Modelling The Number of Unemployment in East Java: Negative Binomial Regression Approach

Zakiatul Wildani[1], Sri Pingit Wulandari[2]
{zakia@its.ac.id[1], sripingitwulandari@gmail.com[2]}

Department of Business Statistics, Institut Teknologi Sepuluh Nopember Kampus ITS Sukolilo-Surabaya 60111, Indonesia[1,2]

Abstract. Unemployment is one of the benchmarks for the success of development in a country and affects sustainable economic growth in an area, including in East Java. The government has made lots of effort to overcome high unemployment, such as holding job fairs every month. However, in East Java, the unemployment rate in 2019 still exceeds the ideal unemployment rate, which is around 2-3 percent. Besides, there is no significant change in the unemployment rate in the last three years during 2017-2019. Therefore, this study aims to model the number of unemployment in East Java by using Negative Binomial regression. In other words, this study investigates how certain factors affect the number of unemployment in East Java. The Negative Binomial regression model is employed in this study as an alternative from the Poisson regression model because the number of unemployment is a count data and, in many cases, is overdispersion. That is, the comparison between the expected value is not the same as the variance. This research will contribute to the East Java Provincial Government or related labor agencies to overcome high unemployment. The finding shows that factors such as regional minimum wage and the number of enterprises significantly affect East Java's unemployment in 2019. Besides, the Negative Binomial regression model with only significant explanatory variables is the best model for modeling the number of unemployment with the lowest AIC value.

Keywords: unemployment, negative binomial regression

1 Introduction

Unemployment nowadays is still considered a significant issue in development and social humanities and economic growth encountered by many countries, including Indonesia. Currently, Indonesia is in the top four of the world's most populous countries, with nearly 280 million population equivalent to 3.51% of the total world population. The high population gives Indonesia on what-so-called bonus demography starting in 2020, which means that the proportion of a working-age population (age between 16-65) will rise and be expected at the peak in 2030 [1]. This bonus demography gives a benefit to the country in terms of boosting economic growth. However, at the same time, it poses some threats. Significantly, it will increase unemployment if the government fails to generate sufficient jobs.

By definition, unemployment is a situation when someone does not have a job and is currently searching for a job; planning for a business; are not looking for a job because they think it is difficult to get a job; and someone who already has a job but has not started the job yet [2]. Research from [3] stated that unemployment has a severe effect on economic progress

AECon 2020, December 19-20, Purwokerto, Indonesia

DOI 10.4108/eai.19-12-2020.2309154

and the country's improvement including the loss of human capital, the strain on families and societies, the primary cause of poverty, the fostering of social and criminal strife, and the hampering of regional growth. Therefore, the government should pay attention to the policies and focus on reviving the labor market and creating jobs as much as possible.

Based on previous research, unemployment is mainly caused by three factors, namely, population, economy, and education. From the population aspects, rapid population growth will impact unemployment if it is not supported by sufficient jobs available in the labor market. In terms of education, the standard of education for the working-age population is still low. Lastly, the economic factor that has not been focused on investment is being unable to absorb adequate labor. There is a standard measure of unemployment termed the unemployment rate, which is the number of unemployed individuals divided by the number of people in the labor force, consisting of both working and unemployed individuals. Indonesia's unemployment rate is relatively high, around 7.1% compared to other countries such as Malaysia and Thailand, which is only 4.6% and 1.9%, respectively. Therefore, in this research, we will focus on investigating factors that significantly affect unemployment in one of the provinces in Indonesia, namely East Java using a regression model.

East Java is widely known as the second-most populous province in Indonesia. Around 14% of the countries' population were residing in this province by the time of 2019. Its capital city Surabaya is recognized as the second-largest city in Indonesia after Jakarta and has become the center of industry and business. In 2019, the unemployment rate in East Java Province was 3.92%, very far from the ideal unemployment rate in developed countries, which is around 2-3 percent [4]. Figure 1 demonstrates there is no significant change in the unemployment rate in East Java for the last three years (2017-2019). In other words, the unemployment problem has not been fully resolved and still be a significant concern in this area. The provincial government, especially The Department of Labour and Transmigration of East Java and labor agencies in several regencies/cities under The Department of Labour and Transmigration of East Java, have held job fairs every month to reduce the high number of unemployment.

