Modelling of Diarrheal Sufferers Percentage and Certified Drink Water Providers in East Java Using Bivariate Binomial Negative Regression Method

N A P Rahmayanti, Y S Yonani and M F F Mardianto
Study Program of Statistics, Department of Mathematics, Airlangga University, Surabaya, Indonesia

E-mail: nindaayui99@gmail.com, m.fariz.fadillah.m@fst.unair.ac.id

Abstract. Diarrheal is a disease with the condition of faeces becoming soft or fluid. In general, diarrheal results from food and drinks exposed to viruses, bacteria or parasites. Diarrheal is one of the health problems of people in poor and developing countries. Diarrheal often occurs due to the lack of clean water that pass health requirements. The method used in this research is bivariate negative binomial regression which is a regression method to model a pair of response variables, each of which has a negative binomial distribution and correlates. This study uses 38 secondary data from the East Java Provincial Health Office in 2016 about the number of diarrheal sufferers and the number of certified drink water providers which pass health requirements. The results of this research, there are 4 factors that significantly influence the model of diarrheal sufferers and drink water providers who pass health requirements, that are the percentage of residents with access to proper sanitation, public places that pass health requirements, population literacy rates and food management places tested.

1. Introduction
Infectious disease is a disease that affects many Indonesians from a long time ago, including intestinal infections (diarrhea). Diarrhea is a clinical symptom of digestive disorders (intestine) which is characterized by increasing frequency of defecation more than usual and repetitive accompanied by changes in shape and consistency of the stool to be soft or liquid which occurs at least three times in 24 hours. In general, diarrheal results from food and drinks exposed to viruses, bacteria or parasites. Diarrhea is to defecate three or more times a day which may be accompanied by vomiting or bloody stools or vomiting.

Diarrhea is one of the health problems of people in poor and developing countries. Diarrhea often occurs due to lack of clean water that meets health requirements. Drinking water is certainly different from water for cooking, bathing and washing. Good drinking water should be sterile and free from harmful substances and bacteria. However, the community cannot ensure sterility or not the water consumed so that many people end up drinking non-sterile water. This consumption of non-sterile water can cause disease, one of which is diarrhea. The climate and soil conditions that are polluted, coupled with the limitations of infrastructure cause clean water to become very scarce and expensive.
The important role of water in human life is no longer negotiable. We need water for bathing, washing, cooking and the main thing is to drink. Clean water is still a serious problem in Indonesia. While the need for clean water is needed. At present the world has experienced a clean water crisis. The amount of clean water in the world is only 1% that can be consumed. Of the 1% of available clean water, not all of them can be easily accessed by the public. World Health Organization found that 663 million people still have difficulties in accessing clean water in 2015. UNESCO said that in connection with this water crisis, it is predicted that by 2025 almost two-thirds of the world's population will live in areas experiencing water shortages.

Previous research [7] regarding the Association between floods and infectious diarrhea and their effect modifiers using the two-stage model method showed that flooding was significantly associated with infectious diarrhea at the provincial level with a cumulative RR of 1.22 (95% CI): 1.05, 1.43. The modeling of the incidence of diarrhea with a spatial regression approach showed that the variables that influence the incidence of diarrhea are distance and drinking water facilities [1].

These studies only modeled the number of diarrhea sufferers and clean drinking water facilities individually, even though medically there is a relationship between the two diseases. Therefore it will be more realistic when modeled together (simultaneously). The relationship between factors that influence the number of diarrhea sufferers and the number of drinking water providers that meet health requirements will be more useful if modeled together (simultaneously). One statistical method that can be used for the number of diarrhea sufferers and the number of drinking water providers that meet the health requirements together is a bivariate negative binomial regression model which is the development of negative binomial regression. Bivariate negative binomial regression method is a regression method to model a pair of response variables in the form of count data, each of which has a negative binomial distribution and correlates with each other. By knowing the shape of this model, it is expected to know factors that significantly influence the number of diarrhea sufferers and the number of drinking water providers that meet the health requirements in each district / city of East Java so that further prevention can be done to reduce the number of diarrhea patients, especially in East Java through socialization programs, prevention, treatment and environmental care.

2. Material and Method

The data used in this study is secondary data. This data was obtained from the East Java Health Office in 2016. The data used consisted of the number of diarrheal sufferers and drinking water providers who met the health requirements along with the factors that were thought to be influential. The observation unit in this study was 38 districts / cities in East Java Province in 2016.

