International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 6 (2018) pp. 3139-3143

Zero Inflated Binomial Model for Infant Mortality Data in Indonesia

Wahyu Bodromurti[1,a)], Khairil Anwar Notodiputro[2,b)], and Anang Kurnia[3,c)]

[1,2,3]Department of Statistics, Bogor Agricultural University, Kampus IPB Darmaga, Bogor, Indonesia.

Abstract

This paper discusses overdispersed binomial models applied to infant mortality data in Indonesia. Overdispersion usually occurs when the data has many zeros, or called as excess zeros. In such cases, binomial models are less fit and the type I error can be inflated or higher false positive rates can be obtained. This problem can be resolved by using zero inflated binomial (ZIB) models. Hall (2000) applied ZIB models by modifying the zero inflated Poisson (ZIP) models developed by Lambert (1992). In the ZIB models, the response variable was assumed to be distributed as a mixture of non-zero value distribution consisted of binomial (n, π) and a distribution of the binary zero-indicator. It was also assumed that the mixing probability was p. The fitness of the model was assessed using ROC curves as well as other criteria such as AIC, AICC, and BIC. The result showed that ZIB model has better fit in terms of overcoming the overdipersed binomial data.

Keywords: excess zeros, overdispersion, infant mortality, zero inflated binomial.

INTRODUCTION

Background

Binary count data with success probability π and upper bound n usually follows binomial (n, π) distribution and usually can be analyzed using binomial models. If the variation is greater than the assumed model then binomial data is called overdispersed (Hinde dan Demetrio 2007). The overdispersion can be caused by excess zeros. Hinde and Demetrio (2007) has claimed that the overdispersion may result in underestimated of standard error which produce underestimated p-values. This means that non-significant association will appear to be significant. Besides that, overdispersion can produces higher false positive rates that affect the validity of inferences. The zero inflated binomial (ZIB) can can be used to overcome the over-dispersion problems. Lambert (1992) was interested to adapt zero inflated Poisson regression (ZIP) models and Hall (2000) modified ZIP into ZIB models.

Infant mortality is a binary event in certain period hence number of infant mortalities in each villages generally follows binomial distribution with probability of death π among n births. Through the Indonesia Demographic and Health Survey (IDHS) the infant mortality data was recorded in five years (2008-2012). Since the number of infant deaths is usually small then this data is very likely to suffer from over-dispersion problem. According to the note by World Health Organization (WHO) in 2015, there were 75% of under-five year deaths occur in the first year of life or around 4.5 billion babies. Based on Indonesia United Nations Children's Emergency Fund (UNICEF Indonesia 2012), the patterns of high infant mortality rate are related to the babies from rural households, babies of mothers who are less educated, delivery place at home or delivery post instead of in health facilities, low birth weight babies (LBWB), birth order 4th until 6th, maternal age at delivery are more than 30 years old, babies are not breastfed or breastfeeding less than one year, and twice birth during last three years. This paper discusses ZIB models for infant mortality data in Indonesia. The response variable is number of infant deaths in each village and the explanatory variables are factors determining the patterns of infant mortality. The fitness of the ZIB model is assessed using Receiver Operating Characteristic (ROC) curves as well as other criteria such as Akaike's Information Criterion (AIC), Akaike's Information Criterion Corrected (AICC), and Bayesian Information Criterion (BIC).

OBJECTIVES

The research objectives are:

1. To understand the application of ZIB models in analyzing infant mortality data related to the death which occurs in the first year among infants in West Java, Indonesia.
2. To investigate the performance of ZIB model and to assess the model using ROC curves as well as other criteria such as AIC, AICC, and BIC.

THEORITICAL REVIEW

Zero Inflated Binomial Model

Overdispersed in GLMs may be due to variability of experimental materials, correlation between individual responses, cluster sampling, aggregation level of data, and omitted unobserved variables (Hinde and Demetrio 2007). In some conditions, the cause of the overdispersion may be recognize from the nature of the data, such as excess zeroes which lead to greater variances than the assumed model. The overdispersed binomial data can be better fit using ZIB models. In ZIB models, the response variable is assumed to be distributed as a mixture of non-zero values distribution as binomial (n, π) and a distribution of the binary zero-indicator, with mixing probability p. Overdispersed binomial data are modeled in ZIB models by Hall (2000), which is the response

variable vector $\boldsymbol{Y} = (Y_1, Y_2, \dots, Y_m)'$ of sample size m in which the probability

$$Y_i \sim \begin{cases} 0, & with\ probability\ p_i \\ binomial(n_i, \pi_i), & with\ probability\ (1-p_i) \end{cases} \quad (1)$$

