PAPER • **OPEN ACCESS**

Poisson gamma model in empirical Bayes of small area estimation (SAE)

To cite this article: Ferra Yanuar *et al* 2019 *J. Phys.: Conf. Ser.* **1317** 012006

IOP Conf. Series: Journal of Physics: Conf. Series **1317** (2019) 012006 doi:10.1088/1742-6596/1317/1/012006

Poisson gamma model in empirical Bayes of small area estimation (SAE)

Ferra Yanuar*, Nadya Cindy Eka Putri, Hazmira Yozza

Mathematics Department, Andalas University, Kampus Limau Manis Padang, Indonesia

* ferrayanuar@sci.unand.ac.id

Abstract. This study aims to describe the empirical Bayes estimator for small area estimation which use Poisson Gamma as prior distribution. The method then apply to generated data and model it with and without indicator variable. Parameter model estimated uses direct method and indirect method (known as empirical Bayes approach; with and without indicator variable). The choice of better estimator is based on MSE with Jackknife method. The criteria of acceptable proposed model are based on Deviance, Scaled Deviance, Pearson ChiSquare and Scaled Pearson Chi-square. This study proves that empirical Bayes in SAE with indicator variable result better estimated values than two other methods. All criteria of acceptable model indicate that proposed model could be accepted. Based on plot between the estimated residual versus predicted values of response variable informed that proposed model is plausible enough.

1. Introduction

Small area typically refers to a small geographic area or a demographic group for which very little information is obtained from the sample surveys. An empirical Bayes (EB) method uses sample survey data in conjunction with relevant supplementary data which are obtained from various administrative sources. The method has been found to be very useful in many applications of small-area estimation and related problems (Butar Butar & Lahiri, 2003; Sundara, Sadik, & Kurnia, 2017).

Direct survey estimators in the area-specifics small sample size, tend to yield unacceptably large standard errors (Jakimauskas & Sakalauskas, 2010). Therefore, it has to borrow the strength from related areas to improve the power of small sample size hope it could provide higher precision of estimates. Such of this estimation method is known as indirect estimators. Those indirect estimators are often based on mixed models and associated empirical Bayes estimators in which random effects represents area-specific effects.

Empirical Bayes estimator for small area places a prior distribution on population- specific area. Many distributions of priors have been used for this purpose, such as log normal, gamma and nonparametric priors (Jeong, 2010; John Quigley, Bedford, & Walls, 2007; Martuzzj & Elliott, 1996). In this study, we describe the Poisson gamma distribution for priors (Sakalauskas, 2010) and then applied it to generated data.

2. Small Area Estimation

Small area estimation are classified into two model, these are basic area level model and basic area level unit. In this paper, we limit the discussion in basic area level model only. A basic area level model assumes that some function $\theta_i = g(\mu_i)$ of the small area mean μ_i is related to the data from specific area $z_i = (z_{1i}, z_{2i}, \cdots, z_{pi})^T$ through a linear model with random effect v_i (Rao, 2003):

$$\theta_i = z_i^T \underline{\beta} + b_i v_{i,} \quad i = 1, \dots, m \tag{1}$$

Which b_i is known positive constant and $\underline{\beta} = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is a p x 1 vector of known regression parameters. Meanwhile v_i is random effects assummed as normal distribution or written as $v_i \sim N(0, \sigma_v^2)$, m is the number of small areas. In this study, we take $g(.)$ as the identity function. In the basic area level model, it is assummed that the direct estimate $\hat{\theta}_i$ is usually design as unbiased estimator, or $E(\hat{\theta}_i) = \theta_i$. It is usual to assume :

$$\hat{\theta}_i = \theta_i + e_i, \quad i = 1, \dots, m. \tag{2}$$

Where e_i is independent normal random variables which denotes the sampling errors associated with the transformed direct estimator $\hat{\theta}_i$. It is assumed $e_i \sim N(0, \psi_i)$

If we combine the equation (1) and (2), we will get this following equation:

$$\hat{\theta}_i = z_i^T \underline{\beta_i} + b_i v_i + e_i \tag{3}$$

This equation is known as Fay-Herriot model, special case of *generalized linear mixed model* in small area estimation that consist of fixed effect, that is β and random effect, that is v_i. Fay-Herriot used this basic area level model to estimate income per capita at any small areas in United State with population less that 1000. In this paper, we are concerned with the basic area-level model of Fay and Herriot (Rao, 2003).

Empirical Bayesian method of small area estimation places a prior distribution on area-specific risks (Clement, 2014). According to properties of this model the distribution of the number of events in populations follows to Poisson law, that intensity parameter is distributed according to Gamma distribution and the rates are estimated as the a posteriori means. Thus in this paper we address the empirical Bayesian estimation techniques for Poisson-Gamma model as well.

