Mathematics and Statistics 10(3): 603-609, 2022 http://www.hrpub.org
DOI: 10.13189/ms.2022.100316

Simulation Study of Bayesian Hurdle Poisson Regression on the Number of Deaths from Chronic Filariasis in Indonesia

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

Department of Statistics, Faculty of Mathematics and Natural Science, Brawijaya University, Indonesia

Received February 13, 2022; Revised April 24, 2022; Accepted May 23, 2022

Cite This Paper in the following Citation Styles

(a): *[1] Nur Kamilah Sa'diyah, Ani Budi Astuti, Maria Bernadetha T. Mitakda , "Simulation Study of Bayesian Hurdle Poisson Regression on the Number of Deaths from Chronic Filariasis in Indonesia," Mathematics and Statistics, Vol. 10, No. 3, pp. 603 - 609, 2022. DOI: 10.13189/ms.2022.100316.*

(b): *Nur Kamilah Sa'diyah, Ani Budi Astuti, Maria Bernadetha T. Mitakda (2022). Simulation Study of Bayesian Hurdle Poisson Regression on the Number of Deaths from Chronic Filariasis in Indonesia. Mathematics and Statistics, 10(3), 603 - 609. DOI: 10.13189/ms.2022.100316.*

Abstract One regression model to explain the relationship between predictor and response variable in the form of count is Poisson regression. In the case of certain Poisson with the presence of many zero values, causing overdispersion can be overcome with the Poisson Hurdle model. There is a good method for estimating the parameters on small sample sizes for all distributions, namely the Bayesian method. The response variable of the original data does not follow Poisson distribution, so parameter will be estimated by Bayesian method. The performance of the Bayesian Hurdle Poisson regression can be seen from simulation data on various sample sizes and overdispersion levels generated based on the parameters of original data showing that the Bayesian Hurdle Poisson regression model proposed in this study is suitable for large sample sizes or with varying levels of overdispersion due $n \geq 30$ or $n \to \infty$ because normal distribution is used as prior. Even though the response variable of the simulation data is generated with a Poisson distribution, it still does not follow a Poisson distribution because it's in accordance with the original data. The parameter estimated based on the simulation data is similar to the parameter estimated on the original data (both the estimator of the MLE Hurdle Poisson regression parameter and the parameter estimator of the Bayesian Hurdle Poisson regression). This indicates that the simulation scenario is appropriate.

Keywords Bayesian, Hurdle, Overdispersion, Poisson

1. Introduction

Poisson regression model explains the relationship between the predictor and the response variable in the form of the count. An important assumption in Poisson regression analysis is that there is equidispersion, where the mean of the distribution is equal to its variance [1].

In certain cases, where response consists of many zero counts, causing the variance to be greater than the mean is known as overdispersion. This might also be the case when regression assumptions were violated. When overdispersion occurs, Poisson regression is less precise, because the model formed will produce a refractive parameter estimation [2]. Overdispersion data can be handled by the Hurdle model.

The method of parameter restoration used in this study is the Bayesian method. The advantage of the Bayesian method is that it can estimate parameters at small extreme observation values and can be used for all distributions.

This study used simulation data ($n = 30,50,100$) based on the number of chronic Filariasis cases in 34 Provinces in Indonesia. Each simulated data is generated with three levels of overdispersion: low $(1 < \chi^2_{Pearson}/df \leq 10)$, moderate $(10 < \chi^2_{Pearson}/df \leq 20)$, and high $(\chi^2_{Pearson}/df > 20)$. This category is designed to test how well the Poisson Hurdle model estimates the parameters

using generated data with the Bayesian method used based on different sample sizes and overdispersion levels.

This research is expected to provide good benefits for statistical users as a reference in using the Hurdle method, where the response variable is in the form of a count containing overdispersion. In addition, it can be a reference that besides classical statistical method there is an alternative way based on data driven, namely Bayesian method.

2. Literature Review

This sub chapter will discuss the literature review analysis that will be used in this research which are: Poisson regression, Hurdle regression, Bayesian.

