Hurdle Regression Modelling on the Number of Deaths from Chronic Filariasis Cases in Indonesia

Nur Kamilah Sa'diyah*, Ani Budi Astuti, and Maria Bernadetha T. Mitakda

Department of Statistics, Faculty of Mathematics and Natural Sciences, Brawijaya University
**Corresponding author. Email: nurkamilahs@student.ub.ac.id*

ABSTRACT

Poisson regression is one of the model to explain the functional relationship between response variable in the form of count and predictor variable. An important assumption in Poisson Regression analysis is equidispersion. In certain cases, where response variable consists of too many zeros, causing the variance to be greater than the mean or called overdispersion that can be overcomed by the Hurdle model. Filariasis disease is caused by filaria worm that led to swelling of the limbs in humans. One province in Indonesia, Papua Barat, reported a quite high death of chronic filariasis cases with a death rate of 459 people. The Hurdle regression model is appropriately to model the number of cases of chronic filariasis death in Indonesia since the data contains overdispersion. This study will compared two regression Hurdle models, namely the Hurdle Poisson regression and Hurdle Negative Binomial Regression. The results showed that the Negative Binomial Hurdle regression model was better than that of the Hurdle Poisson regression model in modeling cases of filariasis in Indonesia with AIC value of 213.263. Based on logit model of Negative Binomial Regression, the percentage of households that have access to proper sanitation (X_5) has a significant effect on the number of cases of death from chronic filariasis in Indonesia.

Keywords: *Filariasis, MLE, Overdispersion, Hurdle regression.*

1. INTRODUCTION

Filariasis (elephant foot disease) is a contagious disease caused by filaria worms and transmitted by the Mansonia mosquitoes, Anopheles, Culex, and Armigeres [1]. Filaria disease infects lymph tissue (lymph nodes) causing swelling in the legs, breasts, arms, and genital organs in humans. Filiariasis has been around since Before Christ (B.C.) as evidenced by some ancient relics that illustrate how people at that time had suffered from chronic filiarisis. Filiarisis was discovered in Indonesia in 1889 in Jakarta by Hacka and Van Eecke.

Population density and proper sanitation facilities (healthy latrines) affect the number of cases of filariasis [2]. One of who's efforts to inhibit filaria's transmision disease is to carry out mass preventive drug administration (POPM) filariasis implemented by District/City endemic filariasis [3]. The study wanted to test several predictor variables (to be discussed in chapter 3) that affect the number of cases of death due to chronic filariasis.

The number of cases of death due to chronic filariasis is discrete so it can be modeled by Poisson regression and Negative Binomial Regression. In certain cases, there are many zero values in the response variables, causing the variance value to be greater than the average value or referred to as overdispersion. In the event of overdispersion, Poisson regression and Negative Binomial regression are less precisely used, because the model formed will result in a refractive parameter presumption [4]. Handling overdispersion in this study using the Hurdle model. Parameter estimation used Maximum Likelihood Estimator (MLE).

The goal of the study was to find out the factors that influence the number of cases of death due to chronic filariasis. The Hurdle regression model used is Poisson Hurdle regression and Negative Binomial Hurdle regression then performs the best model selection.

2. METHOD AND ANALYSIS

In this chapter, we outline the data analysis methods used in this study.

2.1. Multicollinearity Testing

The testing of the relationship between predictor variables used the VIF (Variance Inflating Factor) criteria. If the VIF value of the predictor variables is presented in equation (1) exceeds 10 then the change is said to be very closely related to $(k-1)$ other predictors [5].

$$VIF_j = \frac{1}{1-R_j^2} \tag{1}$$

where R_j^2 is the coefficient of determination to-on j auxiliary regression.

