

Implementation of Negative Binomial Regression to Address Overdispersion in the Analysis of Unemployment Determinants in Sulawesi, 2023

Erna¹, Ermawati^{2*}, Wahidah Alwi³, Sri Dewi Anugrawati⁴

^{1,2,3,4} *Mathematics Study Program, Universitas Islam Negeri Alauddin Makassar, Indonesia*

*Corresponding author: ermawati@uin-alauddin.ac.id

*Submission date: 01 October 2025, Revision: 19 October 2025, Accepted: 27 October 2025

ABSTRACT

This study focuses on the implementation of negative binomial regression as a solution to overcome the problem of overdispersion in the analysis of determinants of unemployment on the island of Sulawesi in 2023. Unemployment is not only viewed as a statistical phenomenon or economic issue, but also as an important indicator that reflects social welfare and the success of development in a region. Sulawesi Island, with its growth in the agricultural and industrial sectors, faces serious challenges in reducing unemployment rates, which have the potential to cause regional disparities if not addressed appropriately. This study aims to develop an appropriate negative binomial regression model to overcome overdispersion and identify the main factors that influence the unemployment rate. The method used is negative binomial regression analysis of district/city unemployment data in Sulawesi Island, which is discrete and shows symptoms of overdispersion. The results of the study produced the negative binomial regression model with significant variables including population size, Human Development Index (HDI), and the number of job placement or fulfillment services. These three factors have been proven to have a significant effect on the number of unemployed people in Sulawesi Island in 2023.

KEYWORDS

Unemployment, Overdispersion, Negative Binomial Regression

1. INTRODUCTION

Unemployment is not merely a statistical figure or an economic issue, but rather a social phenomenon that reflects the level of community welfare and the success of a country's development. High unemployment rates indicate an imbalance between the available labor force and existing job opportunities. This condition not only hinders economic growth but can also lead to social problems such as poverty, crime, and social instability. Therefore, understanding and addressing unemployment is an important step in supporting sustainable development [1]. Based on data from the Central Statistics Agency (BPS), the unemployment rate in Indonesia in 2023 was recorded at 5.32%, indicating that there are still structural challenges in employment, particularly related to the mismatch between workforce skills and the needs of a growing market. When viewed more specifically, the unemployment rate in Sulawesi Island is relatively higher than in other islands, with some provinces such as North Sulawesi reaching more than 6% [2]. This situation is quite concerning, given that Sulawesi Island has great potential in the agricultural and industrial sectors, which should be able to absorb more labor.

In statistical analysis, unemployment data is classified as count data, which is often analyzed using Poisson regression. However, the basic assumption of Poisson regression requires that the variance be equal to the mean, while unemployment data generally exhibits overdispersion, a condition in which the variance is greater than the mean. This situation results in less valid analysis results [3]. To address this issue, negative binomial regression is used as an alternative because it has the ability to add

dispersion parameters, resulting in more accurate estimates [4]. A number of previous studies have also shown the superiority of negative binomial regression in addressing overdispersion, as demonstrated by Elvan Hayat and Sozen Ozden (2023) [5] and Satriani (2022), who found that this model is superior to generalized Poisson regression [6].

Based on this description, this study focuses on the implementation of negative binomial regression to analyze the determinants of unemployment in Sulawesi Island in 2023. The purpose of this study is to construct a negative binomial regression model that can overcome the problem of overdispersion while identifying factors that significantly affect unemployment in Sulawesi Island in 2023.