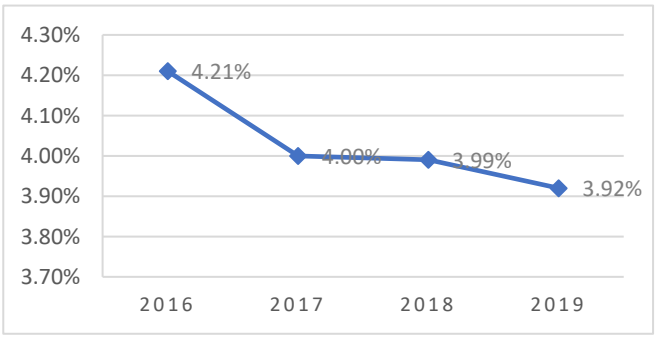

Fig. 1. The unemployment rate in East Java from 2016 to 2019

Many researchers have been modeling the unemployment rate to mitigate this issue. For instance, [5] utilized panel data regression to model the open unemployment rate because the unemployment rate might differ from time to time and region. The study showed that the population aged 15 years and over who worked by the highest education is senior high school or vocational, senior high school's gross participation rate, dependency ratio and Gross Regional Domestic Product (GDP) have a major effect on the unemployment rate in Central Java. Besides, [6] also used the same model to model unemployment in East Java. Another research

was done by [7] employing another model to modeling the unemployment rate in Central Java that is Geographically Weighted regression. Besides, the Geographically Weighted Regression method (NGWR-TS) also can be employed in this case because the unemployment rate has no particular pattern in the regression curve [8]. This research found that factors such as percentage of the low population, percentage of low-educated or elementary school dropouts workforce, economic growth rate, investment ratio, regional minimum wage, the ratio of the amount of large-medium enterprises, percentage of people working in the agricultural sector, and area of agricultural land significantly affect the unemployment rate.

Another model proposed by [9], termed spline regression is conducted in modelling the unemployment rate in Sulawesi. The spline regression is utilised because its ability to the extreme up and down patterns with the dots knots. The research used two predictor variables, labor force participation rate and gross participation rate. Moreover, [10] discussed an extension of the spline regression namely Nonparametric Spline Truncated model to modeling the unemployment rate. The findings presented that labor force participation rate, dependency ratio, average years of schooling and economic growth rate have a significant effect on the unemployment rate in West Java.

Furthermore, [11] mentioned that the industrial sector's growth in labor absorption in Sidoarjo Regency has a positive effect on the employment rate. This study also showed that when the number of industrial sectors is increasing, the labor absorption in the industrial sectors will increase. This result is supported by [12], which said that the Small-Medium Enterprises (SMEs) have an importing rule in creating employment in Indonesia and significantly contribute to the GDP. Besides, SMEs played a significant role in absorbing many workers in Japan, China, India, The United States, Germany, and other countries [3].

All of the presented literature above uses various regression models to model the unemployment rate. However, to the best of our knowledge, not much literature discussed modeling the number of unemployment itself. The number of unemployment is a count data, therefore, we need to use discrete distribution such as Poisson distribution. Thus, this paper presents an approach to modeling the number of unemployment in East Java using Poisson regression. However, in Poisson regression analysis, there is a strict rule where the mean and variance must be similar and it is often rarely fulfilled because of the model's overdispersion. If overdispersion occurs, Poisson regression is not suitable for modeling data, and the estimator of the proposed model will be biased. One of the methods that can be used to overcome overdispersion in Poisson regression is Negative Binomial regression. Therefore, this study compares the Poisson regression model and the Negative Binomial regression model on modeling the number of unemployment in East Java. Furthermore, the comparison is based on the AIC value proposed by [13]. This paper contributes to the literature on alternative modeling of the number of unemployment, helping the government on management and readiness of demographic bonus as well as coping with a high number of unemployment.

2 Method

Before proceeding to the analysis, this section provides an explanation about the theoretical background of the regression model starting with poisson regression and subsequently the negative binomial regression model. The theory in this section is adopted from [14] and [15] if it is not mentioned otherwise.

2.1 Poisson Regression

Poisson regression is a nonlinear regression model that is often used to overcome count data where the response variable follows a Poisson distribution. The count data is a type of data where the values are non-negative, for instance, the number of accidents that happen in Surabaya in 2019, the number of unemployment and so on. The characteristics of the poisson experiments are

1. The number of outcomes are independent
2. Depends on a certain time interval.
3. An event that is included in the counting process.

If the discrete random variable (y) is a Poisson distribution with parameter μ then the probability distribution function of the Poisson distribution can be shown in equation (1).

$$f(y,\mu) = \frac{e^{-\mu}\mu^{y}}{y!}; y = 0,1,2,\ldots,n\,, \tag{1}$$

where μ is the mean Poisson distribution where the mean and variance of y has a value of more than 0. The Poisson regression model can be written in equation (2).

$$\mu = exp\,(\,\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p)\,, \tag{2}$$

where μ is the average number of events that occur within a certain time interval.