2.2. Poisson Distribution Conformity Testing

The Kolmogorov-Smirnov test for Poisson distribution suitability testing is presented in equation (2) [6].

$$D = maksimum\left|F_N\left(y_{(i)}\right) - P\left(y_{(i)}, \lambda\right)\right| \tag{2}$$

where

$F_N\left(y_{(i)}\right)$: cumulative function example, $\frac{frekuensi\ kumulatif\ y_{(i)}}{n}$

$P\left(y_{(i)}, \lambda\right)$: the cumulative chance function of the Poisson distribution, $e^{-\lambda}\left(1+\lambda+\frac{\lambda^2}{2!}+\cdots+\frac{\lambda^k}{k!}\right)$

2.3. Poisson Regression

The regression method for modeling the causal relationship between the discrete response variables and the predictor variables (it can be discrete or continuous) is called Poisson regression. Equation (3) presents Poisson's regression model [7].

$$y_i = \lambda_i + \varepsilon_i \tag{3}$$

where

$$\hat{\lambda}_i = exp\left(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik}\right) \tag{4}$$

$$\ln\left(\hat{\lambda}_i\right) = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + \cdots + \hat{\beta}_k X_{ik} \tag{5}$$

2.4. Overdispersion Testing

In certain cases, there are many zero values in the response variables, causing the variety value to be greater than the average value or referred to as overdispersion. Examination of the occurrence of overdispersion can be done using the Pearson Chi-Square value divided by degrees of freedom of residuals obtained from the results of Poisson regression analysis.

$$\chi^2_{pearson} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \tag{6}$$

If the $\chi^2_{pearson/db}$ value is more than one then it is said that the data contains overdispersion.

2.5. Hurdle Poisson Regression and Hurdle Negative Binomial Regression

One model for overcoming overdispersion is the Hurdle model. The Hurdle Poisson regression model is a combination of the logit model in equation (7) and the truncated Poisson in equation (8).

$$logit\ \pi_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \delta_0 + \sum_{j=1}^{k} x_{ij}\delta_j + \varepsilon_i \tag{7}$$

$$\ln \lambda_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \tag{8}$$

The elaboration results of equation (7) and equation (8) are presented in equation (9) and equation (10).

$$\pi_i = \frac{exp\left(\delta_0 + \sum_{j=1}^{k} x_{ij}\delta_j\right)}{1+exp\left(\delta_0 + \sum_{j=1}^{k} x_{ij}\delta_j\right)} \tag{9}$$

$$\lambda_i = exp\left(\beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j\right) \tag{10}$$

In addition to using Hurdle Poisson, overdispersion can also be overcome by negative binomial Hurdle regression. The Hurdle Negative Binomial model consists of a combination of the logistic regression model in equation (11) and the zero truncated Negative Binomial model in the equation (12) [8].

$$logit\ p_i = \ln\left(\frac{p_i}{1-p_i}\right) = \delta_0 + \sum_{j=1}^{k} x_{ij}\delta_j + \varepsilon_i \tag{11}$$

$$\ln \mu_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \tag{12}$$

2.6. Hurdle Regression Parameter Testing

There are two hurdle regression parameter testing that is simultaneous testing with G test statistics and partially parameter testing with the Wald test.

2.6.1. Simultaneous Test

If you want to find out if the change of predictor variables affects the response variables together, then testing the parameters of the regression model simultaneously [9].
The hypotheses underlying the test are:
H_0: $\beta_j = \delta_j$ *vs*
H_1: There is at least one different.β_j and δ_j
Test statistics are used as in the equation (13).

$$G = -2(l_R - l_F) \sim \chi^2_k \tag{13}$$

Reject H_0 if $G \geq \chi^2_{\alpha,k}$or p value $< \alpha$

2.6.2. Partial Test

Partial testing of parameters to determine the effect of each predictor varibles on the response variables. Partial test results are based on the Wald test [10].

