Forum Statistika dan Komputasi : *Indonesian Journal of Statistics*
ISSN : 0853-8115
Vol. 20 No.2, Oktober 2015, p: 108-115

available online at:
*journal.ipb.ac.id/index.php/statistika*

# SMALL AREA ESTIMATION FOR ESTIMATING THE NUMBER OF INFANT MORTALITY USING MIXED EFFECTS ZERO INFLATED POISSON MODEL

Arie Anggreyani[1], Indahwati[1], Anang Kurnia[1]

[1]Department of Statistics, Bogor Agricultural University, IPB Indonesia
E-mail : anggreyani@gmail.com, indahwati_43@yahoo.co.id,
anangk@apps.ipb.ac.id

## ABSTRACT

*Demographic and Health Survey Indonesia (DHSI) is a national designed survey to provide information regarding birth rate, mortality rate, family planning and health. DHSI was conducted by BPS in cooperation with National Population and Family Planning Institution (BKKBN), Indonesia Ministry of Health (KEMENKES) and USAID. Based on the publication of DHSI 2012, the infant mortality rate for a period of five years before survey conducted is 32 for 1000 birth lives. In this paper, Small Area Estimation (SAE) is used to estimate the number of infant mortality in districts of West Java. SAE is a special model of Generalized Linear Mixed Models (GLMM). In this case, the incidence of infant mortality is a Poisson distribution which has equdispersion assumption. The methods to handle overdispersion are binomial negative and quasi-likelihood model. Based on the analysis results, quasi-likelihood model is the best model to overcome overdispersion problem. However, after checking the residual assumptions, still resulted that residuals of model formed two normal distributions. So as to resolve the issue used Mixed Effect Zero Inflated Poisson (ZIP) Model. The basic model of the small area estimation used basic area level model. Mean square error (MSE) which based on bootstrap method is used to measure the accuracy of small area estimates.*

*Keywords : SAE, GLMM, Mixed Effect ZIP Model, Bootstrap*

## INTRODUCTION

Indonesia Demographic and Health Survey (DHSI) is a national designed survey to provide information regarding the birth rate, mortality rate, family planning and health. DHSI was conducted by BPS cooperation with BKKBN, KEMENKES and USAID. The DHSI 2012 data is compiled based on complex probability sample design which has small sample size and a lot of variables. One publication DHSI 2012 is the mortality rate. Based on the publication of DHSI 2012, infant mortality rate national for a period of five years before survey was 32 deaths per 1000 birth lives [1].

Infant mortality rate is one of indicators to measure the level of health development and quality of life a region or country. The government sometimes needs the indicator to make policy. However, data is only available to national scale or province while to district level, sub district or village is still inadequate and not easy to access. The number of infant mortality estimation is estimated by either direct estimation or design-based estimation. Direct estimation of small area generated large standard error. Therefore, direct estimation is not accurate to predict small area. One of statistics methods that can be used to handle these problems is small area estimation.

In recent years, statistics estimation has grown tremendously. Small area estimation compiled survey data with system registration data or with replenishment covariate. Small area estimation is indirect estimation or another meaning it borrows strength on related area to generate the best precision. The estimation of infant mortality number with small area approach has been done by many researchers. For Example Yadav & Ladusingh [14] at India estimated infant mortality rate with small area estimation through synthetic model, Hajarisman [4] measured infant mortality rate with small area estimation through two-level hierarchical Bayesian Poisson model, and report have been published by BPS and UsAID is small area estimation of nutritional status in Indonesia [2].

The incidence of infant mortality is a response variable which has Poisson distribution. Poisson distribution has equdispersion assumption. If expected value is greater than variance that is indicated overdispersion which leads to the invalid conclusions. The methods to handle overdispersion are binomial negative and approach quasi-likelihood model [3]. Negative binomial distribution (Poisson Gamma) accommodated dispersion parameter so that had large variance more than Poisson distribution [3].

