Modeling and hypothesis testing for the factors affecting infant’s diarrhea using Generalized Poisson Regression

B W Otok1*, C B G Allo1, and Purhadi1
1Institut Teknologi Sepuluh Nopember (ITS), Surabaya 60111, Indonesia

*Corresponding author’s email: dr.otok.bw@gmail.com

Abstract. Infants are weak individuals. Number of infants with Diarrhea is count data. Count data can be modeled using Poisson Regression. Poisson Regression has assumption that must be met. In the real case, overdispersion or underdispersion often occurs in data. This condition causes Poisson Regression cannot be used to model the data. Another alternative used to model the data with violation of assumption in Poisson Regression is Generalized Poisson Regression. This article will estimate the parameters of Generalized Poisson Regression using Generalized Poisson Regression. After getting the estimate parameters, parameters hypothesis testing simultaneously is done using Maximum Likelihood Ratio Test. There are three independent variables. They are percentage of infants who get exclusive breastfeeding, percentage of infants who get complete basic immunization, and percentage of households who have healthy living behavior. Significant parameter used to build the model. So, model for the factors affecting Diarrhea in infants in Pasuruan Regency is a model consisting complete basic immunization and healthy living behavior in the model.

1. Introduction
Poisson Regression is a standard model for count data. The assumption in Poisson Regression, equidispersion, is a condition that the variance is equal with the average of the response variable. If the variance of variable response is greater than the mean, then it is called overdispersion. If the variance of variable response is smaller than the mean, then it is called underdispersion. The development of the Poisson Regression model has been done by many researchers. One of the developments is parameter estimation of Geographically Weighted Multivariate Poisson Regression [1].

Poisson Regression model cannot be used if the equidispersion assumption is not fulfilled. Generalized Poisson Regression have been developed by many researchers. Generalized Poisson Regression can be used when equidispersion assumption is not fulfilled. Poisson Regression (PR) and Generalized Poisson Regression (GPR) are applied in motor insurance claim data in Malaysia [2]. The performance of Poisson, Negative Binomial, and GPR model are compared in the prediction of antenatal care visits in Nigeria [3].

Parameter estimation method is used to get estimator from regression model. There are many parameter estimation methods. There are many parameter estimation methods, like Ordinary Least Square (OLS), Maximum Likelihood Estimation (MLE), Method of Moments (MM), and Generalized Method of Moment (GMM). Maximum Likelihood Estimation (MLE) is used to get estimation of Bivariate Generalized Poisson Regression (BGPR) [4]. They also used Maximum Likelihood Ratio
Test (MLRT) to get hypothesis testing of BGPR. The result of estimate parameter of PR model using MLE and GMM are compared using Acute Respiratory Tract Infection data [5].

Diarrhea is a disease that is familiar to the public. Diarrhea is infectious disease. Diarrhea can be transmitted through food. Children specifically infants affected by Diarrhea need special attention because Diarrhea can cause death. Prevalence of Diarrhea is 8 percent in Indonesia and East Java is in the nineteen position of thirtythree province [6]. Diarrhea is both preventable and treatable. So, it is important to reduce incident of diarrhea in infants.

Parameters of GP are estimated by GMM because the number of moment condition is greater than the parameters. Data of infants affected Diarrhea in Pasuruan Regency is used as application of GPR.

2. Literature review

2.1. Poisson Regression (PR)

A random variable $Y$ is said to have a Poisson probability distribution if only if

$$f(y) = \frac{e^{-\mu} \mu^y}{y!} \quad y = 0, 1, 2, ...$$

with $\mu > 0$ [7]. The mean and variance of $Y$ can be seen in (2).

$$E(Y) = V(Y) = \mu$$

The Poisson Regression (PR) model can be seen in (3). 

$$\mu_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik})$$

2.2. Generalized Poisson Regression (GPR)

Let $Y_j \sim GP(\mu_j, \phi)$ where $j = 1, 2, ..., n$ then $Y_j$ has a probability function:

$$f(y_j | \mu_j, \phi) = \left[ \frac{\mu_j}{1 + \phi \mu_j} \right]^{y_j} \frac{(1 + \phi y_j)^{y_j-1}}{y_j!} \exp \left( -\frac{\mu_j (1 + \phi y_j)}{1 + \phi \mu_j} \right)$$

where:

$$\mu_j = e^{x_j^T \beta}$$

$$x_j = \begin{bmatrix} 1 & x_{j1} & x_{j2} & ... & x_{jk} \end{bmatrix}^T$$

$$\beta = \begin{bmatrix} \beta_0 & \beta_1 & \beta_2 & ... & \beta_k \end{bmatrix}^T$$

