ALTERNATIVE MODELS IN OVERCOMING THE PROBLEM OF OVERDISPERSION IN POISSON REGRESSION

1*Pardomuan Robinson Sihombing, 1Ade Marsinta Arsani, 2Ni Komang Semara Yanti, 2Putu Pande Wahyu Diatmika

1 BPS-Statistics Indonesia
2 Udayana University

*Corresponding Author email: robinson@bps.go.id

Abstract
This study aims to compare various alternative models in overcoming the problem of overdispersion in Poisson regression modeling. The comparative modeling is the Generalized Poisson model, Negative Binomial, and Generalized Negative Binomial. Modeling is applied to modeling the number of poor people in Central Java in 2021 with unemployment, HDI, and GRDP as independent variables. The results obtained by Generalized Poisson are better than Negative Binomial and Generalized Negative Binomial because of the smaller AIC and BIC values and the larger R². For simultaneous tests, it can be concluded that unemployment, HDI, and GRDP significantly affect the number of poor people. Only unemployment and HDI variables partially affect the number of poor people in Central Java. On the other hand, there is not enough evidence that GRDP affects some poor people. There is a need for comprehensive and relevant policies to overcome the number of poor people in an area.

INTRODUCTION

Poverty is still one of the problems in developing a region or country. Even the first goal of the SDGs is to minimize poverty to zero poverty. Many studies related to factors that affect poverty along with other macro variables. Kurniawan and Utami (2021) discuss the impact of unemployment on the poverty rate in Indonesia. The study shows the high unemployment rate contributes to the country's poverty rate. Furthermore, Nugroho & Wulandari (2022) the relationship between unemployment and poverty in Central Java, Indonesia. The results showed that an increase in the unemployment rate significantly increased the poverty rate in the region.

Fitriyani & Sutrisno (2021) analyzed the effect of the Human Development Index on Poverty in Central Java Province. Results show that HDI lowers poverty rates. The same results were obtained by Nasution & Prihartono (2022) in analyzing the Human Development Index's effect on Indonesia's Poverty Rate. Wahyudi (2021) examined the Effect of GRDP, Inflation, and Minimum Wage on Poverty in East Java Province. The results show that there is an influence between GDP and poverty. Where GDP is a region's output or economic strength that can drive the economy. The same thing was obtained by Adiwijaya and Subawa (2022).

In modeling the influence between variables where the dependent variable used is enumerated data, for example, in the case of the number of poor people whose value is a positive integer, the appropriate modeling is Poisson regression (Agresti, 2018). In the Poisson distribution, the values of expectation (mean) and variety are assumed to be equal (equidispersion), namely E[Y]=var[Y]=μ. However, often in its application, there is a violation of these assumptions when the variable value is more significant on average (overdispersion) or when the variable value is smaller than the average (underdispersion). Some of the things that cause overdispersion problems are that there are unobserved sources of diversity (unobserved heterogeneity), missing observations in variable X, outliers in the data so that interaction is needed in the Model, the need to transform varying X or incorrect specifications of connecting functions.

Some alternatives that can be used in overcoming overdispersion in Poisson modeling include Generalized Poisson (Putra et al., 2013), Negative Binomial (, and Generalized Negative Binomial (Consul & Gupta, 1980). (Consul & Gupta, 1980) A Negative Binomial regression model for a nonnegative count dependent variable. In this Model, the count variable is assumed to be equal (equidispersion), namely E[Y]=var[Y]=μ. However, often in its application, there is a violation of these assumptions when the variable value is more significant on average (overdispersion) or when the variable value is smaller than the average (underdispersion). Some of the things that cause overdispersion problems are that there are unobserved sources of diversity (unobserved heterogeneity), missing observations in variable X, outliers in the data so that interaction is needed in the Model, the need to transform varying X or incorrect specifications of connecting functions.

Based on the above problems, Poisson regression modeling often occurs with overdispersion assumptions. Therefore, this study wants to compare various alternatives to solving overdispersion problems using Generalized Poisson, negative binomial, and Generalized Negative Binomial modeling.
was implemented in poverty modeling in Central Java Province in 2021.

**RESEARCH METHODS**

The data used in this study came from the publication of the Statistics Indonesia of Central Java. This research focuses on all regencies in Central Java with a research period 2021. The method used is quantitative analysis to determine the influence between variables (Sugiyono, 2019). The dependent and independent variables in this study can be seen in Table 1.

