Count Regression Models for Analyzing Crime Rates in The East Java Province

Dian Handayani1,*, A F Artari1, W Safitri1, W Rahayu1, and V M Santi1

1Study Program of Statistics, Universitas Negeri Jakarta, Indonesia

*Email: dianh@unj.ac.id

Abstract. Crime rate is the number of reported crimes divided by total population. Several factors could contribute the variability of crime rates among areas. This study aims to model the relationship between crime rates among regencies and cities in the East Java Province (Indonesia) and some potentially explanatory variables based on Statistics Indonesia publication in 2020. The crime rate in the East Java Province was consistently at the top three after DKI Jakarta and North Sumatra during 2017 to 2019. Therefore, it is interesting for us to study further about the crime rate in the East Java. Our preliminary analysis indicates that there is an overdispersion in our sample data. To overcome the overdispersion, we fit Generalized Poisson and Negative Binomial regression. The ratio of deviance and degree of freedom based on Negative Binomial is slightly smaller (1.38) than Generalized Poisson (1.99). The results indicate that Negative Binomial and Generalized Poisson regression, compared to standard Poisson regression, are relatively fit to model our crime rate data. Some factors which contribute significantly ($\alpha=0.05$) for the crime rate in the East Java Province under Negative Binomial as well as Generalized Poisson regression are percentage of poor people, number of households, unemployment rate, and percentage of expenditure.

Keywords: Overdispersion, Negative Binomial, Generalized Poisson, Likelihood

1. Introduction

A standard statistical method that is commonly used to model the relationship between a count response and some explanatory variables is Poisson regression. This method assumes that the count response follows Poisson distribution [1],[2],[3]. There is only one parameter in Poisson distribution namely mean (or rates). Poisson distribution does not have a dispersion parameter. Theoretically, a random variable that follows Poisson distribution have variance which is the same as its mean. This characteristic is also called by an equidispersion.

In real life, the variance of a count response that is observed from sample data could be much greater than the sample mean. This phenomenon is called by overdispersion [4],[5],[6]. The other phenomena that also could be happened in analyzing count response is underdispersion (the sample variance is less than sample mean). However, in practice, overdispersion is more common to be met than underdispersion. There are several things that cause overdispersion phenomena, such as the heterogeneity among observations (units), the correlation between observations, the existence of outliers, the use of improper link function, the use of improper systematic components or the existence of excess zeroes [5],[6].

If Poisson regression model is fitted to the overdispersion data, it could produce the standard error of coefficient regression estimates which is less than expected by Poisson regression model. On the
other hand, if Poisson regression is fitted to underdispersion data, the standard error of the coefficient regression estimates would be greater than expected by Poisson regression model.

An alternative method to deal with overdispersion as well as underdispersion on count response is to specify a Generalized Poisson distribution instead of standard Poisson distribution [7,8,9]. The Generalized Poisson distribution is an extension of the standard Poisson distribution. It does not only have a rate (mean) parameter but also has a dispersion parameter. The Generalized Poisson distribution could have characteristic equidispersion, overdispersion, or underdispersion [7]. It means that the Generalized Poisson distribution is possible to be used for handling overdispersion as well as underdispersion data [10].

Negative binomial distribution is another alternative probability distribution which could also be specified for count response [11],[12],[13]. Like the Generalized Poisson, Negative binomial also has dispersion parameter. However, the variance of random variable that follows negative binomial distribution is always greater or equal to the mean. It is not possible less than the mean. As a result, it is not appropriate for fitting negative binomial distribution on underdispersion data.

The main purpose of this study is to assess goodness of fit of the Generalized Poisson regression and negative binomial regression for modelling the relationship between crime rates in East Java Province (Indonesia) and some potentially explanatory variables. During the period 2017-2019 criminality rates in East Java have always ranked in the top third in Indonesia. Therefore, it is important to identify several factors that affect the crime rate in the East Java Province.

This paper is organized under four sections. Section 2 describes the data and methods that have been used in this paper. Section 3 presents the results and discussion and Section 4 provide concluding remarks.

