Research Article

Journal of Applied Science and Emerging Technology (JASET)
Published by Faculty of Applied Science, KMUTNB, Thailand
doi: 10.14416/JASET.KMUTNB.2023.02.004
Vol. 22, No. 2 [2023]: e250803

# A Comparison of Models for Count Data with an Application to Over-Dispersion Data

Chadarat Tapan[1], Anamai Na-udom[2]* and Jaratsri Rungrattanaubol[3]

[1, 2] Department of Mathematics, Faculty of Science, Naresuan University, Thailand

[3] Department of Computer Science and Information Technology, Faculty of Science, Naresuan University , Thailand

## ABSTRACT

Count models have been widely used in various fields, such as medicine, biology, and public health. The most frequently used count models are Poisson regression, negative binomial regression, and discrete Weibull regression models. The objective of this study was to compare the performance of Poisson, negative binomial, and discrete Weibull regression models using two different sets of data with over-dispersion. The AIC, BIC, and log-likelihood fit statistics were used as the criteria to compare the count models. The results revealed that the negative binomial and discrete Weibull regression were the best fit models as they produced the smallest AIC, BIC, and log-likelihood fit statistics.

**KEYWORDS:** Count models, Over-dispersion data, Poisson regression, Negative regression, Discrete Weibull regression

***Corresponding Author: anamain@nu.ac.th**

**Received:** 15/11/2022; **Revised:** 05/07/2023; **Accepted:** 14/07/2023

## 1. INTRODUCTION

Count data is a statistical data type which describes the frequency of events or items that are occurred within a fixed period of time. Count data can be found in a variety of fields, such as medicine, biological sciences, epidemiology, and public health. For example, the number of heart attacks, the number of students absent during a period of study, the number of men infected with human papillomavirus, the frequency of traffic accidents, the number of cigarettes smoked, and the number of people infected with COVID-19 on a daily basis (Hilbe J. M., 2014; J.-H Lee et al., 2012; Klakattawi H. et al., 2018). In order to extract a crutial information from the count data, a suitable approach such count data model is required for data analysis.

A regression model is the most frequently used as an analytical model. Its objective is to investigate the relationship between a response variable and predictor variables. There are two types of regression analysis: linear regression and non-linear regression. Linear regression is the most popular type of regression analysis in which the line that best fits to the data according to a specified mathematical criterion is found (Kung-Yee Liang & Scott L. Zeger, 1993). When the response variable is count data, the classical regression model may not feasible to use. Thus, Poisson regression is the most popular model for modelling count data. The Poisson regression describes the relationship between predictor variables and the response variable. Poisson distribution, with the conditional mean of occurrence equal to the variance of the response, which is called equi-dispersion (Saputo D. et al., 2021; Wan Tang et al., 2012).

In most cases, the variance of the response variable is greater than its mean, which is known as over-dispersion. The extent of over-dispersion may be determined simply by comparing the sample mean with the variance of the response variable (Cameron A. C. & Trivedi P. K., 2013). In some situation, the over-dispersioned data are caused by the fact that the data contains outliers. A negative binomial regression model or a gamma distribution mixture model is commonly considered the default choice for over-dispersed count data since the variance of the response variable exceeds the mean (Hilbe J. M., 2014). Moreover, in the case of large sample sizes, the discrete Weibull regression model is an attractive alternative to the negative binomial regression model for over-dispersed count data (Klakattawi H. et al., 2018).