2.1. Poisson Regression

The regression method for modeling causal relationships between discrete response variables (Poisson distribution) with one or more predictor variables (it can be discrete or continuous) is called Poisson regression as presented in equation (1)[4].

$$y_i = \lambda_i + \varepsilon_i \quad (1)$$

where:

y_i : the i^{th} observation value of the response variable

λ_i : the i^{th} mean response variable Y that is influenced by the predictor variable values

ε_i : the i^{th} observation error

2.2. Poisson Distribution Conformity

One of the assumptions that must be fulfilled in Poisson regression is that the response variable (Y) follows a Poisson distribution. The assumption $Y \sim \text{Poisson}(\lambda)$ was tested with the Kolmogorov-Smirnov test based on the hypothesis:

H_0: $F_N(y) = P(y, \lambda)$, the response variable follows a Poisson distribution

H_1: $F_N(y) \neq P(y, \lambda)$, the response variable does not follow a Poisson distribution

The Kolmogorov-Smirnov statistical test for the Poisson distribution conformity test is presented in equation (2) [5].

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

where:

$F_N\left(y_{(i)}\right)$: cumulative function of sample,

$P\left(y_{(i)}, \lambda\right)$: cumulative probability function of the Poisson distribution,

$y_{(i)}$: ranking statistics of response variable

k : $0,1,2, \ldots \infty$

Reject H_0 if or $D > D_{(n,\alpha)}$ or $p_{value} < 0.05$.

2.3. Non-Multicollinierity Test

To measure the strength of the relationship between predictor variables, *the VIF* (*Variance Inflating Factor)* criteria are used.

$$VIF_j = \frac{1}{1-R_j^2} \quad (3)$$

where:

R_j^2: coefficient determination of auxiliary regression

As a standard rule, if VIF of a predictor variable exceeds 10, which will occur if the R^2exceeds 0.90, the predictor variable (which serves as the response variable in auxiliary regression) is said to be very closely related to other $(k-1)$ predictors [6].

2.4. Overdispersion

One of the assumptions of Poisson regression analysis is equidispersion (the mean equal the variance) in the response variable. However, not all data fulfill these assumptions because there may also be problems of underdispersion and overdispersion. Overdispersion is a situation in which the variance is greater than mean. One of the causes of overdispersion is the homogenity in observed subjects [7]. Examination of the occurrence of overdispersion can be done using the Pearson Chi-Squared statistic divided by Poisson regression degrees of freedom, mathematically written in equation (4).

$$\chi^2_{Pearson} = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \sim \chi^2_{(n-p)} \quad (4)$$

where:

$\hat{\mu}_i = \hat{\lambda}_i = exp\left(\hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\beta}_j\right)$

n : number of observations

p : number of parameters $(k+1)$

x_{ij} : the value of the j^{th} predictor variable on the i^{th} observation

If $(\chi^2_{Pearson}/(n-p)) > 1$ then it can be concluded that observations contain overdispersion.

2.5. Hurdle Poisson Regression Model

Probability function of the Hurdle model is a combination of probability on logit models and *Poisson truncated* models [8] presented in equation (5) [9].

$$P(Y = y_i) = \begin{cases} (1-\pi_i) & , y_i = 0 \\ \pi_i \frac{e^{-\lambda_i}\lambda_i^{y_i}}{\left(1-e^{-\lambda_i}\right)y_i!} & , y_i > 0 \end{cases} \quad (5)$$

Equation (5) has parameters π and λ that fulfill equation (6) and (7).

$$logit\ \pi_i = \ln(\pi_i/(1-\pi_i)) = \delta_0 + \sum_{j=1}^{k} x_{ij}\delta_j + \varepsilon_i \quad (6)$$

$$\ln \lambda_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j + \varepsilon_i \quad (7)$$

Based on equation (6) and (7), Hurdle Poisson models are derived shown in equation (8) and (9).

$$\pi_i = \frac{exp\left(\delta_0+\sum_{j=1}^{k} x_{ij}\delta_j\right)}{1+exp\left(\delta_0+\sum_{j=1}^{k} x_{ij}\delta_j\right)} \quad (8)$$

$$\lambda_i \quad = exp\left(\beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j\right) \quad (9)$$

In equation (8) provide logit model and equation (9) is truncated model where β and δ are the parameters to be estimated.

