Cek Similarity/9. Paper ICoMCoS (published 26 februri 2021).pdf

By Amin Tohari

WORD COUNT 104 TIME SUBMITTED 14-FEB-2022 09:55AM
PAPER ID 83600912

Modeling the number of confirmed and suspected cases of Covid-19 in East Java using bi-response negative binomial regression based on local linear estimator

Cite as: AIP Conference Proceedings **2329**, 060022 (2021); https://doi.org/10.1063/5.0042288
Published Online: 26 February 2021

Amin Tohari, Nur Chamidah, and Fatmawati

Modeling the Number of Confirmed and Suspected Cases of Covid-19 in East Java Using Bi-response Negative Binomial Regression Based on Local Linear Estimator

Amin Tohari[1, a)], Nur Chamidah[2, b)] and Fatmawati[2, c)]

[1]Faculty of Economic and Business, University of Nusantara PGRI Kediri, Indonesia
[2]Department of Mathematics, Faculty of Science and Technology, Airlangga University

[a)]*amin.tohari-2016@fst.unair.ac.id*
[b)]*Corresponding author: nur-c@fst.unair.ac.id*
[c)]*fatmawati@fst.unair.ac.id*

Abstract. The number of confirmed and suspected cases of Covid-19 are type of count data and they correlate each other. A popular regression model of two response variables for count data is bi-response Poisson regression. However, assumptions violation of Poisson regression that frequently occurs is over-dispersion. Negative binomial regression can overcome this over-dispersion case. The goal of this research is to model the number of confirmed and suspected Covid-19 cases affected by population density using bi-response negative binomial regression based on local linear estimator. The proposed method gave the optimal bandwidth of 609 based on maximum likelihood cross validation criterion, with deviance value of 1.537 which is less than 27.083 of the parametric regression approach. It means that the estimated model of the number of confirmed and suspected cases of Covid-19 affected by population density using bi-response negative binomial regression based on local linear estimator is better than the parametric model approach.

INTRODUCTION

Recently, the world population has panicked at the outbreak of Coronavirus disease in 2019 (Covid-19). Covid-19 allegedly first contacted humans in Wuhan, China on 17 November 2019, then reportedly has begun to spread to Europe, America, and even to the whole world since December 2019. The Covid-19 case was officially announced as a global pandemic by World Health Organization (WHO) on 11 March 2020 [1].

In Indonesia, the finding of the first Covid-19 case was confirmed on 2 March 2020. The Indonesian government immediately follows the pandemic with restricting movement into and abroad to movement between the island and applying a pattern of working from home broadly since 16 March 2020. The researchers also react to a pandemic by conducting various research, such as [2] estimated Covid-19 using Richard's Curve; [3] conducted research a mathematical study on the spread of Covid-19 considering social distancing and rapid assessment; modeling Covid-19 transmissions in Wuhan, Diamond Princess by [4], and Jakarta-cluster; and [5] conducted study to analyze the correlation between weather and Covid-19 pandemic in Jakarta.

East Java Province as one of the region most populous in Indonesia did not escape the terror Covid-19, even since the end of May 2020, East Java province frequently notes the addition of daily cases highest in Indonesia. The increase in confirmed and suspect case in Indonesia especially in East Java cannot be separated from a couple of things, one of which is the density of population in the region of East Java. According to [6], there is a correlation between the population with the epidemics. The population density is closely related to the mobility of the population, the increasing mobility of people, especially if it occurs at the same time will improve and expand the geographical risk of Covid-19. The sheer density of the population in large cities provides an ideal environment for infections to erupt, and fast [7].

