beta)$, $\sigma_v^2 \sim IG(a,b); a \geq 0, b > 0$, parameter β and σ_v^2 are independent of each other.

(5) Stage 5: *The Posterior Distribution*

$$f(\lambda_i, X_i, \beta, \sigma_v^2 | y, W) \propto P(Y_i|\lambda_i) f(\lambda_i|X_i,\beta,\sigma_v^2) P(W_i|X_i) f(x) f(\beta) f(\sigma_v^2)$$

$$= (\prod_{i=1}^{m} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!})(\sigma_v^{-m} \prod_{i=1}^{m} e^{-\lambda_i}\lambda_i^{y_i-1} \exp[-\frac{1}{2\sigma_v^2}\sum_{i=1}^{m}(\log\lambda_i - X_i^t\beta)^2])(\prod_{i=1}^{m}\frac{e^{-X_i}X_i^{w_i}}{w_i!}) f(x) f(\sigma_v^2)$$

$$\propto (\sigma_v^{-m})(\prod_{i=1}^{m} e^{-(\lambda_i+X_i)}\lambda_i^{y_i-1}X_i^{w_i}) \exp[-\frac{1}{2\sigma_v^2}(\prod_{i=1}^{m}(\log\lambda_i - X_i^t\beta)^2] f(x) f(\sigma_v^2) \tag{13}$$

In situation the posterior distribution cannot be determined analytically or via grid approximation, we use simulation to approximate the posterior distribution and its characteristics. The joint posterior distribution representation in equation (13) cannot be obtained in closed form. It is difficult or impossible to simulate directly from a distribution, so we turn to indirect methods. If we use the Gibbs sampler, it requires generating samples from the full conditionals of each step $\lambda, X, \beta, \sigma_v^2$, given the remaining parameters and the data. Because it is not easy to find these full conditional distributions, so posterior distributions of model parameters are computed using the Markov Chain Monte Carlo (MCMC) method using the MH algorithm for estimating the model parameter $\hat{\beta}$ and $\hat{\sigma_v}^2$ with prior distribution from stage 4.

4. SIMULATION STUDY

In this section a simulation study is conducted to better understand how the performance of the HB Poisson-Lognormal measurement error in one covariate including structure of the data, estimation method and accuracy of the estimated parameters. The aim of the simulation study in this study are focused on the model performance, especially about the unbiasedness and accuracy of the parameter estimators. In addition, Bayes' inference based on this simulation study

needs to be evaluated whether the Markov chain has reached the stationarity of the desired posterior distribution or not. This process is carried out through convergence diagnostics using trace plots. Any MCMC scheme aims to produce (dependent) samples from a "target" distribution. If well constructed, the Markov chain is guaranteed to have the posterior as its stationary distribution. But it does not tell us how long we have to run it to convergence. The initial position may have a big influence and the proposal distribution may lead to low acceptance rates. The chain may get caught in a local maximum in the likelihood of surface.

A Markov chain mixes well if it can reach the posterior quickly, and moves quickly around the posterior modes. The behaviour of the Markov chain can do by graphical checks and compare estimators obtained from multiple runs from different initial conditions. The efficiency of the chain can be measured in terms of the variance of estimates obtained by running the chain for a short time. Often start the chain far away from the target distribution because target distribution unknown, so we need to do a visual check for convergence. The first "few" samples from the chain can be a poor representation of the stationary distribution. These are usually thrown away as "burn-in". In this simulation, we use M-H algorithm as a common algorithm for generating samples from a complicated distribution using MCMC. In the M-H algorithm, it start by assuming the distribution we want to sample from, has density proportional to some function f. We also need to pick a "proposal distribution" that changes location at each iteration in the algorithm.