The variables used in this study consist of predictor variables and response variables. The response variable (Y) used in the study consisted of two response variables (Y) namely the percentage of diarrheal patients and the number of certified drink water providers, while the used predictor variable (X) was presented in Table 1 below:

| Table 1. Predictor Variables |
|-----------------------------|
| Variable | Information |
| x₁ | Percentage of population with access to proper sanitation facilities |
| x₂ | Percentage of public places that meet health requirements |
| x₃ | Population literacy rate |
| x₄ | The number of food management places tested |

In this study, we model data on the number of diarrheal sufferers and the number of certified drink water providers in each district / city in East Java Province with bivariate negative binomial regression.
2.1. Pearson Correlation Coefficient

Bivariate regression is regression with response variables as much as two and among response variables must have a correlation between one another. Correlation test is a method used to measure the strength or degree of relationship between two or more variables. Correlation coefficient is an indicator of a value in a linear relationship between two variables. The strength of the weak or the size of a correlation can be known by looking at the size of the coefficient [3]. The Pearson correlation coefficient is defined as follows:

\[
 r_{y_1y_2} = \frac{\sum_{i=1}^{n} (y_i - \bar{y}_1)(y_i - \bar{y}_2)}{\sqrt{\left(\sum_{i=1}^{n} (y_i - \bar{y}_1)^2\right)\left(\sum_{i=1}^{n} (y_i - \bar{y}_2)^2\right)}} 
\]  

(1)

In the correlation coefficient there are two relationships, namely positive and negative. These positive and negative values are due to the value of the correlation ranging from -1 to 1. If the value of the correlation is close to 1, both positive and negative it means that the two variables have a close relationship. Correlation value 0 indicates that the two variables do not have a close relationship. A positive correlation value indicates that the relationship is directly proportional to these 2 variables, while a negative correlation value indicates a relationship that is inversely proportional. Correlation testing between response variables is done with the following hypothesis:

H_0: There is no relationship between Y_1 and Y_2

H_1: There is a relationship between Y_1 and Y_2

with test statistics as follows:

\[
 t = t_{n-2} \frac{n-2}{\sqrt{1-r^2}} 
\]

(2)

The test statistic in equation (2) is distributed t with the degree of freedom is df = n – 2. The critical area is reject H_0 if the value is |t| > t_{\alpha/2}.

2.2. Multicollinearity

Multicollinearity is one of the problems in regression analysis, as well as missing data and outliers. Multicollinearity is a condition where predictor variables are highly correlated. According to [5], one way to identify the existence of multicollinearity cases is by looking at the value of Variance Inflating Factor (VIF) which is more than 10. The VIF value shows how the variance of the estimated parameter results increases due to multicollinearity. The VIF value is formulated by the following equation:

\[
 \text{VIF} = \frac{1}{1-R_j^2} 
\]

(3)

With \( R_j^2 \) is the coefficient of determination between \( x_j \) and other predictor variables.

Multicollinearity problems can be overcome in several ways, including by issuing predictor variables that are highly correlated, transforming data, adding data, using ridge regression or can also use Principal Component Analysis (PCA).

2.3. Over dispersion

Over dispersion in Poisson regression occurs when the data variance value is greater than the average value. If in Poisson regression there are cases of over dispersion, but still using Poisson regression it will cause the estimation of the regression coefficient parameter to remain consistent but not efficient. This will have an impact on the standard error value that becomes underestimate, so that it is invalid. Over dispersion is Pearson chi-square dispersion value and deviance divided by its degree of freedom, obtained values greater than 1. Over dispersion can be written as follows \( \text{Var}(Y) > \text{E}(Y) \)
If over Poisson regression occurs over dispersion, then one alternative that can be used is negative binomial regression [6].

2.4. Bivariate Negative Binomial Distribution

If \( Y_{1i} \) and \( Y_{2i} \) are random variables that are Poisson distributed with the mean \( \mu_{1i} \) and \( \mu_{2i} \), where \( i=1,2,...,n \) is a random variable that follows the Gamma \( (\tau^{-1}, \tau^{-1}) \) distribution. The shared distribution of \( Y_{1i} \) and \( Y_{2i} \) is as follows:

\[
f(y_{1i}, y_{2i}) = \frac{\Gamma(\tau^{-1} + y_{1i} + y_{2i})}{\Gamma(\tau^{-1})\Gamma(y_{1i} + 1)\Gamma(y_{2i} + 1)} \mu_{1i}^{y_{1i}}\mu_{2i}^{y_{2i}}e^{-\tau^{-1}(\tau^{-1} + y_{1i} + y_{2i})} \]

(4)

with is the dispersion parameter [2].