a. Test Model Parameters $log(\mu) = X\beta$

Test of hypothesis-based model parameters $log(\mu) = X\beta$

$H_{01}: \beta_1 = 0 \; vs \; H_{11}: \beta_1 \neq 0$

$H_{02}: \beta_2 = 0 \; vs \; H_{12}: \beta_2 \neq 0$

$\vdots$

$H_{0k}: \beta_k = 0 \; vs \; H_{1k}: \beta_k \neq 0$

Sampling distribution for $\hat{\beta}_j$:

$\hat{\beta}_j \sim N\left(\beta_j, \sigma^2_{\hat{\beta}_j}\right)$

If H_0 true, then $\hat{\beta}_j \sim N\left(0, \sigma^2_{\hat{\beta}_j}\right)$

Test statistics that can be used as in the equation (14).

$$W_j = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)^2 \sim \chi^2_v \qquad (14)$$

Reject H_0 if $W_j \geq \chi^2_{\alpha;v}$or p value $< \alpha$

b. Test Model Parameters $logit(\omega) = X\delta$

Test of hypothesis-based model parameters $logit(\omega) = X\delta$

$H_{01}: \delta_1 = 0$ vs $H_{11}: \delta_1 \neq 0$

$H_{02}: \delta_2 = 0$ vs $H_{12}: \delta_2 \neq 0$

$\vdots$

$H_{0k}: \delta_k = 0$ vs $H_{1k}: \delta_k \neq 0$

The test statistics used are presented in the equation (13).

$$W_j = \left(\frac{\hat{\delta}_j}{SE(\hat{\delta}_j)}\right)^2 \sim \chi^2_v \qquad (15)$$

Reject H_0 if $W_j \geq \chi^2_{\alpha;v}$ or p value$< \alpha$

2.7. Selection of Best Model

One of the criteria for knowing the best model is the AIC *(Akaike Information Criterion)* value is presented in equation (16). Criterion is based on the Maximum Likelihood *Estimator* (MLE) method. The best regression models have the smallest AIC values [11].

$$AIC = e^{\frac{2k}{n}} \sum_{i=1}^{n} \varepsilon_i^2 \qquad (16)$$

where:
$e = 2.718$

3. DATA SOURCE

The data used in this study is sourced from the Indonesian Health Profile 2020. This study discusses the influence of the number of all chronic cases of Filiarisis in Indonesia (X_1), the number of District/City that have succeeded in reducing Mikrophilia <1% (X_2), the number of district/city that is still carrying out Mass preventive drug administration (POPM) Filiarisis (X_3), population density (X_4), percentage of households that have access to proper sanitation (X_5), the number of death due to chronic filiarisis (Y).

4. RESULTS AND DISCUSSION

In this section is outlined related to data description, multicolinarity testing, Poisson regression modeling, overdispersion testing, Hurdle Poisson regression modeling and Negative Binomial Hurdle regression, regression parameter testing, as well as selection of the best model.

4.1. Descriptive Statistical Analysis

Descriptive statistical analysis is diffuse to describe or give an overview of the objects studied. In general, descriptive analysis is in the form of tables, diagrams, and graphs. Table 1 is a descriptive statistic for each variables.

Table 1. Results of Descriptive Statistical Analysis

Variable	Mean	Stdev	Min	Max
X_1	291.35	657.32	2	3615
X_2	3.76	3.28	0	10
X_3	2.18	3.81	0	16
X_4	749.13	2731.2	9.5	16031.4
X_5	79.81	9.96	40.31	96.96
Y	26.20	84.91	0	459

Based on Table 1, the highest number of all chronic cases of Filiarisis in Indonesia at 3615 occurred in Papua. The number of death due to chronic filiarisis has a mean of 26.20 while the range value is 459. This happens because many Provinces in Indonesia do not have cases of death due to chronic filiarisis.

Figure 1 presents a bar chart of the number of death cases from filariasis in every District/City in Indonesia.