2.6. Bayesian Hurdle Poisson Regression

There are three important components of the Bayesian method, (1) the likelihood function of the HPR model, (2) the prior distribution and (3) the posterior distribution. The likelihood function of the HPR model is presented in equation (10).

$$f(Y|\beta,\delta) = \prod_{\substack{i=1\\ y_i=0}}^{n} \frac{1}{1+exp(\boldsymbol{X^T\delta})} \times \prod_{\substack{i=1\\ y_i>0}}^{n} \frac{\left[exp\left(-exp(\boldsymbol{X^T\beta})\right)\right]\left[exp(\boldsymbol{X^T\beta})\right]^{y_i}}{\left(1-\left[exp\left(-exp(\boldsymbol{X^T\beta})\right)\right]\right)y_i!} \quad (10)$$

The prior distribution for β and δ are assumed to normally distributed with mean μ and variance σ^2 as presented in equation (11).

$$f(\beta,\delta) \quad = \prod_{j=0}^{k} \frac{1}{\sigma_\beta\sqrt{2\pi}} exp\left(-\frac{(\beta-\mu_\beta)^2}{2\sigma_\beta^2}\right) \times \quad (11)$$

The posterior distribution is the multiplication between the likelihood function and the prior distribution with the form of the equation as presented in equation (12).

$$f(\beta,\delta|Y) \propto f(Y|\beta,\delta)f(\beta,\delta)$$

$$f(\beta,\delta|Y) \propto \prod_{\substack{i=1\\ y_i=0}}^{n} \frac{1}{1+exp(\boldsymbol{X^T\delta})} \prod_{\substack{i=1\\ y_i>0}}^{n} \frac{\left[exp\left(-exp(\boldsymbol{X^T\beta})\right)\right]\left[exp(\boldsymbol{X^T\beta})\right]^{y_i}}{\left(1-\left[exp\left(-exp(\boldsymbol{X^T\beta})\right)\right]\right)y_i!} \times \prod_{j=0}^{k} \frac{1}{\sigma_\beta\sqrt{2\pi}} exp\left(-\frac{(\beta-\mu_\beta)^2}{2\sigma_\beta^2}\right) \prod_{j=0}^{k} \frac{1}{\sigma_\delta\sqrt{2\pi}} exp\left(-\frac{(\delta-\mu_\delta)^2}{2\sigma_\delta^2}\right) \quad (12)$$

2.7. Bayesian Model Convergence Test

One way to check the accuracy of parameter estimator with the Bayesian method is by conducting a convergence test of model parameters. There are four methods in convergence test [10].

1. Trace Plot

Plot between the estimator that has been simulated on iteration. If the trace plot shows a random pattern then convergence is fulfilled.

2. Autocorrelation Plot

Autocorrelation measures the relationship between simulated values $\left\{x_j^{(t)}\right\}$ and $\left\{x_j^{(t+L)}\right\}$ where L is the amount of lag of separate iterations on two sets of values. If the first lag on the autocorrelation plot is close to one and the next lag is close to zero then convergence is met.

3. Ergodic Mean Plot

Plot between iteration and moving average of parameter estimator, convergence will be fulfilled if after several iterations the ergodic mean plot is stable.

4. Monte Carlo Error (MC Error)

The process of calculating MC error uses the mean batch method by dividing the sample of generation results in K batches. The formula for calculating MC errors, namely:

$$MCE[G(\theta)] = \sqrt{\frac{1}{K(K-1)}\sum_{b=1}^{K}\left(\overline{G(\theta)_b} - \overline{G(\theta)}\right)^2} \quad (13)$$

If the MC error is less than 5% of the standard deviation of each parameter then convergence is fulfilled.

2.8. Best Model Selection

The *model goodness of fit* criteria in the Bayesian method uses *the deviance information criterion* (DIC) criteria. In the Bayesian method, the DIC criterion has advantages over other criteria because it is easily calculated from samples generated by MCMC simulations [11]. The best models have the smallest DIC. The DIC criterion is defined in equation (14).

$$DIC = \overline{D(\theta)} + p_D \quad (14)$$

where:

$p_D = \overline{D(\theta)} - D(\bar{\theta})$

$\overline{D(\theta)}$: Average deviation of posterior distribution θ

$D(\bar{\theta})$: The expected deviation on the posterior $a$verage, θ

3. Research Methodology

The original data of the study was sourced from the 2020 Indonesian Health Profile related to chronic Filariasis cases had 34 observations with five predictor variables and one response variable. Simulation data is generated with various sample sizes, namely 30, 50, and 100 based on the parameters of original data. The steps of data generation simulation are as follows:

1. Generates predictor variable $X_{ij} \sim uniform$ with the minimum and maximum values of as much original data.$n = 30,50,100$
2. Use $\hat{\beta}_{origin}$ and $\hat{\delta}_{origin}$ as a regression coefficient simulation data.
3. Calculating π_i and λ_i where:

$$\pi_i = \frac{exp\left(\hat{\delta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\delta}_j\right)}{1 + exp\left(\hat{\delta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\delta}_j\right)}$$

$$\lambda_i = exp\left(\hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\beta}_j\right)$$

4. Generates $Y_i \sim Poisson(\pi_i + \lambda_i)$ with low level $(1 < \chi^2_{Pearson}/df \leq 10)$, moderate level $(10 < df \leq 20)$, and high level $(\chi^2_{Pearson}/df > 20)$ of overdispersion.

After obtaining the simulation data, the analysis is carried out as follows:

1. Testing the suitability of Poisson's distribution using the Kolmogorov-Smirnov test on equation (2).
2. Testing Non-Multicollinierity based on the VIF criteria in equation (3).
3. Calculate $\chi^2_{Pearson}$ statistics for testing overdispersion according to equation (4).
4. Estimate the parameters of Hurdle Poisson regression model using the Bayesian method corresponds to equation (12) with the stages:
   (1) Determines the initial value of the parameter β^0 and δ^0
   (2) For each iteration.$t = 1,2,\ldots,T$
      a Generates the value of $\beta^{(1)}$ from $f\left(\beta \middle| \delta^{(t-1)}, x\right)$
      b Generates the value of $\delta^{(1)}$ from $f\left(\delta \middle| \beta^{(t)}, x\right)$
      c Repeat Steps (a) and (b) until they are convergence.
5. Examine the convergence of model parameters with the Bayesian method in three ways, namely *trace plot*, autocorrelation plot, and *ergodic mean plot*.
6. The selection of the best model is done according to the criteria of the DIC in equation (14).
7. Interpretation and conclusion.

4. Result and Discussion

The analysis on the simulation data begins with a Poisson distribution conformity test and an overdispersion test to see how similar the simulated data is to the original data.

4.1. Poisson Distribution Conformity Testing

The response variable of simulated data was generated with the Poisson distribution. The results of Poisson distribution conformity test on the response variable are presented in Table 1.

Table 1. Results of Conformity Testing Of Poisson Distribution for Simulation Data

n	Level of Overdispersion	Statistics Kolmogorov Smirnov (D)	p_{value}
30	Low	0.789	0.000
	Moderate	0.831	0.000
	High	0.766	0.000
50	Low	0.960	0.000
	Moderate	0.960	0.000
	High	0.960	0.000
100	Low	0.843	0.000
	Moderate	0.930	0.000
	High	0.930	0.000

Judging from Table 1, all response variable in the simulation data does not follow Poisson distribution because of $p_{value} < 0.05$. This happened because the response variable original data does not follow Poisson distribution, and the pattern of response variable corresponds to the original data pattern (not follow Poisson distribution). Identification of the proper distribution of Y variables is done with EasyFit software. The identification results show that the Poisson distribution ranked third after uniform and geometric distribution (the same as the original data). Therefore, the Bayesian method to estimate the parameters is applied because it has advantages that can be applied to all distributions (*any distribution*).

4.2. Testing Non-Multicollinierity between Predictor Variable

Non-multicollinierity between predictor variable is presented in Table 2.

Table 2. Results of Non-Multicollinierity Testing Between Predictor Variable

n	Level of Overdispersion	VIF_j				
		X_1	X_2	X_3	X_4	X_5
30	Low	1.17	1.11	1.18	1.24	1.32
	Moderate	1.17	1.11	1.18	1.24	1.32
	High	1.11	1.05	1.12	1.17	1.11
50	Low	1.05	1.11	1.12	1.23	1.10
	Moderate	1.05	1.11	1.12	1.23	1.10
	High	1.05	1.11	1.12	1,.23	1.10
100	Low	1.06	1.03	1.03	1.04	1.02
	Moderate	1,02	1,02	1.01	1.01	1.00
	High	1.02	1.02	1.01	1.01	1.00

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

4.3. Overdispersion Testing

Table 3 presents descriptive analysis results (percentage value 0 on response variable) and $\chi^2_{Pearson}/df$ values at various sample sizes and overdispersion levels.