3. Results

3.1. Empirical Bayes Method in Small Area Estimation

Wakefield (2006) introduced Poisson-Gamma model two stage to estimate parameter in rare data. In first stage is assumed that random variable y_i has Poisson distribution or written as $y_i \sim Poisson(e_i \mu_i \theta_i)$ with probability density function is :

$$g(y_i | e_i \mu_i \theta_i) = \frac{e^{-(e_i \mu_i \theta_i)} (e_i \mu_i \theta_i)^{y_i}}{y_i!}, y_i = 0,1, \cdots \tag{4}$$

where $\mu_i = \mu\left(\underline{x_i}, \underline{\beta}\right)$ is regression model, $x_i = \left(x_{1i}, x_{2i}, \cdots, x_{pi}\right)^T$ is a vector of covariates and $\underline{\beta} = \left(\beta_1, \beta_2, \cdots, \beta_p\right)^T$ is regression coefficients. In the second stage, it is assumed the parameter θ_i has Gamma distribution or $\theta_i \sim iid\ Gamma(\alpha, \alpha)$ with probability density function is :

$$k(\theta_i) = \frac{\alpha^\alpha}{\Gamma(\alpha)} e^{-\alpha\theta_i} \theta_i^{\alpha - 1}, \theta_i > 0 \tag{5}$$

Based on equation (1) and (2) is obtained the joint probability density function as follows:

$$h(y_i, \theta_i) = \frac{e^{-(e_i \mu_i \theta_i)} (e_i \mu_i \theta_i)^{y_i}}{y_i!} \frac{\alpha^\alpha}{\Gamma(\alpha)} e^{-\alpha\theta_i} \theta_i^{\alpha - 1}, y_i = 0,1, \cdots; \theta_i > 0 \tag{6}$$

We could also get the marginal probability density function as follows:

$$\begin{aligned} m(y_i) &= \int_0^\infty h(y_i, \theta_i)\, d\theta_i \\ &= \binom{y_i + \alpha - 1}{\alpha - 1} \left(\frac{\alpha}{e_i \mu_i + \alpha}\right)^\alpha \left(1 - \frac{\alpha}{e_i \mu_i + \alpha}\right)^{y_i} \end{aligned} \tag{7}$$

We could identify that the distribution of equation (4) is negative binomial with mean and variance for y_i respectively are as follows:

$$E\left(Y_i|\underline{\beta},\alpha\right) = e_i\mu_i \text{ and } Var\left(Y_i|\underline{\beta},\alpha\right) = e_i\mu_i\left(1+\frac{e_i\mu_i}{\alpha}\right) \quad (8)$$

Then, we will estimate the posterior distribution for θ_i:

$$\pi(\theta_i|y_i,\beta,\alpha) = \frac{h(y_i,\theta_i)}{m(y_i)}$$
$$= \frac{(e_i\mu_i+\alpha)^{y_i+\alpha}}{\Gamma(y_i+\alpha)} e^{-(e_i\mu_i+\alpha)\theta_i}(\theta_i)^{y_i+\alpha-1}, \theta_i > 0 \quad (9)$$

Based on equation (9) we obtain the posterior distribution for θ_i is Gamma, or written as

$$\theta_i|y_i,\beta,\alpha \sim Gamma(y_i+\alpha, e_i\mu_i+\alpha) \quad (10)$$

Thus, the posterior mean and posterior variance are obtained from this Bayes estimate for θ_i, these are:

$$\widehat{\theta_i^B}(\beta,\alpha) = E_B(\theta_i|y_i,\beta,\alpha) = \frac{(y_i+\alpha)}{(e_i\mu_i+\alpha)}$$

And

$$Var_B(\theta_i|y_i,\alpha,v) = \frac{(y_i+\alpha)}{(e_i\mu_i+\alpha)^2} \quad (11)$$

Or it could be presented as (Wakefield, 2007):

$$\widehat{\theta_i^{EB}} = E(\theta_i|y_i,\beta,\alpha) = \widehat{\gamma_i}\widehat{\theta_i} + (1-\widehat{\gamma_i})E[RR_i] \quad (12)$$

Where $\widehat{\gamma_i} = e_i\widehat{\mu_i}/(\hat{\alpha}+e_i\widehat{\mu_i})$, $E[RR_i] = \widehat{\mu_i} \times E(\theta_i) = \widehat{\mu_i} \times 1 = \widehat{\mu_i} = \exp(\underline{x_i^T}, \underline{\hat{\beta}})$ is undirect estimator, $\widehat{\theta_i} = y_i/e_i$ is direct estimator for θ_i, y_i is number of observations and e_i is mean of number of observations.

