Measles Disease Model using Censored Hurdle Negative Binomial Regression in East Java

Liza Nur Aida[1], Ria Dhea Layla Nur Karisma[2]
[1,2] Mathematics Department, Universitas Islam Negeri (UIN) Maulana Malik Ibrahim Malang, Indonesia
Email: [1] lizanuraida@gmail.com, [2] riadhea@uin-malang.ac.id

Abstract- **Measles is an infectious disease caused by measles virus and contagious. In recent years, especially in Indonesia, the number of measles rates have decreased at 2021 then some observations were worth zero. Hurdle Negative Binomial Regression is a method that used to overcome excess zero and over dispersion. Furthermore, count data is a data with non-negative integers that showed the number of event then it unable to use Poisson Regression. The aim of the study is to obtain measles model using HNBR in Eat Java. Based on the result of study, the factors that influence are vitamin A distribution, malnutrition in toddlers, and population density in East Java.**

Keywords— censored data; Hurdle Negative Binomial Regression; Measles

I. Introduction

Measles is a contagious disease caused by measles virus. The signs and symptoms are cough, runny, nose, fever, and red eyes and then causes a rash all over the body. Measles transmission occurs through air that contaminated by the secretions of an infected person. The government efforts to reduce the number of measles is by knowing the pattern of risk factors then the action can be determined [1]. In the recent years, especially in Indonesia, the number of measles rates have decreased, then some observations are worth zero. Count data is non-negative integers that states the number of events [2]. Hurdle Negative Binomial Regression is a method used to overcome excess zro and overdispersion. If overdispersion occurs, the Poisson regression model can't be used, because the Poisson regression must meet the equidispersion assumption [3].

A research by Mullahy, the first introduced the Hurdle count data model which can solve both under dispersion and over dispersion problems [3]. Hurdle Negative Binomial Regression is a modified model that can to overcome over dispersion and under dispersion conditions. Then, it can be used for excess zero events or non-negative integers. In addition, in the Hurdle Negative Binomial Regression has estimation parameter estimation that carried out separately (maximized separately) then it simple to interpret [4]. In measles data that we have, contained zero in most the observations and some values appeared in varying non-negative number then it called censored data. The type of censor data in this study are right-censor while the censor point selected uses the low category as the censor limit. The aim in this study is modelling in censored data in measles disease using Hurdle Negative Binomial especially in East Java.

II. Methods

A. *Hurdle Negative Binomial*

Suppose, Y_i is discrete random variable with i being a non-negative integer, then $i = 1,2,3,\dots,n$ that formed for count data. Furthermore, Y_i is the response variable from Hurdle Negative Binomial Regression model, then the value of the response variable has modeled in two conditions. The first condition is when the zero state and the second condition is a negative binomial state which has a negative binomial distribution [4]. The probability function Y_i of the Hurdle Negative Binomial regression model is:

$$P(Y_i = y_i | x_i, z_i)$$
$$= \begin{cases} p_i & ; y_i = 0 \\ (1-p_i)\dfrac{\Gamma\,(\mathrm{y_i}+k^{-1})}{\Gamma\,(\mathrm{y_i}+1)\Gamma(\mathrm{k}^{-1})}\left(\dfrac{(1+k\mu_i)^{-k^{-1}} - y_i k^{y_i} \mu_i{}^{y_i}}{1-(1+k\mu_i)^{-k^{-1}}}\right) & ; y_i > 0 \end{cases}$$

or

$$\Pr(Y_i = y_i) = \begin{cases} p_i & ; y_i = 0 \\ (1-p_i)\dfrac{g}{1-(1+\ k\mu_i)^{-k^{-1}}} & ; y_i > 0 \end{cases}$$

The value of the dependent variable appears in two different conditions. The first state is called the zero state which occurs on probability p_i. The second state is called Negative Binomial State occurs in $1 - p_i$ with $0 < p_i 1, \mu_i$ is the average of the Negative Binomial distribution with $k > 0$ and insignificance to independent variable.

Let p_i and μ_i depend on vectors in independent variables which is defined as follows:

$$\text{Logit }(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = x_i^T\delta$$

$$\frac{p_i}{1-p_i} = e^{x_i^T\delta}$$

$$p_i = (1-p_i)^{e^{x_i^T\delta}}$$

$$p_i = e^{x_i^T\delta} - p_i e^{x_i^T\delta}$$

$$p_i\left(1 + e^{x_i^T\delta}\right) = \ e^{x_i^T\delta}$$

Then,

$$\frac{e^{x_i^T\delta}}{1+e^{x_i^T\delta}}$$

For μ_i obtained from the linear log model:

$$\log(\mu_i) = x_i^T\beta$$

$$= \mu_i = e^{x_i^T\beta}$$

The probability function of Hurdle Negative Binomial regression model obtained by

