Risk Analysis Forecasting Models of Poisson Regression, Negative Binomial Regression, Poisson GSARIMA, and Negative Binomial GSARIMA (Case Study: Number of Bicycle Sales)

Windya Harieska Pramujati
Department of Statistics, YPPI Rembang University, Rembang, Indonesia
*Corresponding author: windyaharieska@gmail.com

Received: 29 March 2024
Revised: 15 June 2024
Accepted: 19 June 2024

ABSTRACT – The Poisson model is a model that can be applied to count data, where in this research the case study used is the number of bicycle sales. However, there is an equidispersion assumption in the Poisson model, that the response variable has the same mean and variance. A more flexible model is needed if the equidispersion assumption is not met, namely the Negative Binomial model. In this research, two models were applied, namely the regression model and the GSARIMA model, with two different distributions, namely the Poisson distribution and the Negative Binomial distribution. Therefore the models that will be compared are the Poisson Regression, Negative Binomial Regression, Poisson GSARIMA, and Negative Binomial GSARIMA models. The differences in results for each model are due to errors that occur in each model used. Hence, a model with a smaller error can be said to be a model that has a smaller risk than other models. The results of this study show that the error rate in the Negative Binomial GSARIMA ZQ1 model is relatively smaller than other models with a value of AIC = 1058.7. This model is the best model that can be used as a forecasting model in the case of bicycle sales and can minimize the risk of error in a forecasting result.

Keywords– Poisson Regression, Negative Binomial Regression, GSARIMA

I. INTRODUCTION

Bicycle is one of the transportations and can also be used as sports facilities. Based on "Big Data di Tengah Masa Adaptasi Kebiasaan Baru Badan Pusat Statistik", the number of bicycle sales has increased [1]. This is partly due to WHO's recommendation to exercise by bicycle during the pandemic. The number of bicycle sales data is count data, where count data is the data in the discrete domain and a positive integer {0,1,2,3,…} [2]. In count data, the data usually does not spread normally, therefore a forecasting model in the non-Gaussian form must be applied.

McCullagh and Nelder developed the Generalized Linear Models (GLM) model to analyze the relationship between response variables and predictor variables, where the response variable does not have to be normally distributed (Gaussian), but is included in the exponential family [3]. Then Cameron, et al (1998) made a regression analysis of count data [4]. The Poisson distribution is an exponential family distribution, but there is an equidispersion assumption, that the mean value of the response variable is the same as the variance value. The most frequent violation of equidispersion is overdispersion, in which the response variable has a greater variance than the mean. Overdispersion cases can be overcome with other distributions such as Negative Binomial distribution. Benjamin, et al (2003) developed the Generalized Autoregressive Moving Average model for data that follows non-Gaussian distributions such as the Poisson distribution and the Negative Binomial distribution [5].

Previous research regarding the application of the Poisson and Negative Binomial models is the application of the Poisson regression model to large frequency data in PSTP [6]. Funda H, et al (2004) and Haibin Liu, et al (2005) respectively applied the Negative Binomial regression model to data on the number of labor strikes and the number of power outages [7, 8] with Negative Binomial model have realistic results to overcome data overdispersion. Furthermore, Briet, et al (2013) developed the GARMA model for seasonal data with a differencing, namely the GSARIMA model for the number of malaria sufferers in Sri Lanka [2]. In Indonesia, Asrirawan (2015) compared the GSARIMA and SARIMA models on the number of dengue fever sufferers in Surabaya [9]. Then in the same year, Mada Aqil and Agil Desti estimated the parameters of the Negative Binomial and Poisson GARMA forecasting models using the IRLS algorithm [10, 11]. Furthermore, Wardhani, LP et al (2020) researched forecasting the number of theft crimes in the Surabaya Polrestabes area using the Poisson GARMA model and the Negative Binomial GARMA model [12]. In this research, the best model obtained was the GARMA Negative Binomial model. Then in the same year, T Kim, et al applied the Poisson regression model to predict Covid-19 cases in the United States [13].