International Conference on Mathematics, Computational Sciences and Statistics 2020
AIP Conf. Proc. 2329, 060022-1–060022-8; https://doi.org/10.1063/5.0042288
Published by AIP Publishing. 978-0-7354-4073-9/$30.00

Regression analysis can be used to model the number of confirmed cases and suspect cases with population density. Regression analysis is used to analyze the functional relationship between response and predictors variables. For this analysis, not all response variables are continuous, but there are still discrete response variables. For discrete response variables, Poisson regression model with specific assumptions was used, for example, in the standard Poisson regression model mean and variance of response variable is the same. But, in the reality, this assumption is often misleading because the variance could be less than the mean called as under-dispersion, and reverse, called as over-dispersion. The negative binomial regression can be the solution to this over-dispersion problem [8]. Data of the number of confirmed and suspect cases follows the bivariate Poisson distribution. But, a few count data frequently shows over-dispersion conditions. Therefore, in this study, bi-response negative binomial regression model is used to model the number of confirmed and suspect case Covid-19. We know that there are two approaches in regression analysis. The parametric regression approach will be used if the shape of the regression function is identified and the relationship between the response and the predictor variables follows certain curves. It is only assumed to be continuous and differentiable.

Nonparametric regression models based on local linear estimator in the case of bi-responses negative binomial distribution has not been developed in statistical modeling. Until now, researchers have developed regression with parametric approach. Some of these studies include [9–12]. Nonparametric approaches with continuous responses in bi-response and multi-response cases have also been developed by several researchers. They are [13,14], [15–20], and [21–25]. Whereas the nonparametric regression of discrete responses is still limited to the Poisson distribution response variable, such studies include [28] used the kernel estimator; [29] used spline estimator; [30] used kernel estimator; [31] used local linear estimator; and [32] used local polynomial estimator in generalized Poisson regression. Lately, it had developed modeling using negative binomial regression based on local linear approach, i.e. [31,32]. The objective of this study is to estimate the regression function which describes relationship between response variables (the number of confirmed and suspected Covid-19 cases) and predictor variable (the population density) in East Java.

MATERIAL AND METHODS

The data used in this research are secondary data collected from East Java Provincial Health Office, namely confirmed and suspect case Covid-19 dated on 21 July 2020. East Java Province consist-of the 29 districts and 9 cities. In this study, the data are decomposed becomes in-sample and out-sample data. Data in-sample is data from 31 districts or cities in East Java Province which is used for modeling and 7 other districts or cities were used as out-sample data is selected randomly and used for testing models. Steps used to estimate the number of confirmed and suspect cases Covid-19 in East Java with population density as a predictor variable are as follows, the first step is to conduct a correlation test among both response variables. Suppose Y_1 and Y_2 are two response variable, so the correlation between the variables can be expressed as follows:

$$r_{y_1y_2} = \frac{\sum_{i=1}^{n}\left(y_{1i}-\bar{y}_1\right)\left(y_{2i}-\bar{y}_2\right)}{\sqrt{\sum_{i=1}^{n}\left(y_{1i}-\bar{y}_1\right)^2}\sqrt{\sum_{i=1}^{n}\left(y_{2i}-\bar{y}_2\right)^2}}, \tag{1}$$

if the two response variables are significantly correlated, bi-response negative binomial regression modeling can be done.

Then, to estimate bi-response negative binomial regression based on local linear estimator approach. Suppose that the $\left(x_i, y_{1i}, y_{2i}\right)$ data is paired, with $i = 1,2,...,n$ and n is the amount of observation data. Regression models for the response variable y_{1i} and y_{2i} which are discrete random variable type and it is assumed that both response variables have a bivariate negative binomial distribution as follows [33]:

$$f\left(y_{1i}, y_{2i} \mid x_i\right) = \frac{\Gamma\left(\frac{1}{\alpha}+y_{1i}+y_{2i}\right)}{\Gamma\left(\frac{1}{\alpha}\right)\Gamma\left(y_{1i}+1\right)\Gamma\left(y_{2i}+1\right)}\left[\mu_1\left(x_i\right)\right]^{y_{1i}}\left[\mu_2\left(x_i\right)\right]^{y_{2i}} \times \alpha^{\frac{1}{\alpha}}\left(\frac{1}{\alpha}+\mu_1\left(x_i\right)+\mu_2\left(x_i\right)\right)^{-\left(\frac{1}{\alpha}+y_{1i}+y_{2i}\right)} \tag{2}$$

$$\text{where} \quad \mu_1\left(x_i\right) = \exp\left(x_i^T \underset{\sim}{\beta}_1\right) \tag{3}$$