2.5. Bivariate Negative Binomial Regression

According to [4] the bivariate negative binomial regression model (BNB) is stated as follows:

\[
(Y_{1i}, Y_{2i}) \sim \text{BNB}(\mu_{1i}, \mu_{2i}, \tau)
\]

With:

\[
\hat{\mu}_{ji} = e^{X_j\hat{\beta}_j}; j=1,2; i=1,2,...,n
\]

(5)

\( X_i = [1 \ x_{1i} \ x_{2i} \cdots x_{pl}] \) is the observation vector of the predictor variable

\( \beta_j = [\beta_{0j} \ \beta_{1j} \ \beta_{2j} \cdots \beta_{pj}] \) is a vector of parameters of the BNB regression model

\( \mu_{ji} = E(Y_j | X_i) \) is the average value of the \( Y_j \) response variable if the \( X_i \) predictor variable is known.

2.6. Estimation of Bivariate Negative Binomial Regression Parameters

According to [8] to estimate the parameters of the BNB regression model, the Maximum Likelihood Estimator (MLE) method is used. Suppose that \( \delta = (\beta_1, \beta_2, \tau) \), then the likelihood function is obtained as follows:

\[
L(\delta) = \prod_{i=1}^{n} \frac{\Gamma(\tau^{-1} + y_{1i} + y_{2i})}{\Gamma(\tau^{-1})\Gamma(y_{1i} + 1)\Gamma(y_{2i} + 1)} \mu_{1i}^{y_{1i}}\mu_{2i}^{y_{2i}}e^{-\tau^{-1}(\tau^{-1} + y_{1i} + y_{2i})} \]

(6)

from the likelihood function the log-likelihood function is obtained as follows:

\[
\ell(\delta) = \sum_{i=1}^{n} \left[\sum_{k=1}^{y_{1i}+y_{2i}} \ln\left(y_{1i} + y_{2i} + \tau^{-1} - k\right) + y_{1i} \ln \mu_{1i} + y_{2i} \ln \mu_{2i} - \frac{\ln \tau}{\tau} \right]
\]

\[
- \left(\tau^{-1} + y_{1i} + y_{2i}\right) \ln\left(\tau^{-1} + y_{1i} + y_{2i}\right) - \ln\left(y_{1i}!\right) - \ln\left(y_{2i}!\right)
\]

(7)

2.7. Model Goodness Criteria

Akaike Information Criterion (AIC) is a model conformity criterion in estimating a model statistically. In the formation of regression models, it is necessary to have the AIC criteria with the aim of getting the factors that influence the model. The AIC value is formulated as follows:

\[
\text{AIC} = -2\ln L(\hat{\delta}) + 2k
\]

(8)

with is the number of parameters used, and is the maximum value of likelihood.
2.8. Algorithm for modelling the number of diarrheal sufferers and the number of certified drink water providers which pass health requirements

Stages of algorithm as follows:

1. Describe the number of diarrheal sufferers and drinking water providers who meet health requirements in East Java Province and the factors that are thought to be influential.
2. Modelling the number of diarrheal sufferers and drinking water providers who meet health requirements in each district / city in East Java Province with bivariate negative binomial regression.
3. Analysing and interpreting the results of estimation of the number of diarrheal sufferers and drinking water providers who meet health requirements in East Java Province.

3. Result and Discussion

Based on the results of data testing using bivariate negative binomial regression method with software R and software Minitab, the following results were obtained:

3.1. Descriptive Statistics

Descriptive statistics are used to find out the description of the percentage of diarrheal sufferers and the number of drinking water providers who meet health requirements in East Java Province in 2016 are presented in Table 2 below:

|       | N  | Minimum | Maximum | Mean  | Variance |
|-------|----|---------|---------|-------|----------|
| y1    | 38 | 1.614   | 75.562  | 26.316| 420.96   |
| y2    | 38 | 5       | 1199    | 143.63| 50750.2  |

Based on Table 2, the average percentage of diarrheal patients is 26.316% with a maximum value of 75.56% found in Sidoarjo Regency and a minimum value of 1.62% is found in the City of Kediri. The variance value is 420.96 which means that the percentage of diarrheal sufferers owned between cities / regencies in East Java Province has a high enough value. While the average number of drinking water providers that meet health requirements is 144 with a maximum value of 1199 in the City of Surabaya and a minimum value of 5 is in Sampang Regency. The variance value is 50750.2 which means that the number of drinking water providers that meet the health requirements owned between cities / regencies in East Java Province has a high enough value.