DOI: 10.12962/j27213862.v7i2.20255

The model will be applied that involves seasonal elements in one of the count data, namely the number of bicycle sales data. Then a regression model is applied with several factors that influence bicycle sales with two different distributions, namely Negative Binomial distribution and Poisson distribution. Therefore, the models applied are the Poisson Regression, Negative Binomial Regression, Poisson GSARIMA, and Negative Binomial GSARIMA. The application of forecasting to the four models will be compared with the criteria for selecting the best model. Then analyze how errors occur in the model. The model with the smaller error will be chosen as the best model that can minimize the risk of error in a forecasting result.

II. LITERATURE REVIEW

A. Poisson Model and Negative Binomial Model for Count Data

Generalized Linear Models (GLM) is a development of the classical linear model, where in this model the response variable follows an exponential family distribution [3]. In this research, the exponential family distributions used are the Negative Binomial distribution and the Poisson distribution. The Negative Binomial distribution has the following exponential family distribution function as follows [9]:

$$f(y_t;\mu_t,k) = exp\left\{ln\left(\frac{\Gamma(y_t+1/k)}{y_t!\,\Gamma(1/k)}\right)+y_t ln\left(\frac{k\mu_t}{k\mu_t+1}\right)+\frac{1}{k}ln\left(\frac{1}{k\mu_t+1}\right)\right\} \quad (1)$$

The Poisson distribution has an exponential family distribution function as follows:

$$f(y_t;\mu_t) = exp\{y_t ln\mu_t - \mu_t - lny_t!\} \quad (2)$$

where the link function g(.) in the Poisson distribution and Negative Binomial distribution is:

$$g(\mu_t) = \ln(\mu_t) = \mathbf{X}^{\boldsymbol{T}}\boldsymbol{\beta} \quad (3)$$

B. Regression Model

The Poisson regression model has been stated in the research of Famoye et al [14], and then compared with another distribution model, namely the Negative Binomial regression model [15]. Then the parameter estimation is carried out for each model using the Maximum Likelihood (MLE) method. The dispersion parameters in the Negative Binomial regression model are also estimated, and then a forecasting model is formed.

C. Generalized Seasonal Autoregressive Integrated Moving Average Model (GSARIMA)

The GSARIMA model used in this research is a model formulated in the research of Briet et al [2]. In this research, two GSARIMA models were applied, namely the GSARIMA model with ZQ1 transformation and the GSARIMA model with ZQ2 transformation. The value of y_t given in the ZQ1 transformation GSARIMA model is $y_t' = max(y_t, c)$ where $0 < c \leq 1$, where c is the threshold parameter. The GSARIMA model with ZQ2 transformation is a model that adds a c value to each value of y_t.

D. Best Model Selection Criterion

The criteria for selecting the best model used in this research is Akaike's Information Criterion (AIC) [9], where the best prediction model will be selected using this criterion.

III. RESULTS AND DISCUSSIONS

A. Research Variable

The data used in this research is secondary data from publications by "Badan Pusat Statistika" and Grand View Research regarding Bicycle Marker Analysis. The data obtained is the mountain bike sales (Y) with factors that influence it (X) for each month in 2016 - 2020. The independent variables (X) in this study are the inflation rate (X_1), household consumption index for sports equipment (X_2), consumer price index (X_3), and average daily wage (X_4). Before forming the regression model, a multicollinearity test was carried out on each independent variable. The results of the multicollinearity test are presented in the Table 1.

Table 1 Multicollinearity Test Results

Variable	Tolerance	VIF
X_1	0.980	1.021
X_2	0.965	1.036
X_3	0.312	3.208
X_4	0.309	3.233

DOI: 10.12962/j27213862.v7i2.20255

Table 1 shows the result that each independent variable has a $tolerance\ value > 0.1$ and a $VIP\ value < 10$, which means that multicollinearity does not occur.

B. *Determination of a Poisson Regression Model*

In the Poisson regression model, bicycle sales data is assumed to have a Poisson distribution. Then the first parameter significance test, namely the simultaneous test, obtained the value of each $\boldsymbol{deviance} > \chi^2_{(\mathbf{4,0.05})} = \mathbf{9.488}$, then $\boldsymbol{H_0}$ was rejected, in other words, there was at least one significant variable. Then the second parameter significance test which can be partially seen from the value of $|\boldsymbol{Z_{hit}}|$ in Table 2.