$$\mu_2(x_i) = \exp\left(x_i^T \underset{\sim}{\beta}_2\right). \tag{4}$$

The parameter $\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$ in this study will be estimated using the method of locally weighted maximum likelihood estimator. Parameter estimation using the weighted local likelihood function with the kernel function. The model based on the likelihood function in which an estimator for the regression curve is constructed using the maximum likelihood method, Chauduri and Dewanji [36] introduced Maximum Likelihood Cross Validation (MLCV). If $\hat{f}_{-i}(x)$ is an estimate of the regression function at point x_i without including the i^{th} data, then MLCV is a function of the bandwidth h given in the following equation:

$$MLCV_h = \sum_{i=1}^{n} \ln f(y_i, \hat{f}_{-i}(x)) \tag{5}$$

The optimal bandwidth to be used in the best model is the one that gives the largest MLCV value. The optimal bandwidth selection procedure with MLCV has been applied by Staniswalis [37] in the nonparametric regression model with discrete response variables and gives excellent results in selecting the best model.

Bi-respon negative binomial regression estimator obtained by maximizing the locally likelihood function on the following equation:

$$\begin{aligned} L\left(\underset{\sim}{\beta}_1, \underset{\sim}{\beta}_2, \alpha, x_0\right) = \sum_{i=1}^{n} K_h(x_i - x_0) \Bigg\{ & \ln\Gamma\left(\frac{1}{\alpha} + y_{1i} + y_{2i}\right) - \ln\Gamma\left(\frac{1}{\alpha}\right) - \ln\Gamma\left(y_{1i}+1\right) - \ln\Gamma\left(y_{2i}+1\right) + y_{1i}\ln\left(\mu_1\left(\underset{\sim}{x}_i\right)\right) \\ & + y_{2i}\ln\left(\mu_2\left(\underset{\sim}{x}_i\right)\right) - \frac{1}{\alpha}\ln\alpha - \left(\frac{1}{\alpha} + y_{1i} + y_{2i}\right)\ln\left(\frac{1}{\alpha} + \mu_1\left(\underset{\sim}{x}_i\right) + \mu_2\left(\underset{\sim}{x}_i\right)\right)\Bigg\} \end{aligned} \tag{6}$$

The first derivative of parameters $\underset{\sim}{\beta}_1$ and $\underset{\sim}{\beta}_2$ do not provide explicit equations, so for estimating them require an iterative method. Cameroon and Trivedi [38], stated that the commonly used procedure is *iteratively reweighted least square* (IRLS). This procedure is performed by newton raphson method. Newton raphson method is a popular method to solve nonlinear equations in determining approximations to the root of the real function. Newton raphson iteration is an algorithm that utilizes first-order derivative vectors and second-order derivative matrices from maximized likelihood functions. Newton raphson method convergen quickly, especially when iteration begins quite close to the desired root [37].

Goodness of fit used in the bi-response negative binomial regression model is deviance value [36]. The formula used to calculate deviance are as follows:

$$D\left(\hat{\underset{\sim}{\beta}}\right) = 2\left[L\left(\hat{\Omega}\right) - L\left(\hat{\omega}\right)\right], \tag{7}$$

where $L\left(\hat{\Omega}\right)$ is the likelihood function under the complete model involving predictor variables, and $L\left(\hat{\omega}\right)$ is the likelihood function for a simple model without involving predictor variables. If $D\left(\hat{\underset{\sim}{\beta}}\right) < \chi^2_{(\alpha, n-p-1)}$ then the model is appropriate.

RESULT AND DISCUSSION

Bi-response negative binomial regression model can be used if there is a significant correlation between the two responses. The analysis showed that p-values of pearson correlation test between the two responses is less than 0.001, and then H_0 is rejected to all $1\% \le \alpha \le 10\%$. Therefore, the number of Confirmed and suspect cases Covid-19 can be modeled together using bi-response negative binomial regression. Fig. 1 is a plot of two response variables and a predictor variable generated with OSS R. Based on Fig. 1, it can be assumed that the population density variable has a pattern of relationship that spreads to the number of confirmed and suspect case Covid-19 or we can indicate that this does not follow a specific pattern, so the appropriate approach for modeling is to use nonparametric approach.

