The Small Area Estimation by Using Empirical Bayes Method

Nurul Astuty Yensy*

Mathematics Education Study Program, University of Bengkulu
**Corresponding author. Email: nurulastutyyensy@unib.ac.id*

ABSTRACT
The Small Area Estimation (SAE) is useful for estimating subpopulation parameters with small sample size. Various methods have been developed to estimate the area parameters, especially model-based methods. The Empirical Bayes is a method that can be used to estimate small area parameters. A small area is defined as a subpopulation (area) that is the size of a small sample. This Empirical Bayes method is suitable for use in counted data with the Poisson-Gamma model in addition to the Bayesian Hierarchical method. The purpose of this study was to examine the use of the empirical Bayes method in small area statistical estimation based on the Poisson-Gamma model with accompanying variables. The results showed that SAE for discrete data, namely the Empirical Bayes relative risk estimator from the Poisson-Gamma model with accompanying variables, gave estimation results with higher accuracy than the direct estimator of Standardized Mortality Ratio (SMR).

Keywords: *Empirical Bayesian, Small Area Estimation, Poisson-gamma Model.*

1. INTRODUCTION

The Small Area Estimation (SAE) is a statistical technique for estimating the parameters of a subpopulation whose sample size is small. This estimation technique uses data from large domains (i.e, census data, national socio-economic survey data) to estimate the variables of interest in smaller domains. The Small area is defined as a subpopulation whose sample size is small so that direct estimates cannot produce an accurate prediction [1,2,3,4,5]. In Indonesia, the subpopulation can be a province, district or city, and sub-district or village.

In general, there are three approaches to obtain parameter estimators in SAE, namely direct estimation, indirect estimation and composite estimation. If the sample size in the subpopulation is small and even zero, then the direct estimation statistics will have a large error range and even the estimation cannot be done [6,7,8,9]. To overcome this problem, the indirect can be used. Meanwhile, the composite estimation is the approximation that is carried out by weighting the direct estimator with the indirect estimator. This estimation is to balance the bias of the synthetic indirect estimator with the instability of the direct estimator by providing a weighted average for the two estimators [10.11.12.13.14].

The resulting estimator from this indirect estimation is the Empirical Best Linear Unbiased Prediction (EBLUP), Empirical Bayes estimator (EB), and Hierarchical Bayes estimator (HB). The EBLUP method is a method that is applied to linear mixed models, but the linear mixed models are designed for continuous variables so they are not suitable for binary or countable data. So, for binary data, the EB and HB methods are used to estimate the small areas.

One of the applications of the small area estimation for enumeration data is in disease mapping. In disease mapping, the small sample size (number of diseased cases) is a problem that is often faced because the area is very small, disease or both. So that the direct estimation in estimating relative risk, namely Standardized Mortality Ratio (SMR); becomes unreliable. An alternative method to deal with this problem is the Empirical Bayes method, with the model that is often used is the Poisson-Gamma model. The advantages of this empirical Bayes method are among others put forward by [15,16] which can to accommodate information between areas which is intended to reduce the number of error middle squares. Besides, the Empirical Bayes technique is suitable because it produces a relative risk estimator that is more reliable than the estimator of maximum probability [17,18,19,20]. Efforts to improve the estimation of relative risk can be done by entering the companion

variables into the Poisson-Gamma model. This study aims to examine the use of the empirical Bayes method in small area statistical estimation based on the Poisson-Gamma model with accompanying variables.

1.1. Small Area Model

The small area model is basic in estimating a small area. This model is grouped into two, namely the basic area level model and the basic unit level model [21,22,23, 24,25,26] In the basic area level model, it is assumed that the variable of interest is a function of the average response variable $\theta_i = g(\bar{Y}_i)$ for g(.) specific data relating to small area companion data $\underline{z}_i^T = (z_{1i,}..., z_{pi})^T$ and follow a linear model as follows:

$$\theta_i = \underline{z}_i^T \underline{\beta} + b_i v_i \quad i = 1,...m \tag{1}$$

b_i = a constant of known positive values

$\underline{\beta} = (\beta_1,...,\beta_p)^T$ is the vector of the regression coefficient that is sized p x 1

v_i = small area random effect, with $E(v_i) = 0$; Var $(v_i) = \sigma_v^2 > 0$

While the basic unit level model uses a nested error linear regression model as follows:

$$y_{ij} = x_{ij}^T \underline{\beta} + v_i \quad i = 1,...,n; \quad j = 1,...,m \tag{2}$$

y_{ij} = response variable

$x_{ij}^T = (x_{ij1,....,} x_{ijp})^T$

Meanwhile, the estimator of empirical Bayes θ_i are as follows:

$$\theta_i^{EB} = \theta_i^B(\underline{\beta}, \alpha) = \gamma_i \theta_i + (1 - \gamma_i) E(RR_i) \tag{3}$$

$\gamma_i = e_i \mu_i / (\alpha + e_i \mu_i)$

$E(RR_i) = \mu_i x E(\theta_i) = \mu_i x 1 = \mu_i = Exp(x_i^T \underline{\beta})$

= the i^{th} relative risk expected value which is an indirect estimator

$\theta_i = y_i / e_i$ = direct estimator (*standardized mortality ratio*) to θ_i, y_i and e_i which each represents the number of observations and the expected number of cases. [27,28,29].