$$P(Y_i = y_i) = \begin{cases} \dfrac{e^{x_i^T\delta}}{1+ e^{x_i^T\delta}} & ,\text{untuk } \mathrm{y_i} = 0 \\ \dfrac{1}{1+e^{x_i^T\delta}}\dfrac{g}{1-\left(1+ke^{x_i^Tk}\mu\right)^{-k^{-1}}} & .untuk\ y_i > 0 \end{cases}$$

with:

$$g = g(y_i;\ k,\beta) = \frac{\Gamma(\mathrm{y_i} + k^{-1})}{\Gamma(\mathrm{y_i}+1)\Gamma(\mathrm{k}^{-1})}\,(1+\ k\mu_i)^{k^{-1}-y_i}k^{yi}\mu_i^{yi}$$

where $x_i^T$is a vector of dependent Variables $(q+1)\times 1$ and q is the number of independent variables denoted, while parameters and are vectors of coefficient parameters with size $(q+1)\times 1$ [5]. It presents in matrix from as follows:

$$x_i = [\,1 \quad x_{1i} \quad x_{2i} \quad \dots \quad x_{pi}]^T$$

$$\beta = [\,\beta_0 \quad \beta_1 \quad \beta_2 \quad \dots \quad \beta_p\,]^T$$

$$\delta = [\delta_0 \quad \delta_1 \quad \dots \quad \delta_p]^T$$

B. *Measles*

Measles is an infection disease caused by a virus called paramyxovirus. The transmission is saliva through the nose, mouth, and throat of measles sufferers. The incubation period is 10-14 days before symptoms appearing. Symptoms include fever, cough, runny nose, and red patches on the skin, after which red patches appearing on the cheeks and then spread to the face, body, and other body parts. Complications of measles include pneumonia, ear infections, inflammation of the nerves, and the brain, then cause permanent brain damage [8].

Measles prevention is to maintain health through a healthy diet, regular exercise, adequate rest and immunization. Immunization causes active immunity, which aims to prevent measles. People who are susceptible to measles are infants aged over 1 year, infants who have not been vaccinated, and children who have not received a second vaccination [8].

III. APPLICATION

The data used in this study is secondary data obtained from the Central Statistics Agency of East Java Province. The dependent variable used is measles case data and the independent variable used are as follows:

X_1: percentage of vitamin A administration
X_2: percentage of measles immunization
X_3: number of malnourished children under five
X_4: population density
X_5: percentage of families with access to sanitation

The steps taken are as follows:

1. Conduct a descriptive analysis to find out the general description of measles disease data in East Java Province
2. Identify correlations between independent variables
3. Checking for over dispersion using deviance
4. Perform multicollinearity testing using the VIF criteria
5. Modeling Hurdle Negative Binomial regression
6. Test the significance of the parameters on the Hurdle Negative Binomial model simultaneously and partially
7. Interpreting the Hurdle Negative Binomial regression model using the odds ratio

IV. RESULT AND DISCUSSION

A. *Data Description*

The response variable used in this study is the number of measles cases (Y) in East Java Province. It decreased from previous years since a mass immunization campaign was held in Indonesia. There are five explanatory variables that are thought to influence the response variable to the number of measles. The five variables are the percentage of vitamin A (X_1), the percentage of measles immunization (X_2), malnutrition in toddlers (X_4), and the percentage of families with access to sanitation (X_5). Table 1 describes descriptive statistics from variables that used in the study:

Table 1. Descriptive Analysis

Descriptive Analisis			
Variable	N	Mean	Std. Deviasi
Y	38	0,45	0,891
X_1	38	91,4	7,609
X_2	38	87,59	13,46984
X_3	38	4408	3582
X_4	38	1951,4359	2202,60727
X_5	38	92,8474	9,14185

B. *Over dispersion Checking*

Over dispersion in this study using deviance value. The deviance value divided by degree of freedom. The result of deviance value is 1,292. It values above 1, then measles data in East data is over dispersion.

C. *Multicollinearity Checking*

The multicolliniearity in this study use to detect of multicollinearity in the regression model. When, multicolliniearity appears then VIF (Variance of Inflation) is more than 10. Table 2 is multicollinearity checking

Table 2 Multicollinearity Check

	X_1	X_2	X_3	X_4	X_5
VIF value	1,224	1,240	1,176	1,137	1,318
	No multicollinearity				

Table 2 VIF values in each independent variable has lower than 10. Therefore, it be concluded that all variables are used.

D. *Hurdle Negative Binomial Regression Model*

The next step is use Hurdle Negative Binomial Regression Model to measles disease in East Java. In this study, the significant level is 10%. The estimated model parameter presented in Table 3:

Table 3 Parameter Estimation of HNB

Parameter estimates censored data					
Truncated negative binomial			Zero state		
	Estimate	p-value		Estimate	p-value
β_0	0,9708	0,9678	δ_0	13,2022	0,1047
β_1	-0,1019	0,4039	δ_1	-0,1582	0,0233*
β_2	-0,0394	0,7623	δ_2	0,04835	0,1573
β_3	0,00059	0,0002*	δ_3	-0,0002	0,3004
β_4	0,00911	0,0829*	δ_4	-0,0003	0,2304
β_5	0,15827	0,4351	δ_5	-0,0260	0,6021
k	0,01521	0,9738			

The statistical results of the G test is 27.81, which is greater than $\chi^2_{0,1;3} = 6{,}2514$. It shows that there is at least one independent variable that has a significant effect in dependent variable. Table 4 shows that the independent variables that have a partially significant effect are the percentage variable given vitamin A (X_1), malnutrition in toddlers (X_3), and population density in East Java (X_4).