In this simulation study, we assume the true covariate X is generated from Gamma distribution and its surrogate has Poisson distribution with $E(W_i) = X_i$. By using number of MCMC samples 1000000 with thinning 500, the output of the estimated posterior distribution is as follows:

TABLE 1. Posterior statistics summaries of HB estimator of Poisson-Lognormal model with measurement error in one covariate

Number of Area	Parameter	Cov $X_i \sim G(1,5)$		MSPE	Cov $X_i \sim G(1,0.2)$		MSPE
		Mean	Std dev		Mean	Std dev	
$m = 5$	β_0	1.1858	0.5001	1.6565	0.7323	0.5536	1.2799
	β_1	0.4010	0.6422		0.2362	0.1647	
	σ_v^2	0.5074	0.4503		0.3669	0.3451	
$m = 10$	β_0	0.7859	0.4397	1.9992	0.8766	0.5096	2.5211
	β_1	0.4938	0.3480		0.2657	0.1466	
	σ_v^2	0.3496	0.2880		0.9553	0.5710	
$m = 30$	β_0	0.8868	0.1924	3.4091	1.2341	0.1218	7.6777
	β_1	0.3322	0.1184		0.1564	0.0124	
	σ_v^2	0.1483	0.1176		0.1639	0.0800	

Markov Chain Monte Carlo (MCMC) diagnostics are tools that can be used to check whether the quality of a sample generated with an MCMC algorithm is sufficient to provide an accurate approximation of the target distribution. In particular, MCMC diagnostics are used to check; whether a large portion of the MCMC sample has been drawn from distributions that are significantly different from the target distribution and whether the size of the generated sample is too small. The results of the diagnostic convergence for the simulation study were carried out through a trace plot as shown in Figures 1 and 2. The trace plot shows the sampled values of a parameter over time and helps us to judge how quickly the MCMC procedure converges in the distribution or how quickly it forgets its starting values. The parameter trace plots are created to make sure that the prior distribution is well-calibrated which is indicated by parameters having sufficient state changes as the MCMC algorithm runs. It allows us to evaluate convergence and mixing of the chains visually. When the sampler has converged the chains show one horizontal band, as in the Figure 1. In an intuitive sense, stationarity means that the properties of a process generating samples do not change over time. In this case, trace plot is relatively constant between mean and variance.