The probability of outcome k for Y is equal to

$$Y_i = \begin{cases} 0, & with\ probability\ \ p_i + (1-p_i)(1-\pi_i)^{n_i} \\ k, & with\ probability\ (1-p_i)\binom{n_i}{k}{\pi_i}^k(1-\pi_i)^{n_i-k} \end{cases} \quad (2)$$

where $y_i = 0,1,2,\dots,n_i$ and $i = 1,2,\dots,m$ with expectation value $E(Y_i) = (1-p_i)n_i\pi_i$, and variance $var(Y_i) = (1-p_i)n_i\pi_i[1-\pi_i(1-p_in_i)]$. The parameter of ZIB model $\boldsymbol{p} = (p_1, p_2, \dots, p_N)'$ and $\boldsymbol{\pi} = (\pi_1, \pi_2, \dots, \pi_N)'$ are modeled through GLMs with canonical link function $logit(\boldsymbol{\pi}) = \boldsymbol{B}\beta$ and $logit(\boldsymbol{p}) = \boldsymbol{G}\gamma$ for design matrices $\boldsymbol{B}$ dan $\boldsymbol{G}$.

METHODOLOGY

Data

In this study, this research uses Indonesia Demographic and Health Survey (IDHS 2012) consisting of 24 cities/regencies and 93 villages in West Java, Indonesia. The data were recorded during five years (2008-2012). The response variable is number of infant deaths at the first year of life which were also recorded during five years (2008-2012) and explanatory variables as shown in Table 1.

Method

To accomplish the objectives of this study, the following steps are:

1. Explore the IDHS data to gain general insight of the data as well as to identify the explanatory variables potentially affected the response.
2. Fix the explanatory variables to fit the ZIB model.
3. Specify the initial values for the ZIB model.
4. Apply the ZIB model to infant mortality data.
5. Make ROC curve for the ZIB model.

Table 1. Variable used, information, and category

Variable	Information	Category
Y	Number of infant deaths at the first year of life during 2008-2012	-
n	Number of infant births during 2008-2012	-
X1	Percentage of babies from rural area	Urban
		Rural
X2	Percentage of less educated mothers	No education
		Incomplete primary
		Complete primary
		Incomplete secondary
		Complete secondary
		Higher
X3	Percentage of babies which the birth order is 4th-6th	Birth Order 1st
		Birth Order 2nd-3rd
		Birth Order 4th-6th
		Birth Order >= 7th
X4	Percentage of low birth weight babies (LBWB)	Very large
		Larger than average
		Average

Variable	Information	Category
		Smaller than average
		Very small
		Don't know
X5	Percentage of delivery place at home and delivery post	Respondent's home and Delivery Post
		Others
X6	Percentage of mothers who birth twice during last three years	1 birth during 3 years
		2 birth during 3 years
		No birth during 3 years
X7	Percentage of maternal age at delivery are 31-50 years old	Maternal age <= 20
		Maternal age 21-30
		Maternal age 31-40
		Maternal age 41-50
X8	Percentage of babies are not breastfed and breastfed less than one year	Ever breastfed, not currently breastfeeding
		Never breastfed
		Still breastfeeding

RESULTS AND DISCUSSION

There were 28 babies who died at the first year of life during 2008-2012 occurs in 23 of 93 villages in West Java, Indonesia. This implies that there were 75.27% zero values in the infant mortality data. Figure 1 showed that infant deaths in rural and urban areas were relatively the same, i.e. 3.8% infant death in rural area whereas in urban area was 3.9%. The lower the mother's education the higher was infant mortality. In fact, among mothers who did not complete primary education the infant mortality was 9.6%. Similarly, the number of children in a household also increased infant mortality. Note that among babies with birth order between 4th and 6th the mortality was 8.2%. Moreover, among the infants with low birth weights the percentage of deaths was 50%. Delivery baby at home and delivery post resulted in 4.3% of infant death. Finally, close birth spacing such as two births during three years, maternal age at birth such as more than 30 years old, as well as un-breast fed babies were considered potential to affect the infant mortality.

As indicated in Figure 1, babies who were born in rural areas and in urban areas had similar risk of mortality. The figure showed that 3.9% of infant death occurred in both area, hence X1 was not included in the model. Suppose that for every infant births in village *i* (n_i), the probability of an infant would died in the first year of life is π_i. In the case of excess zero data, the ZIB model via canonical link function is as follows

$$logit(\hat{\pi}) = \widehat{\alpha_2} \times X2 + \widehat{\alpha_4} \times X4 + \widehat{\alpha_6} \times X6 + \widehat{\alpha_7} \times X7 + \widehat{\alpha_8} \times X8 \quad (3)$$

$$logit(\hat{p}) = \widehat{\beta_3} \times X3 + \widehat{\beta_5} \times X5 + \widehat{\beta_8} \times X8 \quad (4)$$