There have been a variety of researches conducted to deal with over-dispersion data for count data models. For instance, Ver Hoef and Boveng (2007) proposed a model for over-dispersed count data using quasi-Poisson and negative binomial regression. The data collection was based on aerial surveys of harbor seals. These counts were affected by date, time of day, and time relative to low tide. They provided results from a data set that showed a dramatic increase in harbor seal abundance when using quasi-Poisson versus negative binomial regression. Linden A. and Mantyniemi S. (2011) used negative binomial regression to investigate model over-dispersion in bird migration data sets. According to the findings, the negative binomial regression model could be a reasonable approximation for modeling marginal distributions of independent count data. J.-H. Lee, G. Han, W. J. Fulp, and A. R. Giuliano (2012) demonstrated Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial models for over-dispersion count data. The data set contained the number of incidents involving human papillomavirus infection. The four models produced different statistical findings. Klakattawi H., Vinciotti V., and Yu K. (2018) presented a count data model based on a discrete Weibull regression and compared it to Poisson and negative binomial regression based on over-dispersion data. The result found that the discrete Weibull regression could be applied to over-dispersion data better than other models. Yinglin Xia et al. (2012) analyzed data from HIV preventive intervention studies and compared four popular statistical models: Poisson, negative binomial, zero-inflated Poisson, and zero-inflated negative binomial. They found that a zero-inflated negative binomial was better than other models under the likelihood ratio test, AIC, and BIC criteria.

Melliana A. et al. (2013) used data on the number of cervical cancer cases to compare generalized Poisson and negative binomial regression, which was an indicator of over-dispersion. The negative binomial regression model was the best model since it had the lowest AIC value. Avei E., Alturk S., and Soylu E. (2015) compared over-dispersed algal data count data models. The regression models of Poisson, quasi-Poisson, negative binomial, and COM-Poisson were considered. Because the log likelihood and AIC were the minimum, COM-Poisson regression was the best-fit model. Jasin M., Hussein M., and Hamodi H. (2017) compared models for the number of patients infected with pneumonia. The Geometric, hurdle, and zero inflated-Geometric regressions were compared. The results of log likelihood and AIC indicated the zero inflated-Geometric regression was the best fit. Alebachew A. (2019) collected data from lecturers' publications between November 2015 and 2016 and compared the performances of count data models such as Poisson, negative binomial, zero-inflated negative binomial (ZINB), zero-inflated Poisson, and Poisson hurdle. The ZINB regression model was chosen as the most appropriate and efficient regression model based on AIC value. Durmus B. and Guneri O. I. (2020) studied the generalized Poisson and Poisson regression models for over-dispersion data on the number of strikes between 1984 and 2017. The result found that the generalized Poisson was the best model.

In this paper, we compared count data models using simulation study and two different over-dispersion real data sets. Three models namely Poisson, negative binomial, and discrete Weibull regression were compared in this paper. The main difference between this paper and the other papers published so far is that we aimed to compare count data models for a small sample size using both of real data sets and simulation studies. In the next section, we present materials and methods which cover the details of models used in this study. The parameter estimation method and the details of data sets are also presented. The results and conclusions will be presented in the last section.

## 2. MATERIALS AND METHODS

In this section, we introduce models for count data namely Poisson, negative binomial, and discrete Weibull regression models, as well as methods of parameter estimation, criteria used to validate the model performance and simulation study, respectively. Further the details of data sets used in the study are also presented.

### *2.1 Poisson Regression*

Poisson regression is a popular and fundamental model for modeling count data. In some situations, the response variable represents a count data of some rare event or a count of particulate matter. Hilbe J. M. (2014) introduced the Poisson model where the response variable is a count number, or the response terms must be nonnegative integers. The observations are independent of one another, while the mean and variance of the model are identical. The discrete random variable  is Poisson distribution with parameter $\mu$ ; $\mu > 0$. The probability mass function (pmf) is

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

where mean and variance are $E(Y) = \mu$ and $Var(Y) = \mu$, respectively (Nakagawa T. & Osaki S., 1975).

The maximum likelihood method was used to estimate the parameters in a Poisson distribution. Given a random sample  from a Poisson distribution, the log-likelihood function (Cameron A. C. & Trivedi P. K., 2013) can be written as

$$\ln L(y_i;\mu) = -n\mu - \sum_{i=1}^{n}\ln(y_i!) + \ln\mu\sum_{i=1}^{n}(y_i) \tag{2}$$

Maximum likelihood estimator of $\mu$ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{3}$$

The log-link function is used with the Poisson distribution and can be written as

$$\mu_i = e^{X_i'\beta}; \quad i = 1,2,3,\ldots,n \tag{4}$$

where $X_i' = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ik})$ is a vector of covariates and $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_k)'$ is an unknown vector of regression coefficients (Montgomery et al., 2012).