2.1.1 Parameter Estimation of Poisson Regression Model

One of the methods utilised to estimate Poisson regression parameters is the Maximum Likelihood Estimation (MLE) method. The MLE method is usually used by maximizing the likelihood function. In Poisson regression model, the estimated parameter is denoted by β_k.

2.1.2 Hypothesis Test

After the estimated parameter is obtained, we can proceed to the hypothesis test. There are two hypothesis tests termed simultaneous tests or widely known as overall significance parameter test and partial test. Overall significant parameter test is employed to determine the effect of parameters on the model with a certain level of significance. The parameter significance test of the Poisson regression model is performed by using the Maximum Likelihood Ratio Test (MLRT) method with hypothesis

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1 : There is at least one $\beta_k \neq 0$, where $k = 1, 2, \ldots, p$.

The test statistics used for the parameter significance test is shown in equation (3).

$$D(\hat{\beta}) = -2\ ln\ \left(\frac{L(\hat{\omega})}{L(\hat{\Omega})}\right)\,, \tag{3}$$

where $L(\hat{\omega})$ and $L(\hat{\Omega})$ are the two likelihood functions associated with the regression model. $L(\hat{\omega})$ is the maximum likelihood value for the model without involving predictor variables and $L(\hat{\Omega})$ is the maximum likelihood value for the model involving predictor variables. The decision to reject H_0 if the value of $D(\hat{\beta}) > \chi^2_{\alpha;p}$, it means that at least one parameter has a significant effect on the model. $D(\hat{\beta})$ is the likelihood ratio statistic that follows the Chi-Squared distribution with p degrees of freedom [16]. Then performed partial parameter test to see the significance of the parameters on the model with the hypothesis

$H_0 : \beta_k = 0$
$H_1 : \beta_k \neq 0$, where $j = 1, 2, \ldots, p$.
The test statistics used is following a Z distribution and can be shown in equation (4).

$$Z = \frac{\widehat{\beta}_k}{S\hat{E}(\widehat{\beta}_k)}, \tag{4}$$

where $S\hat{E}(\hat{\beta}_k)$ is standard error value, that obtained from the $(k+1)^{th}$ diagonal element of $(\widehat{\beta})$ where $(\widehat{\beta}) = -E\left(H^{-1}(\hat{\beta})\right)$. The hypothesis H_0 will be rejected if the value of $|Z|$ is greater than the value of $Z_{\frac{\alpha}{2}}$, where α is the level of significance [16].

2.2 Overdispersion

As we have stated before, in the Poisson regression model there is a strict rule where the variance must be similar to the expected value. However, if the variance is greater than the expected value the Poisson regression model is said to be overdispersion. Suppose there is an overdispersion in count data and still use Poisson regression as the method for solving the problems, an invalid conclusion will be obtained because the standard error value will be underestimated. This is because the regression coefficient parameter that generated from the Poisson regression is inefficient even though the regression coefficient is still consistent. The value of Pearson Chi-Square dispersion $D(\hat{\beta})$ can be calculated with formula in equation (2). The value of θ can be found by using the formula in equation (5), where df is the degrees of freedom that can be obtained from $(n - p - 1)$.

$$\theta = \frac{D(\widehat{\beta})}{df}. \tag{5}$$

If $\theta > 1$, it means that there is an overdispersion in Poisson regression. However, if $\theta < 1$, underdispersion is occurred in the model and if $\theta = 1$ means that there is no over/underdispersion or termed as equaldispersion.

2.3 Negative Binomial Regression

Negative Binomial regression model has a probability mass function that is shown in equation (6) as follows.

$$P(y, \mu, \theta) = \frac{\Gamma\left(y+\frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right)y!}\left(\frac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}}\left(\frac{\theta\mu}{1+\theta\mu}\right)^{y}, \tag{6}$$

where $y = 0, 1, 2, \ldots, n$ and $\mu = exp\ exp\ (X_i^T\beta)$ [17]. Negative Binomial regression can be employed to modeling count data where overdispersion occur in the poisson regression because the Negative Binomial distribution is an extension of the gamma poisson distribution with dispersion parameter θ [18]. In other words, this regression model loosens the extremely restrictive presumption that the variance is equal to mean in the poisson regression model. The condition of overdispersion is shown by the value of $\theta > 1$. The Negative Binomial regression model can be rewritten as follows.

$$\mu = exp\ exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p\right) . \quad (7)$$

2.3.1 Parameter Estimation of Negative Binomial Regression Model

The Maximum Likelihood Estimation (MLE) method is employed to estimate the parameter in Negative Binomial regression. The likelihood function of negative binomial regression is shown below.

$$L(\beta, \theta) = \prod_{i=1}^{n} \frac{\Gamma\left(y+\frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right) y!} \left(\frac{1}{1+\theta\mu_i}\right)^{\frac{1}{\theta}} \left(\frac{\theta\mu_i}{1+\theta\mu_i}\right)^{y} . \quad (8)$$

The regression estimation utilizes Newton-Raphson iteration method for maximizing the likelihood function [19].