2. Data and methods

The data that have been analyzed in this study is based on publication of 2020 Crime Statistics Indonesia and 2020 The East Java Province in Figures. The observations are 29 regencies and 9 cities in the East Java Province. The response variable (Y), the explanatory variables \(X_1, X_2, \ldots, X_{11}\) as well as descriptive statistics are described in Table 1.

| Variable | Description | Mean   | Standard Deviation |
|----------|-------------|--------|--------------------|
| Y        | Crime counts | 535.2  | 518.99             |
| X_1      | Population density (per sq.km) | 1897.6 | 2200.32            |
| X_2      | The percentage of poor people (%) | 10.29  | 4.3                |
| X_3      | Number of household | 288307 | 1839               |
| X_4      | The open unemloyment rate (%)  | 3.75   | 1.11               |
| X_5      | The average length of school (year) | 7.74   | 1.58               |
| X_6      | Growth rate of gross regional domestic product (%) | 5.08   | 1.33               |
| X_7      | The regional minimum wages (IDR) | 2254   | 6998               |
| X_8      | The labor force participation rate > 15 years (%) | 69.48  | 3.50               |
| X_9      | Net wage/salary of informal employee (IDR) | 1625   | 4221               |
| X_10     | The percentage of expenditure per capita (%) | 50.26  | 5.77               |

2.1. Poisson regression

Poisson regression is a method that can be used for modelling the relationship between a count response variable Y and p explanatory variables \(X_1, X_2, \ldots, X_p\). The response variable Y is assumed to follow Poisson distribution which its probability function is given by:

\[
f(y_i; \mu_i) = \frac{e^{-\mu_i}(t_i\mu_i)^{y_i}}{y_i!}, \quad y_i = 0,1,2,\ldots; \quad i = 1,2,\ldots,n; \quad \mu_i > 0
\]
where \( \mu_i \) is a mean (rate) parameter, \( n \) is number of sample units and \( t_i \) is an index to represent that events occur over time, space or some other index of size. The expected value and the variance for count response \( Y \) under Poisson probability function (1) is given by:

\[
E(Y_i) = \text{Var}(Y_i) = t_i \mu_i = t_i \exp(x_i^T \beta)
\]

Poisson regression model is a special case of Generalized Linear Model (GLM) which uses log link function to relate the mean of count response (\( \mu \)) at index \( t \) with some explanatory variables. The Poisson regression model can be written in matrix notation as follows:

\[
\ln \left( \frac{\mu_i}{t_i} \right) = x_i^T \beta, \quad i = 1, 2 \ldots n
\]

Alternatively, Poisson regression model (3) also can be rewritten by:

\[
\ln(\mu_i) = \ln(t_i) + (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})
\]

where \( \ln t_i \) is also called by an offset.

### 2.2. Negative binomial regression

Negative Binomial regression model is a special case of the Generalized Linear Model (GLM) which the response variable (random component) follows negative binomial distribution. Negative binomial distribution is a generalization of Poisson distribution by including a gamma noise variable which has a mean of 1 and a scale parameter \( \delta \). The gamma mixture can accommodate overdispersion on Poisson counts. Negative binomial probability function which is the result form Poisson-Gamma mixture can be written as follows [7],[12],[13],[14]:

\[
f(y_i; \mu_i, \alpha) = \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(y_i + 1)\Gamma(\frac{1}{\alpha})} \left( \frac{1}{1 + \alpha \mu_i} \right)^{\frac{1}{\alpha}} \left( 1 - \frac{1}{1 + \alpha \mu_i} \right)^{y_i}; \quad y_i = 0, 1, 2, \ldots
\]  

The parameter \( \mu_i \) is the mean rate of \( y \) per unit of exposure. The parameter \( \alpha(>0) \) in (4) is a dispersion parameter which is defined by \( \alpha = \frac{1}{\delta} \). If \( \alpha \to 0 \), then the negative binomial probability function (4) approaches to Poisson probability function with parameter \( \mu_i \).