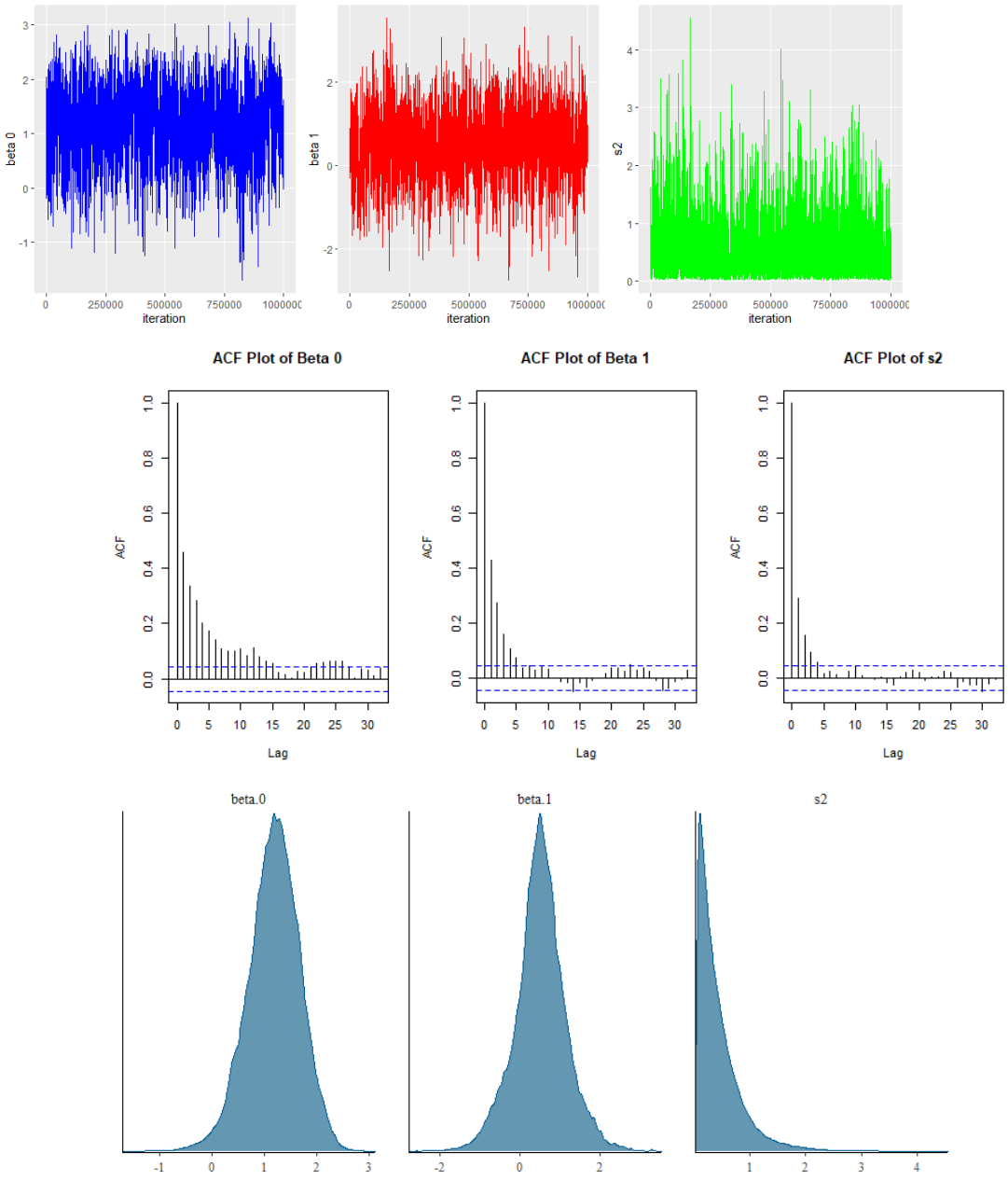

FIGURE 1. Markov Chain Monte Carlo (MCMC) diagnostics for HB estimator of Poisson-Lognormal model with true covariate $X_i \sim Gamma(1,5), i = 1,...,5$

Trace plots for $m = 5$ show that the MCMC simulation process based on the prior gamma distribution has reached convergence to the stationarity of posterior distribution. In the ACF plot for $m = 30$, there is a lot of serial correlation between successive draws. The chain is very slow in exploring the sample space. The sample space has been explored only a few times. In other words, there seem to be few independent observations in our sample or the effective size