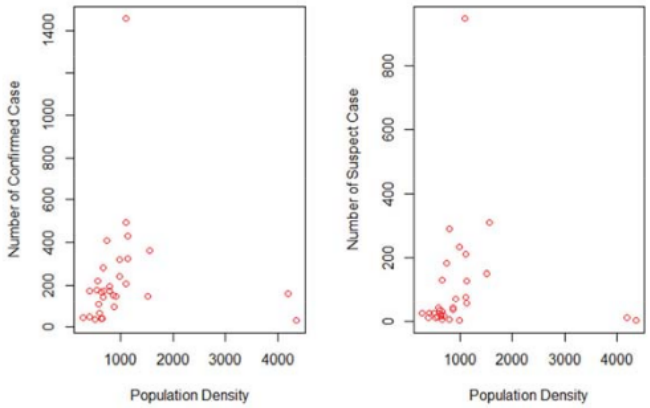

FIGURE 1. Scatter plot between both of response variables with 1 predictor variable

In the estimation of nonparametric regression models based on local linear estimators, optimal bandwidth values are needed. The determination of the optimal bandwidth was based on the maximum value from the Maximum Likelihood Cross-Validation (MLCV) criteria. Table 1 illustrates the bandwidth and MLCV values, which are summarized from the output of R-code.

TABLE 1. Bandwidth and MLCV values

Bandwidth	MLCV
606	12.21393
607	12.21402
608	12.21408
609	**12.21409**
610	12.21405
611	12.21397
612	12.21384

The plot between bandwidth and MLCV was shown in the Fig. 2.

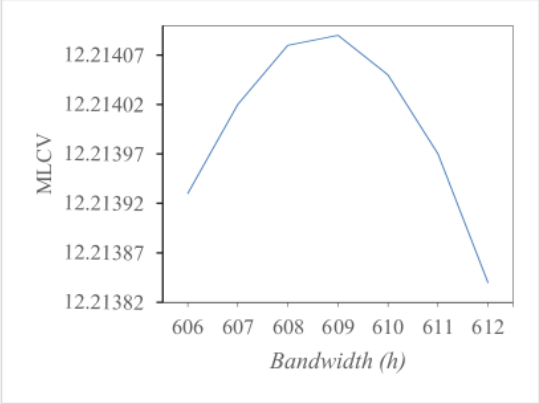

FIGURE 2. Plot between bandwidth and MLCV

Based on Table 1 and Fig. 2, the optimal bandwidth is 609, with a maximum MLCV of 12.21409. After successfully determining the optimal bandwidth value, we can estimate model using bi-response negative binomial regression based on local linear estimator approach. The output model has different coefficients depending on the location. As an illustration, an example of bi-response negative binomial regression model is given for in-sample and out-sample data.

For in-sample of data, the results of estimation for bi-response negative binomial regression model using local linear approach to the number of confirmed and suspect cases Covid-19 in Mojokerto district are as follows:

$$\hat{\mu}_1 = \exp\left(5.818 + 0.00040(x - 1557)\right), 948 < x < 2166 \tag{8}$$

$$\hat{\mu}_2 = \exp\left(5.037 + 0.00042(x - 1557)\right), 948 < x < 2166 \tag{9}$$

Mojokerto district was selected as an example in the interpretation of the model, because Mojokerto is one buffer Surabaya with the highest number of confirmed cases Covid-19 in East Java. Many Mojokerto residents are on the move to and from Surabaya. Based on equation (8), we can interpret for every addition of a hundred population density in district of Mojokerto will give a change in the number of Confirmed cases by 1,040 times compared to the previous case. Based on equation (9), we can interpret for every addition of a hundred population density in district of Mojokerto will provide a change in the number of Suspect cases by 1.042 times compared to the previous case.