The indicator to test the acceptable and the goodness of fit for the better proposed model is identified by calculating the value of Mean square error (MSE). (Rao, 2003) stated that better estimator tend to have smaller value of MSE.

(Ghosh, Maiti, & Roy, 2008) have proven that Bayes method is asymtotically unbiased estimator and empirical Bayes as well. The Jackknife method then be used to obtain this asymtotic unbiased estimator of MSE for $\hat{\theta}_i^{EB}$. Following is the formula for Jackknife estimator :

$$MSE_j\left(\hat{\theta}_i^{EB}\right) = \widehat{m}_{1i} + \widehat{m}_{2i}$$

where $\widehat{m}_{1i} = g_{1i}\left(\underline{\hat{\beta}}, \hat{\alpha}, y_i\right) - \frac{m-1}{m}\sum_{l=1}^{m}\left[g_{1i}\left(\underline{\hat{\beta}}_{-1}, \hat{\alpha}_{-1}, y_i\right) - g_{1i}\left(\underline{\hat{\beta}}, \hat{\alpha}, y_i\right)\right]$

and $\widehat{m}_{2i} = \frac{m-1}{m}\sum_{l=1}^{m}\left(\hat{\theta}_i^{EB} - \hat{\theta}_{i,-1}^{EB}\right)^2$ with $\hat{\theta}_i^{EB} = k_i\left(y_i, \underline{\hat{\beta}}, \hat{\alpha}\right)$ and $\hat{\theta}_{i,-1}^{EB} = k_i\left(y_i, \underline{\hat{\beta}}_{-1}, \hat{\alpha}_{-1}\right)$.

3.2. Application of Empirical Bayes in SAE Using Generated Data.

The theoritical framwork of empirical Bayes for SAE as presented above then applied to a set of generated data. Generate a set data of y_i and x_i represent response variable and indicator variable corresponding with normal distribution for both variables. Other variables such as n_i, e_i and direct estimator for $\widehat{\theta_i}$ are calculated based on the fixed formula. This following Table 1 presents the statistics of the data.

Table 1. Statistics from Generated Data

Statistics	y_i (Observations)	e_i (Expected Values)	Direct Estimator for $\widehat{\theta_i}$
Mean	1.680	1.680	1.071
Minimum	0.000	1.043	0.000
Maximum	6.000	2.508	3.833
Standard deviation	1.464	6.564	1.013

Table 1 informs us that minimum value for $\widehat{\theta_i}$ is 0.000. Standard deviation for expected value (e_i) is 6.564 more than standard deviation for observations, y_i. These results are hard to accepted. Thus, the informations obtained from classical appproaches (direct estimator) could not believed.

Then we did the estimate for parameter $\widehat{\theta_i}$ using proposed empirical Bayes (EB) based on Poisson Gamma model using without and with indicator variables. Table 2 is presenting the comparison results based on direct estimator and indirect estimator using empirical Bayes (EB) in SAE with and without indicator variables.

Tabel 2. The Results of Estimate for parameter $\widehat{\theta_i}$

Mean	Direct Estimator	Indirect Estimate	
		EB Without Variable	EB With Variable
Relative Risk	1.071195	1.005896	1.015143
MSE	0.723369	0.374396	0.287734
Standard error	0.041754	0.040824	0.040824

Based on Table 2, it informs that relative risk for both indirect estimator methods result smaller values than direct estimator as well as values of MSE and standard error. These results indicate that indirect estimator yield better values than direct estimator. Meanwhile for both indirect estimator, EB with variable result better values than EB without variable, based on the result from MSE and standard error.

From this study we can conclude that indirect estimator using empirical Bayes in small area estimation is better estimator than direct estimator. We also obtained than empirical Bayes in SAE with indicator variable tends to result better estimator than without indicator variable, since indicator variable could increase the goodness of fit of proposed model. Table 3 presents the indicator of goodness of fit for empirical bayes in SAE with indicator variable.

Tabel 3. The Criteria of Goodness of Fit for Propsoed Model (EB with Variable)

Criterion	DF	Value	Value/DF
Deviance	23	26.5527	1.1545
Scaled Deviance	23	26.5527	1.1545
Pearson Chi-Square	23	27.7788	1.2078
Scaled Pearson X2	23	27.7788	1.2078
Log Likelihood		-21.1788	

Last column in Table 3 is the values for each criteria of goodness of fit for proposed model. All these values are less than 2 which indicate that proposed model could be accepted. Thus based on this analysis, the equation of proposed model is as follows :

$$ln\left(\frac{\widehat{y_i}}{e_i}\right) = -0{,}2090 + 0{,}0139x_i \quad (13)$$

In this study we also check the estimated residual versus predicted values of response variable to assess the plausibility of the proposed model. Figure 1 presents the plot between these two variables (Yanuar, Ibrahim, & Jemain, 2013). This figure informs that no trends are detected. We could also conclude here that the estimated model which is obtained based on empirical Bayes analysis could be considere adequate and could be accepted.