2. LITERATURE REVIEW

2.1 Negative Binomial Distribution

One commonly used variant of the Negative Binomial distribution is based on a series of Bernoulli trials. This distribution represents the number of Bernoulli trials conducted until k successes are obtained. Each trial is assumed to be independent, with a fixed probability of success of l and a probability of failure of $1-l$. If the random variable X indicates the total number of trials required to achieve k successes, then the probability distribution of X is known as the Negative Binomial distribution. This distribution has a specific probability function that describes the characteristics of the process [7]:

$$Pr(Y = y) = \frac{\Gamma(y+k)}{\Gamma(k)y!} l^k (1-l)^y \quad (1)$$

where:

$$k = \frac{1}{\alpha}$$

$$l = \frac{1}{1 + \alpha \exp(\mathbf{X}_i^T \beta)}$$

$$y = 1, 2, 3, \dots$$

2.2 Multicollinearity Test

Multicollinearity testing aims to assess whether there is a high correlation or linear relationship between the independent variables in a regression model. The test statistic used is [8]:

$$VIF_j = \frac{1}{1 - R_j^2}, j = 1, 2, 3, \dots k \quad (2)$$

The index k indicates the independent variable k .

Where R_j^2 is the coefficient of determination between the j -th independent variable and other independent variables in the model. If the VIF value exceeds 10 [9], this indicates multicollinearity in the variable.

2.3 Overdispersion

In analyzing count data, there is one important assumption that must be met, namely the assumption of equidispersion, where by the variance of the response variable must be equal to its mean. However, in practice, it is often found that the variance of the response variable exceeds or is not proportional to its mean, indicating a violation of this assumption [7]. To ensure that the assumption of equidispersion is met, testing can be performed using the following formula, which is designed to detect discrepancies between the variance and mean in count data.

$$\alpha = \frac{\text{nilai deviance}}{df} \quad (3)$$

$$\text{nilai deviance} = 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right] \quad (4)$$

Parameters α acts as a measure of dispersion, while df represents the degrees of freedom in the model. When the deviation value divided by the degrees of freedom produces a number greater than one, this indicates a symptom of overdispersion in the data [10].

Negative Binomial Regression

Negative binomial regression is a type of regression within the Generalized Linear Models (GLM) framework designed to analyze data with higher variability than assumed by the Poisson model, also known as overdispersion. This model uses the Negative Binomial distribution as the distribution for the response variable, which is theoretically a combination of the Poisson and Gamma distributions. The logarithmic form of the negative binomial regression model can be written as follows [11]:

$$\ln(\mu_i) = \mathbf{X}_i^T \boldsymbol{\beta} \quad (5)$$

$$\mu_i = \exp(\mathbf{X}_i^T \boldsymbol{\beta}) = \exp(\beta_0 + \beta_1 X_{i1} + \cdots + \beta_j X_{ji}), \quad i = 1, 2, \dots, n \quad (6)$$

Where $\mathbf{X}_i^T \boldsymbol{\beta}$ is a form of regression function used in the parametric regression model framework [7].

Estimation of Negative Binomial Regression Parameters

Parameters in the negative binomial regression model, such as $\beta_0, \beta_1, \beta_2, \dots, \beta_j$, are unknown values, so an estimation process is required. To obtain an estimate of these parameters, the Maximum Likelihood Estimation (MLE) method is used [12].

1. Forming the likelihood function

$$\begin{aligned} L(\boldsymbol{\beta}, \alpha) &= \prod_{i=1}^n f(y_i; \boldsymbol{\beta}, \alpha) \\ &= \prod_{i=1}^n \frac{\Gamma(y_i + \frac{1}{\alpha})}{\Gamma(\frac{1}{\alpha}) y_i!} \left(\frac{1}{1 + \alpha \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right)^{\frac{1}{\alpha}} \left(\frac{\alpha \exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \alpha \exp(\mathbf{X}_i^T \boldsymbol{\beta})} \right)^{y_i} \end{aligned} \quad (7)$$

2. Form the log function of the likelihood function that has been obtained

$$\begin{aligned} L(\boldsymbol{\beta}, \alpha) &= \ln L(\boldsymbol{\beta}, \alpha) \\ &= \sum_{i=1}^n \left[\ln \Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln \Gamma\left(\frac{1}{\alpha}\right) - \ln(y_i!) + \frac{1}{\alpha} \ln\left(\frac{1}{1 + \alpha \exp(\mathbf{X}_i^T \boldsymbol{\beta})}\right) + y_i \ln\left(\frac{\alpha \exp(\mathbf{X}_i^T \boldsymbol{\beta})}{1 + \alpha \exp(\mathbf{X}_i^T \boldsymbol{\beta})}\right) \right] \end{aligned} \quad (8)$$

3. Maximizing the function in $L(\boldsymbol{\beta}, \alpha)$ can be done by differentiating each parameter $\boldsymbol{\beta}$ and α , and then setting the result equal to zero. Determining the maximum likelihood estimator requires a relatively complex iterative procedure. In negative binomial regression, the parameter estimation process is usually done using numerical methods, one of which is the Newton–Raphson algorithm.