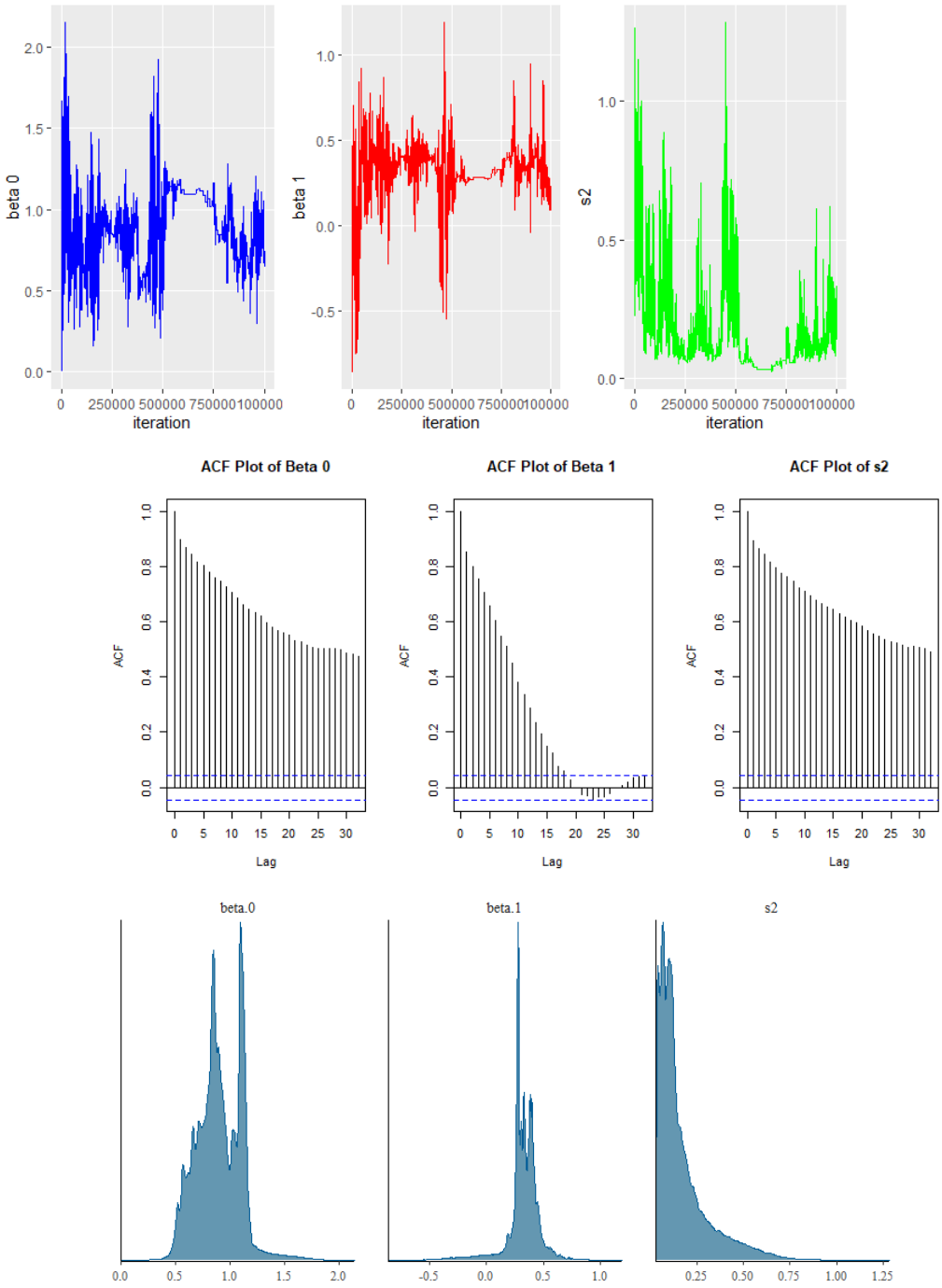

FIGURE 2. Markov Chain Monte Carlo (MCMC) diagnostics for HB estimator of Poisson-Lognormal model with true covariate $X_i \sim Gamma(1, 0.2), i = 1, ..., 30$

of our sample is too small. Based on trace plot is shown in Figure 1, the plot shown for $m = 5$ are quite ideal, compare with $m = 30$. It exhibits rapid up-and-down variation with no long-term trends or drifts. If we were to mentally break up this plot into a few horizontal sections, the trace within any section would not look much different from the trace in any other section. This indicates that the convergence in distribution takes place rapidly. Long-term trends or drifts in the plot indicate slower convergence. "long-term" is relative to the horizontal scale of this plot, which depends on the number of samples. As we take more samples, the trace plot gets squeezed together like an accordion, and slow drifts or trends eventually begin to look like rapid up-and-down variations. The rapid up-and-down motion means that the sampled value at any iteration is unrelated to the sampled value k iterations later, for values of k that are small relative to the total number of samples.

It can be seen that the estimated parameter values for each parameter of the prior gamma distribution are centered around the original generated value for $m = 5$. This shows that the MCMC simulation process based on the prior gamma distribution has reached convergence to the stationarity of posterior distribution. Furthermore, based on the acf plot and the density plot shown in Figure 1, it provides an indication that is consistent with the trace plot results. According to the results of the autocorrelation plot for each observed parameter, it shows almost the same pattern, a rapid decline after the second lag. Likewise, the probability density plot shows a pattern that the distribution of the observed parameter estimates tends to be symmetrical, except for density plot of random effect variance. However, based on the results in Figure 2 for $m = 5$, it shows that the estimated value of the observed parameters has relatively large fluctuations and it is indicated the larger sample size value is relatively less convergence of the posterior distribution. Consequently, when traces show a trend convergence has not been reached and more iterations are necessary. If convergence of Markov chain has not been achieved, it is necessary to increase the 'burn-in' period in the simulation process [24].