2.3.2 Hypothesis Test

The simultaneously parameter significance test of the Negative Binomial regression model using the deviance test with the hypothesis

$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$

H_1 : There is at least one $\beta_k \neq 0$, where $k = 1, 2, \ldots, p$.

The test statistics used for the simultaneously parameter significance test is shown in equation (3) where H_0 will be rejected if the value of $D(\hat{\beta}) > \chi^2_{\alpha;p}$. Furthermore, the partial parameter significance test shows the significance of the parameter individually on the model with the hypothesis

$H_0 : \beta_k = 0$

$H_1 : \beta_k \neq 0$, where $j = 1, 2, \ldots, p$.

The test statistics used for the partially parameter significance test is defined by

$$W_k = \frac{\hat{\beta}_k}{S\hat{E}(\hat{\beta}_k)}, \quad (9)$$

where H_0 will be rejected if the value of W_k or |t| is greater than the value of $t_{\frac{\alpha}{2;n-k}}$. Rejecting H_0 means that the k^{th} parameter significantly affects the respon variable [19].

2.4 Akaike Information Criterion (AIC)

One of indicators for choosing the best model that is generally used in a regression model is the value of Akaike Information Criterion (AIC). This method is based on the MLE method. The formula for calculating the value of AIC has the form in equation (10).

$$AIC = -2\ ln\ L\ (\beta) + 2k\ , \quad (10)$$

where $L(\beta)$ is the value of likelihood and k is the number of parameters. The best regression model is a regression model that has the lowest AIC value [20].

3 Result and Discussion

We consider data obtained from an annual publication published by the Central Bureau of Statistics East Java in 2020. The object of observation is 38 regencies/cities in East Java Province.

3.1 Research variables

The variables used in this study are taken based on the previous researches (see [5], [6], , [7], [8], [11]). More information regarding the variable can be seen in the table below.

Table 1. List of variables.

Variables	Explanation	Measurement Scale
Y	Number of unemployment	Discrete
X_1	Labor's participation rate (%)	Ratio
X_2	Regional minimum wage (IDR)	Ratio
X_3	Percentage of the population age ten years or more with the highest completed high school level and bachelor's education. (%)	Ratio
X_4	Number of enterprises (micro and small enterprises, medium and large enterprises)	Ratio
X_5	High school gross participation rate (%)	Ratio
X_6	Dependency ratio (%)	Ratio

Up until this point, we have discussed the theoretical background of Poisson and Negative Binomial regression and the considered predictor and response variables. To sum up, the following steps describe the modeling procedure for the number of unemployment in East Java.

1. Describe the characteristics of the number of unemployment and the factors that are expected to have a significant effect using descriptive statistics such as mean, median, etc.
2. Estimate the parameter of the Poisson regression model using Newton-Raphson algorithms.
3. Test the hypothesis of the Poisson regression model (simultaneous and partial parameter significance test).
4. Check whether there is overdispersion in the Poisson regression model. Dispersion estimation that is greater than one is evidence that overdispersion occurred.
5. Estimate the parameter of the Negative Binomial regression model using the Poisson gamma distribution algorithms.
6. Test the hypothesis of the Negative Binomial regression model.
7. Compared the proposed models based on their AIC values.
8. Make conclusions.

This chapter will explain the results of the analysis and discussion of modeling the number of unemployment in East Java in 2019 using Poisson regression and Negative Binomial regression.

3.2 Characteristics of the variables

The analysis is started by investigating the characteristics of the variables, for instance, mean, median, variance, etc. The result can be seen in Table 2.

Table 2. Characteristic of the variables.

Variables	Mean	Variance	Median	Minimum	Maximum
Y	22204	335180615	17736	1715	91912
X_1	69.48	12.26	69.12	63.11	79.55
X_2	2254314.00	4.90×10^{11}	1854800.00	1763268.00	3871053.00
X_3	29.01	118.55	24.95	14.77	56.87
X_4	20688	158404288	20447	2618	52616
X_5	85.53	254.90	87.33	55.91	130.59
X_6	44.08	12.66	43.82	35.91	52.21

Table 2 shows the variance of the response variable; where in this case, the number of unemployment is higher than the average. Therefore, this is a sign of overdispersion occurring in the data. Furthermore, the lowest number of unemployment (Y) is 1715 people were in Mojokerto. Meanwhile, Surabaya had the highest unemployment in 2019. The average labor participation rate (X_1) was 69.48%, and the lowest score in Bangkalan, while the highest score in Pacitan.