Table 2 Parameter Estimation of Poisson Regression Model

Parameter	Estimation Value	SE	Z_{hit}	$p\ value$
β_0	-8.376	0.047	-179.559	0.0000
β_1	0.095	0.000	391.601	0.0000
β_2	0.004	0.002	2.393	0.0167
β_3	0.044	0.000	75.526	0.0000
β_4	0.000	0.000	213.613	0.0000

Based on the Table 2, each parameter has a value of $|\boldsymbol{Z_{hit}}| > \boldsymbol{Z_{\alpha/2}} = \mathbf{1.96}$, in other words, each parameter has a partially significant effect on the model. Therefore we get the Poisson regression model as follows

$$\ln(\hat{\mu}) = -8.376 + 0.095X_1 + 0.004X_2 + 0.044X_3 + 0.000\,X_4 \tag{4}$$

The residual deviance value obtained divided by the degrees of freedom from the parameter estimation results is $\frac{\mathbf{107448}}{\mathbf{55}} = \mathbf{1953.6} > \mathbf{1}$, which means the model experiences overdispersion. There is an assumption in Poisson regression, namely that the equidispersion condition must be met. Then another regression model will be formed that is more flexible in conditions such as overdispersion, namely the Negative Binomial regression model.

C. *Determination of a Negative Binomial Regression Model*

In the Negative Binomial regression model, bicycle sales data is assumed to have a Negative Binomial distribution. The results of parameter estimation of the Negative Binomial regression model are presented in the Table 3.

Table 3 Parameter Estimation of Negative Binomial Regression Model

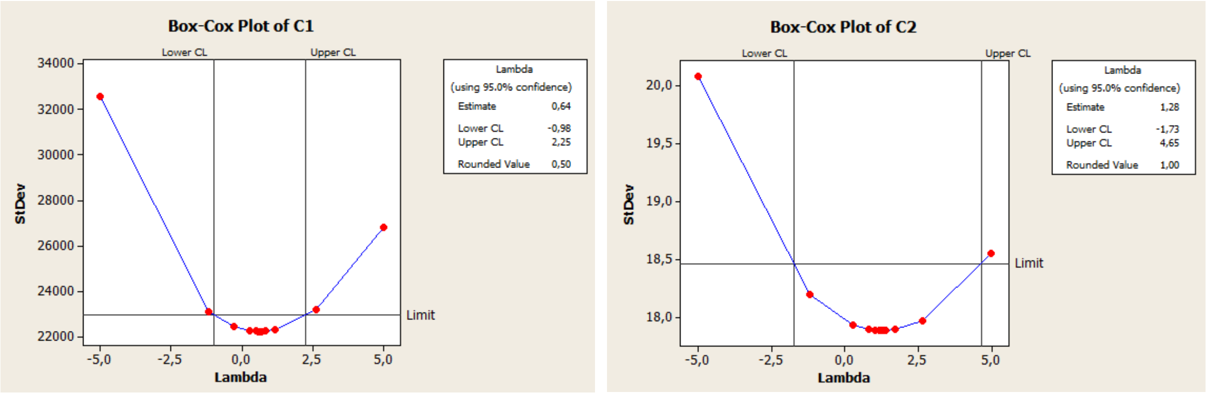
Parameter	Estimation Value	SE	Z_{hit}	P_{value}
β_0	-8.0082	2.074	-3.861	0.000
β_1	0.0947	0.011	9.015	0.000
β_2	-0.0040	0.080	-0.050	0.960
β_3	0.0423	0.026	1.627	0.104
β_4	0.0002	0.000	4.528	0.000

Based on the Table 3, it can be seen that each parameter has a value of $|Z_{hit}| > Z_{\alpha/2} = 1.96$, except β_2. It can be said that the parameters $\beta_0, \beta_1, \beta_3,$ and β_4 have a partially significant effect on the model. Then we get the Negative Binomial regression model as follows

$$\ln(\hat{\mu}) = 8.008 + 0.095\ X_1 - 0.004\,X_2 + 0.042\,X_3 + 0.0002\,X_4 \tag{5}$$