If $\phi = 0$ then equidispersion. It means equation (4) will same with equation (1). If $\phi > 0$ then overdispersion. If $\phi < 0$ then underdispersion [8]. The mean and variance of GP can be seen in (5) and (6). There is no difference between Poisson Regression model and Generalized Poisson Regression model. So, the GPR model can be seen in (3).

$$E(Y_j) = \mu_j$$

$$V(Y_j) = \mu_j \left(1 + \phi \mu_j \right)^2$$

2.3. Method of Moments (MM)

Important thing in Method of Moments is choose the moment condition of model. The moment conditions can be seen in (7) [9]:

$$E(h(y_j, \theta_0)) = 0$$

After getting moment conditions, the sample moment conditions can be searched by (8).
\[ h_n(\theta) = \frac{1}{n} \sum_{j=1}^{n} h(y_j, \theta) \]  

(8)

So, the \( \hat{\theta} \) by Method of Moments can be obtained by solve the equation in (9). The method of moments only applies when the number of moment conditions is equal with the number of parameters.

\[ h_n(\theta) = 0 \]  

(9)

2.4. Generalized Method of Moments (GMM)

Another condition is when the number of moment conditions is greater than the number of parameters. So, we cannot use Method of Moments (MM). First time, GMM is proposed by Hansen (1982) [10]. The method of moments only applies when the number of moment conditions is equal with the number of parameters.

Estimation of weighted matrix as follows:

\[
\hat{W}(\theta^*) = \frac{n}{n-p} \sum_{i=1}^{n-p} \left( \frac{y_i}{b} \right)^T \hat{\Gamma}(s)
\]

(10)

where:

- \( n \) : number of observations
- \( p \) : number of parameters
- \( t \) : kernel function
- \( b \) : bandwidth

\[
\hat{\Gamma}(s) = \begin{cases} 
\frac{1}{n} \sum_{j=1}^{n} h(\theta^*, y_j, x_j) h(\theta^*, y_j, x_{j-s})^T & \text{if } s \geq 0 \\
\frac{1}{n} \sum_{j=1}^{n} h(\theta^*, y_j, x_j) h(\theta^*, y_j, x_j)^T & \text{if } s < 0
\end{cases}
\]

There are five types of kernel functions. They are Truncated, Bartlett, Parzen, Tukey-Hanning, and Quadratic Spectral. Choice of kernel functions does not affect the asymptotic properties of GMM, very little is known about the impacts in finite samples [12]. The best bandwidth can be obtained by minimizing the MSE.

Iterative GMM (ITGMM) proposed by Hansen et al (1996) is development of two stage GMM (2GMM) proposed by Hansen (1982) [13]. The algorithm is as follows [12]:

- Compute \( \theta^{(0)} = \arg \min_{\theta} h_n(\theta)^T h_n(\theta) \)
- Compute the weighted matrix, \( \hat{W}(\theta^{(c-1)}) \). \( \hat{W}(\theta^{(c-1)}) \) is computed by (10).
- Compute \( \theta^{(c)} = \arg \min_{\theta} h_n(\theta)^T \hat{W}(\theta^{(c-1)})^{-1} h_n(\theta) \)
- If \( \| \theta^{(c)} - \theta^{(c-1)} \| < \epsilon_c \) stops where \( \epsilon_c \) as small as we want, else \( \theta^{(c)} = \theta^{(c-1)} \) and go to 2 where \( c = 1, 2, \ldots, C \) and \( C \) is number of iteration.
- Define the ITGMM estimator \( \hat{\theta} \) as \( \theta^{(c)} \).

2.5. Maximum Likelihood Ratio Test (MLRT)

Maximum Likelihood Ratio Test (MLRT) is used to get statistics test in hypothesis testing simultaneously. There are several steps in MLRT. The steps for GPR model are as follows:
Making the hypothesis

\[ H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \]

\[ H_1 : \text{There is at least one } \beta_l \neq 0 \text{ where } l = 1, 2, \ldots, k. \]

Determine the parameters set under population

\[ \Omega = \{\beta_0, \beta_1, \ldots, \beta_k, \phi\} \text{ where } n(\Omega) = k + 2 \]

Find \( \hat{\Omega} \)

Determine the parameters set under null hypothesis

\[ \omega = \{\beta_0, \phi\} \text{ where } n(\omega) = 2 \]

Find \( \hat{\omega} \)

Statistics Test finds the distribution of \( G^2 < c \), where \( G^2 = \frac{L(\hat{\omega})}{L(\hat{\Omega})} \). \( H_0 \) is rejected if only if

\[ G^2 > \chi^2_{\alpha, n} \), where \( \nu = n(\Omega) - n(\omega). \]

3. Methodology

The data is secondary data obtained from thirty-three health centers in Pasuruan Regency, East Java in 2017. The variables used in this study are the same with research of Allo et al (2019) [14]. Step analysis is as follows:

Step 1. Check the equidispersion. If equidispersion occurs in the data then use Poisson Regression, else go to step 2.