The mean of \( y_i \) which follows negative binomial distribution (4) is given by: \( E(y_i) = \mu_i \) and the variance is given by: \( V(y_i) = \mu_i(1 + \alpha \mu_i) \). If there are some explanatory variables \( X_1, X_2, \ldots, X_p \) and exposure \( t_i \), the mean of \( y_i \) will depend on the explanatory variables and the exposure. It is given by:

\[
E(y_i|x_i) = t_i \mu_i = t_i \exp(x_i^T \beta)
\]

where \( x_i^T = (x_{i1}, x_{i2}, \ldots, x_{ip}) \). If logarithm transformation is taken in formula (6), it will be obtained:

\[
\ln \mu_i = \log t_i + \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}
\]

where \( \log t_i \) is also called by an offset. Model (7) is a negative binomial regression model.

### 2.3. Generalized poisson regression

Generalized Poisson regression is a regression analysis in which the count response is assumed to follow Generalized Poisson distribution. The Generalized Poisson distribution is a generalization of Poisson distribution. The random variable \( Y \) which follows generalized poisson distribution with mean parameter of \( \mu \) and dispersion parameter \( \theta \) can be written as follows [8],[9],[10],[15]:
\[ f(y_i; \mu_i; \theta) = \left( \frac{\mu_i}{1 + \theta \mu_i} \right)^{y_i} \frac{\exp\left( -\frac{\mu_i}{1 + \theta \mu_i} \right)}{y_i!}, \quad y_i > 0 \]  

The mean of \( y_i \) which follows negative binomial distribution (8) is given by: \( E(y_i) = \mu_i \) and the variance is given by: \( V(y_i) = \mu_i(1 + \theta \mu_i) \).

In generalized poisson probability function (8), if \( \theta = 0 \), then the function will be reduced to a Poisson probability function. If \( \theta > 0 \), the Generalized Poisson probability function will have variance which is larger than mean. On the other hand, if \( \theta < 0 \), the Generalized Poisson probability function will have variance which is smaller than variance. Because of the possible values of dispersion parameter \( \theta \), Generalized Poisson distribution could have characteristic equidispersion, overdispersion, and underdispersion.

If there are \( p \) explanatory variables \( X_1, X_2, \ldots, X_p \) and exposure \( t_i \), the conditional mean of \( y_i \) is given by

\[ E(y_i | x) = t_i \mu_i = t_i \exp(\mathbf{x}_i^T \beta) \]  

where \( \mathbf{x}_i^T = (x_{1i}, x_{2i}, \ldots, x_{pi})^T \) and \( \beta \) is a \((p + 1)\) vector of regression parameters (Note: there will be \((p + 1)\) vector of regression parameter by including the intercept \( \beta_0 \)).

If logarithm transformation is taken in formula (9), it will be obtained:

\[ \ln \mu_i = \ln t_i + \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} \]  

where \( \ln t_i \) is also called by an offset. Model (10) is a negative binomial regression model.

3. Results and discussion

Figure 1 shows a map of the distribution of crime counts per district/city in the East Java Province. The red colour means that more criminality has occur in the district (or city). On the other hand, the yellow colour means that less criminality has occur in the district or city.

Crime counts among districts or cities in the East Java Province is very heterogenous. This is due to the diversity of characteristics of each district (or city) in the East Java Province. The highest crime counts is 2,404 cases which was occurred in Jember regency whereas the lowest crime counts is 73 cases in Pacitan Regency.

| Variable | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 | X9 | X10 |
|----------|----|----|----|----|----|----|----|----|----|-----|
| VIF      | 5.43 | 5.25 | 2.80 | 3.04 | 4.66 | 3.49 | 6.52 | 1.36 | 11.16 | 6.67 |

The values of the explanatory variables which are considered in our research are measured in different scales. As a result, before conducting further analysis, it is necessary to standardize the values.
Furthermore, it is also important to check the multicollinearity among the explanatory variables. The Variance Inflation Factors (VIF) is a tool which can be used for checking the multicollinearity. If the value of VIF on a certain explanatory variable is greater than 10, it indicates that there is a strong correlation between the variable with other explanatory variables. The value of VIF for each explanatory variable which are considered in our research are shown in Table 2.