D. *Determination of a GSARIMA Model*

The determination of the GSARIMA(p,d,q)(P,D,Q)^s model is obtained from identifying the best SARIMA(p,d,q)(P,D,Q)^s model. The first step is to analyze the stationarity of the data regarding the variance and mean through the Box-Cox plot and ACF plot. Box-Cox Plot and Box-Cox transformation results are presented in Figure 1

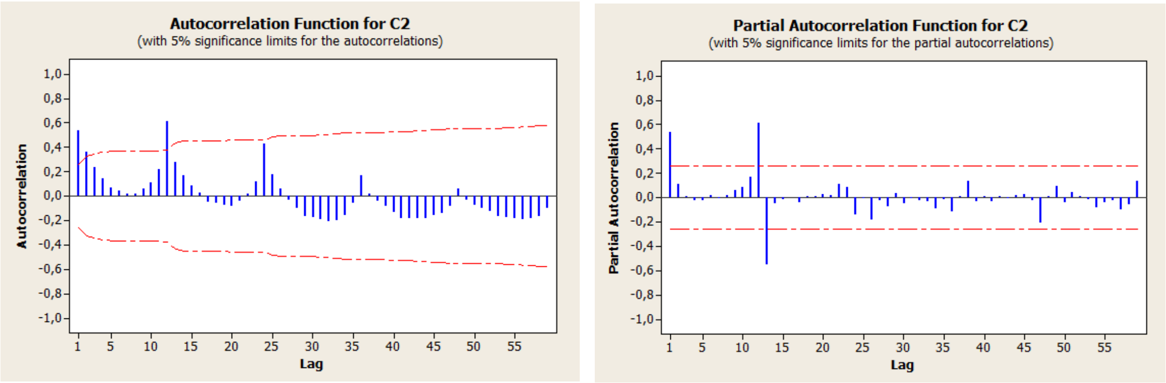


(A.)

(B.)

Figure 1 Box-Cox Plot and Box-Cox Transformation Results of Bicycle Sales Data (A: Box-Cox Plot, B: Box-Cox Transformation)

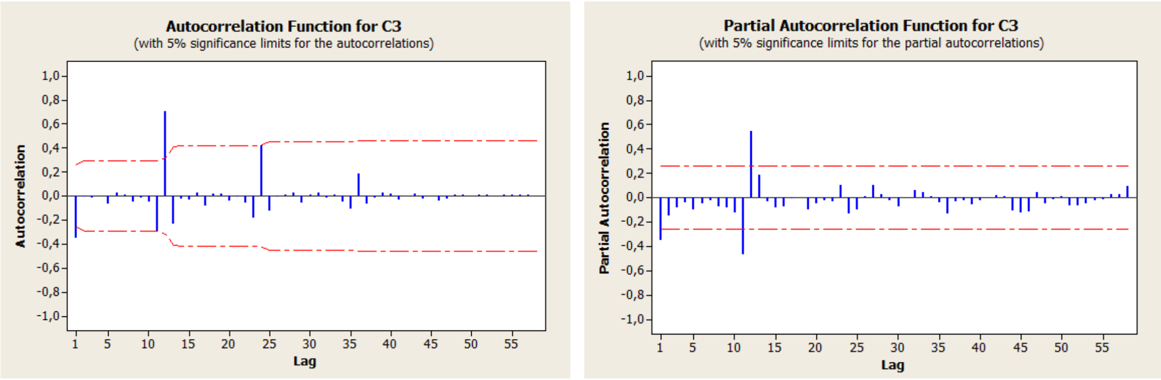
Then the average stationarity analysis is carried out through the ACF plot. Based on Figure 2, it can be seen that the data is not stationary for the mean, then a differencing is made between non-seasonal lag and seasonal lag. The results of differencing between non-seasonal lag and seasonal lag are presented in Figure 3 and Figure 4.



(A.)

(B.)