2. RESEARCH METHODS

The data used in this study was secondary data obtained from the Bengkulu Provincial Health Office (2019). This data was in the form of the number of lung cancer patients (as a response variable) and the expected value of the number of lung cancer patients recorded for six years from 2012 to 2018 (small area) in Bengkulu Province. A companion variable was the percentage of work in agriculture, fisheries and forestry.

The research method used in estimating the relative risk of an area affected by disease was based on the direct standardized mortality ratio (SMR) estimator and the empirical Bayes estimator from the Poisson-Gamma model with accompanying variables which are described as follow:

a. Determine the expected value of the number of lung cancer sufferers.
b. Determine the Standardized Mortality Ratio (SMR)
c. Determine the approximate middle square of the error.
d. Determine α and β.
e. Determine the Empirical Bayes predictor.

Next, determine the Pearson residue, plot the linear predictor with the Pearson remainder and compare the goodness between the SMR and the empirical Bayes predictor from the Poisson-Gamma model with the accompanying variables by looking at the standard error value. The calculation process used SAS Software and Microsoft Excel.

3. RESULTS AND DISCUSSION

The results of descriptive statistics are shown in Table 1. Based on Table 1, it can be seen that there were districts that were not affected by the disease, which is stated by the minimum value of the direct estimator of Standardized Mortality Ratio (SMR). In fact, the district may have a relative risk of contracting the disease. The high expected value of the number of lung cancer patients, which is 88.61, is a result of the large population in the district. Meanwhile, based on the standard deviation value, the expected value of the number of lung cancer patients has a greater diversity than the number of lung cancer patients (observation).

Table 1. Descriptive statistics of lung cancer data in Bengkulu Province (2012-2018)

	Minimum	Maximum	Mean	Standard Deviation
Observation	0	39	9.57	7.91
Expected	1.06	88.61	9.57	13.16
SMR	0	6.53	1.53	1.32

2.1. The Estimation of Relative Risk

In this study, to determine the estimated relative risk of a district contracting lung cancer, namely the direct estimator of SMR and empirical Bayes estimator from the Poisson-Gamma model with accompanying variables. The goodness of the estimation of relative risk is measured from the level of accuracy indicated by the magnitude of the standard error. The graph of the relationship between the relative risk estimators and the default error is presented in Figures 1 and 2.

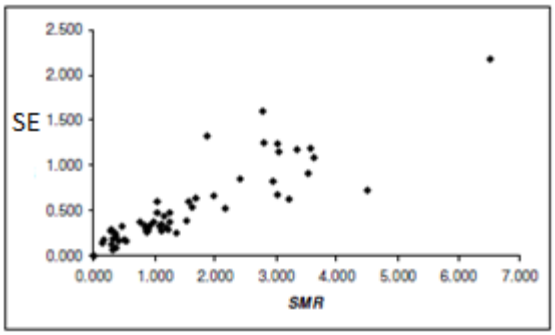

Figure 1.The plot of SMR and standard error.

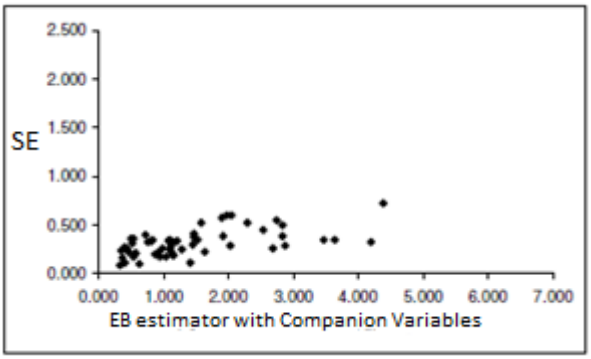

Figure 2.The Empirical Bayes Estimator plot with the companion variables and standard errors.

The Figures 1 and 2 generally show that the greater the relative risk estimator value, the greater the default error. The Empirical Bayes Estimator provides better accuracy than SMR. It can be seen that for the empirical Bayes estimator with companion variables, the greater the relative risk estimator values the smaller the default error when compared to other estimators.