It can be seen from Table 2 that there is only $X_9$ (Net wage/salary of informal employee) which has VIF > 10. It means that there is an existence of multicollinearity between $X_9$ and other explanatory variables. As a result, $X_9$ is dropped or it will not be involved for subsequent analysis.

### Table 3. The model estimates and standard error

| Parameter | Standard Poisson | Negative Binomial | Generalized Poisson |
|-----------|------------------|-------------------|--------------------|
| $\beta_0$ | -7.62            | 6.04              | -7.44              |
|           | $(8.6 \times 10^{-3})$ | $(5.32 \times 10^{-2})$ | $(9.45 \times 10^{-2})$ |
| $\beta_1$ | 0.26             | 0.11              | 0.52               |
|           | $(1.19 \times 10^{-2})$ | $(9.48 \times 10^{-2})$ | $(1.36 \times 10^{-1})$ |
| $\beta_2$ | -0.39            | -0.39             | -0.68              |
|           | $(1.72 \times 10^{-2})$ | $(1.19 \times 10^{-1})$ | $(1.11 \times 10^{-1})$ |
| $\beta_3$ | 0.03             | 0.46              | -0.39              |
|           | $(8 \times 10^{-3})$ | $(7.39 \times 10^{-2})$ | $(1.27 \times 10^{-1})$ |
| $\beta_4$ | 0.08             | 0.16              | 0.28               |
|           | $(1.23 \times 10^{-2})$ | $(7.39 \times 10^{-2})$ | $(5.08 \times 10^{-2})$ |
| $\beta_5$ | 0.08             | -0.09             | -0.39              |
|           | $(1.52 \times 10^{-2})$ | $(1.15 \times 10^{-1})$ | $(1.39 \times 10^{-1})$ |
| $\beta_6$ | -0.04            | -0.05             | -0.12              |
|           | $(1.58 \times 10^{-2})$ | $(9.27 \times 10^{-2})$ | $(9.19 \times 10^{-2})$ |
| $\beta_7$ | -0.19            | 0.03              | 0.32               |
|           | $(1.4 \times 10^{-2})$ | $(7.82 \times 10^{-2})$ | $(9.52 \times 10^{-2})$ |
| $\beta_8$ | -0.28            | -0.18             | -0.03              |
|           | $(1.1 \times 10^{-2})$ | $(6.87 \times 10^{-2})$ | $(5.93 \times 10^{-2})$ |
| $\beta_{10}$ | 0.31             | 0.25              | 0.27               |
|           | $(1.47 \times 10^{-2})$ | $(1.02 \times 10^{-1})$ | $(1.69 \times 10^{-3})$ |

In the preliminary analysis, Poisson regression model is fitted to our sample data because it is the benchmark for modelling the relationship between count response and some explanatory variables. However, it is already known that equidispersion characteristic in Poisson distribution is often violated in real life. Therefore, it is necessary to check whether the equidispersion is satisfied. By fitting Poisson regression model to our data, it can be seen in Table 4, the ratio between residual deviance to its degree of freedom is $37928.28/28 = 1354.581$. This value is much larger than one which is expected if the Poisson regression model were fit to our sample data. Furthermore, based on the descriptive measures which is presented in Table 1, it can be seen that the sample mean is 535.2 and the variance is $(518.99)^2 = 269350.6$. The sample variance is much larger than the mean. It also could indicate that there is an overdispersion in our data. To overcome the overdispersion, we specify the generalized poisson regression and negative binomial regression. The results of fitting poisson regression, generalized poisson regression and negative binomial regression are shown in Table 3.