Table 3. Overdispersion Test Results for Simulation Data

n	Level of Overdispersion	Percentage Value 0 On Response Variable (%)	$\chi^2_{Pearson}/df$
30	Low	46.67	8.67
	Moderate	36.67	14.94
	High	36.67	38.63
50	Low	58.00	4.13
	Moderate	54.00	18.69
	High	58.00	36.33
100	Low	50.00	2.18
	Moderate	55.00	15.31
	High	52.00	31.73

Table 3 shows that a sample size $n = 30$ with a low overdispersion level has the highest percentage value (46.67%) of 0 on response variable compared to moderate and high overdispersion levels at the same sample size and test statistic $\chi^2_{Pearson(24)}$ is 8.66

In sample size 50, low and high overdispersion levels are equal to 58% of 0 on the response variable, and test statistic 4.13 for low overdispersion level and 36.33 for high overdispersion level.

In sample size 100 with a moderate overdispersion level has the highest percentage of 0 at response variable (55%) has $\chi^2_{Pearson(94)}$ is 15.315. This indicates that the percentage value of 0 on the response variable does not guarantee how high the test statistic is. However, the value of predictor variable also plays an important role in the level of overdispersion considering equation (4).

4.4. Results of Estimation of MLE Hurdle Poisson Regression Parameters

The simulation data was generated based on the original Poisson MLE Hurdle regression parameter. If the estimator of Hurdle Poisson regression parameters in both data is not much different, then the simulation built is appropriate. The estimator of the MLE Hurdle Poisson regression parameter in the simulation data is presented in Table 4 and Table 5.

Table 4 and Table 5 show that statistics $\hat{\delta}_j$ and $\hat{\beta}_j$ of simulation data and simulation data are not different. This proves that the data simulations are appropriate.

4.5. Convergence of Bayesian Model Parameters

Methods for parameter convergence tests are carried out on simulation data, namely (1) Trace Plot, (2) Autocorrelation Plot, and (3) Ergodic Mean Plot (4) MC error. In the original data, convergence was achieved when conducted 300,000 iterations and 7 batches (thin).

In simulation data $(n = 30{,}50{,}100)$ with low, moderate, and high levels of overdispersion converged when iterated as many as 300,000 and 7 batches (*thin*). This is in line with the convergence examination of the original data.

Table 4. Results of Parameter Estimation of MLE Hurdle Poisson Regression on Simulation Data for $y_i = 0$

Data	n	Level of Overdispersion	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$	$\hat{\delta}_4$	$\hat{\delta}_5$
Original	34	High	-14.43	-4.87×10^{-3}	0.20	0.25	-2.08×10^{-4}	0.008
Simulation	30	Low	-10.68	-1.03×10^{-3}	0.46	0.53	-3.77×10^{-4}	0.13
		Moderate	-6.76	-9.78×10^{-3}	0.13	0.39	-1.27×10^{-4}	0.10
		High	-1.47	-2.88×10^{-3}	0.66	0.41	-7.00×10^{-5}	0.23
	50	Low	-7.98	-3.06×10^{-3}	0.59	0.48	-6.30×10^{-4}	0.18
		Moderate	-10.92	-1.85×10^{-3}	0.45	0.54	-2.75×10^{-4}	0.16
		High	-10.56	-1.04×10^{-3}	0.24	0.33	-3.66×10^{-4}	0.17
	100	Low	-9.34	-2.34×10^{-3}	0.86	0.70	-3.82×10^{-4}	0.15
		Moderate	-5.46	-1.38×10^{-3}	0.31	0.45	-4.37×10^{-4}	0.009
		High	-6.59	-1.40×10^{-3}	0.35	0.57	-4.34×10^{-4}	0.009

Table 5. Results of Parameter Estimation of MLE Hurdle Poisson Regression on Simulation Data for $y_i > 0$