<table>
<thead>
<tr>
<th>Dependent Variables</th>
<th>Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Poor People</td>
<td>People</td>
</tr>
<tr>
<td>Independent Variables</td>
<td>Unit</td>
</tr>
<tr>
<td>Number of Unemployment</td>
<td>People</td>
</tr>
<tr>
<td>Human Development Index (HDI)</td>
<td>Points</td>
</tr>
<tr>
<td>Gross Regional Domestic Product (GRDP)</td>
<td>Million</td>
</tr>
</tbody>
</table>

**Poisson Distribution**

Poisson distribution is a distribution for events whose probability of occurrence is small, depending on a specific time interval or in a particular area, with observations as discrete variables and between variables mutually complementary. According to Cameron and Trived (2013), a variable Y whose data type is discrete will follow the Poisson distribution if μ it is the average of an event per unit time and I is a certain period of time then the average from y becomes . μiThe mass function of the probability of Poisson spread is given in the following equation:

\[ P(Y_i; \mu_i) = \frac{e^{-\mu_i} \mu_i^{Y_i}}{Y_i!}, \quad Y_i = 0, 1, 2, \ldots, \infty \]  

(1)

This equation calculates the probability of a random variable Y with the mean and variance of Poisson distribution equal or \( E[Y_i] = \mu_{Y_i} = \text{var}[Y_i] \)

**Poisson Regression**

Poisson regression is a regression model that can be used on data with the variable Y spread following the Poisson spread. The Poisson regression model is written as follows:

\[ Y_i = \mu_i + \epsilon_i = \exp(x_i^T \beta + \epsilon_i) = \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_n x_{ni} + \epsilon_i), \quad i = 1, 2, \ldots, n \]  

(2)

where is the number of events and the mean number of events assumed not to change from data to data.

**Dispersion and Overdispersion Parameters**

The dispersion parameter (\( \phi \)) obtained from the formula \( \phi = \frac{\text{ni} \text{al deviance}}{df} \) Where df is the free degree, according to Rashwan and Kamel (2011), the deviance value is defined as:

\[ G^2 = 2 \sum_{i=1}^n y_i \ln (\frac{\gamma_i}{\hat{\gamma}_i}) \]  

(3)

If the value \( \phi > 0 \), then overdispersion occurs, and vice versa. If \( \phi < 0 \), then underdispersion happens. In Poisson regression, there is an assumption of equidispersion, but there are often violations, such as cases of overdispersion or underdispersion. Overdispersion or underdispersion is detected using the Goodness of fit statistic by Pearson divided by the degree of freedom.

**Generalized Poisson Regression**

Generalized Poisson regression uses an extension of the Poisson function where the variance value is not equal to the mean. The parts of the Generalized Poisson are:

\[ P(Y_i; \mu_i) = \left( \frac{\mu_i}{1+k\mu_i} \right)^{\gamma_i} \frac{(1+ky_i)^{\gamma_i-1}}{\gamma_i!}, \quad \gamma_i = 1 \]  

(4)

with \( E[Y_i] = \mu_{Y_i}; \text{var}[Y_i] = \mu_{Y_i}(1 + k\mu_i)^2 \) and \( k \) dispersion parameters (Putra et al., 2013)

**Negative Binomial Regression**

Suppose a \( Y_i \) Negative Binomial distribution will have the following probability function: \( Y_i \)

\[ P(Y_i = y_i | \mu_i) = \frac{\Gamma(y_i + \alpha - 1)}{(\Gamma(\alpha)(\gamma_i)^{(\gamma_i)})(\mu_i)^{(\gamma_i-1)}}(1 + \frac{\mu_i}{\gamma_i})^\alpha \]  

(5)

Where, gamma is function and \( \alpha (\alpha > 0) \) is an overdispersion parameter. For \( \alpha \) close to zero, the binomial model will approach the Poisson model. Unlike Poisson regression which assumes the mean and variance are equal, in a Negative Binomial regression, the conditional variety will exceed the conditional mean \( \Gamma(.)V(y_i | \mu_i) = \alpha + a\mu_i+E \). \( (y_i | \mu_i = \mu_i) \)

**Model Selection Criteria**

In this study, the model selection was based on error criteria used, including AIC (Akaike, 1974) and BIC (Gideon Schwarz, 1978) and coefficient of determination (R²). The best Model is the Model that has the smallest AIC and BIC values (Widarjono, 2007) and the most significant coefficient of determination (Gujarati, 2004). The formula used is:

\[ AIC = -2L(\hat{\theta}) + 2p \]  

(6)

\[ BIC = -2L(\hat{\theta}) + p \ln(n) \]  

(7)

\[ R^2 = \frac{SSR}{SST} = \frac{\sum_{i=1}^n(y_i - \bar{Y})^2}{\sum_{i=1}^n(y_i - \bar{Y})^2} \]  

(8)

Where \( L(\hat{\theta}) \) is the likelihood value, \( p \) is the number of parameters to be estimated, including constants, and \( n \) is the number of samples. Value \( \bar{Y} \) is the predicted value of
the Model’s dependent variable, and \( Y \) is the observation value of the dependent variable.