The basic model of the small area estimation (SAE) used basic area level model. It is based on the availability of supporting data that is only there for a certain area level. SAE is a special model of generalized linear mixed model (GLMM) consists of fixed effect and random effect. This research used Poisson mixed model, binomial negative mixed model, and mixed model approach quasi-likelihood. However, after checking the residual assumptions, still resulted that residuals of model formed two normal distributions. To resolve the issue used Mixed Effect Zero Inflated Poisson (ZIP) Model. The method which was used to estimate predict variable is empirical bayes (EB). Thus account Mean square error (MSE) which was based on bootstrap method was used to measure the accuracy of small area estimates [6, 11].

## STATISTICAL MODELS AND ESTIMATION

### 2.1 Direct Estimation

The classical approach to estimate parameters of an area is based on the application design model sampling (design-based) which is known as direct estimation (direct estimation). Direct estimation method raises two important issues [7]. First, the estimate generated is not a biased estimator but has big variance which is produced from small sample size. Secondly, if on a smaller area-i is not represented in the survey, it is not possible to do direct estimation.

In this research, estimating the number of infant mortality at the level of district in West Java using IDHS data of 2012. The sampling method used in the IDHS 2012 with a three-stage method [1]. Step 1, choose a number of Primary Sampling Units (PSU) of the PSU sample frame in Probability Proportional to Size (PPS). Step 2, select census blocks by PPS. Step 3, choose the number of households in each census block systematically. So that the direct estimation becomes difficult. The methods can used Horvitz Thompson Method [5], Taylor linearization method [15], Jackknife method, etc.

Estimation of total Taylor method is defined as follows:

$$\hat{y}_i = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{j-1}^{m_{hi_i}} \omega_{hij} y_{hij} \qquad (1)$$

With $\omega_{ij}$ is weights to the area-i, household to-j and $y_{ij}$ is number of infant mortality to the area – i, household to – j Estimation variance of total Taylor method is defined as follows:

$$\hat{V}_h(\hat{Y}) = \frac{n_h(1-f_h)}{n_h-1}\sum_{i=1}^{n_k}(y_{hi.}-\bar{y}_{h..}) \qquad (2)$$

$$y_{hi.} = \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} \text{ and } \bar{y}_{h..} = {}^{1}\!/_{n_h}\left(\sum_{i=1}^{n_k} y_{hi.}\right).$$

With $n_k$ is the sample size area-i.

### 2.2 Indirect Estimation

Kurnia [7] and Sadik [12] stated that the sample size of area is small so that produce a great variance or even the areas that are not selected to be an example. Therefore the need to develop a method of indirect estimation. Estimation is not directly used to increase the size of the sample and degrade the effectiveness of variability which makes it more accurate. Estimates of indirect estimation which can "borrowing information" by using variable values of examples in other areas observed. Estimation is known as a small area estimation or Small Area Estimation (SAE). Small area estimation models consist of area-level model (Type-A) and Unit Level Model (Type-B) [11].

Area level models are used when information of auxiliary variable on the unit level is unknown so assumed $\theta_i = g(\bar{Y}_i)$ or $\theta_i = g\left(\sum Y_i\right)$ to *g(.)* Of certain associated with auxiliary variable on the area, is $x_i' = (x_{1i},...,x_{pi})'$ with the linear model is : $\theta_i = x_i'\beta + z_i v_i$, *i = 1, ..., m*. With $v_i \sim N(0,\sigma_v^2)$ is random variable on the area-i. Direct estimator $\hat{\bar{Y}}_i$ assumed to be known to draw conclusions about mean of small area $\bar{Y}_i$, namely $\hat{\theta}_i = \theta_i + e_i$, i = 1, ..., m. Where $e_i$ is the sampling error is normal distribution $v_i \sim N(0,\sigma_e^2)$ and $\sigma_v^2$ known. Both models were combined to obtain a deterministic model in $\theta_i$ the following : $\hat{\theta}_i = g(\bar{Y}_i) = x_i'\beta + z_i v_i + e_i$ , i = 1, ..., m.