Moreover, the average regional minimum wage (X_2) was IDR 2254314.00, with the lowest value in Pacitan (IDR 1763268.00) and the highest value in Surabaya (IDR 3871053.00). Besides, the average of the population aged ten years or more with the highest completed level of education is high school and college (X_3) was 29.01%, with the lowest score in Sampang and the highest score in Madiun. The average number of industries (X_4) was 20688 industries. The average high school gross participation rate (X_5) was 85.53%, with the lowest score in Bangkalan and the highest Sidoarjo score. Finally, the average dependency ratio (X_6) was 44.08% per 100 population with the lowest value in Surabaya and the highest value in Bangkalan.

3.3 Multicollinearity check

Multicollinearity is a condition where the predictor variables have a linear correlation (significant relation). One of the methods used for detecting multicollinearity is the Variance Inflation Factor (VIF) value. If a predictor variable has a linear correlation with other predictor variables, the VIF value will be more than 10. The results of the multicollinearity check in this analysis are shown in Table 3.

Table 3. Multicollinearity check.

Variable	VIF
X_1	1.47
X_2	1.70
X_3	3.25
X_4	1.76
X_5	2.31
X_6	1.69

Table 3 indicates that VIF values for all explanatory variables are less than 10 which means that no multicollinearity or linear relationship between the predictor variables.

3.4 Correlation between the response variable and explanatory variables

Before we proceed to the estimation, we need to check the correlation between the response variable with each explanatory variable. The scatterplot between the number of unemployment in East Java with each predictor variable is depicted in Figure 2.

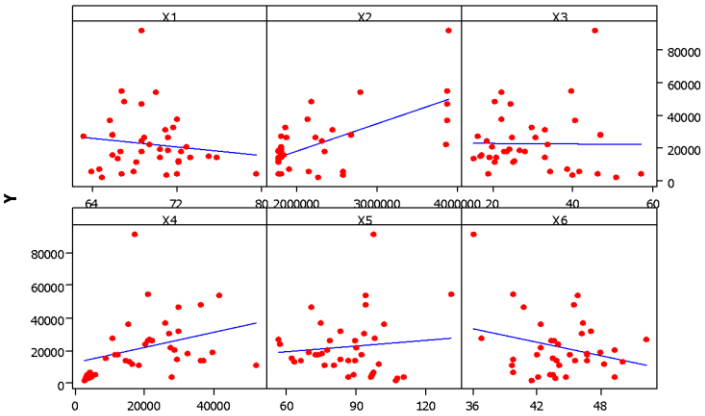

Figure 2. Scatterplot response variable vs. predictor variables.

Figure 2 shows no sign of positive correlation or negative correlation between the response variable and predictor variables except the regional minimum wage (X_2). The graph demonstrates correlation tends to be positive, which means that if the regional minimum wage increases, the number of unemployment will also increase. The correlation coefficient value also supports this result between the number of unemployment and the predictor variables shown in Table 4.

Table 4. Correlation Coefficient

Variable	Correlation Coefficient
X_1	-0.13
X_2	**0.64**
X_3	-0.002
X_4	0.32
X_5	0.10
X_6	-0.27

Table 4 shows a positive correlation coefficient between the number of unemployment with the regional minimum wage, the number of industries, and the high school gross participation rate. On the other hand, the correlation coefficient between the number of unemployment with the labor participation rate, the percentage of population ten years and over with the highest level of education completed is a high school or university, and the dependency ratio is negative. This result is in line with the relationship pattern shown in Figure 2.

3.45 Poisson regression model

The results of Poisson regression analysis in modeling the number of unemployment in East Java can be explained as follows.

3.5.1 Estimation and significance parameters of the Poisson regression model.

The result of the Poisson regression model's estimation and significance parameters for each predictor variables and intercept can be shown in Table 5.

Table 5. Estimation and significance parameters of a Poisson regression model.

Variable	Coefficient	Standard Error	Z	p-value
Intercept	1.23×10^{1}	3.96×10^{-2}	310.56*	0.00*
X_1	-3.86×10^{-2}	4.19×10^{-4}	-92.17*	0.00*
X_2	5.47×10^{-7}	1.73×10^{-9}	315.97*	0.00*
X_3	-1.56×10^{-2}	2.04×10^{-4}	-76.52*	0.00*
X_4	2.25×10^{-5}	1.15×10^{-7}	195.60*	0.00*
X_5	1.63×10^{-3}	9.01×10^{-5}	18.11*	0.00*
X_6	-2.59×10^{-2}	4.33×10^{-4}	-59.88*	0.00*
Deviance	= 219589			
AIC	= 220041			

* indicate the rejection of the null hypothesis at the 5% significance level test

Based on Table 5, the Poisson regression model can be formed as follows.