The Poisson regression is formulated as

$$Y_i = \exp(X_i'\beta) \tag{5}$$

where $Y_i$ is a vector of response variable in the form of  size of vector (Saputo D. et al., 2021).

Maximum likelihood estimation method has been widely used to estimate the parameters in Poisson regression. From Equation (2), log likelihood function can be expressed as follows

$$\ln L(y_i;\beta) = \sum_{i=1}^{n}\left\{y_i X_i'\beta - e^{(X_i'\beta)} - \ln(y_i!)\right\} \tag{6}$$

### *2.2 Negative Binomial Regression*

From the previous section, the assumption of the Poisson regression model was the equality of mean and variance. In most situation, the variance will exceed the mean, $Var(Y) > E(Y)$, and the distribution allows for over-dispersion. Hence, the negative binomial regression would be suggested (Ismail N. & Jemain A., 2007). Hilbe J. M. (2014) summarized the negative binomial regression model's assumptions as follows: the response variable is a count of nonnegative integers.

The negative binomial distribution or the Poisson-gamma mixture distribution (Cameron A. C. & Trivedi P. K., 2013) has the probability distribution function as

$$f(y;\mu,\alpha) = \begin{pmatrix} y_i + \dfrac{1}{\alpha} - 1 \\ \dfrac{1}{\alpha} - 1 \end{pmatrix} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{7}$$

or

$$f(y;\mu,\alpha) = \frac{\Gamma(y+\alpha^{-1})}{\Gamma(y+1)\Gamma(\alpha^{-1})}\left(\frac{\alpha^{-1}}{\alpha^{-1}+\mu}\right)^{\alpha^{-1}}\left(\frac{\mu_i}{\alpha^{-1}+\mu}\right)^{y}$$

where $\alpha \geq 0,\ y = 0,1,2,\ldots$. The function $\Gamma(\cdot)$ is the gamma function. The mean and variance of $Y$ are $E(Y) = \mu$ and $Var(Y) = \mu + \mu^2\alpha = \mu(1+\alpha\mu)$.

In order to fit the negative binomial distribution, the parameter must be estimated by using maximum likelihood method. For a given random sample $y_1, y_2, y_3, \ldots, y_n$ from the negative binomial distribution, the log-likelihood function can be presented in $\mu$ formats as follows

$$L(y;\mu,\alpha) = \sum_{i=1}^{n} y_i \ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}\ln(1+\alpha\mu_i) + \ln\Gamma\left(y_i + \frac{1}{\alpha}\right)$$
$$-\ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right). \qquad (8)$$

The negative binomial regression model used the log-link function (Hilbe J. M., 2014) given by

$$\ln(\mu_i) = \ln\left(E\left(Y \middle| \alpha,\beta\right)\right) = X_i'\beta; \quad 1 \leq i \leq n \qquad (9)$$
$$\text{or} \qquad \mu_i = \exp(X_i'\beta)$$

when $\beta$ is unknown vector of regression coefficients.

Maximum likelihood estimation was used to estimate the parameters of negative binomial regression. The log-likelihood function is given by

$$L(y;\beta,\alpha) = \sum_{i=1}^{n} y_i \ln\left(\frac{\alpha\exp(X_i'\beta)}{1+\alpha\exp(X_i'\beta)}\right) - \frac{1}{\alpha}\ln\left(1+\alpha\exp(X_i'\beta)\right)$$
$$+\ln\Gamma\left(y_i + \frac{1}{\alpha}\right) - \ln\Gamma(y_i+1) - \ln\Gamma\left(\frac{1}{\alpha}\right) \qquad (10)$$

### *2.3 Discrete Weibull (DW) Regression*

Klakattawi et. al. (2018) introduced the discrete Weibull regression model for count data for 3 types of dispersions: over-dispersion, under-dispersion, and covariate-specific dispersion. Most of the research applied the NB regression model to the over-dispersion count data. For this reason, in this study, we compare the DW regression models with the NB regression models. In addition, the DW regression model was compared with the Poisson model to obtain clear study results.