### 2.3 Generalized Linear Mixed Model (GLMM)

Small area estimation model~~s~~ is a special form of the model GLMM. Generalized Linear Mixed Model (GLMM) is a linear model that includes the effect of random and fixed effect in the model with the response variable should not spread into the normal or exponential family. McCulloch and Searle [9] described the estimation of parameters in GLMM if y is free with the distribution of family exponentially and ***v*** distributed with parameter ***D*** :

$$f_{y_i|u}(y \mid v,\beta,\phi) = \exp\left\{\frac{y\theta_i - c(\theta_i)}{a(\phi)} + d(y,\phi)\right\} \qquad (3)$$

$u \sim f_v(v \mid D)$ Where $\theta_i = x_i'\beta + z_i v_i$,
Likelihood function :

$$L(\beta,\phi,D \mid y) = \int \prod_{i=1}^{n} f_{y_i|v}(y \mid v,\beta,\phi) f_v(v \mid D)dv \qquad (4)$$

The likelihood function above normally cannot be evaluated in closed form and has integral with dimensions equal to the number of levels of a random factor ***v***. Thus become an obstacle because of the likelihood function obtained is not simple anymore.

In the modeling problem when distribution of the data is unclear to detect, the likelihood function is not always biased. So the anoher developed approach which is known as quasi likelihood. In McCullagh and Nelder [8] and Pawitan [10] described the concept of quasi likelihood.

$$\sum_{i=1}^{n} \frac{\partial\mu_i}{\partial\beta} V(Y_i)^{-1}(Y_i - \mu_i) = 0 \qquad (5)$$

With assumption :

1. $E[Y_i] = \mu_i(\beta)$
2. $Var[Y_i] = \phi v_i(\beta)$

### 2.4 Overdispersion

Dispersion is a measure of the spread of a set of data to the mean of the data. Small dispersion value indicates a homogeneous range of data, while a large dispersion value indicates heterogeneity in the data. Dispersion values are identified by a ratio and constant. In the Poisson distribution has the characteristics of the average value equal to the value of diversity or equdispersion. However, the condition often occurs in the Poisson distribution is greater than the value of the average variety or overdispersion.

Some things that may cause extra overdispersion is diversity in the random variable that exceeds the range of Poisson random variables and the presence of outliers in the data. In the model GLMM, an event that follows the Y vector Poisson random but follows a certain distribution, the distribution of marginal will show overdispersion behavior.

$$E(Y) = E[E(Y|u)] = E(\mu_i)$$

$$\begin{aligned} Var(Y) &= Var[E(y_i|u)] + E[\text{var}(y_i|u)] \\ &= Var(\mu_i) + E(\mu_i) > \mu = E(\mu_i) \end{aligned}$$

There are several ways that can be used to detect overdispersion, are deviance and Pearson value which are ~~is~~ divided by degrees of freedom. If the obtained value is greater than 1 then indicate there is overdispersion. Where as if the value is less than 1 then indicate there is underdispersion. McCullagh and Nelder [8] and Hoef & Boveng [13], a common way to handle overdispersion with quasi-likelihood approach or negative binomial models.

### 2.5 Mixed Effect Zero Inlated Poisson Mixed Model

In mixture of distributions, the approach taken in the data which consists of two parts, namely the a binary part as a occurrence part and a positive part as an intensity part (Shin, 2012).

The general form:

$$f(y)=\begin{cases}\Pr(Y=0) & y=0,\\ (1-\Pr(Y=0))h(y) & y>0\\ 0 & y<0\end{cases}$$

Where *h(y)* is a probability density when *y > 0*, while $y_{ij}$ is an observation of j, j = 1, 2, ..., $n_i$ measurements on to-i, i = 1, 2, ..., m subject and $y_{ij}$ are all nonnegative.

**First**, Occurrence part was part of events: happen or not happen. The first variable is defined as an incident where:

$$R_{ij}=\begin{cases}0, if\ Y_{ij}=0,\\ 1, if\ Y_{ij}>0\end{cases}$$

Rij is defined as the conditional probability as follows:

$$\Pr(R_{ij}=r_{ij}\mid\theta_1)=\begin{cases}1-p_{ij}(\theta_1), if\ r_{ij}=0,\\ p_{ij}(\theta_1), if\ r_{ij}=1\end{cases}$$

Where $\theta_1=[\alpha_1^{'},b_{1i}]'$ is a vector of fixed occurrence effects $\alpha_1^{'}$ and occurrence effects of random unit $b_{1i}$. The first part of mixed-effects and mixed distribution model is an occurrence part (yij> 0) and involves logistic model for occurrence so that :

$$Logit(p_{ij}(\theta_1))=\log\left(\frac{p_{ij}(\theta_1)}{1-p_{ij}(\theta_1)}\right)=X_{1ij}^{'}\alpha_1^{'}+b_{1i} \qquad (1)$$

Where $X_{1ij}^{'}$ is the vector of covariates for the occurence.