$$\hat{\mu} = exp(\ 1.23 \times 10^{1} - 3.86 \times 10^{-2}X_1 + 5.47 \times 10^{-7}X_2 - 1.56 \times 10^{-2}X_3 + 2.25 \times 10^{-5}X_4 + 1.63 \times 10^{-3}X_5 - 2.59 \times 10^{-2}X_6).$$

Then, we proceed to the hypothesis test, such as the simultaneous parameter significance test, which is widely known as the overall significance test. This test is conducted to know whether all the predictor variables simultaneously have a significant effect on the response variable (the number of unemployment). Using a significant level α of 0.05, we will reject the null hypothesis that all predictor variables are not significantly affecting the response variable if deviance is greater than $\chi^2_{0.05(6)}$. Based on result in Table 5, the deviance is 219589 greater than $\chi^2_{0.05(6)}$ that is 12.59. Therefore, H_0 will be rejected. In other words, there are at least one of the predictor variables that have a significant effect on the number of unemployment.

After we conduct the simultaneous parameter significance test, we proceed to the partial parameter significance test. Unlike the simultaneous parameter significant test, where we measure all predictor variables' effect on the response variable, we measure the effect of each predictor variable on the response variable in this test. Using a significant level α of 0.05, H_0 will be rejected if the |Z| value is greater than $Z_{0.025}$ or if the p-value is less than α. Based on the results shown in Table 5, the value of |Z| for all predictor variables is greater than the value of $Z_{0.025}$ that is 1.96. Also, p-value for all predictor variables is less than α. Consequently, the null hypothesis will be rejected. It means that each predictor variable affects the number of unemployed in East Java considerably.

3.5.2 Overdispersion Check.

Overdispersion check is managed by dividing the deviance by its degrees of freedom. The deviance value of the poisson regression model is 219589, and the degrees of freedom is 31 obtained from (n-p-1), therefore, the ratio is 7083.52 greater than 1. Strictly speaking, there is overdispersion in the data of the number of unemployment in East Java in 2019. The parameter of Poisson regression that had overdispersion will be biased. Therefore, to resolve the overdispersion, the Negative Binomial approach will be used. Firstly, we need to determine the initial value for minimizing the dispersion parameter. The initial value is attained through trial and error until the division of deviance to degrees of freedom is one so that there is no overdispersion. The result of the trial and error for the initial value can be shown in Table 6 below.

Table 6. Trial and error for the initial value.

Initial θ	Deviance	df	Deviance/df
3.5	41.73	31	1.35
3	35.77	31	1.15
2.5	29.81	31	0.96
2.6	31.002	31	1.00006
2.59	30.88	31	0.996
2.595	30.94	31	0.998
2.596	30.95	31	0.9985
2.597	30.97	31	0.9989
2.598	30.98	31	0.9993
2.599	30.99	31	0.9997
2.5991	30.99	31	0.9997
2.5999	31.001	31	1.00003
2.5998	30.999	31	0.99997
2.59985	31	31	1

Table 6 shows that after some trial and error, we can obtain value one if the initial value is 2.59985.

3.6 Negative Binomial Regression

The results of Negative Binomial regression analysis on the modeling of the number of unemployment in East Java by using the initial value 2.59985 can be explained as follows. In addition to the Negative Binomial regression model, the estimation and significance parameters are presented in Table 7.

Table 7. Estimation and significance parameters of Negative Binomial regression model.

Variable	Coefficient	Standard Error	t	p-value
Intercept	1.23×10^{1}	2.97	4.16*	0.00*
X_1	-4.15×10^{-2}	3.26×10^{-2}	-1.27	0.21
X_2	7.10×10^{-7}	1.75×10^{-7}	4.05*	0.00*
X_3	-2.24×10^{-2}	1.56×10^{-2}	-1.44	0.16
X_4	2.99×10^{-5}	9.92×10^{-6}	3.01*	0.01*
X_5	-4.17×10^{-3}	8.96×10^{-3}	-0.47	0.64
X_6	-1.89×10^{-2}	3.44×10^{-2}	-0.55	0.59

Deviance = 31.00
AIC = 815.50
* indicate the rejection of the null hypothesis at the 5% significance level test

Based on Table 7, we can write Negative Binomial regression model as follow.

$$\hat{\mu} = \exp(1.23\times10^{1} - 4.15\times10^{-2}X_1 + 7.10\times10^{-7}X_2 - 2.24\times10^{-2}X_3 + 2.99\times10^{-5}X_4 - 4.17\times10^{-3}X_5 - 1.89\times10^{-2}X_6)$$

$$\ln(\hat{\mu}) = 1.23\times10^{1} - 4.15\times10^{-2}X_1 + 7.10\times10^{-7}X_2 - 2.24\times10^{-2}X_3 + 2.99\times10^{-5}X_4 - 4.17\times10^{-3}X_5 - 1.89\times10^{-2}X_6$$

.