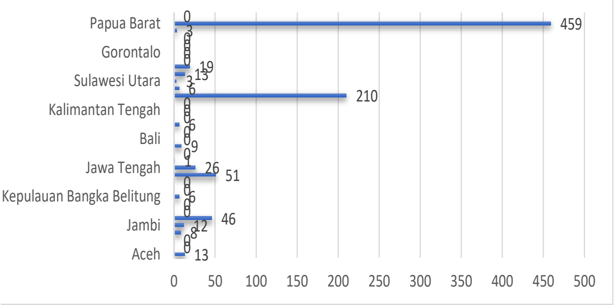

Figure 1 Bar chart of the number of deaths from chronic Filariasis.

Based on Figure 1, West Papua Province, East Kalimantan, and West Java are the three highest provinces that have the highest cases of chronic Filariasis deaths in Indonesia. However, 50% of districts/cities in Indonesia do not have cases of death from Filariasis.

4.2. Multicollinearity testing

One of Poisson's regression assumptions is that there is no relationship between predictor variables (Non-Multicolinierity). Multicollinearity test results with VIF value criteria are presented in Table 2.

Table 2. VIF Values in Predictor Variables

Variable	VIF Value
X_1	4.782
X_2	1.530
X_3	3.872
X_4	1.173
X_5	3.162

Based on Table 2 VIF value of all predictor variables less than 10 it can be said that the assumption of non-multicollinearity is fulfilled.

4.3. Poisson Regression Modeling

After conducting multicollinearity,Poisson regression modeling was done. The parameter estimation of Poisson regression parameters are presented in Table 3.

Table 3. Parameter Estimation Results of Poisson Regression

Variable	Coefficient	Value-*z*	Value-*p*
X_0	-11.79	-14.770	< 2e-16*
X_1	-1.667×10^{-3}	-8.315	<0.01*
X_2	0.1778	8.813	< 2e-16*
X_3	0.5746	26.841	< 2e-16*
X_4	-3.973×10^{-4}	-3.046	0.00232*
X_5	0.159	17.9	< 2e-16*

Based on Table 3 all predictor variables that is the number of all chronic cases of Filiarisis in Indonesia (X_1), the number of district/city that have succeeded in reducing Mikrophilia <1% (X_2) , the number of District/City that is still carrying out Mass preventive drug administration (POPM) Filiarisis (X_3), population density(X_4), percentage of households that have access to proper sanitation against the number of cases of (X_5) affects the number of death due to chronic Filiarisis (Y).

Then the Poisson regression model can be written according to the equation (17).

$$\ln \lambda_i = -11.79 \pm 1.667 \times 10^{-3} X_1 + 0.1778 X_2$$
$$+0.5746 X_3 + 0.5746 X_{i3} X_4 + 0.159 X_5 \qquad (17)$$

4.4. Overdispersion Testing

In certain cases, there are many zero values in the response variables, causing the variety value to be greater than the average value or referred to as overdispersion. In the event of overdispersion, Poisson regression is less appropriately used. The value $\frac{\chi^2_{pearson}}{db} = 212.549$ then can be concluded that the data contains overdispersion.

4.5. Hurdle Poisson Regression Modeling

The parameters generated by the Hurdle model are two, the logit model and the Poisson truncated model. Table 4 provides the results of parameter estimation of logit model parameters for $y_i = 0$ and Table 5 presents Poisson's truncated model for $y_i > 0$.

Table 4. Parameter Estimation Results of Logit Model for Hurdle Poisson Regression

Variable	Coefficient	Value-*z*	Value-*p*
X_0	-14. 43	-2,202	0.0277*
X_1	4.868×10^{-4}	0.247	0.8049
X_2	0.2045	1.322	0.1863
X_3	0.2508	1.021	0.3074
X_4	-2.078×10^{-4}	-0.549	0.5830
X_5	0.1646	2.153	0.0313*

Based on Table 4, the percentage of households that have access to proper sanitation (X_5) has a significant effect on the number of cases of death from chronic filariasis in Indonesia. It can be interpreted that every 1% increase in the percentage of households that have access to sanitation deserves to increase the number of cases of death from chronic filariasis in Indonesia by $e^{0.1646} = 1.18 \approx 1$ life. The logit model for Poisson's Hurdle regression is presented in the equation (18).