3.2. Modelling of Diarrheal Sufferers Percentage and Certified Drink Water Providers

3.2.1. Negative Binomial Test. Based on the results of the negative binomial distribution test, the critical value is 0.21544 with α of 5% so it can be concluded that the percentage of diarrheal patients in East Java has a negative binomial distribution. Besides that, the percentage data of diarrheal sufferers fulfil the assumption of over dispersion, that is the variance value is more than the mean, so the percentage data of diarrheal sufferers are negative binomial distribution. Based on Table 2 above, the data on the number of drinking water providers that meet the health requirements fulfil the assumption of over dispersion, namely the variance value is more than the mean, so the data is are negative binomial distribution.

3.2.2. Correlation Test Between Two Response Variables. Based on the test results, obtained Pearson correlation value of 0.346 with p-value = 0.034 <α = 0.05. Decisions that can be taken are rejecting H0 which means that there is a significant relationship between the percentage of diarrheal sufferers and the number of drinking water providers who meet health requirements.

3.2.3. Detection of Multicollinearity Cases of Predictor Variables. One way to find out about multicollinearity in the predictor variable used is to look at the VIF value in each predictor variable as presented in Table 3 below.
Based on Table 3 shows that none of the VIF values on each predictor variable exceeds the number 10. This means that between predictor variables do not have a correlation with each other.

3.2.4. Modeling of Diarrhea Sufferers Percentage and Certified Drink Water Providers using Bivariate Negative Binomial Regression. From the calculation results, the Likelihood Ratio Test value is 35225.53. This means that there is at least one variable that affects the resulting model. The next step is the partial parameter testing to determine the predictor variables that significantly influence the response variables presented in Table 4 below.

Based on Table 4 it can be seen that the variables that have a significant effect on the model of diarrheal sufferers are the percentage of the population with access to proper sanitation facilities ($X_1$), and the number of food management places is nurtured and tested for quotations ($X_4$) with the model obtained below:

$$\hat{\mu}_1 = \exp\left(7.1940 - 0.5846X_1 + 0.0100X_2 + 0.6948X_3 - 0.0009X_4\right)$$ \quad (9)

Based on Table 5 it can be seen that the variables that significantly influence the number of certified drink water providers are the percentage of the population with access to proper sanitation facilities ($X_1$), the percentage of public places that meet health requirements ($X_2$), the population literacy rate ($X_3$), and the number of food management places is nurtured and tested for quotations ($X_4$) with the model obtained below:

$$\hat{\mu}_2 = \exp\left(0.5474 + 3.6084X_1 - 2.3523X_2 + 0.0177X_3 + 0.7715X_4\right)$$ \quad (10)
Bivariate negative binomial regression model on the percentage of diarrheal patients and the number of certified drink water providers obtained AIC values using all significant predictor variables that is equal to 1563.407. Deviance value in bivariate negative binomial regression model is 15.1714.

3.3. Analysis and Interpretation of Estimated Results

Percentage of diarrheal sufferers and Certified Drink Water Providers

Based on Table 4, the estimated percentage of diarrheal sufferers can be interpreted that every increase of 1% of the population with access to proper sanitation facilities by assuming other predictor variables are constant, resulting in a decrease in the percentage of diarrheal patients by \( \exp(-0.5846) = 0.5573 \). Every increase of 1% of public places that meet health requirements by assuming another predictor variable is constant, resulting in an increase in the percentage of diarrheal patients by \( \exp(0.0100) = 1.01 \). Every 1% increase in literacy rate by assuming other predictor variables are constant, resulting in an increase in the percentage of diarrheal patients by \( \exp(0.6948) = 2.003 \). Every increase of 1 food management unit was nurtured and tested by assuming other predictor variables were constant, resulting in a decrease in the percentage of diarrheal patients \( \exp(-0.0009) = 0.999 \).