3. METHODOLOGY

This study uses a quantitative approach with a focus on testing the formulated hypotheses. The data used is secondary data obtained from official publications of the Central Statistics Agency (BPS) in 2023. The response variable in this study is the number of unemployed people in districts/cities throughout Sulawesi Island. The predictor variables include:

1. Labor Force Participation Rate (X_1),
2. Population (X_2),
3. Number of Job Vacancies (X_3),
4. Human Development Index (X_4),
5. Economic Growth (X_5),
6. Number of Job Placement/Fulfillment Services (X_6).

3.1 Research Procedures

1. Conducting descriptive analysis to describe the characteristics of unemployment data and variables that are estimated to have an influence on unemployment rates in Sulawesi Island in 2023.
2. Standardize data on predictor variables to equalize the scale between each variable.
3. Testing the assumptions underlying the application of negative binomial regression, with the following steps:
 - (a) Perform a negative binomial distribution test.
 - (b) Perform a multicollinearity test between predictor variables.
 - (c) Perform an overdispersion test on the response variable.
4. Modeling the number of unemployed people in Sulawesi Island in 2023 using negative binomial regression analysis was carried out in the following stages:
 - (a) Estimating the parameters of the negative binomial regression model using the Maximum Likelihood Estimation (MLE) approach.
 - (b) Evaluating the suitability of the model through a goodness-of-fit test using the deviance value as a measure of model consistency.
 - (c) Testing the significance of the model parameters as a whole (simultaneous/ overall test) and individually (partial/ subtest).
 - (d) Identifying significant determinants that affect the number of unemployed based on the modeling results obtained.

4. RESULT & DISCUSSION

4.1 Descriptive Analysis

Before applying negative binomial regression, descriptive analysis must first be conducted to gain an initial understanding of the general characteristics of the data in the study. The data analyzed in this study comes from the number of unemployed people in the Sulawesi region in 2023 (Y), which acts as the response variable. Meanwhile, there are six predictor variables used, namely Labor Force Participation Rate (LFPR) (X_1), population (X_2), number of job vacancies (X_3), Human Development Index (HDI) (X_4), economic growth (X_5), and the number of job placement/ fulfillment services (X_6). The following are the results of the descriptive analysis of the variables used in this study:

Table 1. Descriptive Analysis

Variable	Average	Variance	Minimum	Maximum
Y	5,097	63,388,250	334	69,124
X_1	68.44	27.6988	55.49	82.35
X_2	253,999	39,382,790,000	39,000	1,415,960
X_3	588	1,033,869	0	5,979
X_4	72.52	16.91665	66.94	85.51
X_5	4.60	13.69618	0.75	23.04
X_6	764	8,718,944	0	26,373

Based on **Table 1**, it can be seen that the average number of unemployed people in the Sulawesi region in 2023 reached 5,097. Meanwhile, the variance value is relatively high, at 63,388,250. This high variance is due to regional disparities, where some districts/cities have unemployment rates in the tens of thousands, while other areas only record unemployment rates in the hundreds.

4.2 Assumption Testing

4.2.1 Negative Binomial Distribution Test

After conducting descriptive analysis, the next step is to test whether the negative binomial distribution is appropriate, to ensure that the distribution of unemployment data follows a negative binomial distribution pattern. This test is conducted by proposing the following hypothesis:

- H_0 : The number of unemployed people on the island of Sulawesi does not follow a negative binomial distribution pattern
- H_1 : The number of unemployed people on Sulawesi Island follows a negative binomial distribution pattern

Table 2. Kolmogorov–Smirnov Test

Statistic	Value
Sample Size (n)	81
Kolmogorov–Smirnov	0.11077
P-Value	0.254

The calculation results are reflected in **Table 2**. Based on the table, the significance value obtained is 0.254, which is greater than the significance level of 0.05. This indicates that there is sufficient evidence to reject H_0 , so it can be concluded that the distribution of unemployment data is in accordance with the negative binomial distribution.