For out-sample of data, results of estimation the model obtained in the city of Surabaya are as follows:

$$\hat{\mu}_1 = \exp\left(5.412 + 0.00002(x - 8262)\right), 7653 < x < 8871 \tag{10}$$

$$\hat{\mu}_2 = \exp\left(5.061 + 0.00012(x - 8262)\right), 7653 < x < 8871 \tag{11}$$

Surabaya was chosen as an example of interpretation, because Surabaya is a city in East Java with the highest number of confirmed and suspect cases Covid-19. Based on equation (10), we can interpret for every addition of a hundred population density in Surabaya will give a change in the number of confirmed cases by 1,002 times compared to the previous case. Based on equation (11), we can interpret for every addition of a hundred population density in Surabaya will provide a change in the number of Suspect cases by 1.012 times compared to the previous case.Goodness of fit test on bi-response negative binomial regression model by using local linear estimator for the number of Confirmed and suspect cases in East Java with population density as a predictor variable can be seen from the deviance value, that is equal to 1.537 which is smaller than the $\chi_{(\alpha,2*n-(p+1))}$ of 75.624. It means that the model is appropriate. Next, we also estimate the number of Confirmed and suspect cases in East Java by using parametric regression model approach. Thus, we get the estimated model as follows:

$$\hat{\mu}_1 = \exp\left(5.412 + 0.000015x\right) \text{ and } \hat{\mu}_2 = \exp\left(4.520 + 0.00016x\right) \tag{12}$$

based on the estimated model, we obtain deviance value of 27.083.

Based on equations (8) and (9), we can plot the estimated results for the both response variables with the observations shown in Figs. 3, where the red dot is the observation, the blue line is the parametric estimated results, and the green line is nonparametric estimated results.

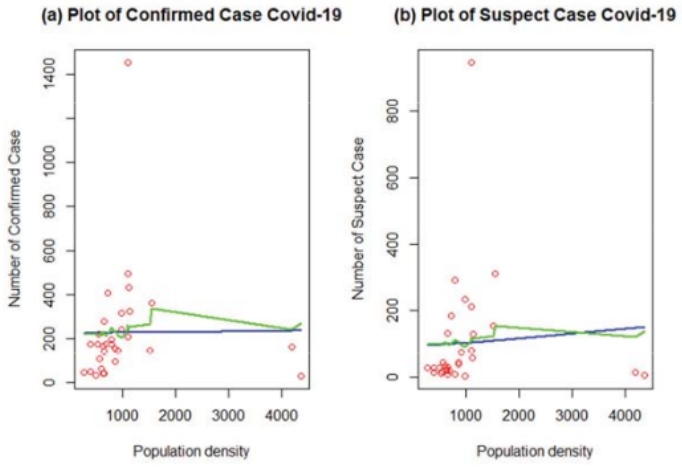

FIGURE 3. Plot of observational data and estimates of Confirmed case and Suspect cases for in-sample data

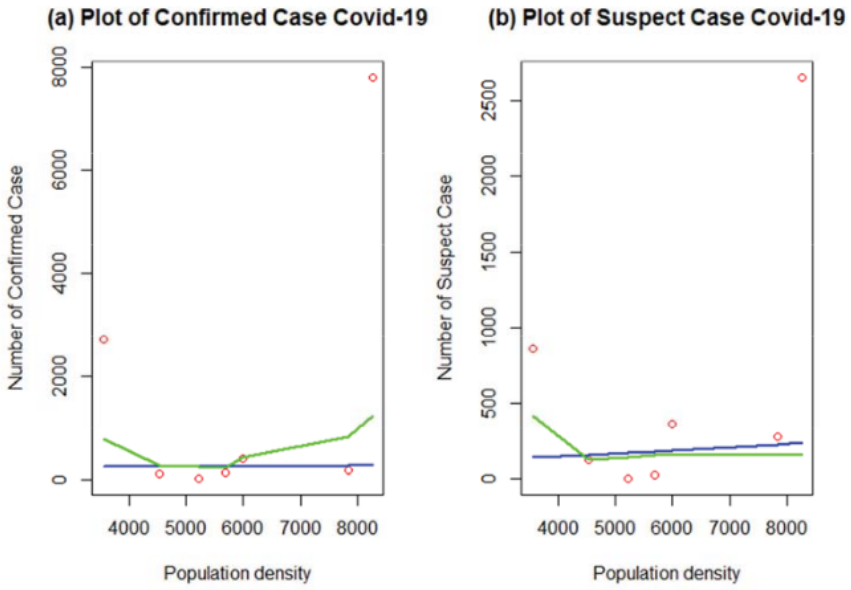

FIGURE 4. Plot of observational data and estimates of Confirmed case and Suspect cases for out-sample data