Hall (2000) stated that although this model consists of two distinct parts, the model components must be fit simultaneously. For modeling the non-zero part in $logit(\hat{\boldsymbol{\pi}})$, this paper used variables X2, X4, X6, X7 and for modeling the whole data in $logit(\hat{\boldsymbol{p}})$, this paper used variables X3, X5, and X8. It based on our experience in selecting variables in order to obtain the model from proc nlmixed using SAS program adapted from Steventon et al. (2005). The initial values were also important to overcome the failure in finding the estimates. The ZIB model used initial values from logistic model. Initial values of parameter in $logit(\pi)$ are from logistic regression between y/n and X variables, and initial values of parameter in $logit(p)$ are from logistic regression between δ and X variables. Degree of freedom of ZIB model as many as 85 are from number of observations, 93, reduce by number of parameter estimates, 8.

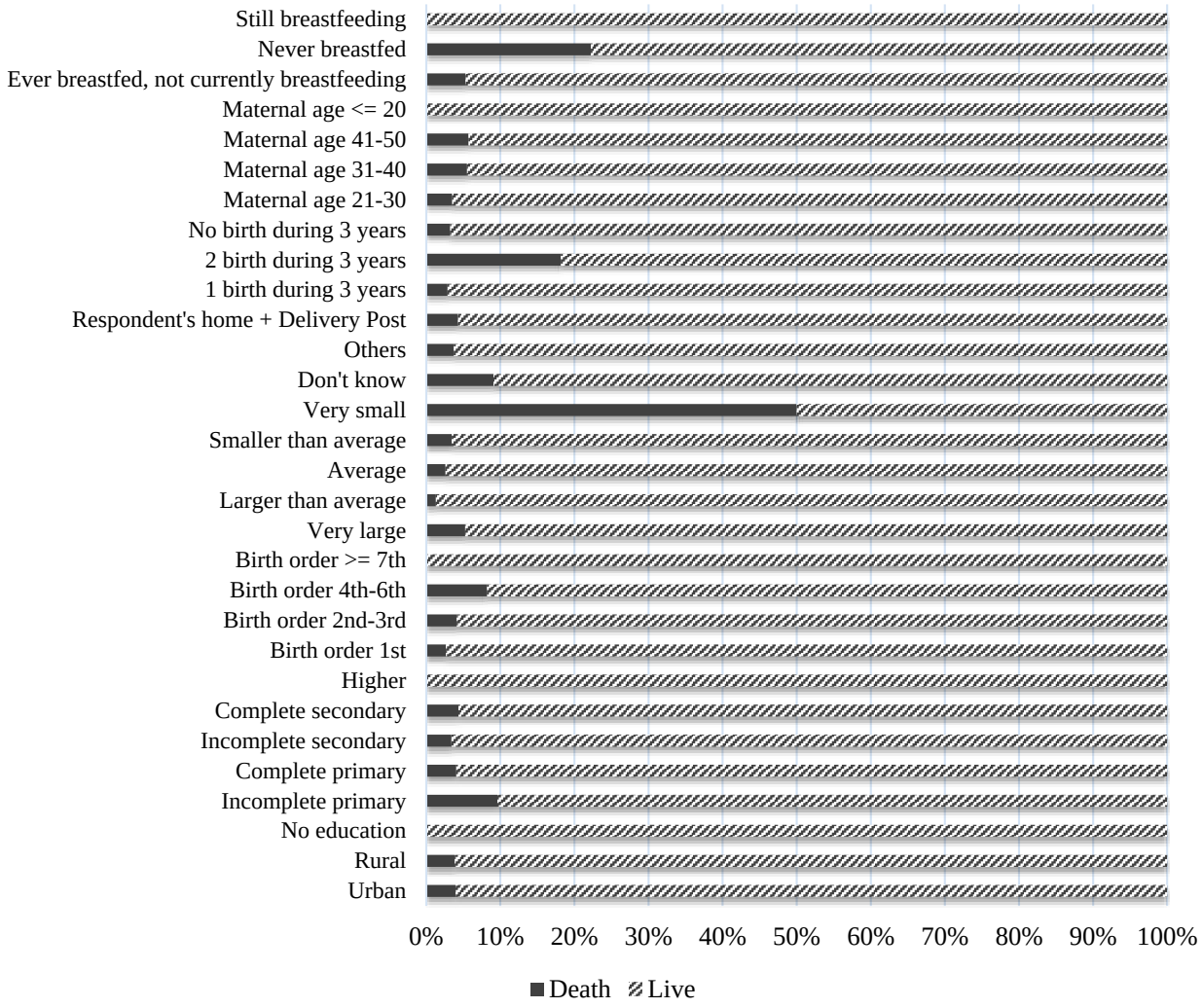

Figure 1. Number of infant lives and deaths based on explanatory variables

Table 2 showed only X8 = percentage of babies without breastfed and breastfed less than one year was significant at α = 5%. The X7 = percentage of maternal age at delivery are 31-50 years old was not significant at α = 5% but significant at α = 10%. Moreover, Table 3 presented the log likelihood, AIC, AICC, and BIC values of the ZIB model (137.6, 153.6, 155.3, 173.8) to measure the model performance. It was shown that the ZIB model has good performance with small values of the log likelihood, AIC, AICC, and BIC.