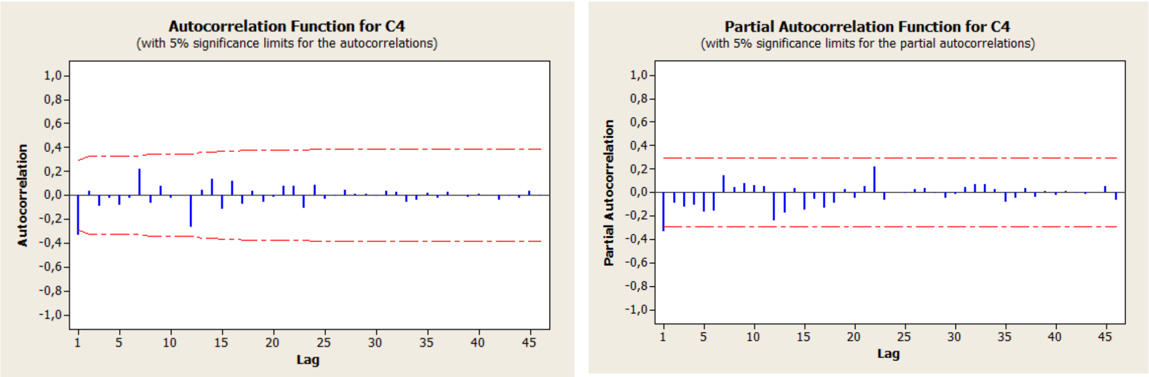
Figure 2 ACF dan PACF Plot of Box-Cox Transformation Results (A: ACF Plot, B: PACF Plot)



(A.)

(B.)

Figure 3 ACF dan PACF Plot of Non-seasonal Lag Differencing Result (A: ACF Plot, B: PACF Plot)



(A.) (B.)

Figure 4 ACF dan PACF Plot of Seasonal Lag Differencing Result (A: ACF Plot, B: PACF Plot)

Based on Figure 4, the estimated temporary model obtained is the SARIMA(1,1,1)(0,1,0)^{12} model. The results of parameter estimation and significance testing with the t-test and $\alpha = 5\%$ are presented in Table 4**.**

Table 4 Estimation Results and Significance Test of SARIMA(1,1,1)(0,1,0)^{12} Model Parameters

Parameter	Coefficient	t_{test}	t_{tabel}	Result
AR 1	0.502	3.18	1,996	Significance
MA 1	0.948	9.19	9.015	Significance

Then a white noise and normal residual test was carried out. It was found that the SARIMA(1,1,1)(0,1,0)^{12} model fulfilled the white noise and normal residual assumptions. Therefore the SARIMA model formed is SARIMA(1,1,1)(0,1,0)^{12}, with $p = 0,\ P = 0,\ q = 1, Q = 1,\ d = 1,\ D = 1,\ S = 12$. So the GSARIMA(1,1,1)(0,1,0)^{12} model with ZQ1 transformation is

$$\ln(\mu_t) = \beta_0 + \ln(y'_{t-1}) + \ln(y'_{t-12}) - \ln(y'_{t-13}) + \phi_1 \ln(y'_{t-1}) - \phi_1 \ln(y'_{t-2}) - \phi_1 \ln(y'_{t-13}) + \phi_1 \ln\,(y'_{t-14}) + \theta_1 \ln\left(\frac{y'_{t-1}}{\mu_{t-1}}\right) \tag{6}$$

or it can also be written as follows

$$\mu_t = exp\left\{\beta_0 + \ln(y'_{t-1}) + \ln(y'_{t-12}) - \ln(y'_{t-13}) + \phi_1 \ln(y'_{t-1}) - \phi_1 \ln(y'_{t-2}) - \phi_1 \ln(y'_{t-13}) + \phi_1 \ln\,(y'_{t-14}) + \theta_1 \ln\left(\frac{y'_{t-1}}{\mu_{t-1}}\right)\right\} \tag{7}$$

the GSARIMA(0,1,1)(0,1,1)^{12} ZQ2 transformation model is as follows

$$\ln(\mu_t) = \beta_0 + \ln(y_{t-1} + c) + \ln(y_{t-12} + c) - \ln(y_{t-13} + c) + \phi_1 \ln(y_{t-1} + c) - \phi_1 \ln(y_{t-2} + c) - \phi_1 \ln(y_{t-13} + c) + \phi_1 \ln\,(y_{t-14} + c) + \theta_1 \ln\left(\frac{y_{t-1} + c}{\mu_{t-1} + c}\right) \tag{8}$$