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = -14.43 + 0.1646 X_5 \qquad (18)$$

Table 5. Parameter Estimation Results of Truncated Poisson on Hurdle Poisson Regression

Variable	Coefficient	Value-z	Value-p
X_0	-4.2095	-4.942	7.71×10^{-7}*
X_1	-0.0027	-13.717	< 2e-16*
X_2	0.1494	6.562	5.3×10^{-11}*
X_3	0.5104	22.558	< 2e-16*
X_4	-0.0004	-3.482	0.0005*
X_5	0.0802	8.501	< 2e-16*

Based on Table 5 all predictor variables that is the number of all chronic cases of Filiarisis in Indonesia (X_1), the number of District/City that have succeeded in reducing Mikrophilia <1% (X_2) , the number of District/City that is still carrying out Mass preventive drug administration (POPM) Filiarisis (X_3), population density(X_4), percentage of households that have access to proper sanitation against the number of cases of (X_5) affects the number of death due to chronic filiarisis (Y). Then the Poisson regression model can be written according to the equation (19).

$$\ln \lambda_i = -4.2095 - 4.2095X_1 + 0.1494X_2 + 0.5104X_3 -0.0004X_4 + 0.0802X_5 \quad (19)$$

4.6. Hurdle Negative Binomial Regression Modeling

The parameters generated by the Hurdle model are two, the logit model and the zero-truncated Negative Binomial model. Table 6 provides the results of the restoration of logit model parameters for $y_i = 0$ and Table 7 presents the zero truncated Negative Binomial model for $y_i > 0$.

Table 6. Parameter Estimation Result of Logit Model on Negative Binomial Hurdle Regression

Variable	Coefficient	Value-z	Value-p
X_0	-14. 43	-2,202	0.0277*
X_1	4.868×10^{-4}	0.247	0.8049
X_2	0.2045	1.322	0.1863
X_3	0.2508	1.021	0.3074
X_4	-2.078×10^{-4}	-0.549	0.5830
X_5	0.1646	2.153	0.0313*

The result of presumption of logit model parameters in Negative Binomial Hurdle regression is the same as Poisson's Hurdle model. The percentage of households that have access to proper sanitation (X_5) has a significant effect on the number of cases of death from chronic filariasis in Indonesia.

Table 7. Parameter Estimation Results of Zero Truncated Negative Binomial on Binomial Negative Hurdle Regression

Variable	Coefficient	Value-z	Value-p
X_0	-8.826	-0.931	0.352
X_1	-4.831×10^{-5}	-0.011	0.991
X_2	0.4077	1.329	0.184
X_3	0.4717	1.592	0.111
X_4	6.174×10^{-5}	0.048	0.961
X_5	0.1138	1.114	0.265

In the zero truncated Negative Binomial model, all predictor variables did not affect the number of cases of death from chronic Filiarisis.

4.8. Selection of the Best Model

After performing three regression modelings (Poisson regression, Poisson Hurdle regression and Negative Binomial Hurdle regression) then the selection of the best model with the criteria of AIC value presented in Table 8.

Table 8. AIC Value Criteria in Three Regression Models

Type of Regression Model	AIC Value
Poisson Regression	1386.2
Poisson Hurdle Regression	958.686
Negative Binomial Hurdle Regression	213.263

Based on Table 8, the best model is the Negative Binomial Hurdle regression model because it has the smallest AIC value of 213.263.