Based on Table 5 the estimation results of the number of drinking water providers that meet health requirements can be interpreted that every increase of 1% of the population with access to proper sanitation facilities by assuming other predictor variables are constant, resulting in an increase in the number of certified drink water providers \( \exp(3.6084) = 36.91 \). Every increase of 1% of public places that meet health requirements by assuming other predictor variables are constant, resulting in a decrease in the number of certified drink water providers \( \exp(-2.3523) = 0.095 \). Every 1% increase in literacy rate by assuming another predictor variable is constant, resulting in an increase in the number of certified drink water providers \( \exp(0.0177) = 1.018 \). Every increase of 1 unit of food management place was fostered and tested using the other predictor variables constant, resulting in an increase in the number of certified drink water providers \( \exp(0.7715) = 2.16 \).

The following will show a plot between the estimation results with observations on the first and second response variables in figure 1 and figure 2 where the circle shape states the observation and the line shape states the estimation results of the response variable.

Based on figure 1 it can be seen that the highest percentage of diarrheal patients in East Java is in Sidoarjo Regency and the lowest percentage of diarrheal sufferers in East Java is in the City of Kediri. Based on figure 2 it can be seen that the number of drinking water providers who meet the highest health requirements in East Java are in Surabaya City and the number of drinking water providers who meet the lowest health requirements in East Java are in Sampang Regency.
4. Conclusion

Percentage of diarrhoea patients have average 26.316% with a maximum value of 75.56% found in Sidoarjo Regency and a minimum value of 1.62% is found in the City of Kediri. The variance value is 420.96 which means that the percentage of diarrhoea sufferers owned between cities / regencies in East Java Province has a high enough value. While the average number of drinking water providers that meet health requirements is 144 with a maximum value of 1199 in the City of Surabaya and a minimum value of 5 is in Sampang Regency. The variance value is 50750.2 which means that the number of drinking water providers that meet the health requirements owned between cities / regencies in East Java Province has a high enough value.

Based on the modelling results, the percentage of diarrhoea sufferers and drinking water providers who meet health requirements in East Java Province using bivariate negative binomial regression obtained the estimation model as follows:

\[
\hat{\mu}_1 = \exp(7.1940 - 0.5846X_1 + 0.0100X_2 + 0.6948X_3 - 0.0009X_4) \\
\hat{\mu}_2 = \exp(0.5474 + 3.6084X_1 - 2.3523X_2 + 0.0177X_3 + 0.7715X_4)
\]

The results of the estimation of the percentage of diarrhoea sufferers with predictor variables that have a significant effect is the percentage of the population with access to adequate sanitation facilities, and the number of food management places are nurtured and tested. While the estimation of the number of drinking water providers that meet the health requirements with a predictor variable that has a significant effect is the percentage of the population with access to adequate sanitation facilities, the percentage of public places that meet health requirements, the literacy rate of the population, and the number of food management places fostered and quotes.

Based on the model obtained, it is obtained deviance values in the negative bivariate binomial regression model of 15.1714, and the AIC value of 1563.407.

Acknowledgment

The authors would like to thank Airlangga University for funding to this research in the academic year.

References

[1] Arumsari N and Sutikno. 2010. Modeling of Diarrhea Events with Spatial Regression Approach. Postgraduate X-ITS National Seminar. 979 545 0270 1
[2] Cheon S, Song S H and Jung B C. 2009. Test of Independence in a Bivariate Negative Binomial Models. Journal of the Korean Statistical Society. 38:185-190.
[3] Draper N R and Smith H. 1992. Applied Regression Analysis, Second Edition. New York: John Wiley and sons inc.
[4] Famoye F. 2010. On the Bivariate Negative Binomial Regression Model. Journal of Applied Statistics. 37(6):969-981.
[5] Gujarati D N. 2004. Basic Econometrics, Fourth Edition. USA: Department of Social Sciences, Naval Academy at West Point.
[6] Hilbe M J. 2011. Negative Binomial Regression, Second Edition. New York: Cambridge University Press.
[7] Liu Z, Zhan F, Li J, Ding G, Zhang C, Liu O and Jiang B. 2018. Science of the total environment. 626 630-637
[8] Park B J and Lord D. 2008. Adjusment for The Maximum Likelihood Estimate of The Negative Binomial Dispersion Parameter. USA: Texas University.