Data	n	Level of Overdispersion	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Original	34	High	-4.21	-2.70×10^{-3}	0.15	0.51	-4.00×10^{-4}	8.02×10^{-2}
Simulation	30	Low	-3.89	-2.32×10^{-3}	0.14	0.45	-3.08×10^{-4}	7.85×10^{-2}
		Moderate	-3.90	-2.54×10^{-3}	0.17	0.50	-3.78×10^{-4}	7.42×10^{-2}
		High	-3.36	-2.65×10^{-3}	0.17	0.48	-3.38×10^{-4}	6.89×10^{-2}
	50	Low	-4.24	-2.58×10^{-3}	0.15	0.52	-3.72×10^{-4}	7.81×10^{-2}
		Moderate	-4.06	-2.35×10^{-3}	0.16	0.53	-3.94×10^{-4}	7.11×10^{-2}
		High	-4.41	-2.53×10^{-3}	0.12	0.53	-3.69×10^{-4}	8.06×10^{-2}
	100	Low	-3.98	-2.68×10^{-3}	0.15	0.50	-3.73×10^{-4}	7.87×10^{-2}
		Moderate	-4.37	-2.65×10^{-3}	0.16	0.49	-3.66×10^{-4}	8.28×10^{-2}
		High	-4.50	-2.67×10^{-3}	0.15	0.51	-3.69×10^{-4}	8.29×10^{-2}

4.6. Parameters of the Bayesian Hurdle Poisson Regression Model 0

After conducting an iteration of 300,000 and 7 *batches* (thin), then obtained the estimator of the Bayesian Hurdle Poisson regression parameter for $y_i = 0$ presented in Table 6 and the estimator of the Bayesian Hurdle Poisson regression parameter for $y_i > 0$ in Table 7.

Table 6. Results of Parameter Estimation of Bayesian Hurdle Poisson Regression on Simulation Data for $y_i = 0$

Data	n	Level of Overdispersion	$\hat{\delta}_0$	$\hat{\delta}_1$	$\hat{\delta}_2$	$\hat{\delta}_3$	$\hat{\delta}_4$	$\hat{\delta}_5$
Original	34	High	16.65	4×10^{-4}	-0.23	-0.30	6×10^{-4}	-0.19
Simulation	30	Low	22.26	2.12×10^{-3}	-0.91	-1.07	7.28×10^{-4}	-0.27
		Moderate	12.34	1.79×10^{-3}	-0.24	-0.73	2.31×10^{-4}	-0.19
		High	29.13	5.96×10^{-3}	-1.29	-0.83	1.44×10^{-4}	-0.47
	50	Low	13.55	5.15×10^{-3}	-0.98	-0.82	1.07×10^{-4}	-0.30
		Moderate	15.44	2.70×10^{-3}	-0.66	-0.78	3.94×10^{-4}	-0.22
		High	14.81	1.51×10^{-3}	-0.34	-0.47	5.35×10^{-4}	-0.24
	100	Low	11.68	2.92×10^{-3}	-0.10	-0.87	4.68×10^{-4}	-0.19
		Moderate	6.43	1.67×10^{-3}	-0.37	-0.53	5.29×10^{-4}	-0.11
		High	7.90	1.68×10^{-3}	-0.41	-0.69	5.28×10^{-4}	-0.11

Table 7. Results of Parameter Estimation of Bayesian Hurdle Poisson Regression on Simulation Data for $y_i > 0$

Data	n	Level of Overdispersion	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$
Original	34	High	-4.24	-2.70×10^{-3}	0.15	0.51	-4.0×10^{-4}	8.05×10^{-2}
Simulation	30	Low	-4.15	-2.35×10^{-3}	0.14	0.45	-3.1×10^{-4}	8.12×10^{-2}
		Moderate	-4.19	-2.57×10^{-3}	0.18	0.50	-3.8×10^{-4}	7.72×10^{-2}
		High	-3.50	-2.67×10^{-3}	0.17	0.49	-3.4×10^{-4}	6.97×10^{-2}
	50	Low	-4.20	-2.57×10^{-3}	0.15	0.52	-3.7×10^{-4}	7.69×10^{-2}
		Moderate	-4.05	-2.35×10^{-3}	0.16	0.53	-4.0×10^{-4}	7.06×10^{-2}
		High	-4.37	-2.52×10^{-3}	0.12	0.53	-3.7×10^{-4}	7.97×10^{-2}
	100	Low	-3.99	-2.70×10^{-3}	0.15	0.50	-3.7×10^{-4}	7.88×10^{-2}
		Moderate	-4.36	-2.65×10^{-3}	0.16	0.49	-3.7×10^{-4}	8.27×10^{-2}
		High	-4.52	-2.67×10^{-3}	0.15	0.51	-3.7×10^{-4}	8.31×10^{-2}

Consider Table 6 and Table 7, where $\hat{\delta}_j$ and $\hat{\beta}_j$ of the simulation data is not much different from the original data. The DIC of simulation data is discussed in sub Chapter 4.6.