**Second**, the second model is an intensity part and involves the non-zero part of the observed values. Where $S_{ij}\equiv[Y_{ij}\mid R_{ij}=1]$ be the intensity variable with probability density function $f(s_{ij}\mid\theta_2)$ for $s_{ij}>0$ and mean $E(S_{ij}\mid\theta_2)=\mu_{S_{ij}}(\theta_2)$ where $\theta_2=[\alpha_2^{'},b_{2i}]'$ is a vector of fixed intensity effects $\alpha_2^{'}$ and occurrence effects of random unit $b_2$. For skewed distribution, lognormal models for intensity is assumed so that:

$$Log(S_{ij}\mid\theta_2)\sim N(X_{2ij}^{'}\alpha_2^{'}+b_{2i},\sigma_e^2) \qquad (2)$$

Where $X_{2ij}^{'}$ is the vector of covariates for intensity.

Random effects were assumed to be in two models (1) and (20 are jointly normal and possibly correlated such that:

$$\begin{bmatrix}b_{1i}\\ b_{2i}\end{bmatrix}\sim BVN\left(\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{bmatrix}\right)$$

Lambert on Shin (2012) approached zero inflated count data with the mixture of two distributions is zero and the Poisson distribution. In the zero inflated Poisson regression (ZIP), $Y=(Y_1,...,Y_m)^T$ are independent.

$$Y_i\sim\begin{cases}0, & with\ \ probability\ 1-p_i\\ Poisson(\lambda_i), & with\ \ probablity\ p_i\end{cases}$$

So that

$$\Pr(Y_i=y_i)\sim\begin{cases}(1-p_i)+p_ie^{(-\lambda_i)}, & y_i=0\\ p_i\dfrac{e^{(-\lambda_i)}\lambda_i^{y_i}}{y_i!}, & y_i>0, 0\le p_i\le 1\end{cases}$$

Model zero inflated Poisson (ZIP) with random effects is based on subject-specific models. The distribution of the response variable is modeled in terms of covariates from each subject, and the variance of the random effects are determined to each of the subject. In the model, $\alpha_1$ and $\alpha_2$ measure the change in the conditional logit $p_{ij}$ and the conditional log $\lambda_{ij}$ with covariates $X_{1i}$ and $X_{2i}$ to measure each subject and illustrated by $b_{1i}$ and $b_{2i}$. And also assumed that the random effect $b_{1i}$ and $b_{2i}$ drawn from a normal distribution and correlated.

### 2.6 Mean Square Error (MSE) with Bootstrap Method

In small area estimation, mean squared error is an outstanding problem. Nonlinear model with complex structure make a close form expression for MSE is difficult. The method for correcting the bias of estimation is Jackknife method [6] and bootstrap method. On the MSE estimation was used jackknife method obtain smallest value MSE. Its cause of no convergence from iterative process. Follows a description of steps bootstrap estimation in simple stratified sampling proposed by Rao and Wu [17].

Steps to calculate MSE of bootstrap method with is as follows:

1. Specify stratum h with sample size $n_k$ household. Resampling sub sample by simple random sampling with replacement. This step is repeated independently for each stratum.
2. Repeat steps 1 r times
3. To obtain the bootstrap estimator off MSE, Calculate estimated $MSE_i = \frac{1}{R}\sum_{i=1}^{R}(\hat{\theta}_i^{boot} - \hat{\theta}_i)^2$ , *i = 1, 2, ... , h* where :

   $\hat{\theta}_i^{boot}$ = the composite estimator obtained with resampling bootstrap