Step 2. Check the multicollinearity between independent variables. If there is multicollinearity between independent variables the handle the multicollinearity, else go to step 3.

Step 3. Estimate the parameters of GPR using GMM.

Step 4. Testing the estimated parameters.

4. Result and discussion

Step 1 check the assumption of PR. Overdispersion checking needs to be checked. Value of deviance divided by degree of freedom of Infants affected Diarrhea in Pasuruan Regency is 143.15. Value of pearson chi-square divided by degree of freedom of Infants affected Diarrhea is 160.81. This result shows that overdispersion occurs in the data because the value of deviance divided by degree of freedom of Infants affected Diarrhea and value of pearson chi-square divided by degree of freedom of Infants affected Diarrhea are greater than one. Step 2 check the multicollinearity using Variance Inflation Factor (VIF). If VIF value is more than 10 then there is multicollinearity between independent variables. Every VIF value in Table 1 is not more than 10.

| Variable | VIF |
|----------|-----|
| X_1      | 1.088 |
| X_2      | 1.049 |
| X_3      | 1.038 |

Step 3 estimate the parameter of GPR using GMM. The moment conditions must satisfy equation (7). So, the moment conditions for GPR model with the data is as follows:
Follow equation (8), we can get sample moment conditions where 
\[ x_j = [1 \ x_{ij} \ x_{2j} \ x_{3j}]^T \]
and 
\[ \mathbf{b} = [\beta_0 \ \beta_1 \ \beta_2 \ \beta_3]^T. \]
To get the estimator, we need numerical iteration. This paper uses Nelder Mead iteration because function does not need to be differentiable. This paper also uses Quadratic Spectral kernel because the choice of kernel selection does not affect the asymptotic properties of GMM [12] and initial value used is the estimation coefficient from PR with \( \varepsilon_{sm} = \varepsilon_{tr} = 10^{-30} \). The result can be seen in Table 2. Every independent variable gives negative impact to dependent variable.

### Table 2. Estimation result of GPR

| Parameter | Estimator | SE  |
|-----------|-----------|-----|
| \( \phi \) | 0.0986    | 0.0148 |
| \( \beta_0 \) | 6.5535   | 0.4851 |
| \( \beta_1 \) | -9.5227e-05 | 0.0036 |
| \( \beta_2 \) | -0.008    | 0.0038 |
| \( \beta_3 \) | -0.0223   | 0.0081 |

Step 4 is testing the hypothesis. The hypothesis for this case is as follows:
\( H_0: \beta_l = \beta_2 = \beta_3 = 0 \)
\( H_1: \) There is at least one \( \beta_l \neq 0 \) where \( l = 1,2,3. \)
The statistics test can be seen in (12).

\[
G^2 = -2 \ln \left( \frac{L(\hat{\phi})}{L(\hat{\Theta})} \right) = 2 \left[ \ln L(\hat{\Theta}) - \ln L(\hat{\phi}) \right]
\]  

(12)

\( G^2 \) follows Chi Square distribution with degree of freedom \( \nu \), where \( \nu = n(\Theta) - n(\phi) \). The null hypothesis is rejected if \( G^2 > \chi^2_{\nu,0.05} \) [15]. For this case, \( G^2 \) is 235.846. If \( G^2 \) is compared with \( \chi^2_{0.05,3} \), then \( G^2 > \chi^2_{0.05,3} \). So, At least one independent variable is different from zero. On the other hand, the GMM estimator converges as \( n \) goes to infinity to the following distribution [12]:
\[
\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{L} N(0, V)
\]

where:
\[
V = E\left(\frac{\partial h(x, \theta_0)}{\partial \theta}\right)^T W(\theta_0)^{-1} E\left(\frac{\partial h(x, \theta_0)}{\partial \theta}\right)
\]

and \(\theta_0\) is true value.