If a random variable $Y$ follows the discrete Weibull distribution (type I) (Nakagawa T. & Osaki S., 1975), then the pmf is given by

$$f(y;q,\alpha) = q^{y^\alpha} - q^{(y+1)^\alpha}; \quad y = 0,1,2,\ldots, \qquad (11)$$

where the parameter $0 < q < 1$ and $\alpha > 0$.

The mean and variance for DW distribution are given by, respectively,

$$E(Y) = \mu = \sum_{y=1}^{\infty} q^{y^\alpha}$$

and

$$Var(Y) = \sigma^2 = 2\sum_{y=1}^{\infty} yq^{y^\alpha} - \mu - \mu^2.$$

We can estimate the parameter for a DW distribution by using the maximum likelihood method. Given that $y_1, y_2, y_3, \ldots, y_n$ are a random sample from a DW distribution, the log-likelihood is given by

$$L(y;q,\alpha) = \sum_{i=1}^{n} \log\left(q^{y_i^\alpha} - q^{(y_i+1)^\alpha}\right). \qquad (12)$$

Klakattawi et.al. (2018) introduced the model of the relationship between a count response variable and a set of covariates when the response $Y_i$ has a DW conditional distribution $f\left(y_i, q(x_i), \alpha \middle| x_i\right)$; $q(x_i)$ is the DW parameter related to the explanatory variables $X_i$ through the link function:

$$\log(-\log(q_i)) = x_i'\beta, \quad x_i'\beta = \beta_0 + x_{i1}\beta_1 + \ldots + x_{ik}\beta_k. \qquad (13)$$

According to Equation (13), $q_i$ can be illustrated as

$$q_i = e^{-e^{x_i'\beta}}, \qquad (14)$$

from which the conditional probability mass function of the response variable $Y_i$ given $X_i$ is as below.

$$f\left(y_i \middle| x_i\right) = \left(e^{-e^{x_i'\beta}}\right)^{y_i^\alpha} - \left(e^{-e^{x_i'\beta}}\right)^{(y_i+1)^\alpha} \qquad (15)$$

Parameter estimation for unknow parameter $\beta$ and $\alpha$ used method of maximum likelihood estimation. The log-likelihood is given by

$$L(y;\beta,\alpha) = \sum_{i=1}^{n} \log\left[\left(e^{-e^{X_i'\beta}}\right)^{y_i^\alpha} - \left(e^{-e^{X_i'\beta}}\right)^{(y_i+1)^\beta}\right]. \qquad (16)$$

### 2.4 Method of Parameter Estimation

The maximum likelihood estimation (MLE) method is the most widely used to estimate the unknown parameters of an assummed probability distribution. For a given observed data, MLE attempts to find the parameters with the highest joint probability by using an optimization method. In this case MLE method aims to find the maximum of the likelihood function $L(\theta; y)$. If the likelihood function is differentiable, the derivative methods can be directly applied to dind the maxima. In most circumstances, numerical methods are necessary to find the maximum of the likelihood function. The computational methods used when the estimate of MLE is not in a closed explicit form are the Newton-Raphson iterative method, the method of scoring, the simplex method, and the EM algorithm (Cameron A. C. & Trivedi P. K., 2013; Garthwaite et al., 2002). In practice, it has been observed that maximization of log-likelihood function $\ln\left[L(\theta; y)\right]$ is much easier than direct maximization of $L(\theta; y)$. In this paper, we presented the log-likelihood function for estimating the parameters of count data models. We used the package in the R program to estimate the parameters or coefficients for count data models. The analysis was carried out in the R program by using the glm () function, glm. nb () function, and the package 'DWreg', respectively.