4.2.2 Multicollinearity Test

After testing shows that the response variable follows a negative binomial distribution, the next step is to perform a multicollinearity test to identify whether there is a high correlation between predictor variables. One approach used to detect the presence of multicollinearity is through the calculation of the Variance Inflation Factor (VIF) value. This test is carried out by proposing the following hypothesis:

- H_0 : $VIF \leq 10$ (there is no significant multicollinearity between predictor variables)
- H_1 : $VIF > 10$ (there is significant multicollinearity between predictor variables)

The calculation results are reflected in **Table 3**.

Table 3. VIF Values of Predictor Variables

Variable	VIF
X_1	1.266339
X_2	2.096258
X_3	2.350984
X_4	1.369181
X_5	1.259355
X_6	3.017928

Based on the calculations in **Table 3**, all VIF values of the predictor variables are less than 10. This indicates that there is insufficient evidence to reject H_0 . This condition indicates that there is no multicollinearity among the predictor variables, so the analysis process can proceed to the overdispersion testing stage.

4.2.3 Overdispersion Test

Overdispersion testing of unemployment data in Sulawesi Island in 2023 can be done by looking at the deviance value. If the result of dividing the deviance by the degree of freedom exceeds 1, then this condition indicates overdispersion. This test is conducted by proposing the following hypotheses:

- H_0 : the unemployment data in Sulawesi Island does not show any overdispersion.
- H_1 : the unemployment data in Sulawesi Island shows overdispersion.

The dispersion value can be calculated using the following formula:

$$\alpha = \frac{\text{deviance value}}{df}$$

$$\alpha = \frac{43,038.45}{74}$$

$$\alpha = 581.6007$$

Based on the calculation results, a dispersion value of 581.6007 was obtained. This indicates that there is insufficient evidence to accept H_0 , so it can be concluded that the unemployment data is overdispersed. This can also be seen from the results of the descriptive analysis in **Table 1**, which shows that the variance value of the unemployment data in this study is higher than the average value. Thus, the appropriate approach to analyze the factors that influence the unemployment rate in Sulawesi Island is through the application of negative binomial regression.

4.3 Negative Binomial Regression Modeling

4.3.1 Estimated Values of Negative Binomial Regression Parameters Using the MLE Method

The results of model parameter estimation using the Negative Binomial distribution approach are presented in **Table 4**:

Table 4. Estimated Results of Negative Binomial Regression Parameters

Parameter	Estimate	Standard Error	$z_{\text{Calculated}}$	P-value
β_0	8.19315	0.04	194.989	$< 2 \times 10^{-16}$
β_1	-0.009228	0.05	-0.194	0.846
β_2	0.677306	0.06	11.069	$< 2 \times 10^{-16}$
β_3	0.092517	0.06	1.427	0.153
β_4	0.347154	0.05	7.020	2.23×10^{-12}
β_5	-0.018401	0.05	-0.388	0.698
β_6	-0.304464	0.07	-4.148	0.000336

Based on the **Table 4**, the negative binomial regression model was developed as:

$$\hat{\mu} = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \hat{\beta}_4 X_4 + \hat{\beta}_5 X_5 + \hat{\beta}_6 X_6)$$

$$\hat{\mu} = \exp(8.19315 - 0.009228X_1 + 0.677306X_2 + 0.092517X_3 + 0.347154X_4 - 0.018401X_5 + (-0.304464X_6))$$

$$\ln(\hat{\mu}) = 8.19315 - 0.009228X_1 + 0.677306X_2 + 0.092517X_3 + 0.347154X_4 - 0.018401X_5 + (-0.304464X_6)$$

Before applying the model, the first step is to test the model's suitability.