The simulation study indicates further studies to be carried out to validate the choice of prior distribution of model parameters. The lacks of the proposed models are the assumption we use in the EBD method for estimating the probability distribution of covariate measured with error is finite discrete support set and the assumption of independence of observed sample covariate that

is not always be satisfied in practice. Following the HB way of thinking, a Poisson-lognormal model could be more appropriate for taking explicitly into account the nature of the count response variable and distribution of covariate measurement error.

5. ILLITERACY RATE PREDICTION OF RIAU ISLAND PROVINCE, INDONESIA

As real data example to illustrate the construction of Poisson-Lognormal model with measurement error in covariate is illiteracy dataset of Riau Island Province Indonesia from the National Socio-Economic Survey (Susenas) in March 2020. In Indonesia, there has been a significant decline in the number of illiterate people, from 7.42% in 2009 to 4% in 2020 based on National Socio-Economic Survey (Susenas) 2020 by the Central Bureau of Statistics (BPS) Indonesia. Before 1993, the achievement of literacy is evaluated every 10 years through a census. After 1993, literacy data is collected annually in the form of Susenas which was held by BPS to provide data on human resources, particularly those that are related to socio-economy characteristics.

To evaluate the implementation of literacy education, a literacy indicator is used and for investigating the respondents' ability to read and write, the survey interviewers ask the respondent to demonstrate his/her ability to read a simple paragraph and to write simple sentences in the Indonesian language. The indicator is the ratio of those aged 15 and over who are literate to the total adult population (aged 15 and over). The literacy rate (LR) of the population aged 15 and over is defined as the number of literates aged 15 and over divide by total population aged 15 and over ($x100\%$). Another indicator used is the illiteracy rate, which refers to the ratio of illiterates among the total population falling into a certain age group. The illiteracy rate of the population aged 15 and above is defined as the number of illiterates aged 15 and over divide by the total population aged 15 and over ($x100\%$). LR can be used as one of people welfare indicators for measuring educational development in a certain area. The indicator can be measured if all variables related is available. Along with the establishment of regional policy, where regional governments had greater power to manage their region, the availability of LR on lower levels to monitor regional educational development is necessary.

For Susenas March 2020, the survey provides data at the city/district level and the data can also be categorized by rural and urban, sex, age groups, and family expenditure. These make

it possible for the country to observe any discrepancy in literacy level among groups. Due to the sampling design of Susenas, accommodated the only estimation on the district level and it will give high variance if it used to estimate on lower sub-district level, although still unbiased. The high variance will result in a broader confidence interval of estimation, which will make the estimation unreliable. One of the methods to obtain accurate estimators from inadequate sample size in a small area is a method of SAE with the count response, Y_i is the number of illiterate people aged 15 and over in sub-district i and from the same survey, we choose the number of people who live in rural areas as covariate measurement error W_i.

TABLE 2. Descriptive statistics of illiterate dataset

Variables	Min	Max	Mean	Std Dev
Y_i	0	1470	230.47	303.37
X_i	0	17067	3257.24	3842.98

Because Y_i is non-symmetrical count data and it has overdispersion problem as in Table 2, for predicting the number of illiterate people aged 15 years old and above, we construct the Poisson-Lognormal model as combination of the small area model and the hierarchical Bayes approach to handle the problem of estimating small areas of counts, which includes taking into account the random effects of the observed area and measurement error in covariate. We use the number of people who lived in rural areas in the same survey as covariate measurement error. As the Poisson-Lognormal model contains three parameters and the Bayes estimates can not be obtained in closed form, the Metropolis-Hastings algorithm is used to generate MCMC samples from the conditional posterior density of each parameter and produced parameter estimates as shown in Table 3 as follows