4.7. Selection of the Best Model

DIC criteria are used as the goodness of Bayesian method, and the best models have the smallest DIC. The DIC of the simulation data is presented in Table 8.

Table 8. DIC of Simulated Data

n	Level of Overdispersion	DIC
30	Low	-6341.0
	Moderate	-10324.1
	High	-13463.2
50	Low	-337508.0
	Moderate	-485,913.4
	High	-208,563.4
100	Low	-46,076.5
	Moderate	-686,547.1
	High	-419,2

Table 8 shows that simulation data for $n = 30$ with a high overdispersion level has the smallest DIC, compared to low and moderate overdispersion levels. The simulation data for $n = 50$ with moderate overdispersion level has the smallest DIC, compared to low and high overdispersion levels. Simulation data for $n = 100$ with moderate overdispersion level has the smallest DIC compared to low and high overdispersion levels. This suggests that the Hurdle Poisson regression model with the Bayesian method performed in this study is suitable for use on $n \geq 30$ or $n \to \infty$ sample sizes with varying levels of overdispersion because the prior distribution a normal. If the sample size is less than 30, it is considered a different prior distribution appropriate to the data behavior.

5. Conclusions

The performance of the Bayesian Hurdle Poisson regression model can be seen from the simulation data $(n = 30, 50, 100)$ generated based on the parameters of original data with low, moderate, and high overdispersion levels. Based on the DIC, the same sample size with different levels of overdispersion does not indicate that certain levels of overdispersion are better than other levels of overdispersion. Likewise, at the same level of overdispersion with different sample sizes, it does not indicate that certain sample sizes are better than other sample sizes. It shows that the Bayesian Hurdle Poisson regression model proposed in this study is suitable for use on $n \geq 30$ or $n \to \infty$ sample sizes with varying levels of overdispersion due to a normal prior distribution. The parameter estimator obtained from the simulation data is similar to the estimators of the original data (both the estimation of the MLE Hurdle Poisson regression parameter and the Bayesian Hurdle Poisson regression). This indicates that the simulation scenario is appropriate. This study has not modeled the best fitted distribution (uniform and geometric). Therefore, other researchers can further model Hurdle uniform regression or Hurdle geometric regression in handling overdispersion.

REFERENCES

[1] F. Famoye, J. T. Wulu, and K. P. Singkh. On The Generalized Poisson Regression Model with an Application to Accident Data, Journal of Data Science, Vol. 2, 287-295, 2004. DOI: https://doi.org/10.6339/JDS.2004.02(3).167

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

[3] Taufiq, A. B. Astuti, and A. A. R. Fernandes. Geographically Weighted Regression in Cox Survival Analysis for Weibull Distributed with Bayesian Approach, IOP Conference Series Material Science and Engineering, 2019.

[4] C. Cameron and K. T. Pravin. Regression Analysis of Count Data, Cambridge University Press, 2013. DOI: https://doi.org/10.1017/CBO9781139013567

[5] F. Antoneli, F. Passos, F. M. Lopes, and R. S. Briones. A Kolmogorov Smirnov Test for the Molecular Clock Based on Bayesian Ensembles of Phylogenies, Cornel University Library, 2018. DOI: https://doi.org/10.1371/journal.pone.0193633

[6] D. N. Gujarati and C.P. Dawn. Dasar Dasar Ekonometrika Edisi 5, Salemba Empat, 2012.

[7] Agresti. Categorical Data Analysis Second Edition, John Wiley and Sons, 2002.

[8] E. Cantoni and A. Zedini. A Robust Version of the Hurdle Model, Journal of Statistical Planning and Inference, Vol. 141, No.3, 1214-1223, 2011. DOI: https://doi.org/10.1016/j.jspi.2010.09.022.

[9] E. Saffari, S. R. Adnan, and W. Greene. Parameter Estimation of Hurdle Poisson Regression Model with Censored Data, Jurnal Teknologi Vol. 57, No. 1, 2012. DOI: https://doi.org/10.11113/jt.v57.1533.

[10] Ntzoufras. Bayesian Modelling Using WinBUGS, John Wiley and Sons, 2011

[11] D. J. Spiegelhalter, N.G. Best, B.P. Carlin, and A.V.D. Linde. Bayesian Measure of Model Complexity and Fit, Royal Statistical Society, Vol. 64, No. 4, 583-639, 2002. DOI: https://doi.org/10.1111/1467-9868.00353.