TABLE 2 Parameter estimates of ZIB model

Logit	Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	95% Confidence Limits	
	X2	0.00	0.03	85	0.04	0.97	-0.05	0.05
	X4	0.00	0.02	85	-0.19	0.85	-0.03	0.03
π	X6	0.01	0.01	85	0.53	0.59	-0.02	0.03
	X7	-0.02	0.01	85	-1.91	0.06	-0.04	0.00
	X8	-0.03	0.01	85	-3.55	0.00	-0.05	-0.01
	X3	-0.02	0.03	85	-0.55	0.58	-0.08	0.05
p	X5	0.01	0.01	85	1.44	0.15	-0.01	0.03
	X8	-0.01	0.01	85	-0.93	0.36	-0.04	0.01

Table 3. Fitness of ZIB model

Fit Statistics	
-2 Log Likelihood	137.6
AIC	153.6
AICC	155.3
BIC	173.8

Besides that, this research also uses Receiver Operating Characteristic (ROC) curves and Area Under The Curve (AUC) to evaluate the fitness or predictive accuracy of the model. The greater AUC values was resulted or the curve shape were close to a straight line at one value on the Y-axis indicated that the model has good performance. Figure 2 presented ROC curve of the ZIB model, and AUC of ZIB model were equal to one. This means ZIB model could model the overdispersed binomial data very well and this model can be used to estimate infant mortality at the unit level (village).

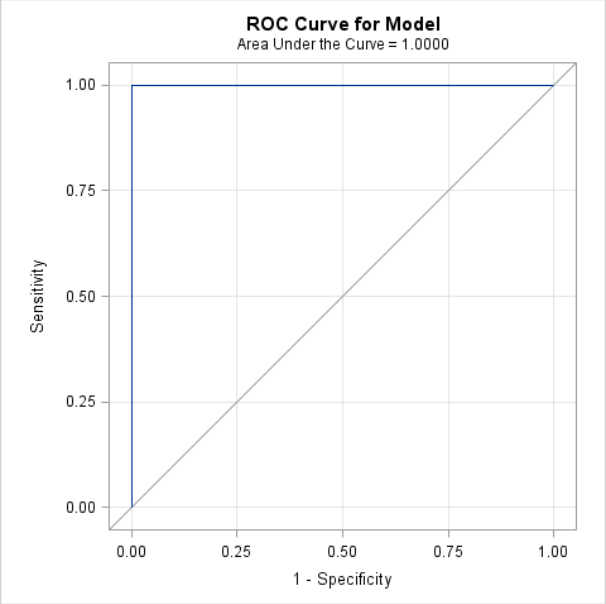

Figure 2. ROC curve of ZIB model

CONCLUSION

This paper discusses ZIB model for infant mortality data in Indonesia with excess zeros. The ZIB model was adapted from Hall (2000), and Steventon et al. (2005) to obtain the ROC curve. Based on several criteria such as the log likelihood, AIC, AICC, BIC values, and ROC curves, the model was successfully overcome overdispersion problem in binomial data. The problem with this model was related to computation which requires definite positive Hessian matrix to obtain the solution of the likelihood function. In this case we observed that the appropriate initialization values and arrangement of explanatory variables included in the models are important things to overcome the failures in algorithms.

REFERENCES

[1] Hall DB. 2000. Zero-Inflated Poisson and Binomial Regression with Random Effects: A Case *Study*. Biometrics 56:1030-1039.

[2] Hinde J, and Demetrio CGB. 2007. Overdispersion: Models and Estimation A Short Course for SINAPE 1998.

[3] Steventon JD, Bergerud WA, and Ott PK. 2005. Analysis of Presence/Absence Data when Absence is Uncertain (False Zeroes): An Example for the Northern Flying Squirrel using SAS. British Columbia: Ministry of Forests and Range Forest Science Program.

[4] [UNICEF Indonesia] Indonesia United Nations Children's Emergency Fund. 2012. Ringkasan Kajian Kesehatan Ibu & Anak [Internet]. [accessed 2016 February 06]. Available from: http://www.unicef.org/indonesia/id/A5_-_B_Ringkasan_Kajian_Kesehatan_REV.pdf.

[5] [WHO] World Health Organization. 2015. Global Health Observatory (GHO) data, Infant Mortality [Internet]. [accessed 2015 December 04]. Available from: http://www.who.int/gho/child_health/mortality/neonatal_text/en/.