Table 2 Estimation of Lung Cancer Data in Bengkulu Province

Average	*SMR*	The Empirical Bayesian estimator with Accompanying Variables
Relative Risk	1.53	1.45
Middle Square of Error	0.47	0.12
Standard Error	0.53	0.32

In Table 2, the information was obtained that on average the inclusion of the companion variables in the Poisson-Gamma model provided the better accuracy with a smaller standard error value than SMR. This was due to the accompanying variables that can be modeled optimally with the variables of interest. The optimum relationship is explained by the value of the goodness of fit and the residue in Table 3, Figure 3 and Figure 4 below:

Table 3. The Criteria for measuring model feasibility

Criterion	*DF*	*Value*	*Value/DF*
Deviance	54	62.30	1.15
Scaled Deviance	54	62.30	1.154
Pearson Chi-Square	54	57,50	1.07
Scaled Pearson X2	54	57.50	1.07
Log Likelihood		770.73	

Based on Table 3, it can be seen that this negative binomial regression model was feasible, because the Value or DF value for each criterion was less than two. Meanwhile Figures 3 and 4 also showed the fulfillment of the feasibility of the model with the regression equation reflecting the distribution of the data and the remainder tends to have no pattern.

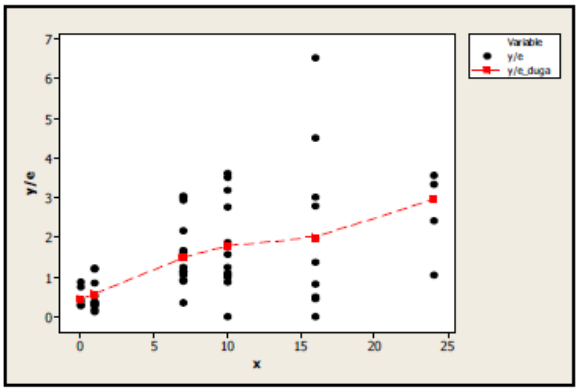

Figure 3.Relationship between relative risk (y / e) and co-variable: percentage of work in agriculture, fisheries and forestry (x).

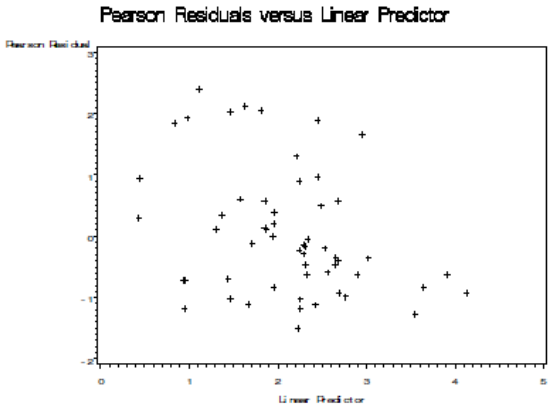

Figure 4. The Linear predictors with Pearson's remainder.

Based on the results and discussion above, it appears that estimation of small area using the Bayes Empirical method has a higher accuracy, this was in accordance with the opinion of [30,31,32,33] who suggest that to increase the effectiveness of sample sizes and reduce standard errors, namely by indirect

estimation, this estimation was to borrow strength from observing examples of adjacent areas by utilizing additional information, namely from data census and administrative records. The estimator that results from this indirect estimation is the best empirical linear unbiased prediction predictor.

4. CONCLUSION

Some of the conclusions obtained from the results of this study are as follow:

1. The Empirical Bayes relative risk estimator from the Poisson-Gamma model with accompanying variables provides predictions with higher accuracy than the standardized mortality ratio (SMR) direct estimator.
2. The Improved estimation by including companion variables in the Poisson-Gamma model produced an Empirical Bayes estimator with increased accuracy if the relationship between the accompanying variables and the variables of interest can be modeled optimally and derived from census data or administrative data.

REFERENCES

[1] Carlin BP, Louis TA. *Bayes and Empirical Bayes Methods for Data Analysis*. New York: Chapman & Hall. 2000.

[2] Dean CB, MacNab YC. Modelling of rates over a hierarchical health administrative structure. *The Canadian Journal of Statistics* . 2011, 29, 405-419.

[3] Fay RE, Herriot RA. Estimates of income for small places: an application of James-Stein procedures to census data. *Journal of the American Statistical Association*. 1999, 74, 269-277.

[4] Ghosh M, Maiti T. Small-area estimation based on natural exponential family quadratic variance function models and survey weights. *Biometrika*. 2014. 91, 95-112.

[5] MacNab YC *et al*. Estimation in Bayesian disease mapping. Biometric. 2004.