5. CONCLUSIONS

Based on the results of the data analysis that has been done, conclusions are obtained, among others:

1. Violations of equidispersion assumptions can be overcome with a proven Hurdle regression model of AIC values in the Poisson Hurdle model and the Negative Binomial Hurdle compared to Poisson regression.
2. The best regression model to model the number of cases of death from chronic filiarisis is the Negative Binomial Hurdle regression model because it has the smallest AIC value compared to the Poisson regression model and regression model. Hurdle Poisson.
3. The results of the restoration of logit model parameters in Negative Binomial Hurdle regression show that the percentage of households that have access to proper sanitation (X_5) has a significant

effect on the number of cases of death from Filariasis. Chronic in Indonesia. While in the zero truncated Negative Binomial model, all predictor variables do not affect the number of cases of death from chronic filariasis in Indonesia.

Suggestions that can be given to the next researcher are:

1. In this study using the Negative Binomial Hurdle model, only one predictor variable affected the number of cases of Filariasis death. Researchers can further add predictor variables that if it affects the number of cases of Filariasis death in Indonesia.
2. The study only compared the Hurdle model for overdispersion handling. Researchers can further add other methods such as Zero-Inflated Poisson and Zero-Inflated Negative Binomial so that they can find out the best regression model for handling overdispersion cases on the data of the number of cases. Death of Filariasis in Indonesia.

AUTHORS' CONTRIBUTIONS

The first author contributes to create research ideas, collecting and analyzing data, as well as journaling. Ani Budi Astuti and Maria Bernadetha T. Mitakda contributes to verifying methods and results of analysis and conducting journal writing reviews.

ACKNOWLEDGMENTS

This journal was written as one of the requirements for thesis writing. The author would like to thank the thesis advisor who has provided guidance and direction to complete this research. The author also thanked the reviewers who had contributed expertise to review and evaluate the manuscript for better quality.

REFERENCES

[1] Masrizal, Penyakit Filariasis, Jurnal Kesehatan Masyarakat, September 2012-2013. Vol. 7, No. 1, pp. 32-38.

[2] S.P.P. Natalia, D. Ispriyanti, Sugito, Penerapan Regresi Zero Inflated Generalized Poisson dan Pengujian Autokorelasi Spasial Pada Kasus Penyakit Filariasis di Jawa Tengah, Jurnal Statistika, Mei 2018 Vol. 6, No. 1, pp. 29-33. DOI: https://jurnal.unimus.ac.id/index.php/statistik/article/view/3420/3252

[3] G. Meliyanie, D. Andiarsa, Lymphatic Filariasis Elimination Program In Indonesia. JHECDs, 2017, Vol. 3, No. 2, pp. 63-70. DOI: https://doi.org/10.22435/jhecds.v3i2.1790

[4] D.W. Osgood, Poisson Based Regression Analysis of Aggregate Crime Rates, Journal of Quantitative Criminology, 2000, Vol.16, No. 1, pp. 21-43. DOI: https://doi.org/10.1023/A:1007521427059

[5] D.N. Gujarati, C . P . Dawn, Dasar-Dasar Ekonometrika, Edisi 5. Terjemahan Raden Carlos Mangunsong. Jakarta: Salemba Empat. 2012.

[6] F. Antoneli, F.M. Passos, L.R. Lopes, R.S. Briones, A Kolmogorov-Smirnov Test for the Molecular Clock Based on Bayesian Ensembles of Phylogenies. Cornell University Library. 2018. DOI: 10.1371/journal.pone.0190826

[7] A.C. Cameron, K.T. Pravin, Regression Analysis of Count Data. Cambridge University Press. 2013.

[8] N. Bhakta, Properties of Hurdle Negative Binomial Models for Zero-Inflated and Overdispersed Count Data. The Ohio University. 2018.

[9] D.W. Hosmer, S. Lemeshow, Applied Logistic Regression Second Edition. New York: John Wiley & Sons. 2000.

[10] A. Carolan, Partially Parametric Testing. Thesis School of Mathematics and Applied Statistics The University of Wollongong. 2000.

[11] H. Akaike, A Bayesian Analysis of The Minimum AIC Procedure. The Institute of Statistical Mathematics. 1978. pp. 9-14. DOI: https://doi.org/10.1007/BF02480194