### 2.5 Criteria

This section presents the criteria for comparing count data models. For fitted models, most papers used the Akaike information criterion (AIC), Bayesian information criterion (BIC), and log-likelihood (LL). In 1973, Hirotsugu Akaike proposed a model selection criterion based on the fitted log-likelihood function. The AIC is the most commonly used fit statistic. Let $L$ be the model likelihood, $p$ is the number of parameters (predictors) in the model. The AIC is $AIC = -2\ln(L) + 2p$ (Cameron A. C. & Trivedi P. K., 2013). In 1978, Gideon schwarz proposed the modification to AIC include the Bayesian information criterion. The formulate of BIC is $BIC = -2\ln(L) + p\ln(n)$ with $p$ is number of parameter and $n$ the number of observations in the model (Hilbe J. M., 2014). When considering the maximum likelihood method, the log-likelihood ($LL$) test can be used for model comparisons. The $LL$ test can be used to determine whether or not there is over-dispersion. Probability ratio statistics is calculated as; $LL = 2(\ln L_1 - \ln L_0)$. Where $L_1$ and $L_0$ are the log-likelihood under the respective hypothesis. When the AIC and BIC values are the least, or the $LL$ value is the greatest, it is possible to conclude that a model is excellent (Montgomery et al., 2012).

### 2.6 Simulation studies

In order to implement the count data models, we simulate the data set and then apply the three count data models with the simulated data set. The details of conducting simulation studies are given as follows:

We set the simple regression model using equation (17)

$$\log \mu_i = \beta_0 + \beta_1 x_i, \quad i = 1, 2, 3, \ldots, n \qquad (17)$$

We set up different sample sizes ($n$) as $n$=20, 50, 100, 500. All the results are based on the average of 1,000 repetitions through the following steps:

1) Simulation of random samples of sizes $n$ to present the predictor variables from the normal distribution, $x \sim N(0,1)$.

2) Set up the parameters $\beta_0 = 1.5, \beta_1 = 1$.

3) Simulation of data from a Poisson inverse Gaussian distribution as the true model with a sample size $n$, $\mu = 1.5$ and $\sigma = 2$.

4) Estimation of parameters by MLEs method.

5) Repeat steps 1-4 for 1,000 times and then compute the average MLEs, bias, and MSEs.

### 2.7 Data Sets

In this study, we also consider two over-dispersion data sets as an application of the count data models. These two data sets have been selected as they are similar in terms of sample sizes and the same over-dispersion. The two data sets consist of the number of deaths caused by accidents per day and the dataset in the Ecdat R package. The details of each data set are given below.

**Table 1** The number of deaths caused in accidents per day from January to June of 2021.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 1 | 3 | 8 | 9 | 15 | 25 | 9 | 23 | 15 | 10 | 10 |
| Count | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | ≥ 21 |
| Frequency | 7 | 2 | 5 | 7 | 4 | 3 | 3 | 1 | 2 | 1 | 1 |

**Table 2** The number of contracts strikes in US manufacturing observed monthly from January 1968 to December 1976.

| Count | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 13 | 15 | 16 | 18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Frequency | 5 | 12 | 14 | 11 | 9 | 14 | 9 | 4 | 7 | 10 | 6 | 1 | 3 | 1 | 1 | 1 |

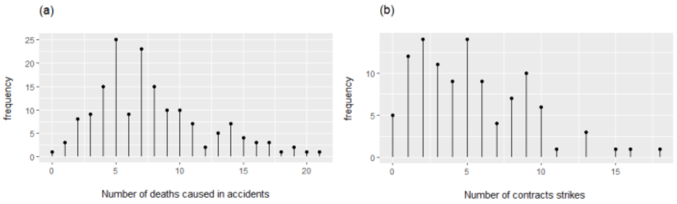

**Figure 1** (a) The distribution of the number of caused in accidents per day and (b) the distribution of the number of contracts strikes in US manufacturing.