**RESULTS AND DISCUSSION**

Before further discussing Poisson Model, a descriptive analysis was carried out on the research variables. Table 1 presents descriptive statistics of each variable. On average, the number of poor people in Central Java is 117422, with the lowest score of 9440 people in Magelang City and the highest of 314950 people in Brebes Regency. The average number of unemployed people in Central Java Province is 32235 people, with the lowest value of 5769 people in Magelang City and the highest of 98718 people in Semarang City. The average HDI of Central Java Province is 72.85 points, with the lowest value of 66.32 points in the Brebes Regency and the highest of 83.60 in the Salatiga Regency. On average, Central Java's GRDP is 28.47 trillion rupiahs, with the lowest value of 6.51 trillion rupiahs in Magelang City and the highest of 144.7 trillion rupiahs in Semarang City.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Standard Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poverty</td>
<td>117422</td>
<td>65987</td>
<td>9440</td>
<td>314950</td>
</tr>
<tr>
<td>Unemployment</td>
<td>32235</td>
<td>22115</td>
<td>5769</td>
<td>98718</td>
</tr>
<tr>
<td>HDI</td>
<td>72.85</td>
<td>4.45</td>
<td>66.32</td>
<td>83.60</td>
</tr>
<tr>
<td>GRDP</td>
<td>28470431</td>
<td>26068461</td>
<td>6513895</td>
<td>144704572</td>
</tr>
</tbody>
</table>

The regression model requires no high relationship/multicollinearity between independent variables, as seen from a Variant Inflation Factor (VIF) value of less than 10. In Table 2, all independent variables had a VIF value of less than ten in this study. This result means all the independent variables used in the Model.

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIFs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployment</td>
<td>2.85</td>
</tr>
<tr>
<td>HDI</td>
<td>1.59</td>
</tr>
<tr>
<td>GRDP</td>
<td>3.14</td>
</tr>
</tbody>
</table>

Following the author's process model Poisson regression, the author performs a dispersion of the model testing before interpreting Poisson regression results. In Table 4. Indicates the presence of overdispersion in the resulting Poisson model. This result is identified by a probability value of 0.000< \( \alpha = 0.05 \). Therefore, the Poisson model becomes inefficient in estimating model parameters. Alternatives used in overdispersing Poisson modeling are Generalized Poisson, negative binomial, and Generalized Negative Binomial models.

<table>
<thead>
<tr>
<th>Test</th>
<th>Chi-Stat</th>
<th>Prob.Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deviance</td>
<td>455.8747</td>
<td>0.0000</td>
</tr>
<tr>
<td>Pearson</td>
<td>435.8716</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Furthermore, the best Model was selected by comparing each method's criteria: the error criteria (AIC and BIC) and the coefficient of determination R². When viewed from the most R² index values in Table 5, the best method is Generalized Poisson. Meanwhile, if assessed from the AIC and BIC values, the Generalized Poisson method has smaller values, namely 360.28 and 368.05. So, it can be said that the Generalized Poisson model is better than the negative binomial and Generalized Negative Binomial because of the smaller AIC and BIC values and the larger R². So, for this case, Generalized Poisson is the best method for overcoming overdispersion in the poison model.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Poisson</th>
<th>Generalized Poisson (GP)</th>
<th>Negative Binomial (NB)</th>
<th>Generalized Negative Binomial (GNB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unemployment</td>
<td>.00001063***</td>
<td>8.106e-06*</td>
<td>.00001299*</td>
<td>.00001299*</td>
</tr>
<tr>
<td>HDI</td>
<td>-.08260087***</td>
<td>-.10032038***</td>
<td>-.09931624***</td>
<td>-.09931624***</td>
</tr>
<tr>
<td>GRDP</td>
<td>8.276e-10</td>
<td>4.160e-09</td>
<td>2.395e-09</td>
<td>2.395e-09</td>
</tr>
<tr>
<td>Const</td>
<td>10.304475***</td>
<td>11.563657***</td>
<td>11.378931***</td>
<td>11.378931***</td>
</tr>
</tbody>
</table>
The results of this study are in line with the research of Abdullah (2021) that concluded that despite the increase in GDP in African countries, non-inclusive policies and lack of access to education and employment opportunities hinder improvements in poverty rates.

**CONCLUSION AND SUGGESTION**

Modeling the effect of unemployment, HDI, and GRDP on the number of poor people in Central Java using Poisson regression experienced a violation of the assumption of overdispersion. Therefore, the Poisson model becomes inefficient in estimating model parameters. The Alternatives used in overdispersion Poisson modeling are Generalized Poisson is better than Negative Binomial and Generalized Negative Binomial because of the smaller AIC and BIC values and the larger R2. For simultaneous tests, it can be concluded that unemployment, HDI, and GRDP significantly affect the number of poor people. Only unemployment and HDI variables partially affect the number of poor people in Central Java. On the other hand, there is not enough evidence that GRDP affects the number of poor people.

Further research adds variables that can potentially affect the number of poor people, such as the Gini ratio, investment, regional income, etc. In terms of modeling, you can add alternative quasi-Poisson models or add data to the panel data model and add random effects to the Model.

**REFERENCE**


Press.