[6] Agresti A. *Categorical Data Analysis*. New Jersey: John Wiley & Sons. 2005.

[7] Bayarri MJ, Berger JO. The interplay of Bayesian and Frequentist analysis. Boca Raton: Chapman & Hall. 2004.

[8] McCullagh P, Nelder JA. *Generalized Linear Models*. London: Chapman & Hall. 2000.

[9] Piegorsch WW. Maximum likelihood estimation for the negative binomial dispersion parameter. *Biometrics*. 2003, 46, 863-867.

[10] Pringle DG. *Disease mapping: A comparative analysis of maximum likelihood and empirical Bayes estimates of disease risk*. 2001. [terhubung berkala]. http://www.nuim.ie/staff/dpringle/ebe.pdf [27 Juli 2019].

[11] Manton *et al*. Empirical Bayes procedures for stabilizing maps of US cancer mortality rates. *Journal of the American Statistical Association.* 2008, 84, 637-650.

[12] Marshall RJ. Mapping disease and mortality rates using empirical Bayes estimators. *Applied Statistics*. 2001, 40, 283-294.

[13] Rao JNK. Some recent advances in model-based small area estimation. *Survey Methodology*. 1999, 25, 175-186.

[14] Rao JNK. *Small Area Estimation*. New Jersey: John Wiley & Sons. 2003.

[15] SAS Institute Inc. SAS/STAT 9.1 User's Guide. 2004. [terhubung berkala]. http://support.sas.com/ [27 Maret 2017].

[16] Stern HS, Cressie N. Posterior predictive model checks for disease mapping models. *Statistics in Medicine*. 2000, 19, 2377-2397.

[17] Tsutakawa RK. Mixed model for analyzing geographic variability in mortality rates. *Journal of the American Statistical Association*. 2001, 83, 37-42.

[18] Wakefield J, Elliott P. Issues in the statistical analysis of small area health data. *Statistics in Medicine*. 2003, 18, 2377-2399.

[19] Wakefield J. Disease mapping and spatial regression with count data. 2006. [terhubung berkala]. http://www.bepress.com/uwbiostat/paper286.pdf [17 Juni 2019].

[20] Yasui *et al*. An empirical evaluation of various priors in the empirical Bayes estimation of small area disease risks. *Statistics in Medicine*. 2000, 19, 2409-2420.

[21] N. G. N. Prasad & J. N. K. Rao. The estimation of the mean squared error of small area estimators. *Journal of the American Statistical Association*, 2005. 85, 163–171. B. F.

[22] Qaqish& J. S. Preisser Resistant fits for regression withcorrelated outcomes: An estimating equations approach. *Journal of Statistical Planning and Inference*, 2006, 75, 415–431.

[23] P. Hall & T. Maiti. On parametric bootstrap methods for small area prediction. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 2006, 68, 221–238.

[24] K. SINHA and J. N. K. RAO. La revue canadienne de statistique Robust small area estimation Sanjoy. *The Canadian Journal of Statistics* Vol. 37 No. 3, 2009, pages 381–399.

[25] Isabel Molina and J. N. K. RAO. La revue canadienne de statistique Small area estimation of poverty indicators. *The Canadian Journal of Statistics* Vol. 38, , No. 20, 2010, pages 369–385.

[26] Y. You & J. N. K. Rao. A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights. Canadian Journal of Statistics, 2002. 30, 431–439.

[27] S. Haslett & G. Jones. Small area estimation using surveys and some practical and statistical issues. Statistics in Transition, 2005, 7, 541–555.

[28] W. Gonzalez-Manteiga, M. J. Lombard ´ ´ia, I. Molina, D. Morales & L. Santamar´ia . Bootstrap mean squared error of a small-area EBLUP. Journal of Statistical Computation and Simulation, 2008, 78, 443–462.
[29] Jiang, J & Lahiri, P. Estimation of finite population domain means: a model-assisted empirical best prediction approach. *Journal of the American Statistical Association*, 2006, 101, 301–311.
[30] Jiang, P & Lahiri, P. Empirical best prediction for small area inference with binary data. Ann. Inst. Statist. Math. 2001, 53, 217–243.
[31] S. Haslett & G. Jones. Small area estimation using surveys and some practical and statistical issues. Statistics in Transition, 2005, 7, 541–555.
[32] W. González-Manteigaa, M.J. Lombardíaa, I. Molinab, D. Moralesc , L. Santamaríac. Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model. *Computational Statistics & Data Analysis*. 2007, 51, 2720 – 2733.
[33] Butar, F.B., Lahiri, P. On measures of uncertainty of empirical Bayes small area estimators. *J. Statist. Plann. Inference*, 2003, 112, 63–76.