TABLE 3. Posterior statistics summary of HB estimator

Parameter estimation	Mean posterior	Std dev	HPD interval	
			25%	95%
$\hat{\beta}_0$	0.0686	1.1419	0.0183	0.1189
$\hat{\beta}_1$	-0.0289	0.9625	-0.0713	0.0135
$\hat{\sigma}_v^2$	92061.6190	16.1051	92060.9095	92062.3284

Based on the results of data analysis and various model diagnostic criteria such as convergence test of MCMC through traceplot, the autocorrelation plot of each parameter estimator and the MSPE value, it is concluded that the Poisson-Lognormal model with covariates containing measurement errors can be used to predict the illiteracy rate at the sub-district level in Kepulauan Riau Province where Suak Midai District has the highest illiteracy rate of 30.88%. In addition, it can also be obtained a prediction of the illiteracy rate of the Kepulauan Riau Province in 2020 is 3.66%. This prediction is greater than the estimated illiteracy rate of the Kepulauan Riau Province in 2020 based on the publication of Riau Island Province in 2021 which is only 1%.

6. CONCLUSIONS

The main goal of this research was to develop the Poisson-Lognormal model with measurement error in its covariate using the HB approach in the context of small area estimation for nonsymmetrical count data because of overdispersion. The covariate measured with error is considered to be a random variable and modeled as a structural measurement error model. Instead of covariate observed directly, we observe a contaminated version of true covariate or surrogate. One of the more significant findings to emerge from this study is that the true or unobserved covariate can be modeled as a structural measurement error model and the density function is estimated using Empirical Bayes Deconvolution (EBD) method. In the EBD method, the probability distribution of an unobserved covariate is considered as the prior distribution of its surrogate which the probability distribution is known.

As in the HB framework, we developed four stages of the HB approach by adding the estimated probability distribution of unobserved covariate as the prior distribution. The findings of this research provide insights into the Poisson-Lognormal model with measurement error in covariate which has a small MSPE value for smaller sample sizes. The research has also shown that the stationarity of posterior distribution of parameters in the MCMC simulation process using the Metropolis-Hastings (MH) algorithm reaches convergence rapidly for smaller sample sizes. This gain in efficiency is also visible when the model is applied to a data set consisting of illiterate data at the sub-district level in Kepulauan Riau Province obtained from Susenas in March 2020 and the number of people who live in rural-area as covariate measured with error.

Generally, this research urges the use of the Poisson-Lognormal model with measurement error in its covariate as an alternative model to small area estimation for nonsymmetrical count data such as the Poisson-Gamma model since it allows the non-conjugate prior (lognormal distribution) and considers covariate measured with error which comes from a survey. For covariate measured with error, estimating the probability distribution by EBD method is an alternative way for prior distribution in HB stages for Poisson-Lognormal model with measurement error in its covariate. The Poisson-Lognormal model with measurement error in covariate based on HB approach frameworks fully relies on the prior distribution. The most important limitation lies in this study is arguments about the choice of prior distributions. Further research could also be conducted to explore the interdependence between measurement error and covariate, and whether this impacts the results in the EBD method.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] J.N.K Rao, I. Molina, Small area estimation, 2nd Eds. John Wiley & Sons, Hoboken, 2015.

[2] R.E. Fay III, R.A. Herriot, Estimates of income for small places: an application of James-Stein procedures to census data, J. Amer. Stat. Assoc. 74 (1979), 269–277. https://doi.org/10.1080/01621459.1979.10482505.

[3] C.R. Henderson, Estimation of genetic parameters, Ann. Math. Stat. 21 (1950), 309–310.

[4] C.R. Henderson, Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. Third Berkeley Symp. Math. Statist. Probab. 1 (1956), 197–206.

