Journal of Advanced Technology and Multidiscipline (JATM)
Vol. 03, No. 01, 2024, pp. 9-12
e-ISSN: 2964-6162

Application of Negative Binomial Regression Model in West Java Tourism

Arip Ramadan[1], Dwi Rantini[2]

[1]*Information System Study Program, School of Industrial and System Engineering, Telkom University, Indonesia*
[2]*Data Science Technology Study Program, Department Engineering, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Indonesia*

***Abstract*— Tourism is the most important sector for potential areas. One province in Indonesia that has quite potential is West Java. To be able to increase this potential, it is necessary to increase the number of tourists. This research aims to increase the number of visitors through several factors that must be considered, namely the number of star hotels, number of stalls, and number of restaurants. The method used to determine the relationship between the number of visitors and these factors is regression. Because the research data is in the form of count data, it uses the Poisson distribution. If the data indicates overdispersion, it can be modeled using a Negative Binomial regression model as a comparison. The results obtained state that the Negative Binomial regression model is better than the Poisson regression model. Using the Negative Binomial regression model, it is found that increasing the number of tourists can be achieved by increasing the number of star hotels and number of restaurants. However, the number of stalls must be reduced so that visitors can increase. Thus, the aspect that must be addressed is to disband illegal stalls and increase the number of hotels and restaurants.**

***Keywords*—Tourism, Negative Binomial regression model, overdispersion, West Java, Poisson.**

I. INTRODUCTION

Several sectors can be a source of regional income. One of the regional income sectors is the tourism sector. Tourism can bring in local and foreign tourists. To be able to increase the number of tourists visiting, there should be several supporting factors. Some factors that may have an influence are the number of star hotels, number of stalls, and number of restaurants [1]. Increasing these factors can enable an increase in tourists [2].

Research conducted by Hartanto et al. stated that digital marketing and attraction can increase the number of tourists [3]. Their research was carried out in the tourism in Rejang Lebong Regency [3]. Similar to previous research, research conducted by Popescu et al. also suggested increasing the number of attractions to attract tourist interest [4]. In research conducted by Javier et al. to increase the number of tourists, they provide wine parties [5]. In research conducted by Insani and Arnan, it was found that there was a relationship between hotel taxes and restaurant taxes on the number of tourists [6]. The method they use is growth rate analysis [6]. In research conducted by Nadra et al. using the basic concepts of gastronomic tourism method to analyze tourism conditions on the island of Bali [7]. In research conducted by Roth and Cords, they used the Negative Binomial distribution to model zoo visitors [8].

In the research mentioned previously, there are several shortcomings, including research conducted by Hartanto et al. and Popescu et al., their research only increases the number of attractions to increase the number of tourists [3][4]. In the research conducted by Insani and Arnan the focus was only on taxes, even though if the number of restaurants were increased then tourists would have more choices [6]. In Nadra's research, she only used methods from the social aspect, not using scientific methods [7]. Then research by Roth and Cords only focused on zoos, not tourist attractions in general [8]. In their research, they compared the Negative Binomial and Binomial models [8]. Also, in their research, they examined the influence of zoo visitors on sleep, displacement activities, and affiliative and aggressive behaviors in captive ebony langurs [8]. Therefore, in our research, we will focus on the tourism sector in general by considering the factors that greatly influence the number of visitors. The factors in question are the number of star hotels, the number of stalls, and the number of restaurants. This research aims to model how the relationship between the number of star hotels, the number of stalls, and the number of restaurants influences the number of tourists visiting. To model it, we use regression. Because the data in this study is count data, the Poisson distribution is used. If our data experiences overdispersion, then the alternative distribution that can be used is the Negative Binomial distribution. We will use this method for tourism data in West Java because West Java province is one of the potential provinces in Indonesia.

Journal of Advanced Technology and Multidiscipline (JATM)
Vol. 03, No. 01, 2024, pp. 9-12
e-ISSN: 2964-6162

II. METHODOLOGY

This chapter provides research data and several literature reviews.

A. Dataset

Table 1. Research dataset

Regency /City	Number of Tourist Visits	Number of Star Hotels	Number of Stalls	Number of restaurants
Bogor	6,319,408	33	47	326
Sukabumi	2,767,167	7	43	21
Cianjur	1,988,585	19	193	42
Bandung	1,030,084	21	93	40
Garut	3,874,577	8	150	10
Tasikmalaya	898,599	1	35	2
Ciamis	1,099,443	0	47	5
Kuningan	3,081,084	6	0	6
Cirebon	683,909	7	205	52
Majalengka	1,050,524	1	78	3
Sumedang	1,815,426	9	89	2
Indramayu	836,785	4	54	0
Subang	5,953,952	7	357	3
Purwakarta	1,731,423	3	84	35
Karawang	2,492,910	18	265	190
Bekasi	2,779,981	33	0	200
Bandung Barat	3,480,347	19	43	152
Kota Bogor	2,652,355	43	120	78
Kota Sukabumi	405,863	8	80	13
Kota Bandung	2,923,284	193	211	263
Kota Cirebon	1,833,129	28	0	178
Kota Bekasi	975,195	26	1,325	819
Kota Depok	3,210,633	5	0	1,300
Kota Cimahi	110,767	1	38	28
Kota Tasikmalaya	1369231	11	111	40
Kota Banjar	68,864	0	47	1