or it can also be written as follows

$$\mu_t = exp\left\{\beta_0 + \ln(y_{t-1} + c) + \ln(y_{t-12} + c) - \ln(y_{t-13} + c) + \phi_1 \ln(y_{t-1} + c) - \phi_1 \ln(y_{t-2} + c) - \phi_1 \ln(y_{t-13} + c) + \phi_1 \ln\,(y_{t-14} + c) + \theta_1 \ln\left(\frac{y_{t-1} + c}{\mu_{t-1} + c}\right)\right\} \tag{9}$$

The estimated parameter values for the Negative Binomial GSARIMA model and the Poisson GSARIMA Model were obtained using Bayesian inference with the MCMC method using R software. The parameter estimation results for the GSARIMA model are presented in Table 5.

Table 5 Parameter Estimation Results of the Negative Binomial GSARIMA Model and the Poisson GSARIMA Model

Model	Parameter	Mean	2,5%	97,5%
Poisson GSARIMA(0,1,1)(0,1,1)^{12} Transformation ZQ1	β_0	0.00003	0.00000	0.00010
	AR1(ϕ_1)	0.00003	0.00000	0.00014
	MA1(θ_1)	0.00006	0.00000	0.00020
Poisson GSARIMA(0,1,1)(0,1,1)^{12} Transformation ZQ2	β_0	0.00002	0.00000	0.00010
	AR1(ϕ_1)	0.00003	0.00000	0.00011
	MA1(θ_1)	0.00005	0.00000	0.00015
Binomial Negatif GSARIMA(0,1,1)(0,1,1)^{12} Transformation ZQ1	β_0	0.00807	0.00061	0.02067
	AR1(ϕ_1)	0.02374	0.00003	0.08300
	MA1(θ_1)	0.04697	0.00216	0.14660
Binomial Negatif GSARIMA(0,1,1)(0,1,1)^{12} Transformation ZQ2	β_0	0.00767	0.00033	0.02061
	AR1(ϕ_1)	0.02164	0.00002	0.10600
	MA1(θ_1)	0.05081	0.00159	0.17240

E. *Analysis of Forecasting Results and Selection of The Best Model*

The best model is selected based on the smallest AIC value presented in the Table 6.

Table 6 The Comparison of AIC Values

Model	AIC
Poisson	108340.1
Negative Binomial Regresion	1406.1
Poisson GSARIMA ZQ1	54396.1
Poisson GSARIMA ZQ2	54395.0
Negative Binomial GSARIMA ZQ1	1058.7
Negative Binomial GSARIMA ZQ2	1059.5

Based on Table 6, the Negative Binomial GSARIMA ZQ1 model has the smallest AIC value = 1058.7, which can be used as the best model for forecasting the number of bicycle sales. The results of comparing forecasting plots with outsample data using the best model are presented in Figure 5.

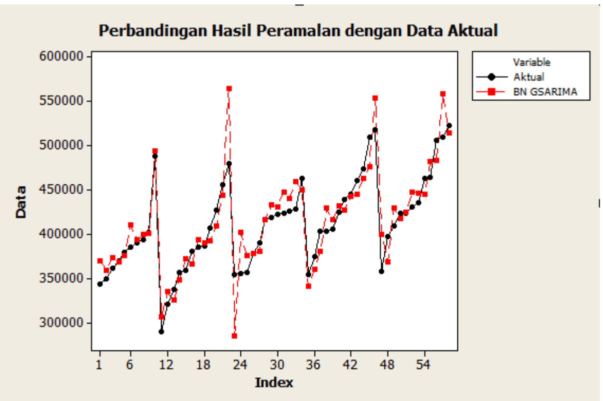


Figure 5 Comparison of Forecasting Results with the Negative Binomial GSARIMA Model with Actual Data

IV. CONCLUSIONS AND SUGGESTIONS

The conclusion obtained in this research is that the Negative Binomial Regression model applied to forecasting the number of bicycle sales has the lowest AIC value. It can be said that this model has the smallest risk than another model. In this study, the estimation method used is Bayesian inference, where this method allows for analysis of the posterior distribution of predictions. This shows that the posterior predictive distribution is much better for the GSARIMA ZQ1 Negative Binomial model. Hence this model can be used as the best model for forecasting the number of bicycle sales in the future period. Therefore, using this forecasting model can minimize errors in the forecasting results. In further research, other estimation methods can be applied to forecasting models for count and seasonal data as a material for comparison and decision-making.