Same procedures as the Poisson regression model above also proceed in this Negative Binomial regression model. In the simultaneous significance test with a significant level α of

0.05, we reject the null hypothesis if the deviance is greater than $\chi^2_{0.05(6)}$. Based on Table 7, the Deviance value is 31.00, greater than 12.59. Therefore, the null hypothesis that there are no explanatory variables that significantly affect the response variable will be rejected. It means there are at least one predictor variable that affects the number of unemployment. Then, the analysis will be continued to the partial parameter significance test. Using the same significant level of α 0.05, H_0 will be rejected if the value of |t| is greater than $t_{0.025;32}$ or if the p-value is less thanα. Table 7 shows there are only two variables that significantly affect the number of unemployment, namely the regional minimum wage (X_2) and the number of enterprises (X_4).

3.7 Negative Binomial Regression using significant predictor variables

The Negative Binomial regression analysis in the previous subsection shows that the regional minimum wage (X_2) and the number of industries (X_4) significantly affect the response variable with AIC value 815.50 while the other predictors are not. Therefore we need to re-modeling the Negative Binomial regression model with only significant predictor variables. Before we do that, we need to determine the initial value θ for minimizing the dispersion parameter. The initial value θ is obtained through trial and error to get the result of the division of Deviance value to degrees of freedom is equal to 1, therefore, there is no overdispersion. The results of the trial and error for the initial value can be shown in the table below.

Table 8. The results of the trial and error for the initial value of Negative Binomial Regression model using significant predictor variables.

Initial $\boldsymbol{\theta}$	Deviance	df	Deviance/df
2.5	34.82	35	0.99
2.6	36.21	35	1.03
2.55	35.51	35	1.01
2.525	35.16	35	1.005
2.51	34.95	35	0.999
2.515	35.02	35	1.0007
2.5125	34.99	35	0.9997
2.5126	34.00	35	0.9997
2.5127	34.992	35	0.9998
2.5129	34.995	35	0.9999
2.5132	34.999	35	0.99997
2.51325	34.999	35	0.9997
2.51327	35	35	1

Table 8 shows that by using the initial value 2.51327, we obtain a division between deviance value and degrees of freedom is 1. Then, modeling Negative Binomial regression using significant predictor variables will be done using the initial value 2.51327. Estimation and significance parameters of Negative Binomial Regression model using significant predictor variables only.

The estimation and significance parameters of the Negative Binomial regression model using significant predictor variables with the initial value 2.51327 can be shown in Table 9.

Table 9. Estimation and significance parameters of Negative Binomial Regression model using significant predictor variables.

Variable	Coefficient	Standard Error	t	p-value
Intercept	7.78	3.39×10^{-1}	22.95*	0.00*
X_2	6.09×10^{-7}	1.28×10^{-7}	4.76*	0.00*

X_4	3.36×10^{-5}	7.12×10^{-6}	4.72*	0.00*
Deviance	= 35.00			
AIC	= 812.87			

* indicate the rejection of the null hypothesis at the 5% significance level test

Based on Table 9, the Negative Binomial regression model can be formed as follows.

$$\hat{\mu} = \exp(7.78 + 6.09\times10^{-7} X_2 + 3.36\times10^{-5} X_4)$$
$$\ln(\hat{\mu}) = 7.78 + 6.09\times10^{-7} X_2 + 3.36\times10^{-5} X_4 \quad .$$

The simultaneous and partial significance test conclude that the null hypothesis is rejected, which means that the predictor variables significantly affect the number of unemployment with AIC values is 812.87.

3.8 Choosing The Best Regression Model

Up until now, we have successfully built three models for modeling the number of unemployment in East Java. Therefore, we need to choose the best model based on the lowest AIC value criteria. The AIC values from each regression model can be shown in Table 10.

Table 10. Choosing the best regression model.

Regression Model	AIC Value	Note
Poisson	219589	Overdispersion
Negative Binomial with all predictor variables	815.50	Equaldispersion
Negative Binomial with significant predictor variables	812.87	Equaldispersion

Table 10 shows that the Negative Binomial regression model with all significant predictor variables has the lowest AIC value, around 812.87, much lower than the other two models. Therefore, we choose this model over two other models. This model suggests that, ceteris paribus: as the regional minimum wage goes up by a rupiah, the number of unemployment in East Java increases around $exp(6.09 \times 10^{-7}) = 1.000000609$. Moreover, as the number of enterprises in East Java goes up, the number of unemployment will rise around $e(\ 3.36 \times 10^{-5}) = 1.000033601$. This result is different from the previous research where the growth in the industrial sector on labor absorption in Sidoarjo Regency has a positive effect on the employment rate (see [3], [11], and [12]).