[5] H.O. Hartley, J.N.K. Rao, Maximum-likelihood estimation for the mixed analysis of variance model, Biometrika. 54 (1967), 93–108. https://doi.org/10.1093/biomet/54.1-2.93.

[6] Y. You, J.N.K. Rao, Hierarchical Bayes estimation of small area means using multilevel models, Survey Methodol. 26 (2000), 173–181.

[7] L.M.R. Ybarra, S.L. Lohr, Small area estimation when auxiliary information is measured with error, Biometrika. 95 (2008), 919–931. https://doi.org/10.1093/biomet/asn048.

[8] X.F. Wang, B. Wang, Deconvolution estimation in measurement error models: the R package decon, J. Stat. softw. 39 (2011), 10.

[9] R.J. Carroll, D. Ruppert, L.A. Stefanski, et al. Measurement error in nonlinear models: a modern perspective, CRC Press, New York, 2006.

[10] L.A. Stefanski, R.J. Carroll, Deconvolving kernel density estimators, Statistics. 21 (1990), 169–184. https://doi.org/10.1080/02331889008802238.

[11] B. Efron, Empirical Bayes deconvolution estimates, Biometrika. 103 (2016), 1–20. https://doi.org/10.1093/biomet/asv068.

[12] B. Narasimhan, B. Efron, deconvolveR: A G-modeling program for deconvolution and empirical Bayes estimation, J. Stat. Softw. 94 (2020), 1–20.

[13] J.P. Buonaccorsi, Measurement error: models, methods, and applications, CRC Press, 2010.

[14] G.Y. Yi, Statistical analysis with measurement error or misclassification, Springer New York, 2017. https://doi.org/10.1007/978-1-4939-6640-0.

[15] S. Arima, G.S. Datta, B. Liseo, Bayesian estimators for small area models when auxiliary information is measured with error, Scand. J. Stat. 42 (2014), 518–529. https://doi.org/10.1111/sjos.12120.

[16] M. Gosh, K. Natarajan, T.W.F Stroud, et al. Generalized linear models for small-area estimation, J. Amer. Stat. Assoc. 93 (1998), 273–282.

[17] G.S. Datta, M. Ghosh, Bayesian prediction in linear models: Applications to small area estimation, Ann. Stat. 19 (1991), 1748–1770. https://www.jstor.org/stable/2241902.

[18] T. Maiti, Hierarchical Bayes estimation of mortality rates for disease mapping, J. Stat. Plan. Inference. 69 (1998), 339–348. https://doi.org/10.1016/s0378-3758(97)00165-1.

[19] M. Trevisani, N. Torelli, Hierarchical Bayesian models for small area estimation with count data, Working Paper 115, Dipartimento di Scienze Economiche e Statistiche, Universita Degli studi di Trieste, (2007).

[20] J.M. Hilbe, W.H. Greene, Count response regression models, in: C.R. Rao, J.P. Miller, D.C. Rao, (eds). Epidemiology and medical statistics, Elsevier Handbook of Statistics Series, London, 2007.

[21] E.G. Tsionas, Bayesian analysis of poisson regression with lognormal unobserved heterogeneity: with an application to the patent-R&D relationship, Commun. Stat. - Theory Methods. 39 (2010), 1689–1706. https://doi.org/10.1080/03610920802491774.

[22] L.F. Miranda-Moreno, L. Fu, F.F. Saccomanno, et al. Alternative risk models for ranking locations for safety improvement, Transport. Res. Record. 1908 (2005), 1–8. https://doi.org/10.1177/0361198105190800101.

[23] S.H. Ong, W.J. Lee, Y.C. Low, A general method of computing mixed Poisson probabilities by Monte Carlo sampling, Math. Computers Simul. 170 (2020), 98–106. https://doi.org/10.1016/j.matcom.2019.09.003.

[24] W.R. Gilks, G.O. Roberts, Strategies for improving MCMC. In: Gilks, W.R., Richardson, S., Spiegelhalter, D.J. (eds.) Markov Chain Monte Carlo in Practice, pp. 89–114. Chapman & Hall, London (1996).