REFERENCES

[1] Badan Pusat Statistika. *Analisis Big Data di Tengah Masa Adaptasi Kebiasaan Baru*. Jakarta: Badan Pusat Statistika. 2020

[2] Briet, J. T. O., Amerasinghe, H. P., dan Vounatsou, P. *Generalized Seasonal Autoregressive Integrated Moving Average Models for Count Data with Application to Malaria Time Series with Low Case Numbers. Malaria Journal*, PLoS ONE, 8(6): e65761. 2013

[3] McCullagh, P. & Nelder, J. A. *Generalized Linear Models.* London: Chapman and Hall. 1989.

[4] Cameron, A.C dan Trivedi,P.K. *Regression Analysis of Count Data*. Cambridge University Press. USA. 1998.

[5] Benjamin, M. A., Rigby, R. A. & Stasinopoulos, D. M. Generalized Autoregressive Moving Average Models. *Journal of the American Statistical Association*, Vol 98, No 461, pp. 214-216. 2003.

[6] June MA & Konstadinos G. Applications of Poisson Regression Models to Actvity Frequency Analysis and Prediction. *Transportation Research Record 1676*. Paper No. 99-0813. 1999

[7] Funda h. *et al*. Poisson Regresyon Modelinde asiri Yayilim Durumu Ve Negatif Binomial Regresyon Analizinin Turkiye Grev Sayilari Uzerine Bir Uygulamasi. *Yonetim, Yil15, Sayi 48, Haziran. 17-25.* 2004

[8] Haibin Liu, *et al*. Negative Binomial Regression of Electric Power Outages in Hurricanes. *Journal Infrastructur System.* 11(4): 258-267. 2005.

[9] Asrirawan. Simulasi Perbandingan Metode Peramalan Model Generalized Seasonal Autoregressive Integrated Moving Average (GSARIMA) dengan Seasonal Autregressive Integrated Moving Average (SARIMA). *Jurnal Dinamika*, 06(1), pp. 61-66. 2015

[10] M. Aqil. Etimasi Parameter pada Model Poisson *Generalized Autoregressive Moving Average* (GARMA) dengan Algoritma IRLS Studi Kasus: Peramalan Jumlah Kecelakaan di Jalan Tol Surabaya-Gempol. *Jurnal Sains dan Seni ITS*, Vol. 7, No. 2. 2019

[11] Agil Desti. Etimasi Parameter pada Model Poisson *Generalized Autoregressive Moving Average* (GARMA) dengan Algoritma IRLS Studi Kasus: Peramalan Jumlah Kecelakaan di Jalan Tol Surabaya-Gempol. *Jurnal Sains dan Seni ITS*, Vol. 8, No. 1. 2019

[12] L P Wardhani *et al*. The Theft Criminality Forecasting In The Surabaya District Police Region Using Poisson GARMA Model and Negative Binomial GARMA Model. *Journal of Physics*, Conf. Ser. 1490 012015. 2020.

[13] T. Kim *et al*. Prediction Regions for Poisson and Over-Dispersed Poisson Regression Models with Applications to Forecasting Number of Deaths during the COVID-19 Pandemic. 2020

[14] Famoye, F., Wulu, J. T., dan Singh, K. P. "On The Generalized Poisson Regression Model with an Application to Accident Data", *Journal of Data Science*, No. 2, pp. 287-295. 2004.

[15] Hilbe, J. *Negative Binomial Regression, Second Edition*. New York: Cambridge University Press. 2011.

DOI: 10.12962/j27213862.v7i2.20255