So, \(\hat{\theta}\) approximately distributed as \(N(\theta_0, \frac{V}{n})\). The result for hypothesis testing partially can be seen in Table 3. Table 3 shows the significant parameters are \(\phi\), \(\beta_0\), \(\beta_1\), and \(\beta_2\) with \(\alpha = 0.05\).

| Parameter | p-value |
|-----------|---------|
| \(\phi\)  | 0.00000 |
| \(\beta_0\) | 0.00000 |
| \(\beta_1\) | 0.48955 |
| \(\beta_2\) | 0.01955 |
| \(\beta_3\) | 0.00286 |

Only significant variables are interpreted in the model. The model for factors affecting Diarrhea in infants in Pasuruan regency can be seen in (14).
\[
\hat{\mu} = \exp(6.5535 - 0.4895X_1 - 0.008X_2 - 0.0223X_3)
\]

Each there is increasing in the \(X_2\), then the average number of infants affected by diarrhea will decrease one with all other variables are constant. Each there is increasing in the \(X_3\), then then the average number of infants affected by diarrhea will decrease 1.02 with all other variables are constant. The result shows how basic immunization and households who healthy living behavior are important. Table 3 shows \(X_1\) is not significant, but exclusive breastfeeding for infants is very important. So, infants can get good immune.

5. Conclusion
Overdispersion occurs in infant’s Diarrhea data in Pasuruan regency. So, Poisson Regression (PR) cannot be used in this data. Generalized Poisson Regression (GPR) can handle the overdispersion in data. In GPR model, the number of moment conditions is greater than the number of parameters. Parameters of GPR are estimated using Generalized Method of Moments. Based on hypothesis testing simultaneously using Maximum Likelihood Ratio Test (MLRT), there are independent variables that significantly affecting Diarrhea in infants. Every coefficient of independent variable is negative. It means all independent variables can reduce Diarrhea in infants. The percentage of infants who get exclusive breastfeeding is not significantly affecting Diarrhea in infants. It can be caused by the high awareness of mothers to give exclusive breastfeeding to their infant. Infants who get complete basic immunization and households who have healthy living behavior can reduce the Diarrhea in infants.

References
[1] Triyanto, Purhadi, Otok B W and Purnami S P 2015 Parameter Estimation of Geographically Weighted Multivariate Poisson Regression Applied Mathematical Sciences 9 4081-93.
[2] Noriszura I and Jemain A A 2005 Generalized Poisson Regression: An Alternative for Risk Classification Jurnal Teknologi 43 39-54.
[3] Yusuf O B and Ugalahi L O 2015 On the Performance of the Poisson, Negative Binomial and Generalized Poisson Regression Models in the Prediction of Antenatal Care Visits in Nigeria American Journal of Mathematics and Statistics 5 128-36.
[4] Wardani D K, Purhadi and Wibowo W 2016 Parameter Estimation and Hypothesis Testing on Bivariate Generalized Poisson Regression Yogyakarta. The 3RDICRIEMS.
[5] Mahpolah, Suharto, Wibowo A and Otok B W 2018 The Estimation of Generalized Method Moment Poisson Regression Model on The Prevalence of Acute Respiratory Tract Infection (RTI) in South Kalimantan CAUCHY-Jurnal Matematika Murni dan Aplikasi 5(3) 161-8.
[6] Kementerian Kesehatan Republik Indonesia 2019 Riset Kesehatan Dara Tahun 2018 (Jakarta: Kementerian Kesehatan Republik Indonesia).
[7] Wackerly D D, Mendenhall W I and Scheaffer R L 2008 Mathematical Statistics with Applications (United States: Thomson Brooks/Cole).
[8] Famoye F, Wulu J T and Singh K P 2004 On The Generalized Poisson Regression Model with An Application to Accident Data Journal of Data Science 2 287-95.
[9] Matyas L 1999 Generalized Method of Moments Estimation (New York: Cambridge University Press).
[10] Hansen L P 1982 Large Sample Properties of Generalized Method of Moments Estimators Econometrica 50 1029-54.
[11] Andrews D W K 1991 Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation Econometrica 59(3) 817-58.
[12] Chausse P 2010 Computing Generalized Method of Moments and Generalized Empirical Likelihood with R Journal of Statistical Software 34(11) 1-35.
[13] Hansen L P, Heaton J and Yaron A 1996 Finite-Sample Properties of Some Alternative GMM Estimators Journal of Business and Economic Statistics 14 262-80.
[14] Caecilia B G Allo, B W Otok and Purhadi 2019 Estimation Parameter of Generalized Poisson Regression Model Using Generalized Method of Moments and Its Application IOP Conference Series: Materials Science and Engineering 546 1-7.
[15] Agresti A 2002 Categorical Data Analysis (New York: John Wiley & Sons).