**Table 3** The coefficients of the Poisson, NB, and DW regression models and model comparison based on the simulation studies.

| $n$ | Model | parameter | Estimates Coefficient | AIC | BIC | LL |
|---|---|---|---|---|---|---|
| 20 | Poisson | $\beta_0$ | 0.64442 | 78.9700 | 80.9615 | -37.4850 |
| | | $\beta_1$ | -0.04653 | | | |
| | NB | $\beta_0$ | 0.64457 | 78.2695 | 81.2566 | -36.1347 |
| | | $\beta_1$ | -0.04934 | | | |
| | DW | $\beta_0$ | -1.3582 | 78.1311 | 81.1183 | -36.0656 |
| | | $\beta_1$ | 0.0854 | | | |
| 50 | Poisson | $\beta_0$ | 0.2071 | 171.3467 | 175.1707 | -83.67334 |
| | | $\beta_1$ | -0.2931 | | | |
| | NB | $\beta_0$ | 0.2145 | 163.3170 | 169.053 | -78.65849 |
| | | $\beta_1$ | -0.2678 | | | |
| | DW | $\beta_0$ | -0.6752 | 163.3486 | 169.0846 | -78.67428 |
| | | $\beta_1$ | 0.2217 | | | |
| 100 | Poisson | $\beta_0$ | 0.2095 | 305.2494 | 310.4598 | -150.6247 |
| | | $\beta_1$ | 0.0362 | | | |
| | NB | $\beta_0$ | 0.20959 | 302.5666 | 310.3822 | -148.2833 |
| | | $\beta_1$ | 0.03696 | | | |
| | DW | $\beta_0$ | -0.87466 | 302.3092 | 310.1247 | -148.1546 |
| | | $\beta_1$ | -0.06101 | | | |
| 500 | Poisson | $\beta_0$ | 0.33585 | 1820.7274 | 1829.1566 | -908.3637 |
| | | $\beta_1$ | 0.03624 | | | |
| | NB | $\beta_0$ | 0.3358 | 1634.8996 | 1647.5434 | -814.4498 |
| | | $\beta_1$ | 0.0399 | | | |
| | DW | $\beta_0$ | -0.6565 | 1635.0194 | 1647.6632 | -814.5097 |
| | | $\beta_1$ | -0.0340 | | | |

**Table 4** The coefficients of the Poisson, NB, and DW regression models and model comparison of the data set 1.

| Coefficients/Criterion | Poisson | | NB | | DW | |
|---|---|---|---|---|---|---|
| | Estimates Coefficient | Standard Error | Estimates Coefficient | Standard Error | Estimates Coefficient | Standard Error |
| Intercept | 1.52032 | 0.0433 | 1.51040 | 0.0608 | -4.2382 | 0.2848 |
| Number of accidents | 0.00465 | 0.0002 | 0.00473 | 0.0004 | -0.0112 | 0.0013 |
| AIC | 1003.381 | | 969.162 | | 970.768 | |
| BIC | 1009.778 | | 972.384 | | 980.363 | |
| LL | -499.690 | | -481.581 | | -482.384 | |

**Table 5** The coefficients of the Poisson, NB, and DW regression models and model comparison of the data set 2.

| Coefficients/Criterion | Poisson | | NB | | DW | |
|---|---|---|---|---|---|---|
| | Estimates Coefficient | Standard Error | Estimates Coefficient | Standard Error | Estimates Coefficient | Standard Error |
| Intercept | 1.6539 | 0.0422 | 1.6538 | 0.0686 | -3.0704 | 0.2625 |
| Economic activity | 3.1342 | 0.8032 | 3.2250 | 1.2841 | -5.2937 | 1.8958 |
| AIC | 627.969 | | 566.597 | | 564.157 | |
| BIC | 633.333 | | 574.643 | | 572.203 | |
| LL | -311.985 | | -280.298 | | -279.079 | |

Data set 1: The number of deaths caused by accidents per day from January to June of 2021, which included 181 observations obtained from the website of government data in Thailand: https://datagov.mot.go.th/. The predictor variable is the number of accidents per day, and the response variable is the number of deaths caused by accidents per day. The sample mean and variance of the response variable are 8.3867 and 38.6829, respectively. This clearly indicates the existence of over-dispersion as the response variance is larger than the mean. The details of data set are shown in Table 1.