4 Conclusion

Based on the analysis, we found that variable regional minimum wage (X_2) and the number of enterprises (X_4), either micro and small enterprises and large and medium enterprises, positively affect the number of unemployment in East Java in 2019. In other words, by ceteris paribus, increasing the regional minimum wage will increase the number of unemployment. This can be caused by various things, especially worker efficiency policies in the industries. Moreover, increasing the number of industries will also increase the number of unemployment in East Java. This finding is different from previous research that can be caused by the fact that a lot of human resources do not meet the qualifications set by the industry even though the number of industries is rising. Besides, the Negative Binomial regression model is the best

model for modeling the number of unemployment in East Java in 2019 with the lowest AIC value. A few limitations as follows. We restrict our study to the six predictor variables. Future studies could extend the regression by including or omitting a variable that is considered in this study. Note that the different predictor variables will produce a different result, seeing as each variable may react differently. In addition, future research is highly encouraged to extend our approach by considering different regression models such as panel regression, dynamic models, etc.

Acknowledgments. The authors would like to thank the Department of Business Statistics, Vocational Faculty, Sepuluh Nopember Institute of Technology for the grants and support to this research.

References

[1] Warsito T. Attaining the Demographic Bonus in Indonesia. J Pajak Dan Keuang Negara 2019;1:8.
[2] Statistik BP. Provinsi Jawa Timur Dalam Angka. Surabaya Badan Pus Stat Propinsi Jawa Timur 2016.
[3] Alia Y. The Effectiveness of Small and Medium Enterprises Adoption as a Strategic Option to Solve Unemployment Problem in the Arab World, an Example of Algeria. IjbssnetCom 2014;5:161–71.
[4] Sadono S. Makro Ekonomi, Edisi Ketiga, PT. Raja Graf Persada Jakarta 2008.
[5] Prasanti TA, Wuryandari T, Rusgiyono A. Aplikasi Regresi Data Panel Untuk Pemodelan Tingkat Pengangguran Terbuka Kabupaten/Kota Di Provinsi Jawa Tengah. Semin Nas Mat Dan Pendidik Mat 2020;4:687–96.
[6] Sari R, Sari RS, Budiantara IN. Pemodelan Pengangguran Terbuka di Jawa Timur dengan Menggunakan Pendekatan Regresi Spline Multivariabel. J Sains Dan Seni ITS 2012;1:D236–41.
[7] Utami TW, Semiparametrik R, Truncated S, Negara D. Pendekatan Regresi Semiparametrik Spline Truncated untuk Pemodelan Tingkat Pengangguran Terbuka di Jawa Tengah. Statistika 2018;6.
[8] Sifriyani, Budiantara IN, Kartiko SH, Gunardi. Evaluation of Factors Affecting Increased Unemployment in East Java Using NGWR-TS Method. Int J Sci Basic Appl Res 2019;46:123–42.
[9] Wahyuni SA, Ratnawati R, Indriyani I, Fajri M. Spline Regression Analysis to Modelling The Open Unemployment Rate in Sulawesi. Nat Sci J Sci Technol 2020;9:2–7.
[10] Kurniawati NA. Pemodelan Tingkat Pengangguran Terbuka di Nonparametrik Spline Truncated 2019;8:2–8.
[11] Purwasih H, Soesatyo Y. Pengaruh Pertumbuhan Sektor Industri Terhadap Penyerapan Tenaga Kerja Di Kabupaten Sidoarjo. J Pendidik Ekon 2017;5:1–6.
[12] Utami RM, Lantu DC. Development Competitiveness Model for Small-Medium Enterprises among the Creative Industry in Bandung. Procedia - Soc Behav Sci 2014;115:305–23.
[13] Sakamoto Y, Ishiguro M, Kitagawa G. Akaike information criterion statistics. Dordrecht, Netherlands D Reidel 1986;81.
[14] Agresti A. Categorical data analysis. vol. 482. John Wiley & Sons; 2003.
[15] Walpole RE, Myers RH. Probability & statistics for engineers & scientists. Pearson Education Limited; 2012.
[16] McCullagh P. Generalized linear models. Routledge; 2018.
[17] Greene W. Functional forms for the negative binomial model for count data. Econ Lett 2008;99:585–90.
[18] Hilbe JM. Negative binomial regression. Cambridge University Press; 2011.
[19] Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression. vol. 398. John Wiley & Sons; 2013.

[20] Bozdogan H. Akaikes Information Criterion and Recent Developments in Information Complexity. J Math Psychol 2000;44:62–91.