It can be seen from Table 3 that generalized poisson regression and negative binomial regression have a larger standard error than Poisson regression. Although the difference of standard error between negative binomial and Generalized Poisson is not too large, it can be seen from Table 3 that both of them have a larger standard error than Poisson regression. The results indicate that Generalized Poisson and negative binomial regression tend to be more capable for handling the overdispersion on our crime rates data.
Based on the AIC value criteria, it can be seen from Table 4, the AIC of Negative Binomial regression is the smallest. However, the AIC of negative binomial is quite similar with the Generalized Poisson regression model. Both of them is much smaller than standard poisson regression model. The ratio between residual deviance to degree of freedom for binomial negative is 38.47/28 = 1.37 whereas the ratio between residual deviance to degree of freedom for generalized Poisson regression is 55.76/28 = 1.99. Based on the AIC and the ratio between residual deviance to degree of freedom, Negative Binomial regression model is chosen for analysing the relationship between crime rates and some explanatory variables in the East Java Province.

Based on the results in Table 3, the fitted negative binomial regression model is given by:

\[
\ln(\mu_i) = \ln(t_i) + 6.04 + 0.11x_{i1} - 0.39x_{i2} + 0.46x_{i3} + 0.16x_{i4} - 0.09x_{i5} - 0.05x_{i6} + 0.03x_{i7} - 0.18x_{i8} + 0.25x_{i10} \tag{11}
\]

where \(\mu_i\) denotes the expected count of crime counts and \(t_i\) denotes the population in regency or city \(i\). The interpretation for coefficient estimates in (11) is a one-unit increase on \(x_j\) \(j = 1, 2, \ldots, p\) will result in a multiplicative effect of \(e^{\beta_j}\) on the rate \(\mu_i\). The fitted values for crime counts in regency or city \(i\) are given by:

\[
\hat{y}_i = \hat{\mu}_i = t_i \exp(6.04 + 0.11x_{i1} - 0.39x_{i2} + 0.46x_{i3} + 0.16x_{i4} - 0.09x_{i5} - 0.05x_{i6} + 0.03x_{i7} - 0.18x_{i8} + 0.25x_{i10}) \tag{12}
\]

The population density (\(X_2\)), number of household (\(X_3\)), the open unemployment rate (\(X_4\)) the regional minimum wages (\(X_7\)), and the percentage of expenditure per capita (\(X_{10}\)) have a positive sign. A positive change in these factors induces an increase in the crime rates. Meanwhile, the percentage of poor people (\(X_2\)), the average length of school (\(X_3\)), growth rate of gross regional domestic product (\(X_6\)), and the labour force participation rate which is greater than 15 years (\(X_9\)) have a negative sign. A positive change in these factors induces a decrease in the crime rates.

The variable (\(X_1\)) has a coefficient estimates \(\beta_1 = 0.11\) which means that for each one-unit increase in population density (\(X_1\)) in \(i^{th}\) regency or city in the East Java Province, the expected log count of crime counts in the regency (or city) increases by 0.11. Alternatively, each one-unit increase on \(X_1\), the variable \(X_2\) has a multiplicative effect of \(e^{0.11} = 1.12\) times on the mean of \(Y\). The interpretation for other coefficients with positive sign is similar to the interpretation for the coefficient estimates \(\beta_1\).

The variable \(X_8\) has a coefficient estimates \(\beta_8 = -0.18\) which means that for each one-unit increase in labor force participation rate > 15 years (\(X_8\)) in \(i^{th}\) regency or city in the East Java Province, the expected log count of crime total in the regency (or city) decreases by 0.18. Alternatively, each one-unit increase on \(X_8\) will result in a multiplicative effect of \(e^{-0.18} = 0.83\) times given other variables are fixed. The interpretation for other coefficients with negative sign is similar to the interpretation for the coefficient estimates \(\beta_8\).

Based on hypothesis testing about significance for each explanatory variable, there are five explanatory variables that have a significant effect on crime counts (\(\alpha = 0.05\)). These variables are the percentage of poor people (\(X_2\)), number of households (\(X_3\)), the open unemployment rate (\(X_4\)), net wage/salary of informal employee (\(X_8\)), and the percentage of expenditure per capita (\(X_{10}\)).
4. Concluding remarks

Based on our sample data, fitting Poisson regression model for analyzing the relationship between crime rates and some explanatory variables in the East Java Province is not suitable because there is an indication of overdispersion. Generalized Poisson regression and Negative Binomial regression model which are known capable to overcome overdispersion are fitted to the sample data. The results indicate that the Akaike Information Criteria (AIC) of Negative Binomial regression is slightly smaller (501.09) than Generalized Poisson (502.29). It means that the Negative Binomial regression model as good as Generalized Poisson regression for modelling data of crime rates in East Java Province than Generalized Poisson regression.