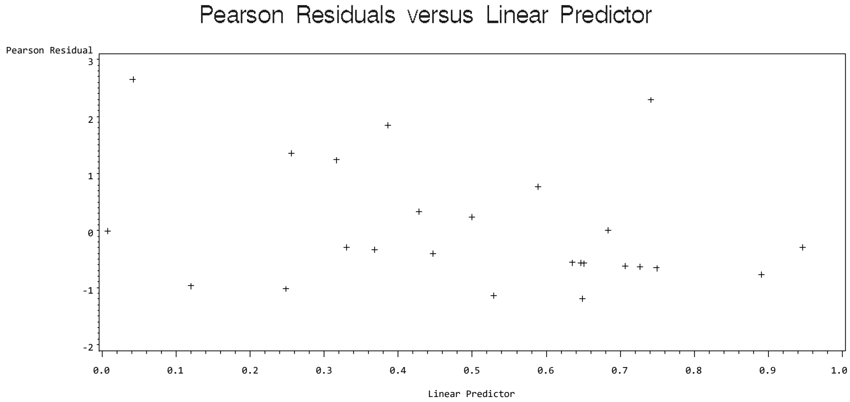

Figure 1. Plot of Residual versus Predicted Response Variable.

4. Discussions

This study described the empirical Bayes estimator for small area estimator which used Poisson Gamma as prior distribution. The method then applied to generated data and model it with and without indicator variable. Parameter model estimated used direct method and indirect method (known as empirical Bayes approach). The choice of better estimator based on MSE with Jackknife method. The criteria of acceptable proposed model were based on Deviance, Scaled Deviance, Pearson ChiSquare and Scaled Pearson Chi-square. Meanwhile to assess the plausibility of the proposed model was examined by plotting the estimated residual versus predicted values of response variable.

References

[1] Butar Butar, F., & Lahiri, P. (2003). On measures of uncertainty of empirical Bayes small-area estimators. *Journal of Statistical Planning and Inference*, *112*(1–2), 63–76. https://doi.org/10.1016/S0378-3758(02)00323-3

[2] Clement, E. P. (2014). Small Area Estimation with Application to Disease Mapping. *International Journal of Probability and Statistics*, *3*(1), 15–22. https://doi.org/10.5923/j.ijps.20140301.03

[3] Ghosh, M., Maiti, T., & Roy, A. (2008). Influence functions and robust Bayes and empirical Bayes small area estimation. *Biometrika*, *95*(3), 573–585. https://doi.org/10.1093/biomet/asn030

[4] Jakimauskas, G., & Sakalauskas, L. (2010). Empirical Bayesian Estimation for Poisson-Gamma Model. *International Conference*, (June).

[5] Jeong, K. M. (2010). Accuracy Measures of Empirical Bayes Estimator for Mean Rates. *Communications of the Korean Statistical Society*, *17*(6), 845–852.

[6] John Quigley, Bedford, T., & Walls, L. (2007). Estimating rate of occurrence of rare events with empirical bayes : A railway application. *Reliability Engineering & System Safety*, *92*, 619–627. https://doi.org/10.1016/j.ress.2006.02.007

[7] Martuzzj, M., & Elliott, P. (1996). Empirical Bayes Estimation of Small Area Prevalence of Non-Rare Conditions. *Statistics In Medicine*, *15*, 1867–1873.

[8] Rao, J. N. K. (2003). *Small Area Estimation. Wiley series in survey methodology*. A John Wiley & Sons. https://doi.org/10.1007/s13398-014-0173-7.2

[9] Sakalauskas, L. (2010). On the empirical bayesian approach for the poisson-gaussian model. *Methodology and Computing in Applied Probability*, *12*, 247–259. https://doi.org/10.1007/s11009-009-9146-2
[10] Sundara, V. Y., Sadik, K., & Kurnia, A. (2017). Cluster information of non-sampled area in small area estimation of poverty indicators using Empirical Bayes. In *AIP Conference Proceedings* (Vol. 020026, p. 020026). https://doi.org/10.1063/1.4979442
[11] Wakefield, J. (2007). Disease mapping and spatial regression with count data. *Biostatistics*, *8*(2), 158–183. https://doi.org/10.1093/biostatistics/kxl008
[12] Yanuar, F., Ibrahim, K., & Jemain, A. A. (2013). Bayesian structural equation modeling for the health index. *Journal of Applied Statistics*, *40*(6), 1254–1269. https://doi.org/10.1080/02664763.2013.785491
[13]