4.3.2 Negative Binomial Regression Model Goodness of Fit Test

The goodness of fit of a negative binomial regression model can be tested by dividing the deviance value by its degrees of freedom. This test is performed by proposing the following hypotheses:

H_0 : there is a mismatch between the model and the observed data

H_1 : the constructed model has a good level of fit with the observed data.

The calculation results are reflected in **Table 5**.

Table 5. Suitability of the Negative Binomial Regression Model

Statistic	Value	Df	Deviance Value/df
Deviance Value	82.95258	74	1.120981

Based on **Table 5**, the deviance value in the negative binomial regression model shows a relatively small number and is close to 1. This indicates that there is sufficient evidence to reject H_0 . This condition indicates that the negative binomial regression model is the most appropriate choice in dealing with the problem of overdispersion.

4.3.3 Simultaneous Significance Test of Negative Binomial Regression Model Parameters (Overall Test) and Partial Test (Subtest)

After the negative binomial regression model has been successfully obtained, the next step is to conduct a simultaneous test and a partial test of each parameter.

The results of simultaneous testing (model feasibility test) on negative binomial regression are obtained based on the following hypothesis testing:

$H_0: \beta_1 = \beta_2 = \dots = \beta_j = 0$ (no significant effect of predictor variables on the number of unemployed)
 H_1 : At least one $\beta_j \neq 0; j = 1, 2, 3, \dots, 6$ (there is at least one predictor variable that has a significant effect on the number of unemployed)

The calculation results are reflected in **Table 6**

Table 6. Simultaneous Test of the Negative Binomial Regression Model

Test Statistic	Value
$D(\hat{\beta})$	149.26
$\chi^2_{(0.05;6)}$	12.592

Based on **Table 6**, it can be seen that the value of $D(\hat{\beta})$ is greater than the value of $\chi^2_{(0.05;6)}$. Therefore, the results of the simultaneous test on the negative binomial regression model show that H_0 is rejected. This means that there is at least one predictor variable that has a significant effect on the number of unemployed people.

After the simultaneous test is performed, the next step is to test the significance of the parameters partially with the following hypotheses:

$H_0: \beta_i = 0$ (predictor variables do not have a significant effect on the number of unemployed)
 $H_1: \beta_i \neq 0$ (predictor variables have a significant effect on the number of unemployed)

Based on **Table 4**, it can be seen that variables X_2 , X_4 , and X_6 have p-values that are smaller than the significance level $\alpha = 0.05$. These results indicate that there is insufficient evidence to uphold the null hypothesis (H_0). Thus, it can be concluded that the variables of population size (X_2), Human Development Index (HDI) (X_4), and number of job placement/ fulfillment services (X_6) have a significant effect on the response variable, namely the number of unemployed.

The results of the negative binomial regression model interpretation show that an increase in population has an impact on the average number of unemployed people, which rises by $\exp(0.677306)$ or 1.97 times, assuming that other variables are not taken into account in the analysis model. Similarly, an increase in HDI also has an impact on the average number of unemployed people, which increases by $\exp(0.347154)$ or 1.415 times, assuming that other variables are not taken into account in the analysis model. Conversely, an increase in the number of job placement/fulfillment services can reduce the average number of unemployed by $\exp(-0.304464)$ or equivalent to 0.74 times, assuming that other variables are not taken into account in the analysis model.

4.4 Determinants of Unemployment Rates

Partial testing shows that the significant variables are X_2 , X_4 , and X_6 . Therefore, negative binomial regression modeling is repeated by only including variables that have been proven to be significant. After conducting the analysis process by considering only three variables, namely X_2 , X_4 , and X_6 , the latest negative binomial regression model was obtained. The model was formulated based on the estimation results and is presented in **Table 7**:

Table 7. Testing Significant Parameters

Parameter	Estimate	Standard Error	p-value
β_0	8.19511	0.04	0
β_1	0.67302	0.06	0
β_2	0.35310	0.05	0
β_3	-0.23735	0.06	0