By fitting Negative Binomial regression model to our data, it can be identified that the explanatory variables that have a significant effect ($\alpha = 0.05$) to the crime rates are the percentage of poor people ($X_2$), number of household ($X_3$), the open unemployment rate ($X_4$), net wage/salary of informal employee ($X_8$), and the percentage of expenditure per capita ($X_{10}$).

The ratio of residual deviance to the degree of freedom for Negative Binomial regression model is still greater than one (1.37). It suggests that it is necessary to find other methods which is more suitable.

There are several other alternative models to overcome the problem of overdispersion, including Quasi-Poisson, Zero-Inflated Poisson (ZIP), etc.

Acknowledgement

This research was supported by a grant from the Public Service Agency (Badan Layanan Umum) of Universitas Negeri Jakarta.

References

[1] Cameron, A. C., and Trivedi, P. K. (1998). Regression Analysis of Count Data. Cambridge University Press.
[2] Agresti, A. (2019). An Introduction to Categorical Data Analysis. John Wiley & Sons. Edition 3rd.
[3] Dobson, A. J. and Barnett, A.G. (2018). An Introduction to Generalized Linear Models. Edition 4th. CRC Press.
[4] Yang, Z., Hardin, J.W., Addy, C.L., Vuong, Q.H. (2007). Testing Approaches for Overdispersion in Poisson Regression versus the Generalized Poisson Model. Biometrical Journal, Vol.49, No. 4, 565-584.
[5] Hinde, J., & Dem´etrio, C. G. B. (1998). Overdispersion: Models and Estimation. UK: Computational statistics & data analysis.
[6] Handayani, D., Kurnia, A., and Sadik, K. (2016). Pendugaan Parameter Overdispersi Dalam Pengepasan Model Pada Data Dengan Respon Banyak Nol (Sparse Data). http://matematika.fmipa.unp.ac.id/semastat16/download/Prosiding%20SEMSTAT%202016.pdf
[7] Hilbe, J. M. (2011). Negative Binomial Regression (Second). New York: Cambridge University Press.
[8] Wang, W. and Famoye, F. (1997). Modeling household Fertility Decisions with Generalized Poisson Regression. Journal of Population Economics, Vol.10, 273-283.
[9] Consul, P.C., and Famoye, F. (1992). Generalized Poisson Regression Model. Communications in Statistics-Theory and Methods, Vol. 22, No.5, 1335-1354.
[10] Famoye, F. (1993). Restricted Generalized Poisson Regression Model. Communications in Statistics-Theory and Methods, Vol. 22, No.5,1335-1354.
[11] Sellers, K.F., Shmueli, G. 2010. Flexible Regression Model For Count Data. The Annals of Applied Statistics, Vol.4, No.2, 943-961.
[12] Maxwell, O. Mayowa, B.A., Chinedu, I.U., Peace, A.F. 2018. Modelling Count Data: A Generalized Linear Model Framework. American Journal of Mathematics and Statistics, Vol. 8, No. 6, 179-183.
[13] Coxe, S., West, S.G., Aiken, L.S. 2009. The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *Journal of Personality Assessment*, Vol.91, No.2, 121–136.

[14] Ismail N & Jemain A.A. 2007. Handling Overdispersion with Binomial Negatif and Generalized Poisson Regression Models. Casualty Actuarial Society Forum: 103-158.

[15] Muchika, I., Ngunyi, A. and Mageto, T. 2020. Modeling Burglar Incidents Data Using Generalized and Quasi Poisson Regression Models: A Case Study of Nairobi City County, Kenya. *American Journal of Theoretical and Applied Statistics*, Vol.9, No.5, 256-262.