   $\hat{\theta}_i$ = the composite estimator obtained data

## RESULTS

The district of Pangandaran is a division of the district of Ciamis. This district was officially separate from Ciamis on October 25, 2012 under UU 21 Th 2012 on the Establishment of District Pangandaran in West Java. So that the number of cities and districts in West Java as much as 27 City / District. But in this study are still using the data Health Profile 2012 before the division with 17 districts and 9 cities. From the 26 cities / regencies in West Java comprises 24 cities / districts are drawn the sample, while two other cities that are not drawn the sample (nir-sample) namely Sukabumi and Banjar. The following table showed the distribution of samples for each district and city:

Based on RPJMN 2004-2009 which mentions several health problems is the disparity in health status, the double burden of disease, the low performance of health services, the behavior of the people who is lack of support a clean and healthy lifestyle, poor environmental health conditions, poor quality, equity and affordability uneven distribution and health low status. In addition to the lack of equity and affordability of health care, socio-economic factors also influence the number of infant deaths. The variables are used the data of West Java provincial health profile in 2012. Based on data from health profile 2012, obtained five variables are correlated : the number of doctors, the number of health centers, the number of households behave clean and healthy, the number of undernourished and poor as well as the amount of weight low birth.

The incidence of infant mortality is the response variable has a Poisson distribution and equidispersion assumptions. The expected value is greater than 1 that indicated there is overdispersion variants which led to the invalid conclusion. Another Methods to deal with are the negative binomial overdispersion and quasi-likelihood. Based on models used namely Poisson mixed model, negative binomial mixed model and Poisson quasi-likelihood model to estimate the fixed effects (β) and random effects ($\sigma^2$).

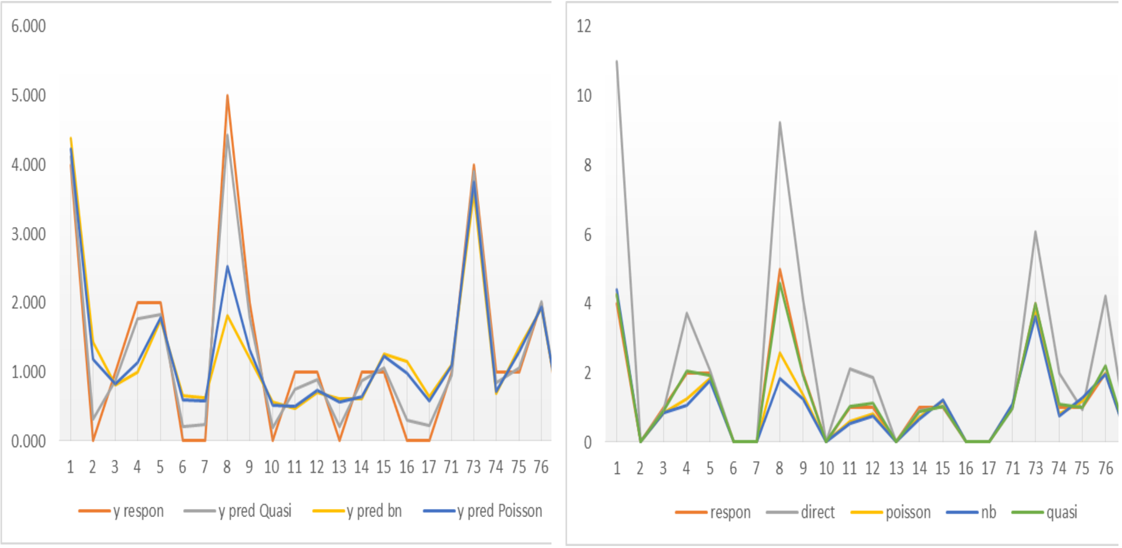
Based on the model based, quasi-likelihood model can solve overdispersion in the case of infant mortality. Although the values obtained of generalized chi-square model of quasi-likelihood were greater than mixture Poisson models and negative binomial mixed model.