The data used in this research is the number of tourist visits in West Java province. This data was taken from the West Java Central Statistics Agency [9]. Number of tourist visits as a response variable Y, then as variables that are thought to be influential are given the number of star hotels, number of stalls, and number of restaurants, denoted by X_1, X_2, and X_3.

B. The Poisson Regression Model

Data obtained by counting is called count data [10]. The appropriate regression for count data is Poisson regression [11]. For $Y \sim Poisson(\mu)$ he Poisson regression model can be written as (1) [12].

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p = \mathbf{x}^T \boldsymbol{\beta} \qquad (1)$$

where $\mathbf{x}^T$ is the vector whose elements are predictor variables, $\boldsymbol{\beta}$ is the vector of parameters, $\mu = E(Y)$, and $g(\square)$ is a link function.

C. Overdispersion Test

Due to heterogeneity in sample values, the existence of outliers, correlated variables, relevant omissions predictor or inflation is zero, causing overdispersion [13]. Overdispersion occurs when the sample variance value is greater than the mean value [14]. For more details regarding the overdispersion test using Pearson's chi-squared test, see [15].

D. The Negative Binomial Regression

The negative binomial regression considers the following link function of the average, μ

$$s(\mu) = \theta = \log\left(\frac{\mu}{k+\mu}\right) = -\log\left(\frac{k}{\mu}+1\right)$$

$$s(\mu) = \eta = \mathbf{x}\boldsymbol{\beta}$$

$$\log\left(\frac{k}{\mu}+1\right) = \eta = \mathbf{x}\boldsymbol{\beta} + \varepsilon$$

where $E(Y) = \mu$, $\text{var}(Y) = \mu + \frac{\mu^2}{k}$, and k^{-1} is called the dispersion parameter [16].

E. Akaike Information Criterion (AIC)

The formula for AIC can be seen in (2) [17].

$$AIC = N \ln\left(\frac{RSS}{N}\right) + 2K \qquad (2)$$

where RSS is the residual sum of squares, N is the number of observations, and K is the number of parameters [17]. The best model is the one with the smallest AIC value.

III. RESULTS AND DISCUSSION

Firstly, in the results of this analysis, we can see the descriptive statistics given in Table 2. From Table 2, it can be seen that the variance is much greater than the mean. Statistically, this is overdispersion.

Table 2. Descriptive statistics of the number of tourist visits

Mean	Median	Variance
2,132,059	1,824,278	2.5308×10^{12}

Visually, to find out the relationship between X_1, X_2, and X_3 against Y, see Figure 1, Figure 2, and Figure 3, respectively. Based on these results, it can be seen that their relationship forms a certain pattern, so there is an indication that X_1, X_2, and X_3 significantly influence Y.

Journal of Advanced Technology and Multidiscipline (JATM)
Vol. 03, No. 01, 2024, pp. 9-12
e-ISSN: 2964-6162