The negative binomial regression model developed in this section is the result of parameter estimation in **Table 7** where:

$$\hat{\mu} = \exp(\hat{\beta}_0 + \hat{\beta}_2 X_2 + \hat{\beta}_4 X_4 + \hat{\beta}_6 X_6)$$

$$\hat{\mu} = \exp(8.19511 + 0.67302X_2 + 0.3531X_4 - 0.23735X_6)$$

$$\ln(\hat{\mu}) = 8.19511 + 0.67302X_2 + 0.3531X_4 - 0.23735X_6$$

Interpretation of the negative binomial regression results shows that an increase in population has an impact on an increase in the average number of unemployed persons by $\exp(0.67302)$ or equivalent to 1.96 times, assuming that other variables are not taken into account in the analysis model. Similarly, an increase in the HDI has an impact on an increase in the average number of unemployed people by $\exp(0.3531)$ or equivalent to 1.42 times, assuming that other variables are not taken into account in the analysis model. Conversely, an increase in the number of job placement/fulfillment services can reduce the average risk of unemployment by $\exp(-0.23735)$ or equivalent to 0.79 times, assuming that other variables are not taken into account in the analysis model.

5. CONCLUSION

The following conclusions can be drawn from the results of this study:

1. The negative binomial regression model used to analyze the number of unemployed people in Sulawesi Island in 2023 can be stated as follows:

$$\ln(\hat{\mu}) = 8.1951 + 0.673X_2 + 0.3531X_4 - 0.2373X_6$$

2. Based on the results of the negative binomial regression analysis, it was found that the variables that had a significant effect on the number of unemployed people on the island of Sulawesi in 2023 were population size, Human Development Index (HDI), and the number of job placement/ fulfillment services.

REFERENCES

- [1] H. Hidayati, H. Nainggolan, R. Erdiansyah, W. Ratri, A. Gorda, N. Prastiwi, S. Kadiman, I. Adnyana, A. Siska, and others, "Ekonomi Sumber Daya Manusia," 2022.
- [2] Rahmawati, "Faktor-faktor yang mempengaruhi pengangguran di provinsi sulawesi selatan tahun 2000–2014," 2016.
- [3] U. Kurniawan, "Penaksiran dan pengujian hipotesis parameter model regresi binomial estimation parameters and testing hypotheses for the bivariate negative binomial regression model: Case study number of infant mortality and maternal," 2015.
- [4] M. Fadil, "Mengatasi overdispersi menggunakan regresi binomial negatif dengan penaksir maksimum likelihood pada kasus demam berdarah di kota makassar," *Pharmacognosy Magazine*, vol. 75, no. 17, pp. 399–405, 2021.
- [5] E. Hayat and A. S. Özden, "Estimating the number of unemployed months for individuals in turkey with the poisson and negative binomial regression models," *Yönetim ve Ekonomi Dergisi*, vol. 30, no. 2, pp. 225–238, 2023.
- [6] Satriani, "Perbandingan model regresi binomial negatif dan generalized poisson untuk data jumlah kematian akibat demam berdarah dengue (dbd) di provinsi sulawesi selatan tahun 2020," pp. 1–80, 2022.
- [7] J. M. Hilbe, *Negative Binomial Regression*, 2011, vol. 11, no. 1.
- [8] A. O. Halim and N. Imro'ah, "Negative binomial regression in overcoming overdispersion poverty data in kalimantan," *Jurnal Forum Analisis Statistik*, vol. 4, no. 1, pp. 1–10, 2024.
- [9] V. M. Santi and Y. Rahayuningsih, "Negative binomial regression in overcoming overdispersion in extreme poverty data in indonesia," *Pattimura International Journal of Mathematics*, vol. 2, no. 2, pp. 43–52, 2023.

- [10] A. Agresti, *An Introduction to Categorical Data Analysis (2nd edn)*. Hoboken, New Jersey: John Wiley & Sons, Inc., 2007.
- [11] N. L. Johnson, *Univariate Discrete Distributions*, 1992.
- [12] R. L. Steiner, “Three estimation methods for the negative binomial parameter k ,” 1993.