**TABLE 1.** Tabulation of sample size, birth, and mortality from DHSI Data 2012

| The Code of District | The Name of District | The Number of Households | The number of Birth 2007-2012 | The number of Infant Mortality |
|---|---|---|---|---|
| 1 | Bogor Regency | 88 | 104 | 4 |
| 2 | Sukabumi Regency | 50 | 56 | 0 |
| 3 | Cianjur Regency | 37 | 41 | 1 |
| 4 | Bandung Regency | 49 | 52 | 2 |
| 5 | Garut Regency | 39 | 42 | 2 |
| 6 | Tasikmalaya Regency | 25 | 27 | 0 |
| 7 | Ciamis Regency | 13 | 13 | 0 |
| 8 | Kuningan Regency | 27 | 30 | 5 |
| 9 | Cirebon Regency | 36 | 40 | 2 |
| 10 | Majalengka Regency | 22 | 22 | 0 |
| 11 | Sumedang Regency | 15 | 16 | 1 |
| 12 | Indramayu Regency | 16 | 18 | 1 |
| 13 | Subang Regency | 24 | 25 | 0 |
| 14 | Purwakarta Regency | 16 | 17 | 1 |
| 15 | Karawang Regency | 34 | 37 | 1 |
| 16 | Bekasi Regency | 43 | 48 | 0 |
| 17 | Bandung Barat Regency | 32 | 32 | 0 |
| 71 | Bogor City | 26 | 26 | 1 |
| 73 | Bandung City | 46 | 55 | 4 |
| 74 | Cirebon City | 15 | 21 | 1 |
| 75 | Bekasi City | 24 | 34 | 1 |
| 76 | Depok City | 38 | 46 | 2 |
| 77 | Cimahi City | 5 | 5 | 0 |
| 78 | Tasikmalaya City | 7 | 7 | 0 |
| Total |  | 727 | 88 | 29 |

**TABLE 2**. Goodness of fits of Poisson Mixed Model, Binomial Negative Mixed Model, and Quasi-Likelihood Model

| Fit Statistics | Poisson | Negative binomial | Quasi-likelihood $V=0.33\mu$ |
|---|---|---|---|
| -2 Res Log Pseudo-Likelihood | 138.19 | 138.78 | 145.10 |
| Generalized Chi-Square | 13.83 | 14.41 | 18.01 |
| Gener. Chi-Square / DF | 0.77 | 0.8 | 1 |



(a)

(b)

**FIGURE 1.** (a) Line plot between Actual value from data and predicted value from models, (b)Line Plot to Compare predicted value and actual value from direct estimation, and composite poisson mixed model, negative binomial model, and quasi-likelihood model

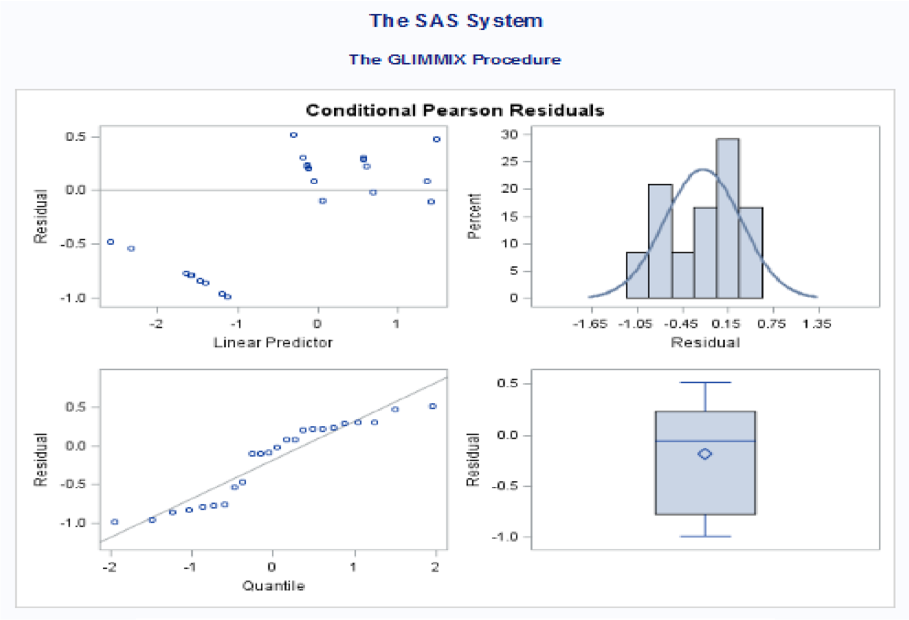


**FIGURE 2.** Residual Plot Poisson Quasi Likelihood Model

The basic model of small area estimation using area level based model was built from Fay and Herriot model. Based on estimates of both design based and model based, were estimated prediction of small area estimation for the district. FIGURE 1 showed that the predicted value model quasi-likelihood makes the same pattern from actual value. Based on overdispersion value and up fatten, the best model is poisson quasi likelihood model.