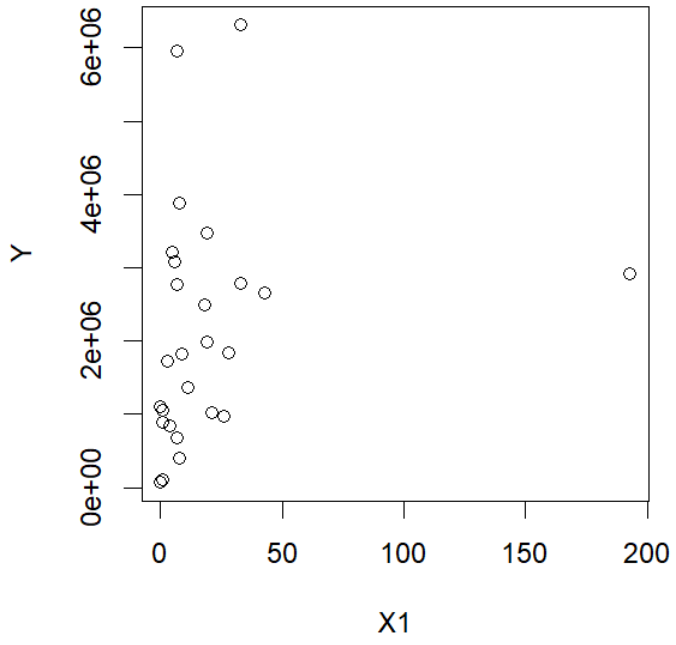

Figure 1. Scatter plot X1 against Y

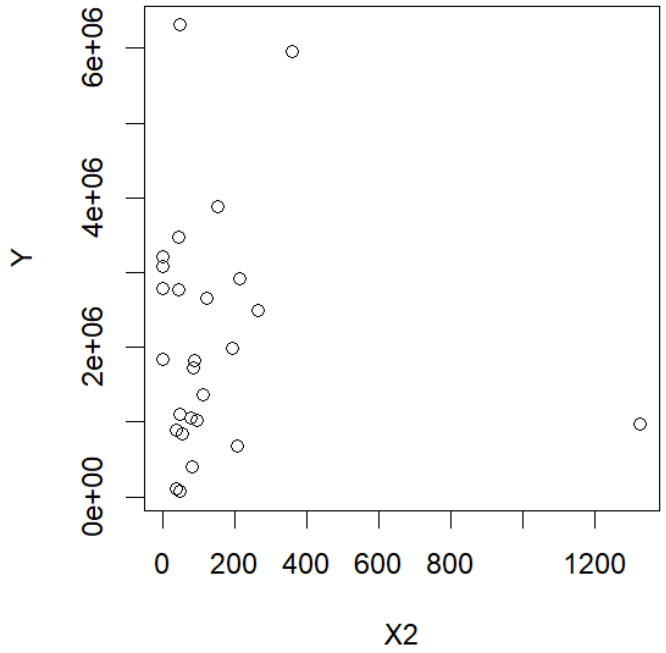

Figure 2. Scatter plot X2 against Y

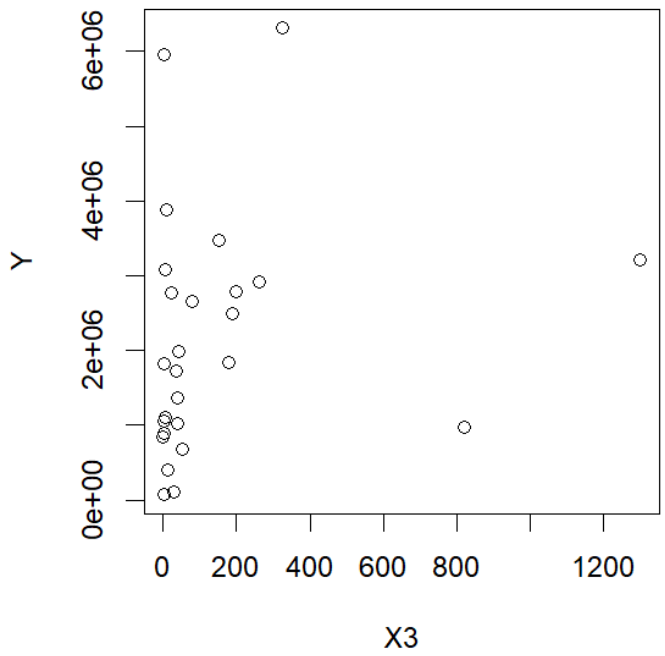

Figure 3. Scatter plot X3 against Y

Continuing from the descriptive statistics in Table 2, the Pearson's Chi-Squared statistical test is given in Table 3. Table 3 shows that the data has overdispersion. This is indicated by a p-value of less than α =5%.

Table 3. Overdispersion test result

dispersion ratio	Pearson's Chi-Squared	p-value
1,325,767.5300	29,166,885.6530	< 0.0010

Table 4. Estimation parameter result of Poisson and Negative Binomial Regression

Estimator	Poisson	Negative Binomial
β_0	14.4800	14.4503
β_1	0.0032	0.0041
β_2	-0.0004	-0.0003
β_3	0.0005	0.0004

The estimation results for the Poisson and Negative Binomial regression models in Table 4 show that there are differences. However, based on Table 5, the Negative Binomial regression model is better because it has a smaller AIC value.

Table 5. AIC value of Poisson and Negative Binomial Regression Model

Poisson	Negative Binomial
26,831,537	815.9800

Journal of Advanced Technology and Multidiscipline (JATM)
Vol. 03, No. 01, 2024, pp. 9-12
e-ISSN: 2964-6162

IV. CONCLUSION

Three conclusions can be drawn from this research. The first conclusion seen from descriptive statistics is that the mean value is not the same as the median, therefore it cannot be modeled with the Normal distribution. Therefore, other distribution alternatives must be sought. Because the response data in this study is count data, the appropriate distribution is Poisson. The second conclusion is that the research data shows overdispersion, whether seen from descriptive statistics or the overdispersion test. Therefore, the data can be modeled with Poisson regression and compared with the Negative Binomial regression model. The third conclusion, from the AIC value, is that the best model is the Negative Binomial regression model. This is because the AIC value of the Negative Binomial regression model is smaller than the Poisson regression model.