Based on Fig. 3 and Fig. 4, the estimated nonparametric by using local linear estimator (green line) is closer to the observations data (red dot) than estimated parametric (blue line). In addition to a visual, comparative estimate of the negative binomial regression models by using parametric and nonparametric regression approaches can also be determined based on deviance value. The comparison of deviance values for parametric and nonparametric regression approaches are shown in Table 2.

TABLE 2. Comparison of deviance for parametric and nonparametric regression approaches

Regression approach	Deviance	$\chi^2_{(0.05,58)}$
Parametric	27.083	75.624
Nonparametric	1.537	

Table 2 presents that based on goodness of fit criterion, i.e., deviance value of the estimated model by using nonparametric regression approach based on local linear estimator of 1.537, which is less than the deviance value of the estimated model by using parametric regression approach of 27.083. It means that bi-response Negative binomial regression by using local linear estimator is more appropriate compared with the estimation results using the parametric regression approach.

CONCLUSION

We founded that estimated the number of confirmed and suspect cases of Covid-19 in East Java affected by population density using bi-response negative binomial regression based on the local linear estimator satisfied the goodness fit of criterion because its deviance value is less than chi-square value. Thus it can be concluded that the estimated model used is appropriate. Also, the estimated model by using nonparametric regression approach based on local linear estimator is better than by using parametric regression approach because it has smaller deviance value.

ACKNOWLEDGEMENT

Authors thank to Director of the Directorate of Research and Public Service, the Directorate General Reinforcing of Research and Development, the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia for financial support of this research through the doctoral dissertation research grant in the fiscal year 2020 with Contract Number: 823/UN3.14/PT/2020