Table 4 Parameter estimation of significant variable

Parameter estimates censored data					
Truncated negative binomial			Zero state		
	Estimate	p-value		Estimate	p-value
β_0	7,32193	0,833	δ_0	9,75295	0,0471*
β_1	-0,09938	0,508	δ_1	-0,10401	0,0586*
β_3	0,0004	0,001*	δ_3	-0,00017	0,4568
β_4	0,00355	0,0432*	δ_4	-0,00031	0,2408

E. *Model Interpretation*

The parameter estimation results from Hurdle Negative Binomial regression model consist of two models. The first is a log model with Negative Binomial process and a logit model with zero state or Zero Hurdle process. In the Truncated Negative Binomial model, using the truncated state process, this model explained cases of measles in a district or city in East Java Province.

$$\mu_i = \exp(0{,}0004\ X_3 + 0{,}00355\ X_4)$$

In Truncated Negative Binomial model, it is known that each addition of one case of malnutrition in toddlers will increase the number of measles cases as much as $\exp(0.0004) = 1.0004$ times from original number of measles cases. Then, other variables are constant. Furthermore, every $1km^2$ increase in population density will be increased the number of measles cases by $\exp(0.00355) = 1.0035$ times. The original number, if other variables are constant.

While, the second model is the zero hurdle which explain the tendency to find cases of measles or not in a district or city in East Java Province.

$$\pi_i = \frac{\exp(9{,}75295 - 0{,}10401\ X_1)}{1 + \exp(9{,}75295 - 0{,}10401\ X_1)}$$

The factor that affect the chance of finding cases of measles is the percentage of vitamin A. It shows that the addition of one percent of vitamin A reduces the chance of finding cases of measles by 10.4% in a district or city. In addition, there are other effects caused by unknown variables.

V. Conclusion

In the logit model, the variables that affect the measles cases are the variable giving vitamin A. At the same time, in the log model the variables that have an effect are malnutrition in toddlers and population density. Every additional case malnutrition in toddlers will increase the number of cases of measles as much as 1.0004 times the original number of measles. Furthermore, every $1km^2$ increase in population density will increase the number of measles cases by 1.0035 times the original number.

References

[1] IDAI, Kampanye Imunisasi Campak Rubella di Pulau Jawa, Jakarta: Kemenkes RI, 2017.

[2] Pontoh, Penerapan Hurdle Negative Binomial pada Data Tersensor, Yogyakarta: UNY, 2015.

[3] Afri, "Pemodelan Regresi Hurdle pada Kasus Penyakit Difteri," *Jurnal Absis*, 2019.

[4] C. M., Generalized Linier Models, London: Chapman and Hall, 1989.

[5] Wulandari, "Konsumsi Rokok Masyarakat Kota Bandung dengan Hurdle Negative Binomial," *Jurnal Aplikasi Statistika dan Komputasi*, 2017.

[6] S. E. Saffari, "Hurdle Negative Binomial Regression Model with Right Censored Count Data," *Journal Statistic and Operation Research Transaction,* pp. 181-194, 2012.

[7] A. Widarjono, Ekonometrika: Teori dan Aplikasi untuk Ekonomi d an Bisnis, Yogyakarta: Ekonisia Fakultas Ekonomi Universitas Islam Indonesia, 2007.

[8] Suparyanto, Tumbuh Kembang dan Imunisasi, Jakarta: EGC, 2014.

[9] A. Bilgic, "Application of a Hurdle Negative Binomial Count Data Model to Demand for Fishing in the Southeastern United States," *Journal of Environmental Management,* pp. 478-490, 2007.

[10] Famoye, "Modeling Household Fertility Dicision with Generalized Poisson Regression," *Jurnal of Papulation Economics,* pp. 273-283, 2004.

[11] Hilbe, Negative Binomial Regression, New York: Cambridge University Press, 2011.

[12] R. Julianda, "Penerapan Data Count dengan Menggunakan Regresi Hurdle Poisson," *Jurnal Matematika,* 2015.

[13] A. McDowell, "From The Help Desk: Hurdle Models," *Stat Corporation,* pp. 178-184, 2003.

[14] M. Pateta, Fitting Poisson Regression Models Using the Genmond Procedure, USA: SAS Institute Ins, 2005.

[15] Agresti, Categorical Data Analysis, New York: John Willey and Sons, 1990.