FIGURE 2 showed that after checking the residual assumptions, still resulted that residuals of model formed two normal distributions and residual plot versus linear predictor make two cluster. Its cause of the more than 30% response value zero. So as to resolve the issue used Mixed Effect Zero Inflated Poisson (ZIP) Model.

## DISCUSSION

Based on the analysis above, it was concluded that quasi-likelihood could overcome overdispersion and prediction on a small area had the predicted-values were close to the actual values of the response. However, after checking the residual assumptions, still resulted that residuals of model formed two normal distributions and negative mean square error. So as to further modeling needs to be done to resolve the issue, for example with mixture distribution and use non parametric bootstrap method to measure mean square error. In estimation of there are no mortality case in some area. So the further research is to assume that there are similarities among particular areas which can be analyzed using clustering technique.

## REFERENCES

BPS, BKKBN, KEMENKES, and USAID, “Survei Demografi dan Kesehatan Indonesia,”. http://www.bkkbn.go.id/litbang/pusdu/Hasil%20Penelitian/SDKI%202012/Laporan%20Pendahuluan%20SDKI%202012.pdf, 2012

Coordinating Ministry for People’s Welfare, World Food Programme, BPS dan AusAID, “Nutrition Map of Indonesia: Small area Estimation of Nutritional status in Indonesia”. 2006.

A.F. Hadi and K.A. Notodiputro. BIAStatistika. **3**, 1, 41-60 (2009).

N. Hajarisman, “Pemodelan Area Kecil untuk menduga angka kematian bayi melalui pendekatan model regresi Poisson bayes berhirarki dua-level,” Disertasi, IPB Bogor, 2013.

D.G. Horvitz, D.J. Thompson, Journal of the American Statistical Association. **47**, 260, 663-685 (1952).

A. Kurnia and K.A. Notodiputro, Penerapan Metode Jackknife dalam Pendugaan Area Kecil. Vol 11, Forum Statistika dan Komputasi, Bogor, 2006

A. Kurnia. “Prediksi Terbaik Empirik untuk Model Transformasi Logaritma di dalam Pendugaan Area Kecil dengan Penerapan pada Data Susenas,” Disertasi, IPB Bogor, 2009.

P. McCullagh and J.A. Nelder, Generalized Linear Models, (Chapman and Hall, London, 1989).

C.E. McCulloch and S.R. Searle. Generalized, Linear, and Mixed Models, (John Wiley & Sons, New York, 2001).

Y. Pawitan. In All Likelihood: Statistical Modelling and Inference Using Likelihood, (Oxford science publications, New York, 2001).

J.N.K. Rao, Small Area Estimation, (John Wiley and Sons, New York, 2003).

K. Sadik, “Metode Prediksi Tak-Bias Linear Terbaik dan Bayes Berhirarki untuk Pendugaan Area Kecil berdasarkan Model State Space,” Disertasi, IPB Bogor, 2009.

J. M. Veor Hoef and P.L. Boveng, “Quasi-Poisson VS. Negative Binomial Regression: How Should We Model Overdispersed Count Data?”, Publications, Agencies, and Staff of the U.S. Departement of Commerce. Paper 142, 2007 see http://digitalcommons.unl.edu/usdeptcommercepub/142

A. Yadav and L. Ladusingh. “District Level Infant Mortality Rate: an Exposition of Small Area Estimation,” Population Association of Amerika – Applied Demography Newsletter Vol. 26 No.1. 2013.

RS. Woodruff. Journal of the American Statistical Association, **66**, 411–414 (1971).

J. Shin, “Mixed-Effects Models for Count Data with Applications to Educational Research”, Dissertation, The Florida State university, 2012.

J. N. K. Rao and C. F. J. Wu. *Journal of the American Statistical Association*, **83**, 401, 231-241 (1988).