REFERENCES

1. Tedros, "Coronavirus confirmed as pandemic by World Health Organization," https://www.bbc.com/news/world-51839944.
2. N. Nuraini, K. Khairudin, and M. Apri, Commun. Biomath. Sci. **3**(1), 1–8 (2020).
3. D. Aldila, S. H. A. Khoshnaw, E. Safitri, Y. R. Anwar, A. R. Q. Bakry, B. M. Samiadji, D. A. Anugerah, M. F. A. GH, I. D. Ayulani, and S. N. Salim, Chaos, Solitons and Fractals **139**, 1–14 (2020).
4. E. Soewono, "On the analysis of Covid-19 transmission in Wuhan, Diamond Princess and Jakarta-cluster," Commun. Biomath. Sci. **3**(1), 9–18 (2020).
5. R. Tosepu, J. Gunawan, D. S. Effendy, L. O. A. I. Ahmad, H. Lestari, H. Bahar, and P. Asfian, Sci. Total Environ. **725**, 1–4 (2020).
6. R. Li, P. Richmond, and B. M. Roehner, Phys. A Stat. Mech. its Appl. **510**, 713–724 (2018).
7. D. Desai, *Urban Densities and the Covid-19 Pandemic : Upending the Sustainability Myth of Global Megacities* (Observer Research Foundation, 2020), (May).
8. J. M. Hilbe, *Negative Binomial Regression*, Second edition (Cambridge University Press, 2011).
9. K. Månsson, Econ. Model. **29**(2), 178–184 (2012).
10. M. M. Husain and M. S. H. Bagmar, Glob. J. Quant. Sci. **2**(4), 22–29 (2015).
11. T. Hall and A. P. Tarko, Accid. Anal. Prev. **128**(March), 148–158 (2019).
12. A. Tohari, N. Chamidah, and Fatmawati, *IOP Conf. Ser. Mater. Sci. Eng.* **546**, 1–6 (2019).
13. N. Chamidah and T. Saifudin, Appl. Math. Sci. **7**(37–40), 1839–1847 (2013).
14. N. Chamidah and B. Lestari, Far East J. Math. Sci. **100**(9), 1433–1453 (2016).
15. N. Chamidah and M. Rifada, Far East J. Math. Sci. **99**(8), 1233–1244 (2016).
16. N. Chamidah and M. Rifada, *AIP Conference Proceedings* (2016), pp. 1–7.
17. B. Lestari, D. Anggraeni, and T. Saifudin, Far East J. Math. Sci. **108**(2), 341–355 (2018).
18. B. Lestari, Fatmawati, I. N. Budiantara, and N. Chamidah, *J. Phys. Conf. Ser.* **1097**, 1–9 (2018).
19. B. Lestari, Fatmawati, and I. N. Budiantara, *Proceeding of Global Conference on Engineering and Applied Science (GCEAS) Sapporo, Hokkaido, Japan* (2019), pp. 71618, 81–93.
20. N. Murbarani, Y. Swastika, A. Dwi, B. Aris, and N. Chamidah, J. Stat. Its Appl. **3**(2), 139–147 (2019).
21. N. Chamidah, B. Zaman, L. Muniroh, and B. Lestari, *Proceeding of Global Conference on Engineering and Applied Science (GCEAS) Sapporo, Hokkaido, Japan* (2019), pp. 71618: 68–78.
22. N. Chamidah, K. H. Gusti, E. Tjahjono, and B. Lestari, TELKOMNIKA (Telecommunication Comput. Electron. Control. **17**(3), 1492 (2019).
23. W. Ramadan, N. Chamidah, B. Zaman, L. Muniroh, and B. Lestari, *IOP Conf. Ser. Mater. Sci. Eng.* **546**, 1–8 (2019).
24. A. Puspitawati and N. Chamidah, *IOP Conf. Ser. Mater. Sci. Eng.* **546**, 1–6 (2019).
25. A. Islamiyati, Fatmawati, and N. Chamidah, Songklanakarin J. Sci. Technol. **42**(4), 1–13 (2020).
26. J. Shim and C. Hwang, J. Korean Stat. Soc. **40**(1), 1–9 (2011).
27. H. Lian, J. Meng, and K. Zhao, J. Multivar. Anal. **141**, 81–103 (2015).
28. C. S. Chee, Comput. Stat. Data Anal. **109**, 34–44 (2017).
29. Darnah, M. I. Utoyo, and N. Chamidah, *IOP Conf. Ser. Earth Environ. Sci.* **243**, 1–7 (2019).
30. E. T. Astuti, I. N. Budiantara, S. Sunaryo, and M. Dokhi, Int. J. Appl. Math. Stat. **33**(3), 92–101 (2013).
31. A. Tohari, N. Chamidah, and Fatmawati, Ann. Biol. **36**(2), 215–219 (2020).
32. A. Tohari, N. Chamidah, and Fatmawati, *AIP Conference Proceedings* (2020), pp. 1–7.

33. S. Cheon, S. H. Song, and B. C. Jung, J. Korean Stat. Soc. **38**(2), 185–190 (2009).
34. P. Chaudhuri and A. Dewanji, Stat. Probab. Lett. **22**(1), 7–15 (1995).
35. J. G. Staniswalis, J. Am. Stat. Assoc. **84**(405), 276–283 (1989).
36. A. C. Cameron and P. K. Trivedi, *Regression Analysis of Count Data* (Cambridge University Press, 1998).
37. J. C. Ehiwario and S. O. Aghamie, IOSR J. Eng. **4**(4), 01–07 (2014).

Cek Similarity/9. Paper ICoMCoS (published 26 februri 2021).pdf

ORIGINALITY REPORT

11%
SIMILARITY INDEX

PRIMARY SOURCES

1	"Rotation effects on coupled heat and mass transfer by unsteady MHD free convection flow in a porous medium past an infinite inclined plate", 'AIP Publishing', 2014 Internet	13 words — 11%

EXCLUDE QUOTES	OFF	EXCLUDE MATCHES	OFF
EXCLUDE BIBLIOGRAPHY	OFF		