Data set 2: The StrikeNb dataset, which was retrived from Ecdat R package at http://CRAN.R-project.org/package=Ecdat. The response variable is the number of contracts strikes in US manufacturing, observed monthly from January 1968 to December 1976, which includes a total of 108 observations. The predictor is the level of economic activity, defined as the cyclical deviation of aggregate production from its trend level. The sample mean and variance of the response variable are 5.2407 and 14.0723. It can be clearly seen that the response variance is larger than mean, indicating the existence of over-dispersion. Table 2 displays data from the second data set.

## 3. RESULTS

The modeling results of the Poisson regression, negative binomial (NB) regression, and discrete Weibull (DW) regression obtained from the simulation study and two data sets are presented in this section. Comparison results based on the AIC, BIC, and LL fit statistics are presented in Tables 3-5, respectively.

Table 3 presents the comparison of the three models obtained from the simulation studies. The results show that the DW regression provides the smallest AIC value at sample sizes of 20 and 100. It is however, when the sample sizes are 50 and 500, the NB regression shows a minimal AIC value.

Table 4 presents the comparison for the 3 models obtained from the first data set, including Poisson, NB, and DW regression models. The results show that the NB regression provides the smallest AIC value with slightly lower than that of DW regression model. When we consider the BIC value, it indicates that NB regression model fits this data set well. Further, NB regression yields the highest value of LL criterion. This indicates that NB regression is the best fit for data set 1 with the coefficients estimated about 1.51040 and 0.00473, respectively. The regression equation for the number of deaths caused by traffic accidents can be expressed as $y = \exp(1.5104 + 0.00473x)$.

Table 5 compares the three models using the second data set. The results illustrate that the DW regression provides the smallest AIC and BIC values and the largest LL value. It can be clearly seen that the differences between the LL values of NB regression and DW regression are very small, hence it could be concluded that DW regression is the best model for data set 2. The coefficients estimated of the DW regression are approximately -3.0704 and -5.2937, respectively. The regression equation for the number of contract strikes is given as $y = \exp\left(-\exp\left(-3.0704 - 5.2937x\right)\right)$.

According to the comparative results among three regression models for overdispersion data using the simulation studies and two data sets as presented in Tables 3-5. We have observed that either NB regression or DW regression models are superior over Poisson regression model with respect to AIC, BIC, and LL criteria. Hence, both of NB regression and DW regression models are recommended to be used for overdispersion data.

## 4. CONCLUSION

This paper presents a comparative study of count data models for over-dispersion data. The count models included in this study are Poisson regression, NB regression, and DW regression, respectively. We study base on the simple regression models by the simulation studies. The data sets used are the number of deaths caused by traffic accidents per day and the strikeNb data set from the Ecdat R package. The criteria for comparing regression models are AIC, BIC, and LL values. The results show that the NB regression and DW regression are the best fit models for over-dispersion data under study. In order to extend the conclusion, more data sets or other models such as Poisson-weighted exponential regression could be further investigated.

## REFERENCES

Alebachew, A. (2019). A Comparison of count regression models on modeling of instructors publication factors: application of Ethiopian public universities. *American Journal of Theoretical and Applied Statistics*, *8*(5), 169-178.

Avcı, E., Altürk, S., & Soylu, E. N. (2015). Comparison count regression models for overdispersed alga data. *International Journal of Research and Reviews in Applied Sciences*, *25*(1), 1-5.

Cameron, A. C., & Trivedi, P. K. (2013). *Regression Analysis of Count Data*. Cambridge, New York: The United States of America.

Cupal, M., Deev, O., & Linnertova, D. (2015). The Poisson regression analysis for occurrence of floods. *Procedia Economics and Finance*, *23*, 1499-1502.