REFERENCES

[1] H. Jiang, L. Mei, Y. Wei, R. Zheng, and Y. Guo, "The influence of the neighbourhood environment on peer-to-peer accommodations: A random forest regression analysis," *Journal of Hospitality and Tourism Management*, vol. 51, pp. 105–118, 2022.

[2] A. Mahdi, T. Tettamanti, and D. Esztergár-Kiss, "Modeling the time spent at points of interest based on google popular times," *IEEE Access*, 2023.

[3] Y. Hartanto, M. A. Firmansyah, and L. Adhrianti, "Implementation digital marketing pesona 88 curup in to build image for the decision of visit tourist attraction," in *4th social and humanities research symposium (SoRes 2021)*, Atlantis Press, 2022, pp. 589–594.

[4] A. Popescu, A. C. Hontus, and M. Caratus-Stanciu, "Trends and changes in tourist flow in Romania in the period 2009-2018.," 2020.

[5] J. Martínez-Falcó, B. Marco-Lajara, P. del C. Zaragoza Sáez, and E. Sánchez-García, "Wine tourism in Spain: The economic impact derived from visits to wineries and museums on wine routes," 2023.

[6] M. N. Insani and S. G. Arnan, "Analysis of the contribution of hotel tax and restaurant tax to local revenue of Bandung District during COVID-19 pandemic," *Jurnal Ekonomi*, vol. 12, no. 01, pp. 103–110, 2023.

[7] N. M. Nadra, M. Suardani, I. A. Elistyawati, I. W. Pugra, and T. R. Pamularsih, "Promoting Bali Culinary as Destination Attraction for Domestic Tourist Market," *International Journal of Multicultural and Multireligious Understanding*, vol. 9, no. 11, pp. 380–387, 2022.

[8] A. M. Roth and M. Cords, "Zoo visitors affect sleep, displacement activities, and affiliative and aggressive behaviors in captive ebony langurs (Trachypithecus auratus)," *Acta Ethol*, vol. 23, pp. 61–68, 2020.

[9] "Badan Pusat Statistik Provinsi Jawa Barat." Accessed: May 28, 2024. [Online]. Available: https://jabar.bps.go.id/subject/16/pariwisata.html#subjekViewTab3

[10] T. Hasebe, "Endogenous switching regression model and treatment effects of count-data outcome," *Stata J*, vol. 20, no. 3, pp. 627–646, 2020.

[11] B. Winter and P. Bürkner, "Poisson regression for linguists: A tutorial introduction to modelling count data with brms," *Lang Linguist Compass*, vol. 15, no. 11, p. e12439, 2021.

[12] L. H. Vanegas and L. M. Rondon, "A data transformation to deal with constant under/over-dispersion in Poisson and binomial regression models," *J Stat Comput Simul*, vol. 90, no. 10, pp. 1811–1833, 2020.

[13] L. P. Fávero, R. de F. Souza, P. Belfiore, H. L. Corrêa, and M. F. C. Haddad, "Count data regression analysis: concepts, overdispersion detection, zero-inflation identification, and applications with R," *Practical Assessment, Research, and Evaluation*, vol. 26, no. 1, p. 13, 2021.

[14] J. Kruppa and L. Hothorn, "A comparison study on modeling of clustered and overdispersed count data for multiple comparisons," *J Appl Stat*, vol. 48, no. 16, pp. 3220–3232, 2021.

[15] H. Li, W. Luo, and E. Baek, "Multilevel modeling in single-case studies with zero-inflated and overdispersed count data," *Behav Res Methods*, pp. 1–17, 2024.

[16] S. García-Bustos, N. Cárdenas-Escobar, A. Debón, and C. Pincay, "A control chart based on Pearson residuals for a negative binomial regression: application to infant mortality data," *International Journal of Quality & Reliability Management*, vol. 39, no. 10, pp. 2378–2399, 2022.

[17] S. Portet, "A primer on model selection using the Akaike Information Criterion," *Infect Dis Model*, vol. 5, pp. 111–128, 2020.

Arip Ramadan is a lecturer at the Information System Study Program, School of Industrial and System Engineering, Telkom University, Indonesia
(e-mail: aripramadan@telkomuniversity.ac.id).

Dwi Rantini is a lecturer at the Data Science Technology Study Program, Department of Engineering, Faculty of Advanced Technology and Multidiscipline, Universitas Airlangga, Indonesia.
(e-mail: dwi.rantini@ftmm.unair.ac.id)