Davide C., Matthijs J. W., & Giuseppe, J. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *18*, 1-24. https://doi.org/10.7717/peerj-cs.623

Durmus, B., & Guneri, O. I. (2020). An application of the generalized Poisson model for over dispersion data on the number of strikes between 1984 and 2017. *Alphanumeric journal*, *8*(2), 249-260.

Emrah, A. (2019). A new model for over-dispersed count data: Poisson quasi-Lindley regression model. *Mathematical Science*, *13*, 241-247.

Garthwaite, P. H., Jolliffe, I. T., & Jones, B. (2002). *Statistical Inference*. Oxford University Press Inc., New York.

Gencturk Y., & Yigiter, A. (2016). Modelling claim number using a new mixture model: negative bionmial gamma distribution. *Journal of statistical computation and simulation*, *86*(10), 1829-1839.

Grine, R., & Zeghdoudi, H. (2017). On Poisson quasi-Lindley distribution and its application. *Journal of Modern Applied Statistical Methods*, *16*(2), 403-417.

Harris, T., Yang, Z., & Hardin, J. W. (2012). Modeling underdispered count data with generalized Poisson regression. *The Stata Journal*, *12*(4), 736-747.

Hilbe, J. M. (2014). *Modeling count data*. Cambridge University, www.cambridge.org/9781107611252

Husain, M., & Bagmar, S. H. (2015). Modeling under-dispersed count data using generalized Poisson regression approach. *Global Journal of Quantitive Science*, *2*(4), 22-29.

Ismail, N., & Jemain, A. (2007). Handling overdispersion with negative binomial and generalized poisson regression model. *Casualty Actuarial Society Forum*, 103-158.

Jasin, M., Hussein, M., & Hamodi, H. (2017). Comparison count regression models for the number of infected of Pneumonia. *Global Journal of Pure and Applied Mathematics*, *13*, 5359-5366.

Lee, J. H., Han, G., Fulp, W. J., & Giuliano, A. R. (2012). Analysis of overdispersed count data: application to the Human Papillomavirus Infection in Men (HIM) Study. *Epidemiology & Infection*, *140*(6), 1087-1094.

Klakattawi, H., Vinciotti, V., & Yu, K. (2018). A simple and adaptive dispersion regression model for count data. *Entropy*, *20*(142).

Liang, K. Y., & Zeger, S. L. (1993). Regression analysis for correlated data. *Annual review of public health*, *14*(1), 43-68.

Linden, A., & Mantyniemi, S. (2011). Using the negative binomial distribution to model overdispersion in ecological count data. *Ecology*, *92*(7), 1414-1421.

Loomis, D., Richardson, D. B., & Elliott, L. (2005). Poisson regression analysis of ungrouped data. *Occup Environ Med*, *62*, 325-329.

Melliana, A. (2013). The comparison of generalized Poisson regression and negative binomial regression method in overcoming overdispersion. *International Journal of Scientific & Technology Research*, *2*, 255-258.

Montgomery, D. C., Peck, E., & Vining, G. G. (2012). *Introduction linear regression analysis*. John Wiley & Sons, Inc, Hoboken, New Jersey.

Nakagawa, T., & Osaki, S. (1975). The discrete Weibull distribution. *IEEE Trans Reliab*, *24*, 300-301.

Saputo, D., Susanti, A., & Pratiwi, N. (2021). The handling of overdispersion on Poisson regression model with the generalized poisson regression model. *AIP Conference Proceedings 2326, 020026*(1-8).

Ver Hoef, J. M., & Boveng, P. L. (2007). Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?. *Ecology*, *88*(11), 2766-2772.

Wan, T., Hua, H., & Xin, M. T. (2012). *Applied Categorical and Count Data Analysis*. Taylor & Francis Group: London New York.

Xia, Y., Morrison-Beedy, D., Ma, J., Feng, C., Cross, W., & Tu, X. (2012). Modeling count outcomes from HIV risk reduction interventions: A comparison of competing statistical models for count responses. *AIDS research and treatment*, *2012*.